Best-Buddy GANs for Highly Detailed Image Super-resolution

Wenbo Li^{1*}, Kun Zhou^{2*}, Lu Qi^{1}, Liying Lu^{1}, Jiangbo Lu^{2}

^{1}The Chinese University of Hong Kong ^{2}Smartmore Technology
{wenboli,luqi,lylu}@cse.cuhk.edu.hk {kun.zhou,jiangbo}@smartmore.com

Abstract

We consider the single image super-resolution (SISR) problem, where a high-resolution (HR) image is generated based on a low-resolution (LR) input. Recently, generative adversarial networks (GANs) become popular to hallucinate details. Most methods along this line rely on a predefined single-LR-single-HR mapping, which is not flexible enough for the ill-posed SISR task. Also, GAN-generated fake details may often undermine the realism of the whole image. We address these issues by proposing best-buddy GANs (Beby-GAN) for rich-detail SISR. Relaxing the rigid one-to-one constraint, we allow the estimated patches to dynamically seek trustworthy surrogates of supervision during training, which is beneficial to producing more reasonable details. Besides, we propose a region-aware adversarial learning strategy that directs our model to focus on generating details for textured areas adaptively. Extensive experiments justify the effectiveness of our method. An ultra-high-resolution 4K dataset is also constructed to facilitate future super-resolution research.

Introduction

The increasing demand for high-quality displays has promoted the rapid development of single image super-resolution (SISR). SISR has been successfully applied to a wide range of tasks, such as medical diagnostic imaging, security imaging and satellite imaging.

A great number of methods were proposed based on insightful image priors and optimization techniques, such as self-similarity (Protter et al. 2008; Glasner, Bagon, and Irani 2009; Yang, Huang, and Yang 2010) and sparsity (Martin et al. 2001; Yang et al. 2008, 2010; Zeyde, Elad, and Protter 2010; Peleg and Elad 2014). In recent years, deep-learning-based methods (Dong et al. 2014; Kim, Kwon Lee, and Mu Lee 2016a; Shi et al. 2016; Tai, Yang, and Liu 2017; Tai et al. 2017; Lim et al. 2017; Haris, Shakhnarovich, and Ukita 2018; Zhang et al. 2018c,b) further advance SISR. Most of them rely on an immutable one-to-one supervision to pursue high PSNR but possibly generate blurry results. For example, the solution of commonly adopted one-to-one MSE/MAE metric approximates mean or median of data (Sønderby et al. 2016). As shown in Figure 1, the HR estimation of RCAN (Zhang et al. 2018b) achieves the highest PSNR, yet lacking high-frequency texture.

To enhance the perceptual quality of recovered images, several methods (Johnson, Alahi, and Fei-Fei 2016; Ledig et al. 2017; Mechrez et al. 2018; Wang et al. 2018b; Zhang et al. 2019; Soh et al. 2019) use adversarial learning and perceptual loss (Johnson, Alahi, and Fei-Fei 2016). It is noted that the issue of excessive smoothing caused by the one-to-one MSE/MAE loss is still not fully addressed. Besides, the training of generative adversarial networks (GANs) (Goodfellow et al. 2014) could be unstable and result in unpleasant visual artifacts (see the recovered tree of ESRGAN (Wang et al. 2018b) in Figure 1). Thus in this paper, we aim to address these issues from two aspects.

It is well-known that SISR is essentially an ill-posed problem since a single low-resolution (LR) patch may correspond to multiple high-resolution (HR) solutions - it is difficult to decide the best. The commonly used one-to-one MSE/MAE loss tends to enforce a rigid mapping between the given LR-HR pair and will penalize the model when the HR estimates do not exactly match the ground truth (GT), even when they are valid solutions. As a result, the strictly constrained HR space makes it difficult to train the network. Relaxing the one-to-one constraint, we propose a novel *best-buddy loss*, an improved one-to-many MAE loss, to allow finding and using HR supervision signals flexibly by exploiting the ubiquitous self-similarity existing in natural images, making the model easy to optimize. Put it differently, an estimated HR patch is allowed to be supervised by different but close-to-ground-truth patches sourced from multiple scales of the corresponding GT image. Additionally, a back-projection constraint is introduced to ensure the validity of the estimated HR signal.

As aforementioned, undesirable artifacts may be produced in images for existing GAN-based methods (Ledig et al. 2017; Mechrez et al. 2018; Wang et al. 2018b; Zhang et al. 2019; Soh et al. 2019). We propose a region-aware adversarial learning strategy to address it. Our network treats smooth and well-textured areas differently, and only performs the adversarial training on the latter. This separation encourages the network to focus more on regions with rich details while avoiding generating unnecessary texture on flat regions (e.g., sky and building). With this improvement, our proposed best-buddy GANs (termed as *Beby-GAN*) is able to

[*] These authors contributed equally.

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

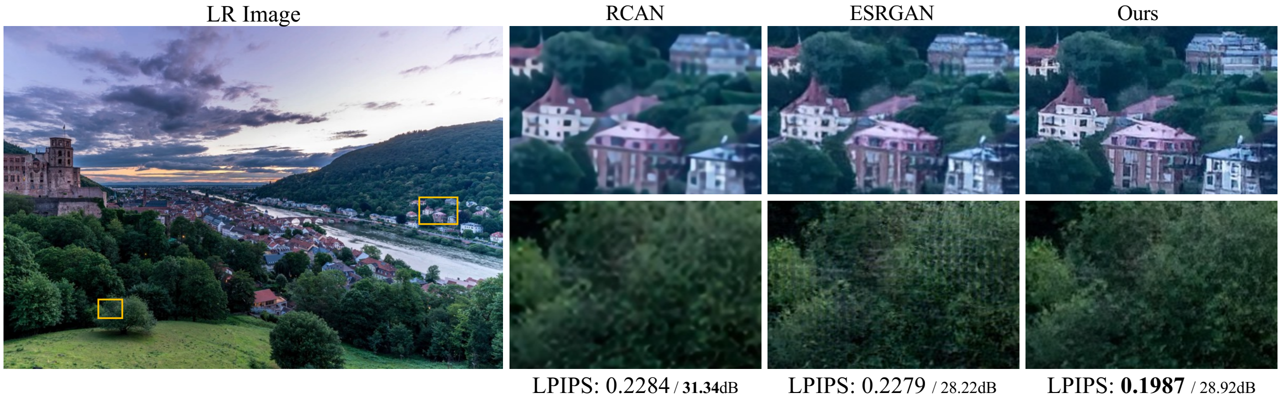

Figure 1: Comparison of our best-buddy GANs, PSNR-oriented RCAN (Zhang et al. 2018b) and perception-oriented ESRGAN (Wang et al. 2018b). The numbers are LPIPS(↓)/PSNR(↑). Richer details are generated in our result with fewer artifacts.

reconstruct photo-realistic high-frequencies with fewer undesirable artifacts (see Figure 1).

