Zero-Shot Image Super-Resolution with Depth Guided Internal Degradation Learning

Xi Cheng, Zhenyong Fu(✉), and Jian Yang(✉)

Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education,
Jiangsu Key Lab of Image and Video Understanding for Social Security,
PCA Lab, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
{chengx,z.fu,csjyang}@njust.edu.cn

Abstract. In the past few years, we have witnessed the great progress of image super-resolution (SR) thanks to the power of deep learning. However, a major limitation of the current image SR approaches is that they assume a pre-determined degradation model or kernel, e.g. *bicubic*, controls the image degradation process. This makes them easily fail to generalize in a real-world or non-ideal environment since the degradation model of an unseen image may not obey the pre-determined kernel used when training the SR model. In this work, we introduce a simple yet effective zero-shot image super-resolution model. Our zero-shot SR model learns an image-specific super-resolution network (SRN) from a low-resolution input image alone, without relying on external training sets. To circumvent the difficulty caused by the unknown internal degradation model of an image, we propose to learn an image-specific degradation simulation network (DSN) together with our image-specific SRN. Specifically, we exploit the depth information, naturally indicating the scales of local image patches, of an image to extract the unpaired high/low-resolution patch collection to train our networks. According to the benchmark test on four datasets with depth labels or estimated depth maps, our proposed depth guided degradation model learning-based image super-resolution (DGDML-SR) achieves visually pleasing results and can outperform the state-of-the-arts in perceptual metrics.

Keywords: Image super-resolution, Zero-shot, Depth guidance

1 Introduction

Single image super-resolution (SR) aims to restore a high-resolution (HR) image from a degraded low-resolution (LR) measurement. Image super-resolution, as an inverse procedure of image downscaling, is an ill-posed problem, in which the internal degradation patterns or kernels followed by images are image-specific and unknown. Most modern image super-resolution models, mainly based on supervised learning techniques such as deep convolutional neural networks (CNNs) [11], rely on massive amounts of high-/low-resolution example pairs for training. In

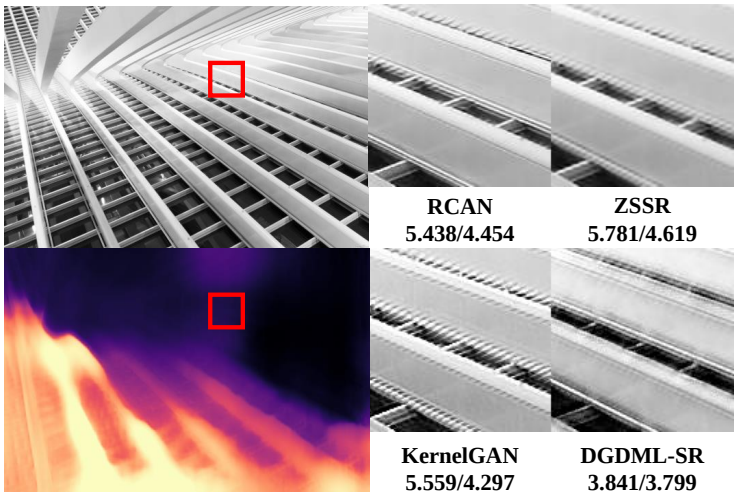


Fig. 1: Our proposed DGDML-SR can achieve better visual quality than the state of the arts. NIQE and PI scores (lower is better) are shown under each image.

reality, collecting a natural pair of high/low-resolution images is difficult; existing SR methods [6, 5, 14, 29, 30] resort to manually designed HR/LR image pairs as a surrogate. In these methods, a given high-resolution image is downscaled to generate a low-resolution counterpart using a simple and pre-determined degradation kernel, e.g. a *bicubic* operation, in order to acquire a pair of HR/LR images. In general, current supervised learning-based SR methods generalize poorly due to the simplified degradation model, especially when dealing with the images with details having not been encountered in the training set.

To overcome the above drawbacks, the key is to model the natural degradation in the images. However, degradation kernels are usually complex and differ greatly; they can be affected by many factors such as luminance, sensor noise, motion blur and compression. Thus, to learn a natural degradation and implement a visually pleasing super-resolution, we should treat each image independently, i.e. zero-shot image degradation and super-resolution. Zero-shot image super resolution is more challenging that needs to learn an image-specific SR model from an image alone, without access to external training sets. The main difficulty of zero-shot SR is to acquire HR/LR image patches for training. Recently, Shocher et al. [19] have proposed a zero-shot super-resolution (ZSSR) approach which extracts local patches from an image and then downscales them using a pre-determined bicubic operation, analogously to other supervised learning-based SR methods. As such, they construct a patch-level training collection composed of high-/low-resolution pairs of local patches from a single image. However, a natural image is seldom degraded by obeying a simple bicubic rule and an unreasonable assumption about the degradation kernel will impede its in-

verse super-resolution procedure. Thus, the problem of the unknown degradation model is still far from solved in existing zero-shot SR works.

