Smart Media Journal / Vol.8, No.4 / ISSN:2287-1322
http://dx.doi.org/10.30693/SMJ.2019.8.4.64

Super-resolution in Music Score Images by Instance Normalization

Minh-Trieu Tran*, Guee-Sang Lee**

Abstract

The performance of an OMR (Optical Music Recognition) system is usually determined by the characterizing features of the input music score images. Low resolution is one of the main factors leading to degraded image quality. In this paper, we handle the low-resolution problem using the super-resolution technique. We propose the use of a deep neural network with instance normalization to improve the quality of music score images. We apply instance normalization which has proven to be beneficial in single image enhancement. It works better than batch normalization, which shows the effectiveness of shifting the mean and variance of deep features at the instance level. The proposed method provides an end-to-end mapping technique between the high and low-resolution images respectively. New images are then created, in which the resolution is four times higher than the resolution of the original images. Our model has been evaluated with the dataset "DeepScores" and shows that it outperforms other existing methods.

Keywords : music score images | super-resolution | residual blocks

I. INTRODUCTION

The super-resolution image (SR), the aim of which is to generate a high-resolution (HR) picture from a corresponding low-resolution (LR) image, is a major issue in computer vision. Some articles on image SR have been published by Nasrollahi [1] and Yang [2]. Here our research will only concentrate on single image super-resolution and do not consider approaches that involve retrieving high-resolution images from diverse images[3,4]. Inspired by papers enhancing accurate optical character recognition (OCR) through the super-resolution technique [5,6], we proposed a technique with which to improve the quality of optical music score images, which can be printed type. We created new networks based on Enhanced Deep Residual Networks (EDSR) [7]. However, our networks are deeper than EDSR [7]. We will mention the details of our networks in the following sections. The datasets that we use for training and testing are DeepScores datasets [8]. Deep neural networks [17,21,22] provide improved performance with respect to peak signal to noise ratio (PSNR) in the super-resolution problem. Most existing SR algorithms have different scale factors with the different independent problems focused on by the authors. Our paper only focuses on scale factors ×4 and applying this method for music score images. Instance normalization (IN) [23] was chosen in our architecture because it restrains instance-specific mean value and covariance shift, make the learning process will be simpler. The network architectures that we proposed are deep and lead to better performance for the image super-resolution task. Music score images are a type of handwritten or printed music notation that uses new musical symbols to denote notes for a song or instrumental. However, since the 1980s, access to music notation has included musical notation being presented on computer screens and the

*Student Member, Graduate Student, **Member, Professor, Dept. of ECE, Chonnam National University

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A3B05049058 & NRF-2017R1A4A1015559) and by a grant (HCRI 19 001-1*H2019-0295) Chonnam National University Hwasun Hospital Institute for Biomedical Science.

Manuscript : 2019. 07. 16
Revised : 2019. 08. 13
Confirmation of Publication : 2019. 11. 19

Corresponding Author : Guee-Sang Lee,
e-mail : gslee@jnu.ac.kr

development of scoring computer programs that can record a song or electronic music score. The use of the term "score" or "sheet" distinguishes the forms of written or printed music from sound recordings, radio or television programs, or recording sessions. In everyday use, "music tracks" (or simply "music") may refer to print editions of commercial music combined with the release of a new movie, TV show, and album, as well as recordings or other special or popular events related to music. The first printed music sheet made with a printer was created in 1473. Sheet music is the standard form in which classical music is symbolized so that it can be learned and performed by soloists or musicians or music groups. Many forms of traditional and popular western music are often "heard by ear" by singers and musicians as opposed to using sheet music, although in many cases, traditional music and pop music may also be available in sheet form. There are many studies about optical music recognition with distortion music score image. Nhat studied the system to recognize non-distortion music score images on a smartphone [27]. Jorge Calvo-Zaragoza used convolution neural network to recognize music symbol in his research [28].

