Unsupervised Real-world Image Super Resolution via Domain-distance Aware Training

Yunxuan Wei[1,*], Shuhang Gu[2,3,*], Yawei Li[3], Radu Timofte[3], Longcun Jin[1], Hengjie Song[1,†]

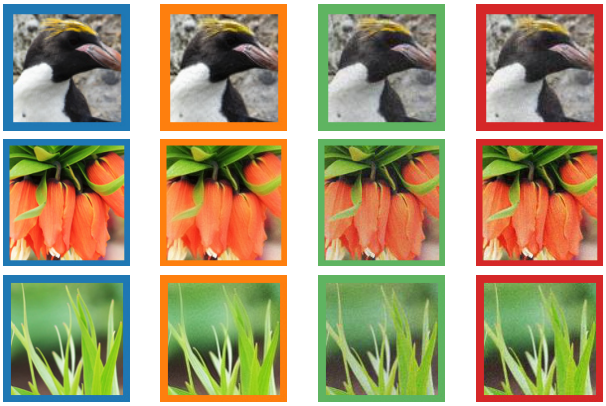
[1] South China University of Technology, [2] The University of Sydney, [3] ETH Zurich
{yunxuanwei, sehjsong}@mail.scut.edu.cn, shuhanggu@gmail.com,
{yaweili, radu.timofte}@vision.ee.ethz.ch

Abstract

These days, unsupervised super-resolution (SR) is soaring due to its practical and promising potential in real scenarios. The philosophy of off-the-shelf approaches lies in the augmentation of unpaired data, i.e. *first generating synthetic low-resolution (LR) images $\mathcal{Y}^g$ corresponding to real-world high-resolution (HR) images $\mathcal{X}^r$ in the real-world LR domain $\mathcal{Y}^r$, and then utilizing the pseudo pairs $\{\mathcal{Y}^g, \mathcal{X}^r\}$ for training in a supervised manner. Unfortunately, since image translation itself is an extremely challenging task, the SR performance of these approaches is severely limited by the domain gap between generated synthetic LR images and real LR images. In this paper, we propose a novel domain-distance aware super-resolution (DASR) approach for unsupervised real-world image SR. The domain gap between training data (e.g. $\mathcal{Y}^g$) and testing data (e.g. $\mathcal{Y}^r$) is addressed with our* ***domain-gap aware training*** *and* ***domain-distance weighted supervision*** *strategies. Domain-gap aware training takes additional benefit from real data in the target domain while domain-distance weighted supervision brings forward the more rational use of labeled source domain data. The proposed method is validated on synthetic and real datasets and the experimental results show that DASR consistently outperforms state-of-the-art unsupervised SR approaches in generating SR outputs with more realistic and natural textures. Codes are available at* https://github.com/ShuhangGu/DASR.

1. Introduction

Single image super-resolution (SR) aims at reconstructing a high-resolution (HR) image from a low-resolution (LR) observation. In the past two decades, SR has been a thriving research topic due to its highly practical value in enhancing image details and textures. A wide variety of models [15, 19, 55, 49, 22] have been suggested to deal with the image SR problem.



Bicubic FSSR [16] DASR(ours) Real-world

Figure 1: Visualization of the domain-gap between the generated LR images by different methods and LR images in the AIM [40] dataset. More details can be found in our main text.

Benefiting from the rapid development of deep convolutional neural networks (CNNs), recent years have witnessed an explosive spread of training CNN models [27, 38, 48, 52, 60, 63, 57, 36] for SR. State-of-the-art SR performance has been boosted by directly training networks to capture the LR-to-HR mapping. Moreover, when combined with adversarial training [20] or perceptual losses [31], SR networks can produce accurate and natural-looking image details.

In spite of their success on benchmark datasets, the poor generalization capacity of discriminatively trained SR networks limits their application in real scenarios. When applied to super-resolve real images, SR networks trained on simulated datasets usually lead to undesired strong artifacts in their SR results. For the pursuit of real image SR, great attempts have been made in the last couple of years. By adjusting the focal length of a digital camera, several works prepared real image SR datasets [9, 61, 8]. But the collections of these datasets are often laborious and costly. Furthermore, SR networks trained on the collected datasets are hard to generalize to images captured in other condi-

*The first two authors contribute equally to this work.
†Corresponding author.

tions. Another category of approaches investigates real-world image SR from an algorithmic perspective. Some works [41, 64, 5, 30] assume the LR and HR images satisfy a parameterized degradation model and propose blind SR algorithms which are able to adapt to the unknown down-sampling kernel in the testing phase. These blind SR algorithms [11, 56, 58, 4] have shown improved generalization capacity over models trained on predetermined synthetic data, but the fixed degradation assumption greatly limits their performances on real data, which are often subject to complex sensor noise and compression artifacts.

Recently, without any assumptions on the degradation model, unsupervised SR approaches have been proposed to leverage unpaired training data. Given a set of real-world LR images $\mathcal{Y}^r = \{y_i^r\}_{i=1,...,N}$, some works [7, 39, 16, 24] proposed to train a degradation network to generate LR observations y_i^g of the available HR images $x_i^r \in \mathcal{X}^r$, and enforcing the same distribution of the generated LR images $\mathcal{Y}^g = \{y_i^g\}_{i=1,...,M}$ with that of real LR images $\mathcal{Y}^r$. With the generated pseudo pairs $\{\mathcal{Y}^g, \mathcal{X}^r\}$, supervised training can be employed to train the SR network. Such unsupervised settings exploit the real training data to learn the complex degradation model and lead to promising SR results on real-world images [40]. However, existing unsupervised SR approaches [16, 7, 39] ignore the domain-gap between $\mathcal{Y}^g$ and $\mathcal{Y}^r$ in the training process of SR networks. In Fig. 1, we visualize the domain gap between the generated and the real LR images. We employed Bicubic downsampling, the trained down-sampling networks by FSSR [16] and our proposed DASR to generate LR images from HR images. Although the trained down-sampling networks are able to generate better LR images that reside in a domain closer to the real LR domain than the bicubically downsampled images, the domain gap still exists between $\mathcal{Y}^g$ and $\mathcal{Y}^r$. The Fréchet Inception Distance (FID) [26] between bicubically downsampled images, FSSR generated LR images, DASR generated LR images and AIM LR images are 37.69, 33.89 and 31.28, respectively. As the four groups of LR images share the same image contents, the FID scores clearly reflect the domain-distance between generated LR images and real-world LR images in the AIM [40] dataset.

