Unsupervised Real-World Super-Resolution: A Domain Adaptation Perspective

Wei Wang[1*], Haochen Zhang[12*], Zehuan Yuan[1†], Changhu Wang[1]
[1]ByteDance AI Lab, [2]UC San Diego
{wangwei.frank, yuanzehuan, wangchanghu}@bytedance.com haz035@ucsd.edu

Abstract

Most existing convolution neural network (CNN) based super-resolution (SR) methods generate their paired training dataset by artificially synthesizing low-resolution (LR) images from the high-resolution (HR) ones. However, this dataset preparation strategy harms the application of these CNNs in real-world scenarios due to the inherent domain gap between the training and testing data. A popular attempts towards the challenge is unpaired generative adversarial networks, which generate "real" LR counterparts from real HR images using image-to-image translation and then perform super-resolution from "real" LR→SR. Despite great progress, it is still difficult to synthesize perfect "real" LR images for super-resolution. In this paper, we firstly consider the real-world SR problem from the traditional domain adaptation perspective. We propose a novel unpaired SR training framework based on feature distribution alignment, with which we can obtain degradation-indistinguishable feature maps and then map them to HR images. In order to generate better SR images for target LR domain, we introduce several regularization losses to force the aligned feature to locate around the target domain. Our experiments indicate that our SR network obtains the state-of-the-art performance over both blind and unpaired SR methods on diverse datasets.

1. Introduction

Single image super-resolution (SISR), aiming to increase the resolution of an image from a single low-resolution (LR) counterpart, has attracted a lot of attentions in computer vision community. In recent years, convolution neural networks (CNNs) have been applied to SISR task[6] and achieved the state-of-the-art performance[5] over traditional arts. In order to construct training image pairs, some SR studies leverage determined operations such as bicubic interpolation, to down-sample the original high-resolution (HR) images. Obviously, such predefined techniques limit the model generalization capability as several blur and noise of unknown types exist in real LR images. As a result, the notorious domain gap, between the training LR images and the real-world testing images, harms the inference performance of these well-trained CNNs in real-world scenarios.

Blind SR is a straightforward attempt to process real LR images, which aims to restore HR images from LR counterparts with unknown degradation parameters. For example, [1, 10, 20] considered arbitrary blur kernels during the training process. Although these blind models have achieved satisfactory performance for a large range of predefined degradations, the performance still will drop drastically in real-world scenerios, as the distribution of real LR images is different from those degraded by manually designed operations anyway. Recently, inspired by the success of generative adversarial network (GAN) [9] in image style translation [48], many studies use CycleGAN[48] framework to train SR networks in an unpaired manner. They usually suppose that only two unpaired datasets are available: a real LR dataset with no predefined degradations and a real HR one. The main idea is directly generating real LR counterparts from real HR images using image-to-image translation. Then a SR network is trained to map the degraded LR outputs to the corresponding HR images in a paired manner. Although these methods have shown their advantages in real-world SR by directly simulating the real LR image distribution, it is still difficult to train an ideal degraded LR image generator to perfectly mimic real images and real-world SR remains a challenging problem so far.

In contrast to existing SISR efforts generating plausible "real" LR by image-to-image transferring, in this paper we reconsider the unpaired real-world SR from a feature-level domain adaptation perspective. Specifically, the source domain includes the real HR dataset and its synthetic LR counterparts while the real LR dataset is regarded as the target domain inputs without labels. As opposed to high-level vision tasks which try to learn a domain-invariant image representation, our goal is to obtain degradation-indistinguishable feature maps and then map these feature maps to HR images.

[*] Authors contributed equally. This work was performed while Haochen Zhang worked as an intern at ByteDance.

[†] Corresponding author.

Inspired by several adversarial-based domain adaptation approaches in learning domain-invariant features[8, 26, 31], we propose a novel framework for unpaired SR training based on feature alignment, shown as in Figure 1 (a). Such feature alignment technique could hopefully make the features from source and target LR images indistinguishable, so that we can obtain the target HR images during the inference stage by using the decoder network trained with only source HR supervision. However, different from high-level domain adaptation where the aligned image representations are of low resolution, the shared feature space in SR task is extremely large due to relatively fewer down-sampling in CNNs. Therefore as described in Figure 1 (b), we also introduce extra constraints to further help align features and preserve more details compared to the traditional domain adaptation task. The whole domain adaptation based framework for unpaired SR training is illustrated in Figure 2.

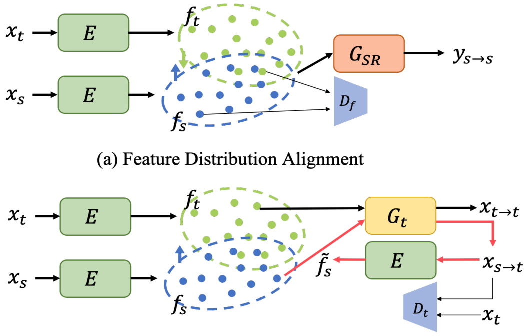
To our best knowledge, this is the first work formulating unpaired SR training as a feature-level domain adaptation problem and our main contributions are summarized as below:

- We propose a novel unpaired SR training framework based on feature distribution alignment.
- We introduce several losses to not only align feature space better but also preserve image details for the downstream SR task.
- Extensive experiments on three challenging datasets show that our proposed method has advantages over the existing unpaired SR training solutions.

2. Related Work

Domain Adaptation. Domain adaptation is a branch of transfer learning where source domain labels are available but target domain labels are not. A popular practice for domain adaptation is to match the feature distributions between domains in order to obtain domain-invariant image representations. Long *et al.* used DAN in [16] to minimize max mean discrepancies (MMD) over the domain-specific layers. Ganin *et al.* [8] used a domain classifier with a gradient reversal layer to encourage the feature extractor to learn domain-invariant features. From then on, methods learning domain-invariant representations in adversarial manner flourished. For example, [26] presented a MADA approach, which captured multi-mode structures to better align different data distributions. Nowadays, domain adaptation approaches have been widely used in high-level tasks, e.g, [4, 24, 34, 36].

