Adversarially Robust Deep Image Super-Resolution using Entropy Regularization

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

Abstract. Image super-resolution has been widely employed in various applications with boosted performance thanks to the deep learning techniques. However, many deep learning-based models are highly vulnerable to adversarial attacks, which is also applied to super-resolution models in recent studies. In this paper, we propose a defense method that is formulated as an entropy regularization loss for model training, which can be augmented to the original training loss of super-resolution models. We show that various state-of-the-art super-resolution models trained with our defense method are more robust against adversarial attacks than their original versions. To the best of our knowledge, this is the first attempt of adversarial defense for deep super-resolution models.

1 Introduction

Image super-resolution, which is a task to obtain an image having higher spatial resolution than the given image, is one of the most actively researched image enhancement techniques in recent days. Notably, development of the deep learning technology brings significant improvement in the performance of super-resolution over conventional image upsampling methods such as bicubic and bilinear upscaling. Consequently, deep learning-based super-resolution has been successfully applied to the real-world applications, including medical imaging, remote sensing, biometric identification, and visual surveillance [1].

While the deep learning shows promising results in various research fields, concerns about the vulnerability of deep learning-based algorithms against malicious attacks have arisen. Many studies have shown that the adversarial attacks, which add unnoticeable small perturbations to the given image, can fool the target classification model to produce wrong results [2, 3]. Recently, such vulnerability has also been observed beyond classification problems. In the deep super-resolution models, the state-of-the-art deep models produce largely deteriorated outputs [4] or lead erroneous results when they are used as pre-processing steps for other computer vision tasks [5]. Defense methods have been proposed for deep classification models to ensure robustness against adversarial attacks [2, 6–9], but not for super-resolution models.

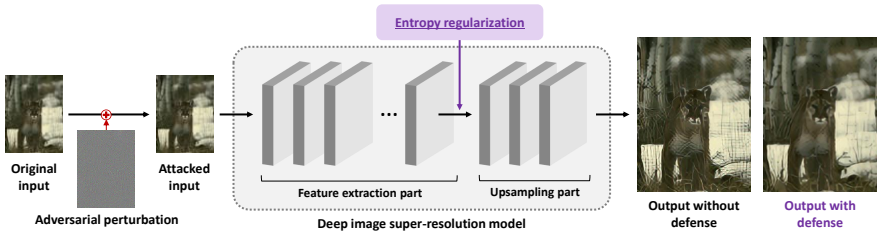


Fig. 1. Illustration of the proposed defense method against adversarial attacks on deep image super-resolution models.

In this paper, we propose a novel defense method, which can be applied to various deep image super-resolution models. Our method aims to reduce the sensitivity of the target super-resolution model to adversarial perturbations by adjusting the activation patterns of an intermediate layer. For this, our method employs a probability density estimator, which is used to obtain the probabilities of the intermediate feature values. Our method then tries to reduce the entropy of the estimated probability distribution by minimizing an entropy regularization loss during model training. As a result, the intermediate features do not change much when an adversarial perturbation is introduced in the input image, and thus undesirable degradation of the super-resolved output image can be prevented effectively. We conduct thorough experimental investigations and ablation studies in order to evaluate the proposed method. In addition, we examine the feasibility of utilizing our method together with the adversarial training strategy. To the best of our knowledge, this is the first approach to defend super-resolution models against adversarial attacks. The idea of our method is illustrated in Fig. 1.

2 Related Work

2.1 Image super-resolution

Recently, many deep learning-based image super-resolution methods have been proposed. One of the earliest approaches is the super-resolution convolutional neural network (SRCNN) model [10], which consists of two convolutional layers. After that, much deeper and more complex models are introduced to achieve better performance. For example, Lim et al. [11] propose the enhanced deep super-resolution (EDSR) model, which employs more than 30 convolutional layers. In addition, EDSR contains residual blocks that have skip connections for better training procedures. Zhang et al. [12] propose the residual channel attention network (RCAN) model, which adds a channel attention mechanism to the residual blocks to handle the intermediate features efficiently. Li et al. [13] develop the multi-scale residual network (MSRN) model that employs convolutional layers having various kernel sizes to utilize image features in a multi-scale manner.

While many approaches including the aforementioned ones aim to achieve high quantitative performance with large networks in terms of peak signal-to-noise ratio (PSNR), some proposals focus on different objectives. For instance, Ledig et al. [14] build a model named SRGAN, which employs a discriminator network of the generative adversarial network (GAN) [15] for their super-resolution method named SRResNet to improve the perceptual quality of the super-resolved outputs. Ahn et al. [16] propose the cascading residual network (CARN) model, which employs cascading residual blocks to reduce the model size without performance degradation. Some other recent approaches employ quantitative score prediction [17], wavelet transform [18], and so on.

2.2 Adversarial attacks

Many researchers reveal the vulnerability of deep image classifications models against various adversarial attacks such as optimization-based [19] and gradient sign-based [2] methods. While most of the studies focus on attacking classification models, the vulnerability of deep models for other tasks has been noted recently. For example, Ganeshan et al. [20] propose the feature disruptive attack (FDA) method, which attempts to perturb the intermediate features of the given deep model. Choi et al. [4] develop an attack method for super-resolution models, which extends the iterative fast gradient sign method (I-FGSM) developed for the classification task. They also extend the attack-agnostic vulnerability measure for classification, named cross Lipschitz extreme value for network robustness (CLEVER) [21], to the super-resolution task.

