Perception-Enhanced Image Super-Resolution via Relativistic Generative Adversarial Networks

Thang Vu, Tung M. Luu, and Chang D. Yoo

Department of Electrical Engineering,
Korea Advanced Institute of Science and Technology (KAIST)
{thangvubk,tungluu2203,cd_yoo}@kaist.ac.kr

Abstract. This paper considers a deep Generative Adversarial Networks (GAN) based method referred to as the Perception-Enhanced Super-Resolution (PESR) for Single Image Super Resolution (SISR) that enhances the perceptual quality of the reconstructed images by considering the following three issues: (1) ease GAN training by replacing an absolute with a relativistic discriminator, (2) include in the loss function a mechanism to emphasize difficult training samples which are generally rich in texture and (3) provide a flexible quality control scheme at test time to trade-off between perception and fidelity. Based on extensive experiments on six benchmark datasets, PESR outperforms recent state-of-the-art SISR methods in terms of perceptual quality. The code is available at https://github.com/thangvubk/PESR.

Keywords: Super-resolution, perceptual quality

1 Introduction

In recent years, Single Image Super Resolution (SISR) has received considerable attention for its applications that includes surveillance imaging [1,2], medical imaging [3,4] and object recognition [5,6]. Given a low-resolution image (LR), SISR aims to reconstruct a super-resolved image (SR) that is as similar as possible to the original high-resolution image (HR). This is an ill-posed problem since there are many possible ways to generate SR from LR.

Recent example-based methods using deep convolutional neural networks (CNNs) have achieved significant performance. However, most of the methods aim to maximize peak-signal-rate-ratio (PSNR) between SR and HR, which tends to produce blurry and overly-smoothed reconstructions. In order to obtain non-blurry and realistic reconstruction, this paper considers the following three issues. First, standard GAN [7] (SGAN) based SISR methods which are known to be effective in reconstructing natural images are notoriously difficult to train and unstable. One reason might be attributed to the fact that the generator is generally trained without taking real high-resolution images into account. Second, texture-rich high-resolution samples that are generally difficult to reconstruct from low-resolution images should be emphasized during training.

EDSR [8], state-of-the-art PSNR Our PESR

Fig. 1. Super-resolution result comparison on image *lenna* from Set14 dataset. Our method exhibits more convincing textures and perceptual quality compared to those of the state-of-the-art PSNR-based method

Third, trading-off between PSNR and perceptual quality at test time with existing methods is impossible without retraining. Exiting methods are commonly trained to improve either PSNR or perceptual quality, and depending on the application, one objective might be better than the other.

To address these issues, this paper proposes a GAN based SISR method referred to as the Perception-Enhanced Super-Resolution (PESR) that aims to enhance the perceptual quality of reconstruction and to allow users to flexibly control the perceptual degree at test time. In order to improve GAN performance, PESR is trained to minimize relativistic loss instead of an absolute loss. While SGAN aims to generate data that looks real, the PESR attempts to generate fake data to be more real than real data. This philosophy is extensively studied in [9] with Relativistic GAN (RGAN). In PESR, valuable texture-rich samples are emphasized in training. It is observed that the texture-rich patches, which play an important role in user-perceived quality, are more difficult to reconstruct and play an important role in user-perceived quality. In training PESR, easy examples with smooth texture are deemphasized by combining GAN loss with a focal loss function. Furthermore, at test time, we proposed a quality-control mechanism. The perceptual degree is controlled by interpolating between a perception-optimized model and a distortion-optimized model. Experiment results show that the proposed PESR achieves significant improvements compared to other state-of-the-art SISR methods.

The rest of this paper is organized as follows. Section 2 reviews various SISR methods. Section 3 presents the proposed networks and the loss functions to train the networks. Section 4 presents extensive experiments results on six benchmark datasets. Finally, Section 5 summarizes and concludes the paper.

2 Related Work

2.1 Single Image Super-Resolution

To address the super-resolution problem, early methods are mostly based on interpolation such as bilinear, bicubic, and Lancroz [10]. These methods are simple and fast but usually produce overly-smoothed reconstructions. To mitigate this problem, some edge-directed interpolation methods have been proposed [11,12]. More advanced methods such as dictionary learning [13,14,15,16], neighborhood embedding [17,18,19] and regression trees [20,21] aim to learn complex mapping between low- and high-resolution image features. Although these methods have shown better results compared to their predecessors, their performances compared to that of recent deep architectures leave much to be desired.

Deep architectures have made great strides in SISR. Dong *et al.* [22,23] first introduced SRCNN for learning the LR-HR mapping in an end-to-end manner. Although SRCNN is only a three-convolutional-layer network, it outperformed previous methods. As expected, SISR also benefits from very deep networks. The 5-layer FSRCNN [24], 20-layer VDSR [25], and 52-layer DRRN [26] have shown significant improvements in terms of accuracy. Lim *et al.* [8] proposed a very deep modified ResNet [27] to achieve state-of-the-art PSNR performance.

