https://doi.org/10.2352/ISSN.2470-1173.2018.05.PMII-245

Texture Enhancement via High-Resolution Style Transfer for Single-Image Super-Resolution

Il Jun Ahn[†] and Woo Hyun Nam[†]; Samsung Electronics Inc. Samsung Research; Seoul, Korea

Abstract

Recently, various deep-neural-network (DNN)-based approaches have been proposed for single-image super-resolution (SISR). Despite their promising results on major structure regions such as edges and lines, they still suffer from limited performance on texture regions that consist of very complex and fine patterns. This is because, during the acquisition of a low-resolution (LR) image via down-sampling, these regions lose most of the high frequency information necessary to represent the texture details. In this paper, we present a novel texture enhancement framework for SISR to effectively improve the spatial resolution in the texture regions as well as edges and lines. We call our method, high-resolution (HR) style transfer algorithm. Our framework consists of three steps: (i) generate an initial HR image from an interpolated LR image via an SISR algorithm, (ii) generate an HR style image from the initial HR image via down-scaling and tiling, and (iii) combine the HR style image with the initial HR image via a customized style transfer algorithm. Here, the HR style image is obtained by down-scaling the initial HR image and then repetitively tiling it into an image of the same size as the HR image. This down-scaling and tiling process comes from the idea that texture regions are often composed of small regions that similar in appearance albeit sometimes different in scale. This process creates an HR style image that is rich in details, which can be used to restore high-frequency texture details back into the initial HR image via the style transfer algorithm. Experimental results on a number of texture datasets show that our proposed HR style transfer algorithm provides more visually pleasing results compared with competitive methods.

(a) Original (b) Bicubic

(c) VDSR (d) HRST (proposed)

Figure 1. Our high-resolution style transfer (HRST) based SISR method compares favorably with a representative related work [19] on texture region (up-sampling factor is 4.).

1. Introduction

The aim of single-image super-resolution (SISR) algorithm is to recover a high-resolution (HR) image from a single low-resolution (LR) image [1]. Although the SISR problem inherently ill-posed, many valuable algorithms have been presented for computer vision and image processing applications such as surveillance imaging, medical imaging, or ultra-high-definition (UHD) image generation where more image details are required. Early methods include simple and fast interpolation-based scheme with bicubic or Lanzcos filter [2]. For better performance, more advanced schemes using statistical image priors [3]-[7] or internal patch recurrence [8], [9] were also introduced.

Meanwhile, sophisticated machine learning based schemes have been widely used to learn the relationship from LR to HR patches. Neighborhood embedding approaches [10], [11] up-sample a given LR image patch by finding similar LR training patches in a low dimensional manifold and combining their corresponding HR patches for reconstruction. Sparse-coding (or dictionary learning) approaches [12]-[14] use a learned compact dictionary on the basis that natural patches can be represented using sparse activations of dictionary atoms. Random forests approaches [15] directly formulate SISR as a regression problem, which can avoid complex and time-consuming training of a sparse dictionary.

Recently, various deep learning-based approaches via convolutionnal neural networks (CNN) were proposed with excellent performance. Dong et al. [16], [17] showed that CNN could be successfully applied for SISR. This CNN method, which we call SRCNN, used a three layer convolutional network and trained in an end-to-end manner to learn a mapping from interpolated LR image to original HR image. To further improve the performance on both accuracy and speed, the authors extended their work to enable the network to learn the mapping from LR to HR image directly, rather than from the interpolated LR image [18]. Since up-sampling is only performed in the last layer of the network, the method can avoid expensive computations in the HR dimension.

Kim et al. [19] presented a highly performant architecture which consists of very deep convolutional network of 20 layers. Since the deep networks lead to enlargement of receptive fields that can take a large image context into account, the method

[†]Both authors contributed equally to this work.