2.3 Defense against adversarial attacks

Defense methods against adversarial attacks have been proposed for the classification models. One of the well-known effective solutions is adversarial training, which uses adversarial examples as training data [2, 7, 8]. Another approach is to pre-process the input images to reduce the amount of perturbations, such as JPEG compression [6] and random resizing [9]. Since super-resolution is one of the image enhancement techniques, it is sometimes used as a defense method for classification models [22]. However, no method has been proposed to defend the super-resolution models themselves against adversarial attacks.

3 Proposed Method

Our defense method for a deep super-resolution model basically aims to make the intermediate activations of the model be insensitive against adversarial perturbations. For this, we build our method with the following two components. First, we train the model in such a way that the entropy of the probability distribution for the intermediate activation values is minimized, so that the distribution of activations remains similar across different input images. Second, random noise

is added to the intermediate activations during training so that the insensitivity is further enhanced.

Consider a super-resolution model denoted by $S(\cdot)$, which outputs a super-resolved high-resolution image $\mathbf{X}_{SR}$ from a low-resolution input image $\mathbf{X}_{LR}$, i.e., $\mathbf{X}_{SR} = S(\mathbf{X}_{LR})$. The main objective of the super-resolution task is to minimize the reconstruction loss function $\mathcal{L}_r$, which calculates the quantitative difference between the output image $\mathbf{X}_{SR}$ and its corresponding ground-truth image $\mathbf{X}_{HR}$. For this, pixel-wise L_1 [11, 12, 16] or L_2 [14] losses are typically used. Other losses can also be appended as part of the reconstruction loss, e.g., adversarial loss [14] and perceptual loss [17]. Our method defines an additional loss function, the entropy regularization loss function $\mathcal{L}_e$, and uses it together with the original reconstruction loss function for model training.

Let $\mathbf{\Phi} \in \mathbb{R}^{W \times H \times D}$ denote the features extracted from a selected intermediate layer of $S(\cdot)$, where W, H, and D are the width, height, and channel dimensions, respectively. Our defense method aims to regulate $\mathbf{\Phi}$ by minimizing the entropy value for each channel. This can be summarized as

$$\mathcal{L}_e = -\frac{1}{WHD} \sum_{d \in D} \sum_{w \in W} \sum_{h \in H} \log_2 p_d\big(\mathbf{\Phi}_{w,h,d}\big) \tag{1}$$

where $p_d(\cdot)$ is the probability density function for the d-th channel and $\mathbf{\Phi}_{w,h,d}$ is the value of $\mathbf{\Phi}$ at (w, h, d)[3]. We add this loss function to the original loss function, i.e.,

$$\mathcal{L} = \mathcal{L}_r + \lambda \mathcal{L}_e \tag{2}$$

where λ is a hyperparameter that controls the amount of the contribution of the entropy regularization.

To calculate $\mathcal{L}_e$, it is necessary to estimate the probability density function of the features. Recently, methods to estimate cumulative distribution using neural networks have been proposed [23–25]. Once the cumulative distribution function $c_d(\cdot)$ is obtained, the corresponding probability density function $p_d(\cdot)$ can be derived as $p_d(x) = \frac{\partial}{\partial x} c_d(x)$. Adopting the recent method in [23], we design a neural network-based cumulative distribution estimation method to enable end-to-end optimization of super-resolution models for minimizing the objective function in (2). Suppose that $c_d(\cdot)$ is obtained by a cascade of K functions, i.e.,

$$c_d(\cdot) = f_{d,K}\Big(f_{d,K-1}\big(f_{d,K-2}(...)\big)\Big) \tag{3}$$

where $f_{d,i}$ takes an $m_{d,i}$-dimensional vector as input and outputs an $m_{d,i+1}$-dimensional vector (i.e., $\mathbb{R}^{m_{d,i}} \rightarrow \mathbb{R}^{m_{d,i+1}}$). Note that $m_{d,1} = 1$ (feature values) and $m_{d,K+1} = 1$ (probability values). Then,

$$p_d(\cdot) = {f'}_{d,K} \cdot {f'}_{d,K-1} \cdot ... \cdot {f'}_{d,1} \tag{4}$$

[3] Note that the entropy is expressed by the sum over the feature elements, not by the sum over the feature values.

where ${f'}_{d,i}$ is the derivative of $f_{d,i}$. Since $p_d(\cdot)$ outputs a probability value, both $c_d(\cdot)$ and $p_d(\cdot)$ must be within $[0, 1]$. In addition, $c_d(\cdot)$ must be monotonically increasing. To meet these constraints, we design the functions $f_{d,i}$ as

$$f_{d,i}(\mathbf{x}) = \begin{cases} \sigma\big(\mathbf{M}_{d,i}\mathbf{x} + \mathbf{b}_{d,i}\big) & \text{if } i = K \\ g_{d,i}\big(\mathbf{M}_{d,i}\mathbf{x} + \mathbf{b}_{d,i}\big) & \text{otherwise} \end{cases} \tag{5}$$

$$g_{d,i}(\mathbf{x}) = \mathbf{x} + \mathbf{a}_{d,i} \circ \tanh(\mathbf{x}) \tag{6}$$

where $\mathbf{M}_{d,i}$ is a matrix, $\mathbf{a}_{d,i}$ and $\mathbf{b}_{d,i}$ are vectors, $\sigma(\cdot)$ is the sigmoid function, $\tanh(\cdot)$ is the hyperbolic tangent function, and $\circ$ denotes the element-wise multiplication. The sigmoid function in $f_{d,K}(\cdot)$ forces the range of $c_d(x)$ to be within $[0, 1]$. Then, the derivatives of the functions are given as

$${f'}_{d,i}(\mathbf{x}) = \begin{cases} \sigma'\big(\mathbf{M}_{d,i}\mathbf{x} + \mathbf{b}_{d,i}\big) \cdot \mathbf{M}_{d,i} & \text{if } i = K \\ \mathrm{diag}\Big({g'}_{d,i}\big(\mathbf{M}_{d,i}\mathbf{x} + \mathbf{b}_{d,i}\big)\Big) \cdot \mathbf{M}_{d,i} & \text{otherwise} \end{cases} \tag{7}$$

$${g'}_{d,i}(\mathbf{x}) = 1 + \mathbf{a}_{d,i} \circ \tanh'(\mathbf{x}). \tag{8}$$