Non-blind SISR. Works in non-blind SR field assume the degradation from HR to LR is known, and in most cases it is bicubic interpolation. The pioneer method was SRCNN[6] and most following approaches [5, 12, 13, 29, 30, 44, 45] focus on proposing powerful network blocks to take better use of internal and external information. For example, [45] proposed a residual dense block to improve the expressive ability of the model. There are also works considering visual quality directly[15, 37, 43]. For example, [15] proposed SRGAN which used feature loss[11] and GAN loss to generate visually pleasing results.



(a) Feature Distribution Alignment

(b) Feature Domain Regularization

Figure 1. Schematic diagram of two key components in our framework. (a) guides the encoder E to learn degradation-indistinguishable feature maps using a feature domain discriminator D_f. (b) makes the encoder preserve more information from target LR domain by forcing decoder G_t to restore target degradation with source and target contents, *i.e.* $x_{s \to t}$ and $x_{t \to t}$ respectively. Both parts are further restricted by HR reconstruction losses, shown in Figure 2, in order to exact super-resolution-helpful features.

Blind SISR. This field often assumes the blur kernels used in down-sampling are unavailable. Thus, many methods[1, 10, 20, 23, 47] choose to estimate the unknown kernel first and then perform a standard SISR with estimated kernel prior. For example, [1] learned the blur kernel distribution using a KernelGAN. IKC[10] proposed a correction network, which was trained with estimation network iteratively. More recently, [20] trained several estimation and SR network pairs in an end-to-end manner. Although the estimated kernel improved performance of blind SR, most existing methods could not handle degradation except blurring well. There are also works like ZSSR[27] and its variants[25, 28], which trained SR with self-similarity. However, these image-specific models required extra training stage for each LR image, which were highly cost in application.

Unpaired SISR Methods in this category address SR training under unsupervised setting where no HR-LR pairs are provided. Inspired by CycleGAN[48] and DualGAN[39], Yuan *et al.* [40] combined two CycleGANs and built a CinCGAN to handle unsupervised SR training. [46] designed a bi-cycle degradation network and used bi-cycle consistency to train both high-to-low GAN and low-

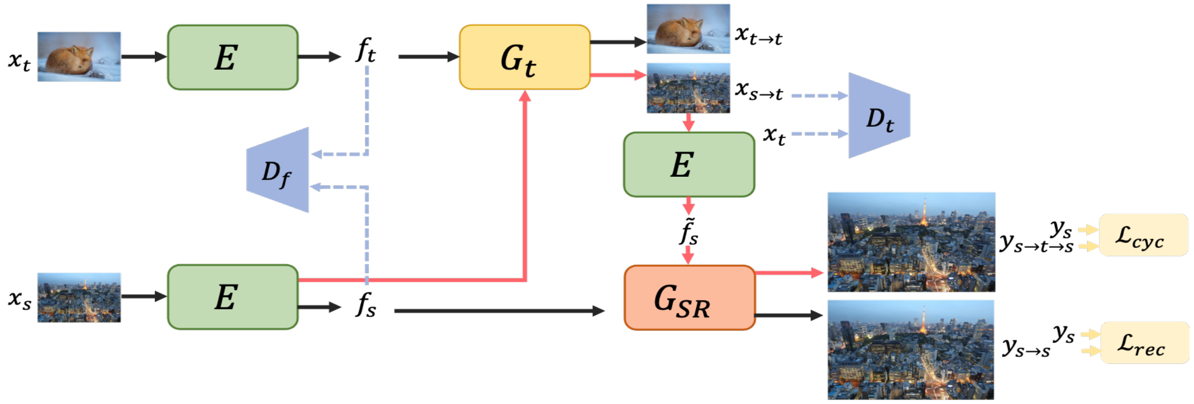


Figure 2. The whole data-flow of our proposed framework. E means three copies of a same encoder and G_t, G_{SR} represent two different decoders. x_t and x_s are input LR images from target and source domain respectively, and f_t, f_s are corresponding feature maps. As the arrows in different colors imply, $x_{t\to t}$ is an image restored from feature f_t which has the same contents and degradation with x_t. $x_{s\to t}$ is generated from feature f_s with contents of x_s but degradation of x_t. Then $x_{s\to t}$ is fed into encoder E to extract feature $\tilde{f}_s$. Finally, the super-resolved images $y_{s\to t\to s}$ and $y_{s\to s}$ is generated from feature maps $\tilde{f}_s$ and f_s respectively. Please refer to Figure 1 for the design principles of each component and Figure 5 for visual inspection

to-high SR networks. [3] and [17] proposed using GANs to simulate degradation process first and then trained SR networks in supervised manner with the generated real degradation pairs. In contrast, [21] used two networks during inference: one produced pseudo-clean LR images from LR images with real degradation and the other was trained as SR network pair-wisely with "pseudo-supervision". All the mention methods translate a fake LR/HR image to a pseudo real one in order to obtain image-level supervision for SR network training, which are quite different from our proposed domain adaptation based framework which applies feature-level domain alignment and regularization.

3. Proposed Method

According to the assumption widely used in unsupervised SR training, we have unpaired real LR and real HR datasets, but no degradation prior. Let $\mathcal{S}$ denote the source domain including a real HR dataset with samples named as y_s. Similarly, let $\mathcal{T}$ denote the target domain consisting of a real LR dataset with samples x_t and no available ground truth. In order to conduct domain adaptation training, we still need source LR images x_s. Fortunately, this requirement is easily satisfied because generating source LR images x_s from source HR images y_s by synthesis is cheap and convenient. Given the conditions above, the proposed domain adaptation based unpaired SR training framework, shown in Figure 2, is introduced in this section.

