**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 models that can have different characteristics in quantitative evaluation, consequences of attacks, and defense strategy. To tackle this, we present 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 in image classification tasks, the performance degradation on image-to-image tasks can largely differ 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.

**1. 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 [9, 25].

Meanwhile, image-to-image models have also been developed. Unlike image classification models, image-to-image models cover much broader task objectives, including colorization [31], super-resolution [16], denoising [30], deblurring [20], and image-to-image translation [33]. 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 how 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.

For image classification models, an adversarial attack tries to enforce the input image to move out of the decision boundary (Figure 1a). Let \( X \) and \( \tilde{X} \) denote the original and attacked input images, respectively, i.e., \( \tilde{X} = X + \Delta \) for a small perturbation \( \Delta \). For a given classification model \( f_c(\cdot) \), which outputs a class label \( y \) for input \( X \), the goal of
the adversarial attack is to achieve \( f_c(\tilde{X}) \neq y \). 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 image \( f_m(X) \) and the attacked output image \( f_m(\tilde{X}) \), where \( f_m(\cdot) \) is the target image-to-image model (Figure 1b), i.e.,

\[
\max_{\Delta} d(f_m(X), f_m(X + \Delta)),
\]

where \( d(\cdot, \cdot) \) 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) cannot be used.

- **What are the characteristics of adversarial attacks for image-to-image models?** Because of the different properties from the image classification 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.

To find answers, we conduct comprehensive and thorough 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 influence and feasibility of defense methods that are conventionally applied to image classification models.

### 2. Related Work

**Adversarial Attack.** Researchers have recently developed adversarial attack methods that can fool various deep image classification models [25]. Szegedy et al. [26] developed an optimization-based approach to make a given classification model produce wrong classification results with minimum amounts of input perturbation. Goodfellow et al. [9] proposed the fast gradient sign method (FGSM), which calculates the perturbation from the sign of the gradients obtained for a given classification model. Kurakin et al. [14] iteratively employed FGSM for better fooling performance, called I-FGSM. The attack methods have also been extended to investigate other vulnerability issues, e.g., finding universal perturbation [19] and measuring transferability among different models [17]. More recently, attack methods applicable for tasks other than classification have been developed. For instance, [7] proposed the feature disruptive attack (FDA) method, which attempts to find perturbation from the intermediate features of a given model. These focus on the image classification task, and an in-depth study on vulnerability of image-to-image models for various tasks has not been conducted.

**Defenses against Adversarial Attack.** Defense methods against attacks 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 [6], bit depth reduction [29], and random resizing [28]. Another effective way is adversarial training, which uses images containing adversarial perturbation as training data [9, 12, 14].

### 3. Method

We consider 16 deep models for five popular image-to-image tasks: colorization (CIC [31]), deblurring (Deep-Deblur [20]), denoising (three DnCNN models [30] having different color channels and amount of noise \( \sigma \), super-resolution (EDSR [16], RCAN [32], CARN [2], SRResNet [15], and SRGAN [15]), and translation (three CycleGAN models [33] having two pathways for each).

To investigate the attacking patterns of the image-to-image models, 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 and gradient-based attack. As the feature-based attack, the FDA method [7] is employed. It 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. As the gradient-based attack, the I-FGSM method [5, 14] is employed. It iteratively finds the attacked input \( \tilde{X} \) 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 set a constraint to limit the amount of perturbation \( \epsilon \) with varying \( \epsilon \) values. See the Appendix section for the details of the employed models and attack methods.

### 3.1. Quantitative Vulnerability Evaluation

As we mentioned earlier, it is not possible to evaluate vulnerability of image-to-image models by directly employing conventional quantitative measures used for image classification models (e.g., success rate or fool rate). To this end, we design an evaluation metric based on the peak signal-to-noise ratio (PSNR) that is one of the most widely used metrics for evaluating the quality of the output images in image-to-image tasks. Because an adversarial attack
measures the degree of vulnerability. It is calculated as:

\[ \text{VI} = \frac{\text{PSNR}_{\text{in}}}{\text{PSNR}_{\text{out}}} \].

When the VI value is larger, the model has higher vulnerability. It allows us to compare the level of vulnerability of multiple models easily. We also employ structural similarity (SSIM) instead of PSNR to measure VI and observe similar trends (see the Appendix section). Using VI, we conduct in-depth analysis of the varying distortion patterns and factors affecting the vulnerability in Section 4.1.