In summary, our contribution is threefold:

- We present Beby-GAN for high-quality image super-resolution. The proposed *one-to-many* best-buddy loss benefits generating richer and more plausible texture. Extensive experiments and a user study justify the effectiveness of the proposed method.
- A region-aware adversarial learning strategy is designed to further enhance the visual quality of images.
- Breaking through the 2K resolution limitation of current SISR datasets, we provide an ultra-high-resolution 4K (UH4K) image dataset with diverse categories to promote future study, which will be made publicly available.

Related Work

Single image super-resolution (SISR) is a classical image restoration task. It is roughly divided into three categories of example-based or prior-based methods (Yang et al. 2008; Yang, Huang, and Yang 2010; Zeyde, Elad, and Protter 2010; Timofte, De Smet, and Van Gool 2013, 2014), PSNR-oriented methods (Dong et al. 2014; Kim, Kwon Lee, and Mu Lee 2016a,b; Shi et al. 2016; Lai et al. 2017; Haris, Shakhnarovich, and Ukita 2018; Zhang et al. 2018c,b) and perception-driven methods (Johnson, Alahi, and Fei-Fei 2016; Ledig et al. 2017; Mechrez et al. 2018; Wang et al. 2018b,c,a; Zhang et al. 2019; Soh et al. 2019).

Example-Based Methods

This line (Yang et al. 2008; Zeyde, Elad, and Protter 2010; Timofte, De Smet, and Van Gool 2014; Yang et al. 2012; Peleg and Elad 2014) learns mapping from low-resolution patches to high-resolution counterparts, where the paired patches are collected from an external database. In this paper, we exploit this idea to search for one-to-many LR-HR mappings to produce visually pleasing results.

PSNR-Oriented Methods

In past years, particular attention is paid to improve the pixel-wise reconstruction measures (e.g., peak-to-noise ratio, PSNR). It is the first time that SRCNN (Dong et al. 2014) introduces a deep convolutional neural network into the SISR task. Afterwards, more well-designed architectures were proposed including residual and recursive learning (Kim, Kwon Lee, and Mu Lee 2016a,b; Tai, Yang, and Liu 2017; Lim et al. 2017), sub-pixel upsampling (Shi et al. 2016), Laplacian pyramid structure (Lai et al. 2017) and dense connecting (Zhang et al. 2018c). Especially, Zhang *et al.* (Zhang et al. 2018b) integrated channel attention modules into a network achieving a significant improvement in terms of PSNR performance.

Perception-Driven Methods

Despite breakthroughs on PSNR, the aforementioned methods still face a challenge that super-resolved images are typically overly-smooth and lack high-frequencies. To tackle this problem, Johnson *et al.* (Johnson, Alahi, and Fei-Fei 2016) proposed a novel perceptual loss. Ledig *et al.* (Ledig et al. 2017) presented SRGAN, which utilizes an adversarial loss and a content loss to push outputs into residing on the manifold of natural images. Thanks to a patch-wise texture loss, EnhanceNet (Sajjadi, Scholkopf, and Hirsch 2017) obtains better performance. ESRGAN (Wang et al. 2018b) marked a new milestone, which consistently generates more realistic texture benefiting from model and loss improvements. Later on, Zhang *et al.* (Zhang et al. 2019) proposed RankSRGAN capable of being optimized towards a specific perceptual metric. However, most of these methods rely on single-LR-single-HR MSE/MAE supervision. Besides, without a region-aware mechanism, the architecture design can not deal with regions differently according to their properties. From these perspectives, we propose best-buddy GANs detailed as follows.

Beby-GAN for Image Super-Resolution

Given a low-resolution (LR) image $\mathbf{I}_{\mathrm{LR}} \in \mathbb{R}^{H \times W}$, single image super-resolution (SISR) is supposed to generate

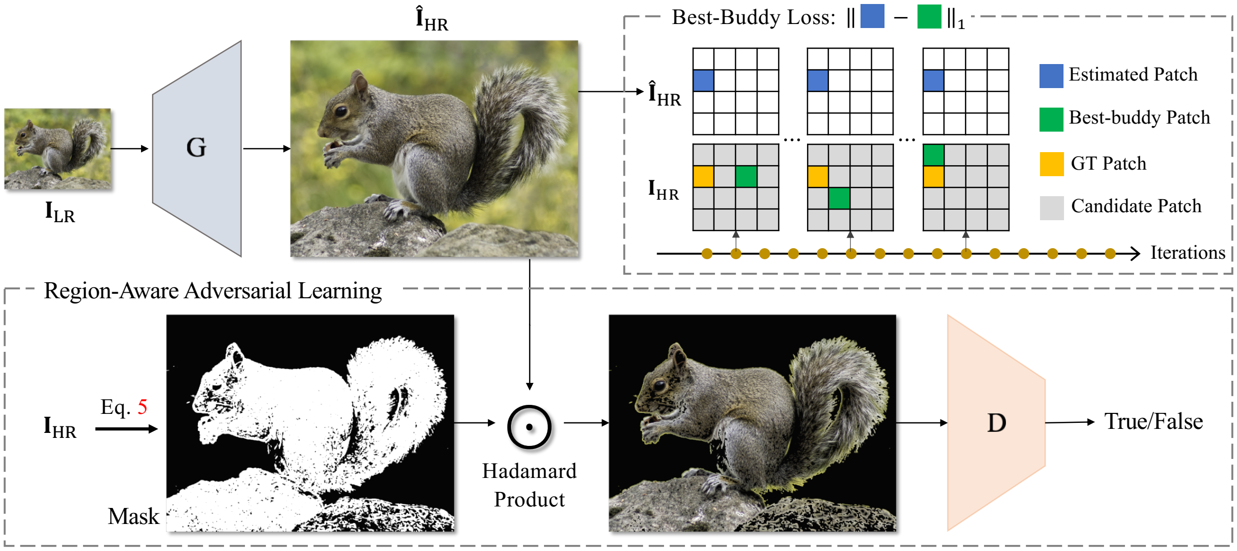

Figure 2: The framework of our Beby-GAN. The best-buddy loss allows the estimated HR patches to be supervised in a dynamic way during training. The region-aware adversarial learning is proposed to make the discriminator focus on rich-texture areas.