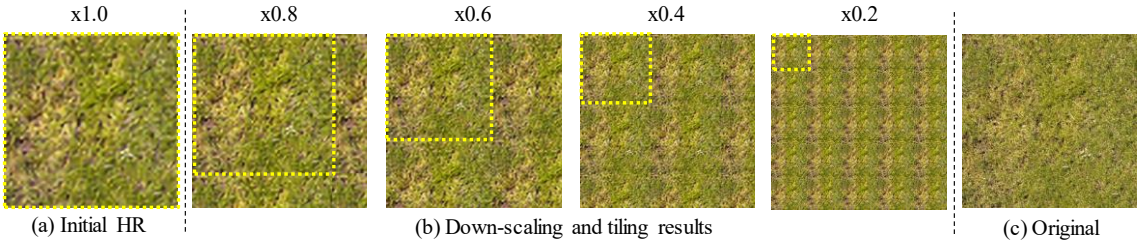

Figure 2. (a) Initial HR image enhanced from the interpolated LR image via a SISR, (b) various results obtained by down-scaling and tiling the initial HR image based on scaling factors ranging from 0.2 to 0.8, and (c) original HR image. In terms of texture detail representation, the result with scaling factor 0.4 is most similar to the original image. The yellow boxes in (b) denote down-scaled ones of initial HR image.

achieved state-of-the art performance with a large margin. They also presented a novel residual learning approach and showed that it is more favorable in training the deep layers than a non-residual based one. Meanwhile, to reduce the number of convolutional parameters while keeping the large receptive fields, the authors proposed a different architecture based on deeply recursive convolutional network [20], which showed comparable performance to [19].

Despite the promising results of the recent SISR algorithms, compared with original HR image, they still show overly smoothed results and/or lack of high-frequency details, especially on texture regions (see Fig. 1(a)-(c).). In an attempt to resolve this problem, Johnson et al. [21] suggested using perceptual loss, instead of conventional mean squared reconstruction error (MSE), and Ledig et al. [22] proposed a notable SR framework combined with generative adversarial network (SRGAN). Even though quantitative performance of these methods, such as peak signal-to-noise-ratio (PSNR) or structural similarity (SSIM), is inferior to competitive methods, they delivered visually improved HR images.

Meanwhile, Gatys el al. [23] presented a very interesting approach on artistic image generation, called style transfer algorithm. Based on high-level feature maps extracted via pre-trained VGG networks [24], this algorithm synthesizes a style of an artwork to a content image, of an arbitrary photograph, while preserving the structure of the content image.

Inspired by this artistic image generation, we were wondering if we might apply this style transfer algorithm to generate the texture-enhanced image. In other words, if we can obtain a satisfactory level of HR texture image, even if the image is different from the original HR image, we can regard the obtained image as a style image and combine it with the input content image to generate a texture-enhanced HR image.

Based on this motivation, we present a novel texture enhancement framework for SISR. We call our proposed method, HR style transfer (HRST) algorithm. As shown in Fig.1, our proposed HRST algorithm provides more visually pleasing results compared to the representative state-of-the-art SISR method [19].

The remainder of this paper is organized as follows. We introduce the observation of our method in Section II. Section III describes the proposed HRST-based SR framework in detail. In Section IV, we provide experimental results with qualitative and quantitative analyses on 100 texture images. We then discuss the robustness of the proposed method and the detailed method for 4K image SR in Section V. Finally, we draw the conclusion in Section VI.

2. Observation

It is a very challenging problem to recover the finer details of texture regions when we super-resolve at a large up-sampling factor (over $\times 4$). As shown in Fig. 1(a)-(c), the current algorithm only sharpens the lines and edges present in the interpolated LR image. However, it cannot effectively restore the high-frequency details lost by the down-sampling used to create the LR image.

To address this issue, we use the observation that texture regions tend to be comprised of visually very similar patches of various sizes. Based on this idea, we down-scale and repetitively tile the input image. This process creates a texture map we call an HR style image that is very similar in feel to the original HR image (See Fig. 2.). Here, to better correlate the image obtained via down-scaling and tiling with the original HR image, SR version of the interpolated LR image via a SISR method, namely initial HR image, may be utilized as an input image, instead of the interpolated LR image itself.