Beside building very deep networks, utilizing advanced deep learning techniques lead to more robust, stable, and compact networks. Kim *et al.* [25] introduced residual learning for SISR showing promising results just by predicting residual high-frequency components in SISR. Tai *et al.* [26] and Kim *et al.* [28] investigated recursive networks in SISR, which share parameters among recursive blocks and show superior performance with fewer parameters compared to previous work. Densely connected networks [29] have also shown to be conducive for SISR [30,31].

2.2 Loss Functions

The most common loss function to maximize PSNR is the mean-squared error (MSE). Other losses such as L1 or Charbonnier (a differentiable variant of L1) have also been studied to improve PSNR. It is well-known that pixel-wise loss functions produce blurry and overly-smoothed output as a result of averaging all possible solutions in the pixel space. As shown in Figure 1, the natural textures are missing even in the state-of-the-art PSNR-based method. In [32], Zhao *et al.* studied Structural Similarity (SSIM) and its variants as a measure for evaluating the quality of the reconstruction in SISR. Although SSIM takes the image structure into account, this approach exposes the limitation in recovering realistic textures.

Instead of using pixel-wise errors, high-level feature distance has been considered for SISR [33,34,5,35]. The distance is measured based on the feature maps which are extracted using a pre-trained VGG network [36]. Blau *et al.* [37] demonstrated that the distance between VGG features are well correlated to human opinion based quality assessment. Relying on the VGG features, a number

of perceptual loss functions have been proposed. Instead of measuring the Euclidean distance between the VGG features, Sajjadi *et al.* [5] proposed a Gram loss function which exploits correlations between feature activations. Meanwhile, Mechrez *et al.* [35] introduced contextual loss, which aims to maintain natural statistics of images.

To enhance training computational efficiency, images are cropped into multiple small patches. However, training samples are usually dominated by a large number of easily reconstructable patches. When these easy samples overwhelm the generator, reconstructed results tend to be blurry and smooth. This is analogous to an observation in dense object detection [38], where the background samples overwhelm the detector. Focal loss which emphasizes difficult examples should be considered for SISR.

2.3 Adversarial Learning

Ever since it was first proposed by Goodfellow *et al.*, GANs [7] have been incorporated for various tasks such as image generation, style transfer, domain adaptation, and super-resolution. The general idea of GANs is that it allows training a generative model G to produce real-like fake data with the goal of fooling a discriminator D while D is trained to distinguish between the generated data and real data. The generator G and the discriminator D compete in an adversarial manner with each other to achieve their individual objectives; thus, the generator mimics the real data distribution. In SISR, adversarial loss was introduced by Ledig *et al.* [34], generating images with convincing textures. Since then, GANs have emerged as the most common architecture for generating photo-realistic SISR [35,5,39,40,41]. Wang *et al.* [41] proposed a conditional GAN for SISR, where the semantic segmentation probability maps are exploited as the prior. Yuan *et al.* [40] investigated the use of cycle-in-cycle GANs for SISR, where HR labels are not available and LR images further degraded by noise, showing promising results. In a recent study, Blau *et al.* [37] have demonstrated that GANs provide a principle way to enhance perceptual quality for SISR.

2.4 Contribution

The four main contributions of this paper are as follows:

1. We demonstrate that stabilizing GAN training plays a key role in enhancing perceptual quality for SISR. When GAN performance is improved, the generated images are closer to natural manifolds.
2. We replace SGAN by RGAN loss function to fully utilize data at training time. A focal loss is used to emphasize valuable examples. The total variance loss is also added to mitigate high-frequency noise amplification of adversarial training.
3. We propose a quality control scheme at test time that allows users to adaptively emphasize between the perception and fidelity.

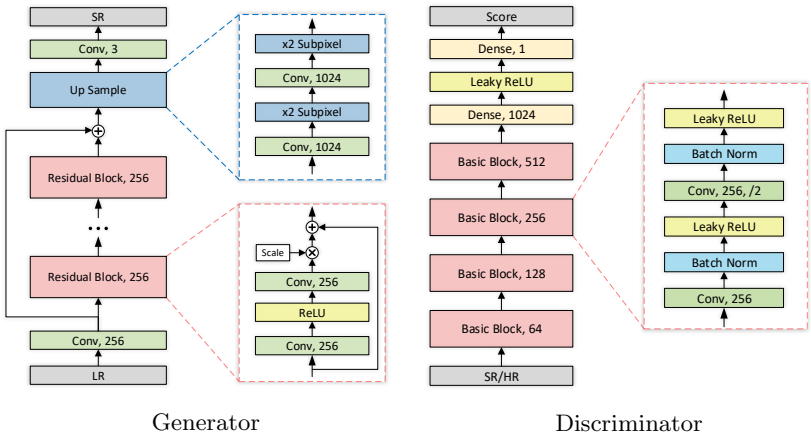

Fig. 2. Architecture of Generator and Discriminator networks.