a high-resolution (HR) counterpart $\hat{\mathbf{I}}_{\text{HR}} \in \mathbb{R}^{Hs \times Ws}$ under an upscale factor s. As shown in Figure 2, the main body of our framework is built upon the generative adversarial networks (GANs) (Goodfellow et al. 2014), where the generator is used to reconstruct HR images and the discriminator is trained to distinguish between recovered results and real images. Following (Wang et al. 2018b), we adopt a classical pretrained RRDB model as our generator since it has demonstrated strong learning ability. In this section, we first describe the proposed best-buddy loss and region-aware adversarial learning strategy, followed by other loss functions.

Best-Buddy Loss

In the super-resolution task, a single LR patch is essentially associated with multiple natural HR solutions, as illustrated in Figure 3(a). Existing methods generally focus on learning immutable single-LR-single-HR mapping using an MSE/MAE loss in the training phase (see Figure 3(b)), which ignores the inherent uncertainty of SISR. As a result, the generated HR images may lack several high-frequency structures (Mathieu, Couprie, and LeCun 2016; Johnson, Alahi, and Fei-Fei 2016; Ledig et al. 2017).

To alleviate this issue, we propose a one-to-many best-buddy loss to enable trustworthy but much more flexible supervision. The key idea is that an estimated HR patch is allowed to be supervised by diverse targets in different iterations (see Figure 2). In this paper, all supervision candidates come from a multi-scale ground-truth image pyramid. As shown in Figure 3(c), for an estimated HR patch $\hat{\mathbf{p}}_i$, we look for its corresponding supervision patch $\mathbf{g}_i^*$ (i.e., *best buddy*) in the current iteration to meet two constraints:

Constraint 1. $\mathbf{g}_i^*$ is required to be close to the predefined ground-truth $\mathbf{g}_i$ in the HR space (1^{st} term in Eq. 2). Relying on the ubiquitous multi-scale self-similarity in natural images (Kindermann, Osher, and Jones 2005; Protter et al. 2008; Yang, Huang, and Yang 2010; Li et al. 2020), it is very likely to find a HR patch that is close to the ground-truth $\mathbf{g}_i$.

Constraint 2. In order to make optimization easy, $\mathbf{g}_i^*$ ought to be close to the estimation $\hat{\mathbf{p}}_i$ (2^{nd} term in Eq. 2). Note that $\hat{\mathbf{p}}_i$ is considered to be a reasonable prediction as our generator is well initialized.

Optimized with these two objectives, the obtained best buddy $\mathbf{g}_i^*$ is regarded as a plausible HR target for supervision. In detail, we first downsample the ground-truth (GT) HR image $\mathbf{I}_{\text{HR}}$ with different scale factors as

$$\mathbf{I}_{\text{HR}}^r = S\left(\mathbf{I}_{\text{HR}}, r\right), \ r = \{2, 4\}, \tag{1}$$

where $S(\mathbf{I}, r) : \mathbb{R}^{H \times W} \to \mathbb{R}^{\frac{H}{r} \times \frac{W}{r}}$ is a bicubic downsampling operator, and obtain a 3-level image pyramid (including the original GT HR image). Then, we unfold the estimated HR image and corresponding GT image pyramid into patches (3×3 in our paper), among which the GT part forms the supervision candidate database $\mathcal{G}$ of this image. For the i-th estimated patch $\hat{\mathbf{p}}_i$, instead of being supervised by the immutable predefined GT patch $\mathbf{g}_i$, it is allowed to find the best buddy $\mathbf{g}_i^*$ in the current iteration as

$$\mathbf{g}_i^* = \underset{\mathbf{g} \in \mathcal{G}}{\arg\min}\ \alpha \|\mathbf{g} - \mathbf{g}_i\|_2^2 + \beta \|\mathbf{g} - \hat{\mathbf{p}}_i\|_2^2, \tag{2}$$

where $\alpha \geq 0$ and $\beta \geq 0$ are scaling parameters. The best-buddy loss for this patch pair $(\hat{\mathbf{p}}_i, \mathbf{g}_i^*)$ is represented as

$$\mathcal{L}_{\text{BB}}\left(\hat{\mathbf{p}}_i, \mathbf{g}_i^*\right) = \|\hat{\mathbf{p}}_i - \mathbf{g}_i^*\|_1. \tag{3}$$

Notice that when $\beta \ll \alpha$, the proposed best-buddy loss degenerates into the typical one-to-one MAE loss.

Besides, we enforce another back-projection constraint on the estimation $\hat{\mathbf{I}}_{\text{HR}}$. The super-resolved images after downscaling must match the fidelity expected at the lower resolution. We introduce an HR-to-LR operation (bicubic downsampling) to ensure that the projection of the estimated HR

