Abstract

We present a simple and effective image super-resolution algorithm that imposes an image formation constraint on the deep neural networks via pixel substitution. The proposed algorithm first uses a deep neural network to estimate intermediate high-resolution images, blurs the intermediate images using known blur kernels, and then substitutes values of the pixels at the un-decimated positions with those of the corresponding pixels from the low-resolution images. The output of the pixel substitution process strictly satisfies the image formation model and is further refined by the same deep neural network in a cascaded manner. The proposed framework is trained in an end-to-end fashion and can work with existing feed-forward deep neural networks for super-resolution and converges fast in practice. Extensive experimental results show that the proposed algorithm performs favorably against state-of-the-art methods.

1. Introduction

Single image super-resolution (SR) aims to estimate a high resolution (HR) image from a low resolution (LR) image. It is a classical image processing problem and has received active research efforts in the vision and graphics community within the last decade. The renewed interest is due to the widely-used high-definition devices in our daily life, such as iPhoneXS (2436×1125), Pixel3 (2960×1440), iPad Pro (2732×2048), SAMSUNG Galaxy note9 (2960×1440), and 4K UHDTV (4096×2160). There is a great need to super-resolve existing LR images so that they can be pleasantly viewed on high-definition devices.

Recently significant progress has been made by using convolutional neural networks (CNNs) in a regression way. For example, numerous methods [3, 10, 11, 22, 15, 14, 4, 35] develop feed-forward networks with advanced network architectures (e.g., residual network [7], attention model [33]) or optimization strategies to learn the LR-to-HR mapping. These methods are efficient and outperform conventional hand-crafted prior-based methods by large margins. However, as the SR problem is highly ill-posed, using feed-forward networks may not be sufficient to estimate the LR-to-HR mapping. In particular, the reconstructed HR images often do not strictly satisfy the image formation model of SR.

To address this issue, several methods improve feed-forward networks with feedback schemes, such as re-implementing iterative back-projection method [9] by deep CNNs [6], using deep CNNs as image priors to constrain the solution space in a variational setting [31], using the image formation model in a feedback step to constrain the training process [18]. However, these algorithms all regenerate LR images from the reconstructed intermediate HR results. The downsampling operation leads to information loss and thus makes these algorithms hard to estimate the details and

Image Formation Model Guided Deep Image Super-Resolution

Jinshan Pan1 Yang Liu2 Deqing Sun3 Jimmy Ren4 Ming-Ming Cheng5 Jian Yang1 Jinhui Tang1
1Nanjing University of Science and Technology 2Dalian University of Technology 3Google Research 4SenseTime Research 5Nankai University

Figure 1. Super-resolution result (×4). Our algorithm uses the image formation of super-resolution to constrain a deep neural network via pixel substitution, which generates the images satisfying the image formation model and better recovers structural details.
structures (e.g., Figure 1(f)).

We note that the LR image is usually assumed to be obtained by a convolution followed by a downsampling process on the HR image. Under this assumption, at the undecimated positions, the LR image should have the same pixel values as the blurred HR image which is obtained by applying a convolution operation to the clear HR image. Thus, we should impose this image formation constraint in the network architecture to generate high-quality images.

However, it is challenging to apply the hard image formation constraint to deep neural networks, because it requires a feedback loop. To this end, we propose a cascaded architecture to efficiently learn the network parameters. The algorithm first generates an intermediate HR image by a deep neural network and then uses the LR image to update the intermediate HR image based on the image formation process. The updated intermediate HR image is further refined by the same deep neural network. Extensive experiments show that the proposed algorithm based on this cascaded manner converges quickly and can generate high-quality images with clear structures.

2. Related Work

We briefly discuss methods most relevant to this work and refer interested readers to [29] for comprehensive reviews.

Dong et al. [3] are the first to develop a CNN method for SR, named as SRCNN. Kim et al. [10] show that the SR-CNN algorithm is less effective at recovering image details and propose a residual learning algorithm using a 20-layer CNN. In [11], Kim et al. introduce a deep recursive convolutional network (DRCN) using recursive-supervision and skip connections. The recursive learning algorithm is further improved by Tai et al. [23], where both global and local learning are used to increase the performance. However, these methods usually upscale LR images to the desired spatial resolution using bicubic interpolation as input to a network, which is less effective for the details restoration as the bicubic interpolation method usually removes details [22].