In this paper, we propose a Domain-distance Aware Super-resolution (DASR) framework for real-world image super-resolution. Different from previous unsupervised methods [7, 39, 16, 23] which rely on the generation of pseudo pairs for supervised training, our DASR takes into consideration the domain gap between the generated and real LR images, *i.e.* $\mathcal{Y}^g$ and $\mathcal{Y}^r$, and solves the SR problem with both of them under a domain adaptation setting. Our DASR method addresses the domain gap issue through two training strategies: ***domain-gap aware training*** and ***domain-distance weighted supervision***. **Firstly**, with the domain-gap aware training, DASR employs both the generated pseudo pairs $\{\mathcal{Y}^g, \mathcal{X}^r\}$ and real LR images $\mathcal{Y}^r$ to train the SR network. Besides the supervised loss on the pseudo pairs $\{\mathcal{Y}^g, \mathcal{X}^r\}$, DASR also imposes adversarial constraints on the HR estimation $\hat{\mathcal{X}}^{r\to r}$ of real-world data $\mathcal{Y}^r$. Incorporating $\mathcal{Y}^r$ into training informs the network of the target domain, greatly improves its SR performance on real-world data. **Secondly**, besides the domain-gap aware training, a domain-distance weighted supervision strategy is also proposed for advanced exploitation of the generated pseudo pairs. As shown in Fig. 1, some generated LR samples reside closer to the real-world domain while the others are relatively far away from it. We therefore adjust the importance of each pair $\{y_i^g, x_i^r\}$ according to the domain distance between y_i^g and $\mathcal{Y}^r$. Samples that are relatively closer to the real-world domain are assigned with larger weights in the training phase; while unrealistic samples are only allowed to make a limited contribution to the training.

In addition to the above strategies, which are the major contributions of this paper, we also improve previous methods by employing better architecture of the down-sampling network and better adversarial loss in the wavelet domain.

Our contributions can be summarized as follows:

- A domain distance aware super-resolution (DASR) framework is proposed to solve the real-world image SR problem. DASR addresses the domain gap between generated LR images and real images with the proposed domain-gap aware training and domain-distance weighted supervision strategies.
- We provide detailed ablation studies to analyze and validate our contributions. Experimental results on synthetic and real datasets clearly demonstrate the superiority of DASR over the competing approaches.

2. Related Works

2.1. Single Image Super-Resolution with CNNs

Nowadays CNN-based methods are the mainstream in the single image SR field. In the pioneering work of Dong *et al.* [12], the first CNN-based SR method to reach competitive results was introduced. They proposed SRCNN, a 3 layers CNN, to directly learn the mapping function between LR and HR image pairs. After that, a surge of network architectures, such as a deep network with residual learning [27], network with residual blocks [38], densely connected network [52], have been designed to solve the SR task. The SR performance on benchmark datasets have been continuously improved by newly proposed network architectures [2, 13, 25, 29, 32, 34, 63, 3, 28, 51]. Besides investigating more powerful network architecture, perceptual-driven approaches explore better loss functions to improve the perceptual quality of SR results. Johnson *et al.* [31] proposed a perceptual loss which measures the error of two images in the feature space instead of pixel space. Ledig *et*

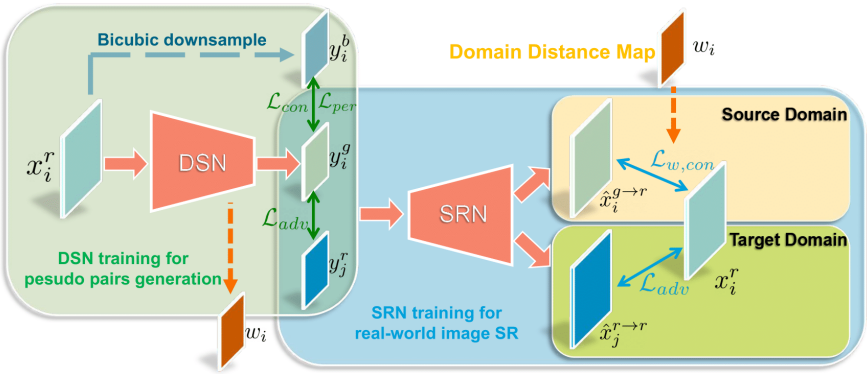


Figure 2: Illustration of our DASR framework. DASR firstly trains a down-sampling network (DSN) which aims to generate LR images y^g in the real LR domain y^r from HR images x^r: $y^g = DSN(x^r)$. Then, DASR take the domain-gap between generated LR images y^g and real-world LR images y^r into consideration, utilize the generated LR-HR pairs $\{y^g, x^r\}$, domain distance maps w and real-world LR images y^r to train super-resolution network (SRN).

al. [35] firstly introduced the adversarial loss to favor outputs residing on the manifold of natural images. Inspired by these pioneer works, different training criteria [52, 60, 44] have been suggested to promote the visual quality of SR results.

Although significant advances have been made, all the aforementioned approaches are trained and evaluated on simulated datasets which assume simple and uniform degradation. These days, the real-world image super-resolution problem has attracted increasing attention due to its high practical values. A branch of work [30, 41, 64, 5, 11] assumes the degradation model between LR and HR images can be characterized by an unknown blur kernel and the subsequent downsampling operation. These blind SR works explicitly estimate the unknown blur kernel at the testing time and take the estimated kernel as an input variable for kernel adaptive SR networks [58, 64] to adapt to different degradation hyper-parameters. There are also works [4, 37] attempting to use the test image for training or fine-tuning the SR network in the testing phase. However, both approaches still rely on a known degradation model during training. To deal with more general real-world SR task, some recent works consider an unsupervised setting which does not rely on the degradation assumption. Given a group of LR images, Yuan *et al*. [56] firstly learned a mapping to transfer the original input images to the clean image domain and applied SR in the clean image domain. Other unsupervised approaches [16, 7, 39, 24] proposed to learn a downsampling process to generate paired data and train SR network with the generated data in a supervised manner. The advantage of these unsupervised SR methods is that they do not rely on the degradation assumption, and therefore are capable of generalizing to very challenging real-world images. However, as image translation itself is an extremely challenging task, the generated LR images are often not consistent with the real LR images. Such a domain gap between training and testing data will deteriorate the final SR performance in the testing phase.