4. We evaluate the proposed method using recently-proposed quality metric [37] that encourages the SISR prediction to be close to natural manifold. We quantitatively and qualitatively show that the proposed method achieves better perceptual quality compared to other state-of-the-art SISR algorithms.

3 Proposed method

3.1 Network Architecture

The proposed PESR method utilizes the SRGAN architecture [34] with its generator replaced by the EDSR [8]. As shown in Figure 2, a low-resolution image is first embedded by a convolutional layer, before being fed into a series of 32 residual blocks. The spatial dimensions of the residual blocks are maintained until the very end of the generator such that the computational cost is kept low. The output of the 32 residual blocks is summed with the embedded input. Then it is upsampled to the high-resolution space, after which it is reconstructed.

The discriminator is trained to discriminate between generated and real high-resolution image. An image is fed into four basic blocks, each of which contains two convolutional layers followed by batch normalization and leaky ReLU activations. After the four blocks, a binary classifier, which consists of two dense layers, predicts whether the input is generated or real.

The generator and discriminator are trained by alternating gradient update based on their individual objectives which are denoted as $\mathcal{L}_G$ and $\mathcal{L}_D$ respectively. To enhance the stability and improve texture rendering, the generator loss is a linear sum of three loss functions: focal RGAN loss $\mathcal{L}_{FRG}$, content loss $\mathcal{L}_C$, and total variance loss $\mathcal{L}_{TV}$, shown as below:

$$\mathcal{L}_G = \alpha_{FRG}\mathcal{L}_{FRG} + \alpha_C\mathcal{L}_C + \alpha_{TV}\mathcal{L}_{TV}. \quad (1)$$

Here α_{FRG}, α_C, and α_{TV} are trade-off parameters. The three loss functions are described in more detail in the following subsections.

3.2 Loss Functions

Focal RGAN Loss. In the GAN setting, the input and output of the generator and the real samples are respectively the low-resolution image I^{LR}, generated super-resolved image I^{SR} and the original high-resolution image I^{HR}. As in SGAN, a generator G_θ and a discriminator D_φ are trained to optimize a min-max problem:

$$\min_{\theta}\max_{\varphi}\mathbb{E}_{I^{HR}\sim\mathbb{P}^{HR}}\log D_\varphi(I^{HR})+\mathbb{E}_{I^{LR}\sim\mathbb{P}^{LR}}\log(1-D_\varphi(G_\theta(I^{LR}))). \tag{2}$$

Here $\mathbb{P}^{HR}$ and $\mathbb{P}^{LR}$ are the distributions of real data (original high-resolution image) and fake data (low-resolution image), respectively. This min-max problem can be interpreted as minimizing explicit loss functions for the generator and the discriminator $\mathcal{L}_{SG}$ and $\mathcal{L}_{SD}$ respectively as follows:

$$\mathcal{L}_{SG}=-\mathbb{E}_{I^{LR}\sim\mathbb{P}^{LR}}\log(D_\varphi(G_\theta(I^{LR})), \tag{3}$$

and

$$\mathcal{L}_{SD}=-\mathbb{E}_{I^{HR}\sim\mathbb{P}^{HR}}\log D_\varphi(I^{HR})-\mathbb{E}_{I^{LR}\sim\mathbb{P}^{LR}}\log(1-D_\varphi(G_\theta(I^{LR})). \tag{4}$$

It is well known that SGAN is notoriously difficult and unstable to train, which results in low reconstruction performance. Furthermore, Eq. 3 shows that the generator loss function does not explicitly depend on I^{HR}. In other words, the SGAN generator completely ignores high-resolution image in its updates. Instead, the loss functions of both generator and discriminator should exploit the information provided by both the high-resolution and fidelity of the synthesized image. The proposed method considers relative discriminative score between the I^{HR} and I^{SR} such that training is easier. This can be achieved by increasing the probability of classifying the generated high-resolution image as being real and simultaneously decreasing the probability of classifying the original high-resolution image as being real. Inspired by RGAN [9], the following loss functions for the generator and discriminator can be considered,

$$\mathcal{L}_{RG}=-\mathbb{E}_{(I^{LR},I^{HR})\sim(\mathbb{P}^{LR},\mathbb{P}^{HR})}\log\left[\sigma(C_\varphi(G_\theta(I^{LR}))-C_\varphi(I^{HR}))\right], \tag{5}$$

and

$$\mathcal{L}_{RD}=-\mathbb{E}_{(I^{LR},I^{HR})\sim(\mathbb{P}^{LR},\mathbb{P}^{HR})}\log\left[\sigma(C_\varphi(I^{HR})-C_\varphi(G_\theta(I^{LR})))\right]. \tag{6}$$

Here C_φ which is referred to as the critic function [42] is taken before the last sigmoid function σ of the discriminator.