As a remedy, the sub-pixel convolutional layer [22] or deconvolution layer [4] are developed based on SRCNN. In [13], the Laplacian Pyramid Super-Resolution Network (LapSRN) is proposed to predict sub-band residuals on various scales progressively. Based on the sub-pixel convolutional layer, several algorithms develop the networks with advanced architectures and strategies, e.g., dense skip connection [27, 35], dual-state recurrent models [5], residual channel attention method [33]. These algorithms are effective for super-resolving LR images but usually tend to smooth some structural details. To generate more realistic images, Generative Adversarial Networks (GANs) with both pixel-wise and perceptual loss functions have been used to solve the SR problem [14, 19]. Recent work [2] first uses GANs to generate more realistic training images then trains GANs with the generated training images for SR. Motivated by the generative network in [14], Lim et al. [15] remove some unnecessary non-linear active functions in the generator [14] and propose an Enhanced Deep Super-Resolution (EDSR) network to super-resolve images. However, all these methods directly predict the nonlinear LR-to-HR mapping based on feed-forward networks. They do not explore the domain knowledge of the SR problem and tend to fail at recovering fine image details.

To generate high-quality images that satisfy the image formation constraint, Wang et al. [28] propose a sparse coding network (SCN) based on the sparse representation prior. In [31], Zhang et al. learn a CNN as an image prior to constrain the iterative back-projection algorithm [9]. More recently, the deep neural networks with feedback schemes have been used in SR. Haris et al. [6] improve the conventional iterative back-projection algorithm using CNNs. Pan et al. [18] propose a GAN model with an image formation constraint for image restoration. However, these algorithms need to regenerate low-resolution images in the feedback step which accordingly increase the difficulty for the details and structures restoration. Moreover, the image formation in these methods is used as a soft constraint, which does not directly help the SR results [18]. Using the image formation as a hard constraint is first introduced by Shan et al. [21] in the variational framework. This method [20] uses the pixel substitution to ensure that the generated SR results satisfy the image formation of SR in a hard way. However, it cannot effectively recover the details and structures as only the sparsity of gradient prior is used.

In this work, we revisit the idea of pixel substitution to impose the hard image formation constraint in a deep neural network. The proposed algorithm explores the information from both HR images and LR inputs by a deep neural network in a regression way and is able to generate the results satisfying the image formation model, thus facilitating the high-quality image restoration.

3. Image Formation Process

We first describe the image formation process of the SR problem and then derive the image formation constraint. Given a HR image $I$, the process of generating the LR image $L$ is usually defined as

$$L = \downarrow^s (k \otimes I),$$

where $k$ denotes the blur kernel, $\otimes$ denotes the convolution operator, and $\downarrow^s$ denotes the downsampling operation with a scale factor $s$. Mathematically, this image formation process can be rewritten as

$$L = DKI,$$

where $K$ denotes the filtering matrix corresponding to the blur kernel $k$, $D$ denotes the downsampling operation; $L$ and $I$ denote the vector forms of $L$ and $I$. 

$\text{\copyright}$ 2018 ACM. This is an author-produced, peer-reviewed version of the work published in Proceedings of the ACM SIGGRAPH Asia 2018 Conference on Graphics Interface. This version includes any corrections that were made after the publication of the version in SIGGRAPH Asia 2018.

We note that the LR image is usually assumed to be obtained by a convolution followed by a downsampling process on the HR image. Under this assumption, at the undecimated positions, the LR image should have the same pixel values as the blurred HR image which is obtained by applying a convolution operation to the clear HR image. Thus, we should impose this image formation constraint in the network architecture to generate high-quality images.

However, it is challenging to apply the hard image formation constraint to deep neural networks, because it requires a feedback loop. To this end, we propose a cascaded architecture to efficiently learn the network parameters. The algorithm first generates an intermediate HR image by a deep neural network and then uses the LR image to update the intermediate HR image based on the image formation process. The updated intermediate HR image is further refined by the same deep neural network. Extensive experiments show that the proposed algorithm based on this cascaded manner converges quickly and can generate high-quality images with clear structures.

2. Related Work

We briefly discuss methods most relevant to this work and refer interested readers to [29] for comprehensive reviews.

Dong et al. [3] are the first to develop a CNN method for SR, named as SRCNN. Kim et al. [10] show that the SR-CNN algorithm is less effective at recovering image details and propose a residual learning algorithm using a 20-layer CNN. In [11], Kim et al. introduce a deep recursive convolutional network (DRCN) using recursive-supervision and skip connections. The recursive learning algorithm is further improved by Tai et al. [23], where both global and local learning are used to increase the performance. However, these methods usually upscale LR images to the desired spatial resolution using bicubic interpolation as input to a network, which is less effective for the details restoration as the bicubic interpolation method usually removes details [22].