To make the range of $p_d(x)$ be within $[0, 1]$, we replace $\mathbf{M}_{d,i}$ and $\mathbf{a}_{d,i}$ with

$$\mathbf{M}_{d,i} = \mathrm{softplus}\big(\hat{\mathbf{M}}_{d,i}\big) \tag{9}$$

$$\mathbf{a}_{d,i} = \tanh\big(\hat{\mathbf{a}}_{d,i}\big). \tag{10}$$

This also ensures the monotonicity of $c_d(\cdot)$. To summarize, K and $m_{d,i}$ are the hyperparameters and $\hat{\mathbf{M}}_{d,i}$, $\hat{\mathbf{a}}_{d,i}$, and $\mathbf{b}_{d,i}$ are the parameters to be trained.

From the cumulative distribution $c_d(\cdot)$ obtained by our neural network-based estimator, we estimate the probability $p_d(\cdot)$ by considering feature values within $\mathbf{\Phi}_{w,h,d} - (\delta/2)$ and $\mathbf{\Phi}_{w,h,d} + (\delta/2)$ to be similar to $\mathbf{\Phi}_{w,h,d}$, i.e.,

$$p_d\big(\mathbf{\Phi}_{w,h,d}\big) = c_d\Big(\mathbf{\Phi}_{w,h,d} + \frac{\delta}{2}\Big) - c_d\Big(\mathbf{\Phi}_{w,h,d} - \frac{\delta}{2}\Big) \tag{11}$$

where δ determines the range of the similarity.

In addition, to further enhance the insensitivity of the feature values to perturbations by the attack, we apply random uniform noise to $\mathbf{\Phi}$ i.e.,

$$\mathbf{\Phi} \leftarrow \mathbf{\Phi} + \mathbf{\Gamma} \tag{12}$$

where $\mathbf{\Gamma}$ is the uniform noise in a range of $[-\delta/2, \delta/2]$. Because $\mathbf{\Gamma}$ changes at every training iteration, the super-resolution model is trained to consider similar feature values as the same in order to generate outputs consistently. Note that the random uniform noise is applied only during training.

4 Experiments

We conduct experiments with state-of-the-art deep super-resolution models to evaluate the proposed method. This section provides the experimental settings, including target models, evaluation methods, and evaluation conditions.

4.1 Super-resolution models

Our method can be applied to various deep super-resolution models. We select six representative models for our experiments, including EDSR [11], SRResNet [14], SRGAN [14], RCAN [12], MSRN [13], and CARN [16]. All these models have similar structures, i.e., the input low-resolution image is processed through convolutional layers, and upsampling is performed at the final stage, as illustrated in Fig. 1. EDSR consists of several residual blocks, whose structure is widely used in many other state-of-the-art super-resolution models. In this paper, we employ the baseline version of EDSR. SRResNet employs batch normalization and parametric ReLU [26]. SRGAN is an extended version of SRResNet, which employs a discriminator network to improve the perceptual quality of the upscaled images. RCAN employs the so-called "residual in residual" structure and a channel attention mechanism to utilize channel-wise features more thoroughly. MSRN consists of convolutional layers having different kernel sizes from 1×1 to 5×5. CARN is a lightweight super-resolution model in terms of the model size, and it is reported that CARN is one of the most robust super-resolution models against adversarial attacks due to its small model size [4].

According to the training procedures specified in the original papers, we train EDSR, RCAN, MSRN, and CARN on the DIV2K dataset [27], and SRResNet and SRGAN on a 350k subset of the ImageNet dataset [28]. The L_1 loss function with the pixel value range of $[0, 255]$ is used for training the EDSR, RCAN, MSRN, and CARN models. The L_2 loss function with the pixel value range of $[-1, 1]$ is used for training the SRResNet and SRGAN models. We consider super-resolution at a scaling factor of 4 in all experiments.

4.2 Attack methods

We employ two types of attack methods that are applicable to super-resolution: feature-based attack and gradient-based attack.

For the feature-based attack, we employ the feature disruptive attack (FDA) [20], which aims to reduce the variance of the activation at intermediate layers of the target model. For a given intermediate feature $\mathbf{\Phi}$, the objective is to maximize the following function:

$$\log\left(\left|\left|\{\mathbf{\Phi}_{whd}|\mathbf{\Phi}_{whd} < C(w,h)\}\right|\right|_2\right) - \log\left(\left|\left|\{\mathbf{\Phi}_{whd}|\mathbf{\Phi}_{whd} > C(w,h)\}\right|\right|_2\right) \quad (13)$$

where $C(w,h)$ is the mean values across the channel dimension. Since it does not depend on the final output, this attack method can be applied to various deep models in addition to classification models. The perturbations in the input image are found iteratively while the L_∞ norm of the perturbations is kept smaller than a constant ϵ. We conduct our experiment with various values of ϵ. The number of iterations (nb_{iter} in [20]) is set to 50. The amount of the perturbations at each iteration (ϵ_{iter} in [20]) is set to ϵ/nb_{iter}.

