This CVPR 2020 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Unpaired Image Super-Resolution using Pseudo-Supervision

Shunta Maeda
Navier Inc.
shunta@navier.co.jp

Abstract

In most studies on learning-based image super-resolution (SR), the paired training dataset is created by downscaling high-resolution (HR) images with a predetermined operation (e.g., bicubic). However, these methods fail to super-resolve real-world low-resolution (LR) images, for which the degradation process is much more complicated and unknown. In this paper, we propose an unpaired SR method using a generative adversarial network that does not require a paired/aligned training dataset. Our network consists of an unpaired kernel/noise correction network and a pseudo-paired SR network. The correction network removes noise and adjusts the kernel of the inputted LR image; then, the corrected clean LR image is upscaled by the SR network. In the training phase, the correction network also produces a pseudo-clean LR image from the inputted HR image, and then a mapping from the pseudo-clean LR image to the inputted HR image is learned by the SR network in a paired manner. Because our SR network is independent of the correction network, well-studied existing network architectures and pixel-wise loss functions can be integrated with the proposed framework. Experiments on diverse datasets show that the proposed method is superior to existing solutions to the unpaired SR problem.

1. Introduction

Image super-resolution (SR) is a fundamental ill-posed problem in low-level vision that reconstructs a high-resolution (HR) image from its low-resolution (LR) observation. Recent progress in the on deep learning-based methods has significantly improved the performance of SR, increasing attention from the practical perspective. However, in many studies, training image pairs are generated by a predetermined downscaling operation (*e.g.*, bicubic) on the HR images. This method of dataset preparation is not practical in real-world scenarios because there is usually no HR image corresponding to the given LR one.

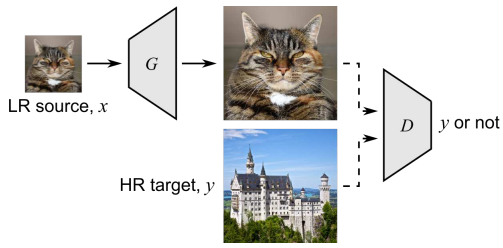
Some recent studies have proposed methods to overcome the absence of HR–LR image pairs, such as blind SR methods [39, 12, 57] and generative adversarial network (GAN)-based unpaired SR methods [51, 4, 56, 32]. Blind SR aims to reconstruct HR images from LR ones degraded by arbitrary kernels. Although recent studies have achieved "blindness" for limited forms of degradation (*e.g.*, blur), real LR images are not always represented with such degradation; thus, they perform poorly on the images degraded by not expected processes. By contrast, GAN-based unpaired SR methods can directly learn a mapping from LR to HR images without assuming any degradation processes.



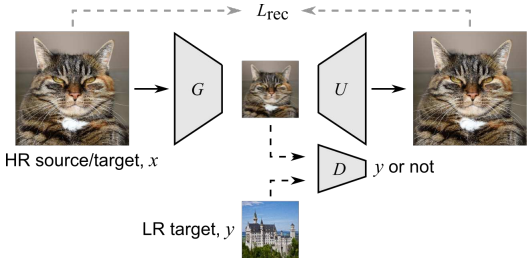
Figure 1: **Unpaired SR results on diverse datasets.** From left to right: $\times 4$ SR result for a synthetically degraded LR image from DIV2K realistic-wild set [42], $\times 4$ SR result for a real-world LR face image from Widerface [48], and $\times 2$ SR result for a real-world LR aerial image from DOTA [47]. Zoom in for better view.

GANs learn to generate images with the same distribution as the target domain through a minimax game between a generator and discriminator [11, 37]. GAN-based unpaired SR methods can be roughly classified according to whether they start from an LR image (direct approach; Fig. 2a) or an HR image (indirect approach; Fig. 2b).

Direct approach. In this approach, a generator upscales source LR images to fool an HR discriminator [51]. The main drawback of this approach is that the pixel-wise loss functions cannot be used to train the generator, *i.e.*, SR network. In paired SR methods, the pixel-wise loss between reconstructed images and HR target images plays a crucial role not only in distortion-oriented methods but also in perception-oriented methods [28, 2].



(a) Direct approach.



(b) Indirect approach.

Figure 2: **Two distinct approaches to unpaired SR using GANs.** (a) Generator directly upscales LR images. (b) Generator first downscales HR images and uses the generated LR images to train SR network U.

Indirect approach. In this approach, a generator downscales source HR images to fool an LR discriminator [4, 32]. The generated LR images are then used to train the SR network in a paired manner. The main drawback of this approach is that the deviation between the generated LR distribution and the true LR distribution causes train–test discrepancy, degrading the test time performance.

Our approach. The main contribution of this work is that we simultaneously overcome the drawbacks of the above two approaches by separating the entire network into an unpaired kernel/noise correction network and a *pseudo-paired* SR network (Fig. 3). The correction network is a CycleGAN [58]-based unpaired LR $\leftrightarrow$ *clean* LR translation. The SR network is a paired *clean* LR $\rightarrow$ HR mapping, where the *clean* LR images are created by downscaling the HR images with a predetermined operation. In the training phase, the correction network also generates *pseudo-clean* LR images by first mapping the *clean* LR images to the true LR domain and then pulling them back to the *clean* LR domain. The SR network is learned to reconstruct the original HR images from the *pseudo-clean* LR images in a paired manner. With the following two merits, our method achieves superior results to state-of-the-arts ones: (1) Because our correction network is trained on not only the generated LR images but also the true LR images through the bi-directional structure, the deviation between the generated LR distribution and the true LR distribution does not critically degrade the test time performance. (2) Any existing SR networks and pixel-wise loss functions can be integrated because our SR network is separated to be able to learn in a paired manner.