### 3.2. Transferability of Universal Perturbations

An adversarial perturbation is basically found for each image, but it is also possible to find an adversarial perturbation that can affect any input image for a given model [19]. This so-called universal perturbation 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 [17], which refers to the applicability of the universal perturbation found for a model to another model. By measuring transferability of different universal perturbations, it is possible to determine whether some models have similar characteristics or not in terms of vulnerability.

We investigate the transferability of universal perturbations in two-fold (Section 4.2). We examine the transferability of universal perturbations from one image-to-image model to another. In addition, we further investigate the transferability of universal perturbations between image-to-image and image classification models, including VGG16 [24], ResNet-101 [11], and MobileNetV2 [22]. 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. To compare the perturbation on the same resolution across different models, we adjust all the input images to have a resolution of 224×224 pixels.

### 3.3. Characteristics of Adversarial Examples

Image-to-image models can 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 [10, 23]. 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

\[ \tilde{\Delta}^{(i)} = \text{IDCT}(M \circ \text{DCT}(\Delta^{(i)})), \]

where DCT(\( \cdot \)) and IDCT(\( \cdot \)) are the 2-D discrete cosine transform (DCT) and inverse DCT, respectively, \( \circ \) denotes the element-wise multiplication, and \( M \in \mathbb{R}^{W \times H \times C} \) is a binary mask that extracts certain frequency components of the perturbation. For attacking the low-frequency subspace, the value of \( M \) at \((w, h, c)\) is set to one when \( w \leq \epsilon W \) and \( h \leq \epsilon H \), and zero otherwise. The parameter \( \epsilon \in [0, 1] \) 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 \( X + \tilde{\Delta}^{(i)} \). We also investigate the effectiveness of attacking the high-frequency subspace by setting the value of \( M \) at \((w, h, c)\) to one when \( w \geq (1 - \epsilon) W \) and \( h \geq (1 - \epsilon) H \), and zero otherwise. The results are presented in Section 4.3.

### 4. Experimental Results

#### 4.1. Vulnerability Evaluation

Figure 2 compares the performance of the models in terms of VI, when \( \epsilon \) is set to 8 (see the Appendix section for different \( \epsilon \) values). We also employ random uniform noise as a baseline attack method, where the bound of noise is set to \([\epsilon, \epsilon] \). The VI values for the random noise are near 1, meaning that the input and output images have deteriorated 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 Figure 3: 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 go deeper into examining degradation patterns along with the VI values.

While any successful attacks on image classification models result in incorrect classification, attacks on image-to-image models result in visual distortions in the output images, where distortion patterns can largely differ depending on various factors. In this regard, we investigate characteristics of vulnerability with respect to different tasks, attack approaches, model structures, and datasets.

**Relation to Tasks.** Figure 3 shows example images obtained with and without the adversarial attacks. For all tasks, the attacks add hardly perceivable perturbation to the input images. However, significant quality deterioration can be found in the output images with various deterioration textures except denoising (as shown in Figure 2). The failure to attack the denoising models is because they are trained to identify and remove the elements unrelated to the original visual content in a given image. The models for the other tasks, on the contrary, 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. Because the visual components that each model tries to recover differ, different patterns of destruction are observed in the output images of different models. We discover more details by investigating transferability of the adversarial examples in Section 4.2.

**Relation to Attack Approaches.** In the classification, the attack results are usually represented only as success rates regardless of attack methods. In the image-to-image tasks, however, characteristics of attack methods can directly affect the patterns of the quality degradation in the output images. This can be observed in Figure 3. Overall, the output images under I-FGSM look more corrupted than those under FDA. It is also shown as larger VI values for I-FGSM in Figure 2. Figure 4 shows another examples when DnCNN is employed. When FDA is used, the sharp textures are largely removed 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.

**Relation to Model Structure.** Several image-to-image models employ generative adversarial networks (GANs) [8] 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 Figure 2, DeepDeblur and CycleGAN, which employ GANs, show relatively larger VI values. This is also confirmed when SRResNet and SRGAN are compared in Figure 5. 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 [4]. Thus, the perturbation in the input image can also be intensified in the output image.

**Relation to Training Datasets.** The translation models share the same model structure, but they show different characteristics in terms of vulnerability and distortion patterns. In Figure 2, 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. Figure 6 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. The same applies to the models converting horses to zebras and Van Gogh’s paintings to photos (see the Appendix section). 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.

