ICASSP 2022 Poster #2710

Learning Adjustable Image Rescaling with Joint Optimization of Perception and Distortion

Zhihong Pan

Baidu Research (USA)

zhihongpan@baidu.com

Abstract

The performance of image super-resolution (SR) have been greatly advanced by deep learning techniques recently. Most models are only optimized for the ill-posed upscaling task while assuming a predefined downscaling kernel for low-resolution (LR) inputs. Additionally, there exists a conflict between the objective and perceptual qualities of upscaled outputs for optimizing these models. To achieve an effective trade-off between these two qualities, the current methods are either inflexible as the model is optimized for a fixed trade-off, or inefficient as it needs to interpolate weights or images from two separately trained models. Based on the invertible rescaling net (IRN) which learns image downscaling and upscaling together, we propose a joint optimization method to train just one model that could achieve adjustable trade-off between perception and distortion for upscaling at inference time. Additionally, it's shown in experiments that this jointly optimized model could produce results with better accuracy while maintaining high perceptual quality compared to one optimized for perceptual quality only.

Contributions

1. As far as we know, it is the first solution to train one image rescaling model with joint optimization for both objective and perceptual qualities.
2. A mixture loss conditional to auxiliary latent variable z is proposed for joint optimization to achieve adjustable trade-off by modulating the random sampling of z at inference.

Motivations

The powerful deep learning techniques have led to recent developments of single image super resolution (SR) models with impressive performances. While it is ideal to restore a SR image which is both accurate and photo-realistic, there is a trade-off between the ability to achieve low MSE and high perceptual quality.In other words, for methods minimizing MSE, the restored outputs are often blurred which lack sharp details as in real images. while for methods aim to improve the perceptual image quality using adversarial training, the outputs are sharper but are subjected to lower accuracy when compared to GT references. ESRGAN [1] proposed to train two separate networks which enhance the objective and perceptual quality respectively and combine them using weights interpolation.

Furthermore, these models are often trained for upscaling reconstruction only without taking the image downscaling method into consideration together. Recently, Xiao *et al.* [2] proposed a invertible rescaling net (IRN) that has set the state-of-the-art (SOTA) for learning based bidirectional image rescaling. Based on the invertible neural network (INN), IRN is optimized for both downscaling and inverse upscaling jointly so it is able to restore HR image more accurately. Similarly, it is subject so similar trade-off between perception and distortion and two models are needed to train for each objective separately. The goal of this work is to train the model one time to achieve adjustable trade-off between perception and distortion at inference.

Methods

Problem Formulation

As shown in Fig. 1, the input HR image x is split to low-frequency component x_L and high-frequency component x_H using an $\times 2$ wavelet transformation, before transforming into a downscaled LR image y and an auxiliary latent variable z using an invertible neural network. This forward downscaling process is described as $(y, z) = f(x)$. As both steps are invertible transformations, the inverse upscaling process $x = f^{-1}(y, z)$ is determined once f is known. In previous IRN work, it forces z to be a case-agnostic random variable thus the inverse upscaling process can reconstruct an HR image $\hat{x}$ to either best objective quality like x_d or perceptual quality like x_p using different mixtures of losses as shown below

$$L = \lambda_1 L_r + \lambda_2 L_g + \lambda_3 L_d + \lambda_4 L_p. \tag{1}$$

Here L_r is the $L1$ reconstruction loss for upscaled HR output and L_g is the $L2$ guidance loss for downscaled LR output. For L_d, it is either implemented as distribution regulation of of latent variable z to help train a model (IRN) with minimal distortion, or as an adversarial loss from a co-trained discriminator to train a model (IRN+) generating photo-realistic HR images. The perceptual loss L_p is only used for IRN+ when the goal is best perceptual quality. While the proposed losses work well for either optimal objective or perceptual qualities, it is not an efficient solution. Not only it requires training of two separate models, the random sampling of z from normal distribution $N(0, 1)$ does little in increasing diversity of upscaled outputs. Here we propose a joint optimization method to train just one model that can generate multiple HR outputs, one with high accuracy, or high perception quality, or an adjustable trade-off between the two by sampling z differently.