The generator loss can be further enhanced to emphasize texture-rich patches which tend to be difficult samples to reconstruct with high loss $\mathcal{L}_{RG}$. Emphasizing difficult samples and down-weighting easy samples will lead to better texture

reconstruction. This can be achieved by minimizing the focal function with a focusing parameter of γ:

$$\mathcal{L}_{FRG} = -\sum_i (1-p_i)^\gamma \log(p_i), \tag{7}$$

where $p_i = \sigma(C_\varphi(G_\theta(I_i^{LR})) - C_\varphi(I_i^{HR}))$.

Content Loss. Beside enhancing realistic textures, the reconstructed image should be similar to the original high-resolution image which is ground truth. Instead of considering pixel-wise accuracy, perceptual loss that measures distance in a high-level feature space [33] is considered. The feature map, denoted as ϕ, is obtained by using a pre-trained 19-layer VGG network. Following [34], the feature map is extracted right before the fifth max-pooling layer. The content loss function is defined as,

$$\mathcal{L}_C = \sum_i \|\phi(I_i^{HR}) - \phi(I_i^{SR})\|_2^2. \tag{8}$$

Total Variance Loss. High-frequency noise amplification is inevitable with GAN based synthesis, and in order to mitigate this problem, the total variance loss function [43] is considered. It is defined as

$$\mathcal{L}_{TV} = \sum_{i,j,k} \left(\left| I_{i,j+1,k}^{SR} - I_{i,j,k}^{SR} \right| + \left| I_{i,j,k+1}^{SR} - I_{i,j,k}^{SR} \right| \right). \tag{9}$$

4 Experiments

4.1 Dataset

The proposed networks are trained on DIV2K dataset [44], which consists of 800 high-quality (2K resolution) images. For testing, 6 standard benchmark datasets are used, including Set5 [17], Set14 [16], B100 [45], Urban100 [46], DIV2K validation set [44], and PIRM self-validation set [47].

4.2 Evaluation Metrics

To demonstrate the effectiveness of PESR, we measure GAN training performance and SISR image quality. The Fréchet Inception Distance (FID) [48] is used to measure GAN performance, where lower FID values indicate better image quality. In FID, feature maps $\boldsymbol{\psi}(I)$ are obtained by extracting the *pool_3* layer of a pre-trained Inception V3 model [49]. Then, the extracted features are modeled under a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. The FID $d(\boldsymbol{\psi}(I^{SR}), \boldsymbol{\psi}(I^{HR}))$ between generated features $\boldsymbol{\psi}(I^{SR})$ and real features $\boldsymbol{\psi}(I^{HR})$ is given by [50]:

$$d^2(\boldsymbol{\psi}(I^{SR}), \boldsymbol{\psi}(I^{HR})) = \left\|\boldsymbol{\mu}^{SR} - \boldsymbol{\mu}^{HR}\right\|_2^2 + \mathrm{Tr}\left(\boldsymbol{\Sigma}^{SR} + \boldsymbol{\Sigma}^{HR} - 2\left(\boldsymbol{\Sigma}^{SR}\boldsymbol{\Sigma}^{HR}\right)^{1/2}\right). \tag{10}$$

To evaluate SISR performance, we use a recently-proposed perceptual metric in [37]:

$$\text{Perceptual index} = \frac{(10 - \text{NRQM}) + \text{NIQE}}{2}, \tag{11}$$

where NRQM and NIQE are the quality metrics proposed by Ma *et al.* [51] and Mittal *et al.* [52], respectively. The lower perceptual indexes indicate better perceptual quality. It is noted that the perceptual index in Eq. 11 is a non-reference metric, which does not reflect the distortion of SISR results. Therefore, the conventional PSNR metric is also used as a distortion reference.

4.3 Experiment Settings

Throughout the experiments, LR images are obtained by bicubically down-sampling HR images with a scaling factor of $\times 4$ using MATLAB *imresize* function. We pre-process all the images by subtracting the mean RGB value of the DIV2K dataset. At training time, to enhance computational efficiency, the LR and HR images are cropped into patches of size 48×48 and 196×194, respectively. It is noted that our generator network is fully convolutional; thus, it can take arbitrary size input at test time.

We train our networks with Adam optimizer [53] with setting $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. Batchsize is set to 16. We initialize the generator using L1 loss for 2×10^5 iterations, then alternately optimize the generator and discriminator with our full loss for other 2×10^5 iterations. The trade-off parameter for the loss function is set to $\alpha_{FRG} = 1, \alpha_C = 50$ and $\alpha_{TV} = 10^{-6}$. We use a focusing parameter of 1 for the focal loss. The learning rate is initialized to 10^{-4} for pretraining and 5×10^{-5} for GAN training, which is halved after 1.2×10^5 batch updates.

Our model is implemented using Pytorch [54] deep learning framework, which is run on Titan Xp GPUs and it takes 20 hours for the networks to converge.