3.1. Feature Distribution Alignment

Shown as in Figure 1 (a), feature alignment between source and target domain is the first key part of the proposed framework. In this part, we define an encoder-decoder architecture that inputs LR images x and outputs the reconstructed HR images, *i.e.* SR results. The encoder E includes several convolution layers and we denote its parameters as θ_E, *i.e.* $f = E(x;\theta_E)$. Then the feature f is mapped by a decoder G_{SR} to the reconstructed HR images, and we denote the parameters of G_{SR} as $\theta_{G_{SR}}$. Finally, we define a discriminator D_f with the parameters θ_{Df}.

Feature alignment loss. During the training stage, we want to obtain a degradation-indistinguishable feature f. To achieve this, we need the two distributions: $\mathcal{S}(f) = \{E(x;\theta_E)|x \sim \mathcal{S}(x)\}$ and $\mathcal{T}(f) = \{E(x;\theta_E)|x \sim \mathcal{T}(x)\}$ to be similar. A straightforward way is to adopt a GAN structure to reduce the distribution shift of these two feature distributions. In detail, we use two copies of an encoder network E to generate the source feature maps f_s as well as the target feature maps f_t. Then the discriminator D_f is trained to distinguish the domain for each feature map, while encoder E is trained to fool D_f. The optimization of E and D_f is achieved via the adversarial way. Since in SR task, feature maps f_s and f_t are not representation vectors but 3D tensors, we use LSGAN[22] here:

$$\begin{aligned}\min_{\theta_E} \mathcal{L}_{align}(E) = \mathbb{E}_{x_t\sim\mathcal{T}(x)}\left[(D_f(E(x_t)) - 0.5)^2\right] \\ + \mathbb{E}_{x_s\sim\mathcal{S}(x)}\left[(D_f(E(x_s)) - 0.5)^2\right]\end{aligned} \tag{1}$$

$$\begin{aligned}\min_{\theta_{Df}} \mathcal{L}_{align}(D_f) = \mathbb{E}_{x_t\sim\mathcal{T}(x)}\left[(D_f(E(x_t)) - 0)^2\right] \\ + \mathbb{E}_{x_s\sim\mathcal{S}(x)}\left[(D_f(E(x_s)) - 1)^2\right]\end{aligned} \tag{2}$$

SR reconstruction loss. At the same time, we also aim to reconstruct the HR images in the source domain. Therefore we feed the aligned source feature maps f_s into the decoder network G_{SR} and obtain the reconstructed result $y_{s\to s}$ by $y_{s\to s} = G_{SR}(f_s;\theta_{G_{SR}})$. Thus, the encoder E and decoder G_{SR} can be optimized in a supervised way by minimizing the MAE and widely used perception loss shown as Eq.(3). Here, $\mathcal{L}_{fea}$ measures the feature difference based on VGG network[11] and $\mathcal{L}_{adv}$ is adversarial loss calculated with LSGAN[22].

$$\mathcal{L}_{rec}(E,G_{SR}) = \|y_s - y_{s\to s}\|_1 + \alpha\mathcal{L}_{fea}(y_s,y_{s\to s}) + \beta\mathcal{L}_{adv}(y_s,y_{s\to s}) \tag{3}$$

In this way, the encoder closes the gap between the source and target feature domains and the decoder can map the features in the shared feature space to vivid super-resolved images. As a result, during the inference stage, we can obtain the reconstructed HR images in the target domain by feeding the aligned target feature maps f_t into the decoder G_{SR}.

3.2. Feature Domain Regularization

In the previous subsection, we picture a skeleton about using domain adaptation to conduct unpaired SR training. However, it is not enough to just align the feature maps in low-level vision tasks, such as SR. We need to analyze the specialty of SR task further.

Intuitively, in high-level domain adaptation, encoder can just discard some textures not related to classification in order to obtain domain-invariant image representation. In contrast, SR network could not discard textural information but needs to recover high-frequency details. Moreover, the 3D tensor feature space in SR is too large to guarantee good aligning performance. Therefore, we need extra regularizations for the aligned feature space. Considering our final objective is to restore HR images in target domain, we want the encoder E to preserve more information from target LR images. Inspired by the success of CycleGAN[48] in transferring domain on image level, we adopt a similar strategy in feature domain to make the shared feature space closer to target feature domain. To achieve this, we need another decoder G_t with parameters denoted as θ_{G_t}.

Target LR restoration loss. Shown as the black flow in Figure 1 (b), starting from the shared feature space, we feed the target feature maps f_t into the decoder G_t to restore target LR input itself, *i.e.* $x_{t\to t} = G_t(f_t;\theta_{G_t})$. This loss works in two aspects: 1) It forces the encoder E to keep target domain information as much as possible when extracting features. Therefore, it regularizes the shared feature space closer to target domain. 2) It requires the decoder G_t to preserve image contents while generating target domain images. This term is guaranteed by a pixel-wise loss $\mathcal{L}_{res}$.

$$\mathcal{L}_{res}(E,G_t) = \|x_t - x_{t\to t}\|_1 \tag{4}$$

Target degradation style loss. At the mean time, we also force the decoder G_t to generate source LR input with target domain degradation, $x_{s\to t} = G_t(f_s;\theta_{G_t})$. In the other word, we want $x_{s\to t}$ to have the contents of x_s but the degradation of x_t. We achieve this by re-using the encoder E, as shown in Figure 1 (b) the red data-flow. Firstly, as most of CycleGAN based methods do, we adopt another discriminator D_t with parameters θ_{D_t} and an adversarial loss to guarantee x_t and $x_{s\to t}$ have the similar distribution:

$$\min_{\theta_E,\theta_{G_t}} \mathcal{L}_{sty}(E,G_t) = \mathbb{E}_{f_s\sim\mathcal{S}(f)}\left[(D_t(G_t(f_s)) - 1)^2\right] \tag{5}$$

$$\begin{aligned}\min_{\theta_{D_t}} \mathcal{L}_{sty}(D_t) = \mathbb{E}_{f_s\sim\mathcal{S}(f)}&\left[(D_t(G_t(f_s)) - 0)^2\right] \\ &+ \mathbb{E}_{x_t\sim\mathcal{T}(x)}\left[(D_t(x_t) - 1)^2\right]\end{aligned} \tag{6}$$