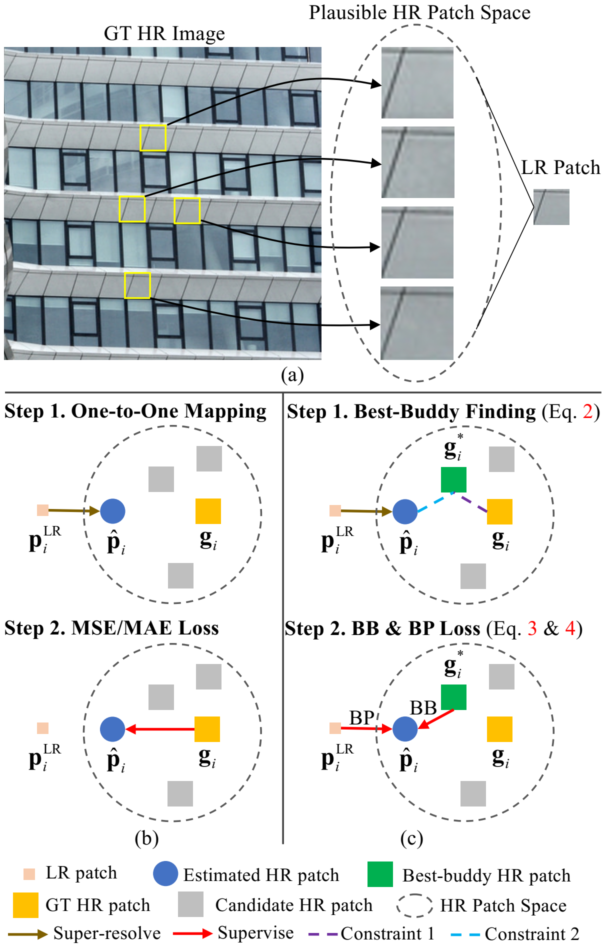

Figure 3: Comparison of the MSE/MAE and the best-buddy (BB) loss with a back-projection (BP) constraint. (a) Depiction of one-to-many nature in the SISR task. (b) MSE/MAE loss. (c) BB and BP loss. Variables $\mathbf{p}_i^{\mathrm{LR}}$, $\hat{\mathbf{p}}_i$, $\mathbf{g}_i$ and $\mathbf{g}_i^*$ indicate the LR patch, estimated HR patch, ground-truth HR patch and best-buddy HR patch in current iteration.

image onto the LR space is still consistent with the original LR one. The back-projected result is supervised by

$$\mathcal{L}_{\mathrm{BP}} = \left\| S\left(\hat{\mathbf{I}}_{\mathrm{HR}}, s\right) - \mathbf{I}_{\mathrm{LR}} \right\|_1 , \tag{4}$$

where s is the downscale factor. From Figure 4, we notice that this back-projection loss plays an essential role in maintaining content and color consistency.

Region-Aware Adversarial Learning

As shown in Figure 1, previous GAN-based methods sometimes produce undesirable texture, especially in flat regions. Thus, as illustrated in Figure 2, we propose to differentiate the rich-texture areas from smooth ones according to local pixel statistics, and only feed the textured content to the discriminator since smooth regions can be easily recovered without adversarial training. Our strategy is to first unfold the ground-truth HR image (i.e., $\mathbf{I}_{\mathrm{HR}} \in \mathbb{R}^{Hs \times Ws}$) into

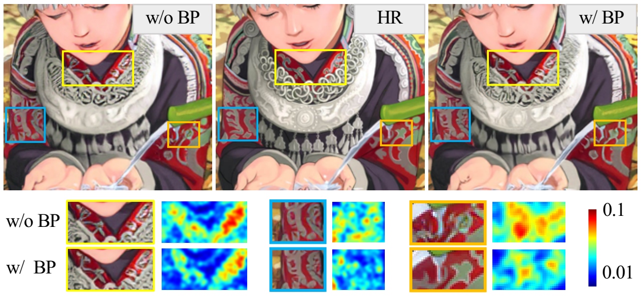

Figure 4: Comparison between with (w/) and without (w/o) the back-projection (BP) loss. We visualize the L2 error heatmaps between the estimated results and ground truth.

patches $\mathbf{B} \in \mathbb{R}^{Hs \times Ws \times k^2}$ with size k^2, and then compute standard deviation (std) for each patch. After that, a binary mask is obtained as

$$\mathbf{M}_{i,j} = \begin{cases} 1, & \operatorname{std}\left(\mathbf{B}_{i,j}\right) \geq \delta \\ 0, & \operatorname{std}\left(\mathbf{B}_{i,j}\right) < \delta \, , \end{cases} \tag{5}$$

where δ is a predefined threshold and (i, j) is the patch location. The highly textured regions are marked as 1 while flat regions as 0. Then the estimated result $\hat{\mathbf{I}}_{\mathrm{HR}}$ and ground-truth $\mathbf{I}_{\mathrm{HR}}$ are multiplied with the same mask $\mathbf{M}$ yielding $\hat{\mathbf{I}}_{\mathrm{HR}}^{\mathbf{M}}$ and $\mathbf{I}_{\mathrm{HR}}^{\mathbf{M}}$, which are next processed by the following discriminator. Though more sophisticated strategies can be used at the cost of more computation, we show that this simple masking method already works very well. It directs our model to generate realistic fine details for textured areas.

Other Loss Functions

Perceptual Loss. Apart from computing pixel-wise distances in image space, several works (Estrach, Sprechmann, and LeCun 2016; Dosovitskiy and Brox 2016; Johnson, Alahi, and Fei-Fei 2016) use feature similarity to enhance the perceptual quality of recovered images. Following this idea, we set the perceptual loss as

$$\mathcal{L}_{\mathrm{P}} = \sum_i \eta_i \left\| \phi_i\left(\hat{\mathbf{I}}_{\mathrm{HR}}\right) - \phi_i\left(\mathbf{I}_{\mathrm{HR}}\right) \right\|_1 , \tag{6}$$

where ϕ_i represents the i-th layer activation of a pretrained VGG-19 (Simonyan and Zisserman 2015) network and η_i is a scaling coefficient. To capture feature representations at different levels, we take into consideration three layers including $conv_{3\_4}$, $conv_{4\_4}$ and $conv_{5\_4}$ and set scaling coefficients to $\frac{1}{8}$, $\frac{1}{4}$ and $\frac{1}{2}$ empirically.

Adversarial Loss. The discriminator in our network is implemented based on Relativistic average GANs (RaGANs) (Jolicoeur-Martineau 2018), which estimate the probability that a ground-truth HR image is more realistic than a generated one. It has been shown that RaGANs are more stable and can produce results of higher quality (Jolicoeur-Martineau 2018; Wang et al. 2018b; Soh et al. 2019). The loss functions are formulated as

$$\mathcal{L}_{\mathrm{D}} = -\mathbb{E}_{x_r}\left[\log\left(\bar{D}\left(x_r\right)\right)\right] - \mathbb{E}_{x_f}\left[\log\left(1 - \bar{D}\left(x_f\right)\right)\right] , \tag{7}$$

$$\mathcal{L}_{\mathrm{G}} = -\mathbb{E}_{x_r}\left[\log\left(1-\bar{D}\left(x_r\right)\right)\right] - \mathbb{E}_{x_f}\left[\log\left(\bar{D}\left(x_f\right)\right)\right] , \tag{8}$$

where

$$\bar{D}\left(x\right) = \begin{cases} \text{sigmoid}\left(C\left(x\right) - \mathbb{E}_{x_f} C\left(x_f\right)\right) , & \text{x is real} \\ \text{sigmoid}\left(C\left(x\right) - \mathbb{E}_{x_r} C\left(x_r\right)\right) , & \text{x is fake} . \end{cases} \tag{9}$$

In Eq. 9, x_r denotes the masked real data and x_f is the masked fake data (i.e., generated data) and $C\left(x\right)$ is the non-transformed discriminator output.