2. Related Work

The training data, network architecture, and objective function are three essential elements of a learning deep network. Paired image SR is aimed at optimizing the network architecture and/or objective function to improve performance under the assumption that ideal training data exists. However, in many practical cases, there is a lack of training data (*i.e.*, target HR images corresponding to source LR images). This problem has been addressed by recent studies on blind and unpaired image SR. As another approach, a few recent works [7, 54, 5] have built real paired SR datasets using specialized hardware and data correction processes, which are difficult to scale.

2.1. Paired Image Super-Resolution

In most SR studies, the paired training dataset is created by downscaling HR images with a predetermined operation (*e.g.*, bicubic). Since the first convolutional neural network (CNN)-based SR network [9], various SR networks have been proposed to improve LR-to-HR reconstruction performance. Early studies [20, 30] found that a deeper network performs better with residual learning. A proposed residual channel attention network (RCAN) [55] achieved further improved depth and performance. Upscaling strategies have also been studied, such as progressive upscaling of LapSRN [26] and iterative upscaling and downscaling of DBPN [13]. In these studies, a simple L1 or L2 distance was used as the objective function, but it is known that these simple distances alone result in blurred textures. To improve the perceptual quality, SRGAN [28] introduced perceptual loss [18] and adversarial loss [11], realizing more visually pleasing results. ESRGAN [44], which is an enhanced version of SRGAN, is one of the state-of-the-art perception-oriented models.

2.2. Blind Image Super-Resolution

Relatively less research attention has been paid to blind image SR despite its importance for practical applications. Studies on blind SR usually focus on models that are only blind to the blur kernels [34, 38, 39, 12, 57]. For instance, ZSSR [39] exploits the recurrence of information inside a single image to upscale images with different blur kernels, and IKC [12] uses the intermediate outputs to iteratively correct the mismatch of blur kernels. Few studies on blind SR have addressed the combined degradation problem (*i.e.*, additive noise, compression artifacts, etc.) beyond the blur blind SR, whereas several blind methods have been proposed for specific degradation problems, such as denoising [23] and motion deblurring [35, 24].

2.3. Unpaired Image Super-Resolution

A few recent works have addressed the SR problem without using a paired training dataset. Different from the un-

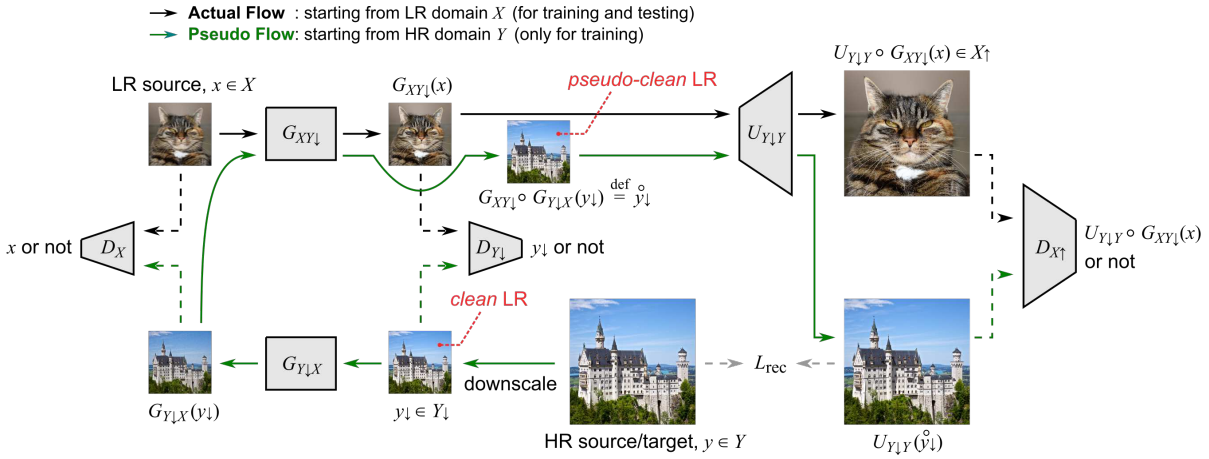


Figure 3: **Data-flow diagram of proposed method.** SR network $U_{Y_\downarrow Y}$ can be learned in a paired manner through $\mathcal{L}_{rec}$, even if the training dataset $\{X, Y\}$ is not paired. The whole network is end-to-end trainable.

paired translation methods, such as CycleGAN [58] and DualGAN [49], unpaired SR aims to upscale source LR images while preserving style and local structure. Bulat *et al*. [4] and Lugmayr *et al*. [32] first trained a high-to-low degradation network and then used the degraded outputs to train a low-to-high SR network. Yuan *et al*. [51] proposed a cycle-in-cycle network to simultaneously learn a degradation network and an SR network. Different from our method, the degradation network of Yuan *et al*. is deterministic, and the SR network is incorporated with the bi-cycle network; thus, the usable loss function is limited. Zhao *et al*. [56] also jointly stabilized the training of a degradation network and SR network by utilizing a bi-directional structure. Similar to Yuan *et al*., the SR network of Zhao *et al*. has a limited degree of freedom to select the loss function.

3. Proposed Method

Our goal is to learn a mapping F_{XY} from an LR source domain X to an HR target domain Y based on the given unpaired training samples x ($\in X$) and y ($\in Y$). Here, we define "*clean* LR," *i.e*., HR images downscaled with a predetermined operation, as $y_\downarrow$ ($\in Y_\downarrow$). The downscaling operation $Y \to Y_\downarrow$ used is a combination of Gaussian blur with $\sigma = (\text{scale factor})/2$ and bicubic downscaling. The mapping F_{XY} of our model is a combination of the two mappings $G_{XY_\downarrow}$ and $U_{Y_\downarrow Y}$, where $G_{XY_\downarrow}$ is a mapping from X to $Y_\downarrow$, and $U_{Y_\downarrow Y}$ is an upscaling mapping from $Y_\downarrow$ to Y. Figure 3 illustrates the proposed framework.

