Degradation Model Learning for Real-World Single Image Super-resolution

Jin Xiao[1], Hongwei Yong[1,2], and Lei Zhang[1,2⋆]

[1] Department of Computing, The Hong Kong Polytechnic University, Hong Kong
[2] DAMO Academy, Alibaba Group, Hangzhou, China
{csjxiao, cshyong, cslzhang}@comp.polyu.edu.hk

Abstract. It is well-known that the single image super-resolution (SISR) models trained on those synthetic datasets, where a low-resolution (LR) image is generated by applying a simple degradation operator (e.g., bicubic downsampling) to its high-resolution (HR) counterpart, have limited generalization capability on real-world LR images, whose degradation process is much more complex. Several real-world SISR datasets have been constructed to reduce this gap; however, their scale is relatively small due to laborious and costly data collection process. To remedy this issue, we propose to learn a realistic degradation model from the existing real-world datasets, and use the learned degradation model to synthesize realistic HR-LR image pairs. Specifically, we learn a group of basis degradation kernels, and simultaneously learn a weight prediction network to predict the pixel-wise spatially variant degradation kernel as the weighted combination of the basis kernels. With the learned degradation model, a large number of realistic HR-LR pairs can be easily generated to train a more robust SISR model. Extensive experiments are performed to quantitatively and qualitatively validate the proposed degradation learning method and its effectiveness in improving the generalization performance of SISR models in practical scenarios.

1 Introduction

Single image super-resolution (SISR) aims to recover a high-resolution (HR) image from its low-resolution (LR) observation, which is a highly valuable technique for improving the resolution and quality of digital photography. As a typical ill-posed inverse problem, SISR has been widely studied during the past decades [1–6], yet it is still a challenging and active research topic. The traditional methods generally utilize powerful image priors [7–12] for SISR, and have made remarkable progresses. However these handcrafted image priors are limited in representing the complex image textures.

Benefitting from the rapid development and great success of deep convolutional neural networks (CNNs) [13], recently SISR has witnessed significant progresses by employing deep CNNs [14–24]. Most of the existing CNN based

⋆ Corresponding author. This work is supported by the Hong Kong RGC GRF grant (PolyU 152216/18E).

SISR models are trained on synthetic HR-LR image pairs, which are generated by applying a simple degradation model (e.g., bicubic downsampling) to the HR images [14, 15, 18, 19, 21, 23, 24]. However, the authentic HR to LR image degradation process is much more complicated than these simple uniform downsample operators. As a result, the SISR networks trained on such synthetic datasets have low generalization capability to real-world LR images, largely limiting their value in practical applications.

Efforts have been made to address the generalization problem of SISR models [16, 25–27]. Zhang *et al.* [16] proposed to use multiple Gaussian kernels together with additive white Gaussian noise to increase the diversity of HR-LR pairs, yet the selection and combination of these kernels are very sensitive. Very recently, researchers have started to construct real-world datasets by using digital cameras to capture images of the same scene under different focal lengths [25–27]. Particularly, Cai *et al.* [27] carefully designed a registration algorithm to obtain pixel-wise aligned HR-LR image pairs. The so-called RealSR dataset enables supervised learning of SISR models, and the learned models demonstrate better performance than previous ones on real-world scenarios. However, constructing such datasets of real-world HR-LR pairs is laborious and costly, and the existing datasets of this kind [25–27] are all limited in number of image pairs, diversity of scenes and illuminating conditions. For example, the RealSR dataset contains only 559 scenes in total, limiting the generalization capability of trained SISR models to a wider range of scenarios.

While constructing real-world datasets of HR-LR image pairs, researchers have also proposed to learn the image degradation process from unpaired HR and LR images, and use the learned degradation model to generate HR-LR image pairs for SISR model learning [28–31]. All these methods employ the Generative Adversarial Network (GAN) [32] to learn the degradation process by differentiating the distribution between generated LR and real LR images. Unfortunately, training such a GAN with unpaired data is very difficult and may not converge to the desired result. Moreover, using a network to model the degradation from HR to LR images makes it hard to interpret the degradation process, ignoring some prior knowledge on the image formation.

In this paper, we model the image degradation process by using spatially variant degradation kernels instead of a network, and propose to learn this model from the HR-LR image pairs in the RealSR dataset instead of the unpaired HR and LR images. It is widely agreed in literature [2, 1, 33, 16, 3] that the LR image formation process can be formulated as first blurring the HR image with a degradation kernel, followed by downsampling and noise addition, while in real scenarios the degradation kernel is spatially variant, relating to the depth and local content in the scene. Clearly, the pixel-wise degradation kernels are the key to model the degradation process. One may propose to learn a network to directly map the HR image to LR image, or propose to learn a network to directly predict the pixel-wise degradation kernel. However, the learning space of those two proposals can be too big for modeling the degradation process, while they ignore the common knowledge of image degradation. Considering the fact

that blurring kernels in an optical imaging system can be generally described as bell-shaped smooth functions [34], we argue that the plausible degradation kernels distribute in a small subspace, which can be approximated as a linear combination of a group of basis kernels. Therefore, we propose to learn a group of basis kernels as well as a weight prediction network to predict the combination coefficients at each pixel.

An end-to-end learning scheme is designed to learn the basis kernels and the weight prediction network from the RealSR dataset [27]. Once learned, our degradation model takes an HR image as input, predicts the spatially variant kernels at each location, and outputs the degraded LR image. In this way, we can easily generate a large amount of realistic HR-LR image pairs using the HR images on hand. Finally, we can train SISR models by using these synthetic yet realistic HR-LR pairs. Experimental results show that the trained SISR models achieve better generalization performance than the models trained only on the RealSR dataset, owing to the enlarged training data of realistic HR-LR image pairs. Our main contributions are summarized as follows:

- We propose to learn the LR image degradation process in a supervised manner from a set of real-world HR-LR image pairs. Specifically, we learn spatially variant degradation kernels by learning a group of basis kernels as well as a pixel-wise weights prediction network.
- By using the learned degradation model to generate realistic HR-LR image pairs, more robust SISR models can be trained, which exhibit higher generalization performance than previous SISR models and produce promising visual quality for real-world LR images.