**4.2. Universal Perturbations and Transferability**

Figure 7 shows the universal perturbation for each attack method and each task. First of all, the patterns of the perturbations differ depending on the attack method because of the differences in the attack objectives. While FDA tries to find a perturbation 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

![Figure 2](image-url) **Figure 2.** Performance comparison in terms of VI for different attack methods employed with $\epsilon = 8$.}
Figure 3. Visual showcases of the input and output images obtained from the models (CIC, DeepDeblur, DnCNN (color, multiple $\sigma$), RCAN, and CycleGAN (horse→zebra)). The attacked results are obtained with $\epsilon = 8$. Each perturbation is magnified 8 times for better visualization.

Figure 4. Images obtained from DnCNN (grayscale, multiple $\sigma$). FDA and I-FGSM are employed with $\epsilon = 16$.

Figure 5. Visual comparison of the output images obtained from SRResNet and SRGAN, when I-FGSM is employed with $\epsilon = 8$.

噪声样式的扰动，获得于其他任务，倾向于被去除在输出图像中由于去噪过程。它令人感兴趣的是比较于从图像到图像模型中所获得的扰动与从图像分类模型中所获得的扰动。前者包含更多粒状纹理，而后者具有更粗的纹理。一个可能的原因是由于不同的架构：图像分类模型使用空间聚合操作（例如，最大池化，平均池化）来降低中间特征的数量，而图像到图像模型通常不使用这些操作来最小化影响输出质量的信息的损失。

图8比较了VI对于通用攻击的结果。这些结果类似于图2c所获得的图像特定攻击的结果。尽管它稍微低效于的
In the previous sections, we evaluated vulnerability of image-to-image models in various aspects. The next question is: Are the defense methods for image classification models applicable to image-to-image models? To investigate this, we examine the effectiveness of several possible defense methods for the image-to-image tasks, including image transformation and adversarial training, which are popular approaches in image classification tasks (as we mentioned in Section 2). For transformation-based approaches, we consider JPEG compression [6] (with a quality level of 75), random resizing [28], and bit reduction...
[29] (to 4 bits for each channel). In addition, we consider geometric self-ensemble that was introduced in the super-resolution [16] to improve quality of the output images without additional training. For adversarial training, adversarial examples are obtained by I-FGSM.

Figure 12 compares the performance of the defense methods in terms of VI, when I-FGSM is employed ($\epsilon = 8$). The VI values obtained without applying any defense methods are also shown with diamond markers. Overall, the defense methods reduce influences of adversarial perturbations to some extent, i.e., the VI values are reduced. The JPEG compression shows the best effectiveness. Nevertheless, the VI values are still larger than 1, and this implies that deterioration in 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 for unattacked input images. Figure 13 shows that the defense methods can 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
Figure 12. Performance comparison of the frequency-aware attacks in terms of VI for I-FGSM with $\epsilon = 8$ and $r = 1/4$.

Figure 13. Comparison of the original performance under defense in terms of PSNR. Diamond markers refer to the original PSNR values without defense.

Figure 14. Examples of the defenses for EDSR.

to-image models, i.e., significant deterioration is introduced in the output images when the input image is not attacked. The transformations directly affect the original textures of the input image. While the geometric self-ensemble method is originally for improving the original performance, the other transformation-based methods do not consider improving or preserving the original performance. Furthermore, these defense methods are not so effective when an adversarial attack is applied; the deterioration is still visible even though the amount of deterioration is reduced.

These results give us the following insights. First, existing transformation-based approaches that are used in image classification models are not suitable for image-to-image tasks. When devising a defense method to manipulate the input image, it should be carefully designed to not damage the original performance. Second, improving the robustness of the models directly (e.g., adversarial training) may be a more promising solution to defend image-to-image models.
against adversarial attacks.

6. Conclusion

We presented in-depth investigations 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 even transferable across tasks. However, the attack results have significantly different characteristics depending on the task type, attack method, model structure, and dataset. Furthermore, we found that the vulnerability patterns characterized by the perturbation can also differ depending on the aforementioned factors. Finally, our results showed that many of the existing defense methods used for classification models are not suitable to image-to-image models and revealed the necessity of considering the original performance.

A. Appendix

Attack Methods. We consider two types of attack methods that are applicable to image-to-image models: a feature-based method and a gradient-based method.