In the general field of study about super-resolution image, Tai [9] incorporates a side-by-side super-resolution technique based on a previously described gradient profile [10] with the advantages of detailed learning-based synthesis. Zhang and colleagues [11] proposed a method to capture the same patch image at different scales. Yue [12] took the image correlation with similar content from the internet and proposed a standard that fits the cognitive structure to be used for alignment. There is currently a remarkable research direction with convolutional neural networks (CNN). Super-resolution algorithms based on a convolutional neural network platform provide substantially high performance. One typical algorithm is Learner Iterative Shrinkage and Threshold Algorithm (LISTA) [13]. Dong [14,15] has used the BICUBIC technique to expand the resolution of an input image along with a full depth network used to train from start to finish, creating a super-resolution image with very good efficiency. In image super-resolution technique, calculating the image upscaling through a simple method is the first step. Typically, in this step, the BICUBIC filter and bilinear filter are used. Neural networks will study a way to solve BICUBIC filtering mistakes and generate a higher characteristic super-resolution image. In research of Duchon [16], different models that operate similarly are shown. This model uses new sub-pixel layers, as such a way of seeing is more effective than some models before. Several filters are better than BICUBIC filter as it can be studied during the training process and does not need to do in the high-resolution space, so all tasks can be solved effectively in space of the low-resolution.

Super-resolution applications in optical character recognition systems are a new topic. Ankit Lat [5] has used super-resolution to enhance OCR accuracy. In this paper, we used super-resolution method to make a clear music score image.

The rest of this paper is arranged with five sections. We will present instance normalization [23] in section II. And in section III, the proposed method for image super-resolution is shown. Section IV discusses our experiments in comparison to other methods. Finally, the conclusion is given in section V.

II. INSTANCE NORMALIZATION

There are several commonly used types of normalization such as batch normalization (BN) [24], group normalization [25], and instance normalization [23]. In the batch normalization [24], the mean and standard deviation are calculated along the batch axis (T), and spatial axes (H *and* W) which mean that all pixels of the same channel are normalized together. For instance normalization, the value of mean and standard deviation are calculated along the (H, W) axes for each image and each channel. Normalization Instance is a special case of group normalization when the number of channels per group (Z) equals one. With group normalization, the mean and standard deviation will be calculated along the (H, W) axes and along with a group of Z channels. We selected instance normalization [23] because it restrains instance-specific mean value and covariance shift, make the learning process will be simpler. That leads to the result that it is effective in image dehazing [18].

We assume that x is a tensor of input, tensor x contains a batch of T photos, mean value is denoted by μ, σ is the standard deviation, ε is added to variance to avoid dividing by zero, d and c are the span spatial dimensions, the feature channel is b,

y_{abcd} is the output normalization, x_{abcd} denotes its $abcd$ −th element. Index of the image in the batch is denoted by a. In batch normalization [24], the mean and variance could be denoted as:

$$\mu_b = \frac{1}{HWT}\sum_{a=1}^{T}\sum_{l=1}^{W}\sum_{m=1}^{H} x_{ablm} , \tag{1}$$

$$\sigma_b^2 = \frac{1}{HWT}\sum_{a=1}^{T}\sum_{l=1}^{W}\sum_{m=1}^{H}(x_{ablm} - m \cdot \mu_b)^2, \tag{2}$$

$$y_{abcd} = \frac{x_{abcd} - \mu_b}{\sqrt{\sigma_b^2 + \varepsilon}}. \tag{3}$$

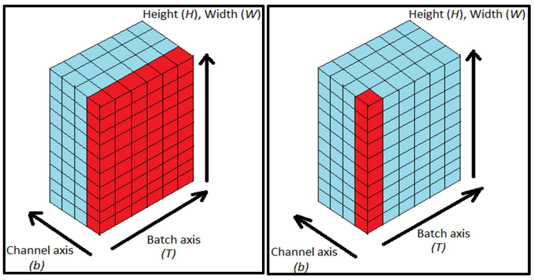
Moreover, the mean and variance of IN are shown as:

$$\mu_{ab} = \frac{1}{HW}\sum_{l=1}^{W}\sum_{m=1}^{H} x_{ablm} , \tag{4}$$

$$\sigma_{ab}^2 = \frac{1}{HW}\sum_{l=1}^{W}\sum_{m=1}^{H}(x_{ablm} - m \cdot \mu_{ab})^2 , \tag{5}$$

$$y_{abcd} = \frac{x_{abcd} - \mu_{ab}}{\sqrt{\sigma_{ab}^2 + \varepsilon}}. \tag{6}$$

The important difference between batch normalization and instance normalization is shown in Fig. 1. In Fig. 1(a), BN applies the normalization technique to all batches of images(red) as contradictory to a single image(red) in IN of Fig. 1(b).