As a remedy, the sub-pixel convolutional layer [22] or deconvolution layer [4] are developed based on SRCNN. In [13], the Laplacian Pyramid Super-Resolution Network (LapSRN) is proposed to predict sub-band residuals on various scales progressively. Based on the sub-pixel convolutional layer, several algorithms develop the networks with advanced architectures and strategies, e.g., dense skip connection [27, 35], dual-state recurrent models [5], residual channel attention method [33]. These algorithms are effective for super-resolving LR images but usually tend to smooth some structural details. To generate more realistic images, Generative Adversarial Networks (GANs) with both pixel-wise and perceptual loss functions have been used to solve the SR problem [14, 19]. Recent work [2] first uses GANs to generate more realistic training images then trains GANs with the generated training images for SR. Motivated by the generative network in [14], Lim et al. [15] remove some unnecessary non-linear active functions in the generator [14] and propose an Enhanced Deep Super-Resolution (EDSR) network to super-resolve images. However, all these methods directly predict the nonlinear LR-to-HR mapping based on feed-forward networks. They do not explore the domain knowledge of the SR problem and tend to fail at recovering fine image details.

To generate high-quality images that satisfy the image formation constraint, Wang et al. [28] propose a sparse coding network (SCN) based on the sparse representation prior. In [31], Zhang et al. learn a CNN as an image prior to constrain the iterative back-projection algorithm [9]. More recently, the deep neural networks with feedback schemes have been used in SR. Haris et al. [6] improve the conventional iterative back-projection algorithm using CNNs. Pan et al. [18] propose a GAN model with an image formation constraint for image restoration. However, these algorithms need to regenerate low-resolution images in the feedback step which accordingly increase the difficulty for the details and structures restoration. Moreover, the image formation in these methods is used as a soft constraint, which does not directly help the SR results [18]. Using the image formation as a hard constraint is first introduced by Shan et al. [21] in the variational framework. This method [20] uses the pixel substitution to ensure that the generated SR results satisfy the image formation of SR in a hard way. However, it cannot effectively recover the details and structures as only the sparsity of gradient prior is used.

In this work, we revisit the idea of pixel substitution to impose the hard image formation constraint in a deep neural network. The proposed algorithm explores the information from both HR images and LR inputs by a deep neural network in a regression way and is able to generate the results satisfying the image formation model, thus facilitating the high-quality image restoration.

3. Image Formation Process

We first describe the image formation process of the SR problem and then derive the image formation constraint. Given a HR image $I$, the process of generating the LR image $L$ is usually defined as

$$L = \downarrow^s (k \otimes I),$$

where $k$ denotes the blur kernel, $\otimes$ denotes the convolution operator, and $\downarrow^s$ denotes the downsampling operation with a scale factor $s$. Mathematically, this image formation process can be rewritten as

$$L = DKI,$$

where $K$ denotes the filtering matrix corresponding to the blur kernel $k$, $D$ denotes the downsampling operation; $L$ and $I$ denote the vector forms of $L$ and $I$. 

$\text{\copyright}$ 2018 ACM. This is an author-produced, peer-reviewed version of the work published in Proceedings of the ACM SIGGRAPH Asia 2018 Conference on Graphics Interface. This version includes any corrections that were made after the publication of the version in SIGGRAPH Asia 2018.
4. Proposed Algorithm

The analysis above inspires us to use the image formation process to constrain the deep neural networks for SR. Specifically, we first generate an intermediate HR image \( I \) from a LR image \( L \) by a deep neural network. Then we apply the convolution kernel to \( I \) and use pixel substitution (Section 4.2) to enforce the image formation constraint in the feedback step, as shown in Figure 2. In the following, we will explain the details of the proposed algorithm.

4.1. Intermediate HR image estimation

The effectiveness of using deep CNNs to super-resolve images has been extensively validated in SR problems. Our goal here is not to propose a novel network structure but to develop a new framework to constrain the generated SR results using the image formation process. Thus, we can use an existing network architecture, such as EDSR [15], SRCNN [3], and VDSR [10]. In this paper, we use similar network architecture by [15] as our HR image estimation sub-network. Figure 2 shows the proposed network architecture for one stage of the proposed cascaded approach. The parameters of the network are shown in Table 1.

4.2. Pixel substitution

Let \( I \) be the output of the HR image estimation sub-network. If \( I \) is the ground truth HR image, the equality in the SR formation model (3) strictly holds. Thus, to enforce the intermediate HR image \( I \) to be close to the ground truth HR image, we adopt the pixel substitution operation [21]. Specifically, we first obtain the upsampled image \( D^\top L \) by applying the upsampling matrix \( D^\top \) to the LR image \( L \) and blurred intermediate HR image \( B \) and by applying the blur kernel \( K \) to the intermediate HR image \( I \), respectively. Then the output of the pixel substitution operation is

\[
\hat{B}(x) = \begin{cases} 
D^\top L(x), & x = ds, \\
B(x), & \text{otherwise.}
\end{cases}
\]  

(5)

Empirically, we find that the approximation scheme for image formation process converges well, as shown in Figure 8.

