# MULTIPLE-IMAGE SUPER RESOLUTION USING BOTH RECONSTRUCTION OPTIMIZATION AND DEEP NEURAL NETWORK

*Jie Wu, Tao Yue, Qiu Shen, Xun Cao, Zhan Ma*

School of Electronic Science and Engineering, Nanjing University

## ABSTRACT

We present an efficient multi-image super resolution (MISR) method. Our solution consists of a $L1$-norm optimized reconstruction scheme for super resolution (SR), and a three-layer convolutional network for artifacts removal, in a concatenated fashion. Such a two-stage method achieves excellent performance, which outperforms the existing state-of-the-art SR methods in both subjective and objective measurements (e.g., 5 to 7 dB improvements on popular image database using PSNR metric).

***Index Terms*—** Super-resolution, Multi-image, Reconstruction, Convolutional network

## 1. INTRODUCTION

Image super-resolution (SR) aims at recovering a high resolution image with more details from a single (single-frame SR) or a series of low resolution images (multi-frame SR). There are two main categories of existing SR algorithms, i.e., reconstruction-based methods and learning-based methods.

**Reconstruction-based methods.** Reconstruction-based algorithms try to mimic the inverse process of down-sampling to reconstruct the high resolution images from series of slightly different observations. However, the reconstruction-based SR method suffers from the ill-posedness due to the loss of high frequency components. This issue can be tackled to a certain degree by introducing some regularizers into the objective function, like $L1$- or $L2$-Norms. Theoretically, adding $L1$-norm/$L2$-norm regularizer is equivalent to adding the Laplacian/Gaussian distribution prior information. These kinds of methods have been explored for decades. For instance, [1] recovered the high-resolution counterpart particularly based on the locally linear embedding, and [2] found the connection between soft edge smoothness and a soft cut metric on an image grid. However, these distribution based regularizers cannot resolve the ill-posed problem completely, especially for the cases of high enlarge factors.

**Learning-based methods.** The learning-based algorithms learn a mapping from low resolution images to higher ones from either internal information [4, 5, 6] (i.e., exploiting internal similarities of the same image) or external information [7, 8, 9, 10, 11] (i.e., learning mapping functions from external low and high-resolution exemplar pairs). Recently, the deep learning based methods achieve impressive performance for SR. A three-layer convolutional network that takes the low-resolution image as the input and outputs the high-resolution one is proposed in [12]. And [13] presented a deep network to generate a high-resolution image inspired by VGG-net used for ImageNet classification [14]. A twenty-layer deep network was applied to learn the residuals for decent performance. These learning based methods directly guess the high resolution details according to the low resolution input and the learned mapping functions, which may cause the incorrect results (different from the real cases), although these recovered images are of good visual quality.

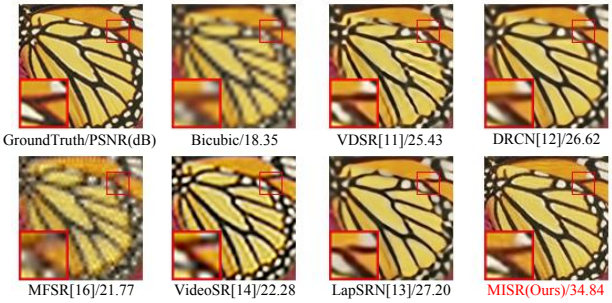


**Fig. 1**. 4x SR for "butterfly" of popular SR algorithms. Our MISR provides 7 dB PSNR improvement over the latest LapSRN [3].

