LOCAL- AND HOLISTIC- STRUCTURE PRESERVING IMAGE SUPER RESOLUTION VIA DEEP JOINT COMPONENT LEARNING

Yukai Shi[1], *Keze Wang*[1], *Li Xu*[2], *Liang Lin*[1⋆]

[1]Sun Yat-Sen University, Guangzhou, China
[2]SenseTime Group Limited

shiyk3@mail2.sysu.edu.cn, kezewang@gmail.com, xuli@sensetime.com, lianglin@ieee.org

ABSTRACT

Recently, machine learning based single image super resolution (SR) approaches focus on jointly learning representations for high-resolution (HR) and low-resolution (LR) image patch pairs to improve the quality of the super-resolved images. However, due to treat all image pixels equally without considering the salient structures, these approaches usually fail to produce visual pleasant images with sharp edges and fine details. To address this issue, in this work we present a new novel SR approach, which replaces the main building blocks of the classical interpolation pipeline by a flexible, content-adaptive deep neural networks. In particular, two well-designed structure-aware components, respectively capturing local- and holistic- image contents, are naturally incorporated into the fully-convolutional representation learning to enhance the image sharpness and naturalness. Extensively evaluations on several standard benchmarks (*e.g.*, *Set5*, *Set14* and *BSD200*) demonstrate that our approach can achieve superior results, especially on the image with salient structures, over many existing state-of-the-art SR methods under both quantitative and qualitative measures.

Index Terms— Image super-resolution; Deep neural network; Deconvolutional process; Low-level computer vision

1. INTRODUCTION

Owing to various practical reasons such as cost of camera, storage limitation and limited bandwidth, images with low resolution are inevitable. To solve this problem, single image super resolution (SR), with the goal of increasing the resolution of the image from the single input, has been drew considerable attention from different research communities.

Techniques for single image SR can be roughly categorized as reconstruction-, example- and interpolation- based approaches. The reconstruction-based approaches [1] assume that the registered frames are in compliance with a global blur degradation model, which can be formulated by explicitly modeling deconvolution [2] or by allowing blind super resolution with unknown blur degradation [3]. Due to the inverse problem with inaccurate blur kernels, reconstruction-based approaches may introduce ringing artifacts around salient structures [3].

Opening up a new data-driven direction for single image SR, example-based approaches [4, 5, 6, 7] advance in the use of internal or external patch data to increase the resolution by a large factor with synthesized artificial details. However, the synthesized details are not necessarily consistent with real details. Specifically, it is not well grounded that the synthesized patch does bring real details of original optical images into the low resolution version.

Though having the advantage in efficiency and balanced performance, interpolation-based approaches, such as bilinear / bicubic / spline, are prone to mix colors along the main edges, especially when encountering large upsampling factors. Most traditional interpolation-based approaches, which use single fixed kernel to interpolate the whole image, are inevitable to blur the main structure, e.g., [8, 9] propose a kind of adaptive-interpolation and attempt to resolve this problem. The pursuit of bringing higher visual quality to human visual system is still the fundamental goal of many practical SR approaches. As a matter of fact, human visual system is more sensitive to color and structural changes than absolute color values. Preserving the contrast and sharpness of the salient structural edge is thus crucial for generating visually plausible results. In contrast, some SR approaches [10, 11] take interpolation operation first. This may bring negative influences on the main structure because of not good enough initialization. It is beneficial if we can incorporate some techniques to learn multiply edge-preserving kernels for interpolation from sufficient image data.

Deep convolutional neural network has been successfully applied to various computer vision tasks including rain/dirt removal [12], image deconvolution [13], noise removal [14] and image in-painting [15]. More recently, [16, 10, 11, 17, 18] have achieved promising results by incorporating deep learning for SR. In particular, example-based approaches with deep learning framework [17, 11] propose to capture the mapping between low- and high- resolution images to obtain the

Corresponding author is Liang Lin.