Then, we take the HR style image and combine it feature-wise with the initial HR image to generate a texture-enhanced HR image. This is different from a simple texture mapping process which just overlays one image onto a different image. Instead, our method searches both the initial HR image and the HR style image from low-level to high-level feature space for similar features. If matches are found, it strengthens them. These features are similar in terms of the correlation in feature space that is invariant to the spatial location, scale or rotation.

3. Proposed Algorithm

Based on the observation detailed above, we propose a texture enhancement framework for SISR, HRST algorithm. As shown in Fig. 3, the proposed framework consists of three steps: (i) initial enhancement to generate an initial HR image from an interpolated LR image via an SISR algorithm, (ii) HR style image generation via down-scaling and tiling, and (iii) texture enhanced HR image generation by combining the HR style image with the initial HR image via a customized style transfer algorithm. In the initial enhancement step, an existing state-of-the art SISR algorithm [19] is adopted to obtain the best initial HR image.

The detail procedures based on the interpolated LR and initial HR images are presented as follow.

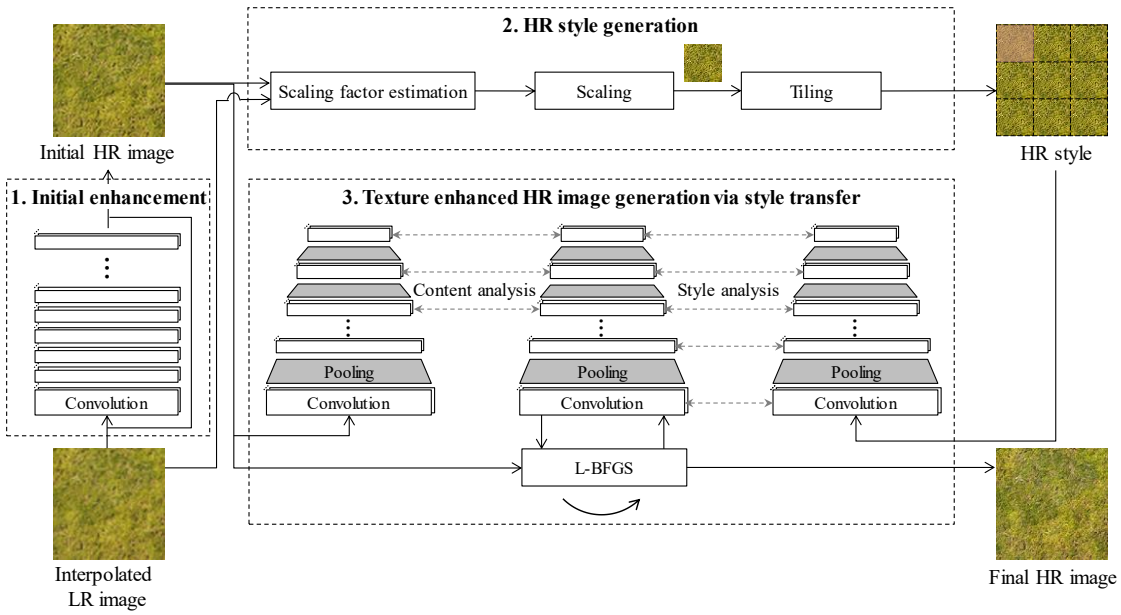

Figure 3. Overall diagram of the proposed texture enhancement framework.

As shown in Fig. 3, the HR style image is generated by down-scaling the initial HR image and then repetitively tiling it into an image of the same size as the HR image. By using the HR style, we generate the final HR image with improved HR texture details, while maintaining the global characteristics of the initial HR image such as location and shape of major structures. To realize this, we adopt the style transfer algorithm [23], and customize it for better performance.

In the customized style transfer algorithm, we mainly perform two adjustments: (i) increase the number of layers used for content loss calculation, and (ii) utilize the initial HR image as an initial estimate for the final HR image.