Overall Loss. The overall loss of generator is formulated as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\mathrm{BB}} + \lambda_2 \mathcal{L}_{\mathrm{BP}} + \lambda_3 \mathcal{L}_{\mathrm{P}} + \lambda_4 \mathcal{L}_{\mathrm{G}} , \tag{10}$$

where $\lambda_1 = 0.1$, $\lambda_2 = 1.0$, $\lambda_3 = 1.0$ and $\lambda_4 = 0.005$.

Experiments

Datasets

Our network is trained on DIV2K (Agustsson and Timofte 2017) (800 images) and Flickr2K (Timofte et al. 2017) (2650 images) datasets. Apart from the widely used testing benchmark including Set5 (Bevilacqua et al. 2012), Set14 (Zeyde, Elad, and Protter 2010), BSDS100 (Martin et al. 2001) and Urban100 (Huang, Singh, and Ahuja 2015), we also adopt the 100 validation images in DIV2K to evaluate the performance of our model.

Besides, we propose an ultra-high-resolution 4K (UH4K) dataset to perform a more challenging and complete study on the single image super-resolution (SISR) task. The images are collected from YouTube with resolution 3840 × 2160. The dataset contains over 400 images featuring four categories, i.e., animal, city, nature and sports (see supplementary). Compared with existing benchmark datasets, ours has higher resolution, higher variety and richer texture/structure. In this paper, our 4K dataset is *only* used for evaluation.

Perceptual Metrics

PSNR and SSIM (Wang et al. 2004) (the higher, the better ideally) have already been shown to correlate weakly with human perception regarding image quality (Ledig et al. 2017; Sajjadi, Scholkopf, and Hirsch 2017). Thus, following (Jo, Yang, and Kim 2020; Wang et al. 2018b), we mainly utilize a ground-truth-based perceptual metric LPIPS (Zhang et al. 2018a) and a no-reference perceptual metric PI (Blau et al. 2018) (the lower, the better) for evaluation. The LPIPS results are calculated based on the VGG model. We also conduct a user study for better comparison.

Training Details

All experiments are carried out on NVIDIA GeForce RTX 2080 Ti GPUs under the ×4 setting. The mini-batch size is set to 8. We adopt Adam as the optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. There are 3 periods in our training, each with 200K iterations. The learning rate for every period is set to 1×10^{-4} initially in accompany with a warm-up and a cosine decay. The images are augmented with random cropping, flipping and rotation. The input size is 48 × 48 and the rotation is 90° or −90°. The α and β are both set to 1.0 from empirical experiments. The kernel size k and δ are set to 11 and 0.025 (for normalized images). The calculation of bset-buddy loss costs 5.9ms for images of size 196 × 196. Our method costs 193.6ms to obtain a 1280 × 720 HR image.

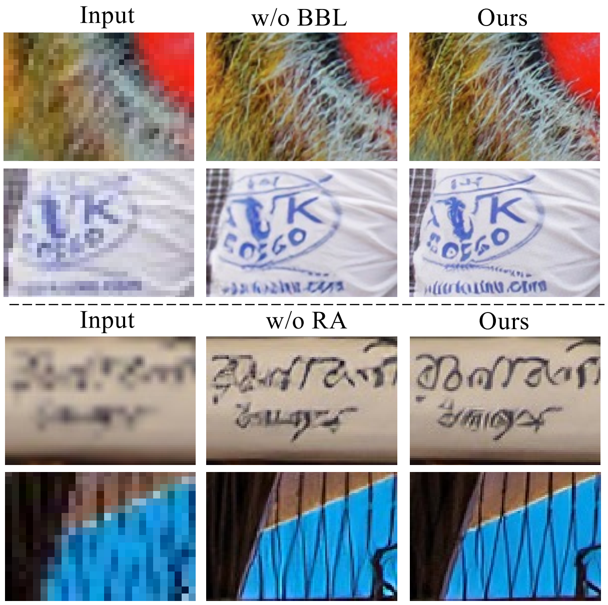

Figure 5: Visual comparison for ablation study. "Ours" is our proposed Beby-GAN. "w/o BBL" and "w/o RA" means removing best-buddy loss and region-aware learning.

Dataset	Ours	w/o BBL	w/o RA
Set14	**0.2202/3.08**	0.2343/3.21	0.2272/3.17
BSDS100	**0.2385/2.44**	0.2514/2.48	0.2469/2.48

Table 1: Results (LPIPS↓/PI↓) for ablation study. "Ours" is our proposed Beby-GAN. "w/o BBL" and "w/o RA" indicate removing best-buddy loss and region-aware learning.

Ablation Study

In this part, we investigate how each design affects the perceptual quality of super-resolved images. Starting from our best-buddy GANs (Beby-GAN), we ablate the best-buddy loss and region-aware learning strategy, respectively. We show a visual comparison in Figure 5. Also, we evaluate the LPIPS and PI performance in Table 1 because they are more consistent with human perception than PSNR/SSIM (Ledig et al. 2017; Sajjadi, Scholkopf, and Hirsch 2017). More ablation study experiments are described in the supplementary file. All these results verify the effectiveness of our method.