Domain transfer in LR. We use a CycleGAN [58]-based model for the domain transfer in LR. Two generators, $G_{XY_\downarrow}$ and its inverse mapping $G_{Y_\downarrow X}$, are simultaneously learned to enforce cycle consistency, *i.e*. $G_{XY_\downarrow} \circ G_{Y_\downarrow X}(y_\downarrow) \approx y_\downarrow$[1]. The training of the generator $G_{XY_\downarrow}$ ($G_{Y_\downarrow X}$) requires a discriminator $D_{Y_\downarrow}$ (D_X) that aims to detect translated samples from the real examples $y_\downarrow$ (x).

Mapping from LR to HR. The upscaling mapping $U_{Y_\downarrow Y}$ is learned to reconstruct HR image y from a *pseudo-clean* LR image $G_{XY_\downarrow} \circ G_{Y_\downarrow X}(y_\downarrow)$ in a paired manner. Thus, any pixel-wise loss functions can be used to train $U_{Y_\downarrow Y}$. Hereafter, we denote $G_{XY_\downarrow} \circ G_{Y_\downarrow X}(y_\downarrow)$ as "$\mathring{y}_\downarrow$".

Adjustment with HR Discriminator. While $\mathring{y}_\downarrow$ is used to train $U_{Y_\downarrow Y}$, the actual input at test time is $G_{XY_\downarrow}(x)$. Accordingly, $\mathring{y}_\downarrow \sim G_{XY_\downarrow}(x)$ is required to minimize the train–test discrepancy. Although this requirement is satisfied to some extent by the normal CycleGAN, we introduce an additional discriminator $D_{X_\uparrow}$, which takes the output of $U_{Y_\downarrow Y}$ as input so that $U_{Y_\downarrow Y}(\mathring{y}_\downarrow)$ gets closer to $U_{Y_\downarrow Y} \circ G_{XY_\downarrow}(x)$. Here, we define $X_\uparrow$ as a domain consisting of $U_{Y_\downarrow Y} \circ G_{XY_\downarrow}(x)$. Thus, $X_\uparrow$ is an unfixed domain that shifts during training. Note that $D_{X_\uparrow}$ updates the parameters of the two generators, and $U_{Y_\downarrow Y}$ is simply used as an amplifier of local image features.

3.1. Loss Functions

Adversarial loss. We impose an adversarial constraint [11] on both generators $G_{XY_\downarrow}$ and $G_{Y_\downarrow X}$. As a specific example, an adversarial loss for $G_{XY_\downarrow}$ and $D_{Y_\downarrow}$ is expressed as

$$\begin{aligned}\mathcal{L}_{adv}(G_{XY_\downarrow}, D_{Y_\downarrow}, X, Y_\downarrow) = \mathbb{E}_{y_\downarrow \sim \mathrm{P}_{Y_\downarrow}}[\log D_{Y_\downarrow}(y_\downarrow)] \\ + \mathbb{E}_{x \sim \mathrm{P}_X}[\log(1 - D_{Y_\downarrow}(G_{XY_\downarrow}(x)))],\end{aligned} \tag{1}$$

where P_X ($\mathrm{P}_{Y_\downarrow}$) is the data distribution of the domain X ($Y_\downarrow$). $G_{XY_\downarrow}$ and $D_{Y_\downarrow}$ simultaneously optimized each other through a mini-max game between them, *i.e*., $\min_{G_{XY_\downarrow}} \max_{D_{Y_\downarrow}} \mathcal{L}_{adv}(G_{XY_\downarrow}, D_{Y_\downarrow}, X, Y_\downarrow)$. Similar to the CycleGAN framework, the inverse mapping $G_{Y_\downarrow X}$ and its corresponding discriminator D_X are also optimized: $\min_{G_{Y_\downarrow X}} \max_{D_X} \mathcal{L}_{adv}(G_{Y_\downarrow X}, D_X, Y_\downarrow, X)$.

In our framework, the two generators are also optimized

[1] $F \circ G(x) := F(G(x))$.

through an HR discriminator $D_{X_\uparrow}$:

$$\begin{aligned}&\mathcal{L}_{adv}((G_{XY_\downarrow}, G_{Y_\downarrow X}), D_{X_\uparrow}, Y_\downarrow, X_\uparrow)\\ &= \mathbb{E}_{x\sim \mathrm{P}_X}[\log D_{X_\uparrow}(U_{Y_\downarrow Y} \circ G_{XY_\downarrow}(x))]\\ &+ \mathbb{E}_{y_\downarrow \sim \mathrm{P}_{Y_\downarrow}}[\log(1 - D_{X_\uparrow}(U_{Y_\downarrow Y}(\mathring{y}_\downarrow)))].\end{aligned} \tag{2}$$

The optimization process of Eq. 2 is expressed as $\min_{G_{XY_\downarrow}, G_{Y_\downarrow X}} \max_{D_{X_\uparrow}} \mathcal{L}_{adv}((G_{XY_\downarrow}, G_{Y_\downarrow X}), D_{X_\uparrow}, Y_\downarrow, X_\uparrow)$.

Cycle consistency loss. The normal CycleGAN learns one-to-one mappings because it imposes cycle consistency on both cycles (*i.e.*, $X \to Y \to X$ and $Y \to X \to Y$). We relax this restriction by requiring cycle consistency for only one side:

$$\mathcal{L}_{cyc}(G_{Y_\downarrow X}, G_{XY_\downarrow}) = \|G_{XY_\downarrow} \circ G_{Y_\downarrow X}(y_\downarrow) - y_\downarrow\|_1. \tag{3}$$

Under the above one-side cycle consistency, the mapping $G_{Y_\downarrow X}$ is allowed to be one-to-many. Consequently, our framework can deal with various noise types/distributions of the LR source domain X.