4.3. Cascaded training

As the proposed algorithm consists of both intermediate HR image estimation and pixel substitution, we perform these two steps in a cascaded manner. Let \( \Theta_t \) denote the model parameters at stage (iteration) \( t \), and \( \{L^n, I^n_{gt}\}_{n=1}^N \) denote a set of \( N \) training samples. We learn the stage-dependent model parameters \( \Theta_t \) from \( \{L^n, I^n_{gt}\}_{n=1}^N \) by

![Figure 2. An overview of the proposed method. The image formation constraint is enclosed in the dotted red box, which is used to constrain a deep CNN for super-resolution. At each stage, our algorithm first generates an intermediate HR image \( I \) by a deep CNN model and updates the intermediate HR image \( I \) according to the image formation of SR by the pixel substitution (5). The obtained HR image \( \hat{B} \) is then taken as an input for the next stage. The network is solved in a cascaded manner and generates better high-quality images.](image-url)
minimizing the cost function
\[
J(\Theta_t) = \sum_{n=1}^{N} \left\| I^n_t - \hat{I}^n_{gt} \right\|_1,
\]
where \( I^n_t \) is the output of the network at the \( t \)-th stage. Following [15], we use \( L_1 \) norm as the loss function. We minimize (6) to learn the model parameters \( \Theta_t \) stage by stage from \( t = 1, \ldots, T \).

5. Experimental Results

We examine the proposed algorithm using publicly available benchmark datasets and compare it to state-of-the-art single image SR methods.

5.1. Parameter settings and training data

In the learning process, we use the ADAM optimizer [12] with parameters \( \beta_1 = 0.9, \beta_2 = 0.999 \), and \( \epsilon = 10^{-4} \). The minibatch size is set to be 1. The learning rate is initialized to be \( 10^{-4} \). We use a Gaussian kernel in (3) with the same settings used in [21]. We empirically set \( T = 3 \) as a trade-off between accuracy and speed. In the first stage, we use the same upsampling layer as [15] to upsample the features before the Conv3 layer.

For fair comparisons, we first follow standard protocols adopted by existing methods (e.g., [6, 15, 34, 35]) to generate LR images using bicubic downsampling from the DIV2K dataset [25] for training and use the Set5 [1] as the validation test set. Then, we evaluate the effectiveness of our algorithm when LR images are obtained with different image formation models of SR in Section 6. We implement our algorithm based on the PyTorch version of [15]. The code will be made publicly available on the authors' website.

5.2. Comparisons with the state of the art

To evaluate the performance of the proposed algorithm, we compare it against state-of-the-art algorithms including A+ [26], SRCNN [3], FSRCNN [4], VDSR [10], LapSRN [13], MemNet [24], DRCN [11], DRRN [23], EDSR [15], RDN [35], and DBPN [6]. We use the benchmark datasets: Set5 [1], Set14 [30], B100 [16], Urban100 [8], Manga109 [17], and DIV2K (validation set) [25] to evaluate the performance. These datasets contain different image diversities, e.g., the Set5, Set14, and B100 datasets consist of natural scenes; Urban100 mainly