4.4 GAN Performance Measurement

To avoid underestimated FID values of the generator, the number of samples should be at least 10^4 [48], hence the images are cropped into patches of 32×32. The proposed method is compared with standard GAN (SGAN) [7], least-squares GAN (LSGAN) [55], Hinge-loss GAN (HingeGAN) [56], and Wassertein GAN improved (WGAN-GP) [57]. All the considered GANs are combined with the content and total variance losses. Table 1 shows that LSGAN performs the worst at FID of 18.5. HingeGAN, WGAN-GP, and SGAN show better results compared to LSGAN. Our method relied on RGAN shows the best performance.

4.5 Ablation Study

The effectiveness of the proposed method is demonstrated using an ablation analysis. As reported in Table 2, the perceptual index of L1 loss training is

Table 1. FID comparison of RGAN with other GANs on DIV2K validation set.

SGAN	LSGAN	HingeGAN	WGAN-GP	RGAN
6.83	18.5	6.97	7.02	6.63

limited to 5.41, and after training with the VGG content loss, the performance is improved dramatically to 3.32. When adversarial training (RGAN) is added, the performance is further improved to 2.28. The total variance loss and focal loss show slightly perceptual index improvement. The proposed method with the default setting (e) obtains the best performance of 2.25.

The effect of each component in the proposed loss function is also visually compared in Figure 3. As expected, L1 loss shows blurry and overly-smooth images. Although VGG loss improves perceptual quality, the reconstruction results are still unnatural since they expose square patterns. When RGAN is added, the reconstruction results are more visually pleasing with more natural texture and edges, and no square patterns are observed.

Fig. 3. Effect of each component in our loss function on B100 dataset (images 163085, 38082, 19021, 351093 from top to bottom rows). Each column from (a) to (e) represents the setting described in Table 2

Table 2. Ablation analysis in terms of perceptual index on B100 dataset.

Setting	L1	VGG	TV	RGAN	Focal	PI
(a)	✓					5.41
(b)		✓				3.32
(c)		✓	✓			3.31
(d)		✓	✓	✓		2.28
(e) default		✓	✓	✓	✓	2.25

4.6 Comparison with State-of-the-art SISR Methods

In this subsection, we quantitatively and qualitatively compare our PESR with other state-of-the-art SISR algorithms. Here, PESR is benchmarked against SRCNN [23], VDSR [25], DRCN [28], EDSR [8], SRGAN [34], ENET [5], and CX [35]. The performance of bicubic interpolation is also reported as the baseline. The results of SRGAN is obtained from a Tensorflow implementation[1]. For CX, the source codes for super-resolution task was unavailable; however, the authors of CX provided the generated images at our request. For the others methods, the results were obtained using publicly available source codes.

Table 3. Perceptual index comparison of the proposed PESR with recent state-of-the-art SISR methods. RED and BLUE indicate best and second best results, respectively.

Dataset	Set5	Set14	B100	Urban100	PIRM2018	DIV2K
Bicubic	7.32	6.97	6.94	6.88	6.80	6.94
SRCNN [23]	6.79	6.03	6.04	5.94	5.94	5.92
VDSR [25]	6.45	5.77	5.70	5.54	5.65	5.62
DRCN [28]	6.45	5.94	5.89	5.79	5.77	5.71
EDSR [8]	6.00	5.52	5.40	5.14	5.08	5.37
SRGAN [34]	3.18	2.80	2.59	3.30	2.30	3.30
ENET [5]	2.93	3.02	2.91	3.47	2.69	3.50
CX [35]	3.29	2.76	2.25	3.39	2.13	3.16
PESR (ours)	3.42	2.66	2.25	3.41	2.13	3.13