For the gradient-based attack, we employ the iterative fast gradient sign method (I-FGSM), which is first introduced for the classification task [8] and recently extended for the super-resolution task [4]. It iteratively finds the attacked

input $\widetilde{\mathbf{X}}_{LR}$ by

$$\widetilde{\mathbf{X}}_{LR}^{(t+1)} = \widetilde{\mathbf{X}}_{LR}^{(t)} + \frac{\alpha}{T} \cdot \mathrm{sgn}\Big(\nabla \big|\big| S(\widetilde{\mathbf{X}}_{LR}^{(t)}) - S(\mathbf{X}_{LR}) \big|\big|_2\Big) \tag{14}$$

where α and T are the hyperparameters that controls the amount of the perturbations and $\mathrm{sgn}(\cdot)$ is the sign function. We set $T = 50$ and use various values of α to evaluate our proposed method.

We also employ other gradient-based methods (see the supplementary material), which show similar results as I-FGSM.

4.3 Evaluation conditions

Our probability density estimator is attached to the last layer right before the upsampling part for each super-resolution model as shown in Fig. 1. For probability estimation, we set $K = 4$ and $m_{d,2} = m_{d,3} = m_{d,4} = 3$. We set $\lambda = 1$ for the EDSR, RCAN, MSRN, and CARN models and $\lambda = 0.1$ for the SRResNet and SRGAN models[4]. The effect of the value of λ is also examined (Section 5.3). We set $\delta = 1$, and the effect of changing this value is also examined (Section 5.5). For the attacks, we employ $\{\epsilon, \alpha\} \in [1/255, 2/255, 4/255, 8/255, 16/255, 32/255]$ for FDA and I-FGSM.

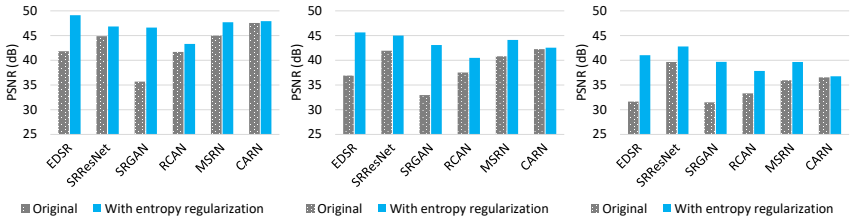
We evaluate the effectiveness of our defense method against the adversarial attack methods primarily in terms of PSNR[5]. PSNR is one of the most widely used metrics for evaluating quality of the super-resolved images. To conduct comprehensive analysis of our proposed method, we measure the PSNR values in two-fold: for low-resolution images and super-resolved images. The PSNR for low-resolution images is measured between the original input $\mathbf{X}_{LR}$ and the attacked one $\widetilde{\mathbf{X}}_{LR}$, which quantifies the amount of the added perturbations. The PSNR for super-resolved images is measured between the original output $\mathbf{X}_{SR}$ and the output obtained from the attacked input $\widetilde{\mathbf{X}}_{SR}$. If a super-resolution model is more robust than another model, the attack method would have difficulty to successfully attack the model, and thus attempt to add a larger amount of perturbations. Therefore, the PSNR values of the low-resolution and high-resolution images would become smaller and larger, respectively, than those for the less robust model.

In addition, we also employ CLEVER [21], which is an attack-independent vulnerability measure originally developed for classification models. It is also extended for super-resolution [4]. It draws N_s random perturbations having values within $[-\alpha_c, \alpha_c]$ and calculates

$$\max_j \Big|\Big| \nabla \big|\big| S(\mathbf{X}_{LR} + \boldsymbol{\Delta}^{(j)}) - S(\mathbf{X}_{LR}) \big|\big|_2 \Big|\Big|_1 \tag{15}$$

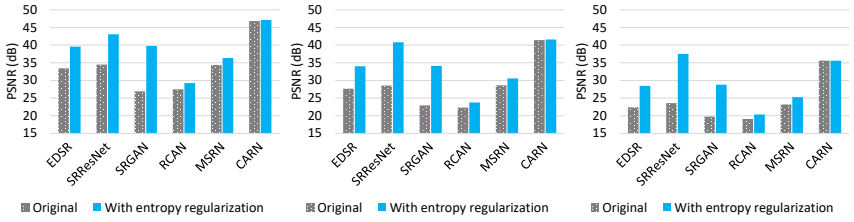
[4] We use a smaller λ value for these models because of the different loss function and range of the pixel values, as explained in Section 4.1.

[5] We observed that the results in terms of structural similarity (SSIM) also show similar tendency to those in terms of PSNR.



$\epsilon = 1/255$ $\epsilon = 2/255$ $\epsilon = 4/255$

(a) FDA



$\alpha = 1/255$ $\alpha = 2/255$ $\alpha = 4/255$

(b) I-FGSM

Fig. 2. PSNR values of the super-resolved images for different models trained only with the original reconstruction loss (gray colors) and with both the reconstruction and entropy regularization losses (blue colors). A larger PSNR indicates better robustness.

where $\mathbf{\Delta}^{(j)}$ is the j-th random perturbation. A larger value of the CLEVER index means higher vulnerability. We set $N_s = 1024$ and $\alpha_c = 1/255$ as in [4].