In this paper, we present a simple yet effective zero-shot SR method without assuming a pre-defined degradation kernel. Instead, we learn the image-specific degradation model in a self-supervised manner. In our method, we sidestep the difficulty of acquiring the patch-level HR/LR training data by leveraging the image depth information. The depth information indicates the distance of each image region relative to the camera. Depth information can be easily computed using a pre-trained depth estimation model [7] or obtained from datasets with depth labels [25,9, 20]. Also, the depth or time-of-flight (TOF) camera is becoming increasingly popular on mobile phones, simplifying the acquisition of depth information. In our method, we view the short-distance local regions as the HR patches, while the distant local regions as the LR patches. After acquiring the HR/LR patch collection, we design two fully-convolutional and image-specific networks: degradation simulation network (DSN), responsible for imitating the unknown degradation kernel of the image, and super-resolution network (SRN), in charge of performing the SR task on the image. Since we have no paired HR/LR local patches but the unpaired HR/LR patch collection, we design a bicycle training strategy to learn our degradation simulation network and super-resolution network simultaneously. Guided by the image-specific degradation model internally learned by DSN with the clue of depth information, our zero-shot or image-specific SR network can achieve a satisfactory SR result for a single image, without using any external training set except the image itself. As depicted in Fig. 1, our method can achieve the best NIQE and PI scores, recover more natural and clear textures, and have fewer artifacts.

Our contributions are three-fold: (1) we propose a zero-shot image SR model that does not require the high-resolution labels; (2) our method leverages the depth information of an image and can learn the internal degradation model of the image in a self-supervised manner; and (3) our method can outperform the state-of-the-art in perceptual metrics. Compared with the latest zero-shot and supervised SR methods (e.g. KernelGAN [2]), our approach is average 0.555 better in NIQE [17] and 0.284 better in PI [4] according to the benchmark.

2 Related Work

In the past several years, deep convolutional neural networks (CNN) based image SR models have been proposed [6, 10, 12, 30, 5]. Compared with traditional methods [8, 23], CNN-based methods are superior in terms of peak signal-to-noise ratio (PSNR). Since the pioneering work of SRCNN [6], CNN-based SR models have been boosted with deeper structures [10, 21] by using a progressive upsampling way [12] or a dense structure [22, 30]. Although the CNN-based SR methods can achieve excellent PSNR results, their results are not visually pleasing, since they typically use the Mean Squared Error (MSE) loss, inherently leading to a blurry high-resolution result. To overcome this problem, some new loss functions have been proposed to replace MSE [12, 16, 15]. Recently, genera-

tive adversarial networks (GANs) based SR models [13, 24] have been shown to produce more realistic high-resolution results with finer details.

A fundamental limitation of the aforementioned methods is that they unrealistically assume a pre-defined degradation model, e.g. bicubic, to be used in image SR. In reality, the degradation model is unknown and more complex than bicubic, often accompanied by severe distracting factors. Thus, existing supervised SR methods generally fail to obtain a satisfactory SR result in a natural environment outside the training condition. RCAN [30] and SRMD [28] added a variety of conditions (e.g. noise and blur) to the degradation model and can improve the SR result in the natural environment. Xu et al. [26] shot photos on real scenes and used raw images from the digital camera sensors to train an image super-resolution model to fit the natural environment. However, they still fail to address the problem of the unknown degradation model.

To mitigate the deficiency of supervised learning-based SR models in real-world environments, CincGAN [27] learned the degradation model on tasks such as noise reduction in an unsupervised manner, but it still used the bicubic downscaling as an intermediate state. Moreover, CincGAN [27] needs a large external training dataset, which is usually unavailable in real-world environments. Zero-shot image super-resolution [19], aiming to learn an SR model from a single image alone and then apply it to super-resolve the image, has drawn considerable attention recently. Zero-shot super-resolution (ZSSR) [19] used a fixed degradation method such as bicubic as a degradation model for local patches. Bell et al. [2] proposed a novel SR method named KernelGAN that used a deep linear generator to learn the downscaling kernel from a single image. Then, they applied ZSSR to perform the super-resolution on the image with the downscaling kernel learned by KernelGAN. Their method greatly improved ZSSR but cannot model the complex and superposed degradation in reality. Moreover, learning the degradation model via KernelGAN and learning the SR model via ZSSR are separated, thus often resulting in a suboptimal SR result.

3 Approach

A natural image is self-explained in that similar local patches tend to recur across positions and scales within the image [8]. Moreover, similar patches in the original scenery will be rescaled during the imaging process due to the changes of depth. The depth measures the distance of each patch in an image relative to the camera. The image patches near the camera will be enlarged, while the patches with similar appearance but away from the camera will be shrunk. In other words, the distant patches captured in an image are more blurred and smaller in appearance than the short-distance patches. We call it depth guided self-similarity prior in images. Fig. 2 gives an example of this prior, in which the relative distance of an image patch can be continuously measured by the depth; the image patches with similar textures become more blurred when their depths are deeper. In this work, we exploit the depth guided information in single image super-resolution.