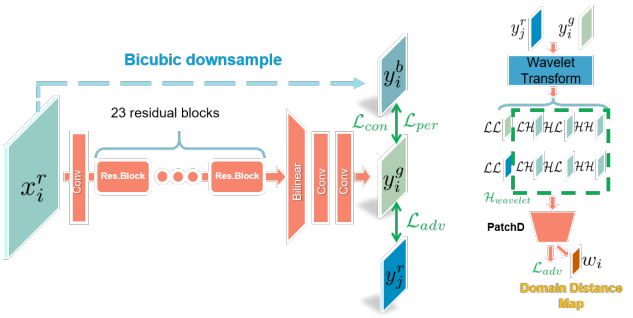
2.2. Domain Adaptation

Domain adaptation aims to utilize a labeled source domain to learn a model that performs well on an unlabeled target domain. It is a classical machine learning problem [17, 14, 46]. Recently, with the explosive spread of using CNN models to solve computer vision tasks, domain adaption has received increasing attention. It has been deployed in many tasks for levering synthetic data or data from other datasets. Early domain adaptation works in the computer vision field focus on solving the domain bias issue in high-level classification tasks [42, 10, 18, 21, 43]. Recently, domain adaptation has also been adopted in more challenging dense estimation tasks such as semantic segmentation [62, 45]. With appropriate adaptation strategies, models trained on synthetic datasets have achieved comparable performance to models trained with real labeled data [6, 17, 45, 54]. In this paper, we utilize domain adaptation to improve SR performance on real data.

3. DASR for Unsupervised Image SR

3.1. Methodology Overview

Given two domains described by two sets of unpaired LR images $\mathcal{Y}^r = \{y_i^r\}_{i=1,...,N}$ and HR images $\mathcal{X}^r = \{x_i^r\}_{i=1,...,M}$, we aim to learn an SR network (SRN) to enlarge the size of an image from the LR domain and simultaneously ensure the HR estimation lies in the real HR domain. To attain this goal, we follow the previous state-of-the-art methods [16, 7, 39] and propose a two-stage approach. Firstly, we train a down-sampling network (DSN) to generate LR images in the real-world LR domain from HR images: $y_i^g = DSN(x_i^r)$. Then, we utilize the generated LR-HR pairs $\{y_i^g, x_i^r\}_{i=1,...,M}$ for training the SRN. In contrast to previous works [16, 7, 39] which simply employ the generated pseudo pairs to train SRN in a supervised



(a) (b)

Figure 3: (a) Proposed DSN architecture and train losses. (b) Adversarial loss in wavelet high-frequency space.

manner, our DASR framework considers the domain bias between $\mathcal{Y}^g$ and $\mathcal{Y}^r$ and adopts domain-gap aware training and domain-distance weighted supervision strategies to take full advantage of real-world LR images as well as the generated pairs. An illustration of the proposed DASR framework is shown in Fig. 2.

In the remaining parts of this section, we firstly introduce how we train a DSN to generate synthetic LR-HR pairs. Then, we present our domain-gap aware training strategy and domain-distance weighted supervision strategy.

3.2. Training of Down-Sampling Network

Network architecture. Different down-sampling networks [16] have been trained in previous unsupervised SR works to generate synthetic real-world LR images from HR images. To avoid changing the image sizes between the input to output, existing approaches adopt a bicubic down-sampling operation as the pre-processing step. Therefore, the degradation networks only need to translate the bicubic downsampled images to the real image domain. In contrast, our DSN takes the HR image as input and captures the whole degradation process with the network directly. Thus, without losing information in the bicubic down-sampling step, all the information in HR images can be exploited for generating better synthetic LR images. Our detailed network architecture can be found in Fig. 3 (a). DSN utilizes 23 residual blocks to extract information from the HR image, each residual block contains two convolutional layers (with kernel size 3×3 and channel number 64) and a ReLU activation in between. Then, a bilinear resize operator and two convolutional layers are adopted to reduce the spatial resolution of features and project the features back to the image domain.

Losses. We train our DSN with a combination of multiple loss functions. To keep the content of generated LR image consistent with the input HR image, we apply content loss $\mathcal{L}_{con}$ and perceptual loss $\mathcal{L}_{per}$ to constrain the distance between generated LR image $y_i^g = DSN(x_i^r)$ and bicubic downsampled HR image y_i^b:

$$\begin{aligned} \mathcal{L}_{con} &= \mathbb{E}_{x^r}||y_i^b - DSN(x_i^r)||_1, \\ \mathcal{L}_{per} &= \mathbb{E}_{x^r}\|\phi(y_i^b) - \phi(DSN(x_i^r))\|_1; \end{aligned} \tag{1}$$

where $y_i^b = \mathcal{B}(x_i^r)$ is the bicubic downsampled HR image, and $\phi(\cdot)$ denotes the VGG [47] feature extractor. In our implementation, we follow ESRGAN [52] and calculate perceptual loss on VGG-19 [47] features from *conv5_3* convolutional layer. While to achieve the goal of domain translation, we impose adversarial losses between image samples in $\mathcal{Y}^g$ and $\mathcal{Y}^r$. We adopt a similar idea with FSSR [16], which only imposes adversarial loss in the high-frequency space. But we use Haar wavelet transform to extract more informative high-frequency components. Concretely, denote the four sub-bands decomposed by Haar wavelet transform as $\mathcal{LL}$, $\mathcal{LH}$, $\mathcal{HL}$ and $\mathcal{HH}$, we stack the $\mathcal{LH}$, $\mathcal{HL}$ and $\mathcal{HH}$ components as the input to the discriminator. Compared with the high-frequency extractor used in FSSR [16], our wavelet-based extractor also exploits direction information to better characterize image details. The GAN loss for generator (*i.e.* our DSN) is defined as:

$$\mathcal{L}_{adv}^G = -\mathbb{E}_{x^r}[log\,(D\,(\mathcal{H}_{wavelet}\,(DSN(x^r))))]; \tag{2}$$

and the GAN loss for training the discriminator is in a symmetrical form:

$$\begin{aligned} \mathcal{L}_{adv}^D = &- \mathbb{E}_{y^r}[log(D(\mathcal{H}_{wavelet}(y^r)))] \\ &- \mathbb{E}_{x^r}[log\,(1 - D\,(\mathcal{H}_{wavelet}\,(DSN(x^r))))]. \end{aligned} \tag{3}$$

$\mathcal{H}_{wavelet}(\cdot)$ in Eqs. (2) and (3) represents extracting $\mathcal{LH}$, $\mathcal{HL}$ and $\mathcal{HH}$ subbands with Haar wavelet transform and concatenating the three variables. Imposing the adversarial loss in the high-frequency domain enables us to ignore the low-frequency content which is less relevant to the SR task [16] and focus more on the image details. Moreover, conducting adversarial training in lower-dimension space also reduces the difficulty of GAN training [33, 53].