Feature identity loss. Then we need an identity loss to keep the content of $x_{s\to t}$ unchanged. Different from most CycleGAN based methods do, we here apply identity loss on feature level. Specifically, we feed $x_{s\to t}$ into the encoder E and obtain its feature maps $\tilde{f_s} = E(x_{s\to t};\theta_E)$. Noting that x_s has source contents and source degradation, while $x_{s\to t}$ has source contents and target degradation. Both of them are mapped by the same encoder E into a shared feature space which is degradation-indistinguishable. Therefore ideally, the feature maps f_s and $\tilde{f_s}$ should be same. The identity loss should be in form of pixel-wise:

$$\mathcal{L}_{idt}(E,G_t) = \left\|f_s - \tilde{f_s}\right\|_1 \tag{7}$$

Cycle loss. Finally, as shown in Figure 2, we re-use the decoder G_{SR} to further guarantee the shared feature space contains useful information for reconstructing HR images. It is similar to the cycle loss in CycleGAN framework while it also could be viewed as a feature-level data augmentation in our framework. In detail, we feed $\tilde{f_s}$ into the decoder G_{SR} to obtain another super-resolved result $y_{s\to t\to s}$. The loss functions applied on $y_{s\to t\to s}$ is exactly same with those applied on $y_{s\to s}$, *i.e.* both VGG networks and discriminator are shared.

$$\mathcal{L}_{cyc}(E,G_t,G_{SR}) = \|y_s - y_{s\to t\to s}\|_1 + \alpha\mathcal{L}_{fea}(y_s,y_{s\to t\to s}) + \beta\mathcal{L}_{adv}(y_s,y_{s\to t\to s}) \tag{8}$$

Full objective. With the aforementioned regularization losses, we make the encoder E more concentrated on extracting and expressing the similarities of patches from target domain with useful information for HR reconstruction. As a result, the encoder E and decoder G_t, G_{SR} are trained with the following objective function, Eq.(9), in an end-to-end manner. The discriminator D_f and D_t as well as the discriminators in Eq.(3) and Eq.(8) are trained with their corresponding loss in an alternate way.

$$\begin{aligned}\mathcal{L}_{train} = {}&\lambda_{align}\mathcal{L}_{align}(E) + \lambda_{rec}\mathcal{L}_{rec}(E,G_{SR}) \\ &+ \lambda_{res}\mathcal{L}_{res}(E,G_t) + \lambda_{sty}\mathcal{L}_{sty}(E,G_t) \\ &+ \lambda_{idt}\mathcal{L}_{idt}(E,G_t) + \lambda_{cyc}\mathcal{L}_{cyc}(E,G_t,G_{SR})\end{aligned} \tag{9}$$

AIM 2019				NTIRE 2020			
Method	LPIPS↓	PSNR↑	SSIM↑	Method	LPIPS↓	PSNR↑	SSIM↑
†Bicubic	0.673	22.36	0.614	†Bicubic	0.632	25.52	0.671
†MadDeamon(Winner)	0.403	21.00	0.504	†Impressionism(Winner)	0.227	24.83	0.672
ZSSR[27]	0.639	22.21	0.603	ZSSR[27]	0.620	24.93	0.642
KernelGAN[1]+ZSSR[27]	0.613	22.40	0.611	KernelGAN[1]+ZSSR[27]	0.598	25.34	0.661
DnCNN[41]+K.[1]+Z.[27]	0.607	22.40	0.614	DnCNN[41]+K.[1]+Z.[27]	0.438	25.84	0.722
DnCNN[41]+IKC[10]	0.614	22.26	0.596	DnCNN[41]+IKC[10]	0.384	26.50	0.748
*Maeda *et al.* [21]	0.454	22.88	0.661	SRResCGAN[35]	0.335	25.05	0.676
DASR[38]	0.346	21.79	0.577				
Ours	0.340	22.60	0.622	Ours	0.252	25.40	0.707

Table 1. Quantitative comparison with state-of-the-art blind/unsupervised methods on unpaired dataset. † means the results are taken from the official website[1]. * means the results are taken from Maeda *et al.* [21]. Please note LPIPS is the most important metric here while PSNR and SSIM are provided for reference. Although improving visual quality, the combinations of blind restoration methods are clearly inferior to unsupervised methods in real-world SR setting. Then among all the unpaired SR methods, our proposed one wins the top performance on both datasets.

where $\lambda_{rec}, \lambda_{align}, \lambda_{res}, \lambda_{sty}, \lambda_{idt}, \lambda_{cyc}$ are loss weights, representing the contributions of each objective.

4. Experiments

4.1. Training Details

Different from previous unsupervised methods (such as CinCGAN[40], DASR[38]) that require two training stages, our framework is optimized through a single training step in an end-to-end mode. For both generators and discriminators, we use Adam optimizer[14] with $\beta_1 = 0.9, \beta_2 = 0.99$, and an initial learning of 1×10^{-4}. The learning rates is halved at 250k, 350k, 450k, and 550k iterations. In each iteration, we train the whole framework with a mini-batch size of 8 and the patch size of the LR image is 128×128. Then, data augmentation of random flip and rotation is performed during training. For simplicity, we divide the hyper-parameters in Eq.(9) into two groups according to their functions. One group $(\lambda_{rec}, \lambda_{res}, \lambda_{cyc})$ controls image reconstruction and the other $(\lambda_{align}, \lambda_{sty}, \lambda_{idt})$ controls domain alignment. We simply set the weights in the same group to be the same and then adjust the ratio between two groups. The coefficients in Eq.(3) and Eq.(8) is fixed as $\alpha = 0.01, \beta = 0.01$ according to previous works[7]. More details can be found in the released codes.