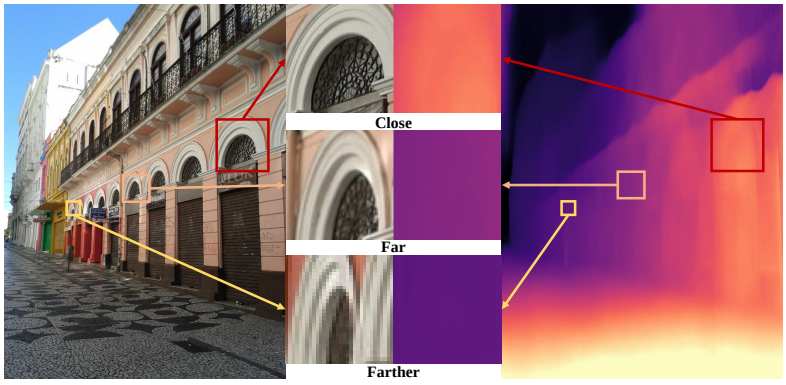


Fig. 2: An example of the image self-similarity and its relation with depth information. We select different patches with different levels of depth (Close, Far, Farther). These patches shares similar textures and the close patch is clearer than the farther patch.

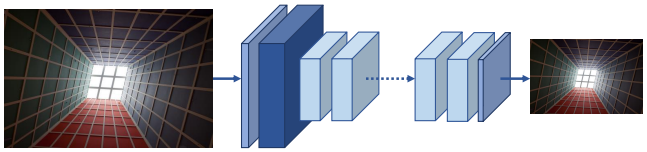
3.1 Depth Guided Training Data Generation

To learn the internal degradation model in an image, we use the depth information to construct a training set from the image. In a nutshell, we treat the distant patches as the low-resolution image patches and the short-distance patches as the high-resolution image patches. Note that we do not use a given degradation kernel as in ZSSR [19], where the image patches need to be downscaled first using a pre-determined degradation operation, i.e., the bicubic. Instead, we only extract the image patches with different depths, hoping to learn a more realistic degradation model from these image patches. Thus, our method could be flexible.

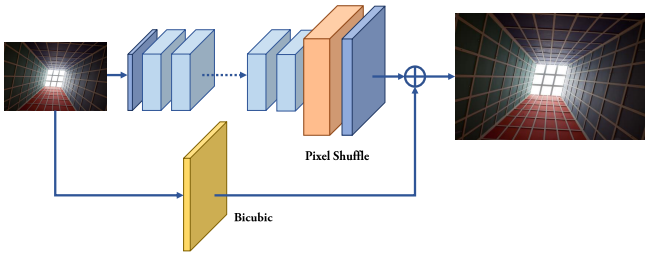
Formally, for a given low-resolution image I, we first convert it from the RGB color space into the YCbCr color space and take the Y, i.e. luminance, channel to calculate the contrast of each patch. Contrast information, reflecting the texture details contained in an image, plays an important clue for building our HR and LR sample patches. Patches with low contrast often contain few image details and thus are useless for training our model. We process the contrast measurement as follows:

$$C = \frac{Y_{max} - Y_{min}}{255}, \tag{1}$$

where C is the range of luminance (i.e. contrast), Y_{max} and Y_{min} denote the 99% and 1% value of the Y channel, respectively. In our work, we set the threshold for C as 0.05; if the range of brightness spans less than the threshold, we define the patch as low contrast and will eliminate it from our training data.



(a) Degradation Simulation Network (DSN)



(b) Super-Resolution Network (SRN)

Fig. 3: The proposed structure of the generators in degradation simulation network (DSN) and super-resolution network (SRN).

Then, we calculate the global mean depth value $\bar{d}$ over the entire depth map as follows:

$$\bar{d} = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} D_{ij}, \tag{2}$$

where H and W denote the height and width of the depth map. Next, we choose a larger image patch (e.g. 64×64 pixels) with a mean depth smaller than $\bar{d}$ as a high-resolution image patch; similarly, we choose a smaller image patch (e.g. 32×32 pixels) with a mean depth bigger than $\bar{d}$ as a low-resolution image patch. As such, we will get a patch-level training collection, (I_{HR}, I_{LR}), from the test image, in which $I_{HR} = \{x_i^h\}$ consists of the high-resolution image patches and $I_{LR} = \{y_i^l\}$ consists of the low-resolution image patches. Notably, (I_{HR}, I_{LR}) are *unpaired* since for each HR patch in I_{HR}, we have no corresponding LR patch in I_{LR}, and vice versa.

3.2 Network Structure

With the depth guided training data (I_{HR}, I_{LR}) extracted from the low-resolution test image I, we train a lightweight and image-specific Super-Resolution Network (SRN), denoted as G^H, for I from scratch. G^H is fully-convolutional and can super-resolve I to $I \uparrow s = G^H(I)$ of arbitrary size, where s is the desired SR scale factor. However, learning G^H from the unpaired data (I_{HR}, I_{LR}) is challenging as the training objective will be highly under-constrained. As suggested in [31], we pair the super-resolution network G^H with another image-specific

Degradation Simulation Network (DSN), denoted as G^L, aiming to learn the internal degradation model—how the high-resolution patches are degraded to low-resolution patches during the imaging process—of a specific image I. We will detail the structures of these two networks in the following.