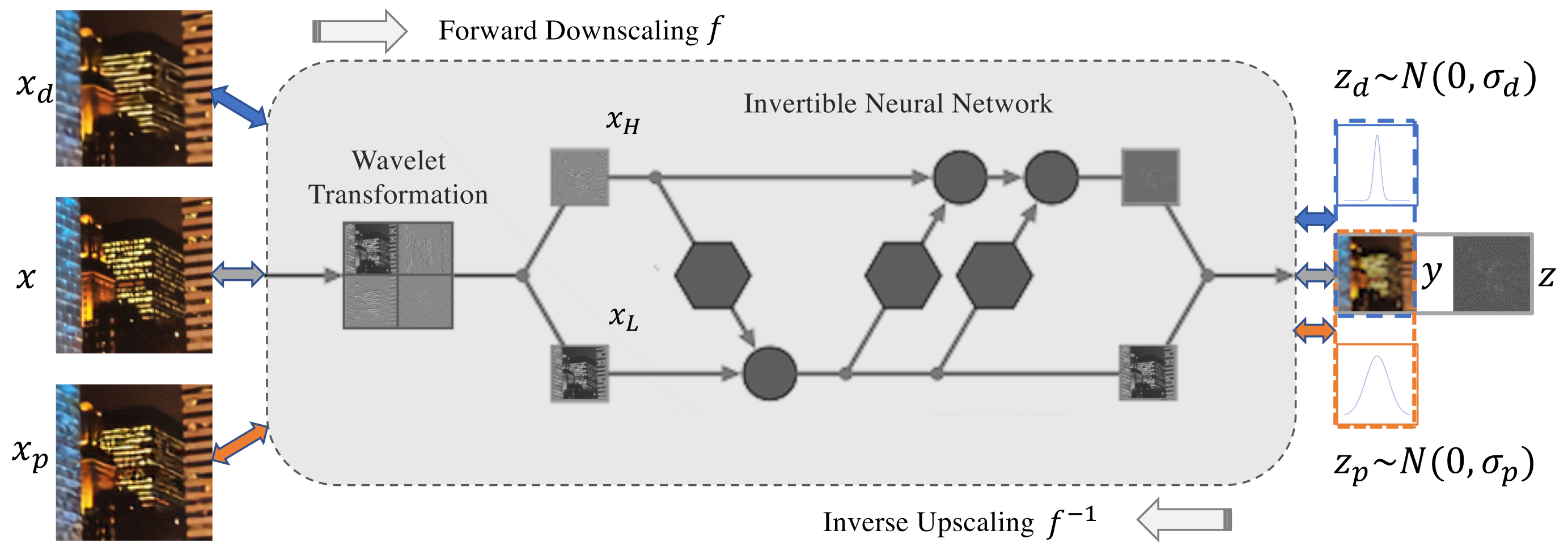

Forward Downscaling f
x_d
x
x_p
Wavelet Transformation
Invertible Neural Network
x_H
x_L
$z_d \sim N(0, \sigma_d)$
y
z
$z_p \sim N(0, \sigma_p)$
Inverse Upscaling f^{-1}

Figure 1: Network architecture for invertible image rescaling with adjustable trade-off between perception and distortion.

Joint Optimization

For learned image rescaling problems, it is required that only the quantized image of downscaled y can be used as known input for inverse upscaling. For IRN, as latent variable z is also needed as input for upscaling, a case-agnostic random variable is used at inference. To achieve our goal of flexible trade-off between perception and distortion with one model, additional information from z must be utilized. For randomly sampled z, there are only two associated parameters μ and σ. On the other hand, it is known in information theory that the differential entropy of a normal distribution is $\ln(\sigma\sqrt{2\pi e})$, which depends on the standard deviation σ solely. In other words, The entropy of normally distributed z is higher when z has higher variance, a larger σ. Lastly, it is known that a restored a HR image x_d with little distortion is more blurry while x_p, which is optimized for better perception, has sharper details. In other words, x_p has higher entropy than x_d.

Inspired by these observations, a joint optimization method is proposed to train an IRN model using the following loss

$$L = (1 - \sigma)\lambda_1 L_r + \lambda_2 L_g + \sigma(\lambda_3 L_d + \lambda_4 L_p). \tag{2}$$

Here the four individual losses, including associated weights λ_i, are the same ones used for IRN+ training. The forward downscaling process is the same as in IRN for our method. For upscaling, z is randomly sampled from $N(0, \sigma)$. Empirically, we limit σ in the range of (0, 1) during training. Thus for smaller σ, latent variable z has lower entropy, the model is biased towards minimizing reconstruction loss L_r, which leads to restored HR output with less distortion and relatively lower entropy. When σ is larger and z has higher entropy, the model is biased towards perception related losses L_d and l_p.

After the joint optimization completed, based on desired trade-off between perception and distortion, the upscaled image can be flexibly generated as $x_\sigma = f^{-1}(y, z_\sigma)$ when z_σ is randomly sampled from $N(0, \sigma)$. Here σ could be set as 0 for the least distortion, or as 1 for best perception, or any value in between according to the desired trade-off. This is done at inference time of upscaling and only one model is needed.