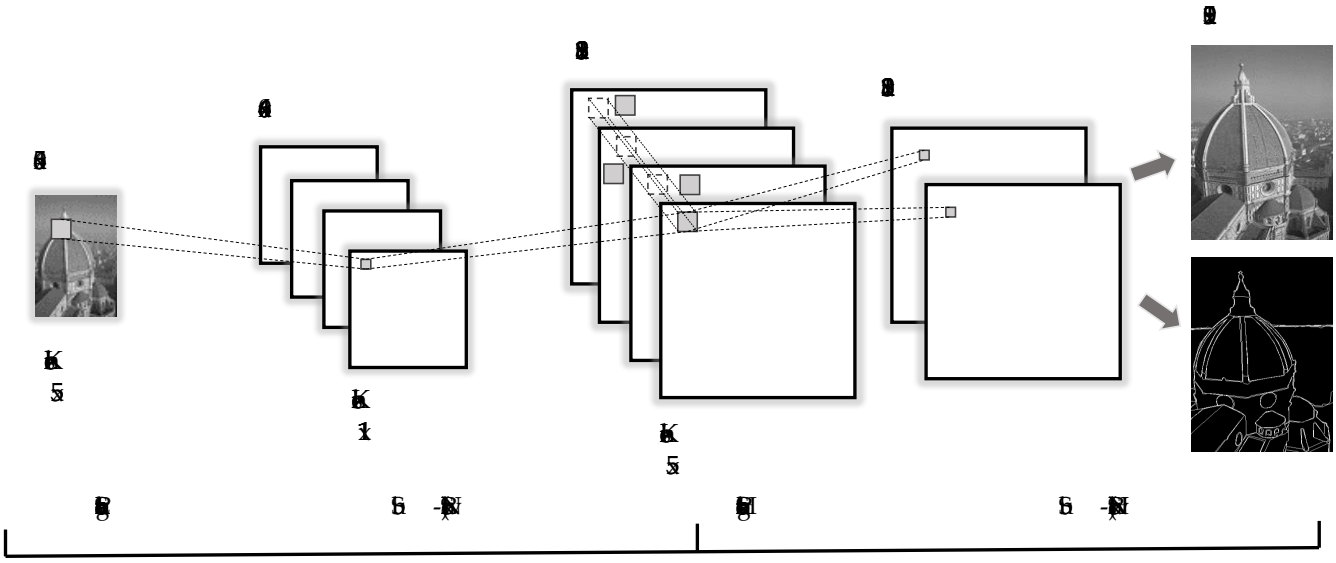


Fig. 1: The architecture of the proposed local- and holistic- structure preserving neural networks. The neural networks are stacked by a convolutional layer, a deconvolution layer, pixel placement operation and a convolution layer. The first two layers and the operation forms the local structure preserving sub-network (LSP), while the final convolutional forms the holistic structure preserving sub-network (HSP). The LSP upsamples the LR patches via deconvolution with local structure preserving displacement (Pixel Placement). Then the HSP further refines the output of LSP via convolution with a multi-task of holistic structure preserving objective.

state-of-the-art performance. Cui *et al.*[17] advocated the use of cascade networks for super resolution, with a local auto-encoder architecture. Dong *et al.* [11] applied the similar FCN in super resolution. However, these approaches are less optimal for that they tend to interpolate image with pixels equally treated and actual contrast of border/texture ignored during the training procedure.

In this paper, we present a local- and holistic- structure preserving neural networks, aiming for salient structural edge preservation. As illustrated in Figure 1, the proposed model is stacked by two component named Local Structure Preserving sub-network (LSP) and Holistic Structure Preserving sub-network (HSP). The LSP upsamples the input low resolution patches via a deconvolution network with local structure preserving displacement (Pixel Placement). The HSP further refines the output of LSP via a fully convolutional layer with a multi-task of structure preserving objective.

The **main contributions** of this paper are organized as follows: (1) We propose a novel deep neural network for super-resolution, which achieves state-of-the-art performance with a small architecture on public *Set5* [19], *Set14* [20] and *BSD200* [21] benchmarks. (2) We make an attempt to re-design the classical interpolation pipeline by replacing building blocks with content-adaptive and structure-aware neural networks. (3) Our study demonstrates that the proposed joint local- and holistic- structure preserving can significantly benefit super resolution.