Degradation simulation network (DSN) The degradation simulation network G^L is lightweight and fully-convolutional, containing five convolutional layers. The structure of G^L is shown in Fig. 3. G^L maps a high-resolution image patch to a low-resolution counterpart. The degradation simulation network G^L in our method indeed learns a specific degradation kernel encoded inside the image, specifying how the imaging process changes the resolutions of patches in the image. Our proposed method will degenerate to models like ZSSR [19] if we use a handcrafted degradation kernel (e.g. a bicubic downscaling) to replace our DSN—a data-driven degradation model. A comparative experiment on this aspect will be detailed in Sec. 5.5. Specifically, G^L is defined as:

$$\tilde{x}_i^l = G^L(x_i^h) = F^{out}((F_5 \cdots F_1(F^{in}(x_i^h))) \downarrow), \tag{3}$$

where x_i^h means a high-resolution image patch, $\tilde{x}_i^l$ is the generated low-resolution patch from x_i^h, and F denotes the convolution layers in the degradation simulation network. F^{in} and F^{out} denote the convolution layers mapping the channels to the desired sizes.

Super-resolution network (SRN) The super-resolution network G^H also uses a lightweight design, in which we stack ten convolutional layers for feature extraction. We apply sub-pixel convolution [18] to upsample the extracted features and predict the high-frequency details. To reduce the computational cost and the number of model parameters, we use the bicubic interpolation for upsampling low-resolution features to generate the low-frequency and blurred HR images. Finally, we apply a global residual learning to merge two branches together to synthesize a visually pleasing high-resolution image as follows:

$$\tilde{y}_i^h = G^H(y_i^l) = F^{out}((F_{10} \cdots F_1(F^{in}(y_i^l))) \uparrow^p) + y_i^l \uparrow^b, \tag{4}$$

where $\tilde{y}_i^h$ is the high-resolution image patch generated by SRN, $\uparrow^p$ and $\uparrow^b$ denote the subpixel shuffle and bicubic interpolation, respectively. Fig. 3 shows the details of the network structure of SRN.

3.3 Bi-cycle training

To learn the degradation simulation network (DSN) and the super-resolution network (SRN) for producing realistic LR and HR image patches respectively, we further equip these two networks with two discriminator networks: D^L for DSN and D^H for SRN. Since the local patches in our depth guided training collection are unpaired, we propose a bi-cycle training strategy to learn these four lightweight networks (G^L, D^L, G^H and D^H) for an image. The concrete learning process contains four steps: (1) our SRN maps the LR patches to the

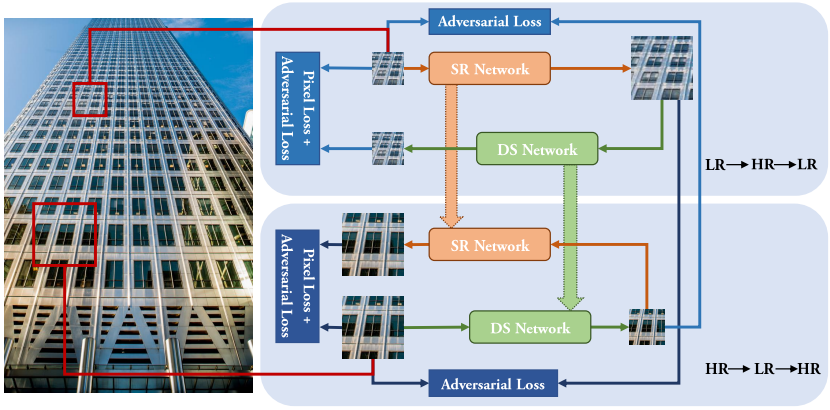


Fig. 4: The proposed structure of bi-cycle training. The first cycle maps LR to HR then back to LR and the second cycle maps HR to LR then back to HR.

fake HR patches, learning to super-resolve the images; (2) the synthesized HR patches are remapped back to their LR patches through DSN; (3) we map the HR image patches to the fake LR counterparts using DSN, simulating the image degradation during the imaging process; and (4) the simulated LR patches is then regenerated back to their HR patches through SRN.