Experiment Results

As shown in Fig.2a, the perception-distortion trade-off curve is first plotted to compare the effects of σ sampling strategies and it shows that 3-point (0, 0.5 and 1) sampling is the best overall, capable of equivalent individual peak performance while enjoying greatly improved trade-off comparing to 2-point sampling. From Fig.2b, it is shown that the jointly optimized IRN$^\sigma$ is able to achieve equivalent performance as the combination of IRN and IRN+. It is also able to achieve higher PSNR for outputs of high perceptual quality, as demonstrated in the visual examples (g) and (h) in Fig. 3.

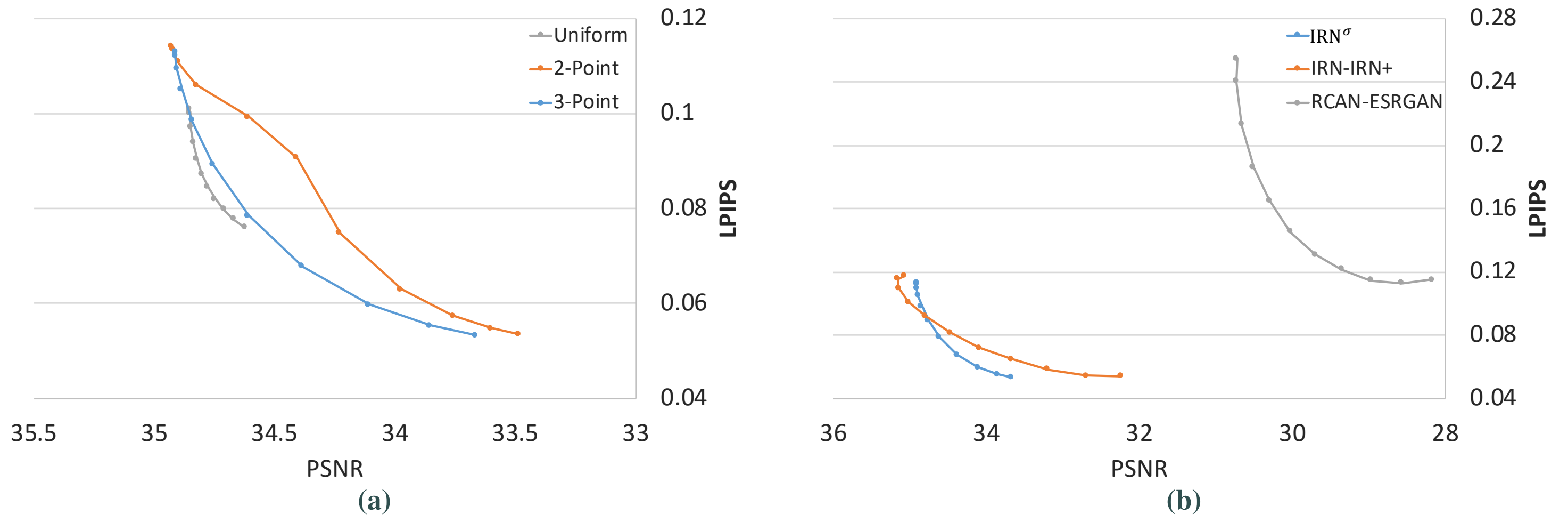
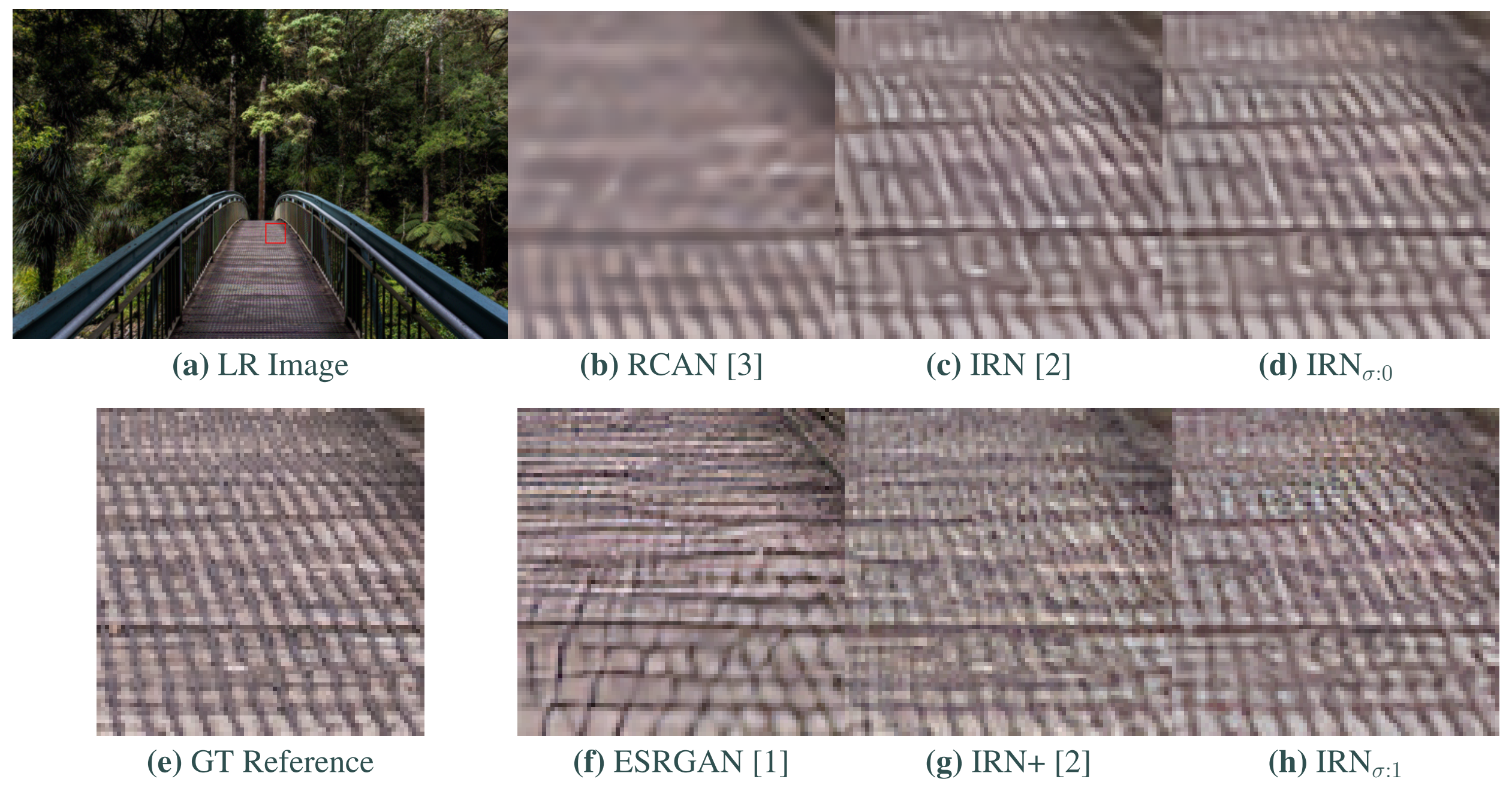