In our implementation, we adopt a similar strategy as CycleGAN [65], which imposes GAN loss on each patch. Concretely, we utilize a 4 layer fully convolutional discriminator, the patch discriminator has a valid receptive field of 23 × 23. The PatchGAN strategy helps to derive the patch-level dense domain distance map, which will be utilized in the subsequent training phase of SRN. We refer to our suppl. material for more details of our patch discriminator.

Training Details. Our DSN is trained using the loss:

$$\mathcal{L}_{DSN} = \alpha\mathcal{L}_{con} + \beta\mathcal{L}_{per} + \gamma\mathcal{L}_{adv}^G. \tag{4}$$

To stabilize our training, we pre-train our DSN network with content loss. After a pre-train process of 25000 iterations, the α, β and γ in Eq. (4) are set as 0.01, 1 and 0.0005, respectively. We train the DSN networks with 192 × 192 HR crops, the batch size is set as 16. The initial learning rate is 0.0001, and we halve it every 10000 iterations. We train the model for 50000 iterations.

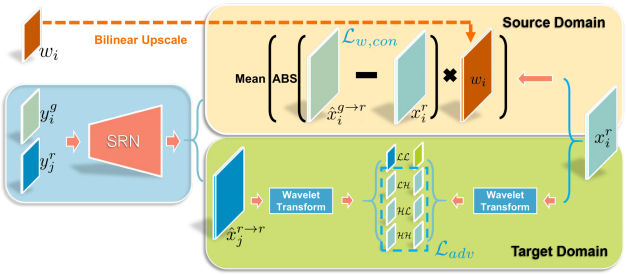


Figure 4: Domain-distance aware training of our SRN.

3.3. Domain distance aware training of SRN

With the aforementioned DSN, we are able to generate synthetic paired data $\{y_i^g, x_i^r\}_{i=1,...,M}$. However, as shown in Fig. 1, a domain gap still exists between generated LR images $\mathcal{Y}^g$ and real LR images $\mathcal{Y}^r$. When the SR network trained on synthetic data is applied to super-resolve real-world LR images, such a domain gap between training and testing data will lead to a performance drop. To alleviate the domain bias issue, we consider a domain adaptation setting and incorporate both source domain labeled data $\{\mathcal{Y}^g, \mathcal{X}^r\}$ and target domain unlabeled data $\mathcal{Y}^r$ in the training of our SR network. The core of our adaptation strategy consists of two parts which include domain-gap aware training and domain-distance weighted supervision.

Domain-gap aware training. Given training samples from source and target domains, we utilize different losses in the two domains to take full advantage of the training data (see Fig. 4). For the data in the source domain, which have supervised labels, we deploy losses to train the network in a supervised manner. While, for target domain data, which do not have labels, we impose adversarial losses to align the distribution of their outputs $\hat{\mathcal{X}}^{r\to r} = SRN(\mathcal{Y}^r)$ and the distribution of real HR images $\mathcal{X}^r$. The same as our DSN training, we introduce GAN losses in the wavelet space.

$$\begin{aligned}\mathcal{L}^G_{target,adv} &= -\mathbb{E}_{y^r}[log(D(\mathcal{H}_{wavelet}(SRN(y^r))))];\\ \mathcal{L}^D_{target,adv} &= -\mathbb{E}_{x^r}[log(D(\mathcal{H}_{wavelet}(x^r)))]\\ &\quad -\mathbb{E}_{y^r}[log(1-D(\mathcal{H}_{wavelet}(SRN(y^r))))].\end{aligned} \tag{5}$$

Besides introducing $\mathcal{L}_{target,adv}$ to guide network training with target domain information, making rational use of information in the source domain is of equal importance for obtaining a good SRN. In the following part, we introduce how the domain distance information of each sample can be used for adaptively supervised training of SRN.

Domain-distance weighted supervision. As shown in Fig. 1, each sample in $\mathcal{Y}^g$ has a distinct distance to the real-world image domain $\mathcal{Y}^r$. More specifically, since the difference between images from different domains only lies in their low-level details, each area of generated images may possess diverse domain distances to the real-world image domain. When being applied as source domain data to train target domain SRN, different areas should be endowed with various importance based on their respective distance to the target domain. We therefore propose a weighted supervision strategy which utilizes a dense domain distance map to adaptively adjust the losses for each pair $\{y_i^g, x_i^r\}$. The weighted supervised losses in the source domain can be written as:

$$\begin{aligned}\mathcal{L}_{source,con} &= \mathbb{E}_{y_i^g,x_i^r}\|w_i \odot ((SRN(y_i^g) - x_i^r)\|_1,\\ \mathcal{L}_{source,per} &= \mathbb{E}_{y_i^g,x_i^r}\|w_i \odot (\phi(SRN(y_i^g)) - \phi(x_i^r))\|_1;\end{aligned} \tag{6}$$

where w_i is the domain distance map for y_i^g, and $\odot$ denotes the point-wise multiplication. We utilize the discriminators obtained during the training process of DSN to evaluate the domain distance map for each sample. Note that the discriminator is trained to distinguish the generated patches from the real-world LR patches and the discriminator output denotes the possibility that the input comes from the target domain. Thus, the larger the discriminator output, the higher the possibility that the input comes from the target real-world LR domain and the less the distance to the target domain. We directly utilize bilinear resize to adjust the spatial size of discriminator outputs, and utilized the resize weight map to weigh the importance of each local area.

Training details. In summary, with our domain-distance aware training strategy, SRN is trained through minimizing the following losses:

$$\mathcal{L}_{SRN} = \alpha\mathcal{L}_{source,con} + \beta\mathcal{L}_{source,per} + \gamma\mathcal{L}_{target,adv}. \tag{7}$$

The same as our training schedule for DSN, we pretrain our SRN with content loss in the source domain. After 25000 iterations of pretraining, we employ all the losses in Eq. (7) with weights $\alpha = 0.01$, $\beta = 1$ and $\gamma = 0.005$ to train the network for another 50000 iterations. We initialize the learning rate as 0.0002, and halve it every 10000 iterations.

