Dual Recovery Network with Online Compensation for Image Super-Resolution

Sifeng Xia[1], Wenhan Yang[1], Jiaying Liu[1,*] and Zongming Guo[1,2]
[1]Institute of Computer Science and Technology, Peking University, Beijing, China
[2]Cooperative Medianet Innovation Center, Shanghai, China

***Abstract*—Image super-resolution (SR) methods essentially lead to a loss of some high-frequency (HF) information when predicting high-resolution (HR) images from low-resolution (LR) images without using external references. To address this issue, we additionally utilize online retrieved data to facilitate image SR in a unified deep framework. A novel dual high-frequency recovery network (DHN) is proposed to predict an HR image with three parts: an LR image, an internal inferred HF (IHF) map (HF missing part inferred solely from the LR image) and an external extracted HF (EHF) map. In particular, we infer the HF information based on both the LR image and similar HR references which are retrieved online. For the EHF map, we align the references with affine transformation and then in the aligned references, part of HF signals are extracted by the proposed DHN to compensate for the HF loss. Extensive experimental results demonstrate that our DHN achieves notably better performance than state-of-the-art SR methods.**

I. Introduction

Image super-resolution (SR) aims to estimate a high-resolution (HR) image from low-resolution (LR) observations. In essence, due to the information loss in the image degradation process, SR is an ill-posed problem. The earliest works, image interpolation, estimate the HR image based on local statistics of the LR image. Typical methods include bilinear, bicubic and new edge directed interpolation that predict the HR pixels by utilizing the spatial relationship between LR and HR pixels. Later on, many successive works [1], [2] regard the image SR as a Maximum-a-posteriori estimation and propose to impose various priors to constrain the inverse estimation of image SR. In these methods, priors and constraints are typically achieved in a heuristic way. Thus, it is insufficient to represent the diversified patterns of natural images.

Learning based methods obtain a mapping between LR and HR images based on a large training set with dynamic learned prior knowledge. Sparse representation based methods such as [3] learn the map by building an LR and HR patch mapping dictionary. Neighbor embedding (NE) methods linearly combine the HR neighbors to infer the HR image. Timofte *et al.* [4] proposed an adjusted anchored neighborhood regression method for image SR. Li *et al.* [5] proposed a neighbor preserving based method which specially utilizes HR reference patches only in reconstructing the high frequency region of LR images. Recently, deep-learning based methods [6], [7], [8], [9], [10] are proposed. SRCNN is the first method [6] that utilizes a three-layer convolutional network for image SR. In [7], the sparse prior is incorporated into the network. Then, the residual learning [8] and sub-band recovery with edge guidance [9] networks are constructed to recover HF signal and offer state-of-the-art performance.

Despite impressive results achieved by the learning-based methods, some HF information has still been lost because of the ill-posed nature of the image SR and the problem that mean squared error leads to *regression to mean* [11]. As a result, a few methods have recently been proposed, which additionally compensate for HF information loss with online retrieved HR references. Yue *et al.* [12] directly utilized the references to enhance the SR result by patch matching and patch blending. Li *et al.* [13] used the retrieved HR image patches to learn more accurate sparse distribution. Liu *et al.* [14] utilized a group-structured sparse representation to further use the nonlocal dependency information of HR references. However, in these methods there are still several important issues not being fully considered. For example, their fusion methods do not effectively extract external HF information for compensation, which may even bring artifacts. Besides, they did not make full use of the internal redundancy to benefit the recovery of HF information.

To address the aforementioned issues, we propose a unified deep network that additionally utilizes online retrieved data to facilitate image SR. Our work can efficiently extract an HF map from multiple HR references that are retrieved based on the intermediately inferred SR image.

Contributions of this paper are as follows: 1) It is the first work that efficiently extracts high-frequency information from the HR reference and successfully compensate for the HF information loss of the SR result with the deep framework. 2) We show the proposed method is capable to model internal and external images jointly, achieving a more accurate and robust fusion of internal and external information for HF information recovery. 3) Compared with both previous deep learning-based methods and online compensation SR methods, our approach has offered new state-of-the-art performance.

The rest of the article is organized as follows. Sec. II illustrates our DHN network. Details of utilizing the EHF map for compensation are introduced in Sec. III. Experimental results are shown in Sec. IV and concluding remarks are given in Sec. V.

*Corresponding author
This work was supported by National Natural Science Foundation of China under contract No.U1636206. We also gratefully acknowledge the support of NVIDIA Corporation with the GPU for this research.