As in [23], we performs the joint minimization of style and content losses in feature space to obtain the final HR image, $\hat{\mathbf{x}}$ by combining the HR style image with the initial HR image, which can be written as,

$$\hat{\mathbf{x}} = \min_{\mathbf{x}} \alpha L_{\text{style}}(\mathbf{x},\mathbf{s}) + \beta L_{\text{content}}(\mathbf{x},\mathbf{c}). \quad (4)$$

Here, α and β are the weighting factors of style and content losses, L_{style} and L_{content}, respectively. $\mathbf{s}$ and $\mathbf{c}$ are the HR style and the content image respectively for L_{style} and L_{content} calculation. $\mathbf{x}$ denotes the intermediate result image for the final HR image.

For clear description of L_{style} and L_{content} in feature space, we define the feature maps in the l-th layer from an image, $\mathbf{z}$ as $F_{\mathbf{z}}^{l} \in R^{C_l \times N_l}$. Here, C_l is the number of feature maps and N_l is the size of a feature map. $F_{\mathbf{z},ik}^{l}$ (or $F_{\mathbf{z},jk}^{l}$) denotes the i-th (or j-th) feature map at a pixel position k for image $\mathbf{z}$. To extract the feature maps, we utilize the pre-trained VGG-16 network [24], which consists of 13 convolutional and 5 pooling layers.

In L_{style}, to analyze the style of $\mathbf{z}$ in feature space, we utilize the Gram matrix, which can measure correlations between two arbitrary feature maps at a certain layer, written as

$$G_{\mathbf{z},ij}^{l} = \sum_{k} F_{\mathbf{z},ik}^{l} F_{\mathbf{z},jk}^{l}. \quad (5)$$

To force the Gram matrix of $\mathbf{x}$, $G_{\mathbf{x},ij}^{l}$ similar to that of $\mathbf{s}$, $G_{\mathbf{s},ij}^{l}$, the energy functional for the l-th layer can be formulated as below,

$$E_{l,\text{style}} = \frac{1}{4C_l^2 N_l^2} \sum_{i,j} \left(G_{\mathbf{x},ij}^{l} - G_{\mathbf{s},ij}^{l}\right)^2. \quad (6)$$

Since $E_{l,\text{style}}$ compares $G_{\mathbf{x},ij}^{l}$ and $G_{\mathbf{s},ij}^{l}$, unlike $F_{\mathbf{x},ik}^{l}$ and $F_{\mathbf{s},ik}^{l}$, it can allow $F_{\mathbf{x},ik}^{l}$ (or $F_{\mathbf{x},jk}^{l}$) to be different to $F_{\mathbf{s},ik}^{l}$ (or $F_{\mathbf{s},jk}^{l}$), while making only the correlation between $F_{\mathbf{x},ik}^{l}$ and $F_{\mathbf{x},jk}^{l}$ similar to that between $F_{\mathbf{s},ik}^{l}$ and $F_{\mathbf{s},jk}^{l}$. This property can be advantageously exploited in a way that the HR style image obtained through the down-scaling and tiling process is not spatially consistent with the original HR image. Since $E_{l,\text{style}}$ does not make $F_{\mathbf{x},ik}^{l}$ (or $F_{\mathbf{x},jk}^{l}$) similar to $F_{\mathbf{s},ik}^{l}$ (or $F_{\mathbf{s},jk}^{l}$), it can implicitly prevent to transfer spatially-corresponding but unwanted image patterns of the $\mathbf{s}$ to the $\mathbf{x}$. Instead, the $E_{l,\text{style}}$ makes it possible to enhance the corresponding features of $\mathbf{x}$ if only similar correlations in feature space, irrespective of spatial location, scale or rotation, are found in both $\mathbf{x}$ and $\mathbf{s}$.