(a) BN (b) IN

Fig. 1. Tensor of the feature maps

III. PROPOSED METHOD

We proposed the use of the super−resolution method with music score images in this section. The resolutions of the output images are four times larger than those of the input low−resolution images, and the details of the notes, symbols, and staff−lines are clear, which is very important in optical music recognition. There are certain methods that can be used to resolve the problem super−resolution well. For example, Ledig et al. [17] applied the ResNet architecture with SRResNet, while Bee Lim et al [7] used the structure with EDSR. In EDSR method, authors removed unnecessary modules from original ResNet architecture, they increased performance while generating their model compact. We made the upsampling part by applying sub−pixel convolution layers. Low−resolution feature maps are created after the last residual block. The image super−resolution is produced by upscaling the low−resolution feature map [26]. The research was focused on the music score image. We proposed a new type of residual blocks, as shown in Fig. 2.

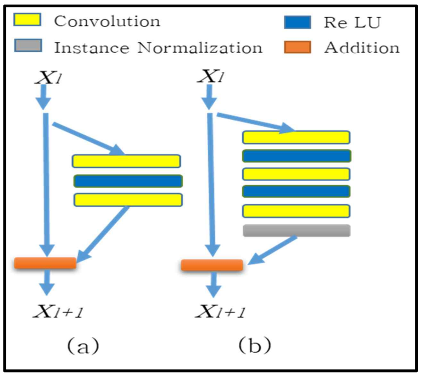


Fig. 2. (a) Residual block of EDSR [7]
(b) Modified residual block

In this figure, we compare the building blocks of model EDSR [7] and our proposed residual block. We added an instance normalization layer, a convolution layer, and a ReLU activation layer for each residual block. Our network has 32 residual blocks and an upscaling factor of ×4, as shown in Fig. 4. The output patch images are shown in Fig. 3, Fig. 5, and Fig. 6. This study only has an effect on the printed music score, we chose x4 scale because this is the popular scale in super−resolution, if we choose larger scale like x8, the risk of information loss will be greater, if we choose smaller scale like x2, recovery task does not make much sense.

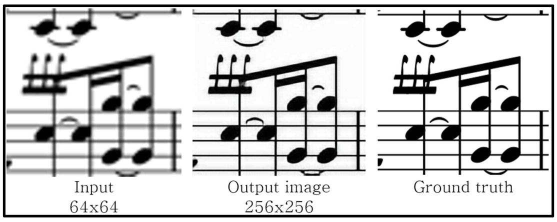


Fig. 3. Input, output, and ground truth images

The $L1$ loss function is chosen because the performance result presents that $L1$ has more advantages than the $L2$ [7], it computes the sum of all differences between predicted (output) value and true (ground truth) value.

$$L1 = \sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\left|A(i,j) - y_{predict\ \ value}(i,j)\right|, \tag{7}$$

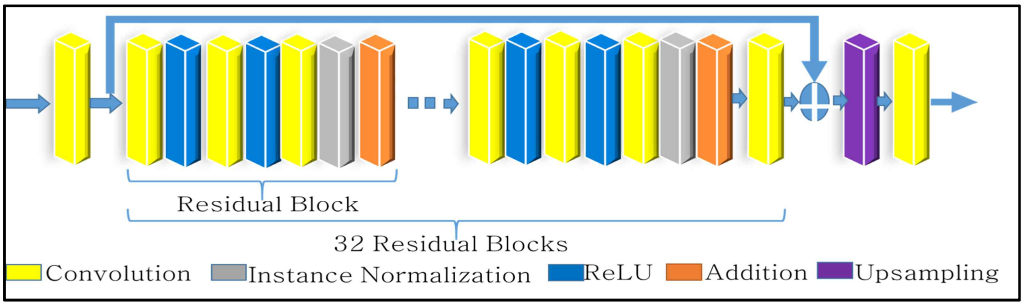


Fig. 4. Architecture of the proposed network

where image *A* has a size of $m \times n$. Our models are evaluated by peak signal−noise ratio (*PSNR*) metric. *PSNR* is defined by the mean squared error (*MSE*) and *MSE* is defined as:

$$M\,SE = \frac{1}{m.n} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [A(i,j) - B(i,j)]^2, \tag{8}$$

$$PSN\,R_{dB} = 10.\,bg_{\;10} \left(\frac{M\,ax\_pixel_{\;A}}{M\,SE}\right)^2, \tag{9}$$