Identity mapping loss. An identity mapping loss was introduced in the original CycleGAN to preserve color composition for a task of painting $\to$ photo. We also impose the identity mapping loss for $G_{XY_\downarrow}$ to avoid color variation:

$$\mathcal{L}_{idt}(G_{XY_\downarrow}) = \|G_{XY_\downarrow}(y_\downarrow) - y_\downarrow\|_1. \tag{4}$$

Geometric ensemble loss. Geometric consistency, which was introduced in a recent work [10], reduces the space of possible translation to preserve the scene geometry. Inspired by geometric consistency, we introduce a simple geometric ensemble loss that requires the flip and rotation for the input images not to change the result:

$$\mathcal{L}_{geo}(G_{XY_\downarrow}) = \|G_{XY_\downarrow}(x) - \sum_{i=1}^{8} T_i^{-1}(G_{XY_\downarrow}(T_i(x)))/8\|_1, \tag{5}$$

where the operators $\{T_i\}_{i=1}^8$ represent eight distinct patterns of flip and rotation. Note that using $\mathcal{L}_{geo}$ increases the total training time by a factor of approximately $3/2$.

Full objective. Our full objective for the two generators and three discriminators is as follows:

$$\begin{aligned}\mathcal{L}_{trans} = &\ \mathcal{L}_{adv}(G_{XY_\downarrow}, D_{Y_\downarrow}, X, Y_\downarrow)\\ &+ \mathcal{L}_{adv}(G_{Y_\downarrow X}, D_X, Y_\downarrow, X)\\ &+ \gamma\mathcal{L}_{adv}((G_{XY_\downarrow}, G_{Y_\downarrow X}), D_{X_\uparrow}, Y_\downarrow, X_\uparrow)\\ &+ \lambda_{cyc}\mathcal{L}_{cyc}(G_{Y_\downarrow X}, G_{XY_\downarrow})\\ &+ \lambda_{idt}\mathcal{L}_{idt}(G_{XY_\downarrow})\\ &+ \lambda_{geo}\mathcal{L}_{geo}(G_{XY_\downarrow}),\end{aligned} \tag{6}$$

where the hyperparameters λ_{cyc}, λ_{idt}, λ_{geo}, and γ weight the contributions of each objective.

While the SR network $U_{Y_\downarrow Y}$ is independent of the generators and discriminators, it is used as an amplifier of the local features of images to be inputted to $D_{X_\uparrow}$. Thus, we jointly update the SR network during the training of the correction network. We use L1 loss to reconstruct an HR image from a *pseudo-clean* LR image $\mathring{y}_\downarrow$:

$$\mathcal{L}_{rec} = \|U_{Y_\downarrow Y}(\mathring{y}_\downarrow) - y\|_1. \tag{7}$$

We again note that any pixel-wise loss (*e.g.*, perceptual loss, texture loss, and adversarial loss) can be used as $\mathcal{L}_{rec}$ in our formulation.

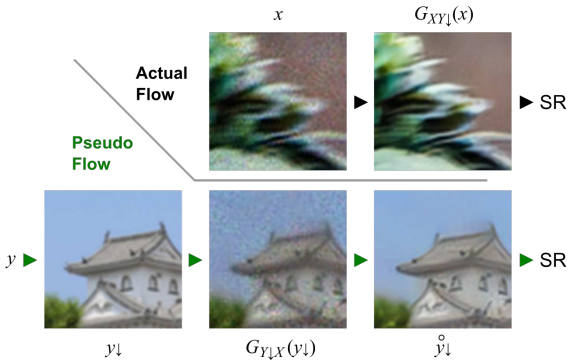


Figure 4: **Intermediate images of proposed method.** x is image "0886" from the DIV2K realistic-wild validation set, and y is image "0053" from the DIV2K training ground-truth set.

3.2. Network Architecture

$G_{XY_\downarrow}$ and $U_{Y_\downarrow Y}$. We utilize an RCAN [55]-based architecture as $G_{XY_\downarrow}$ and $U_{Y_\downarrow Y}$. The RCAN is a very deep SR network realized by a residual in residual structure with short and long skip connections. The RCAN consists of 10 residual groups (RGs), where each RG contains 20 residual channel attention blocks (RCABs). Our $G_{XY_\downarrow}$ ($U_{Y_\downarrow Y}$) is a reduced version of the RCAN consisting of five RGs with 10 (20) RCABs. Note that the final upscaling layer included in the original RCAN is omitted for $G_{XY_\downarrow}$.

$G_{Y_\downarrow X}$. For the generator $G_{Y_\downarrow X}$, we use several residual blocks with 5×5 filters and several fusion layers with 1×1 filters, where each convolution layer is followed by batch normalization (BN) [16] and LeakyReLU. The two head modules, including the one residual block, independently extract the features of an inputted RGB image and single-channel random noise $\mathcal{N}(0, 1)$ that simulates the randomness of distortions. Then, the two extracted features are concatenated to be inputted to a main module consisting of six residual blocks and three fusion layers.

D_X, $D_{Y_\downarrow}$ and $D_{X_\uparrow}$. For the LR discriminators D_X and $D_{Y_\downarrow}$, we use five convolution layers with strides of 1. The convolution layers, except for the last layer, are followed by LeakyReLU without BN. A similar architecture is also used for the HR discriminator $D_{X_\uparrow}$ but with a different stride in the initial layers. For the case of scale facter $= 2$ (4), strides of 2 are used for the first (and second) layer(s) of $D_{X_\uparrow}$. We use PatchGAN [29, 17] for all the discriminators.