978-1-5386-4881-0/18/$31.00 ©2018 IEEE

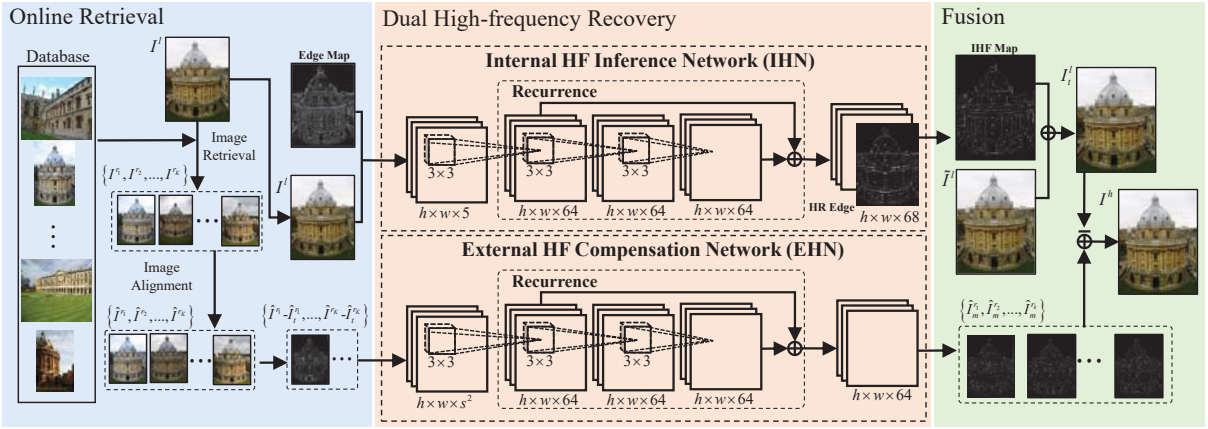


Fig. 1. Framework of the proposed SR method based on the dual high-frequency recovery network (DHN) with online compensation. $h \times w \times *$ means size of the convolution layer and $*$ represents channel numbers. s is the magnification factor. Image $\tilde{I}^l$ is simply up-sampled form I^l. I_t^l is the intermediate image derived by IHN and I^h is the further enhanced result image.

II. Dual High-Frequency Recovery Network

Given an LR image I^l, we predict the HR image I^h from I^l with the reference of K retrieved HR reference images $\{I^{r_1}, I^{r_2}, ..., I^{r_K}\}$ by our dual high-frequency recovery network (DHN). Architecture of the proposed DHN has been illustrated in Fig. 1. DHN consists of two components called internal high-frequency inference network (IHN) and external high-frequency compensation network (EHN), respectively. IHN infers missing HF information of I^l merely based on internal data in I^l. Then, the intermediate SR image I_t^l is generated by combining the internal inferred HF (IHF) map and the simply up-smapled LR image $\tilde{I}^l$. EHN further enhances the final SR result I^h by adding the external extracted HF (EHF) map obtained from the aligned retrieved HR reference images $\{\hat{I}^{r_1}, \hat{I}^{r_2}, ..., \hat{I}^{r_K}\}$ to the intermediate image I_t^l.

A. Internal High-Frequency Inference Network

The first component IHN proposed by [9] is utilized to initially reconstruct the LR image I^l with its own information. As shown in Fig.1, I^l and its edge map, which is extracted by applying a hand-crafted edge detector, are utilized as the input of IHN. Then, the recurrent network of IHN estimates the IHF map from the above input. IHN also predicts an HR edge map, which is used to further guide the HF map estimation.

With the inferred IHF map, the intermediate result image I_t^l is then generalized as follows:

$$I_t^l = \tilde{I}^l \oplus \varphi(I^l), \tag{1}$$

where $\oplus$ is the sum operation and $\varphi(I^l)$ represents the process that IHN infers IHF map from LR image I^l. $\tilde{I}^l$ is the image that simply up-sampled from I^l. We then define the loss of IHN as the combination of loss of the predicted HR edge and I_t^l. The loss is measured by the mean squared error (MSE) with the ground truth signal.

B. External High-Frequency Compensation Network

IHN works well in predicting the HF map from an LR image. However, during this process not all HF information can be well recovered. This inspires us to construct EHN to further extract the significant EHF map $\hat{I}_m^r$ from each HR reference $\hat{I}^r$. Note that during training process $\hat{I}^r$ is generated from the ground truth HR image.