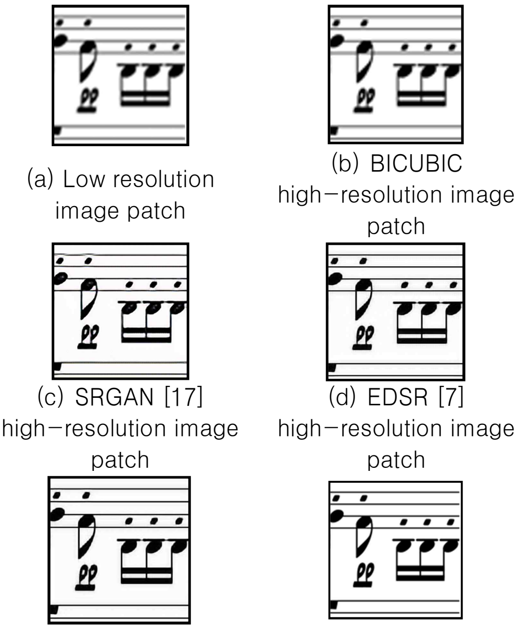
where *B* is a noisy approximation, $M\,ax\_pixel_{\;A}$ is the largest possible pixel value of the photo *A*. Additionally, structural similarity index measurement metric (*SSIM*) [19] is used for evaluating our results. The *SSIM* value is computed on different windows of an image *A*, it is based on three comparison measurements. There are luminance, contrast, and structure corresponding with three windows.

$$SSIM_{\;a,b} = \frac{(2.\mu_a\mu_b + c_1)(2.\sigma_{ab} + c_2)}{(\mu_a^2 + \mu_b^2 + c_1)(\sigma_a^2 + \sigma_b^2 + c_2)}, \tag{10}$$

where a and b are two windows have the same size, the average value for window a is μ_a and for window b is μ_b, σ_{ab} is the covariance for windows a and window b. The variance values for windows a and window b are σ_a^2 and σ_b^2 respectively. With the task of stabilizing the division with a weak denominator, c_1 and c_2 are used.

IV. EXPERIMENTAL RESULTS

In this section, we discuss the datasets, baseline methods, and experimental results regarding the evaluation of our proposed super−resolution model. For the training task, we used color (Red−Green−Blue RGB) input patches. The size of the patch was 64 × 64, we took it from a low−resolution image. The data for training is augmented in two ways. The first one is a horizontal flip and the second is rotation. Our model was trained using the ADAM optimizer [20], and the configuration of parameters was as follows: $\beta 1$ = 0.9000, $\beta 2$=0.9999, $\varepsilon=10^{-8}$, mini−batch size was 16, and learning rate was 10^{-4}. All the databases contain printed music score images. The training set includes 2460 image patches and the data we used for testing contains 534 image patches about notes, staff line, and music symbols. These images are based on DeepScores [8] datasets. About the training time, it takes around 80 hours for training our network structure. The training time can be reduced with a stronger system. We implement the following baseline models that train the data using the SRGAN [17] and EDSR [7] methods. Our architectures show better performance than SRGAN [17] and EDSR [7].