	Method	PSNR	SSIM
Blind denoising/deblurring + Bicubic upscaling	Bicubic (for reference)	19.99	0.4857
	NC [27] + Bicubic	20.03	0.5049
	RL-restore [50] + Bicubic	20.18	0.5119
	RL-restore [50] + SRN-Deblur [40] + Bicubic	20.13	0.5173
	RL-restore [50] + DeblurGAN-v2 [25] + Bicubic	20.21	0.5158
Blind denoising/deblurring + non Blind SR method	DBPN [13] (for reference)	19.82	0.4572
	RL-restore [50] + DeblurGAN-v2 [25] + DBPN [13]	20.25	0.5198
Blind denoising/deblurring + Blind SR method	ZSSR [39]	19.91	0.4835
	ZSSR [39] w/ KernelGAN [1]	19.45	0.4493
	IKC [12]	19.62	0.4251
	RL-restore [50] + DeblurGAN-v2 [25] + ZSSR [39]	20.19	0.5217
	RL-restore [50] + DeblurGAN-v2 [25] + ZSSR [39] w/ KernelGAN [1]	19.83	0.5137
	RL-restore [50] + DeblurGAN-v2 [25] + IKC [12]	20.26	0.5140
	Our method	21.32	0.5541

Table 1: **Numerical comparison with state-of-the-art blind methods on DIV2K realistic-wild validation set (SR scale $\times 4$).** The best and second-best results are highlighted in red and blue, respectively. We use the officially provided evaluation script[2](validation stage setting). Throughout this paper, the real configuration is used for ZSSR, and the Inception backbone model is used for DeblurGAN-v2.



Figure 5: **Qualitative comparison with state-of-the-art blind methods on DIV2K realistic-wild validation set (SR scale $\times 4$).** Our method reconstructs the fine details while removing artifacts, yielding the most visually pleasing results.

4. Experiments

4.1. Network Training

We used the Adam optimizer [21] with $\beta_1 = 0.5$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$ to train the generators and discriminators. The SR network $U_{Y_\downarrow Y}$ was similarly trained but with a different β_1 ($= 0.9$). The learning rates of all networks were initialized to 1×10^{-4}. Then, the learning rates of the networks other than $U_{Y_\downarrow Y}$ were halved at 100k, 180k, 240k, and 280k iterations. We trained our networks for more than 3×10^5 iterations with a mini-batch size of 16. In each iteration, LR patches of 32×32 and HR patches of corresponding size were extracted as inputs in an unaligned manner. Then, data augmentation of random flip and rotation was performed on each training patch. We used PyTorch [36] to conduct all the experiments.

4.2. Experiments on Synthetic Distortions

DIV2K realistic-wild dataset. We used the realistic-wild set (Track 4) of the NTIRE 2018 Super-Resolution Challenge [42]. The realistic-wild set was generated by degrading DIV2K [41], consists of 2K resolution images that are diverse in their content. DIV2K has 800 training images. The realistic-wild set simulates real "wild" LR images via $\times 4$ downscaling, motion blurring, pixel shifting, and noise addition. The degradation operations are the same within a single image but vary from image to image. Four degraded LR images are generated for each DIV2K training image (*i.e.*, 3,200 LR training images in total). We trained our model using the above 3,200 LR and 800 HR paired images but with "unpaired/unaligned" sampling. We evaluated the results on the 100 realistic-wild validation images because the ground truths of the testing images were not provided.

Hyperparameters. We used the loss hyperparameters $\lambda_{cyc} = 1$, $\lambda_{idt} = 1$, $\lambda_{geo} = 1$, and $\gamma = 0.1$ throughout the experiments in this subsection. The SR factor was $\times 4$.

Intermediate images. Figure 4 shows visual examples of the intermediate images of the proposed method. The degradation network $G_{Y_\downarrow X}$ degrades the input *clean* LR image $y_\downarrow$ so that the output $G_{Y_\downarrow X}(y_\downarrow)$ reproduces the noise distribution of the real degraded image x. The reconstruction network $G_{XY_\downarrow}$ is effective in removing the noises of

Username	PSNR	SSIM
xixihaha	24.12	0.56
yyuan13	24.07	0.56
Hot_Milk	23.90	0.56
yifita	23.87	0.56
cskzh	23.55	0.55
JSChoi	23.20	0.53
enoch	23.04	0.52
assafsho	22.93	0.51
hyu_ss	22.57	0.49
cr2018	22.52	0.49
Ours	21.32	0.5541
Ours$^+$	21.35	0.5560

Table 2: **Comparison with NTIRE 2018 baselines.** Top 10 validation results from NTIRE 2018 realistic-wild challenge website are compared as pair-trained upper bounds. Ours$^+$ is an enhanced version of Ours using a standard self-ensemble technique [43].

Method	PSNR	SSIM
Ours	21.32	0.5541
Ours - w/o $D_{X_\uparrow}$	21.29	0.5532
Ours - trained on $y_\downarrow$	21.09	0.5312
Ours - trained on $G_{Y_\downarrow X}(y_\downarrow)$	20.84	0.5500
Ours - trained on $G_{Y_\downarrow X}(y_\downarrow)$ - original RCAN	20.78	0.5482

Table 3: **Ablation study.** Some other variants of our network were compared to verify the proposed method.

both the real (x) and fake ($G_{Y_\downarrow X}(y_\downarrow)$) degraded images.

Comparison with state-of-the-art blind methods. Because the blind SR method for multiple degradations has not been studied sufficiently, we took a benchmark by combining the SR method with the blind restoration methods (Tab. 1, Fig. 5). We first explored the state-of-the-art blind denoising methods: a patch-based method NC [27] and a CNN-based method RL-restore [50]. RL-restore performed better than NC. Then, we compared two CNN-based blind deblurring methods, SRN-Deblur [40] and DeblurGAN-v2 [25], based on the output of RL-restore. The performances of these deblurring methods were almost equivalent, but DeblurGAN-v2 ran faster. Finally, three state-of-the-art SR methods were combined with RL-restore and DeblurGAN-v2: a non-blind SR method DBPN [13] and two blind SR methods ZSSR [39] and IKC [12]. We further combined ZSSR with the recently proposed kernel estimation method KernelGAN [1]. Our method outperformed all of the above methods by a large margin; however, the comparison was not completely fair because the compared methods were not trained on the dataset used here.