It's common for an LR image and its reference image to have illumination and color differences. Moreover, there is much useless low-frequency information in the reference that may affect HF information extraction. Therefore we take different measures to improve the robustness of the process of extracting $\hat{I}_m^r$. First, contrast of the label images is additionally adjusted to simulate the common illumination and color differences in training process. Besides, we alternatively utilize the difference image between $\hat{I}^r$ and its intermediate SR image $\hat{I}_t^r$ as the input of EHN, rather than directly input the information of $\hat{I}^r$. $\hat{I}_t^r$ is obtained through up-sampling the down-sampled image of $\hat{I}^r$ by IHN. The difference image is chosen because of its high efficiency in reducing illumination and color differences and removing redundant low-frequency information.

Then, EHN extracts the EHF map from the input by the recurrent network. Final reconstructed result I^h is derived by:

$$I^h = I_t^l \bar{\oplus} \psi(\hat{I}^r - \hat{I}_t^r), \tag{2}$$

where ψ is the formulation of the process that EHN extracts the HF map $\hat{I}_m^r$. The operation $\bar{\oplus}$ represents the combination of the intermediate image I_t^l and $\hat{I}_m^r$. During the training process, $\hat{I}_m^r$ is directly added to I_t^l. In the testing process, $\hat{I}_m^r$ is utilized based on patch matching results, which is elaborated in Sec. III-B. Loss of EHN is defined as MSE between I^h and the raw ground truth image.

III. Online Compensation

Different with the training process, we retrieve HR reference images $\{I^{r_1}, I^{r_2}, ..., I^{r_K}\}$ online for compensation with the method proposed in [14] during the testing process. Then, the aligned HR references $\{\hat{I}^{r_1}, \hat{I}^{r_2}, ..., \hat{I}^{r_K}\}$ are derived by aligning each I^r to I_t^l and the HF maps $\{\hat{I}_m^{r_1}, \hat{I}_m^{r_2}, ..., \hat{I}_m^{r_K}\}$ are later extracted from the aligned references. As pixels in

978-1-5386-4881-0/18/$31.00 ©2018 IEEE

each aligned reference $\hat{I}^r$ are still not exactly corresponding to the pixels at the same position of I_t^l ,extracted feature values of $\hat{I}_m^r$ can not be directly added to the intermediate up-sampled image I_t^l. Thus patch matching is used to guide the combination of $\hat{I}_m^r$ and I_t^l.

A. Patch Matching

There are usually significant differences on illumination, color and resolution between the intermediate SR image I_t^l and each aligned HR reference $\hat{I}^r$. As a result, for the purpose of better matching results we first utilize the intermediate SR reference image $\hat{I}_t^r$ mentioned in Sec. II-B that shares similar resolution-level with I_t^l for matching. Then, we adjust $\hat{I}_t^r$ to reduce the effect of illumination difference:

$$\hat{I}_t^{r'} = (\hat{I}_t^r - \tau(\hat{I}_t^r))\frac{\sigma(I_t^l)}{\sigma(\hat{I}_t^r)} + \tau(I_t^l), \tag{3}$$

where $\hat{I}_t^{r'}$ is the transform result, $\tau(\cdot)$ and $\sigma(\cdot)$ are the mean and standard deviation values of all pixels of the image, respectively. Then, I_t^l is split into overlapped query patches of size $\sqrt{n} \times \sqrt{n}$ at the step size 4. And we search for the corresponding patches of the query patches within a search window in $\hat{I}_t^{r'}$.

Since small patches contain little structural information of raw images, patch matching results at small patch size are not accurate. Thus we perform patch matching between I_t^l and $\hat{I}_t^{r'}$ with large patches. Considering it is impossible for each patch in I_t^l to have an exact corresponding large patch in $\hat{I}_t^{r'}$, a method that adaptively adjusts patch sizes according to patch difference [12] is adopted for more accurate patch matching.

Let $\mathbf{P}_i$ denote the query patch of size $\sqrt{n} \times \sqrt{n}$ in I_t^l centered at position i and $\mathbf{Q}_j^i$ denote the $\sqrt{n} \times \sqrt{n}$ candidate patch in $\hat{I}_t^{r'}$ centered at j. We search for the best matching candidate patch of $\mathbf{P}_i$ within the search window of size $3\sqrt{n} \times 3\sqrt{n}$ centered at i in $\hat{I}_t^{r'}$. The patch distance between $\mathbf{P}_i$ and $\mathbf{Q}_j^i$ is defined as:

$$d(\mathbf{P}_i, \mathbf{Q}_j^i) = ||\mathbf{P}_i - \mathbf{Q}_j^i||_2^2 + \rho||\nabla(\mathbf{P}_i) - \nabla(\mathbf{Q}_j^i)||_2^2, \tag{4}$$

where ∇ is the operation that calculates the gradient of the patches and ρ is the weighting parameter, which is set to be 10 in this paper. Besides, DC components of the patches are removed before distance computation.