2 Related work

2.1 Single Image Super-resolution

Single image super-resolution (SISR) is an active topic in low-level vision, and a plenty of works have been proposed in the past decades, including interpolation-based [35], model-based [12, 10] and learning-based methods [14, 15, 17–21, 23, 24]. Traditional methods are usually limited in representing the complex image local structures, while the recently developed deep CNN have shown great advantages in image structure representation and consequently improved much the SISR performance [14, 15, 36, 19, 24, 23]. For example, Kim *et al.* [15] employed the residual learning strategy to design the VDSR model with 20 convolutional layers. Liu *et al.* [19] proposed to utilize contextual information by exploiting the image non-locally correlation. Zhang *et al.* [23] proposed a very deep CNN with over 400 layers, and improved much the SISR performance. Despite the great success, most of the CNN based SISR models are trained on synthetic datasets, where the LR images are generated by applying simple operators such as bicubic downsampling to the HR images [14, 15, 17–21, 23, 24]. Unfortunately, the real-world image degradation process is far more complex than bicubic downsampling. Such a gap between synthetic data and real data makes the trained deep SISR models hardly be generalized to real-world LR images.

2.2 Real-world SISR

To solve the problem of real-world SISR, one intuitive way is to use a more complex degradation process to simulate LR images. Zhang *et al.* [16] proposed to use multiple Gaussian kernels with additive white Gaussian noise to simulate LR images, whereas the selection of suitable kernels is difficult and ad hoc for practical applications. Another recently popular solution is to employ the generative adversarial network (GAN) [32] with unsupervised learning. E.g., SRGAN [22] is proposed to utilize adversarial loss to improve the perceptual quality of images. While the GAN-based methods show some interesting results on SISR, their results are not stable and often exhibit some unnatural visual artifacts.

Instead of simulating HR-LR image pairs, recently efforts have been devoted to construct real-world SISR datasets. Qu *et al.* [37] proposed to use a beam splitter to acquire paired HR-LR images. Kóhler *et al.* [38] used hardware binning on camera sensor to generate LR images. However, these two datasets contain very limited scenes, 31 in [37] and 14 in [38]. Very recently, DSLR cameras have been used to construct real-world SISR datasets by capturing the same scene under different focal lengths. Chen *et al.* [26] collected 100 image pairs of printed postcards. Zhang *et al.* [25] constructed the SR-RGB dataset with 500 scenes, whereas the image pairs are not strictly aligned. To enable pairwise learning, an image registration algorithm is proposed in [27] to carefully handle the misalignment between HR and LR images caused in the data collection process. The so-called RealSR dataset contains a set of aligned real-world HR-LR image pairs, which allow direct pairwise training of SISR models. However, the collection and processing of such a dataset is laborious and costly, and the scale and diversity of RealSR dataset is relatively limited (559 scenes in total).

2.3 Degradation Model Learning

To diminish the domain gap between synthetic and real HR-LR image pairs, another line of work aims to learn the image degradation process and uses it to generate more realistic HR-LR image pairs. Bulat *et al.* [29] proposed to use a generator to learn how to degrade from HR to LR, and a discriminator to distinguish between synthetic LR and real LR images. Manuel *et al.* [28] further improved the generator to learn on image high frequency layers. However, training a GAN is very difficult and may not always converge to the desired result, and the above GAN based degradation learning methods do not exploit the prior knowledge of image formation process in an optical imaging system. In this paper, we model the image degradation process by spatially variant degradation kernels, and propose a supervised learning scheme to learn the degradation model from existing real-world SISR datasets.

3 The Proposed Method

In this section, we first formulate the LR image degradation model based on the real-world LR image formation process. We then present how to learn the pixel-

wise degradation models. Finally, we present how to use the learned degradation models to generate realistic HR-LR datasets for training real-world SISR models.

3.1 Formulation of Image Degradation Model

Denote by $\mathbf{I}^H$ an HR image and by $\mathbf{I}^L$ its LR counterpart. In literature [2, 1, 33, 16, 3], the image degradation from an HR image to an LR image can be generally represented as

$$\mathbf{I}^L = (\mathbf{I}^H * \mathbf{k})\downarrow_d + \mathbf{v}, \tag{1}$$

where "$*$" is the convolution operator, $\mathbf{k}$ is the degradation kernel, $\downarrow_d$ is the downsampling operator, and $\mathbf{v}$ is the random observation noise. The goal of SISR is to recover the underlying HR image $\mathbf{I}^H$ given its LR observation $\mathbf{I}^L$.

Most of existing SISR works [14, 15, 18, 19, 21, 23, 24] assumes that the degradation kernel $\mathbf{k}$ is uniform, i.e., spatially invariant, over the whole image. Particularly, they apply the bicubic kernel to HR images to simulate the HR-LR image pairs, and then use those pairs to train SISR models. Whereas in real-world SISR problems, the degradation kernel is much more complex, correlating with the depth and local content of the scene [27]. Therefore, the degradation kernel is typically non-uniform and spatially variant. At each location (i, j), the kernel may vary, and we use $\mathbf{k}_{i,j}$ to denote the per-pixel degradation kernel. The spatially variant image degradation from HR to LR can be formulated as:

$$\mathbf{I}^L(i, j) = \mathbf{I}^H_{i,j} \odot \mathbf{k}_{i,j} + \mathbf{v}(i, j), \tag{2}$$

where $\mathbf{I}^H_{i,j}$ denotes a local image window centered at (i, j) with the same size as kernel $\mathbf{k}_{i,j}$, and "$\odot$" is the inner product operator.