Comparison with NTIRE 2018 baselines. Table 2 shows a comparison with NTIRE 2018 baselines from the validation website[3], where the dataset and evaluation script used were the same as in our experiment. Note that although the NTIRE 2018 competition provides a paired training dataset, we trained our network in an unpaired manner. Thus, the NTIRE 2018 baselines can be regarded as pair-trained upper bounds. Our result is inferior to the upper bounds in PSNR, but the result of the more sophisticated indicator SSIM [46] is comparable to the upper bounds. Because PSNR overestimates slight differences in global brightness and/or color that do not significantly affect the perceptual quality [45], we believe our method shows practically equivalent performance to the pair-trained upper bounds.

[2] https://data.vision.ee.ethz.ch/cvl/DIV2K/

[3] https://competitions.codalab.org/competitions/18026

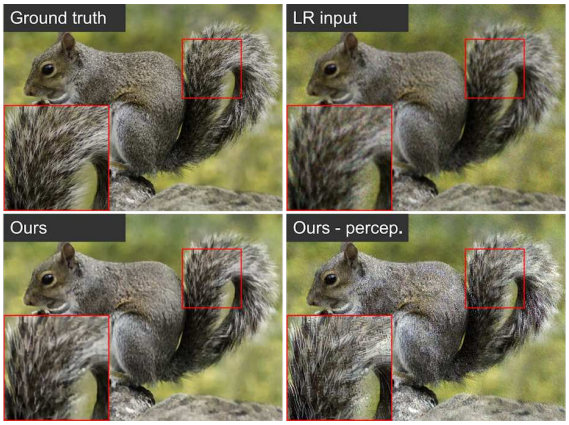


Figure 6: **Example image with perception-oriented training.** "0810" from DIV2K realistic-wild validation set is displayed.

Ablation study. To investigate the effectiveness of the proposed method, we designed some other variants of our network: (1) Ours - w/o $D_{X_\uparrow}$, where the HR discriminator $D_{X_\uparrow}$ is removed (*i.e.* $\gamma = 0$), (2) Ours - trained on $y_\downarrow$, where the SR network $U_{Y_\downarrow Y}$ is trained on $y_\downarrow$ instead of $\mathring{y}_\downarrow$, which is equivalent to a simple combination of a style translation network and SR network, and (3) Ours - trained on $G_{Y_\downarrow X}(y_\downarrow)$, where the SR network $U_{Y_\downarrow Y}$ is trained on $G_{Y_\downarrow X}(y_\downarrow)$ instead of $\mathring{y}_\downarrow$ and only $U_{Y_\downarrow Y}$ is used at testing time, which is equivalent to the indirect approach illustrated in Fig. 2b. For completeness, the variant (3) was validated using the original RCAN model as the SR network, which is larger than our total testing network $U_{Y_\downarrow Y} \circ G_{XY_\downarrow}$. These variants underperformed compared to the proposed method (Tab. 3). In particular, our full model outperformed variant (3), meaning that the proposed *pseudo-supervision* is effective at reducing the train–test discrepancy.

Perception-oriented training. We also trained our model with a perception-oriented reconstruction loss following ESRGAN [44] to demonstrate the versatility of our method. We replaced Eq. 7 with a combination of perceptual loss, relativistic adversarial loss [19], and content loss as in ESRGAN, while the other loss functions and training procedure were unchanged. The perceptually trained model gives a more visually pleasing result than the normal model trained with L1 reconstruction loss (Fig. 6).

4.3. Experiments on Realistic Distortions I

Large-scale face image dataset. In this subsection, we follow the experimental procedure described by Bulat *et*

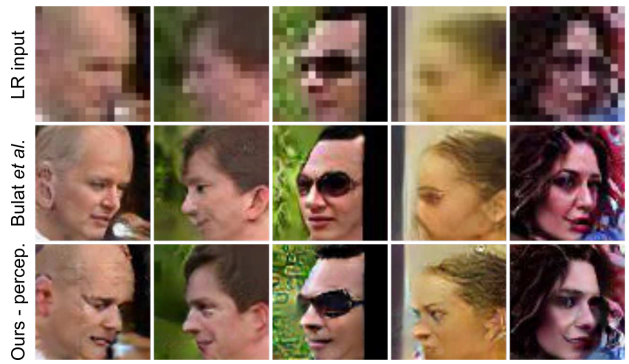


Figure 7: **Qualitative comparison with a state-of-the-art GAN-based unpaired SR method (SR scale $\times 4$).** Input LR images are from the LR test set provided by Bulat *et al*.

Method	FID
SRGAN [28]	104.80
CycleGAN [58]	19.01
DeepDeblur [35]	294.96
Wavelet-SRNet [15]	149.46
FSRNet [8]	157.29
Bulat *et al*. [4]	14.89
Ours - perceptual	13.57

Table 4: **FID-based performance comparison with state-of-the-art methods.** The dataset and evaluation script provided by Bulat *et al*. were used. Lower scores indicate better results.

al. [4]. We used the dataset and evaluation script they provided[4]. They collected 182,866 HR face images from the Celeb-A [31], AFLW [22], LS3D-W [3], and VGGFace2 [6]. They also collected more than 50,000 real-world LR face images from Widerface [48] that are diverse in degradation types. 3,000 images were randomly selected from the LR dataset and kept for testing. Then, all HR and LR face images were cropped in a consistent manner using the face detector [53]. The cropped HR and LR training images were 64×64 and 16×16 patches, respectively.