2. FRAMEWORK

Figure 1 illustrates the architecture of our proposed local- and holistic- structure preserving framework, which consists of two components called Local Structure Preserving sub-network (LSP) and Holistic Structure Preserving sub-network (HSP). The LSP is stacked by a convolutional layer, a deconvolutional layer and our proposed pixel placement operation in order to preserve the local structure of the image. Focusing on the holistic structure of the image, the HSP employs one convolutional layer to further refine the result of LSP by considering non-local and boundary information.

2.1. Local Structure Preserving sub-network (LSP)

Different from existing SR approaches that typically adopting edge-blurring interpolation techniques in the initialization step [11, 4], our proposed LSP incorporates local deconvolution and structure preserving pixel placement into the interpolation phase.

Deconvolution. Traditional interpolation assumes that the LR pixels are evenly placed in the HR grid. In this regard, interpolation is a linear transform invariant (LTI) operation. Specifically, denote the intensity of each pixel in LR and HR image by x_i and y_j, respectively. The interpolation, e.g., bi-linear, can be expressed as

$$y_j = \sum_{n_i \in \Omega} \omega_{n_i} \cdot x_{n_i}, \tag{1}$$

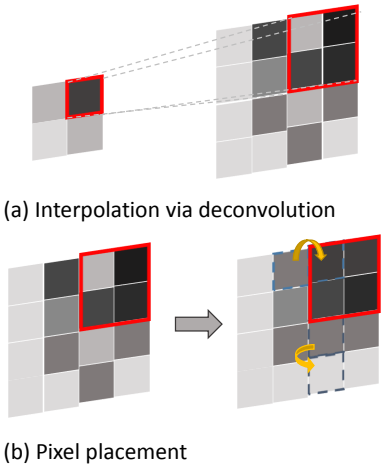


Fig. 2: Illustration of the processing of interpolation by deconvolution.

where Ω is the local window, n_i indexes the pixels and ω_{n_i} indicates the fixed bilinear weights. As one can see that, the formulation to calculate y_j is similar to the deconvolution operation, which is defined as

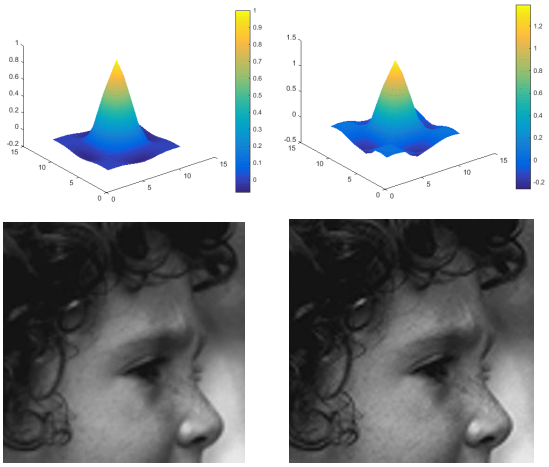
$$h^l = \sigma(W_1 * h_1^{l-1} + \cdots W_n * h_n^{l-1} + b), \tag{2}$$

where $[W_1, W_2, ..., W_n]$ indicates the kernel weight mapping from $(l-1)^{th}$ to the l^{th}, b is the bias. h^l and h^{l-1} mean output and input, respectively. $\sigma(\cdot)$ is the nonlinear function. This process is illustrated in Figure 2a. Hence, the interpolation can be accomplished via a deconvolution layer. Moreover, the deconvolution kernel can be adaptively learned from sufficient training data.

We propose to learn a local deconvolution kernel from sufficient image pairs for interpolation. Figure 3 visualizes some motivating results via the learned kernels from corresponding images. Compared with bicubic, the result reveals that the learned kernel is more extraordinary and its performance is superior to the traditional fast interpolation approach.

Pixel Placement. As mentioned above, the single local deconvolution for interpolation is a linear translation invariant (LTI) operator. Therefore, when applying the LTI filter to pixels with large contrast in LR image, the color mixing is unavoidable because of the actual interaction range between pixels, i.e., image edges are blurred in this case. To overcome this problem, we propose an edge-preserving method called Pixel Placement, which slightly moves the original position of LR pixels in its HR grid to make HR grid homogeneous (see Figure 2b for more details). Figure 4a illustrates a LR sample from the original HR image. Comparing with the result from pure bilinear interpolation in Figure 4b, refining LR pixels in the HR grid (e.g., moving LR pixels farther from the structural edge) contributes to the sharper edge in HR (see Figure 4c).