| Algorithms | Scale | Set5 PSNR/SSIM | Set14 PSNR/SSIM | B100 PSNR/SSIM | Urban100 PSNR/SSIM | Manga109 PSNR/SSIM | DIV2K (validation) PSNR/SSIM |
|------------|-------|----------------|----------------|----------------|-------------------|-------------------|--------------------------|
| Bicubic    | ×2    | 35.60/0.9299  | 30.24/0.8688  | 29.36/0.8431  | 26.88/0.8403     | 30.80/0.9339      | 32.45/0.9640             |
| A+ [26]    | ×2    | 36.54/0.9544  | 32.28/0.9056  | 31.21/0.8863  | 29.20/0.8938     | 35.37/0.9680      | 34.56/0.9330             |
| SRCNN [3]  | ×2    | 36.66/0.9542  | 32.45/0.9067  | 31.36/0.8879  | 29.50/0.8946     | 35.60/0.9663      | 34.59/0.9320             |
| FSRCNN [4] | ×2    | 37.05/0.9560  | 32.66/0.9090  | 31.53/0.8920  | 29.88/0.9020     | 36.67/0.9710      | 34.74/0.9340             |
| VDSR [10]  | ×2    | 37.53/0.9590  | 33.05/0.9130  | 31.90/0.8960  | 30.77/0.9140     | 37.22/0.9750      | 35.43/0.9410             |
| LapSRN [13]| ×2    | 37.52/0.9591  | 33.08/0.9130  | 31.08/0.8950  | 30.41/0.9101     | 37.27/0.9740      | 35.31/0.9442             |
| MemNet [24]| ×2    | 37.78/0.9597  | 33.28/0.9142  | 32.08/0.8978  | 31.31/0.9195     | 37.72/0.9740      | NA/NA                    |
| DRCN [11]  | ×2    | 37.63/0.9588  | 33.04/0.9118  | 31.85/0.8942  | 30.75/0.9133     | 37.57/0.9730      | 35.45/0.9400             |
| EDSR [15]  | ×2    | 38.10/0.9669  | 33.69/0.9195  | 32.35/0.9019  | 29.70/0.9358     | 32.77/0.9783      | 36.56/0.9485             |
| RDN [35]   | ×2    | 38.24/0.9614  | 34.01/0.9212  | 32.34/0.9017  | 29.80/0.9353     | 33.19/0.9780      | 36.52/0.9483             |
| DBPN [6]   | ×2    | 38.09/0.9600  | 33.85/0.9190  | 32.77/0.9000  | 32.55/0.9324     | 38.09/0.9775      | 36.37/0.9475             |
| Ours       | ×3    | 38.26/0.9614  | 33.99/0.9205  | 32.37/0.9021  | 33.09/0.9365     | 39.26/0.9783      | 36.60/0.9487             |

For fair comparisons, we first follow standard protocols adopted by existing methods (e.g., [6, 15, 34, 35]) to generate LR images using bicubic downsampling from the DIV2K dataset [25] for training and use the Set5 [1] as the validation test set. Then, we evaluate the effectiveness of our algorithm when LR images are obtained with different image formation models of SR in Section 6. We implement our algorithm based on the PyTorch version of [15]. The code will be made publicly available on the authors’ website.
contains urban scenes with details in different frequency bands; Manga109 is a dataset of Japanese manga; DIV2K (validation set) contains 100 natural images with 2K resolution. We use the PSNR and SSIM to evaluate the quality of each recovered image.

Table 2 summarizes the quantitative results on these benchmark datasets for the upsampling factors of 2, 3, and 4. Overall, the proposed method performs favorably against the state-of-the-art methods.

Figure 3 shows some SR results with a scale factor of 3 by the evaluated methods. The results by the feed-forward models [3, 4, 10, 13, 15, 35] do not recover the structures well. The EDSR algorithm [15] simplifies and improves the network architectures in [14]. However, the structures of the super-resolved images are not sharp (Figure 3(i1) and (i2)). Although the proposed network is based on the network structure of EDSR [15], using pixel substitution to enforce the image formation constraint generates high-quality images.

Figure 4 shows SR results with a scale factor of 4 by the evaluated methods. The recent DBPN algorithm [6] adopts a feedback network to super-resolve images using information from the LR images. However, this method needs to regenerate LR features from intermediate HR features. Consequently, the information at un-decimated pixels would get lost, which makes it hard to estimate the details and structures. The results in Figure 4(j1) and (j2) show that the structures of the images super-resolved by the DBPN method are not recovered well. In contrast, the proposed method recovers finer image details and structures than the state-of-the-art algorithms.

Real examples. We further evaluate our algorithm using real images (Figure 5). Our algorithm generates much clearer images with better detailed structures than those by the state-of-the-art methods [10, 3, 13, 15]. For example, all the four letters in our result are legible, especially “A” (Figure 5(i2)).

6. Analysis and Discussions

We have shown that enforcing the image formation constraint using pixel substitution leads to an algorithm that outperforms state-of-the-art methods. To better understand the proposed algorithm, we perform further analysis, compare it with related methods, and discuss its limitations.

Effectiveness of the image formation constraint. As our cascaded architecture uses a basic SR network several times, one may wonder whether the performance gains merely come from the use of a larger network. To answer this question, we remove the pixel substitution step from our cascaded network architecture for fair comparisons. The
Figure 4. Visual comparisons for $4 \times$ SR on two examples from the Urban100 dataset. The proposed algorithm generates much better results with fine detailed structures.

Figure 5. Results on real examples ($\times 4$). The proposed method recovers much clearer images with better detailed structures.

comparisons in Figure 6(d) and (f) demonstrate the benefit of using the image formation constraint in generating clearer images with finer details and structures. We note that there is little performance improvement by simply cascading a basic SR network several times to increase the network capacity (Figure 6(d)). The results in Table 3 show that using the image formation constraint of SR consistently improves SR results, which further demonstrates the effec-
ever, this method cannot effectively restore the edges and structures estimation of this constraint.