Hyperparameters. For the experiments on realistic distortions, we found that it is better to take x instead of $y_\downarrow$ as an argument of the identity mapping loss. Thus, we used the modified identity mapping loss

$$\mathcal{L}_{\overline{idt}}(G_{XY_\downarrow}) = \|G_{XY_\downarrow}(x) - x\|_1 \tag{8}$$

instead of Eq. 4 in the following. We used the loss hyperparameters $\lambda_{cyc} = 1$, $\lambda_{\overline{idt}} = 2$, $\lambda_{geo} = 1$, and $\gamma = 0.1$. We first upscaled all the 16×16 LR patches by a factor of two using a bicubic method because the original size was too small. Then, our network was trained on the 32×32 LR patches and 64×64 HR patches with an SR factor of $\times 2$.

Comparison with state-of-the-art methods. We numerically and qualitatively compared our method with the state-of-the-art GAN-based unpaired method proposed by Bulat *et al*. [4]. Our method was also numerically compared

[4] https://github.com/jingyang2017/Face-and-Image-super-resolution



Figure 8: **One-to-many degradation examples.** Examples of different LR images generated by our degradation network for different random noise input.

with five related state-of-the-art methods: image SR method SRGAN [28], face SR methods Wavelet-SRNet [15] and FSRNet [8], unpaired image translation method CycleGAN [58], and deblurring method DeepDeblur [35]. Please see Ref. [4] for a more detailed explanation of each method.

Table 4 shows a numerical comparison with the related state-of-the-art methods. We assessed the quality of the SR results with the Fréchet inception distance (FID) [14] because there were no corresponding ground-truth images. CycleGAN, Bulat *et al*.'s, and our method, which are GAN-based unpaired approaches, largely outperformed all other methods. Besides, our method showed better performance than CycleGAN and Bulat *et al*.'s. For completeness, we calculated PSNR between the bicubically upscaled LR test images and its SR results. The calculated PSNRs for the results of Bulat *et al*. and our method were 20.28 dB and 21.09 dB, respectively. These numerical results indicate that our method produces perceptually better results than Bulat *et al*.'s while maintaining the characteristics of the input images. A qualitative comparison is shown in Fig. 7.

One-to-many degradation examples. Visual examples expressing the various noise intensities/types learned by our degradation network $G_{Y_\downarrow X}$ shown in Fig. 8. The one-sided cycle consistency allows the mapping $G_{Y_\downarrow X}$ to become one-to-many, reproducing the various noise distributions of the real LR images.

4.4. Experiments on Realistic Distortions II

DOTA and Inria aerial image dataset. We used two aerial image datasets with different ground sample distances (GSD) as source and target. We sampled 62 LR source images with the GSD in the range [55cm, 65cm] from a training set of DOTA [47] (a large-scale aerial image dataset for object detection collected from different sensors and platforms). For the HR target, we used the Inria aerial image labeling dataset [33], which contains scenes from several different cities but with the same GSD (30 cm). Note that we used only the images of Vienna city (36 images) so that the qualities of the target images are constant.

Hyperparameters. We used the loss hyperparameters $\lambda_{cyc} = 1$, $\lambda_{\overline{idt}} = 10$, $\lambda_{geo} = 100$, and $\gamma = 0.1$. The SR factor was $\times 2$. In aerial images, the pixel sizes of objects such as vehicles and buildings are rather small, thus we used larger $\lambda_{\overline{idt}}$ and λ_{geo} than in the other experiments



"P2331" from DOTA | LR input | RL-restore + Bicubic | RL-restore + DBPN | ZSSR | Ours

Figure 9: **Qualitative comparison with state-of-the-art blind methods on DOTA validation set (SR scale $\times 2$).** Zoom in for better view.



LR input | w/o $\mathcal{L}_{geo}$ | w/ $\mathcal{L}_{geo}$

Figure 10: **Effect of geometric ensemble loss.** "P2768" from DOTA validation set is displayed as example.

to maintain the local characteristics of the images. We gradually elevated λ_{geo} in the early stages of training to avoid a mode where the entire image is uniform.

Comparison with state-of-the-art methods. We only provide a qualitative comparison in this subsection (Fig. 9) because there are no ground-truth HR images. The input LR image was sampled from the DOTA validation set with the GSD in the range of [55cm, 65cm]. As the benchmark, a CNN-based blind denoising method RL-restore [50] was first tested because the input LR images contained visible artifacts. RL-restore successfully removed the artifacts, but the fine details of the inputted images were removed as well. The over-smoothed output of RL-restore is slightly enhanced by applying a state-of-the-art SR method DBPN [13]. A state-of-the-art blind SR method ZSSR [39] was also tested, but the artifacts were not completely removed. Unlike the above methods, our method super-resolves the fine details while removing the artifacts, yielding the most visually reasonable results.

Effect of geometric ensemble loss. We compared the SR results with and without the geometric ensemble loss $\mathcal{L}_{geo}$ to confirm the effectiveness of $\mathcal{L}_{geo}$. Example images for visual comparison are shown in Fig. 10. It can be seen that the method without $\mathcal{L}_{geo}$ produces geometrically inconsistent results. By enforcing geometry consistency through $\mathcal{L}_{geo}$, our method results in more reasonable mapping, preserving geometrical structures of the input LR image.

4.5. Additional Experiment

We conducted an additional experiment on the dataset provided in the recent AIM 2019 Real-World Super-Resolution Challenge, where no training HR–LR image pairs are available. We focused on Track 2 of the challenge which is a more general setting than Track 1 (see the competition website[5] for more details). We compared our method with Lugmayr *et al.* [32] which is a study of a GAN-based unpaired SR method (indirect approach; Fig. 2b) recently published by the organizers of the challenge. As shown in Tab. 5, our method achieved superior scores in both the distortion metrics (PSNR, SSIM) and the perception metric (LPIPS [52]; lower is better). The visual results are provided in the supplemental material.