Our bi-cycle training consists of two closed processing cycles: in the first cycle, step (1) and (2), we map the real LR patches to fake HR patches and then remap the synthesized fake HR patches back to LR patches; and in the second cycle, step (3) and (4), we map the real HR patches to fake LR patches and then remap the generated fake LR patches back to HR patches. In each cycle in our model, we consider the adversarial loss to penalize the distribution mismatching and the pixel-wise reconstruction loss of patches as our learning objectives. More concretely, the step (1) optimizes the following Wasserstein GAN [1] objective:

$$L_{GAN}^{SRN} = \mathbb{E}_{y^l}[D^H(G^H(y^l))] - \mathbb{E}_{x^h}[D^H(x^h)], \tag{5}$$

where $y^l \sim I_{LR}$ is the sampled low-resolution image patch and $x^h \sim I_{HR}$ is the sampled high-resolution image patch. In the step (2), we jointly optimize the Wasserstein GAN objective and the cycle-consistent loss based L-1 norm as below:

$$L_{cycle}^{SRN} = \mathbb{E}_{y^l}[D^L(G^L(G^H(y^l)))] - \mathbb{E}_{y^l}[D^L(y^l)] + \mathbb{E}_{y^l}[||G^L(G^H(y^l)) - y^l||_1]. \tag{6}$$

Similarly, in step (3) and (4), we optimize the following two objectives, respectively:

$$L_{GAN}^{DSN} = \mathbb{E}_{x^h}[D^L(G^L(x^h))] - \mathbb{E}_{y^l}[D^L(y^l)]. \tag{7}$$

and

$$L_{cycle}^{DSN} = \mathbb{E}_{x^h}[D^H(G^H(G^L(x^h)))] - \mathbb{E}_{x^h}[D^H(x^h)] + \mathbb{E}_{x^h}[||G^H(G^L(x^h)) - x^h||_1]. \tag{8}$$

After completing the training, we input the entire low-resolution image I into the super-resolution network (SRN) to produce a high-resolution image $G^H(I)$.

4 Discussion

Difference to SelfExSR SelfExSR [8] is a searching based method while our DGDML-SR is learning-based. Although both two methods leverage the internal self-similarity of images, their exact manners are different. SelfExSR searches for similar patches within the image and applies the clear patches to recover the similar but blurred ones. Our DGDML-SR uses the internal patches extracted according to the image depth information to learn the image-specific degradation model and super-resolution model, simultaneously.

Difference to ZSSR Both of ZSSR [19] and our DGDML-SR are zero-shot image super-resolution methods. ZSSR selects patches from the image randomly and uses the bicubic or other pre-determined degradation kernels to downscale the patches. After that ZSSR learns the SR model using the downscaled patches and the original patches; the learned SR model is subsequently used to super-resolve the entire image. Our DGDML-SR does not rely on a pre-determined degradation model; instead we learn the degradation model using a neural network. Thus, our method will not suffer from the deficiency of using a pre-determined degradation kernel. More importantly, even though in the same experimental environment (i.e. similar hyperparameters and paired data generated with bicubic downsampling), our method is still better than ZSSR. For example, on Set5 [3], the PSNR score of our method is 0.21dB higher than ZSSR.

Difference to KernelGAN KernelGAN [2] learns the degradation kernel using a deep linear network. However, their degradation kernel learning is independent of the SR model they used; in other words, KernelGAN is not an end-to-end SR model. KernelGAN needs to learn the degradation kernel first and then use this kernel to generate the HR/LR patches for training a ZSSR network as the final SR model. In comparison, our DGDML-SR is an end-to-end model that could simultaneously learn the degradation kernel and the super-resolution network, using the unpaired image patches from the test image alone.

5 Experiment

In this section, we conduct experiments to evaluate the performance of the proposed zero-shot SR method based on depth guided internal degradation learning. Sec. 5.1 introduces the datasets used in our experiment and also the experimental setup. Then we show the quantitative and qualitative comparisons with the

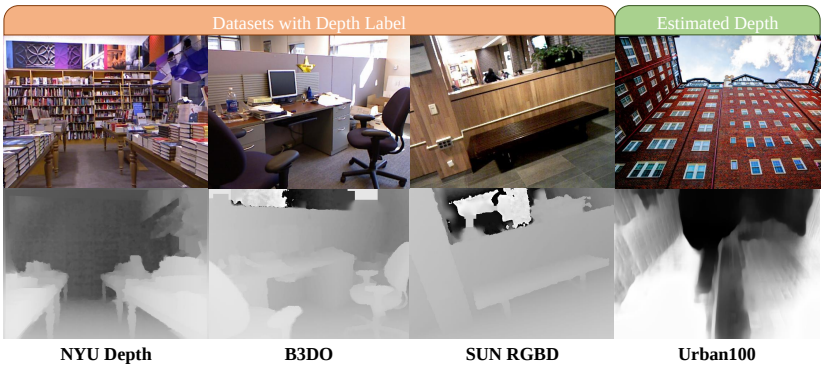


Fig. 5: Examples of images and depth maps from NYU Depth [20], B3DO [9], SUNRGBD [25] and Urban100 [8].

state of the arts in Sec. 5.2 and Sec. 5.3, respectively. In Sec. 5.4, we present two examples of zero-shot super-resolution using the estimated depth information. In the ablation study in Sec. 5.5, we evaluate the performance of our proposed method with and without learning the degradation model.