The value of $d(\mathbf{P}_i, \mathbf{Q}_j^i)/(\sqrt{n} \times \sqrt{n})$ is defined as gradient mean square error (GMSE) and G_i^{min} is set as minimum GMSE value between the query patch $\mathbf{P}_i$ and the candidate patch $\mathbf{Q}_j^i$. Patch matching is performed at initial size 21×21 and changed to a smaller size if the value of G_i^{min} is too large according to Eq. 5.

$$\sqrt{n} = \begin{cases} 21, & G_i^{min} <= 200, \\ 17, & 200 < G_i^{min} <= 500, \\ 13, & 500 < G_i^{min} <= 800, \\ 9, & G_i^{min} > 800. \end{cases} \tag{5}$$

The sliding step of patch matching is set to be $\sqrt{n}/3$. Then, a closest candidate patch $\mathbf{Q}_{j_0}^i$ is found. However, a large step size may result in missing a better matching patch in $\hat{I}_t^{r'}$. Thus we further search a candidate patch of the same size as $\mathbf{Q}_{j_0}^i$ within a $(2 \times \sqrt{n}/3 - 1)^2$ size search window centered at position j_0 in $\hat{I}_t^{r'}$, with the step size of 1.

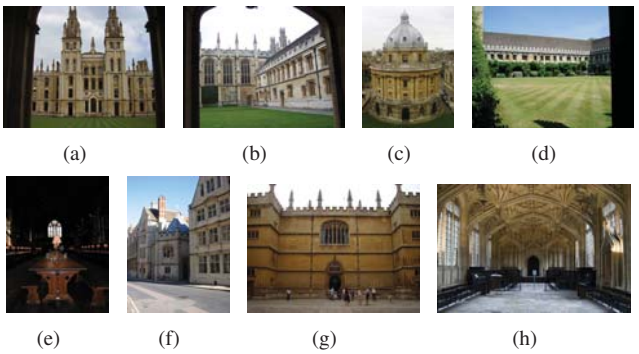


Fig. 2. Testing images from (a) to (h).

B. External High-Frequency Information Utilization

After patch matching, pixels at the same position in the matched patches between I_t^l and $I_t^{r'}$ are matched. Then, the EHF maps are combined with I_t^l based on the pixel-wise matching correlation. For each pixel $\mathbf{p}$ in I_t^l, we define the set of its matching pixels in K EHF maps as $\Omega_{\mathbf{p}}$. Then, the final fused external HF map I_m^l that can be directly added to I_t^l is obtained by:

$$I_{m,\mathbf{p}}^l = \begin{cases} \dfrac{\sum\limits_{\mathbf{q} \in \Omega_{\mathbf{p}}} \hat{I}_{m,\mathbf{q}}^r \cdot e^{\frac{-d(\mathbf{p},\mathbf{q})}{100}}}{\sum\limits_{\mathbf{q} \in \Omega_{\mathbf{p}}} e^{\frac{-d(\mathbf{p},\mathbf{q})}{100}}}, & |\Omega_{\mathbf{p}}| \neq 0, \\ 0, & |\Omega_{\mathbf{p}}| = 0. \end{cases} \tag{6}$$

$|\Omega_{\mathbf{p}}|$ represents the number of elements in set $\Omega_{\mathbf{p}}$. $d(\mathbf{p}, \mathbf{q})$ is the GMSE value between the patches that p and q belong to.

Finally the result SR image is obtained by directly adding the final extracted HF map I_m^l to the intermediate reconstructed SR image I_t^l as $I^h = I_t^l \oplus I_m^l$.