4.2. Experiments on Unpaired Dataset

Dataset We mainly experiment on two unpaired SR datasets. Both are official datasets provided in AIM 2019[19] (Track 2) and NTIRE 2020[18] (Track 1) Real-World Super-Resolution Challenge respectively. Since the generation and partition of the two datasets are similar, here only gives a brief introduction for NTIRE2020 dataset as an example. More details please refer to their paper, [19] and [18], respectively. As said in [18], Lugmayr *et al.* design a degradation operator generating structured artifacts which were commonly produced by image processing deployed on very low-end devices. Since this type of degradation is undisclosed and very different from what has previously been used, the degraded images could be regarded as real LR at least experimentally. Typically in this dataset, training images are divided into two subsets. One includes 2,650 Flickr2K[37] images with the aforementioned degradation but no down-sampling. The other has 800 clean HR images from DIV2K dataset[32]. For validation and test, DIV2K images within corresponding splits are first down-sampled and then degraded.

Settings To generate source LR images, we add down-sampling, gaussian noise and gaussian blur to clean HR images. We train our networks using the provided unpaired training datas and the synthesized source LR images. we evaluate SR performances on the validation set which contains 100 images. Moreover, the loss weights are set as: $\lambda_{rec} = 1, \lambda_{res} = 1, \lambda_{cyc} = 1, \lambda_{align} = 0.01, \lambda_{sty} = 0.01, \lambda_{idt} = 0.01$.

Comparison with state-of-the-art blind methods. Since there is not a popular benchmark for unsupervised SR, we make a simple one by collecting public results and running state-of-the-art blind/unsupervised SR testing codes provided by the authors on AIM 2019 and NTIRE 2020 competition datasets. It also includes some combinations of state-of-the-art restoration and blind SR methods. The results are reported in Table 1. Since in real-world SR setting perception is the main objective and contradictory to distortion[2], LPIPS[42] is the most important metric in Table 1 while PSNR and SSIM are provided for reference. Firstly, among methods consisting of cascade restorations, 'DnCNN[41] + KernelGAN[1] + ZSSR[27]'

[1] https://competitions.codalab.org/competitions/22220#learn_the_details

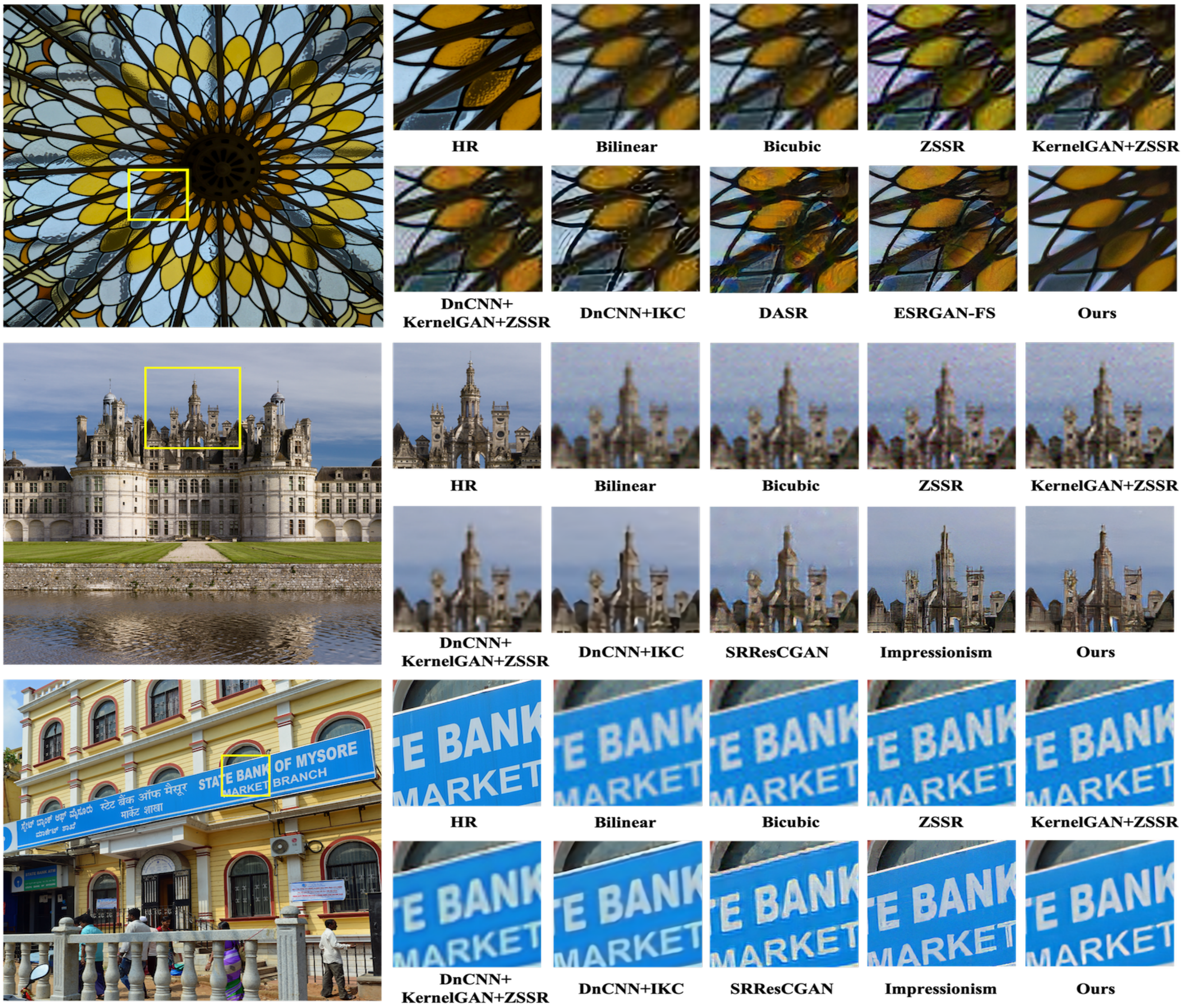


Figure 3. Visual inspection for state-of-the-art methods. The first row is image “0821” from AIM 2019 validation set and the others are image “0830” and “0890” respectively from NTIRE 2020 (Track 1) validation set. As can be seen, among the results without artifacts, ours is the sharpest, and those baselines having sharp results as ours generate obvious artifacts.