From Eq. 2, one can see that the key to model the real-world image degradation process is how to predict the pixel-wise degradation kernel $\mathbf{k}_{i,j}$. One intuitive idea is to learn a CNN from the available HR-LR pairs (e.g., the RealSR dataset [27]) to predict the kernel $\mathbf{k}_{i,j}$; however, the learning space of a CNN can be too big for the kernels and the network can be over-fitted by the limited training data. On the other hand, the predicted kernel may have poor interpretability since they may not accord with our prior knowledge on the image degradation process (please refer to our ablation study in Sec. 4.3 for more discussions). It is commonly agreed that the degradation kernels in an optical imaging system can be generally described as bell-shaped smooth functions [34]. This means that the plausible degradation kernels are not arbitrary but actually fall into a small subspace, which can be spanned by a group of basis kernels. Denote by $\mathbf{\Phi} = \{\boldsymbol{\phi}_1, , \boldsymbol{\phi}_M\}$ the set of M basis kernels. We propose to approximate the pixel-wise degradation kernel $\mathbf{k}_{i,j}$ as a weighted combination of $\mathbf{\Phi}$ as follows:

$$\mathbf{k}_{i,j} \approx \sum\nolimits_{m=1}^{M} \mathbf{C}_{i,j}(m)\boldsymbol{\phi}_m, \tag{3}$$

where $\boldsymbol{\phi}_m$ is the m^{th} basis kernel and $\mathbf{C}_{i,j}$ represents the combination weight vector at location (i, j). The above formulation constrains the kernels in a subspace which can be more easily learned, especially when the available training dataset (e.g., RealSR) is not very big.

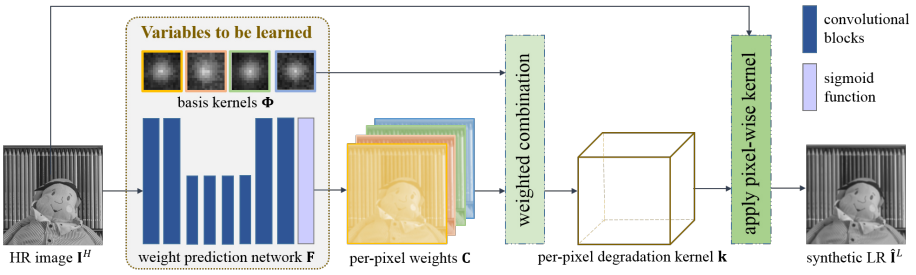

Fig. 1. Overview of the proposed approach for degradation model learning. A group of basis kernels $\mathbf{\Phi}$ are learned together with a weight prediction network $\mathbf{F}$, which are used to generate the pixel-wise degradation kernels. The LR image is obtained by applying the pixel-wise degradation kernels to the HR image.

3.2 Degradation Model Learning

From Eq. (3), one can see that the learning of pixel-wise kernels $\mathbf{k}_{i,j}$ is turned into the learning of basis kernels $\mathbf{\Phi}$ and the weight vectors $\mathbf{C}_{i,j}$. The basis kernels are global to all image regions, while the weights depend on the image local contents. We propose to use a network to predict the weights and learn it simultaneously with the basis kernels from some real-world HR-LR dataset.

Our degradation model learning (DML) approach is illustrated in Fig. 1. With the HR image $\mathbf{I}^H$ as input, a CNN $\mathbf{F}$ with parameters $\mathbf{\Theta}$ is learned to predict the weights, i.e., $\mathbf{C} = \mathbf{F}(\mathbf{I}^H|\mathbf{\Theta})$, where $\mathbf{C}$ is the set of weight vectors $\mathbf{C}_{i,j}$. The basis kernels $\boldsymbol{\phi}_m$ are also learned so that the kernels $\mathbf{k}_{i,j}$ can be predicted according to Eq. (3). The predicted degradation kernels are applied to the HR image $\mathbf{I}^H$ to output the predicted LR image, denoted by $\hat{\mathbf{I}}^L$. Suppose there are N pairs of HR-LR training images, the learning objective can be formulated as

$$\min_{\mathbf{\Phi},\mathbf{\Theta}} \sum_{n=1}^{N} ||\hat{\mathbf{I}}_n^L - \mathbf{I}_n^L||_2^2. \tag{4}$$

We learn the basis kernels $\mathbf{\Phi}$ and weight prediction network $\mathbf{F}$ in an end-to-end manner by using the RealSR dataset [27].

We design the weight prediction network $\mathbf{F}$ following an encoder-decoder structure. It takes an HR image as input and outputs a weight vector at each location. To embrace large receptive field, we use a max pooling layer for feature down-sampling, and employ the bilinear upsampling layer to increase the feature resolution and ensure pixel-wise outputs. Convolutional layer with filters of size 3×3 is used, and ReLU is used as the activation function. To output the per-pixel weights, we use sigmoid function after the last convolutional layer for normalization. The whole network can be easily optimized by the SGD or ADAM optimizer. Examples of the learned kernels, the visualization of the predicted weight maps and more discussions will be provided in the ablation study (see Section 4.3).

3.3 SISR Model Learning

Once the basis kernels $\mathbf{\Phi}$ and the weight prediction network $\mathbf{F}$ are learned by using the DML approach presented in Section 3.2, we can use them to synthesize HR-LR image pairs by using a set of collected HR images as inputs. However, directly using the synthesized LR images to train SISR models is problematic. As described in Eqs. (1) and (2), the real-world LR images are usually corrupted by a certain amount of noise. However, the training objective in Eq. (4) encourages to generate a noise-free LR image since the random noise is hard to predict. If we use the synthesized clean LR images to train the SISR model and then apply the model to real-world noisy LR data, the noise will be exaggerated and lead to unpleasant visual artifacts.