(a) Low resolution image patch

(b) BICUBIC high−resolution image patch

(c) SRGAN [17] high−resolution image patch

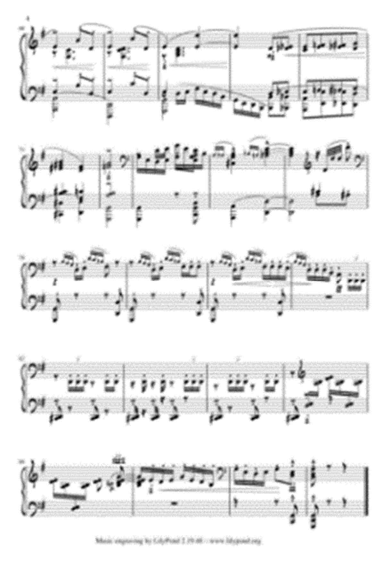
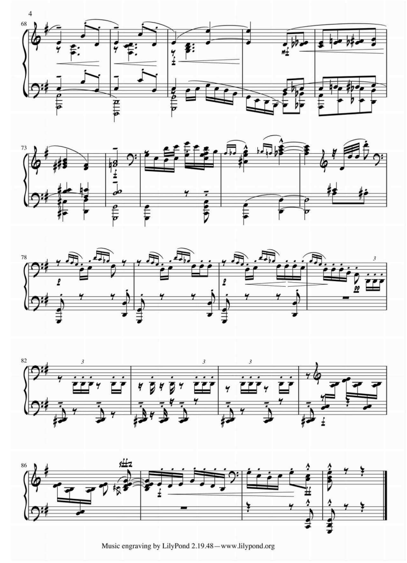
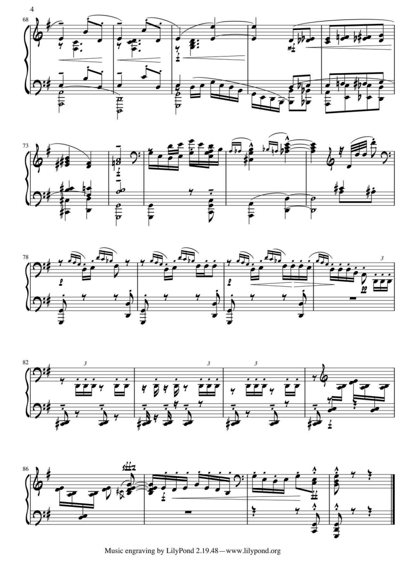
(d) EDSR [7] high−resolution image patch

(e) Our proposed method

(f) Ground truth image patch

Fig. 5. Results of the sample image patch

Table 1. Low−resolution image of the music score, the output image of the music score (upscaling factor is ×4), and ground truth image.

Low resolution	Output image	Ground truth
		

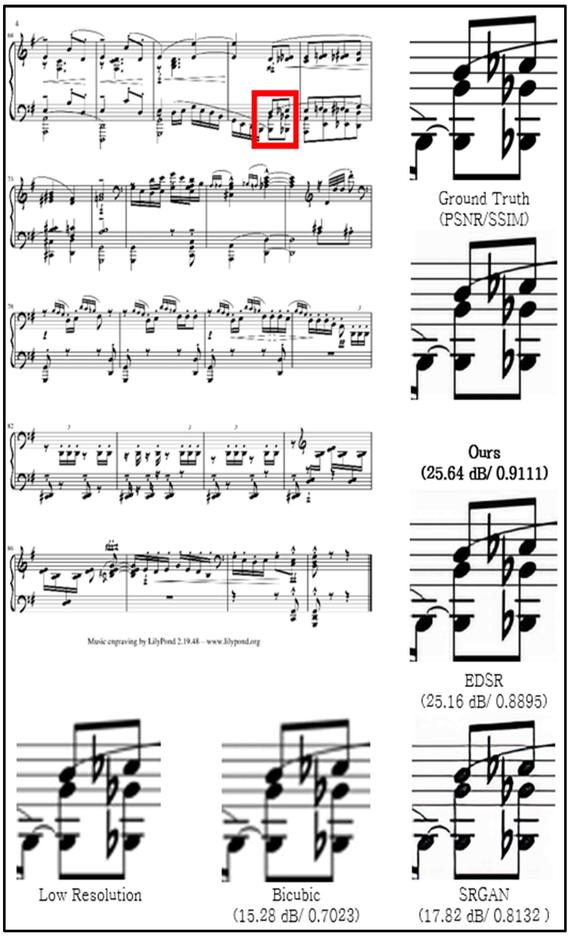


Fig. 6. Super−resolution of our method compared with those of existing algorithms (EDSR [7], SRGAN [17], BICUBIC)

Table 2. Performance comparison between architecture on test set (PSNR (dB)/ SSIM)

Methods	PSNR	SSIM
BICUBIC	18.75 dB	0.8347
SRGAN [17]	21.51 dB	0.8173
EDSR [7]	28.64 dB	0.9327
Our Method	30.88 dB	0.9605

V. CONCLUSION

In this paper, we presented an effective approach to making super−resolution music score images. By adding instance normalization layer in each residual block, we achieve better results than other methods like SRGAN, EDSR. Our proposed model has achieved on printed music score image compared to previous approaches[29]. The resolution images were created with resolutions four times higher than those of the input images, which particularly help improve the quality of the image, thereby increasing the processing performance identification. The music symbols become clearer and no longer blurry. This research opens up many positive opportunities for practical applications in the present as well as in the future and it will contribute to the development of document image processing[30,31].