Our adaptation strategy is applicable to diverse network architectures. In this paper, we directly adopt the architecture used in ESRGAN [52] as our SRN.

4. Experimental Results on Synthetic Datasets

4.1. Experimental Setting

In this section, we evaluate the proposed DASR method on the AIM dataset, which was used in the AIM Challenge on Real World SR at ICCV 2019 [40]. The dataset was simulated by applying synthetic but realistic degradations to clean high-quality images. We follow the experimental setting of *target domain super resolution* in the Challenge. The training set consists of 2650 noisy and compressed images with unknown degradation from the Flickr2K dataset [1], and 800 clean HR images from the DIV2K [50] dataset. We conduct our experiments on the validation dataset of the AIM challenge, which has paired data for quantitative comparison. The validation dataset contains 100 images with

DSN Input	$\mathcal{L}_{adv}$ for DSN	LPIPS↓	PSNR↑	SSIM↑
Bicubic LR	GBFS	0.110	25.258	0.8081
HR	RGB	0.138	25.204	0.8153
HR	GBFS	0.101	25.474	0.8156
HR	WFS	0.067	26.007	0.8097

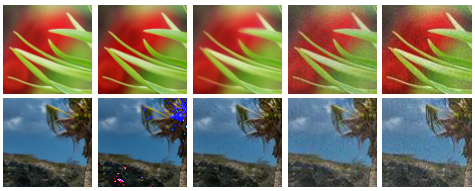


Table 1: LR image generation with different DSN architectures and adversarial training criterion. Details of the experiments are described in section 4.2.

Source domain	$\{\mathcal{Y}^b, \mathcal{X}^r\}$	$\{\mathcal{Y}^b, \mathcal{X}^r\}$	$\{\mathcal{Y}^g, \mathcal{X}^r\}$	$\{\mathcal{Y}^g, \mathcal{X}^r\}$	$\{\mathcal{Y}^g, \mathcal{X}^r\}$	$\{\mathcal{Y}^g, \mathcal{X}^r\}$
Target domain	-	$\mathcal{Y}^r$	-	$\mathcal{Y}^r$	-	$\mathcal{Y}^r$
Domain gap aware	✗	✓	✗	✓	✗	✓
Weighted sup.	✗	✗	✗	✗	✓	✓
PSNR↑	21.382	20.820	21.910	21.805	21.452	21.600
SSIM↑	0.5478	0.5103	0.5555	0.5615	0.5304	0.5640
LPIPS↓	0.543	0.390	0.378	0.359	0.348	0.336
MOS↓	3.16	2.87	2.41	2.37	2.31	1.94

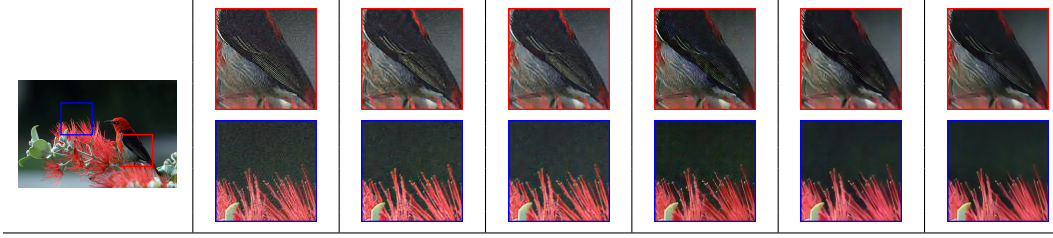


Table 2: Ablation study on the AIM dataset [40]. We evaluate our *Domain gap aware training* and *Domain distance weighted supervision* strategies in different conditions. Details of the experiments are described in 4.2.

the same type of degradation as the training LR images. Since the GAN approaches focus on the perceptual quality of the recovered image, Learned Perceptual Image Patch Similarity (LPIPS) [59] and Mean Opinion Score (MOS) are used as the primary metrics to evaluate different methods. A user study is conducted to calculate the MOS for different methods. The test candidates were shown a side-by-side comparison of a sample result and the corresponding ground-truth. The final MOS of a specific image is the average score of different candidates' opinion: 0 - 'the same', 1 - 'very similar', 2 - 'similar', 3 - 'not similar' and 4 - 'different'. For all the MOS values reported in the paper, we have the same 26 candidates to perform the user study. In addition to the perceptual metrics, the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) are provided for reference.

4.2. Ablation study

Before comparing DASR with state-of-the-art unsupervised real-world image SR methods, we conduct ablation experiments to analyze our DASR model. We firstly analyze our design choice for DSN training. Then, we provide experimental results to demonstrate the effectiveness of the proposed Domain-gap aware training and Domain-distance weighted supervision strategies.

Better down-sampling network for synthetic paired data generation. Our DSN improves upon previous down-sampling networks [16, 7, 39] by directly estimating LR image from un-preprocessed HR image and adopting better adversarial loss in wavelet space. In order to evaluate the effect of our modifications, we train downsampling networks with different settings, and use these models to generate LR images from the HR images in the AIM validation dataset. We compare the generated LR images with the original LR images in the datasets, the quantitative metrics achieved by different models are reported in Table 1 side by side with some visual examples of the LR images generated. In the table, HR/Bicubic LR denotes the respective inputs used by different down-sampling networks. While, Gaussian Blur Frequency Separation (GBFS), Wavelet Frequency Separation (WFS) and RGB indicate the model conducts adversarial training in different spaces: GBFS uses the residual between original and Gaussian blurred images to extract high-frequency component, our WFS approach adopts Wavelet transform to obtain high-frequency component, RGB means we introduce GAN loss directly on RGB images. The results in Table 1 show that both the proposed architecture of DSN and adversarial loss in the wavelet space are beneficial for generating better LR images, which are more similar to the real images in the target domain.