obtains the best LPIPS performance on AIM 2019 and the second place on NTIRE 2020 dataset while ‘DnCNN[41] + IKC[10]’ is the best on NTIRE 2020. Although improving visual quality, these combinations are clearly inferior to unsupervised methods in real-world SR setting. Among all the unsupervised methods on AIM 2019 dataset, ours wins the best LPIPS performance and the most pleasant visual quality, shown in Figure 3. On NITRE 2020 dataset, our methods outperform all the published baselines. Compared to the winner method, ours achieves similar performance in visual quality, but better PSNR and SSIM. Figure 3 show some examples for visual inspection.

Visual inspection on NTIRE 2020 Track2 dataset. Figure 4 also gives some visual comparisons on NTIRE 2020 Track 2 dataset whose settings are exactly real-world scenarios. The training and testing images for this track are both captured by smartphone, containing artifacts generated from image enhancement operations deployed on smartphone. The goal is to reconstruct clean HR images with the reference of unpaired high quality images. The visual quality is the only measurement. As seen from the figure, our method could generate high-frequency details with the least artifacts. This result shows that our model can suppress the artifacts better than existing unpaired SR methods.

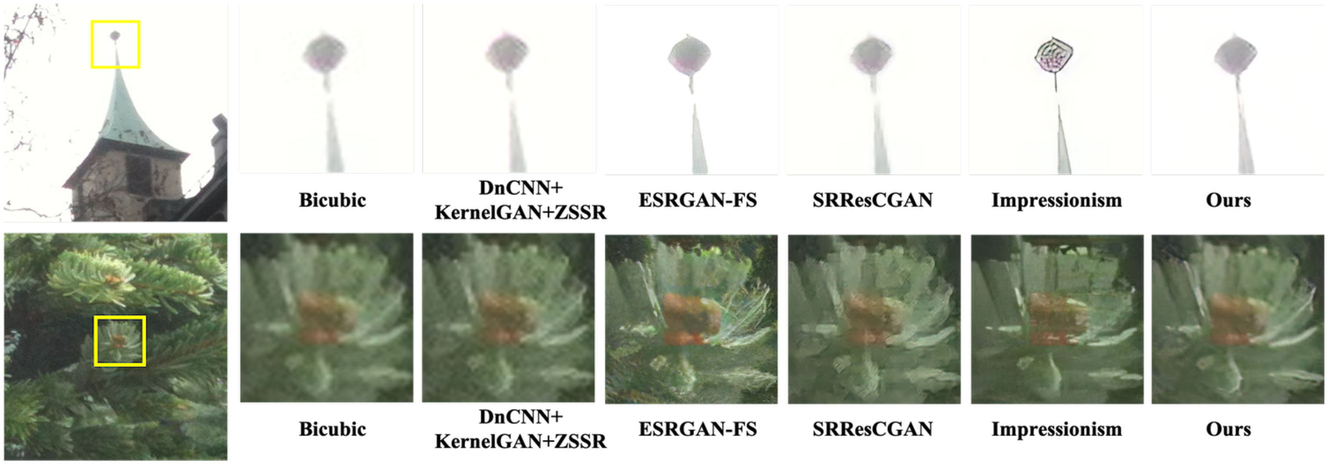


Figure 4. Visual inspection for NTIRE20 (Track 2) dataset. Patches of image "00022" and "00040" are shown here. As we can see, our SR network could generate HR details at the mean time reduce artifacts.

Username	PSNR↑	SSIM↑
ZSSR[27]	22.17	0.472
KernelGAN[1]+ZSSR[27]	22.27	0.475
DnCNN[41]+K.[1]+Z.[27]	22.52	0.489
DnCNN[41]+IKC[10]	21.56	0.433
*Maeda *et al*. [21]	21.32	0.554
Ours	23.39	0.537

Table 2. Quantitative comparison with state-of-the-art blind methods on paired dataset. * means the results are taken from Maeda *et al*. [21]. As can be seen from this table, our proposed method outperforms all the blind restoration combinations as well as the state-of-the-art unpaired SR method in term of PSNR.

4.3. Additional Experiments on Paired Dataset

Dataset. We conduct some additional experiments on the synthetic paired dataset, DIV2K realistic-wild dataset. This set is the Track 4 dataset of the NTIRE 2018 Super-Resolution Challenge[33]. Specifically, it simulates 'real-wild' LR image via 4 times downs-sampling, motion blurring, pixel shifting and additive noise. The degradation operations are image-specific which means the degradation is same within a single image, but different from one to one. Timofte *et al*. totally generate 3,200 LR and 800 HR training samples by degrading each DIV2K training image four times. Following [21], we train our model with "unpaired/unaligned" sampling. Again, we evaluate SR performances on the realistic-wild validation set since the ground truths of the testing images are unavailable. Because this competition evaluates all methods from the perspective of PSNR/SSIM, we do not use visual quality loss of Eq.(3) and Eq.(8). Here, we use hyperparameters $\lambda_{rec} = 10, \lambda_{res} = 10, \lambda_{cyc} = 10, \lambda_{align} = 1, \lambda_{sty} = 1, \lambda_{idt} = 1$.