To address this issue and further diminish the gap between synthetic and real LR images, we add random noise to the synthesized LR image $\hat{\mathbf{I}}_n^L$ according to the LR image formulation process described in Eq. (1). Without additional information on the imaging system (e.g., sensors, lens), we simply assume additive white Gaussian noise (AWGN) and empirically set the noise level as $\sigma = 5$.

Finally, we collect a set of high quality images as the HR set, and use the learned degradation model together with AWGN to generate synthetic yet realistic HR-LR image pairs. These image pairs are used to train the SISR model. In this paper, we adopt two representative SISR network architectures, a lightweight network VDSR [15] and a deeper network RCAN [23], to validate the proposed DML method.

4 Experimental Results

4.1 Experiment setup

We carry out both quantitative and qualitative experiments to demonstrate the effectiveness of our proposed DML method for SISR model training. Considering that there are a few issues to be validated and explained, here we summarize how we set up the experiments for a better understanding of our work.

- In Section 4.2, we introduce the training dataset and the testing dataset in our experiments, as well as some implementation details of our algorithm.
- Section 4.3 conducts some ablation studies. First, we discuss the selection of the number of basis kernels in DML. Then we compare our DML with another two potential solutions to synthesize HR-LR pairs. One is to learn a CNN to directly map an HR image to an LR one, and another is to learn a CNN to predict the pixel-wise degradation kernel.
- In Section 4.4 we demonstrate that our DML can result in more robust real-world SISR performance. We first use the RealSR dataset [27], where aligned real-world HR-LR pairs are available so that PSNR/SSIM/LPIPS indices can be computed, to perform quantitative experiments. We then use other real-world data out of the training dataset to perform qualitative experiments, which are to demonstrate that our DML can improve the robustness and generalization performance of real-world SISR models.

4.2 Datasets and implementation details

Datasets. There are three types of datasets required to validate the performance of DML in degradation process learning and SISR model training.

- The first one is the RealSR [27] dataset (version 2), which contains aligned HR-LR image pairs of 559 scenes collected by two cameras with 3 zooming factors: $\times 2$, $\times 3$ and $\times 4$. We follow [27] to split the RealSR dataset into 459 scenes for training and the remaining 100 for testing. We use the training part of this dataset to train our degradation model by the method described in Section 3.2, and use the testing part to quantitatively evaluate the performance of DML and its application to real-world SISR.
- Once the degradation model is learned, we can apply it to an HR image dataset to generate synthetic HR-LR pairs. We construct an HR dataset by combining the Flickr2K dataset [24] and Internet images, containing 3150 images in total. The Flickr2k dataset has 2650 high quality images of various scenes, whose resolution is mostly 1500×2000. To diminish the effect of compression artifacts, we downsample those Flickr2k images by a factor of 2 after Gaussian smoothing (with scale $\sigma = 1$). We also download 500 raw images of 4K resolution from [39], and then apply the PhotoShop CameraRaw tool to them so that uncompressed high quality RGB images of 4K resolution are obtained.
- The third dataset is to validate the effectiveness of DML for real-world SISR. We use the SR-RGB dataset [25] which consists of real-world LR images and their unaligned HR counterparts obtained by optical zoom of DSLR. Since the HR and LR images are not aligned, the PSNR/SSIM/LPIPS measures can not be calculated but the HR images can be used as references for visual comparison.

Implemention Details. We set the size of basis kernels to be learned as 15×15 for all zooming scales $\times 2$, $\times 3$, and $\times 4$. The basis kernels are randomly initialized, and then normalized to have summation 1 for further updating. The weight prediction network is initialized using the Xavier initializer [40]. In the training of both DML and SISR networks, we convert the RGB images to YCbCr color space, and train or test on the Y channel. Images are cropped into 192×192 patches for training of all models. Left-right and up-down flips are used for data augmentation. The Adam optimizer [41] with the default parameter setting ($\beta_1 = 0.9$, $\beta_2 = 0.999$) is used as the optimizer. We train DML and SISR models using fixed learning rate of $1e^{-4}$ for 100K and 300K iterations, respectively. The batch size is set as 16 in DML training. As for SISR models, we adopt two representative network architectures: VDSR [15] and RCAN [23]. We implement RCAN with 100 convolutional layers. The batch size is set as 16 and 2, respectively, for training VDSR and RCAN models.

4.3 Ablation study

We conduct ablation studies to investigate the following two issues of DML: (1) selection of the number of basis kernels in DML; and (2) comparison of DML

Table 1. Evaluation of the quality of generated LR images and super-resolved HR images by using the RealSR [27] dataset. The best and second results are highlighted in red and blue, respectively.

Method	Generated LR						Super-resolved HR					
	×2		×3		×4		×2		×3		×4	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DML (N=4)	37.82	0.9862	36.46	0.9848	35.61	0.9840	33.23	0.9544	30.09	0.9150	28.50	0.8856
DML (N=8)	37.93	0.9864	36.54	0.9850	35.75	0.9842	33.32	0.9552	30.18	0.9157	28.60	0.8864
DML (N=16)	37.90	0.9863	36.51	0.9849	35.73	0.9841	33.28	0.9548	30.16	0.9153	28.58	0.8859
DirectNet	37.70	0.9864	36.33	0.9843	35.50	0.9838	33.13	0.9539	30.01	0.9144	28.42	0.8853
DirectKPN	37.77	0.9863	36.35	0.9844	35.56	0.9836	33.16	0.9545	30.06	0.9147	28.48	0.8860

with the other two potential HR-LR pair synthesis approaches. We train our DML and its variants on the training set (459 image pairs) of RealSR [27], and use the testing set of RealSR to evaluate the quality of generated LR images and the quality of super-resolved HR images. PSNR and SSIM are used as the quantitative metrics.