Domain-gap Aware Training. As one of our major contributions, domain-gap aware training is of vital importance for the success of our model. In Table 2, we present experimental numerical and visual results to show the advantage of our domain-gap aware training. We conduct experi-

Methods	AIM [40] 4×				RealSR [8] 4×				RealSR [8] 2×				CameraSR [9] 1×			
	PSNR↑	SSIM↑	LPIPS↓	MOS↓	PSNR↑	SSIM↑	LPIPS↓	MOS↓	PSNR↑	SSIM↑	LPIPS↓	MOS↓	PSNR↑	SSIM↑	LPIPS↓	MOS↓
ZSSR	22.327	0.6022	0.630	3.10	26.007	0.7482	0.386	3.42	30.563	0.8787	0.1756	2.51	-	-	-	-
P.T. ESRGAN	21.382	0.5478	0.543	3.16	25.956	0.7468	0.415	3.08	30.397	0.8725	0.1720	2.46	-	-	-	-
CinCGAN	21.602	0.6129	0.461	3.56	25.094	0.7459	0.405	3.24	28.099	0.8665	0.1663	2.23	-	-	-	-
FSSR	20.820	0.5103	0.390	2.41	25.992	0.7388	0.265	2.45	30.397	0.8737	0.1421	1.89	23.781	0.7566	0.180	3.14
DASR(ours)	21.600	0.5640	0.336	1.94	26.782	0.7822	0.228	2.05	29.887	0.8670	0.1291	1.43	25.769	0.8312	0.151	2.60
S.T. ESRGAN	-	-	-	-	25.704	0.7487	0.199	1.35	30.648	0.8802	0.0970	1.25	25.346	0.8036	0.111	1.18

Table 3: Quantitative comparison on different datasets. More experimental details can be found in section 4 and 5. Please note that supervisely trained ESRGAN (S.T. ESRGAN) is trained with paired training data while the other methods are trained without paired training data.

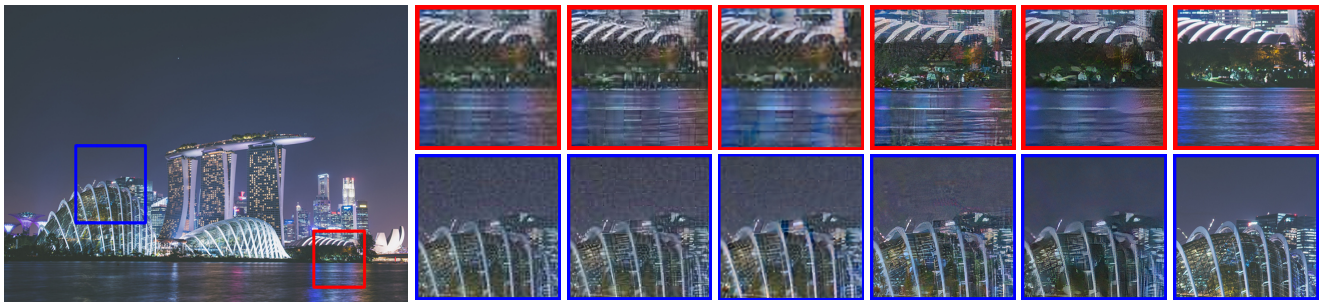


Figure 5: SR results by different methods on testing images from AIM Challenge on Real World SR at ICCV 2019 [40].

ments with our DSN generated synthetic pairs $\{\mathcal{Y}^g, \mathcal{X}^r\}$ or bicubic downsampled LR-HR pairs $\{\mathcal{Y}^b, \mathcal{X}^r\}$ as source domain data. Our domain-gap aware training strategy (see section 3.2) introduces extra adversarial loss in the target domain. For a fair comparison, in Table 2, the models without domain-gap aware training introduce the adversarial loss in the source domain. The same strategy has been widely adopted in previous unsupervised methods [52, 35, 60]. In both settings, the proposed domain-aware training strategy consistently improves the final SR performance. It helps the SRN to generate high-quality HR estimations with better MOS as well as a better LPIPS index. Because the MOS is achieved by comparing the subject image with its corresponding reference image, images with different visual quality may be categorized as the same class, *i.e.* similar or not similar. Therefore, the MOS could not thoroughly reflect the advantage of our domain-gap aware training strategy. On the other hand, the LPIPS index can validate the effectiveness of our domain-gap aware training clearly. By introducing target domain data in the training process of SRN, even the model trained with bicubic downsampled LR-HR pairs $\{\mathcal{Y}^b, \mathcal{X}^r\}$ generalize well on real-world LR images.

Domain-distance weighted Supervision. Besides the domain-gap aware training, we also proposed a domain-distance weighted supervision strategy to make better use of source domain data. The experimental results in Table 2 clearly show the advantage of domain-distance weighted supervision. By introducing weights to adaptively exploit paired training data, we are able to achieve better SRN over the baseline models. In addition, the proposed two strategies are complementary, when combining the two strategies together, DASR achieves significant improvement over the models which only adopts one of the two strategies.

4.3. Comparison with State-of-the-Arts.

In this section, we compare our method with other super-resolution method on the AIM dataset [40]. The competing approaches include Zero-Shot SR (ZSSR) [4] and unpaired learning approaches Frequency Separation for Super Resolution (FSSR) [16] and cycle-in-cycle generative adversarial networks (CinCGAN) [56]. ZSSR applies a Zero-Shot learning strategy in the testing phase to adapt to the image-specific degradation model. CinCGAN and FSSR are recently proposed unsupervised SR approaches, FSSR is the winner of the AIM Challenge on Real World SR at ICCV 2019 [40]. The code of FSSR [16] is provided by the paper authors, and CinCGAN model [56] is implemented by ourselves. Moreover, we also provide the results by pre-trained ESRGAN (denote as P.T. ESRGAN) for reference, the pre-trained ESRGAN model was trained on a synthetic dataset with bicubic downsampled LR images. The quantitative results achieved by different methods are shown in Table 3, while in Fig. 5, we also provide some visual SR results. The quantitative metrics as well as visual examples clearly demonstrate that our proposed DASR approach is superior to the competing models. The degradation assumptions by ZSSR and Pre-trained ESRGAN can not reflect the complex degradation adopted in the AIM challenge, both two approaches generate strange artifacts in the HR estimation. FSSR generates better synthetic data which have similar characteristic with a real-world image to train the model and is able to deliver better SR results than the ZSSR and pre-trained ESRGAN approach. But FSSR does not consider