5.1 Dataset and Training Setup

In the experiments, we select images from NYU depth V2 [20], B3DO [9], Xtion of SUN RGBD [25] and Urban100 [8] dataset. The first three datasets consist of only low-resolution RGB images and low-resolution depth images, while their high-resolution counterparts remain unknown. Among these datasets, NYU has the best image quality and the complete depth information. The image quality of B3DO and SUN RGBD is worse than NYU. In addition to the lower-resolution images and depth maps, they have JPEG compression with an unknown level and more sensor noise. Moreover, their depth information is often incomplete and even incorrect, making these two datasets more challenging. The last dataset only has RGB images without depth labels and we estimate the depth with a pre-trained monocular depth estimation model [7]. On all four datasets, we convert the images from RGB to YCbCr color space and the Y channel is taken out for training and testing. We set the HR (with the bigger region) and LR (with the smaller region) sliding windows on each of the images as mentioned in Sec. 3.1. We use the 64×64 (HR) and 32×32 (LR) sliding windows to extract the image patches for ×2 scaling. For scaling ×4, we use 128×128 (HR) and 32×32 (LR) patch sizes. We then rotate and flip the image patches to augment these patches, so that the number of patches is increased by eight times.

We implement the networks proposed in this paper using PyTorch1.2. We conduct all experiments on one Nvidia RTX2080Ti GPU card. We use RMSprop as the optimizer; the initial learning rate is 0.0001; the batch size is 64; and the learning rate is reduced by 10 times after each iteration of 60 epoch, for a total

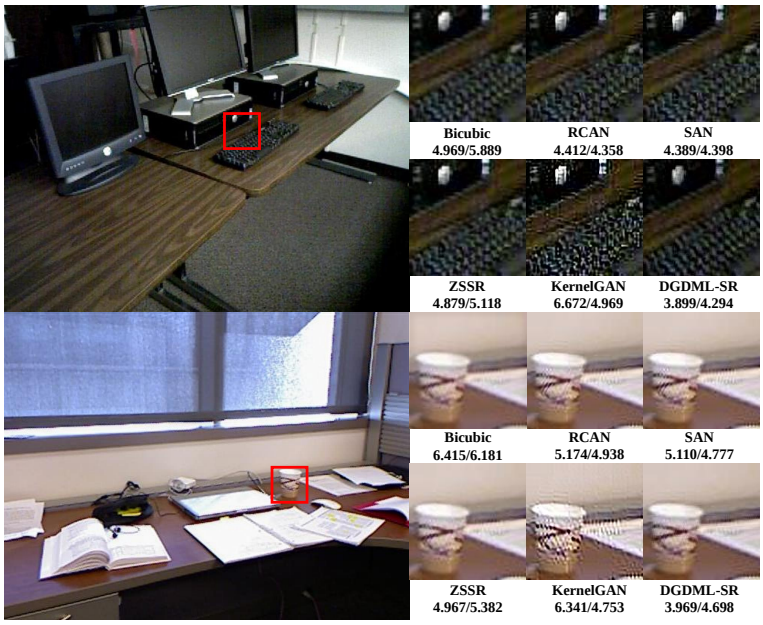


Fig. 6: We compare the visual quality of the super-resolved images from our proposed DGDML-SR with Bicubic, RCAN [30], SAN [5], ZSSR [19], KernelGAN [2]. The NIQE and PI score for these results are shown under each image.

of 150 epochs. In our environment, it takes less than 10 seconds to train a lightweight super-resolution network in each epoch.

5.2 Comparison with the state of the arts

In this section, we compare our method with the state of the arts. In this task we do not have high-resolution labels, thus we use non-reference image quality assessment methods, including NIQE [17] and PI [4] as the comparison metrics. Lower PI and NIQE scores mean better visual quality. In Table. 1, we compare our proposed DGDML-SR with the state of the art zero-shot methods as well as supervised methods including Bicubic, ZSSR [19], RCAN [30], SAN [5] and KernelGAN [2]. The best result is highlighted.

As shown in Table. 1, the recent proposed deep learning based zero-shot methods show great advantages against the supervised methods according to NIQE on the NYU depth dataset. On the other two datasets with worse quality and unknown JPEG compressions, KernelGAN usually generates over-sharped results and amplifies the distracting artifacts. The supervised methods, including RCAN and SAN, overly smooth the details and cannot cope with the compression artifacts either. Among the above methods, our proposed DGDML-SR can achieve almost the best NIQE and PI scores.