Best-buddy loss. In contrast to the commonly used one-to-one MSE/MAE loss, our best-buddy loss allows the network to learn single-LR-multiple-HR mapping. As illustrated in Figure 5, BBL (see "Ours") recovers richer texture and sharper edges compared with one-to-one MAE (see " w/o BBL"). The whiskers have more high-frequency details and the text is clearer. Also, we notice that the super-resolved images are more natural and visually pleasing. As

Dataset	Metric	SRResNet	RRDB	RCAN	SRGAN	ESRGAN-PI	ESRGAN	RankSRGAN[†]	Beby-GAN-PI	Beby-GAN
Set14	LPIPS↓	0.3043	0.2934	0.2922	0.3162	0.2771	0.2372	0.2545	0.2537	**0.2202**
	PI↓	5.36	5.27	5.30	2.87	2.61	2.93	2.61	**2.57**	3.08
	PSNR↑	28.57	28.95	28.97	25.90	26.39	26.28	26.57	26.56	**26.96**
	SSIM↑	0.7834	0.7912	0.7913	0.6942	0.7021	0.6985	0.7052	0.7061	**0.7282**
BSDS100	LPIPS↓	0.3437	0.3341	0.3320	0.3387	0.2801	0.2599	0.2790	0.2777	**0.2385**
	PI↓	5.34	5.30	5.19	2.62	2.27	2.48	2.15	**2.13**	2.44
	PSNR↑	27.61	27.84	27.84	25.38	25.72	25.32	25.57	25.56	**25.81**
	SSIM↑	0.7376	0.7453	0.7456	0.6423	0.6638	0.6514	0.6492	0.6536	**0.6781**
DIV2K	LPIPS↓	0.2991	0.2863	0.2862	0.3109	0.2741	0.2222	0.2368	0.2352	**0.1991**
	PI↓	5.40	5.28	5.33	3.25	**2.95**	3.27	3.00	3.02	3.33
	PSNR↑	30.49	30.90	30.86	27.16	27.80	28.16	28.01	28.12	**28.71**
	SSIM↑	0.8391	0.8478	0.8469	0.7600	0.7653	0.7752	0.7652	0.7688	**0.7923**
UH4K	LPIPS↓	0.2304	0.2225	0.2208	0.3346	0.2694	0.2160	0.2745	0.2711	**0.2009**
	PI↓	5.53	5.50	5.56	3.91	2.93	3.42	**2.87**	2.89	3.54
	PSNR↑	32.11	32.45	32.46	27.85	28.94	29.43	28.99	29.13	**30.02**
	SSIM↑	0.8691	0.8756	0.8751	0.7706	0.7874	0.8052	0.7850	0.7892	**0.8214**

Table 2: Quantitative comparison of PSNR-oriented (on the left) and GAN-based methods (on the right) on benchmarks. '↓' means the lower, the better. '↑' indicates the higher, the better. '†' means that the results of RankSRGAN (Zhang et al. 2019) are from multiple models optimized by different objectives. Best results in GAN-based methods are shown in bold.

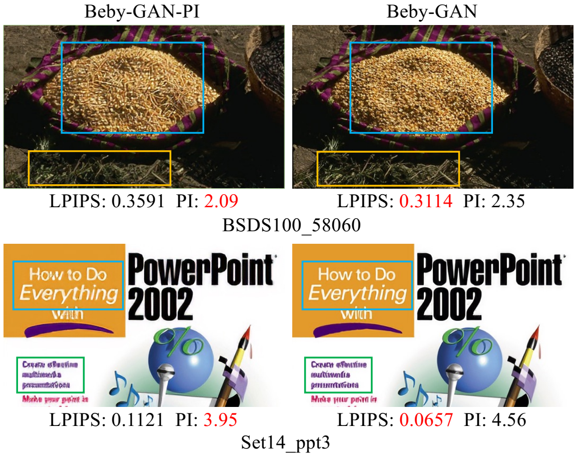

Figure 6: Comparison of Beby-GAN-PI and Beby-GAN. Better results are shown in red. Ground-truth-based LPIPS is more representative and robust than no-reference-based PI. Zoom in for better visual comparison.

shown in Table 1, the best-buddy loss brings about considerable improvement on LPIPS and PI results.

Region-Aware Adversarial Learning. As shown in the Figure 5, there exist unpleasant artifacts near the characters and railing in the results without region-aware learning (see "w/o RA"). After differentiating between rich-texture and flat areas, this problem is alleviated as shown in the 3^{rd} column (see "Ours"). The separation allows the network to know "where" to conduct the adversarial training and yields two major advantages. On the one hand, it leads to easier training since the network only needs to focus on regions of high-frequency details. On the other hand, the network produces less unnatural texture. The results in Table 1 also demonstrate the effectiveness of this strategy.

Comparison with State-of-the-Art Methods

We compare our Beby-GAN with start-of-the-art methods of two categories. They are PSNR-oriented methods including SRResNet (Ledig et al. 2017), RRDB (Wang et al. 2018b), RCAN (Zhang et al. 2018b), and perception-driven methods including ESRGAN (Wang et al. 2018b) and RankSRGAN (Zhang et al. 2019). We use Set14 (Zeyde, Elad, and Protter 2010), BSDS100 (Martin et al. 2001), DIV2K validation (Agustsson and Timofte 2017) and a subset of our UH4K for quantitative evaluation while more datasets (Bevilacqua et al. 2012; Huang, Singh, and Ahuja 2015; Matsui et al. 2017) for qualitative analysis.

Quantitative Results Following ESRGAN (Wang et al. 2018b) and RankSRGAN (Zhang et al. 2019) that provide different models for quantitative evaluation, we prepare two models, named Beby-GAN-PI and Beby-GAN. The former is obtained using network interpolation as ESRGAN-PI (Wang et al. 2018b). As shown in Table 2, GAN-based methods obtain better performance on perceptual metrics with lower PSNR/SSIM scores.

As for GAN-based methods, our Beby-GAN performs best on PSNR/SSIM measures. Also, our method yields new state of the art in terms of LPIPS on all benchmarks. In terms of PI, our PI-based model achieves superior performance on Set14 and BSDS100, as well as comparable results on DIV2K and UH4K. We notice that there is a relatively large disparity between the ground-truth-based LPIPS and no-reference-based PI. As shown in Figure 6, LPIPS is more consistent with human perception. In this case, PI is only used for reference.