To reflect all the texture information of the $\mathbf{s}$ from low-level to high-level feature space, we measure the $E_{l,\text{style}}$ for each layer, and take the weighted summation of the energy functionals for L_{style}. This is summarized as,

$$L_{\text{style}}(\mathbf{x},\mathbf{s}) = \sum_{l \in \Omega_{\text{style}}} w_{l,\text{style}} E_{l,\text{style}}. \quad (7)$$

Figure 4. SR results on a forest image with up-sampling factor of 4. (a) The original image, and resultant images (b-g) via (b) bicubic interpolation, (c) Dong, et al.'s [16], [17], (d) Wang, et al. [28].'s, (e) Kim et al.'s [19], (f) the proposed HRST based SR algorithm. (g) the HR style image at a scaling factor of 0.4 used for (f).

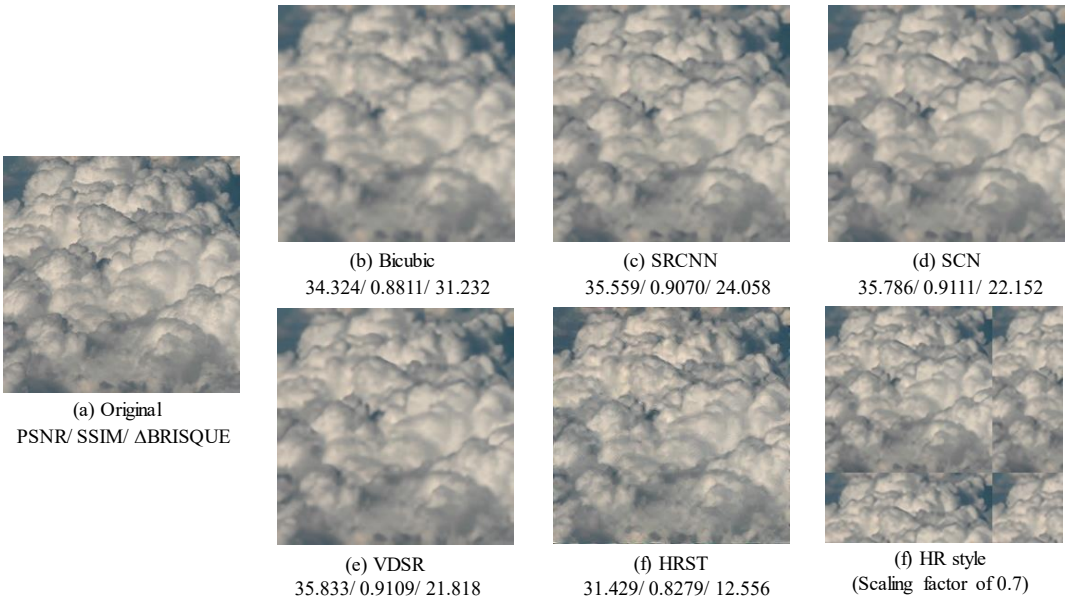

Figure 5. SR results on a forest image with up-sampling factor of 4. (a) The original image, and resultant images (b-g) via (b) bicubic interpolation, (c) Dong, et al.'s [16], [17], (d) Wang, et al. [28].'s, (e) Kim et al.'s [19], (f) the proposed HRST based SR algorithm. (g) the HR style image at a scaling factor of 0.7 used for (f)

Here, $w_{l,\text{style}}$ is the weight factor to adjust contribution of each layer to the style loss, and Ω_{style} is set of layers for style loss calculation.

Meanwhile, to explicitly prevent that $\mathbf{x}$ becomes quite different from $\mathbf{c}$, which is the initial HR image, the energy functional for l-th layer can be formulated as below,

$$E_{l,\text{content}} = \frac{1}{2}\sum_{i,k}\left(F^l_{\mathbf{x},ik} - F^l_{\mathbf{c},ik}\right)^2 , \tag{8}$$

Here, Ω_{style} denotes set of layers for content loss calculation.