Number of basis kernels. We first study the suitable number of basis kernels in our DML. By using the training part of the RealSR dataset, we learn N=4, 8, 16 basis kernels and their associated weight predict networks. We then apply the learned models to the HR images in the testing part of the RealSR dataset to generate LR images. By comparing the synthesized and real LR images, we compute and list the PSNR/SSIM results in Table 1. One can see that by increasing the number from N=4 to N=8, better LR generation performances can be achieved, whereas the performance of using N=16 basis kernels is slightly worse than N=8. This means that the underlying degradation process can be well approximated by using N=8 basis kernels.

We visualize the learned 8 basis kernels for different zooming factors in Fig. 2. One can see that with the increase of zooming factor from 2 to 4, the kernels becomes more dispersed and complex, which are in accordance with our common knowledge of image degradation process. We also visualize the basis coefficients predicted by our weight prediction network in Fig. 3. One can see that the learned network can adaptively assign different weights to the kernels according to the scene content and image local structure to generate realistic LR images.

Since our final goal is to improve the SISR performance via DML, it is also necessary to test the effect of N on the final SISR results. We apply the learned DML models to our collected HR image dataset (see Sec. 4.2) to synthesize 3150 HR-LR images pairs, which are then used to train a VDSR super-resolution model. By applying the trained VDSR model to the LR images in the RealSR testing set, we compute the PSNR/SSIM indices of the super-resolved HR images. Table 1 lists the results. One can see that N=8 again achieves the best results for real-world SISR. Therefore, we set N=8 for DML in our experiments.

Comparison with other HR-LR pair synthesis strategies. Besides the proposed DML, there are two other intuitive strategies to synthesize HR-LR image pairs. One is to learn a CNN that directly maps an HR image to an LR one, denoted as DirectNet, and the other is to learn a kernel prediction network

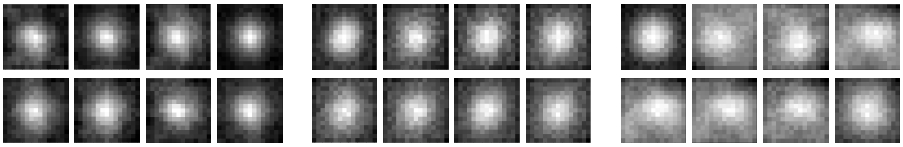

Fig. 2. Visualization of the learned degradation basis kernels by our DML (N=8) model. The left, middle and right 4 columns represent the basis kernels for SR zooming factors $\times 2$, $\times 3$ and $\times 4$, respectively.

Fig. 3. Visualization of the predicted combination weights of the basis kernels by our DML method for zooming factor $\times 2$. The leftmost image is the input HR image, and the right 8 images visualize the predicted weights corresponding to each basis kernels (refer to Fig. 2 for the 8 kernels). The brighter intensity denotes larger weight. One can see that our weight prediction network can adaptively assign different weights according to the scene content and local structures.

[42] to predict the degradation kernel, denoted as DirectKPN. To validate the advantages of our proposed DML method, we implement these two strategies by using the same backbone (with the same hyper-parameters) of the weight prediction network in our DML for fair comparison. For DirectNet, we implement it using the residual learning strategy [15] for better convergence. All the three competitors are trained on the training set of RealSR [27], and tested on the RealSR testing set. PSNR and SSIM are used as quantitative measures.

We first evaluate the performance of the three strategies on LR image generation. The results are listed in Table 1. One can see that DML performs constantly better than DirectNet or DirectKPN on all the three zooming factors, with an improvement of $0.23dB$ and $0.20dB$ in PSNR, respectively. This shows that DML can generate more realistic LR images, owing to our proposed strategy of learning basis kernels and predicting pixel-wise combination weights. Besides, it is observed that DirectKPN performs slightly better than DirectNet. This shows that by taking into account the image degradation process, better LR generation performance can be achieved by learning to predict pixel-wise kernels than directly predicting LR image pixels.

We then evaluate their effectiveness on improving SISR. We apply the three LR image generation models to the collected HR image dataset, synthesizing 3150 HR-LR images pairs by each model. We add small AWGN to those HR-LR pairs (refer to Section 3.3 for details), and train three VDSR models. Finally, we apply these three VDSR models to the LR images in the testing part of the RealSR dataset, and obtain the super-resolved HR images. The PSNR/SSIM results are listed in Table 1. One can see that the VDSR network trained on synthetic HR-LR pairs generated by our DML method, performs constantly better

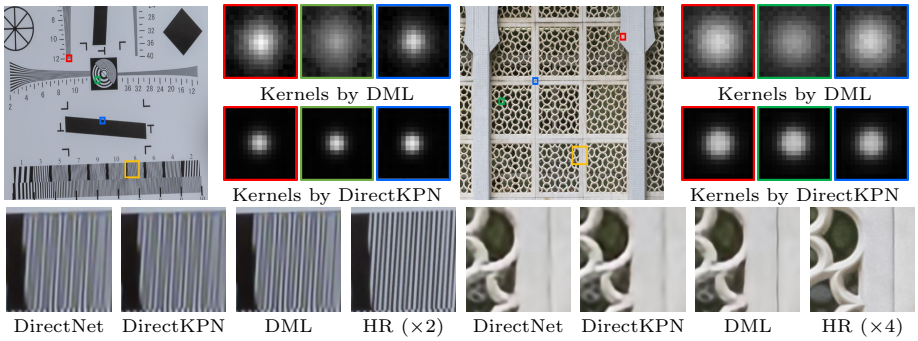

Fig. 4. Visualization of predicted degradation kernels by DML and DirectKPN. One can see that the degradation kernels predicted by DML vary with the image local content, whereas the kernels predicted by DirectKPN are simple and rather uniform across the whole image. We also show the SISR results of the VDSR models trained on the synthetic HR-LR pairs by DML, DirectNet and DirectKPN. One can see that the model based on DML can recover more details with less artifacts.