In this paper, we propose an innovative multi-image super-resolution method (MISR), cascading a reconstruction-based SR and a three-layer deep neural network (DNN) based artifacts removal filter. Specifically, the reconstruction-based method takes multiple images with sub-pixel offsets as input and outputs one high-resolution image. Then, a three-layer convolutional neural network is applied to remove the artifacts caused by the ill-posedness of reconstruction problem and to further sharpen the edges. In the proposed algorithm, we use the $L1$-norm regularization term to constrain the reconstruction process, and the conjugate-gradient algorithm is used for fast convergence. Ringing artifacts are unavoidable

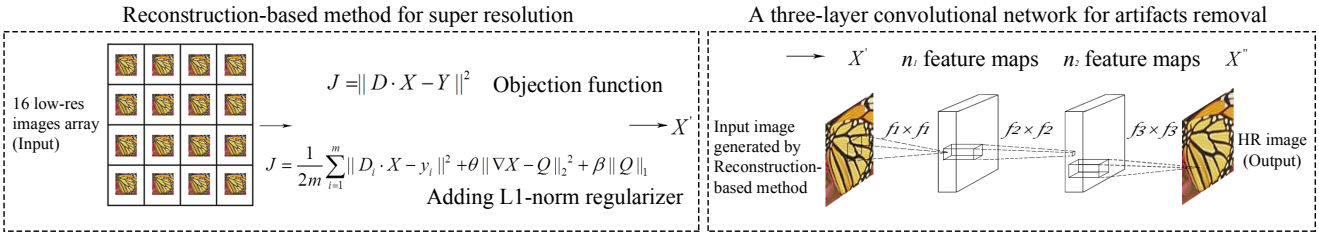


**Fig. 2**. The structure of our proposed MISR method: the left module is the reconstruction-based SR and the right is a three-layer convolutional network for artifacts removal.

for these reconstructed high resolution images because of the ill-posedness of the super resolution. We devise a neural network with three-layer convolutional layers to suppress the ringing artifacts and recover the sharp edges. The proposed method benefits from both the multiple low resolution inputs and information learned from the three-layer neural network, thus it achieves superior accuracy which outperforms the state-of-the-art SR approaches, as shown in Fig. 1.

The rest of the paper is organized as follows. Section 2 details the implementation of our proposed MISR algorithm followed by the experiments conducted in Section 3 to demonstrate the efficiency of MISR in comparison to other state-of-the-art methods. Finally, concluding remarks is given in Section 4.

## 2. TWO-STAGE MULTIPLE IMAGE SUPER RESOLUTION

We propose a novel multi-image super resolution framework, which takes both the sub-pixel measurements and exemplar information into consideration. The framework concatenates two stages namely the reconstruction optimization and a three-layer convolutional neural network. The complete pipeline of our approach is illustrated in Fig. 2, where we take 16 low-resolution images to reconstruct one high-resolution image denoted as $X^{'}$, and then a three layer DNN is applied upon $X^{'}$ to have $X^{''}$ with better quality in both subjective and objective measurements. Even though we exemplified our MISR using $4\times$ SR scaling factor, it could be directly extended to other SR scaling factors.

### 2.1. Reconstruction-Based SR using Multiple Images

We denote the high-resolution image as $X \in \mathbb{R}^{M_s \times N_s}$, the low-resolution images as $\{Y_i\}_{i=1}^{s^2} \in \mathbb{R}^{M \times N}$, and $\{D_i\}_{i=1}^{s^2} \in \mathbb{R}^{MN \times MNs^2}$ as the down-sampling matrix, where we assume the size of the low-resolution image $Y_i$ is $M \times N$, and the scaling factor is $s$. Thus, the high-resolution image is $M_s \times N_s$. Their relationship can be described as,

$$[Y_1, Y_2, ..., Y_{s^2}]^T = [D_1, D_2, ..., D_{s^2}]^T \cdot X. \quad (1)$$

Our goal is to recover the high-resolution image $X$ from the low resolution observations $\{Y_i\}_{i=1}^{s^2}$. Considering the ill-posedness, we translate it into an optimization problem with $L1$-norm regularization, i.e.,

$$X = \arg\min_X ||D \cdot X - Y||^2 + \beta ||\nabla X||_1, \quad (2)$$

where $X$ is the high-resolution image, $\{Y_i\}_{i=1}^{s^2}$ are sixteen low-resolution images, $\{D_i\}_{i=1}^{s^2}$ is the counterpart down-sampling matrices from $X$ to $Y_i$, $\nabla$ is the gradient operator, and $\beta$ is the weight of the regularization term. In this paper, we empirically set $\beta = 0.1$. To solve Eq.(2), an auxiliary variable $Q$ is introduced,