Figure 2: Perception-distortion trade-off comparisons: a) different σ sampling strategies during training; b) different image SR and rescaling models (DIV2K $\times 4$).



(a) LR Image (b) RCAN [3] (c) IRN [2] (d) IRN$_{\sigma:0}$

(e) GT Reference (f) ESRGAN [1] (g) IRN+ [2] (h) IRN$_{\sigma:1}$

Figure 3: Visual examples of upscaled ($\times 4$) images with distortion (top row) and perception (bottom row) preferences.

Method	BSD100		Urban100		DIV2K	
	LPIPS↓	PSNR/SSIM↑	LPIPS↓	PSNR/SSIM↑	LPIPS↓	PSNR/SSIM↑
RCAN [3]	0.3589	27.74/0.742	0.1967	26.75/0.806	0.2547	30.72/0.844
ESRGAN [1]	0.1630	25.29/0.649	0.1239	24.35/0.732	0.1150	28.17/0.775
IRN [2]	0.1654	31.63/0.881	0.0836	31.40/0.915	0.1174	35.07/0.931
IRN+ [2]	0.0749	28.93/0.818	0.0550	28.24/0.867	0.0541	32.24/0.891
IRN$_{\sigma:0}$	0.1626	31.48/0.879	0.0798	31.20/0.913	0.1130	34.91/0.929
IRN$_{\sigma:1}$	0.0778	30.17/0.848	0.0487	30.19/0.896	0.0534	33.66/0.910

Table 1: Comparison of objective and perceptual qualities for upscaled $\times 4$ images, with the best two results highlighted in red and blue. IRN$_{\sigma:0}$ and IRN$_{\sigma:1}$ refer to results from IRN$_\sigma$ while fixing σ at 0 and 1 during inference respectively..

References

[1] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 63–79, 2018.

[2] Mingqing Xiao, Shuxin Zheng, Chang Liu, Yaolong Wang, Di He, Guolin Ke, Jiang Bian, Zhouchen Lin, and Tie-Yan Liu. Invertible image rescaling. In *European Conference on Computer Vision*, pages 126–144. Springer, 2020.

[3] 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 (ECCV)*, pages 286–301, 2018.