(around $0.15dB$ in PSNR) than those trained on pairs generated by DirectNet or DirectKPN. This validates the superiority of DML to DirectNet and DirectKPN on improving SISR performance. Our DML method can generate realistic LR images with a smaller gap to real-world LR images, therefore leading to better SISR results than DirectNet and DirectKPN.

We visualize the pixel-wise degradation kernels predicted by our DML and DirectKPN in Fig. 4 (note that DirectNet does not predict kernels). One can see that predicted degradation kernels by DML vary with the image local content, whereas the degradation kernels predicted by DirectKPN are simple and rather uniform across the whole image. This is probably because when we directly learn the pixel-wise degradation kernel, the solution space is too large so that DirectKPN can only converge to a simple solution, resulting in uniform kernels for an input image. In contrast, our DML strategy can effectively reduce the kernel space and thus result in a more robust adaptive degradation kernel prediction model. We also visualize the SISR results by the three degradation models in Fig. 4. It can be seen that our DML based SISR method exhibits better visual quality with more details and less artifacts.

4.4 Experiments on Real-World SISR

As discussed in the Introduction section, the goal of our DML is to synthesize realistic HR-LR image pairs to supplement the limited number of real-world HR-LR pairs so that more robust SISR models can be trained. To validate whether this goal is achieved by our DML method, in this section we use VDSR [15] (20 layers) and RCAN [23] (100 layers) as two representative SISR models to perform extensive SISR experiments. By using the HR image dataset we collected, we synthesized 3150 HR-LR pairs via the learned DML model, and denote this

Table 2. Evaluation of SISR performances on the RealSR [27] dataset by models trained using different training data. The best, second and third results for each SISR network architecture are highlighted in red, blue and yellow, respectively.

SISR model	Training dataset	LPIPS ↓			PNSR ↑			SSIM ↑		
		×2	×3	×4	×2	×3	×4	×2	×3	×4
VDSR	RealSR	0.141	0.224	0.291	33.60	30.53	28.92	0.957	0.919	0.887
	Syn-DSGAN	0.145	0.240	0.309	32.47	29.57	27.20	0.949	0.908	0.851
	Syn-DML	0.137	0.218	0.284	33.32	30.18	28.60	0.955	0.916	0.886
	RealSR+Syn-DSGAN	0.151	0.234	0.289	33.35	30.13	28.56	0.954	0.915	0.885
	RealSR+Syn-DML	0.124	0.198	0.267	33.50	30.37	28.86	0.957	0.918	0.889
RCAN	RealSR	0.141	0.227	0.283	33.91	30.86	29.26	0.960	0.924	0.896
	Syn-DSGAN	0.148	0.239	0.319	32.45	29.78	27.95	0.948	0.916	0.877
	Syn-DML	0.131	0.210	0.265	33.38	30.29	28.66	0.956	0.918	0.887
	RealSR+Syn-DSGAN	0.143	0.230	0.288	33.50	30.56	28.80	0.956	0.920	0.888
	RealSR+Syn-DML	0.123	0.195	0.242	33.73	30.61	28.99	0.958	0.921	0.891

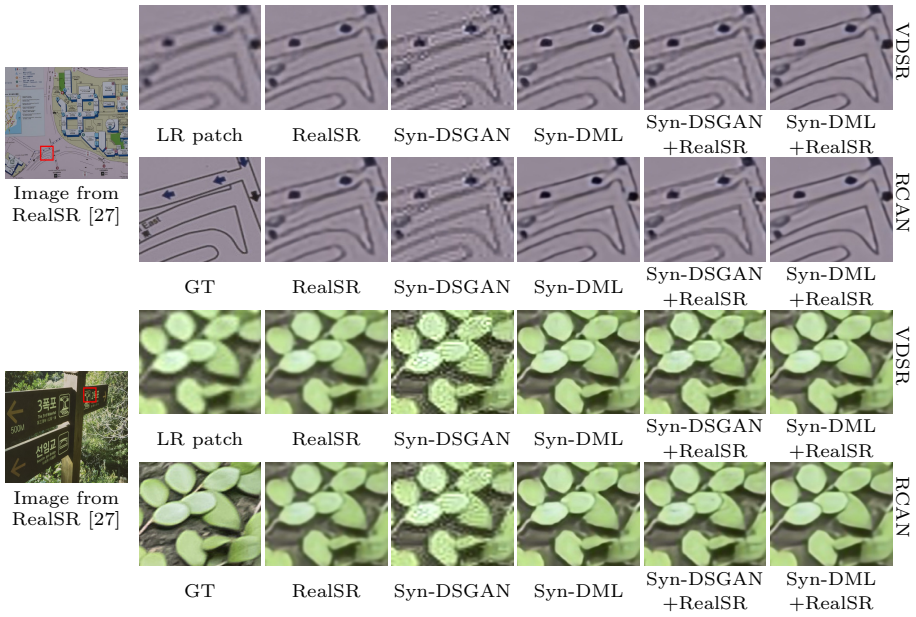

Fig. 5. Visual comparison of the competing SISR models on RealSR [27] dataset with SR scale ×4. The first and second rows of each example are super-resolved patches by VDSR and RCAN networks, respectively, which are trained on different training data.

dataset by Syn-DML. Note that recently a GAN based HR-LR pair synthesis method called DSGAN [28] was developed. We finetuned this model on the RealSR dataset, and applied it to our HR image dataset to synthesize another dataset of HR-LR pairs, denoted by Syn-DSGAN. Therefore, we can train variants of VDSR/RCAN models by using: only RealSR, only Syn-DSGAN, only Syn-DML, the combination of RealSR and DSGAN, and the combination of RealSR and Syn-DML dataset, resulting in a total of 10 SISR models.