$$J = \frac{1}{2s^2}\sum_{i=1}^{s^2} ||D_i \cdot X - Y_i||^2 + \theta ||\nabla X - Q||_2{}^2 + \beta ||Q||_1 \quad (3)$$

where $\theta$ is the weight of auxiliary term and varies for each iteration to accelerate the convergence. In our experiment, $\theta$ is set to be 0.001 initially, and times 0.99 after each iteration. Eq.(3) can be easily solved by the two-step iterative optimization method, where the specific details are described as follows:

*X-step:* In this step, we only optimize the terms with variable $X$, and leave $Q$ fixed. Then the objective function becomes,

$$J_1 = \frac{1}{2s^2}\sum_{i=1}^{s^2} ||D_i \cdot X - Y_i||^2 + \theta ||\nabla X - Q||_2{}^2 \quad (4)$$

We use Conjugate Gradient method to quickly find the optimal value. The reason for choosing Conjugate gradient method is that it only uses first derivative information, and is capable of overcoming the drawback of the steepest descent method which converges very slowly. And the most important thing is that conjugate gradient does not require any external parameters, but offers fast and stable convergence rate.

*Q-step:* In this step, we focus on the optimization of $Q$,

$$J_2 = \theta ||\nabla X - Q||_2^2 + \beta ||Q||_1. \quad (5)$$

Here we take $X$ from aforementioned X-step, into Eq.(5). It then can be solved by using a simple shrinkage operation (see [16] for details). By iteratively optimizing $X$ and $Q$, we can get the optimal solution of Eq.(2) fast and stable, as shown in Alg. 1.

Alg. 1: L1-norm reconstruction algorithm

**Inputs:**

1. $S = 4$ -- Upscaling factor.
2. $\{Y_i\}_{i=1}^{s^2}$ -- Sixteen arrayed LR images.
3. $\theta = 0.001$ -- the weight of auxiliary term.
4. $\beta = 0.1$ -- the weight of the regularization term.
5. $ep = 1e-6$ -- the convergence threshold.
6. $\{D_i\}_{i=1}^{s^2}$ -- The down-sampling matrix.

**Outputs:**

1. $X^{'}$ -- The reconstructed image.

**Process:**

Compute $X_0$, an initial interpolated version of $Y_1$

**While** $norm(X_{k+1} - X_k, \inf) < ep$ **and** $i \le 5$ **do**

- $Q$ step: Compute $Q_i$ by shrinkage method
- Update $\theta = \theta \times 0.99$
- $X$ step: Compute $X_i$ by Conjugate Gradient method

**For** $k = 1$ **to** $5$ **do**

$r_0 := Y - DX_0$
$p_0 := r_0$
$k := 0$
**repeat**
$\alpha_k := \frac{r_k^T r_k}{p_k^T D p_k}$
$X_{k+1} := X_k + \alpha_k p_k$
$r_{k+1} := r_k - \alpha_k D p_k$
$\beta_k := \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}$
$p_{k+1} := r_{k+1} + \beta_k p_k$
$k := k + 1$
**End repeat**
**end**

**end**

$X^{'} = X_{k+1}$

### 2.2. Artifacts Removal and Enhancement Using DNN

It is unavoidable that the reconstruction images $X^{'}$ generated by first module may have some artifacts like ringing effects. We then introduce a three-layer full-convolutional neural network to suppress these artifacts to improve the final image quality. The configuration is outlined in the right part of Fig. 2. For better understanding, we present a detailed DNN structure in Fig. 3. The artifacts removal DNN has three layers, except the first layer, the other two layers are all convolutional layers with 128 feature maps and the filters size are $9 \times 9$, $5 \times 5$ separately. The first layer operates on the input image. The last layer is used for image construction with filter size of $1 \times 7 \times 7$. The batch size for training is 128. We shuffle the train data to avoid the effects of image contents. We apply the Rectified Linear Unit ($ReLU$, max(0,x)) to add the nonlinear mapping. We use Adam optimizer and set our training epoch number as 100 (76029 iterations). The whole training process takes roughly one hour on GPU Tesla P100-PCIE-16GB.