We employ three popular image datasets, which are Set5 [29], Set14 [30], and BSD100 [31]. The results for BSD100 are reported in this paper. The results for the other datasets are reported in the supplementary material.

5 Results

5.1 Model comparison

We first compare the performance of the super-resolution methods in terms of PSNR for the FDA and I-FGSM attacks. Fig. 2 shows the PSNR values of the super-resolved images for different amounts of perturbations. Overall, the performance decreases when the amount of perturbations (i.e., ϵ and α) increases since there is more room to conceal malicious perturbations in the input images. Among the six super-resolution models trained without our defense method, CARN shows the highest robustness, which can be observed as the highest PSNR values except for the case of FDA with $\epsilon = 4/255$. It is also observed that I-FGSM is stronger than FDA, lowering the PSNR values more significantly.

The results in Fig. 2 show that our defense method is effective for various deep super-resolution models and attack methods. For example, for FDA with

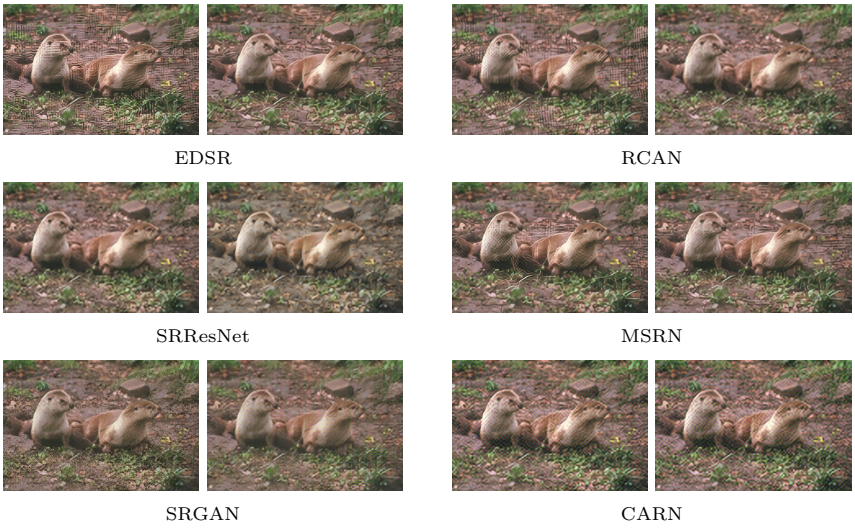


Fig. 3. Images obtained from the super-resolution models trained without (left panels) and with (right panels) entropy regularization. FDA ($\epsilon = 8/255$) is employed as the attack method.

$\epsilon = 4/255$, our method enhances the quality of the output images of EDSR from 31.65 dB to 41.03 dB. All of the models trained with our entropy regularization achieve higher PSNR values than those trained without entropy regularization, except for the case of CARN attacked by I-FGSM with $\alpha = 4/255$.

Fig. 3 depicts example super-resolved images under the FDA attack. The left-side and right-side images represent outputs obtained from the models trained without and with our entropy regularization method, respectively. Without defense, the models tend to output super-resolved images having undesirable textures, which come from the perturbations included in the input images. In contrast, our defense method reduces the quality degradation significantly.

We investigate the effect of our method on the intermediate features of the super-resolution models. Fig. 4 shows the features (averaged along the channel dimension) and their distribution at the intermediate layer of the EDSR model where the entropy regularization is applied (i.e., the layer before the upsampling part). When the proposed defense method is not employed, the intermediate layer tends to output the features that are emphasized on the edges, which usually contain high-frequency information. In addition, the distribution of the features is more dispersed than that obtained from the model trained with the entropy regularization, which can be observed in the histograms. When the perturbations, which contain high-frequency components, are introduced in the low-resolution input image by the attack, both the edge regions and the perturbations are amplified, producing corrupted features. Accordingly, the distribution of the feature values is also affected; it is more dispersed than that for the unattacked input.

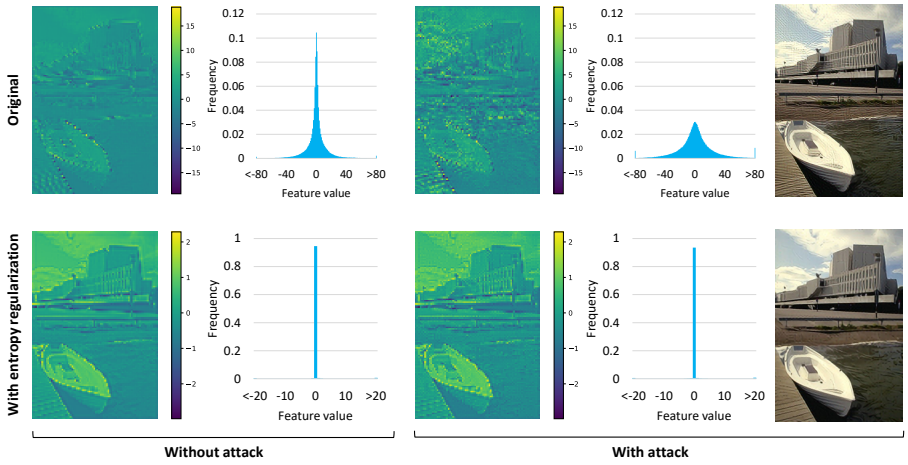
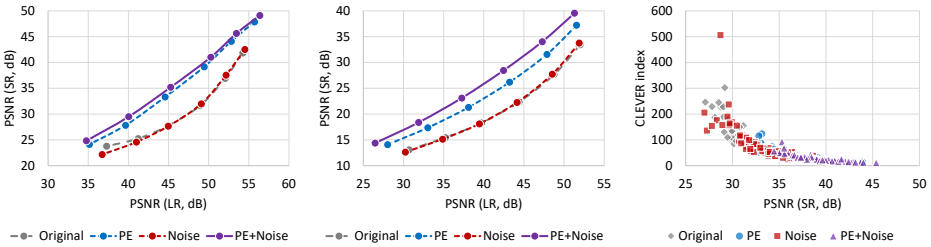