The local structure preserving interpolation is performed by combining above mentioned deconvolution and pixel-placement inside the neural networks. It should be noted that when the pixel is not evenly placed, the interpolation can not be approximated by one deconvolution. The unevenly distributed pixels may have different influence on pixels, which makes the operation no longer LTI. The shepard interpolation theory [24] addressed the problem by normalizing via an indicator map with value one on those unevenly placed sparse LR pixels. The deconvolution is applied both on the HR grid with LR pixels and the indicator map. The normalization is conducted by element-wise division as:



(a) Bicubic kernel PSNR 32.71 (b) Learned kernel PSNR 33.10

Fig. 3: Results of the bicubic and learned interpolation operators.

$$h_{LSP}(x_i) = \frac{d(f_l(x_i))}{f_l(\mathbf{1}(x_i))}; \tag{3}$$

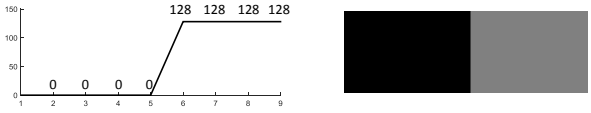
where x_i denotes the input LR image, $f_l(\cdot)$ denotes the deconvolution network, $d(\cdot)$ denotes the image grid guided by the proposed pixel placement and $\mathbf{1}(x_i)$ denotes the indicate map with value on those unevenly placed sparse pixels of x_i.

2.2. Holistic Structure Preserving Sub-network (HSP)

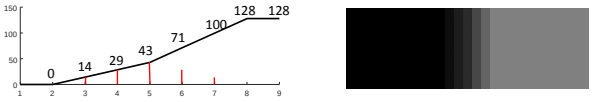
The LSP stands by itself as a single image super resolution network. We aim to improve the performance of SR on the main structural edges, with the help of human labeled boundary guidance. Structure edges and contours are of great importance for human vision system, although they are not necessarily associated with good quantitative results. The HSP first takes the LSP output h_{LSP} as input, i.e., LSP and HSP are combined together in such an end-to-end fine-tuning manner. Moreover, we additionally add an auxiliary objective for global boundary predictions, which guides the network to obtain the ability of preserving the global structure. Specifically, we train our model on a set of (LR, HR, boundary) triplets. In general, suppose we generate N samples in the training set with each sample containing one LR image x_i ,

Test set	Scale	Bicubic		NELLE [22]		SCIP [5]		ANR [23]		SRCNN [11]		Ours	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Set5	2	33.66	0.9299	35.77	0.9489	32.20	0.9511	35.83	0.9499	36.43	0.9515	**36.90**	**0.9547**
	3	30.39	0.8677	31.84	0.8946	32.28	0.9033	31.92	0.8958	32.56	0.9049	**32.79**	**0.9090**
	4	28.42	0.8099	29.61	0.8391	30.03	0.8541	29.69	0.8408	30.31	0.8587	**30.51**	**0.8636**
Set14	2	30.23	0.8689	31.76	0.8992	32.11	0.9026	31.80	0.9004	32.39	0.9042	**32.68**	**0.9079**
	3	27.54	0.7742	28.60	0.8080	28.94	0.8132	28.65	0.8096	29.13	0.8163	**29.31**	**0.8208**
	4	26.00	0.7026	26.81	0.7334	27.14	0.7419	26.85	0.7355	27.40	0.7486	**27.51**	**0.7523**
BSD200	2	29.43	0.8538	30.57	0.8879	31.23	0.8997	30.61	0.8886	31.49	0.9055	**31.78**	**0.9071**
	3	27.18	0.7621	27.89	0.7948	28.13	0.8014	27.92	0.7962	28.33	0.8093	**28.47**	**0.8112**
	4	25.92	0.6955	26.47	0.7240	26.63	0.7284	26.50	0.7258	26.84	0.7384	**26.94**	**0.7408**

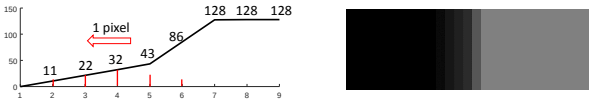
Table 1: Comparison between our models and other methods on the PSNR/SSIM indexes. We use the **bold face** to label the first place in each track.