Comparison with state-of-the-art blind methods. Since there is not a widely used benchmark for blind SR on multiple degradations problem, we combine different blind SR methods with blind restoration as baselines like [21]. We report Maeda *et al*.'s results in Table 2 as well. As seen in the table, our SR network improves the PSNR performance significantly and achieves second best SSIM with small gap between the first one.

4.4. Ablation Study

All the ablation studies conducted in this section are on dataset of NTIRE 2020[18] (Track 1).

Intermediate products. Firstly, we visualize some intermediate product examples to provide an intuitive verification that our pipeline works as expectation. As shown in Figure 5, the source domain LR image x_s has different degradation from the target domain LR image x_t. Then our decoder G_t decodes features in the shared space to $x_{s\to t}$ and $x_{t\to t}$ which consist of different contents but same degradation as target domain. This result shows our proposed losses indeed force the encoder E to preserve more information from target LR images.

Objective functions. Then we conduct the ablation experiments to investigate the contributions of each proposed component. Therefore, we design some variants of our proposed network. (1) Base model. This model only includes an encoder E and decoder G_{SR}. In the training stage, we train this model in the paired manner using source domain data and the loss is just $\mathcal{L}_{rec}$. Then we perform inference on test LR images which have same degradation as target domain. (2) Feature align model. we add discriminator D_f and feature alignment loss $\mathcal{L}_{align}$ to the Base model. This variant represents the vanilla domain adaptation framework which is exactly Figure 1 (a). (3) Full model without

Method	LPIPS↓	PSNR↑	SSIM↑
(1) Base model	0.436	25.65	0.666
(2) only $\mathcal{L}_{align}$	0.448	25.20	0.653
(3) Full model w/o $\mathcal{L}_{align}$	0.373	25.53	0.663
(4) Full model	0.296	25.14	0.690

Table 3. Ablation study for objective functions on NTIRE 2020 dataset. The improvement between the variant (4) and (3) demonstrates the effectiveness of our feature distribution alignment. Comparing variant (2) and (3) shows the importance of our feature domain regularization.

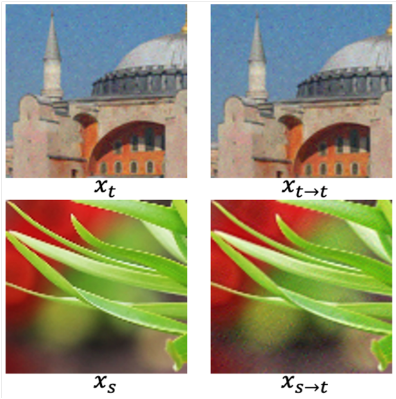


Figure 5. Intermediate images of proposed method. x_s is the bicubic down-sampled result of image "0803" from the DIV2K validation ground-truth set, and x_t is image "0890" from the NTIRE 2020 competition validation set.

$\mathcal{L}_{align}$. This model can be seen as a CycleGAN like framework with extra two identity constrains: target LR restoration loss $\mathcal{L}_{res}$ and feature identity loss $\mathcal{L}_{idt}$. (4) Full model. All networks and losses proposed in Section 3 are used in this model. The performance of each variant is shown in Table 3. Firstly, the Base model does not performance well because of the domain gap between training and testing LR data. Secondly, the variant (2) demonstrates it is difficult to do feature alignment in a shared feature space as large as 3D tensors. Thirdly, comparing the variant (2) and the variant (4) shows our feature domain regularization could shrink the shared space to the target domain. Finally, the improvement between the variant (3) and (4) highlights the effectiveness of our feature distribution alignment.

Source LR synthesis. Would the synthesis of source LR images affect the performance? Here are experiments giving the answer. In this part, we synthesize LR images in two way: (1) One is simple bicubic down-sampling; (2) The other is complex degradation as described in Section 4.2. We train aforementioned 'Base model' (denoted as base model) and 'Full model' (denoted as ours) in Figure 6 on these two kinds of source LR images. Specifically, we use hyperparameters $\lambda_{rec} = 10, \lambda_{res} = 10, \lambda_{cyc} = 10, \lambda_{align} = 1, \lambda_{sty} = 1, \lambda_{idt} = 1$ for simple bicubic data since the domain gap in two settings are quit different. We report their performances in Figure 6. Firstly, we can find the 'Base model' trained on the complex degradation pairs performs better than the one trained on simple bicubic down-sampled pairs. This result means the second source LR domain is closer to the target one. Secondly, compared the 'Base model' and 'Full model' on both degradation, we can see our proposed framework improves the SR performance by aligning feature maps between domains, no matter how source LR images are synthesized. Thirdly, please pay attention to the improvement on two kinds of source LR images. The proposed method gains larger improvement in the case where the gap between source and target domain are larger. That's to say, our methods would be more useful in real-world applications where it is impossible to simulate LR degradation manually. All these results confirm the effectiveness of our proposed method in learning degradation-indistinguishable but super-resolution-helpful features.

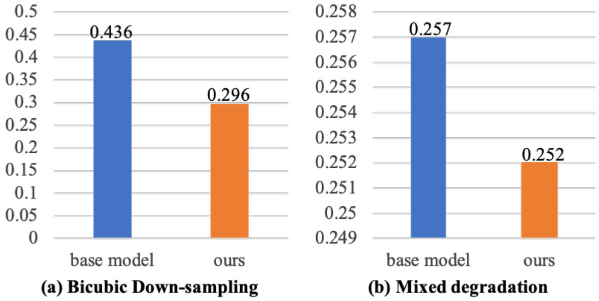


Figure 6. LPIPS performance of the proposed method and base model. (a) source LR images are generated by bicubic down-sampling. (b) extra random blur and noise are added to the source LR images.