Fig. 4. Intermediate features, histograms of the intermediate feature values, and output images of the EDSR models trained without and with entropy regularization. FDA with $\epsilon = 8/255$ is employed as the attack method.

On the other hand, the model trained with our defense method shows a different mechanism of finding information useful for upsampling from the given input, which results in a different pattern of the intermediate features. Since the activation of the intermediate layer is not heavily concentrated on the edge regions, unlike the model trained without entropy regularization, the perturbations are not significantly amplified. In addition, the distribution of the feature values for the attacked input remains similar to that for the original input. Thanks to the reduced sensitivity against adversarial attacks, a relatively small difference between the features obtained from the original and attacked images is observed in the defended model. Because of these effects, the model trained with the entropy regularization is much more robust against the adversarial attack, as shown in the rightmost images in the figure.

5.2 Ablation study

Our method consists of two components, i.e., probability estimation of intermediate features for entropy regularization and random noise injection to the features. We examine the effects of these two. Fig. 5 compares the performance of the EDSR models trained with only one of the two components or both. In Figs. 5(a) and 5(b), a lower low-resolution (LR) PSNR value means that the adversarial attack adds a larger amount of perturbations to the given input, and a higher super-resolved (SR) PSNR value means that the output image is more similar to the original output. Thus, a model with its graph closer to the upper left corner can be considered to have higher robustness. In Fig. 5(c), we plot the CLEVER index with the SR PSNR value for each image attacked by I-FGSM



(a) PSNR (FDA) (b) PSNR (I-FGSM) (c) CLEVER

Fig. 5. Performance comparison in terms of PSNR and CLEVER index values for the EDSR models trained without any defense (Original), with probability estimation (PE), with random noise (Noise), and both (PE+Noise). Six data points of each case in (a) and (b) correspond to six different values of ϵ and α, respectively. Different data points of each case in (c) correspond to different input images.

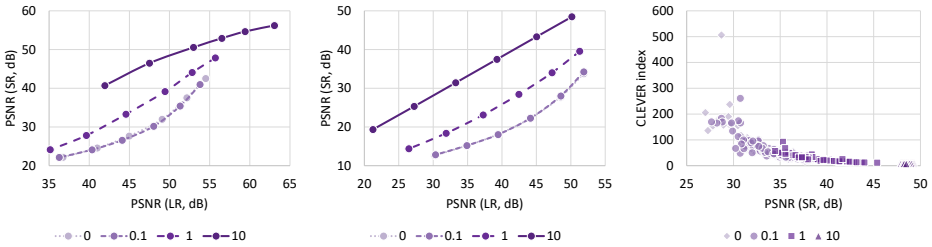
with $\alpha = 1/255$. In this graph, a model having data points closer to the lower right corner is considered to be more robust.

Overall, the models employing the probability estimation show better performance in terms of both the PSNR values and CLEVER indices. Employing the random uniform noise along with the probability estimation provides even more improved performance than employing only the probability estimation. However, employing only the random uniform noise does not improve robustness. In this case, we observed that the range of the intermediate feature values increases. This indicates that the model is trained to produce feature values that have sufficiently large differences so that the differences exceed the magnitude of the noise. These results support that estimating and reducing the entropy of the intermediate features is beneficial to defend against adversarial attacks. In addition, adding random noise does not directly improve the robustness but can boost the effectiveness of the probability estimation.

5.3 Adjusting λ

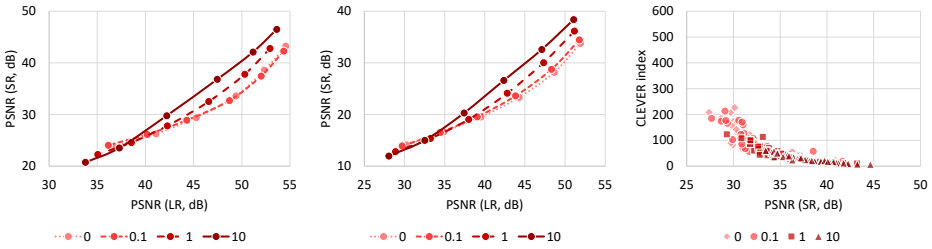
As explained in Section 3, the hyperparameter λ in (2) adjusts the relative contributions of the reconstruction loss and the entropy regularization loss. Hence, we can expect that a larger λ value will improve the robustness against adversarial attacks but possibly at the cost of reconstruction quality reduction. To examine this, we train the EDSR models with different values of λ.

Fig. 6 shows the results. As expected, the models trained with larger λ values show better performance. The PSNR values for the original low-resolution images (without attack) are measured as 27.54, 27.53, 27.28, and 26.41 dB for $\lambda = 0$, 0.1, 1, and 10, respectively. Therefore, the tradeoff relationship exists between the reconstruction loss and the entropy regularization loss, but the amount of performance degradation for unattacked input images is marginal. This indicates that our method efficiently improves the robustness against adversarial attacks with preserving the original performance.