(a) One dimension HR sample



(b) Bilinear kernel without LSP-displacement



(c) Bilinear kernel with LSP-displacement

Fig. 4: Illustration of the the pixel placement problem. (a) the original HR sample; (b) the result of bilinear upsampling; (c) By moving LR pixel away from the edge in HR grid, the interpolated edge becomes much clearer.

one groundtruth HR image y_i, and one boundary image b_i. Denote the parameters of LSP and HSP by ω_l and ω_h, respectively, our model can be formulated as:

$$\min_{\omega_l,\omega_h} \frac{1}{N}\sum_{i=1}^{N}\left([y_i, \alpha\cdot b_i] - f_h(h_{LSP}(x_i;\omega_l);\omega_h)\right)^2 \quad (4)$$

where $f_h(\cdot)$ denotes two outputs (i.e., 2 feature maps) of the proposed HSP with the multi-task objective. $[y_i, \alpha\cdot b_i]$ denotes the fitting targets and α controls weight of preserving global structure.

2.3. Model Training and Testing

As our proposed model seamlessly integrates local structure preserving and global structure preserving, the standard back propagation algorithm is applicable to optimize the model parameters $\{\omega_l, \omega_h\}$. The only remark is that the pixel placement operation is non-differentable, thus we need to save the original location of the moved pixels for the back propagating from HSP to LSP. As for testing the image x_i, the predict result is inside $f_h(h_{LSP}(x_i;\omega_l);\omega_h)$.

3. EXPERIMENTS

3.1. Experiment Setting

Datasets. To justify the effectiveness of our proposed model, we conduct extensive evaluations on three public benchmarks, i.e., the *Set5* [19], *Set14* [20] and *BSD500* [21] dataset. The *BSD500* dataset consists of 500 images, we use its training / validation set (300 images) for training, the rest 200 images for testing (*BSD200*). Besides, the *Set5* and *Set14* datasets are conducted by following the same experiment setting as other state-of-the-art methods [11, 5, 23].

Implementation Details. For all above datasets, we first convert their images into YCbCr colorspace and only consider the luminance channel. Then we generate (32 x 32) sub-images from each HR images in the training set by stride of 12 pixel, and get the corresponding sub-images from boundary annotations of the dataset in the same way. We generate LR sub-images from HR sub-images by sub-sampling as [11]. To demonstrate that our model can handle image blur, we also introduce blurring into LR sub-images. As a result, we obtain the training set with 273600 triplets. Our proposed model is trained with the batch size 32 and fixed learning rate $1e-12$ under each scaling factor of $\{\times 2, \times 3, \times 4\}$.

Evaluation Metric and Compared Methods. We adopt the widely used PSNR[1] and SSIM as our evaluation metrics. We compare our proposed joint component learning network of LSP and HSP (ours) with sparse coding and image prior (SCIP) [5], neighbor embeding and locally linear embedding (NELLE) [22], anchored neighborhood regression (ANR) [25] and a three layers CNN (SRCNN) [11].