Qualitative Results As illustrated in Figure 7, PSNR-oriented methods (i.e., RCAN (Zhang et al. 2018b)) tend to generate overly-smooth results. Although existing GAN-

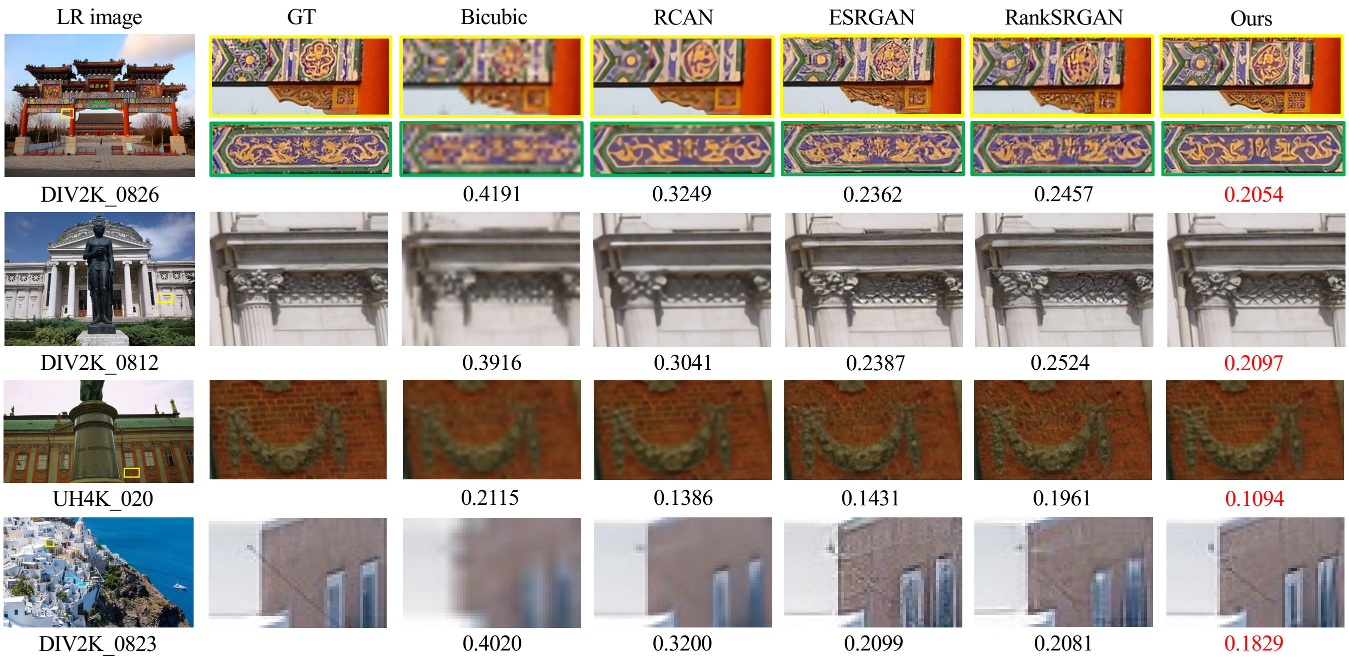

Figure 7: Visual comparison of our Beby-GAN with other methods on ×4 scale. The values beneath images represent LPIPS measures. Red: best quantitative results. It is clear that our Beby-GAN obtains the best visual performance.

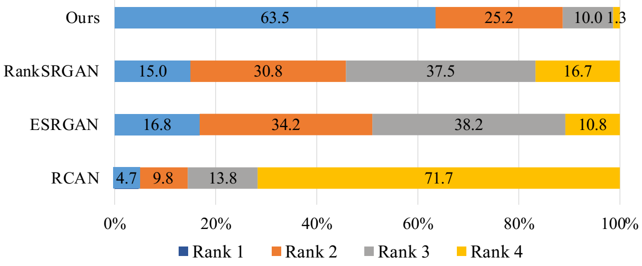

Figure 8: The ranking results of user study involving 30 participants. The values represent the percentages of rankings.

based methods (i.e., ESRGAN (Wang et al. 2018b) and RankSRGAN (Zhang et al. 2019)) can recover some details, they possibly generate unpleasing visual artifacts and color inconsistency (see UH4K_020 and DIV2K_0823) .

In contrast, our Beby-GAN is capable of producing more realistic results. From Figure 7, it is clear that our method reconstructs richer and more compelling patterns as well as sharper structures. Also, fewer artifacts are produced. Especially, our Beby-GAN outperforms others by a large margin on the 4K dataset (see supplementary). In the following, we further present a comprehensive user study to evaluate the human visual quality of reconstructed images.

User Study

In addition to our method, we take into consideration RCAN (Zhang et al. 2018b), ESRGAN (Wang et al. 2018b) and RankSRGAN (Zhang et al. 2019). We prepare testing cases from three sources: (1) *Low-resolution images* stemming from the commonly used benchmark including Set5 (Bevilacqua et al. 2012), Set14 (Zeyde, Elad, and Protter 2010), BSDS100 (Martin et al. 2001) and Urban100 (Huang, Singh, and Ahuja 2015). There are a total of 219 images. (2) *2K resolution* images from the validation subset of DIV2K (Agustsson and Timofte 2017). 100 images are included. (3) *4K resolution* images in our UH4K dataset. There are over 400 images from 4 categories.

Every time we randomly display 30 testing cases and ask the participant to rank 4 versions of each image: RCAN, ESRGAN, RankSRGAN and our Beby-GAN. To make a fair comparison, we follow (Zhang et al. 2019) to zoom in one random small patch for each image.

We invite 30 participants to our user study. As shown in the Figure 8, most of our results rank in the first place while the remaining of ours still gets high rankings. Besides, ESRGAN and RankSRGAN achieve better performance than RCAN. The user study not only demonstrates the superiority of our Beby-GAN, but also explains that existing evaluation measures and human perception are diverse to some extent.

Conclusion

In this paper, we have presented best-buddy GANs (Beby-GAN) for highly detailed image super-resolution. By virtue of the proposed best-buddy loss and region-aware adversarial learning, our Beby-GAN is able to recover realistic texture while maintaining the naturalness of images. Extensive experiments manifest the effectiveness of our method.

References

Agustsson, E.; and Timofte, R. 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, 126–135.

Bevilacqua, M.; Roumy, A.; Guillemot, C.; and line Alberi Morel, M. 2012. Low-Complexity Single-Image Super-Resolution based on Nonnegative Neighbor Embedding. In *Proceedings of the British Machine Vision Conference*, 135.1–135.10. BMVA Press.

Blau, Y.; Mechrez, R.; Timofte, R.; Michaeli, T.; and Zelnik-Manor, L. 2018. The 2018 pirm challenge on perceptual image super-resolution. In *ECCV*, 0–0.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2014. Learning a deep convolutional network for image super-resolution. In *ECCV*, 184–199. Springer.