We evaluate the 10 VDSR/RCAN models on two real-world datasets. One is the testing set of RealSR [27]. Since the aligned HR-LR pair are available, we can

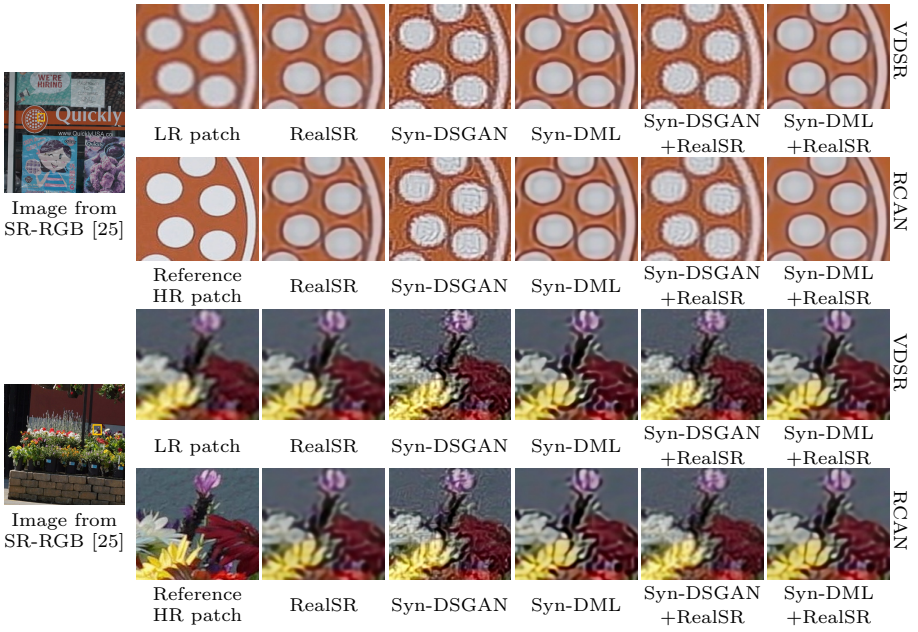

Fig. 6. Qualitative comparison of competing SISR methods on the SR-RGB [25] dataset with SR scale ×4. The first and second rows of each example show the results of VDSR and RCAN models, respectively, trained on different datasets.

compute the PSNR/SSIM/LPIPS indices to perform quantitative evaluation. Another is the SR-RGB dataset [25], which consists of many LR images and their unaligned HR counterparts. Qualitative visual comparisons can be made on it for the different SISR models. Wed like to stress that the testing on the second dataset is more important (though qualitative) because it is independent of the RealSR dataset, part of whose samples are used to train the DML and VDSR/RCAN models. The testing results on the SR-RGB [25] dataset can more faithfully reflect the generalization capability of competing SISR models than those on the RealSR dataset.

Results on the RealSR dataset [27]. We apply the competing VDSR/RCAN models to the testing set of RealSR, and the PSNR/SSIM/LPIPS indices are shown in Table 2. Note that LPIPS is a perceptual index that measures the perceptual quality of images (lower the better). We can have the following findings. First, the VDSR/RCAN models trained on Syn_DML achieve better LPIPS score in all cases than the models trained on RealSR. This validates the effectiveness of our model in improving perceptual quality by synthesizing realistic HR-LR image pairs. Second, the VDSR/RCAN models trained on the Syn-DML dataset achieve comparable but slightly inferior PSNR/SSIM indices to the models trained on RealSR. This is not a surprise because the training and testing data for the latter model are from the same source. Third, SISR models trained on Syn-DML perform significantly better (about 1dB) than those trained on

Syn-DSGAN, which demonstrates the superiority of our DML method to the GAN based DSGAN [28]. Last, by combining RealSR with the synthetic dataset for training, better quantitative results can be achieved than training using only synthetic dataset. Particularly, the VDSR model ($\times 4$) trained on RealSR+Syn-DML achieves even high SSIM scores than the model trained on RealSR.

In Fig. 5, we compare the visual quality of super-resolved HR images by the ten SISR models. One can see that models trained on Syn-DML and RealSR+Syn-DML can effectively recover more image details with more pleasant perceptual quality than the trained using only the RealSR dataset. In particular, the models trained on RealSR+Syn-DML achieve the best visual quality. This validates that our DML method can largely improve the generalization performance of real-world SISR models by synthesizing realistic HR-LR pairs for training.

Results on the SR-RGB dataset [25]. The SR-RGB dataset contains real-world HR and LR images of the same scene, which are however not well aligned. Though it is hard to compute PSNR/SSIM metrics, the HR images in this dataset can be well used a reference for visual comparison of SISR methods. Since the SR-RGB dataset was constructed independently of the RealSR dataset by using different cameras and lens, the results can more fairly demonstrate the generalization capability of an SISR model to real-world scenarios.

In Fig. 6, we visualized the super-resolved HR images on SR-RGB dataset [25] by the ten VDSR/RCAN models trained on different training datasets. One can see that models trained on RealSR dataset can only moderately recover some details. Models trained on Syn-DSGAN produce severe artifacts. Benefitting from the enlarged realistic training data, SISR models trained on Syn-DML can produce visually pleasing results with more fine-grained details. Particularly, the models trained on combined RealSR+Syn-DML deliver the best perceptual quality of super-resolved HR images. The experiments on SR-RGB dataset demonstrate that the SISR models trained by our DML method can be effectively generalized to real-world applications. More visual comparisons can be found in our **supplementary file**.