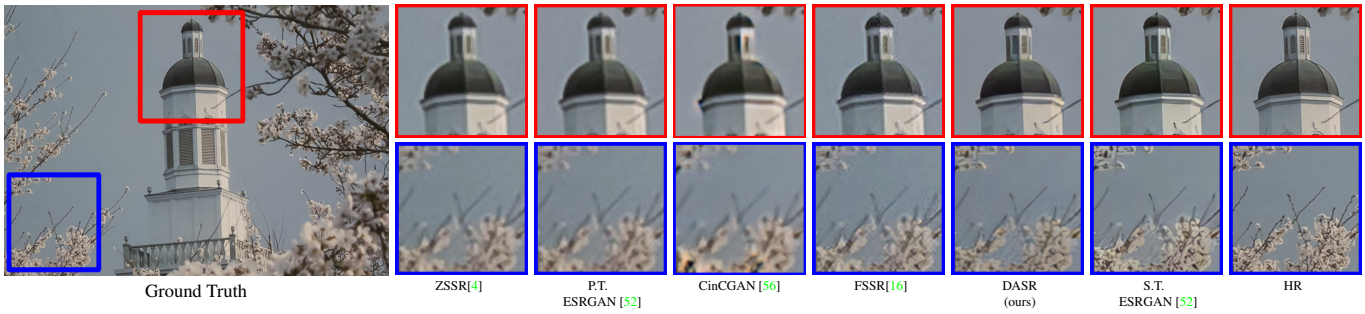


Figure 6: SR results by different methods on testing images from RealSR [8].

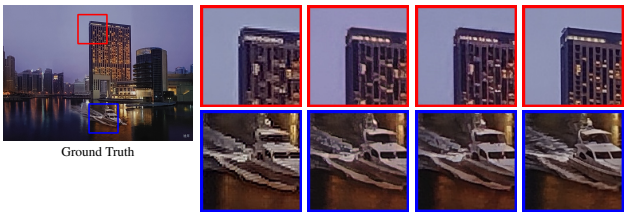


Figure 7: SR results by different methods on a testing image from CameraSR [9].

the domain gap between generated and real LR images, still create artifacts in the final output. Our novel DASR exploits information in the target domain in the training phase, is able to generate high-quality SR estimations which have visually pleasant textures and fewer artifacts. More visual examples can be found in our supplementary material.

5. Experimental Results on Real Images

In this section, we evaluate the proposed DASR model on two real image datasets: RealSR [8] and CameraSR [9]. The two datasets contain real LR-HR image pairs collected by adjusting the focal length of digital cameras. We adopt the LR images in the two datasets and HR images in the DIV2K [50] dataset to deploy our unsupervised training, and evaluate our models on the validation sets which have paired data for quantitative evaluation.

5.1. Experimental Results on RealSR Dataset

RealSR [8] is a recently collected real image SR dataset. The authors utilize a Canon and a Nikon camera to collect 595 real LR-HR image pairs by adjusting the focal length of the cameras, and adopt an image registration algorithm to achieve aligned image pairs. In our experiments, we utilize the 200 LR images collected by the Canon camera as our real-world LR images, and the 800 HR images in the DIV2K [50] as our HR images. We train our DASR model as well as FSSR [16] and CinCGAN [56] models with the same data. After unsupervised training, we employ our model to super-resolve LR images in the validation set of RealSR [8], which consists of 100 LR-HR pairs. The SR results generated by our model and the competing approaches are shown in Table 3. Besides the ZSSR [4], FSSR [16] and pre-trained ESRGAN [52], we also provide the results by supervisely trained ESRGAN (denote as S.T. ESRGAN) for reference, which utilizes the real paired data in the training set to train the ESRGAN model in a fully supervised manner. DASR significantly outperforms other blind super-resolution methods in both LPIPS and MOS. Compared with the Supervised ESRGAN, DASR achieves comparable LPIPS indexes. Some visual examples by different approaches are shown in Fig. 6, more visual examples can be found in our supplementary file.

5.2. Experimental Results on CameraSR

We also compare different approaches on the CameraSR [9] dataset. CameraSR contains 100 LR-HR pairs captured by an iPhoneX and a Nikon Camera, respectively. We test our method on the iPhoneX subset. As the LR and HR images in the dataset are of the same spatial size, we remove the down-sampling and up-sampling operations in our framework as well as the FSSR model. Similar to our experiments on the RealSR dataset, we use 100 LR images in the CameraSR training set and 800 HR images in DIV2K [50] to train our model and FSSR. The SR results by different methods are shown in Table 3. DASR outperforms FSSR by a large margin. Visual examples are shown in Fig. 7, more results can be found in our supplementary file.

6. Conclusions

We propose a novel DASR framework for unsupervised real-world image SR. Given only unpaired data, DASR firstly trains a down-sampling network to generate synthetic LR images in the real-world LR distribution. Then, the generated synthetic pairs and real LR images are exploited to train the SR network under a domain adaptation setting. We proposed a domain-gap aware training strategy to introduce an adversarial loss in the target domain, and a domain-distance weighted supervision strategy to take better advantage of synthetic data in the source domain. Our experimental results on synthetic and real-world datasets demonstrate the effectiveness of our approach for real-world SR.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, 2017. 5

[2] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *ECCV*, 2018. 2

[3] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Image super-resolution via progressive cascading residual network. In *Proc. CVPR*, 2018. 2

[4] Michal Irani Assaf Shocher, Nadav Cohen. "zero-shot" super-resolution using deep internal learning. In *Proc. CVPR*, 2018. 2, 3, 7, 8

[5] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. In *NIPS*, 2019. 2, 3

[6] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proc. CVPR*, 2017. 3

[7] 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. 2, 3, 6

[8] 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 *Proc. CVPR*, 2019. 1, 7, 8

[9] Chang Chen, Zhiwei Xiong, Xinmei Tian, Zheng-Jun Zha, and Feng Wu. Camera lens super-resolution. *CoRR*, 2019. 1, 7, 8

[10] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. *Proc. CVPR*, 2018. 3

[11] CornillèreVictor, DjelouahAbdelaziz, YifanWang, Sorkine-HornungOlga, and SchroersChristopher. Blind image super-resolution with spatially variant degradations. *ACM Transactions on Graphics*, 2019. 2, 3

[12] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, 2014. 2

[13] Yuchen Fan, Honghui Shi, Jiahui Yu, Ding Liu, Wei Han, Haichao Yu, Zhangyang Wang, Xinchao Wang, and Thomas S. Huang. Balanced two-stage residual networks for image super-resolution. *CVPRW*, 2017. 2