The network takes the reconstruction results, denoted as $X^{'}$, as the input. Given a training datasets $\{X_i^{'}, X_i\}_{i=1}^{s^2}$,

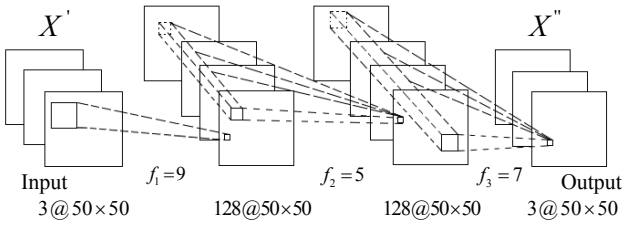


**Fig. 3**. A three-layer convolutional network used for artifacts removal.

where $X$ is the corresponding ground truth image, our goal is to learn a mapping $f$ that predicts values $X^{''} = f(X^{'})$, where $X^{''}$ is an estimated high-resolution image with less artifacts. We minimize the mean squared error $\frac{1}{2}||X^{''} - X^{'}||^2$ averaged over the training set to train a network suppressing the artifacts.

Although the training process needs input images of fixed size, our network is fully convolutional, so it can handle images of arbitrary sizes without any pre- or post-processess.

## 3. EXPERIMENTS

In this section, we evaluate the performance of the proposed and state-of-the-art SR methods on several datasets to verify the effectiveness of the proposed algorithm.

### 3.1. Data Generation

We use Train2014 datasets [20] to train our model, since it contains 1331 natural images which is big enough for our training process. Specifically, we use 100 natural images from the datasets and compute the down-sampled images with different sub-pixel offsets as the input of proposed method. To train the refinement network, we reconstruct the artifacts corrupted high resolution images by using the reconstruction-based algorithm (Alg. 1). Then, we clip these images into $50 \times 50$ patches with a stride of 15, generating 95037 sub-images pairs. Besides, we take $20\%$ training data as the validation set.

**Test datasets:** For benchmark, we use three datasets, i.e., Set5 [19], Set14 [15], and Urban100 [6] which are commonly chosen for demonstrating the SR algorithm to perform the comparison with the state-of-the-art works [10, 11, 21].

### 3.2. Comparisons to State-of-the-Arts

In this section, we show the qualitative results of our method in comparison to state-of-the-art methods. In our paper, we adopt the traditional PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity) indices [22], and PSNR is measured by dB (decibel). Specifically, we compare our method with two multi-frame super resolution methods and two deep


**Fig. 4**. Super-resolution results of three samples from Set14 [15] and Urban100 [6] with $4\times$ scale factor. Obviously, the proposed method recovers sharp lines.

| Dataset | Bicubic PSNR/SSIM | VDSR[13] PSNR/SSIM | LapSRN[3] PSNR/SSIM | MFSR[17] PSNR/SSIM | VideoSR[18] PSNR/SSIM | MISR(Ours) PSNR/SSIM |
|---|---|---|---|---|---|---|
| Set5 | 28.423/0.810 | 30.289/0.871 | 31.522/0.885 | 27.627/0.811 | 28.450/0.845 | **38.200/0.963** |
| Set14 | 26.101/0.704 | 27.166/0.763 | 27.168/0.744 | 25.617/0.740 | 26.121/0.774 | **32.808/0.910** |
| Urban100 | 23.152/0.659 | 24.178/0.736 | 25.201/0.755 | 23.171/0.704 | 21.605/0.659 | **32.073/0.939** |

**Table 1**. Average PSNR/SSIM for $4\times$ scale factor on datasets Set5 [19], Set14 [15] and Urban100 [6]. The bold indicates the best performance. The proposed method outperforms the state-of-the-arts by a large margin.