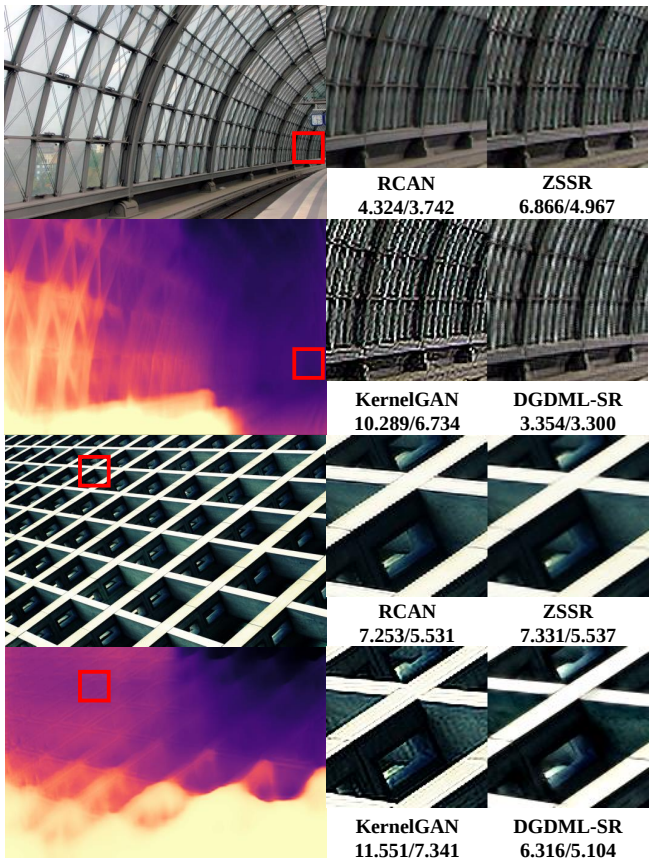


Fig. 7: Quality comparison of RCAN [29], ZSSR [19], KernelGAN [2] and our DGDML-SR on img_002 and img_043 from Urban100 [8] with estimated depth map. NIQE and PI scores are shown under each image.

5.3 Visual Comparison

In this section, we compare the visual quality of the high-resolution images generated by our method with those generated by early developed methods, including Bicubic, ZSSR [19], RCAN [30], SAN [5] and KernelGAN [2]. Fig. 6 shows the zoomed results. Two examples we choose are the img_044 from NYU Depth V2 and the img_0089 from B3DO. The red squares indicate where the patches are taken out. The method name and its NIQE score are shown under the image patch. Similar to the results shown in Sec. 5.2, zero-shot methods can generally generate more details. Due to the linear degradation structure in KernelGAN [2], it cannot handle the multiple degradations and usually generates poor results for the image from B3DO. Among the above methods, our proposed DGDML-SR could generate sharper edges and more high-frequent details with no extra

	Methods	Scale	NYU NIQE/PI	B3DO NIQE/PI	SUN NIQE/PI
Zero-Shot	Bicubic	×2	6.378/6.570	5.786/6.203	5.284/5.931
		×4	8.876/8.086	7.885/6.526	7.542/6.006
	ZSSR [19]	×2	5.753/6.139	5.041/5.360	4.362/5.327
		×4	—/—	—/—	—/—
	KernelGAN [2]	×2	5.620/**4.896**	6.859/**4.751**	6.613/4.847
		×4	6.888/6.591	6.500/5.899	6.457/6.012
	Our DGDML-SR	×2	**4.824**/5.454	**4.281**/4.884	**4.008**/**4.734**
		×4	**6.712**/**6.280**	**5.996**/**5.779**	**5.473**/**5.590**
Supervised	RCAN [29]	×2	5.868/6.126	5.108/4.955	4.813/4.911
		×4	8.387/7.972	6.524/6.404	6.458/6.571
	SAN [5]	×2	6.141/6.258	5.163/5.073	4.713/4.908
		×4	8.399/7.975	6.544/6.516	6.248/6.482

Table 1: Performance comparison of our proposed DGDML-SR with the state of the art zero-shot methods including bicubic, ZSSR [19], KernelGAN [2] and supervised methods including RCAN [29] and SAN [5] in terms of NIQE (lower is better) and PI (lower is better).

high-resolution training datasets, which shows great advantages compared with the state of the arts.

5.4 Super-resolving image with estimated depth

In this section, we use a pre-trained depth estimation model [7] to calculate the depth information for an image without the ground-truth depth label map. Fig. 7 shows the zoom-in details of the high-resolution images (img_002 and img_043 from Urban100 [8]) generated by RCAN [30], ZSSR [19], KernelGAN [2] and our method. The NIQE and PI indexes of each method are shown under the zoom-in image patches. In Fig. 7, the results of RCAN are over-smoothed, losing the high-frequent details. KernelGAN's results are over-sharped while ZSSR's results are blurred at the dense textures. Among these methods, our proposed DGDML-SR can recover sharper edges and more high-frequent details of the image and can also achieve the highest score under the quantitative index.

5.5 Ablation Study

In this section, we first evaluate our depth guided (DG) strategy for collecting the training HR/LR patches. DG generates the unpaired data with the guidance of depth information that helps the network learn a more natural degradation kernel and meanwhile reduce the number of training patches. Without DG, we have to adopt a trivial manner to select the unpaired HR/LR training patches: we randomly select a large region as an HR patch and a small region as an LR patch.

Methods	NYU	B3DO	SUN
None	5.929	5.193	4.926
DG	5.809	4.676	4.687
BCN	5.499	4.328	4.253
DG+BCN	**4.824**	**4.281**	**4.008**

Table 2: Performance comparison with and without the depth guided internal degradation learning (DG) and bi-cycle training (BCN) in terms of NIQE (lower is better).