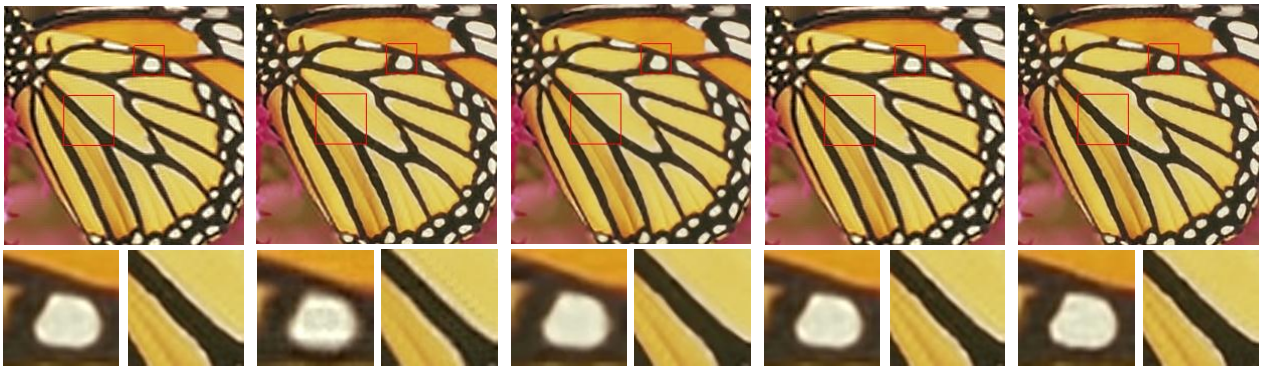
3.2. Empirical Results and Analysis

Table 1 indicates that our proposed model surpasses the former state-of-the-art methods both in the PSNR and SSIM indexes across all the scaling factors. From the results of PSNR metric, one can see that our model outperforms the compared methods by a large margin on *Set5* (factor 2) and *Set14*. A similar tend can also be observed for the SSIM evaluation metrics. To be specific, Table 1 has shown the aver-

[1]Peak signal-to-noise ratio

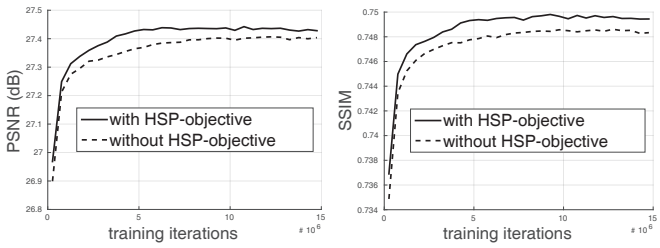
(a) NELLE [22] (b) SCIP [6] (c) ANR [23] (d) SRCNN [11] (e) Ours

Fig. 5: The PSNR / SSIM indexes BSD200 '146074' image from *Set5*



(a) NELLE [22] (b) SCIP [6] (c) ANR [23] (d) SRCNN [11] (e) Ours

Fig. 6: 'Butterfly' image from *Set5*.



(a) PSNR curves (b) SSIM curves

Fig. 7: The PSNR / SSIM curves generated by models trained with and without HSP-objective.

age gains, achieved by our model, are 0.47dB, 0.23dB, 0.2dB higher than second best method SRCNN [11], which is also a deep learning method.

Figure 5 and 6 visualize some promising examples. We interpolate the Cb and Cr chrominance channels by the bicubic method to generate color images for better views. To clearly shown up the difference, we choose two patches from each group and attach them below. With the help of LSP, we can obtain multiple edge-preserving kernels. By means of these novel kernels, we achieve one better structure interpolation image. Compared to other methods, it gives rise to our results have sharper and clearer boundaries. We suggest the reader zooming in the figures to find more details.

To justify the contribution of the proposed HSP, we train our model with/without HSP-objective and plot the PSNR and SSIM curves. In Figure 7, it seems that the HSP-objective not only accelerates the convergence, but also provides a better initialization, which helps the network to converge at a better local optimal.

4. CONCLUSION AND FUTURE WORK

In this work, we propose a novel structure preserving image super-resolution approach from both local and holistic perspectives. Extensive experiments demonstrate that our model not only achieve state-of-the-art performance on popular evaluation metrics, but also have a better visual quality. There are several directions in which we intend to extend this work. First, we only consider to estimate the pixel-place in the local gradient, we will explore more possibility in different elements and situations in the follow-up work. Second, we plan to extend our model for higher level vision tasks such as face hallucination.

Acknowledgements

We would like to thank Liliang Zhang for his assistance in the experiments.This work was supported in part by Guangdong Natural Science Foundation under Grant S2013050014548 and 2014A030313201, in part by Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (the second phase), and in part by Fundamental Research Funds for the Central Universities.