[14] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Proc. CVPR*, 2013. 3

[15] William T Freeman, Egon C Pasztor, and Owen T Carmichael. Learning low-level vision. *International journal of computer vision*, 2000. 1

[16] Manuel Fritsche, Shuhang Gu, and Radu Timofte. Frequency separation for real-world super-resolution. *ArXiv*, 2019. 1, 2, 3, 4, 6, 7, 8

[17] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. *ArXiv*, 2014. 3

[18] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *ECCV*, 2016. 3

[19] Daniel Glasner, Shai Bagon, and Michal Irani. Super-resolution from a single image. In *2009 IEEE 12th international conference on computer vision*, 2009. 1

[20] 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

[21] Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *2011 international conference on computer vision*, 2011. 3

[22] Shuhang Gu, Wangmeng Zuo, Qi Xie, Deyu Meng, Xiangchu Feng, and Lei Zhang. Convolutional sparse coding for image super-resolution. In *ICCV*, 2015. 1

[23] Zhen Han, Enyan Dai, Xu Jia, Shuaijun Chen, Chunjing Xu, Jianzhuang Liu, and Qi Tian. Unsupervised image super-resolution with an indirect supervised path. *arXiv preprint arXiv:1910.02593*, 2019. 2

[24] Z. Han, Enyan Dai, Xu Jia, Xiaoying Ren, Shuaijun Chen, Chunjing Xu, J. Liu, and Q. Tian. Unsupervised image super-resolution with an indirect supervised path. 2020. 2, 3

[25] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. *Proc. CVPR*, 2018. 2

[26] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, G. Klambauer, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. *ArXiv*, abs/1706.08500, 2017. 2

[27] Saifuddin Hitawala, Yao Li, Xian Wang, and Dongyang Yang. Image super-resolution using vdsr-resnext and srcgan. *ArXiv*, 2018. 1, 2

[28] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proc. CVPR*, 2015. 2

[29] Yiwen Huang and Ming Qin. Densely connected high order residual network for single frame image super resolution. *ArXiv*, 2018. 2

[30] Xiaozhong Ji, Yun Cao, Ying Tai, Chengjie Wang, Jilin Li, and Feiyue Huang. Real-world super-resolution via kernel estimation and noise injection. In *CVPRW*, 2020. 2, 3

[31] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 1, 2

[32] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. *Proc. CVPR*, 2015. 2

[33] Szymon Knop, Marcin Mazur, Jacek Tabor, Igor T. Podolak, and Przemyslaw Spurek. Sliced generative models. *ArXiv*, 2019. 4

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

[35] Christian Ledig, Lucas Theis, Ferenc Huszar, José Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. *Proc. CVPR*, 2016. 3, 7

[36] Yawei Li, Vagia Tsiminaki, Radu Timofte, Marc Pollefeys, and Luc Van Gool. 3D appearance super-resolution with deep learning. In *Proc. CVPR*, 2019. 1

[37] Yudong Liang, Radu Timofte, Jinjun Wang, Yihong Gong, and Nanning Zheng. Single image super resolution - when model adaptation matters. *ArXiv*, 2017. 3

[38] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017. 1, 2

[39] Andreas Lugmayr, Martin Danelljan, and Radu Timofte. Unsupervised learning for real-world super-resolution. *ArXiv*, 2019. 2, 3, 6

[40] Andreas Lugmayr, Martin Danelljan, Radu Timofte, Manuel Fritsche, Shuhang Gu, Kuldeep Purohit, Praveen Kandula, Maitreya Suin, AN Rajagopalan, Nam Hyung Joon, et al. Aim 2019 challenge on real-world image super-resolution: Methods and results. *ArXiv*, 2019. 1, 2, 5, 6, 7

[41] Tomer Michaeli and Michal Irani. Nonparametric blind super-resolution. In *ICCV*, pages 945–952, 2013. 2, 3

[42] Jian Ren, Ilker Hacihaliloglu, Eric A Singer, David J Foran, and Xin Qi. Adversarial domain adaptation for classification of prostate histopathology whole-slide images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2018. 3

[43] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proc. CVPR*, 2018. 3

[44] Mehdi SM Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proc. CVPR*, 2017. 3

[45] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Unsupervised domain adaptation for semantic segmentation with gans. *ArXiv*, 2017. 3

[46] Ozan Sener, Hyun Oh Song, Ashutosh Saxena, and Silvio Savarese. Learning transferrable representations for unsupervised domain adaptation. In *NIPS*, 2016. 3

[47] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, 2014. 4

[48] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *ICCV*, 2017. 1

[49] Radu Timofte, Vincent De Smet, and Luc Van Gool. Anchored neighborhood regression for fast example-based super-resolution. In *ICCV*, 2013. 1

[50] Radu Timofte, Shuhang Gu, Jiqing Wu, Luc Van Gool, Lei Zhang, Ming-Hsuan Yang, Muhammad Haris, et al. Ntire 2018 challenge on single image super-resolution: Methods and results. In *CVPRW*, 2018. 5, 8

[51] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In *Proc. CVPR*, 2017. 2

[52] 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 *ECCVW*, 2018. 1, 2, 3, 4, 5, 7, 8

[53] Jiqing Wu, Zhiwu Huang, Dinesh Acharya, Wen Li, Janine Thoma, Danda Pani Paudel, and Luc Van Gool. Sliced wasserstein generative models. In *Proc. CVPR*, 2019. 4

[54] Jin Xiao, Shuhang Gu, and Lei Zhang. Multi-domain learning for accurate and and few-shot color constancy. In *Proc. CVPR*, 2020. 3

[55] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 2010. 1

[56] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. *CVPRW*, 2018. 2, 3, 7, 8

[57] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *Proc. CVPR*, 2017. 1

[58] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *Proc. CVPR*, 2018. 2, 3

[59] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. CVPR*, 2018. 6

[60] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *ICCV*, 2019. 1, 3, 7

[61] Xuaner Cecilia Zhang, Qi feng Chen, Ren Ng, and Vladlen Koltun. Zoom to learn, learn to zoom. *Proc. CVPR*, 2019. 1

[62] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *ICCV*, 2017. 3

[63] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. *Proc. CVPR*, 2018. 1, 2

[64] R. Zhou and S. Süsstrunk. Kernel modeling super-resolution on real low-resolution images. In *ICCV*, 2019. 2, 3

[65] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. CVPR*, 2017. 4