By ignoring the depth or scale information, this trivial strategy will be highly prone to select a short-distance region as an LR patch and a distant region as an HR patch. Another important aim of DG is to filter out the low-quality local patches such that we can decrease the computational burden significantly. The second important aspect of our method we also evaluate is the bi-cycle training strategy, denoted as BCN. No-BCN means we remove the cycle-consistent loss from our training objectives. BCN, the bi-cycle training strategy, ensures that our training process will be well-constrained. We conduct ablation study on DG and BCN, and show the results in Table. 2. From top to the bottom are the model without DG and BCN, models containing one of the two strategies and the model contains both of them. The results show that the method with DG+BCN has a lower NIQE score, indicating better perceptual quality.

6 Conclusion

In this work, we have proposed a novel zero-shot image super-resolution method, in which we have designed a degradation simulation network (DSN) to learn the internal degradation model from a single image. With the help of DSN, our image-specific super-resolution network can produce satisfactory zero-shot SR results. More specifically, to extract the effective unpaired HR/LR patches from the image, we exploit the depth information to extract the natural HR/LR patches. Our zero-shot SR model can decrease the NIQE score at least 0.912 among the evaluation datasets. Compared with those recently proposed methods with pre-determined degradation kernels, our work can learn a more natural degradation model without relying on extra high-resolution training images and achieve better performance not only in quantitive comparison but also in visual quality on NYU Depth, B3DO, SUN and Urban100 datasets.

Acknowledgement

This work was supported by the NSFC (No. U1713208 and 61876085), Program for Changjiang Scholars and CPSF (No. 2017M621748 and 2019T120430).

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International conference on machine learning. pp. 214–223 (2017)
2. Bell-Kligler, S., Shocher, A., Irani, M.: Blind super-resolution kernel estimation using an internal-gan. arXiv preprint arXiv:1909.06581 (2019)
3. Bevilacqua, M., Roumy, A., Guillemot, C., Alberi-Morel, M.L.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding (2012)
4. Blau, Y., Mechrez, R., Timofte, R., Michaeli, T., Zelnik-Manor, L.: The 2018 pirm challenge on perceptual image super-resolution. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops (September 2018)
5. Dai, T., Cai, J., Zhang, Y., Xia, S.T., Zhang, L.: Second-order attention network for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11065–11074 (2019)
6. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence **38**(2), 295–307 (2015)
7. Godard, C., Aodha, O.M., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3828–3838 (2019)
8. Huang, J.B., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5197–5206 (2015)
9. Janoch, A., Karayev, S., Jia, Y., Barron, J.T., Fritz, M., Saenko, K., Darrell, T.: A category-level 3d object dataset: Putting the kinect to work. In: Consumer depth cameras for computer vision, pp. 141–165. Springer (2013)
10. Kim, J., Kwon Lee, J., Mu Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1646–1654 (2016)
11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
12. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 624–632 (2017)
13. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4681–4690 (2017)
14. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 136–144 (2017)
15. Mechrez, R., Talmi, I., Shama, F., Zelnik-Manor, L.: Maintaining natural image statistics with the contextual loss. In: Asian Conference on Computer Vision. pp. 427–443. Springer (2018)
16. Mechrez, R., Talmi, I., Zelnik-Manor, L.: The contextual loss for image transformation with non-aligned data. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 768–783 (2018)
17. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal Processing Letters **20**(3), 209–212 (2012)

18. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1874–1883 (2016)
19. Shocher, A., Cohen, N., Irani, M.: "zero-shot" super-resolution using deep internal learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3118–3126 (2018)
20. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: European Conference on Computer Vision. pp. 746–760. Springer (2012)
21. Tai, Y., Yang, J., Liu, X.: Image super-resolution via deep recursive residual network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3147–3155 (2017)
22. Tai, Y., Yang, J., Liu, X., Xu, C.: Memnet: A persistent memory network for image restoration. In: Proceedings of the IEEE international conference on computer vision. pp. 4539–4547 (2017)
23. Timofte, R., De Smet, V., Van Gool, L.: A+: Adjusted anchored neighborhood regression for fast super-resolution. In: Asian conference on computer vision. pp. 111–126. Springer (2014)
24. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 0–0 (2018)
25. Xiao, J., Owens, A., Torralba, A.: Sun3d: A database of big spaces reconstructed using sfm and object labels. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1625–1632 (2013)
26. Xu, X., Ma, Y., Sun, W.: Towards real scene super-resolution with raw images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1723–1731 (2019)
27. Yuan, Y., Liu, S., Zhang, J., Zhang, Y., Dong, C., Lin, L.: Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 701–710 (2018)
28. Zhang, K., Zuo, W., Zhang, L.: Learning a single convolutional super-resolution network for multiple degradations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3262–3271 (2018)
29. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 286–301 (2018)
30. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2472–2481 (2018)
31. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)