5. REFERENCES

[1] Michal Irani and Shmuel Peleg, "Improving resolution by image registration," *CVGIP: Graphical models and image processing*, vol. 53, no. 3, pp. 231–239, 1991. 1

[2] Qi Shan, Zhaorong Li, Jiaya Jia, and Chi-Keung Tang, "Fast image/video upsampling," in *ACM Transactions on Graphics (TOG)*, 2008, vol. 27, p. 153. 1

[3] Tomer Michaeli and Michal Irani, "Nonparametric blind super-resolution," in *ICCV*, 2013, pp. 945–952. 1

[4] Daniel Glasner, Shai Bagon, and Michal Irani, "Super-resolution from a single image," in *CVPR*, 2009, pp. 349–356. 1, 2

[5] Kwang In Kim and Younghee Kwon, "Single-image super-resolution using sparse regression and natural image prior," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 6, pp. 1127–1133, 2010. 1, 4

[6] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma, "Image super-resolution via sparse representation," *Image Processing, IEEE Transactions on*, vol. 19, no. 11, pp. 2861–2873, 2010. 1, 5

[7] Gilad Freedman and Raanan Fattal, "Image and video upscaling from local self-examples," *ACM Transactions on Graphics (TOG)*, vol. 30, no. 2, pp. 12, 2011. 1

[8] Jinyu Chu, Ju Liu, Jianping Qiao, Xiaoling Wang, and Yujun Li, "Gradient-based adaptive interpolation in super-resolution image restoration," in *Signal Processing, 2008. ICSP 2008. 9th International Conference on*. IEEE, 2008, pp. 1027–1030. 1

[9] Stéfan J van der Walt and BM Herbst, "A polygon-based interpolation operator for super-resolution imaging," *arXiv preprint arXiv:1210.3404*, 2012. 1

[10] Jimmy SJ. Ren, Li Xu, Qiong Yan, and Wenxiu Sun, "Shepard convolutional neural networks," in *Advances in Neural Information Processing Systems*. 2015. 1

[11] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Learning a deep convolutional network for image super-resolution," in *Computer Vision–ECCV 2014*, pp. 184–199. Springer, 2014. 1, 2, 4, 5

[12] David Eigen, Dilip Krishnan, and Rob Fergus, "Restoring an image taken through a window covered with dirt or rain," in *ICCV*. IEEE, 2013, pp. 633–640. 1

[13] Li Xu, Jimmy S Ren, Ce Liu, and Jiaya Jia, "Deep convolutional neural network for image deconvolution," in *Advances in Neural Information Processing Systems*, 2014, pp. 1790–1798. 1

[14] Viren Jain and Sebastian Seung, "Natural image denoising with convolutional networks," in *Advances in Neural Information Processing Systems*, 2009, pp. 769–776. 1

[15] Junyuan Xie, Linli Xu, and Enhong Chen, "Image denoising and inpainting with deep neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 341–349. 1

[16] Zhaowen Wang, Ding Liu, Jianchao Yang, Wei Han, and Thomas Huang, "Deep networks for image super-resolution with sparse prior," *arXiv preprint arXiv:1507.08905*, 2015. 1

[17] Zhen Cui, Hong Chang, Shiguang Shan, Bineng Zhong, and Xilin Chen, "Deep network cascade for image super-resolution," in *ECCV*, pp. 49–64. 2014. 1, 2

[18] K. Zhang, B. Wang, W. Zuo, and H. Zhang, "Joint learning of multiple regressors for single image super-resolution," *IEEE Signal Processing Letters*, vol. 23, no. 1, pp. 1–1, 2015. 1

[19] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," 2012. 2, 4

[20] Roman Zeyde, Michael Elad, and Matan Protter, "On single image scale-up using sparse-representations," in *Curves and Surfaces*, pp. 711–730. Springer, 2012. 2, 4

[21] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011. 2, 4

[22] Hong Chang, Dit-Yan Yeung, and Yimin 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. 4, 5

[23] Radu Timofte, Vincent De, and Luc Van Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1920–1927. 4, 5

[24] Donald Shepard, "A two-dimensional interpolation function for irregularly-spaced data," in *ACM national conference*. ACM, 1968, pp. 517–524. 3

[25] Chih-Yuan Yang and Ming-Hsuan Yang, "Fast direct super-resolution by simple functions," in *Proceedings of IEEE International Conference on Computer Vision*, 2013. 4