5 Conclusions

In this paper we proposed to tackle the generalization problem of real-world SISR models by synthesizing realistic HR-LR pairs. To achieve this goal, we first learned an image degradation model from real-world HR-LR image pairs. Specifically, we learned a set of basis degradation kernels together with a weight prediction network. The degradation kernel at any location was estimated as the linear combination of the basis kernels using the weights predicted by the weight prediction network. The learned degradation model was then used to synthesize 3150 HR-LR image pairs covering various scenes for SISR model training. Our extensive analyses and experiments showed that the proposed degradation model learning method can effectively improve the generalization performance of SISR models to real-world applications.

References

1. Yang, C.Y., Ma, C., Yang, M.H.: Single-image super-resolution: A benchmark. In: European Conference on Computer Vision, Springer (2014) 372–386
2. Park, S.C., Park, M.K., Kang, M.G.: Super-resolution image reconstruction: a technical overview. IEEE signal processing magazine **20** (2003) 21–36
3. Yang, W., Zhang, X., Tian, Y., Wang, W., Xue, J.H., Liao, Q.: Deep learning for single image super-resolution: A brief review. IEEE Transactions on Multimedia **21** (2019) 3106–3121
4. Timofte, R., Agustsson, E., Van Gool, L., Yang, M.H., Zhang, L.: Ntire 2017 challenge on single image super-resolution: Methods and results. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. (2017) 114–125
5. Cai, J., Gu, S., Timofte, R., Zhang, L.: Ntire 2019 challenge on real image super-resolution: Methods and results. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2019) 0–0
6. Lugmayr, A., Danelljan, M., Timofte, R., Fritsche, M., Gu, S., Purohit, K., Kandula, P., Suin, M., Rajagoapalan, A., Joon, N.H., et al.: Aim 2019 challenge on real-world image super-resolution: Methods and results. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), IEEE (2019) 3575–3583
7. Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A.: Non-local sparse models for image restoration. In: 2009 IEEE 12th international conference on computer vision, IEEE (2009) 2272–2279
8. Dong, W., Zhang, L., Shi, G., Li, X.: Nonlocally centralized sparse representation for image restoration. IEEE transactions on Image Processing **22** (2012) 1620–1630
9. Wang, S., Zhang, L., Liang, Y.: Nonlocal spectral prior model for low-level vision. In: Asian Conference on Computer Vision, Springer (2012) 231–244
10. Gu, S., Zuo, W., Xie, Q., Meng, D., Feng, X., Zhang, L.: Convolutional sparse coding for image super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 1823–1831
11. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. IEEE transactions on image processing **19** (2010) 2861–2873
12. Dong, W., Zhang, L., Shi, G., Wu, X.: Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. IEEE Transactions on image processing **20** (2011) 1838–1857
13. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016)
14. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: European conference on computer vision, Springer (2014) 184–199
15. 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. (2016) 1646–1654
16. 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. (2018) 3262–3271
17. 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. (2016) 1874–1883

18. 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. (2017) 3147–3155
19. Liu, D., Wen, B., Fan, Y., Loy, C.C., Huang, T.S.: Non-local recurrent network for image restoration. In: Advances in Neural Information Processing Systems. (2018) 1680–1689
20. 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. (2019) 11065–11074
21. 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. (2018) 2472–2481
22. 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. (2017) 4681–4690
23. 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). (2018) 286–301
24. 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. (2017) 136–144
25. Zhang, X., Chen, Q., Ng, R., Koltun, V.: Zoom to learn, learn to zoom. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 3762–3770
26. Chen, C., Xiong, Z., Tian, X., Zha, Z.J., Wu, F.: Camera lens super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 1652–1660
27. Cai, J., Zeng, H., Yong, H., Cao, Z., Zhang, L.: Toward real-world single image super-resolution: A new benchmark and a new model. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 3086–3095
28. Fritsche, M., Gu, S., Timofte, R.: Frequency separation for real-world super-resolution. arXiv preprint arXiv:1911.07850 (2019)
29. Bulat, A., Yang, J., Tzimiropoulos, G.: To learn image super-resolution, use a gan to learn how to do image degradation first. In: Proceedings of the European conference on computer vision (ECCV). (2018) 185–200
30. Han, Z., Dai, E., Jia, X., Chen, S., Xu, C., Liu, J., Tian, Q.: Unsupervised image super-resolution with an indirect supervised path. arXiv preprint arXiv:1910.02593 (2019)
31. Lugmayr, A., Danelljan, M., Timofte, R.: Unsupervised learning for real-world super-resolution. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), IEEE (2019) 3408–3416
32. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. (2014) 2672–2680
33. Romano, Y., Isidoro, J., Milanfar, P.: Raisr: Rapid and accurate image super resolution. IEEE Transactions on Computational Imaging **3** (2016) 110–125
34. Chaudhuri, S.: Super-resolution imaging. Volume 632. Springer Science & Business Media (2001)

35. Zhang, L., Wu, X.: An edge-guided image interpolation algorithm via directional filtering and data fusion. IEEE transactions on Image Processing **15** (2006) 2226–2238
36. 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. (2017) 624–632
37. Qu, C., Luo, D., Monari, E., Schuchert, T., Beyerer, J.: Capturing ground truth super-resolution data. In: 2016 IEEE International Conference on Image Processing (ICIP), IEEE (2016) 2812–2816
38. Köhler, T., Batz, M., Naderi, F., et al.: Bridging the simulated-to-real gap: benchmarking super-resolution on real data. Arxiv: 180906420 [Cs] (2018)
39. wesaturate: Photo sharing. http://www.wesaturate.com (2016)
40. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. (2010) 249–256
41. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
42. Mildenhall, B., Barron, J.T., Chen, J., Sharlet, D., Ng, R., Carroll, R.: Burst denoising with kernel prediction networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 2502–2510