learning methods, i.e., MFSR [17], VideoSR [18], VDSR [13] and LapSRN [3]. From Fig. 4 and Table 1, we can see that the proposed method achieves better results both objectively and subjectively than the state-of-the-art SR methods. Specifically, the PSNR of our method is higher than LapSRN [3], the best SR algorithm nowadays, by $7dB$ on Urban100 [6]. It is worth noting that both MFSR [17] and VideoSR [18] are specified for video super resolution, and it seems that their results have a slight offsets compared with ground truth images. Therefore, although results of MFSR [17] and VideoSR [18] are of pretty good visual quality, their PSNR and SSIM [22] are not very good in this case, as shown in Fig. 4.

## 4. CONCLUSION

In this work, we have presented a novel multi-image super resolution method, concatenating a reconstruction-based super resolution and a three-layer deep neural network based filtering for artifacts removal. We have demonstrated that our method has superior performance to the state-of-the-art works with a significant performance margin both subjectively and objectively. More specifically, we have demonstrated that 5 to 7 dB PSNR improvements can be obtained over the state-of-the-art LapSRN [3] with our method. As the future step, we will integrate our MISR method with the array system that using multiple off-the-shelf cameras (i.e., low-end) to achieve super high resolution image/video captures.

## 5. ACKNOWLEDGMENT

Dr. Z. Ma is partially supported by National Science Foundation for Young Scholar of Jiangsu Province, China (Grant # BK20140610) and the National Science Foundation of China (Grant # 61571215). We also would like to acknowledge funding from NSFC Projects #61371166, and #61422107.

## 6. REFERENCES

[1] H. Chang, D. Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Computer Vision and*

*Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*. IEEE, 2004, vol. 1, pp. I–I.

[2] S. Dai, M. Han, W. Xu, Y. Wu, and Y. Gong, "Soft edge smoothness prior for alpha channel super resolution," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[3] W. S. Lai, J. B. Huang, N. Ahuja, and M. H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," *arXiv preprint arXiv:1704.03915*, 2017.

[4] Z. Cui, H. Chang, S. Shan, B. Zhong, and X. Chen, "Deep network cascade for image super-resolution," in *European Conference on Computer Vision*. Springer, 2014, pp. 49–64.

[5] G. Freedman and R. Fattal, "Image and video upscaling from local self-examples," *ACM Transactions on Graphics (TOG)*, vol. 30, no. 2, pp. 12, 2011.

[6] J. B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5197–5206.

[7] D. Dai, R. Timofte, and L. Van Gool, "Jointly optimized regressors for image super-resolution," in *Computer Graphics Forum*. Wiley Online Library, 2015, vol. 34, pp. 95–104.

[8] K. Jia, X. Wang, and X. Tang, "Image transformation based on learning dictionaries across image spaces," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 2, pp. 367–380, 2013.

[9] J. Kim, J. Kwon Lee, and K. Mu Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1637–1645.

[10] R. Timofte, V. De Smet, and L. Van Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1920–1927.

[11] R. Timofte, V. De Smet, and L. Van Gool, "A+: Adjusted anchored neighborhood regression for fast super-resolution," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 111–126.

[12] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[13] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1646–1654.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[15] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *International conference on curves and surfaces*. Springer, 2010, pp. 711–730.

[16] Z. Wen, W. Yin, D. Goldfarb, and Y. Zhang, "A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization, and continuation," *SIAM Journal on Scientific Computing*, vol. 32, no. 4, pp. 1832–1857, 2010.

[17] Z. Ma, R. Liao, X. Tao, L. Xu, J. Jia, and E. Wu, "Handling motion blur in multi-frame super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5224–5232.

[18] R. Liao, X. Tao, R. Li, Z. Ma, and J. Jia, "Video super-resolution via deep draft-ensemble learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 531–539.

[19] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," 2012.

[20] T. Y. Lin, M. Maire, S. Belongie, and et al. Hays, J., "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755.

[21] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European Conference on Computer Vision*. Springer, 2014, pp. 184–199.

[22] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.