As a feature-based attack, the FDA method [7] is considered. It tries to reduce the variance of the activation function in the target model by maximizing the following function:

\[
\log \left( \left\| \{ \Phi_{w,h,d} | \Phi_{w,h,d} < C(w,h) \} \right\|_2 \right) - \log \left( \left\| \{ \Phi_{w,h,d} | \Phi_{w,h,d} > C(w,h) \} \right\|_2 \right),
\]

where \( \Phi \) is one of the intermediate features and \( C(w,h) \) is the mean values across the channel dimension. Because it does not rely on the model output to find a perturbation, it can be used to attack image-to-image models as well as classification models. The perturbation for an input image is obtained iteratively while the \( L_\infty \) norm of the perturbation is kept smaller than a constant \( \epsilon \). The number of iterations \( T \) is set to 50 and the amount of the perturbation at each iteration is kept smaller than a constant \( \epsilon \) obtained iteratively while the classification models. The perturbation for an input image is obtained iteratively while the model output to find a perturbation, this method can be applied to the image-to-image models. We set \( \tilde{X}^{(0)} = X \). As in FDA, the number of iterations \( T \) is set to 50.

For all the attack methods, we set \( \epsilon \in \{1, 2, 4, 8, 16, 32\} \) in the pixel value scale of \([0, 255]\).

Models and Datasets. We consider 16 deep models for five popular image-to-image tasks: colorization, deblurring, denoising, super-resolution, and translation. Table 1 summarizes deep learning-based image-to-image models that are examined in this paper. For colorization, CIC [31] is used. For deblurring, DeepDeblur [20] is used. For denoising, three DnCNN models [30] are used, which are trained on grayscale images having various levels of Gaussian noise with \( \sigma \in [0, 55] \), grayscale images having various levels of Gaussian noise (\( \sigma \in [0, 55] \)), and RGB images having various levels of Gaussian noise (\( \sigma \in [0, 55] \)). For super-resolution, EDSR [16], RCAN [32], CARN [2], SRResNet [15], and SRGAN [15] are used. For translation, three CycleGAN models [33] trained on three pairs of datasets in different domains (apple ↔ orange, horse ↔ zebra, and Van Gogh’s paintings ↔ photos) are used. The models are trained with the procedures reported in the original paper. The training datasets are also the same as those used in the original papers.

Evaluation is performed using the datasets that are commonly used for evaluating each task in the literature. For colorization, we use the 1,000 images of the validation split of the ImageNet dataset [21] after cropping at the center regions and resizing to 224×224 pixels (i.e., the same size as the training images). For deblurring, we use the Kôhler dataset [13]. For denoising, we employ 68 images in the BSD500 dataset [3] that are not used for training the models, as in [30]. For super-resolution, we employ the BSDL100 dataset [18]. For translation, we employ the validation dataset provided in [33] for each translation pair.

Implementation Details. We conduct our experiments on various CPUs (e.g., Intel Xeon CPU E5-1660v3, Intel Core i7-7700) and GPUs (e.g., NVIDIA GTX 1080, NVIDIA GTX 2080Ti). We implement our code by using TensorFlow with Python. We will make our code publicly available on GitHub.

Performance Comparison in Terms of PSNR. We first show the performance comparison in terms of VI for different attack methods employed with various \( \epsilon \) values in Figure 15. In addition, we show the performance comparison in terms of the PSNR values for the input (PSNR_in) and the output (PSNR_out) in Figure 16. A curve closer to the lower right corner means that the corresponding model is more vulnerable, i.e., a small amount of perturbation in

\[
\tilde{X}^{(i+1)} = \tilde{X}^{(i)} + \frac{\epsilon}{T} \text{sgn} \left( \nabla \left\| f_m(\tilde{X}^{(i)}) - f_m(X) \right\|_2 \right),
\]
the input image results in a large amount of deterioration in the output image.

**Output Degradation on Other Translation Models.** To supplement the results in Figure 6, we show additional example images obtained from the CycleGAN models converting horses to zebras and Van Gogh’s paintings to photos, when FDA is employed, in Figure 17.