As the proposed network architectures are similar to those used in [15], the proposed algorithm with only one stage would reduce to the feed-forward model [15] to some extent. Both the quantitative evaluations in Table 3 and comparisons in Figure 6(c) show that only using one feed-forward model does not generate high-quality HR images.

We further note that an alternative approach is to add the image formation model (1) to the loss function to constrain the network training instead of using feedback scheme, where the new loss function is defined as

$$\ell_p(I; I_{gt}; L) = \|I - I_{gt}\|_1 + \lambda \|\nabla^s (k \otimes I - L)\|_1,$$

(7)

where $\lambda$ is a weight parameter. We empirically set $\lambda = 0.01$ for fair comparisons in this paper. We quantitatively evaluate the feed-forward network with (7) on the benchmark datasets. Both the quantitative results in Table 3 and visual comparison (Figure 6(e)) demonstrate that adding the image formation loss to the overall loss function does not always improve the performance.

**Closely-related methods.** Several notable methods [6, 21] improve the back-projection algorithm [9] for single image SR. The DBPN algorithm [6] extends the back-projection method [9] using a deep neural network. It needs a down-sampling operation after obtaining intermediate HR images in the feedback stage. As the information at un-decimated pixels of the intermediate HR images may be lost due to the down-sampling operation, DBPN is less effective at recovering details and structures (Figure 7(e)). The method [21] first proposes pixel substitution to enforce the image formation constraint in an iterative optimization scheme. However, this method cannot effectively restore the edges and textures (Figure 7(d)) because only the sparsity of gradient prior is used. In contrast, our algorithm uses pixel substitution to constrain the deep CNN. Both the edges and textures are well recovered (see Figure 7(f)).

We further examine whether the estimated HR images satisfy the image formation constraint. To this end, we apply the image formation to the estimated HR images to generate the LR images and use the PSNR and mean squared error (MSE) as the metrics. The MSE values in Table 4 indicate that the results generated by the proposed method satisfy the image formation model well.

**Robustness to general degradation models of SR.** We have shown that using the image formation constraint can make the deep CNNs more compact thus facilitating the SR problem when the degradation model is approximated by the Bicubic downsampling operation in Section 5. We further evaluate our method on the other degradation models [32, 34]. One degradation model is based on (1), where the blur kernel is Gaussian (denoted as GD). We use this model to generate the LR images using 800 images from DIV2K for training. The size of the Gaussian kernel used for generating LR images ranges from $3 \times 3$ to $17 \times 17$ pixels. Table 5 demonstrates that the proposed algorithm performs favorably against state-of-the-art methods due to the use of the image formation constraint.

We then evaluate the proposed algorithm when the degradation model is approximated by the Bicubic downsampling operation. The image formation plays an important role for the details and structures estimation.

---

**Table 3. Effectiveness of the image formation constraint in SR with the scale factor 2.**

| Dataset     | Set5 PSNR/SSIM | Set14 PSNR/SSIM | B100 PSNR/SSIM | Urban100 PSNR/SSIM | Manga109 PSNR/SSIM |
|-------------|----------------|-----------------|----------------|--------------------|--------------------|
| w/o conv. & (5) | 38.20/0.9612 | 33.96/0.9195 | 32.33/0.9017 | 32.78/0.9347 | 39.07/0.9780 |
| w/ only one stage | 38.19/0.9609 | 33.92/0.9195 | 32.35/0.9019 | 32.97/0.9358 | 39.20/0.9783 |
| Stage 1 | 38.22/0.9612 | 33.84/0.9167 | 32.33/0.9014 | 32.83/0.9351 | 39.00/0.9777 |
| Stage 2 | 38.27/0.9614 | 33.98/0.9201 | 32.37/0.9021 | 33.04/0.9326 | 39.26/0.9784 |
| Stage 3 | 38.26/0.9614 | 33.99/0.9205 | 32.37/0.9021 | 33.09/0.9365 | 39.26/0.9783 |

---

**Table 4. Evaluations of the regenerated LR images (×2).**

| Dataset     | Set5 PSNR/MSE | B100 PSNR/MSE | Urban100 PSNR/MSE | Manga109 PSNR/MSE |
|-------------|---------------|---------------|-------------------|-------------------|
| DBPN [6]  | 61.60/0.0362 | 60.07/0.0704 | 59.00/0.0944 | 60.41/0.0658 |
| Ours  | 72.06/0.0045 | 66.29/0.0264 | 65.04/0.0360 | 67.56/0.0189 |

---

Figure 6. Effectiveness of the image formation constraint (×4). The image formation plays an important role for the details and structures estimation.