[1] https://github.com/tensorlayer/srgan

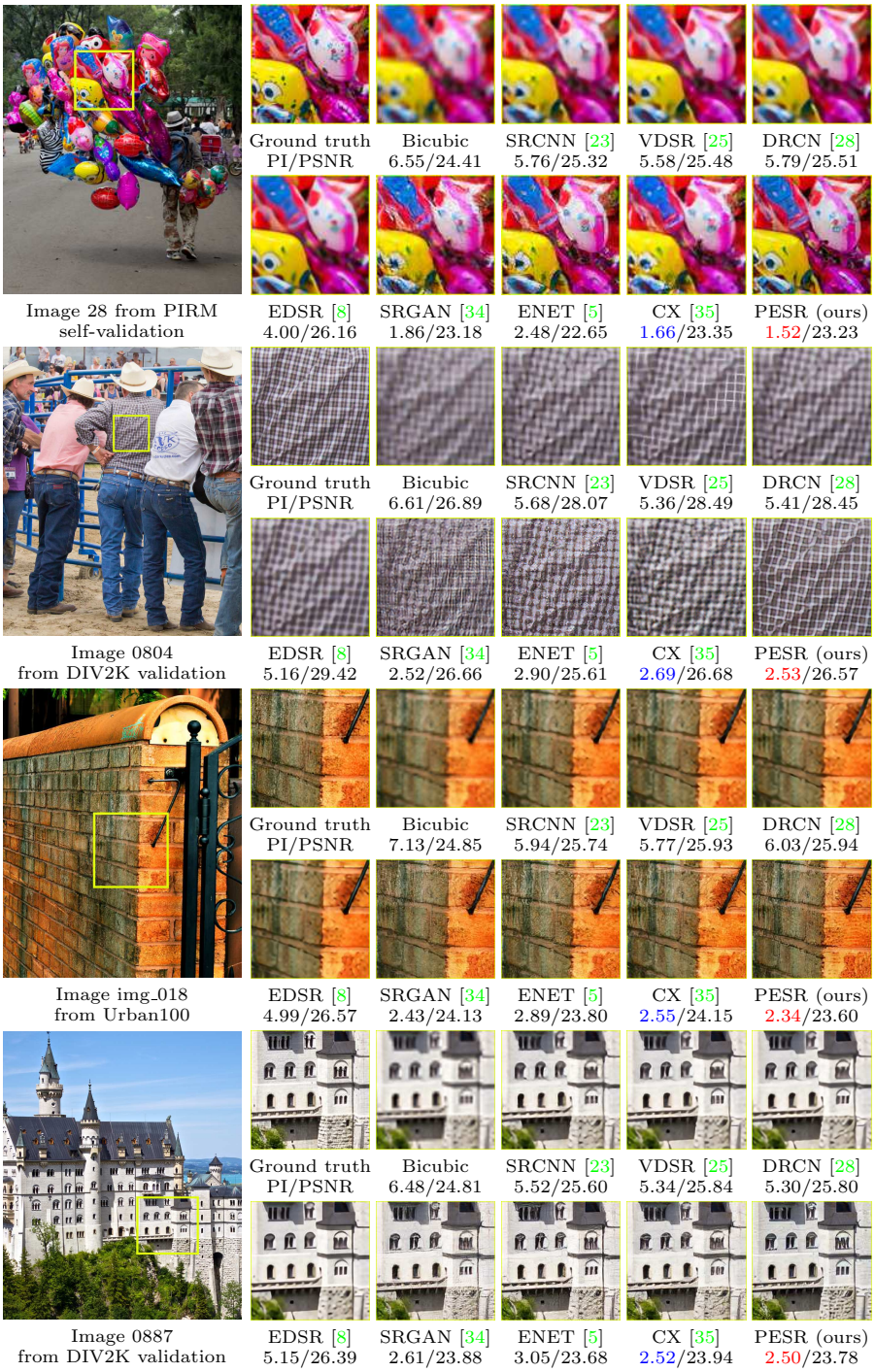

Fig. 4. Qualitative comparison between our PESR and the others. RED and BLUE indicate best and second best perceptual index.

Quantitative results Table 3 illustrates the perceptual indexes of PESR and the other seven state-of-the-art SISR methods. As expected, GAN-based methods, including SRGAN [34], ENET [5], CX [35], and the proposed PESR, outperform the PSNR-based methods in term of perceptual index with a large margin. Here, SRGAN and ENET methods have the best results in Set5 and Urban100 dataset, respectively; however, their performances are relatively limited in the other datasets. It is noted that ENET are trained on 200k images, which is much more than those of other methods (at most 800 images). Our PESR achieves the best performance in 4 out of 6 benchmark datasets.

Qualitative results. The visual comparison of our PESR with other state-of-the-art SISR methods are illustrated in Figure 4. Overall, PSNR-based methods produce blurry and smooth images while GAN-based methods synthesize a more realistic texture. However, SGRAN, ENET, and CX exhibit limitation when the textures are densely and structurally repeated as in image 0804 from DIV2K dataset. Meanwhile, our PESR provides sharper and more natural textures compared to the others.

4.7 Perception-Distortion Control at Test Time

In a number of applications such as medical imaging, synthesized textures are not desirable. To make our model robust and flexible, we proposed a quality control scheme that interpolates between a perception-optimized model G_{θ_P} and a distortion-optimized model G_{θ_D}. The G_{θ_P} and G_{θ_D} models are obtained by training our network with the full loss function and L1 loss function, respectively. The perceptual quality degree is controlled by adjusting the parameter λ in the following equation:

$$I^{SR} = \lambda G_{\theta_P}(I^{LR}) + (1 - \lambda) G_{\theta_D}(I^{LR}). \tag{12}$$

Here, the networks attempt to predict the most accurate results when $\lambda = 0$ and synthesize the most perceptually-plausible textures when $\lambda = 1$.

We demonstrate that flexible SISR method is effective in a number of cases. In Figure 5, two types of textures are presented: a wire entanglement with sparse textures, and shutter with dense textures. The results show that high perceptual quality weights provide more plausible visualization for the dense textures while reducing the weight seems to be pleasing for the easy ones. We also compare our interpolated results and the others, as shown in Figure 6. It is clear that we can obtain better perceptual quality with the same PSNR, and vice versa, compared to the other methods.