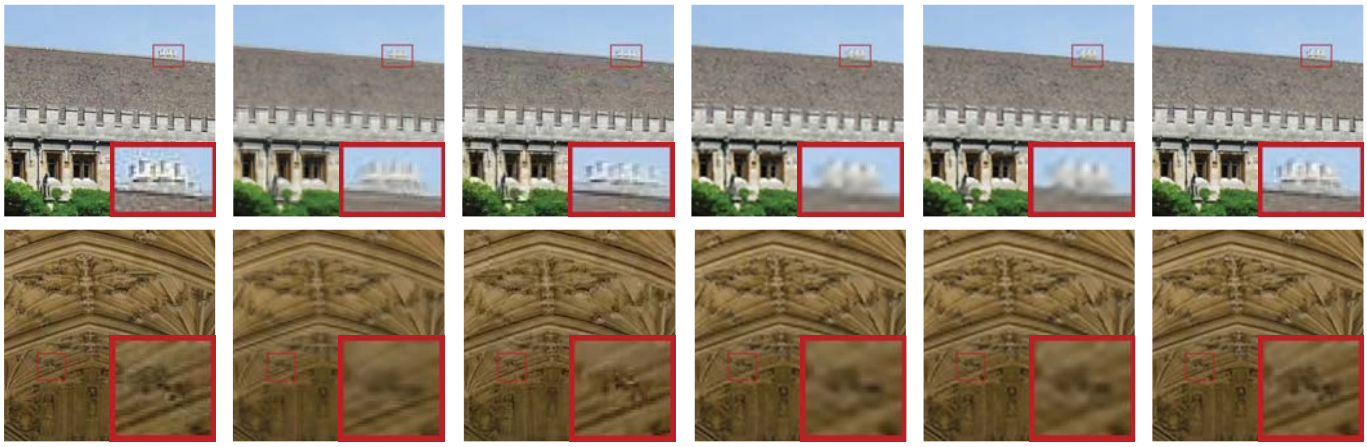
IV. Experimental Results

A. Experimental Settings

We train our DHN based on 91 images in [3] and 200 training images in *BSD500* [15]. The images are first transferred to YC_bC_r color space and only utilize the Y channel. The chrominance channels are later simply up-sampled by the bicubic method in the testing process. Then, we generate sub-images at the size of 32×32 from images in the dataset with the stride step of 16 pixel. Down-sampling method in [16] is utilized that images are first blurred and then down-sampled with factors of 2, 3 and 4. As a result, around 10 thousand sub-images are obtained for training. The learning rate is initially set as 10^{-4} and drops to 10^{-5} after 50,000 iterations.

We compare our algorithm with different SR methods including a typical learning-based SR method [5] (denoted as NE) and two online compensation methods [12], [14] (respectively denoted as Landmark and GSSR). For fair comparison,

978-1-5386-4881-0/18/$31.00 ©2018 IEEE



(a) Ground Truth (b) NE [5] (c) Landmark [12] (d) GSSR [14] (e) Baseline [9] (f) Proposed Method

Fig. 3. Subjective results of different methods with magnification factor 3 for testing images Fig. 2(d) and Fig. 2(h). Some regions that have HF signal have been marked in red rectangle and enlarged for comparison.

TABLE I
PSNR and SSIM values of different methods. $(\cdot)$ denotes performance gain of the proposed method compared with other methods.

Scale	Metrics	NE	Landmark	GSSR	Baseline	Proposed
2	PSNR	28.40	30.41	31.39	32.56	**33.66**
		(5.26)	(3.25)	(2.27)	(1.10)	-
	SSIM	0.822	0.860	0.894	0.922	**0.937**
		(.115)	(.077)	(.043)	(.015)	-
3	PSNR	27.25	29.31	29.20	29.48	**30.93**
		(3.69)	(1.63)	(1.74)	(1.45)	-
	SSIM	0.796	0.826	0.840	0.849	**0.884**
		(.088)	(.058)	(.044)	(.035)	-
4	PSNR	25.61	27.71	27.69	27.85	**29.35**
		(3.74)	(1.64)	(1.67)	(1.50)	-
	SSIM	0.740	0.786	0.785	0.791	**0.835**
		(.095)	(.049)	(.051)	(.044)	-

we add the retrieved HR reference image to the training set of learning-based method NE. Besides, the intermediate results derived by IHN [9] are also shown as the baseline. The baseline is one of the newest deep based SR methods without using external references. The testing images are chosen from the Oxford Building dataset[1] and the online retrieval is also performed over it. There are totally 8 testing images named from (a) to (h) for comparison, as shown in Fig. 2. We set $K = 4$ for the number of reference images. More experimental results can be found on our website[2].

B. Experimental Results and Analysis

Table I shows objective results of 8 chosen images. Our proposed method obtains the best average PSNR and SSIM values in all cases.