(a) PSNR (FDA) (b) PSNR (I-FGSM) (c) CLEVER

Fig. 6. Performance comparison in terms of PSNR and CLEVER index values for the EDSR models trained with different values of λ.



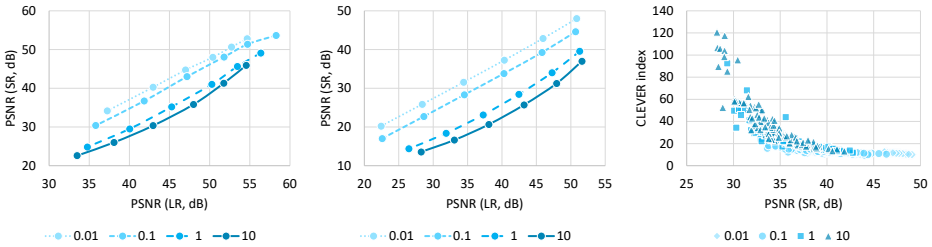
(a) PSNR (FDA) (b) PSNR (I-FGSM) (c) CLEVER

Fig. 7. Performance comparison in terms of PSNR and CLEVER index values for the EDSR models trained with different values of λ, where the entropy regularization loss is applied to the first convolutional layer.

5.4 Target layer for entropy regularization

While we use the entropy regularizer for the last layer of the feature extraction part, the same mechanism can also be applied to any other intermediate layers. We test the case where the regularization is performed at the first layer. The results are shown in Fig. 7 for comparison with Fig. 6.

It is observed that although the defense is slightly successful for small amounts of perturbations (i.e., the region with large PSNR values for low-resolution images), the entropy regularization is much less effective for improving robustness in comparison to the results in Fig. 6. There could be two reasons. First, the functional complexity of only one convolutional layer is not sufficiently high to adjust the features for entropy minimization, whereas several layers can co-operate to adjust the features at the last layer. Second, even though the first layer reduces the effect of the attack, small perturbations that still exist in the features can be undesirably amplified through the subsequent layers. Therefore, employing our defense method in the latter part of a model is more effective than using it in the former part.



(a) PSNR (FDA) (b) PSNR (I-FGSM) (c) CLEVER

Fig. 8. Performance comparison in terms of PSNR and CLEVER index values for the EDSR models trained with different values of δ.

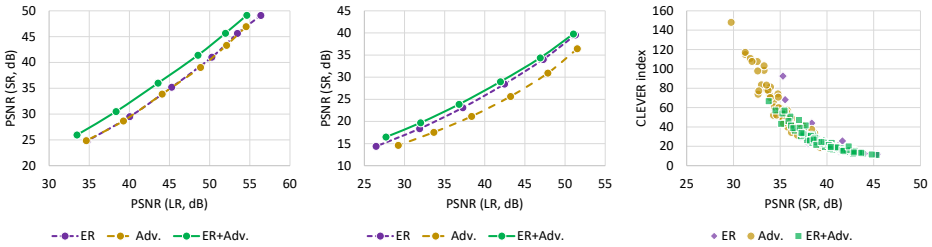
5.5 Adjusting δ

To examine the effect of δ, we train EDSR with different values of δ, where λ is set to 1. Fig. 8 shows the results. It is observed that as δ gets smaller, the robustness against the attacks is improved. Meanwhile, the PSNR values for the original (unattacked) low-resolution images are 26.89, 27.13, 27.28, and 27.34 dB for $\delta = 0.01$, 0.1, 1, and 10, respectively. As explained in Section 3, the hyperparameter δ adjusts the similarity range in calculating the probabilities of the features. A smaller value of δ improves the robustness against the attacks by forcing the feature values to be more similar through entropy minimization, but slightly reduces the reconstruction quality, similarly to a larger value of λ.

5.6 Combining with adversarial training

There is no prior work on defending super-resolution models against adversarial attacks, and it is not straightforward to adopt most defense methods developed particularly for classification tasks due to the different characteristics between the two tasks. As explained in Section 2.3, there are two popular defense approaches in classification: pre-processing of input images and adversarial training. Unlike classification, however, pre-processing degrades the original performance of the super-resolution methods because it inevitably introduces quality degradation of the input low-resolution image. For example, when the random resizing method [9] is applied to the input image, PSNR of the super-resolved output image by EDSR is significantly reduced from 27.54 dB to 26.32 dB.

On the other hand, the adversarial training approach is applicable to the super-resolution tasks by generating adversarial examples from the attack methods for super-resolution, e.g., FDA or I-FGSM. Therefore, we develop a simple adversarial training method and investigate whether our entropy regularization method can collaborate with it to further improve the robustness of super-resolution models. To generate adversarial training examples, we choose I-FGSM in (14) as an attack method, where α and T are set to 8/255 and 1, respectively. For each training iteration, an adversarial perturbation for each input low-resolution image is calculated. Then, both the original low-resolution images



(a) PSNR (FDA) (b) PSNR (I-FGSM) (c) CLEVER

Fig. 9. Performance comparison in terms of PSNR and CLEVER index values for the EDSR models trained with entropy regularization (ER), adversarial training (Adv.) and both (ER+Adv.).

and the attacked low-resolution images serve as inputs, where the corresponding ground-truth high-resolution images remain the same.