5. Conclusion

In this paper, we formulate unpaired SR training as a domain adaptation problem. By using our proposed feature distribution alignment loss and feature domain regularization losses, the encoder of our SR network could map LR images from different domains into a shared degradation-indistinguishable feature space which is relatively closer to target feature domain. Then trained with two SR reconstruction losses, the decoder of our SR network could reconstruct vivid HR images from features in that shared feature space. Extensive experiments on diverse datasets demonstrate the effectiveness of our proposed framework. Although free from artifacts, our method tends to produce smoother results. We will address this problem in our future work.

References

[1] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. pages 284–293, 2019. 1, 2, 5, 7

[2] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *CVPR*, pages 6228–6237, 2018. 5

[3] 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*, pages 185–200, 2018. 3

[4] Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. In *ICCV*, pages 2090–2099, 2019. 2

[5] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *CVPR*, pages 11065–11074, 2019. 1, 2

[6] 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):295–307, 2015. 1, 2

[7] Manuel Fritsche, Shuhang Gu, and Radu Timofte. Frequency separation for real-world super-resolution. In *ICCV Workshops*, pages 3599–3608, 2019. 5

[8] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189, 2015. 2

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1

[10] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *CVPR*, pages 1604–1613, 2019. 1, 2, 5, 6, 7

[11] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711, 2016. 2, 4

[12] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, pages 1646–1654, 2016. 2

[13] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *CVPR*, pages 1637–1645, 2016. 2

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[15] 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*, pages 4681–4690, 2017. 2

[16] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, pages 97–105, 2015. 2

[17] Andreas Lugmayr, Martin Danelljan, and Radu Timofte. Unsupervised learning for real-world super-resolution. In *ICCV Workshops)*, pages 3408–3416, 2019. 3

[18] Andreas Lugmayr, Martin Danelljan, and Radu Timofte. Ntire 2020 challenge on real-world image super-resolution: Methods and results. In *CVPR Workshops*, pages 494–495, 2020. 5, 7

[19] Andreas Lugmayr, Martin Danelljan, Radu Timofte, Manuel Fritsche, Shuhang Gu, Kuldeep Purohit, Praveen Kandula, Maitreya Suin, AN Rajagoapalan, Nam Hyung Joon, et al. AIM 2019 challenge on real-world image super-resolution: Methods and results. In *ICCV Workshops*, pages 3575–3583, 2019. 5

[20] Zhengxiong Luo, Yan Huang, Shang Li, Liang Wang, and Tieniu Tan. Unfolding the alternating optimization for blind super resolution. pages 1–12, 2020. 1, 2

[21] Shunta Maeda. Unpaired image super-resolution using pseudo-supervision. In *CVPR*, pages 291–300, 2020. 3, 5, 7

[22] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, pages 2794–2802, 2017. 3, 4

[23] Tomer Michaeli and Michal Irani. Nonparametric blind super-resolution. In *ICCV*, pages 945–952, 2013. 2

[24] Kwanyong Park, Sanghyun Woo, Dahun Kim, Donghyeon Cho, and In So Kweon. Preserving semantic and temporal consistency for unpaired video-to-video translation. In *ICMM*, pages 1248–1257, 2019. 2

[25] Seobin Park, Jinsu Yoo, Donghyeon Cho, Jiwon Kim, and Tae Hyun Kim. Fast adaptation to super-resolution networks via meta-learning. pages 754–769, 2020. 2

[26] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Multi-adversarial domain adaptation. In *AAAI*, volume 32, 2018. 2

[27] Assaf Shocher, Nadav Cohen, and Michal Irani. “zero-shot” super-resolution using deep internal learning. In *CVPR*, pages 3118–3126, 2018. 2, 5, 7

[28] Jae Woong Soh, Sunwoo Cho, and Nam Ik Cho. Meta-transfer learning for zero-shot super-resolution. In *CVPR*, pages 3516–3525, 2020. 2

[29] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *CVPR*, pages 3147–3155, 2017. 2

[30] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *ICCV*, pages 4539–4547, 2017. 2

[31] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016. 2

[32] 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 Workshops*, pages 114–125, 2017. 5

[33] Radu Timofte, Shuhang Gu, Jiqing Wu, and Luc Van Gool. NTIRE 2018 challenge on single image super-resolution: Methods and results. In *CVPR Workshops*, pages 852–863, 2018. 7

[34] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, pages 7472–7481, 2018. 2

[35] Rao Muhammad Umer, Gian Luca Foresti, and Christian Micheloni. Deep generative adversarial residual convolutional networks for real-world super-resolution. In *CVPR Workshops*, pages 438–439, 2020. 5

[36] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, pages 2517–2526, 2019. 2

[37] 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 Workshops*, pages 1–16, 2018. 2, 5

[38] Yunxuan Wei, Shuhang Gu, Yawei Li, and Longcun Jin. Unsupervised real-world image super resolution via domain-distance aware training. *arXiv preprint arXiv:2004.01178*, 2020. 5

[39] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. DualGAN: Unsupervised dual learning for image-to-image translation. In *ICCV*, pages 2849–2857, 2017. 2

[40] 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 Workshops*, pages 701–710, 2018. 2, 5

[41] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017. 5, 6, 7

[42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 5

[43] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *ICCV*, pages 3096–3105, 2019. 2

[44] 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*, pages 286–301, 2018. 2

[45] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, pages 2472–2481, 2018. 2

[46] Tianyu Zhao, Wenqi Ren, Changqing Zhang, Dongwei Ren, and Qinghua Hu. Unsupervised degradation learning for single image super-resolution. *arXiv preprint arXiv:1812.04240*, 2018. 2

[47] Ruofan Zhou and Sabine Susstrunk. Kernel modeling super-resolution on real low-resolution images. In *ICCV*, pages 2433–2443, 2019. 2

[48] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017. 1, 2, 4