**Performance Comparison in Terms of SSIM.** Along with the evaluation results in terms of PSNR, we also include the performance comparison in terms of structural similarity (SSIM) [27], which is not included in the main sections. Similar to the performance comparison in terms of VI derived from PSNR, we calculate the vulnerability index calculated on SSIM, i.e., $\text{VI}_{\text{SSIM}} = \frac{\text{SSIM}_{\text{in}}}{\text{SSIM}_{\text{out}}}$, where $\text{SSIM}_{\text{in}}$ and $\text{SSIM}_{\text{out}}$ are the SSIM values for the input and output images, respectively.

The performance comparison is shown in Figure 18. The overall trend is similar to that shown in Figure 2. One noticeable difference is that $\text{VI}_{\text{SSIM}}$ penalizes the super-resolution models more than the translation models, while the translation models appear more vulnerable in terms of VI calculated on PSNR. This is because although the pixel-wise changes in the output images are larger for the translation models than for the super-resolution models, which makes VI higher for the former, the changes for the translation models tend to alter the image content (e.g., textures) and be less perceptually annoying than noise-like distortion in the super-resolution models, which is considered by $\text{VI}_{\text{SSIM}}$.

In addition, the performance comparison in terms of $\text{SSIM}_{\text{in}}$ and $\text{SSIM}_{\text{out}}$ is shown in Figure 19.

**Frequency-Aware Attack with Different Values of $r$.** In Section 4.3, we show the frequency-aware attack, which finds the perturbation only in a low-frequency or high-frequency subspace. While the results obtained with $r=1/4$ are included in that section, we also report the results obtained with different values of $r$ in this appendix. The quantitative performance comparison in terms of VI is shown in Figures 20 and 21. In addition, example input and output images are also shown in Figure 22.

**Defenses against Adversarial Attacks.** In Section 5, we discuss the effectiveness of conventional defense approaches in the image-to-image tasks. In this appendix, we show additional example results obtained from the models of five image-to-image tasks. Figure 23 shows visual examples of the transformation-based defenses, including JPEG compression, random resizing, bit reduction, and geometric self-ensemble. Figure 24 shows visual examples of the adversarial training-based defense. Adversarial training shows slightly better defense results than the image transformation methods in terms of both preserving the original performance and improving robustness against adversarial attacks.
| Task            | Model          | With GAN | Training dataset | Evaluation dataset |
|-----------------|----------------|----------|------------------|--------------------|
| **Colorization**| CIC [31]       | -        | ImageNet [21]    | ImageNet [21] subset |
| **Deblurring**  | DeepDeblur [20] | Yes      | GOPRO [20]       | Kohler [13]        |
| **Denoising**   | DnCNN [30] (grayscale, single $\sigma$) | - | BSD500 [3] subset | BSD500 [3] subset |
|                 | DnCNN [30] (grayscale, multiple $\sigma$) | - | BSD500 [3] subset | BSD500 [3] subset |
|                 | DnCNN [30] (color, multiple $\sigma$) | - | BSD500 [3] subset | BSD500 [3] subset |
| **Super-resolution** | EDSR [16] | - | DIV2K [1] | BSD100 [18] |
|                 | RCAN [32]      | - | DIV2K [1] | BSD100 [18] |
|                 | CARN [2]       | - | DIV2K [1] | BSD100 [18] |
|                 | SRResNet [15]  | - | BSD500 [3] subset | BSD100 [18] |
|                 | SRGAN [15]     | Yes | ImageNet [21] subset | BSD100 [18] |
| **Translation** | CycleGAN [33] (apple ↔ orange) | Yes | ImageNet [21] subset | ImageNet [21] subset |
|                 | CycleGAN [33] (horse ↔ zebra) | Yes | ImageNet [21] subset | ImageNet [21] subset |
|                 | CycleGAN [33] (vangogh ↔ photo) | Yes | Wikiart, Flickr [33] | Wikiart, Flickr [33] |

Table 1. Details of the examined image-to-image models and datasets.

Figure 15. Performance comparison in terms of VI for different attack methods employed with $\epsilon \in \{1, 2, 4, 8, 16, 32\}$. 
Figure 16. Performance comparison in terms of $\text{PSNR}_{\text{in}}$ and $\text{PSNR}_{\text{out}}$. Six points of each curve correspond to six different values of $\epsilon$.

Figure 17. Images obtained from CycleGAN trained using different datasets. FDA is employed with $\epsilon = 8$. 
Figure 18. Performance comparison in terms of $V_{\text{SSIM}}$ for different attack methods employed with $\epsilon = 8$.