We train the EDSR models with our entropy regularization method, the aforementioned adversarial training, and both. Fig. 9 compares the performance of the models. The entropy regularization and adversarial training show similar performance against the FDA method. However, for the I-FGSM attack, the entropy regularization shows better performance than the adversarial training. Interestingly, the model trained with both defense methods achieves the best performance for both attack methods. This proves that 1) our proposed method itself is effective in defending the super-resolution models, and 2) our method can collaborate with the adversarial training method to further improve the robustness against adversarial attacks.

6 Conclusion

We proposed a defense method designed for deep super-resolution models to defend against adversarial examples. Our method manipulates the intermediate features of the given model by estimating their probability density and regularizing their entropy value, where the random noise is also utilized to boost effectiveness of the regularization. The experimental results showed that the proposed method can significantly improve the robustness of the state-of-the-art deep super-resolution models without significant degradation of the original performance. We also showed the synergy when our method is combined with adversarial training.

Acknowledgement

This work was supported by the NRF grant funded by the Korea government (MSIT) (NRF-2020R1F1A1070631), and the Artificial Intelligence Graduate School Program (Yonsei University, 2020-0-01361).

References

1. Yue, L., Shen, H., Li, J., Yuan, Q., Zhang, H., Zhang, L.: Image super-resolution: The techniques, applications, and future. Signal Processing **128** (2016) 389–408
2. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: Proceedings of the International Conference on Learning Representations. (2015)
3. Su, D., Zhang, H., Chen, H., Yi, J., Chen, P.Y., Gao, Y.: 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) 631–648
4. Choi, J.H., Zhang, H., Kim, J.H., Hsieh, C.J., Lee, J.S.: Evaluating robustness of deep image super-resolution against adversarial attacks. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 303–311
5. Yin, M., Zhang, Y., Li, X., Wang, S.: When deep fool meets deep prior: Adversarial attack on super-resolution network. In: Proceedings of the ACM International Conference on Multimedia. (2018) 1930–1938
6. Das, N., Shanbhogue, M., Chen, S.T., Hohman, F., Li, S., Chen, L., Kounavis, M.E., Chau, D.H.: Shield: Fast, practical defense and vaccination for deep learning using JPEG compression. In: Proceedings of the ACM International Conference on Knowledge Discovery & Data Mining. (2018) 196–204
7. Huang, R., Xu, B., Schuurmans, D., Szepesvári, C.: Learning with a strong adversary. In: Proceedings of the International Conference on Learning Representations. (2016)
8. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial machine learning at scale. In: Proceedings of the International Conference on Learning Representations. (2017)
9. Xie, C., Wang, J., Zhang, Z., Ren, Z., Yuille, A.: Mitigating adversarial effects through randomization. In: Proceedings of the International Conference on Learning Representations. (2018)
10. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Proceedings of the European Conference on Computer Vision. (2014) 184–199
11. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: Proccedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2017) 136–144
12. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision. (2018) 286–301
13. Li, J., Fang, F., Mei, K., Zhang, G.: Multi-scale residual network for image super-resolution. In: Proceedings of the European Conference on Computer Vision. (2018) 517–532
14. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A.P., Tejani, A., Totz, J., Wang, Z., 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) 4681–4690
15. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proceedings of the Advances in Neural Information Processing Systems. (2014) 2672–2680
16. Ahn, N., Kang, B., Sohn, K.A.: Fast, accurate, and lightweight super-resolution with cascading residual network. In: Proceedings of the European Conference on Computer Vision. (2018) 252–268

17. Choi, J.H., Kim, J.H., Cheon, M., Lee, J.S.: Deep learning-based image super-resolution considering quantitative and perceptual quality. Neurocomputing **398** (2020) 347–359
18. Deng, X., Yang, R., Xu, M., Dragotti, P.L.: Wavelet domain style transfer for an effective perception-distortion tradeoff in single image super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 3076–3085
19. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: Proceedings of the International Conference on Learning Representations. (2014)
20. Ganeshan, A., Babu, R.V.: FDA: Feature disruptive attack. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 8069–8079
21. Weng, T.W., Zhang, H., Chen, P.Y., Yi, J., Su, D., Gao, Y., Hsieh, C.J., Daniel, L.: Evaluating the robustness of neural networks: An extreme value theory approach. In: Proceedings of the International Conference on Learning Representations. (2018)
22. Mustafa, A., Khan, S.H., Hayat, M., Shen, J., Shao, L.: Image super-resolution as a defense against adversarial attacks. IEEE Transactions on Image Processing **29** (2019) 1711–1724
23. Ballé, J., Minnen, D., Singh, S., Hwang, S.J., Johnston, N.: Variational image compression with a scale hyperprior. In: Proceedings of the International Conference on Learning Representations. (2018)
24. Magdon-Ismail, M., Atiya, A.: Density estimation and random variate generation using multilayer networks. IEEE Transactions on Neural Networks **13** (2002) 497–520
25. Magdon-Ismail, M., Atiya, A.F.: Neural networks for density estimation. In: Proceedings of the Advances in Neural Information Processing Systems. (1999) 522–528
26. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 1026–1034
27. Agustsson, E., Timofte, R.: 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) 126–135
28. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: ImageNet large scale visual recognition challenge. International Journal of Computer Vision **115** (2015) 211–252
29. Bevilacqua, M., Roumy, A., Guillemot, C., Alberi-Morel, M.L.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In: Proceedings of the British Machine Vision Conference. (2012) 1–10
30. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: Proceedings of the International Conference on Curves and Surfaces. (2010) 711–730
31. Martin, D., Fowlkes, C., Tal, D., Malik, J.: 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) 416–423