To extend the constraints to all ranges from mid-level to high-level feature space, instead of using (8) for L_{content} as in [23], we define L_{content} as below,

$$L_{\text{content}}(\mathbf{x},\mathbf{h};l) = \sum_{l\in\Omega_{\text{content}}} w_{l,\text{content}} E_{l,\text{content}} . \tag{9}$$

Here, $w_{l,\text{content}}$ is the weight factor to adjust contribution of each layer to the content loss, and Ω_{content} is set of layers for content loss calculation.

To find the optimum solution, $\hat{\mathbf{x}}$ in (4), we adopt a representative gradient-based optimization, L-BFGS method [25], which provides the solution with a fast convergence. In this method, the gradient of (4) with respect to x can be determined based on $\partial L_{\text{style}} / \partial \mathbf{x}$ and $\partial L_{\text{content}} / \partial \mathbf{x}$ obtained via a standard error back-propagation [26].

Meanwhile, it should be emphasized that we initialize **x** with **c**, instead of Gaussian random noise that is used in the existing style transfer algorithm [23]. This is an important detail not only to prevent the final HR image from being generated differently each time but also to help to preserve the original structure better. Furthermore, it leads to fast convergence.

4. Experimental Results

We first describe the parameter settings used to obtain the proposed results. The results are obtained by utilizing the style feature maps on 5 convolutional layers, 'conv1', 'conv3', 'conv5', 'conv8', and 'conv11' ($w_{l,\text{style}}$ = 1/5 for those layers), while utilizing the content feature maps on 3 convolutional layers, 'conv7', 'conv10', and 'conv13' ($w_{l,\text{content}}$ = 1/3 for those layers). Using intermediate and higher level layers for the content feature maps helps to maintain the global and apparent structures, while allowing fine structures to be enhanced by the style feature map. The ratio α / β and the number of iterations are set to 1×10^4 and 300, respectively. To verify the performance of the proposed texture enhancement algorithm, we prepared 100 texture images. These images were cropped to a size of 256×256 pixels from 4K images. The scaling factors determined by (3) ranged from 0.4 to 0.75.

We should mention that the images in this section are not same to those used in the subsection 3. All experiments are performed with a down-sampling factor of 4.

For performance comparison, we emphasize that the goal of this work is not to replicate the results of state-of-the-art PSNR or SSIM, but instead to demonstrate the perceptually improved visual quality. To quantify the visual improvement, we measure the difference of BRISQUE [27] metric compared with original HR image, notated as Δ BRISQUE, the metric, which is known to have a high correlation with human subjective evaluation.

We compared the performance of the proposed HRST method to the bicubic-interpolation and the three different methods: the super-resolution CNN (SRCNN) [18], deep networks for super-resolution with sparse prior (SCN) [28], and very deep CNN-based super-resolution (VDSR) [19], which are currently the best performing CNN-based approaches, among the algorithms that have publically released the available code. In addition, for reference, we show the generated HR style images used for the proposed HRST.

Visual comparison of the super-resolved images is given in Figs. 4 and 5. For the images, the selected scaling factors were 0.4 and 0.7, respectively. We can note that the existing SISR algorithms poorly restore fine and detail textures, and generally provide overly-smoothed results, although they successfully enhance coarse and apparent structures. In contrast, the proposed algorithm provides finer and sharper texture representations without introducing noticeable artifacts.

5. Conclusion

In this paper, we present a novel texture enhancement framework for SISR via HR style transfer algorithm. We effectively improve the spatial resolution on the texture regions as well as edge and line regions, which is yet unresolved by existing state-of-the-art SISR algorithms. For the texture enhancement, we first obtain an initial HR image from the interpolated LR image, and then generate the HR style image from the initial HR image via down-scaling and tiling process. By properly combining semantic information of both the HR style and the initial HR images via the customized style transfer algorithm, we finally generate the texture-enhanced HR image. Experimental results demonstrate that the proposed algorithm can provide realistic and more visually pleasing SR images with finer and sharper textures, compared to the existing SR algorithms, without introducing undesirable artifacts.

References

[1] G. Freedman and R. Fattal, "Image and video upscaling from local self-examples," *ACM Trans. Graph*., vol. 30, no.2, pp. 12.1-12.11, 2011.