Figure 7. Comparisons of the results by different back-projection methods (×4). The DBPN method [6] based on the iterative back-projection algorithm [9] is less effective for the edges restoration due to the additional downsampling operation.
Table 5. Comparisons of the results (×2) by different methods with the GD model.

| Dataset     | Set5     | B100     | Urban100  | Manga109  |
|-------------|----------|----------|-----------|-----------|
| EDSR [15]   | 32.26/0.9218 | 29.05/0.8444 | 25.96/0.8457 | 28.74/0.9274 |
| RDN [35]    | 32.21/0.9212 | 29.08/0.8445 | 25.91/0.8455 | 28.79/0.9275 |
| RCAN [34]   | 32.30/0.9219 | 29.12/0.8446 | 26.16/0.8459 | 28.87/0.9280 |
| Ours        | 32.38/0.9223 | 29.18/0.8450 | 26.20/0.8462 | 28.90/0.9282 |

Table 6. Results (×2) on “Set5” with noisy input images.

| Noise level | 1%  | 2%  | 3%  | 4%  |
|-------------|-----|-----|-----|-----|
| EDSR [15]   | 38.19/0.9610 | 35.96/0.9387 | 35.03/0.9272 | 34.30/0.9179 |
| RDN [35]    | 38.16/0.9609 | 35.69/0.9382 | 35.02/0.9272 | 34.20/0.9176 |
| Ours        | 38.21/0.9611 | 36.01/0.9389 | 35.06/0.9275 | 34.34/0.9184 |

Figure 8. Quantitative evaluation of the convergence property on the super-resolution dataset (Set5, ×2).

downsampling with noise. To generate LR images for training, we add Gaussian noise to each LR image used in Section 5.1, where the noise level ranges from 0 to 10%. Table 6 shows that our algorithm is robust to image noise due to the cascaded optimization method.

All above results on both synthetic and real-world images demonstrate that the proposed algorithm can generalize well even though the image formation constraint is based on known blur kernels.

Convergence property. To quantitatively evaluate the convergence properties of our algorithm, we evaluate our method on the benchmark dataset Set5. Figure 8 shows that the network converges after 250 epochs in terms of the average PSNR values. We further note that using 2-stage cascaded model would generate better results and using more stages does not significantly improve the performance.

Figure 9 shows some intermediate HR images from the proposed method. We note that the structural details are better recovered with more stages. This further demonstrates that using the image formation constraint in a deep CNN helps the restoration of the structural details.

Running time performance. As our algorithm uses a cascaded architecture, it increases the computation. We examine the running time of the proposed algorithm and compare it with state-of-the-art methods on the Set5 dataset, as shown in Table 7. The proposed algorithm takes slightly more running time compared with the feed-forward models, e.g., [10, 15]. The proposed algorithm is about 3 times faster than the feedback DBPN method [6].

Limitations. As our algorithm uses the known image formation of SR to approximate the unknown degradation model of SR, it is less effective when this approximation does not hold. Figure 10 shows an example with significant JPEG compression artifacts, where the image formation model of SR does not approximate the degradation caused by the image compression well. Our algorithm exacerbates the compression artifacts, while the results by the feed-forward models have few artifacts. Building the compression process into the network architecture is likely to reduce these artifacts.

7. Concluding Remarks

We have introduced a simple and effective super-resolution algorithm that exploits the image formation constraint. The proposed algorithm first uses a deep CNN to estimate an intermediate HR image and then uses pixel substitution to enforce the intermediate HR image satisfy the image formation model at the un-decimated pixel positions. Our cascaded architecture can be applied to existing feed-forward super-resolution networks. Both quantitative and
qualitative results show that the proposed algorithm performs favorably against state-of-the-art methods.

References

[1] M. Bevilacqua, A. Roumy, C. Guillemot, and M. Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In BMVC, pages 1–10, 2012. 4

[2] A. Bulat, J. Yang, and G. Tzimiropoulos. To learn image super-resolution, use a GAN to learn how to do image degradation first. In ECCV, pages 187–202, 2018. 2

[3] C. Dong, C. C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. In ECCV, pages 184–199, 2014. 1, 2, 3, 4, 5, 6

[4] C. Dong, C. C. Loy, and X. Tang. Accelerating the super-resolution convolutional neural network. In ECCV, pages 391–407, 2016. 1, 2, 4, 5

[5] W. Han, S. Chang, D. Liu, M. Yu, M. Witbrock, and T. S. Huang. Image super-resolution via dual-state recurrent networks. In CVPR, pages 1654–1663, 2018. 2