Method	PSNR	SSIM	LPIPS
ZSSR [39]	22.42	0.61	0.5996
ESRGAN [44]	20.69	0.51	0.5604
Lugmayr *et al.* [32]	21.59	0.55	0.4720
Ours	22.88	0.6612	0.4539
Ours$^+$	23.01	0.6655	0.4567

Table 5: **Additional experiment on AIM 2019 Real-World Super-Resolution Challenge dataset (Track 2).** An officially provided evaluation script was used for PSNR and SSIM calculations. A version of the LPIPS script used to evaluate our method is v0.1 (v0.0 outputs lower LPIPS value). We used the loss hyperparameters $\lambda_{cyc} = 1$, $\lambda_{\overline{idt}} = 5$, $\lambda_{geo} = 1$, and $\gamma = 0.1$.

5. Conclusion

We investigated the SR problem in an unpaired setting where the aligned HR–LR training set is unavailable. Our network produces *pseudo-clean* LR images as the intermediate products from ground-truth HR images, which are then used to train the SR network in a paired manner (referred to as "*pseudo-supervision*" in this paper). In this sense, the proposed method bridges the gap between the well-studied existing SR methods and the real-world SR problem without paired datasets. The effectiveness of our method was demonstrated by extensive experiments on diverse datasets: synthetically degraded natural images (Sec. 4.2, 4.5), real-world face images (Sec. 4.3), and real-world aerial images (Sec. 4.4).

While the proposed method is applicable to diverse datasets, hyperparameter tuning is necessary for each case to maximize the performance. Making the network more robust against the hyperparameters will be future work.

[5] https://competitions.codalab.org/competitions/20164

References

[1] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. *arXiv preprint arXiv:1909.06581*, 2019. 5, 6

[2] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *ECCV Workshop*, 2018. 1

[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *ICCV*, 2017. 7

[4] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. To learn image super-resolution, use a gan to learn how to do image degradation first. In *ECCV*, 2018. 1, 2, 3, 7

[5] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *ICCV*, 2019. 2

[6] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *FG*, 2018. 7

[7] Chang Chen, Zhiwei Xiong, Xinmei Tian, Zheng-Jun Zha, and Feng Wu. Camera lens super-resolution. In *CVPR*, 2019. 2

[8] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *CVPR*, 2018. 7

[9] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2), 2015. 2

[10] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, and Dacheng Tao. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *CVPR*, 2019. 4

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 1, 2, 3

[12] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *CVPR*, 2019. 1, 2, 5, 6

[13] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *CVPR*, 2018. 2, 5, 6, 8

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017. 7

[15] Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. In *ICCV*, 2017. 7

[16] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 4

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 4

[18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 2

[19] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018. 6

[20] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016. 2

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[22] Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *ICCV workshop*, 2011. 7

[23] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void-learning denoising from single noisy images. In *CVPR*, 2019. 2

[24] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *CVPR*, 2018. 2

[25] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. *arXiv preprint arXiv:1908.03826*, 2019. 5, 6

[26] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*, 2017. 2

[27] Marc Lebrun, Miguel Colom, and Jean-Michel Morel. The noise clinic: a blind image denoising algorithm. *Image Processing On Line*, 5, 2015. 5, 6

[28] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 1, 2, 7

[29] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *ECCV*, 2016. 4

[30] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPR Workshop*, 2017. 2

[31] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 7

[32] Andreas Lugmayr, Martin Danelljan, and Radu Timofte. Unsupervised learning for real-world super-resolution. *arXiv preprint arXiv:1909.09629*, 2019. 1, 2, 3, 8

[33] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *IGARSS*, 2017. 7

[34] Tomer Michaeli and Michal Irani. Nonparametric blind super-resolution. In *ICCV*, 2013. 2

[35] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017. 2, 7

[36] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS Workshop*, 2017. 5

[37] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NIPS*, 2016. 1

[38] Wen-Ze Shao and Michael Elad. Simple, accurate, and robust nonparametric blind super-resolution. In *ICIG*, 2015. 2

[39] Assaf Shocher, Nadav Cohen, and Michal Irani. zero-shot super-resolution using deep internal learning. In *CVPR*, 2018. 1, 2, 5, 6, 8

[40] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *CVPR*, 2018. 5, 6

[41] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *CVPR Workshop*, 2017. 5

[42] Radu Timofte, Shuhang Gu, Jiqing Wu, and Luc Van Gool. Ntire 2018 challenge on single image super-resolution: Methods and results. In *CVPR Workshop*, 2018. 1, 5

[43] Radu Timofte, Rasmus Rothe, and Luc Van Gool. Seven ways to improve example-based single image super resolution. In *CVPR*, 2016. 6

[44] 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 *ECCV Workshop*, 2018. 2, 6, 8

[45] Zhou Wang and Alan C Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine*, 26(1), 2009. 6

[46] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4), 2004. 6

[47] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *CVPR*, 2018. 1, 7

[48] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *CVPR*, 2016. 1, 7

[49] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, 2017. 3

[50] Ke Yu, Chao Dong, Liang Lin, and Chen Change Loy. Crafting a toolchain for image restoration by deep reinforcement learning. In *CVPR*, 2018. 5, 6, 8

[51] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *CVPR Workshop*, 2018. 1, 3

[52] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 8

[53] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. In *ICCV*, 2017. 7

[54] Xuaner Zhang, Qifeng Chen, Ren Ng, and Vladlen Koltun. Zoom to learn, learn to zoom. In *CVPR*, 2019. 2

[55] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 2, 4

[56] Tianyu Zhao, Changqing Zhang, Wenqi Ren, Dongwei Ren, and Qinghua Hu. Unsupervised degradation learning for single image super-resolution. *arXiv preprint arXiv:1812.04240*, 2018. 1, 3

[57] Ruofan Zhou and Sabine Süsstrunk. Kernel modeling super-resolution on real low-resolution images. In *ICCV*, 2019. 1, 2

[58] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2, 3, 7