[2] C. E. Duchon, "Lanczos filtering in one and two dimensions," *J. Appl. Meteorol.*, vol. 18, no. 89, pp. 1016-1022, 1979.

[3] D. Dai, R. Timofte, and L. Van Gool, "Jointly optimized regressors for image super-resolution," *Eurographics*, vol. 34, no. 2, pp. 95-104, 2015.

[4] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael, "Learning low-level vision," *Int. J. Comput. Vis.*, vol. 40, no. 11, pp. 25-47, 2000.

[5] S. Schulter, C. Leistner, and H. Bischof, "Fast and accurate image upscaling with super-resolution forests," in *proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 3791-3799.

[6] R. Timofte, V. De Smet, and L. Van Gool, "Anchored neighborhood regression for fast example-based super- resolution," in *proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 1920-1927.

[7] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang, "Coupled dictionary training for image super-resolution," *IEEE Trans. Image Process.*, vol. 21, no. 11, pp. 3467-3478, 2012.

[8] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 349-356.

[9] C. Y. Yang, J. B. Huang, and M. H. Yang, "Exploiting self-similarities for single frame super-resolution," in *Proc. IEEE Asia Conf. Comput. Vis.*, 2010, pp. 497-510.

[10] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding, in *proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2004, pp. 275-282.

[11] M. Bevilacqua, A. Roumy and M.-L. A. Morel. Low- complexity single-image super-resolution based on nonnegative neighbor embedding," in *proc. British Mach. Vis. Conf.*, 2012, pp. 135.1-135.10.

[12] R. Timofte, V. De Smet, and L. Van Gool., "A+: Adjusted anchored neighborhood regression for fast super-resolution," in *Proc. IEEE Asia Conf. Comput. Vis.*, 2014, pp. 111-126.

[13] J. Yang, J. Wright, T. S. Huang, and Y. Ma., "Image superresolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861-2873, 2010.

[14] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *proc. Int. Conf. Curves Surf.*, 2012, pp. 711-730.

[15] S. Schulter, C. Leistner, and H. Bischof, "Fast and accurate image upscaling with super-resolution forests," in *proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 3791-3799.

[16] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 184-199.

[17] C. Dong, C. C. Loy, K. He, and X. Tang, "Image superresolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295-307, 2015.
[18] 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 *proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 1874-1883.
[19] J. Kim, K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 1646-1654.
[20] J. Kim, K. Lee, and K. M. Lee. "Deeply-recursive convolutional network for image super-resolution," *arXiv preprint arXiv:1511. 04491*, 2015.
[21] J. Johnson, A. Alahi, and F. Li, "Perceptual losses for real-time style transfer and super- resolution," *arXiv preprint arXiv:1603.08155*, 2016.
[22] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," *arXiv preprint arXiv:1609.04802*, 2016.
[23] L. A. Gays, A. S. Ecker, and M Bethge. "Image style transfer using convolutional neural networks, in *proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 2414-2423.
[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[25] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization," *ACM Trans. on Math. Softw.*, vol. 23, no. 4, pp. 550-560, 1997.
[26] Y. A. LeCun, L. Bottou, G. B. Orr, and K. R. Muller, "Efficient backprop," *Neural networks: Tricks of the trade*, 2012.
[27] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp.4695-4708, 2012.
[28] Z. Wang, D. Liu, J. Yang, W. Han, T. Huang, "Deep networks for image super-resolution with sparse prior," in *proc IEEE International Conference on Computer Vision* , 2015, pp. 370-378.

Author Biography

Il Jun Ahn *received the Ph.D. in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, in 2016. Since 2016, he has been with the Samsung Research, Samsung Electronics, where he is currently a Senior engineer.*

Woo Hyun Nam *received the Ph.D. in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, in 2013. Since 2013, he has been with the Samsung Research, Samsung Electronics, where he is currently a Senior engineer.*