4.8 PIRM 2018 Challenge

The Perceptual Image Restoration and Manipulation (PIRM) 2018 challenge aims to produce images that are visually appealing to human observers. The

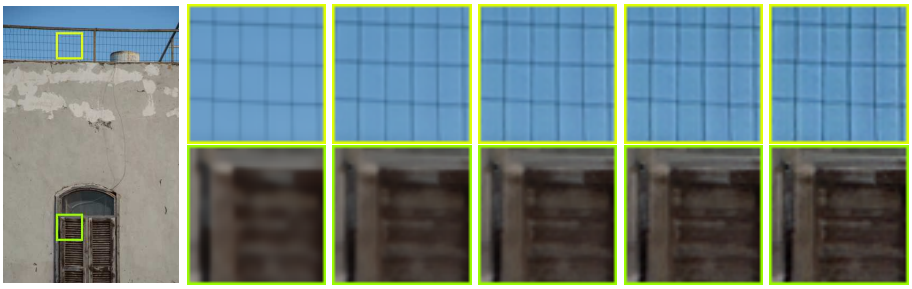

Fig. 5. Perception-distortion trade-off with different perceptual quality weights.

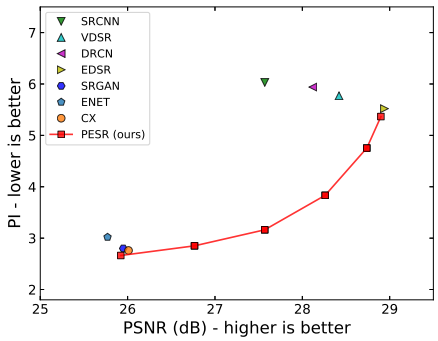

Fig. 6. Our interpolated results in comparison with the others on Set14 dataset. Left- and right-most triangle markers indicate λ being 1 and 0, respectively.

authors participated in the Super-resolution challenge to improve perceptual quality while constraining the root-mean-squared error (RMSE) to be less than 11.5 (region 1), between 11.5 to 12.5 (region 2) and between 12.5 and 16 (region 3).

Our main target is region 3, which aims to maximize the perceptual quality. We ranked 4th with perceptual index 0.04 lower than the top-ranking teams. For region 1 and 2, we use interpolated results without any fine-tuning and ranked 5th and 6th, respectively. We believe further improvements can be achieved with fine-tuning and more training data.

5 Conclusion

We have presented a deep Generative Adversarial Network (GAN) based method referred to as the Perception-Enhanced Super-Resolution (PESR) for Single Image Super Resolution (SISR) that enhances the perceptual quality of the reconstructed images by considering the following three issues: (1) ease GAN training

by replacing an absolute by relativistic discriminator (2) include in a loss function a mechanism to emphasize difficult training samples which are generally rich in texture, and (3) provide a flexible quality control scheme at test time to trade-off between perception and fidelity. Each component of proposed method is demonstrated to be effective through the ablation analysis. Based on extensive experiments on six benchmark datasets, PESR outperforms recent state-of-the-art SISR methods in terms of perceptual quality.

References

1. Zou, W.W., Yuen, P.C.: Very low resolution face recognition problem. IEEE Transactions on Image Processing **21**(1) (2012) 327–340
2. Jiang, J., Ma, J., Chen, C., Jiang, X., Wang, Z.: Noise robust face image super-resolution through smooth sparse representation. IEEE transactions on cybernetics **47**(11) (2017) 3991–4002
3. Shi, W., Caballero, J., Ledig, C., Zhuang, X., Bai, W., Bhatia, K., de Marvao, A.M.S.M., Dawes, T., ORegan, D., Rueckert, D.: Cardiac image super-resolution with global correspondence using multi-atlas patchmatch. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2013) 9–16
4. Ning, L., Setsompop, K., Michailovich, O., Makris, N., Shenton, M.E., Westin, C.F., Rathi, Y.: A joint compressed-sensing and super-resolution approach for very high-resolution diffusion imaging. NeuroImage **125** (2016) 386–400
5. Sajjadi, M.S., Schölkopf, B., Hirsch, M.: Enhancenet: Single image super-resolution through automated texture synthesis. In: Computer Vision (ICCV), 2017 IEEE International Conference on, IEEE (2017) 4501–4510
6. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. arXiv preprint arXiv:1807.02758 (2018)
7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. (2014) 2672–2680
8. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: The IEEE conference on computer vision and pattern recognition (CVPR) workshops. Volume 1. (2017) 4
9. Jolicoeur-Martineau, A.: The relativistic discriminator: a key element missing from standard GAN. ArXiv e-prints (July 2018)
10. Duchon, C.E.: Lanczos filtering in one and two dimensions. Journal of applied meteorology **18**(8) (1979) 1016–1022
11. Allebach, J., Wong, P.W.: Edge-directed interpolation. In: Image Processing, 1996. Proceedings., International Conference on. Volume 3., IEEE (1996) 707–710
12. Li, X., Orchard, M.T.: New edge-directed interpolation. IEEE transactions on image processing **10**(10) (2001) 1521–1527
13. Wang, S., Zhang, L., Liang, Y., Pan, Q.: Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 2216–2223
14. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. IEEE transactions on image processing **19**(11) (2010) 2861–2873