Dosovitskiy, A.; and Brox, T. 2016. Generating images with perceptual similarity metrics based on deep networks. In *NIPS*, 658–666.

Estrach, J. B.; Sprechmann, P.; and LeCun, Y. 2016. Super-resolution with deep convolutional sufficient statistics. In *ICLR*.

Glasner, D.; Bagon, S.; and Irani, M. 2009. Super-resolution from a single image. In *ICCV*, 349–356. IEEE.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NIPS*, 2672–2680.

Haris, M.; Shakhnarovich, G.; and Ukita, N. 2018. Deep back-projection networks for super-resolution. In *CVPR*, 1664–1673.

Huang, J.-B.; Singh, A.; and Ahuja, N. 2015. Single image super-resolution from transformed self-exemplars. In *CVPR*, 5197–5206.

Jo, Y.; Yang, S.; and Kim, S. J. 2020. Investigating loss functions for extreme super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 424–425.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 694–711. Springer.

Jolicoeur-Martineau, A. 2018. The relativistic discriminator: a key element missing from standard GAN. *arXiv preprint arXiv:1807.00734*.

Kim, J.; Kwon Lee, J.; and Mu Lee, K. 2016a. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 1646–1654.

Kim, J.; Kwon Lee, J.; and Mu Lee, K. 2016b. Deeply-recursive convolutional network for image super-resolution. In *CVPR*, 1637–1645.

Kindermann, S.; Osher, S.; and Jones, P. W. 2005. Deblurring and denoising of images by nonlocal functionals. *Multiscale Modeling & Simulation*, 4(4): 1091–1115.

Lai, W.-S.; Huang, J.-B.; Ahuja, N.; and Yang, M.-H. 2017. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*, 624–632.

Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 4681–4690.

Li, W.; Tao, X.; Guo, T.; Qi, L.; Lu, J.; and Jia, J. 2020. MuCAN: Multi-Correspondence Aggregation Network for Video Super-Resolution. *arXiv preprint arXiv:2007.11803*.

Lim, B.; Son, S.; Kim, H.; Nah, S.; and Mu Lee, K. 2017. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 136–144.

Martin, D.; Fowlkes, C.; Tal, D.; and Malik, J. 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, volume 2, 416–423. IEEE.

Mathieu, M.; Couprie, C.; and LeCun, Y. 2016. Deep multi-scale video prediction beyond mean square error. In *ICLR*.

Matsui, Y.; Ito, K.; Aramaki, Y.; Fujimoto, A.; Ogawa, T.; Yamasaki, T.; and Aizawa, K. 2017. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20): 21811–21838.

Mechrez, R.; Talmi, I.; Shama, F.; and Zelnik-Manor, L. 2018. Maintaining natural image statistics with the contextual loss. In *ACCV*, 427–443. Springer.

Peleg, T.; and Elad, M. 2014. A statistical prediction model based on sparse representations for single image super-resolution. *TIP*, 23(6): 2569–2582.

Protter, M.; Elad, M.; Takeda, H.; and Milanfar, P. 2008. Generalizing the nonlocal-means to super-resolution reconstruction. *TIP*, 18(1): 36–51.

Sajjadi, M. S.; Scholkopf, B.; and Hirsch, M. 2017. Enhancenet: Single image super-resolution through automated texture synthesis. In *ICCV*, 4491–4500.

Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 1874–1883.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*.

Soh, J. W.; Park, G. Y.; Jo, J.; and Cho, N. I. 2019. Natural and realistic single image super-resolution with explicit natural manifold discrimination. In *CVPR*, 8122–8131.

Sønderby, C. K.; Caballero, J.; Theis, L.; Shi, W.; and Huszár, F. 2016. Amortised map inference for image super-resolution. *arXiv preprint arXiv:1610.04490*.

Tai, Y.; Yang, J.; and Liu, X. 2017. Image super-resolution via deep recursive residual network. In *CVPR*, 3147–3155.

Tai, Y.; Yang, J.; Liu, X.; and Xu, C. 2017. Memnet: A persistent memory network for image restoration. In *ICCV*, 4539–4547.

Timofte, R.; Agustsson, E.; Van Gool, L.; Yang, M.-H.; and Zhang, L. 2017. Ntire 2017 challenge on single image super-resolution: Methods and results. In *CVPRW*, 114–125.

Timofte, R.; De Smet, V.; and Van Gool, L. 2013. Anchored neighborhood regression for fast example-based super-resolution. In *ICCV*, 1920–1927.

Timofte, R.; De Smet, V.; and Van Gool, L. 2014. A+: Adjusted anchored neighborhood regression for fast super-resolution. In *Asian conference on computer vision*, 111–126. Springer.

Wang, X.; Yu, K.; Dong, C.; and Change Loy, C. 2018a. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, 606–615.

Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; and Change Loy, C. 2018b. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCV*, 0–0.

Wang, Y.; Perazzi, F.; McWilliams, B.; Sorkine-Hornung, A.; Sorkine-Hornung, O.; and Schroers, C. 2018c. A fully progressive approach to single-image super-resolution. In *CVPRW*, 864–873.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *TIP*, 13(4): 600–612.

Yang, C.-Y.; Huang, J.-B.; and Yang, M.-H. 2010. Exploiting self-similarities for single frame super-resolution. In *ACCV*, 497–510. Springer.

Yang, J.; Wang, Z.; Lin, Z.; Cohen, S.; and Huang, T. 2012. Coupled dictionary training for image super-resolution. *IEEE transactions on image processing*, 21(8): 3467–3478.

Yang, J.; Wright, J.; Huang, T.; and Ma, Y. 2008. Image super-resolution as sparse representation of raw image patches. In *CVPR*, 1–8. IEEE.

Yang, J.; Wright, J.; Huang, T. S.; and Ma, Y. 2010. Image super-resolution via sparse representation. *TIP*, 19(11): 2861–2873.

Zeyde, R.; Elad, M.; and Protter, M. 2010. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, 711–730. Springer.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018a. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, W.; Liu, Y.; Dong, C.; and Qiao, Y. 2019. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *ICCV*, 3096–3105.

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018b. Image super-resolution using very deep residual channel attention networks. In *ECCV)*, 286–301.

Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; and Fu, Y. 2018c. Residual dense network for image super-resolution. In *CVPR*, 2472–2481.