Figure 19. Performance comparison in terms of $\text{SSIM}_{\text{in}}$ and $\text{SSIM}_{\text{out}}$. Six points of each curve correspond to six different values of $\epsilon$. 
Figure 20. Performance comparison in terms of VI for the low-frequency attack with $\epsilon = 8$.

Figure 21. Performance comparison in terms of VI for the high-frequency attack with $\epsilon = 8$.

Figure 22. Images obtained from (a) DeepDeblur and (b) EDSR, where the low-frequency and high-frequency attacks are employed with $\epsilon = 8$ and $r \in \{1/8, 2/8, 3/8, 4/8\}$. 
|                | Original outputs | I-FGSM ($\epsilon = 8$) |
|----------------|------------------|-------------------------|
| No defense     | JPEG             | Resizing                | Bit reduction | Self-ensemble |
| No defense     | JPEG             | Resizing                | Bit reduction | Self-ensemble |

(a) Colorization (CIC)

(b) Deblurring (DeepDeblur)

(c) Denoising (DnCNN (color, multiple $\sigma$))

(d) Super-resolution (RCAN)

(e) Translation (CycleGAN, Van Gogh→Photo)

Figure 23. Example output images of the transformation-based defenses.

|                | Original outputs | I-FGSM ($\epsilon = 8$) |
|----------------|------------------|-------------------------|
| No defense     | With adversarial training | No defense | With adversarial training |
| No defense     | With adversarial training | No defense | With adversarial training |

(a) Colorization (CIC)

(b) Deblurring (DeepDeblur)

(c) Denoising (DnCNN (color, multiple $\sigma$))

(d) Super-resolution (CARN)

(e) Translation (CycleGAN, Apple→Orange)

(f) Translation (CycleGAN, Van Gogh→Photo)

Figure 24. Example output images without and with the adversarial training.
References

[1] Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017. 11

[2] 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, 2018. 2, 9, 11

[3] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(5):898–916, 2010. 9, 11

[4] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 4

[5] Jun-Ho Choi, Huan Zhang, Jun-Hyuk Kim, Cho-Jui Hsieh, and Jong-Seok Lee. Evaluating robustness of deep image super-resolution against adversarial attacks. In Proceedings of the IEEE International Conference on Computer Vision, 2019. 2, 9

[6] Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M. Roy. A study of the effect of JPG compression on adversarial images. arXiv:1608.00853, 2016. 2, 6

[7] Aditya Ganeshan and R. Venkatesh Babu. FDA: Feature disruptive attack. In Proceedings of the IEEE International Conference on Computer Vision, 2019. 2, 9

[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, 2014. 4

[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] Chuan Guo, Jared S. Frank, and Kilian Q. Weinberger. Low frequency adversarial perturbation. In Proceedings of the Conference on Uncertainty in Artificial Intelligence, 2019. 3

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 3

[12] Ruitong Huang, Bing Xu, Dale Schuurmans, and Csaba Szepesvári. Learning with a strong adversary. In Proceedings of the International Conference on Learning Representations, 2016. 2

[13] Rolf Köhler, Michael Hirsch, Betty Mohler, Bernhard Schölkopf, and Stefan Harmeling. Recording and playback of camera shake: Benchmarking blind deconvolution with a real-world database. In Proceedings of the European Conference on Computer Vision, 2012. 9, 11

[14] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In Proceedings of the International Conference on Learning Representations, 2017. 2, 9

[15] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan 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. 2, 9, 11

[16] 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, 2017. 1, 2, 7, 9, 11

[17] 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, 3

[18] 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, 2001. 9, 11

[19] 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, 2017. 2, 3

[20] Seungjun Nah, Tae Hyun Kim, and Kyoung 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. 1, 2, 9, 11

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 115(3):211–252, 2015. 9, 11

[22] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 3

[23] Yash Sharma, Gavin Weiguang Ding, and Marcus A. Brubaker. On the effectiveness of low frequency perturbations. In Proceedings of the International Joint Conference on Artificial Intelligence, 2019. 3

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014. 3

[25] 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, 2018. 1, 2

[26] 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

[27] Zhou Wang, Alan C. Bovik, Hamid R Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility
to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[28] 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.

[29] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Proceedings of the Annual Network and Distributed System Security Symposium*, 2018.

[30] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017.

[31] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *Proceedings of the European Conference on Computer Vision*, 2016.

[32] 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*, 2018.

[33] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.