Subjective results are shown in Fig. 3. The edge-preserving based method NE successfully obtains more sharp edge but fails to reconstruct other more detailed HF signals. Although Landmark has successfully combined some HF signals of HR references, artifacts sometimes are brought by incorrect patch matching results or inappropriate patch blending. Sparse-based method GSSR did not consider position feature of the reference patches. While there are many similar reference patches, more noise are brought into GSSR's SR reults. Edge feature combined baseline method [9] has also well reconstructed some HF signal. However, without information from HR references, it fails to reconstruct the detail in complex regions. On the contrary, our method achieves the best result in HF information reconstruction.

[1]http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/

[2]http://www.icst.pku.edu.cn/struct/Projects/DualSR.html

TABLE II
PSNR and SSIM values of VDSR and the proposed method.

Metrics	VDSR			Proposed Method		
	2	3	4	2	3	4
PSNR	33.12	29.90	28.34	**33.94**	**31.28**	**30.04**
	0.81	**1.38**	**1.70**	-	-	-
SSIM	0.931	0.860	0.803	**0.942**	**0.892**	**0.856**
	0.011	**0.032**	**0.053**	-	-	-

We also compare with one of state-of-the-art methods, VDSR[8]. Due to the different bicubic down-sampling configuration, we have retrained our network by utilizing VDSR as the IHN under the new configurat. The results have been shown in Table II. Our method still obtains the gain over VDSR.

V. Conclusion

In this paper, we propose a deep online compensation network for image super-resolution. With the IHF map estimated by IHN, we initially obtain an intermediate SR result by combining the IHF map with a simply up-sampled LR image. Then, the EHF maps are further extracted from online retrieved HR references for compensation. The final SR result is obtained by adding the fused EHF map to the intermediate SR result. Extensive experimental results demonstrate that the proposed method can robustly extract external HF maps from the reference images and significantly improve the SR results based on the compensation brought by the EHF maps.

978-1-5386-4881-0/18/$31.00 ©2018 IEEE

REFERENCES

[1] J. Sun, J. Sun, Z. Xu, and H. Y. Shum, "Gradient profile prior and its applications in image super-resolution and enhancement," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1529–1542, 2011.

[2] A. Marquina and SJ. Osher, "Image super-resolution by TV-regularization and bregman iteration," *Journal of Scientific Computing*, vol. 37, no. 3, pp. 367–382, 2008.

[3] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

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

[5] Y. Li, J. Liu, W. Yang, and Z. Guo, "Neighborhood regression for edge-preserving image super-resolution," in *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*, 2015.

[6] C. Dong, C. Chen, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. European Conference on Computer Vision*, 2014.

[7] D. Liu, Z. Wang, B. Wen, J. Yang, W. Han, and T. S. Huang, "Robust single image super-resolution via deep networks with sparse prior," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3194–3207, 2016.

[8] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2016.

[9] W. Yang, J. Feng, J. Yang, F. Zhao, J. Liu, Z. Guo, and S. Yan, "Deep edge guided recurrent residual learning for image super-resolution," *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 5895 – 5907, 2017.

[10] W. Yang, J. Feng, G. Xie, J. Liu, Z. Guo, and S. Yan, "Video super-resolution based on spatial-temporal recurrent residual networks," *Computer Vision and Image Understanding*, 2017.

[11] R. Timofte, VD. Smet, and LV. Gool, "Semantic super-resolution: When and where is it useful?," *Computer Vision and Image Understanding*, 2016.

[12] H. Yue, X. Sun, J. Yang, and F. Wu, "Landmark image super-resolution by retrieving web images," *IEEE Transactions on Image Processing*, vol. 22, no. 12, pp. 4865–4875, 2013.

[13] Y. Li, W. Dong, G. Shi, and X. Xie, "Learning parametric distributions for image super-resolution: Where patch matching meets sparse coding," in *Proc. IEEE Int'l Conf. Computer Vision*, 2015.

[14] J. Liu, W. Yang, X. Zhang, and Z. Guo, "Retrieval compensated group structured sparsity for image super-resolution," *IEEE Transactions on Multimedia*, vol. 19, no. 2, pp. 302–316, 2017.

[15] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 898–916, 2011.

[16] Y. Li, W. Dong, G. Shi, and X. Xie, "Learning parametric distributions for image super-resolution: Where patch matching meets sparse coding view document," in *Proc. IEEE Int'l Conf. Computer Vision*, 2015.

978-1-5386-4881-0/18/$31.00 ©2018 IEEE