[6] M. Haris, G. Shakhnarovich, and N. Ukita. Deep back-projection networks for super-resolution. In CVPR, pages 1664–1673, 2018. 1, 2, 4, 5, 6, 7, 8

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, pages 770–778, 2016. 1, 3

[8] J.-B. Huang, A. Singh, and N. Ahuja. Single image super-resolution from transformed self-exemplars. In CVPR, pages 5197–5206, 2015. 4

[9] M. Irani and S. Peleg. Improving resolution by image registration. CVGIP: Graphical Model and Image Processing, 53(3):231–239, 1991. 1, 2, 7

[10] J. Kim, J. K. Lee, and K. M. Lee. Accurate image super-resolution using very deep convolutional networks. In CVPR, pages 1646–1654, 2016. 1, 2, 3, 4, 5, 6, 8

[11] J. Kim, J. K. Lee, and K. M. Lee. Deeply-recursively convolutional network for image super-resolution. In CVPR, pages 1637–1645, 2016. 1, 2, 4, 6

[12] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. CoRR, abs/1412.6980, 2014. 4

[13] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In CVPR, pages 624–632, 2017. 2, 4, 5, 6

[14] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In CVPR, pages 105–114, 2017. 1, 2, 5

[15] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee. Enhanced deep residual networks for single image super-resolution. In CVPR, pages 1132–1140, 2017. 1, 2, 3, 4, 5, 6, 7, 8

[16] D. R. Martin, C. C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In ICCV, pages 416–425, 2001. 4

[17] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamazaki, and K. Aizawa. Sketch-based manga retrieval using manga109 dataset. Multimedia Tools Appl., 76(20):21811–21838, 2017. 4

[18] J. Pan, Y. Liu, J. Dong, J. Zhang, J. S. J. Ren, J. Tang, Y.-W. Tai, and M.-H. Yang. Physics-based generative adversarial models for image restoration and beyond. CoRR, abs/1808.00605, 2018. 1, 2

[19] M. S. M. Sajjadi, B. Scholkopf, and M. Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In ICCV, pages 4491–4500, 2017. 2

[20] Q. Shan, J. Jia, and A. Agarwala. High-quality motion de-blurring from a single image. ACM TOG, 27(3):73:1–73:10, 2008. 2

[21] Q. Shan, Z. Li, J. Jia, and C. Tang. Fast image/video upsampling. ACM TOG, 27(5):153:1–153:7, 2008. 2, 3, 4, 7

[22] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In CVPR, pages 1874–1883, 2016. 1, 2

[23] Y. Tai, J. Yang, and X. Liu. Image super-resolution via deep recursive residual network. In CVPR, pages 3147–3155, 2017. 2, 4, 6

[24] Y. Tai, J. Yang, X. Liu, and C. Xu. Memnet: A persistent memory network for image restoration. In ICCV, pages 4539–4547, 2017. 4

[25] R. Timofte, E. Agustsson, L. V. Gool, M. Yang, L. Zhang, and et al. NTIRE 2017 challenge on single image super-resolution: Methods and results. In CVPR Workshops, pages 1110–1121, 2017. 4

[26] R. Timofte, V. D. Smet, and L. V. Gool. A+: adjusted anchored neighborhood regression for fast super-resolution. In ACCV, pages 111–126, 2014. 4, 5, 6

[27] T. Tong, G. Li, X. Liu, and Q. Gao. Image super-resolution using dense skip connections. In CVPR, pages 4799–4807, 2017. 2

[28] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang. Deep networks for image super-resolution with sparse prior. In CVPR, pages 370–378, 2015. 2

[29] C.-Y. Yang, C. Ma, and M. Yang. Single-image super-resolution: A benchmark. In ECCV, pages 372–386, 2014. 2

[30] R. Zeyde, M. Elad, and M. Protter. On single image scale-up using sparse-representations. In The 7th International Conference on Curves and Surfaces, pages 711–730, 2010. 4

[31] K. Zhang, W. Zuo, S. Gu, and L. Zhang. Learning deep CNN denoiser prior for image restoration. In CVPR, pages 2808–2817, 2017. 1, 2

[32] K. Zhang, W. Zuo, and L. Zhang. Learning a single convolutional super-resolution network for multiple degradations. In CVPR, pages 3262–3271, 2018. 7

[33] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu. Image super-resolution using very deep residual channel attention networks. In ECCV, pages 286–301, 2018. 1, 2

[34] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu. Image super-resolution using very deep residual channel attention networks. In ECCV, pages 294–310, 2018. 4, 7, 8

[35] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu. Residual dense network for image super-resolution. In CVPR, pages 2472–2481, 2018. 1, 2, 4, 5, 6, 8