15. Yang, J., Wang, Z., Lin, Z., Cohen, S., Huang, T.: Coupled dictionary training for image super-resolution. IEEE transactions on image processing **21**(8) (2012) 3467–3478
16. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: International conference on curves and surfaces, Springer (2010) 711–730
17. Bevilacqua, M., Roumy, A., Guillemot, C., Alberi-Morel, M.L.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding. (2012)
18. Timofte, R., De Smet, V., Van Gool, L.: Anchored neighborhood regression for fast example-based super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision. (2013) 1920–1927
19. Timofte, R., De Smet, V., Van Gool, L.: A+: Adjusted anchored neighborhood regression for fast super-resolution. In: Asian Conference on Computer Vision, Springer (2014) 111–126
20. Salvador, J., Perez-Pellitero, E.: Naive bayes super-resolution forest. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 325–333
21. Schulter, S., Leistner, C., Bischof, H.: Fast and accurate image upscaling with super-resolution forests. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 3791–3799
22. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence **38**(2) (2016) 295–307
23. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: European conference on computer vision, Springer (2014) 184–199
24. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: European Conference on Computer Vision, Springer (2016) 391–407
25. Kim, J., Kwon Lee, J., Mu Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 1646–1654
26. Tai, Y., Yang, J., Liu, X.: Image super-resolution via deep recursive residual network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Volume 1. (2017) 5
27. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
28. Kim, J., Kwon Lee, J., Mu Lee, K.: Deeply-recursive convolutional network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 1637–1645
29. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR. Volume 1. (2017) 3
30. Tong, T., Li, G., Liu, X., Gao, Q.: Image super-resolution using dense skip connections. In: Computer Vision (ICCV), 2017 IEEE International Conference on, IEEE (2017) 4809–4817
31. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018)
32. Zhao, H., Gallo, O., Frosio, I., Kautz, J.: Loss functions for image restoration with neural networks. IEEE Transactions on Computational Imaging **3**(1) (2017) 47–57

33. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision, Springer (2016) 694–711
34. 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: CVPR. Volume 2. (2017) 4
35. Mechrez, R., Talmi, I., Shama, F., Zelnik-Manor, L.: Learning to maintain natural image statistics. arXiv preprint arXiv:1803.04626 (2018)
36. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
37. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: CVPR. (2018)
38. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. IEEE transactions on pattern analysis and machine intelligence (2018)
39. Wang, Y., Perazzi, F., McWilliams, B., Sorkine-Hornung, A., Sorkine-Hornung, O., Schroers, C.: A fully progressive approach to single-image super-resolution. In: CVPR. (2018)
40. Yuan12, Y., Liu134, S., Zhang, J., Zhang, Y., Dong, C., Lin, L.: Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In: CVPR. (2018)
41. Wang, X., Yu, K., Dong, C., Loy, C.C.: Recovering realistic texture in image super-resolution by deep spatial feature transform. In: CVPR. (2018)
42. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International Conference on Machine Learning. (2017) 214–223
43. Aly, H.A., Dubois, E.: Image up-sampling using total-variation regularization with a new observation model. IEEE Transactions on Image Processing **14**(10) (2005) 1647–1659
44. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: CVPRW. Volume 3. (2017) 2
45. 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: ICCV. Volume 2., IEEE (2001) 416–423
46. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 5197–5206
47. Blau, Y., Mechrez, R., Timofte, R., Michaeli, T., Zelnik-Manor, L.: 2018 PIRM Challenge on Perceptual Image Super-resolution. ArXiv e-prints (September 2018)
48. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems. (2017) 6626–6637
49. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 2818–2826
50. Dowson, D., Landau, B.: The fréchet distance between multivariate normal distributions. Journal of multivariate analysis **12**(3) (1982) 450–455
51. Ma, C., Yang, C.Y., Yang, X., Yang, M.H.: Learning a no-reference quality metric for single-image super-resolution. Computer Vision and Image Understanding **158** (2017) 1–16
52. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a" completely blind" image quality analyzer. IEEE Signal Process. Lett. **20**(3) (2013) 209–212

53. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR. (2014)
54. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. (2017)
55. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Smolley, S.P.: Least squares generative adversarial networks. In: Computer Vision (ICCV), 2017 IEEE International Conference on, IEEE (2017) 2813–2821
56. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957 (2018)
57. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: Advances in Neural Information Processing Systems. (2017) 5767–5777