REFERENCES

[1] Nasrollahi; Moeslund; "Super−resolution: A comprehensive survey," *Machine Vision and Applications*, vol.25, pp.1423−1468, 2014

[2] Yang; Ma; Yang; "Single−image super−resolution: A benchmark," European Conference on Computer Vision (ECCV), vol.4, pp.372−386, 2014

[3] Borman; Stevenson; "Super−Resolution from Image Sequences − A Review," *Midwest Symposium on Circuits and Systems*, pp.374−378, 1998

[4] Farsiu; Robinson; Elad; Milanfar; "Fast and robust multiframe super−resolution," *IEEE Transactions on Image Processing*, vol.13, pp.1327−1344, 2004

[5] Ankit Lat; C. V. Jawahar; "Enhancing OCR Accuracy With Super Resolution," International Conference on Pattern Recognition (ICPR), pp.3162−3167, 2018

[6] S. Baker; T. Kanade; "Limits on super−resolution and how to break them," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.24, pp.1167−1183, 2002

[7] Lim; Bee; Son; Sanghyun; Kim; Heewon; Nah; Seungjun; Lee; Kyoung Mu; "Enhanced Deep Residual Networks for Single Image Super−Resolution," IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp.1132−1140, 2017

[8] Lukas Tuggener; Ismail Elezi; Jurgen Schmidhuber; Marcello Pelillo; Thilo Stadelmann; "DeepScores − a dataset for segmentation, detection and classification of tiny objects," International Conference on Pattern Recognition, pp.3704−3709, 2018

[9] Y.−W. Tai; S. Liu; M. S. Brown; S. Lin; "Super−resolution using Edge Prior and Single Image Detail Synthesis," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2400−2407, 2010

[10] J. Sun; Z. Xu; H.−Y. Shum; "Image super−resolution using gradient profile prior," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008

[11] K. Zhang; X. Gao; D. Tao; X. Li; "Multi−scale dictionary for single image super−resolution," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1114−1121, 2012

[12] H. Yue; X. Sun; J. Yang; F. Wu; "Landmark image super−resolution by retrieving web images," *IEEE Transactions on Image Processing*, vol.22, pp.4865−4878, 2013

[13] K. Gregor; Y. LeCun; "Learning fast approximations of sparse coding," In Proceedings of the 27th International Conference on Machine Learning, pp.399−406, 2010

[14] C. Dong; C. C. Loy; K. He; X. Tang; "Learning a deep convolutional network for image super−resolution," European Conference on Computer Vision, pp.184−199 vol.8692, 2014

[15] C. Dong; C. C. Loy; K. He; X. Tang; "Image super−resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.38, pp.295−307, 2016

[16] C. E. Duchon; Lanczos; "Filtering in One and Two Dimensions," *Journal of Applied Meteorology*, vol.18, pp.1016–1022, 1979

[17] Christian Ledig; Lucas Theis; Ferenc Huszar; Jose Caballero; Andrew Cunningham; Alejandro Acosta; Andrew Aitken; Alykhan Tejani; annes Totz; Zehan Wang; Wenzhe Shi Twitter; "Photo−Realistic Single Image Super−Resolution Using a Generative Adversarial Network," Conference on Computer Vision and Pattern Recognition (CVPR), pp.105−114, 2017

[18] Zheng Xu; Xitong Yang; Xue Li; Xiaoshuai Sun; "Strong Baseline for Single Image Dehazing with Deep Features and Instance Normalization," 29th British Machine Vision Conference, 2018

[19] Zhou Wang; A.C. Bovik; H.R. Sheikh; E.P. Simoncelli; "Image Quality Assessment: From Error Visibility To Structural Similarity," *IEEE Transactions on Image Processing*, vol.13, pp.600−612, 2004

[20] D. Kingma; J. Ba; “Adam: A method for stochastic optimization,” International Conference on Learning Representations (ICLR), vol.abs/1412.6980, 2014

[21] J. Kim; J. Kwon Lee; and K. M. Lee; “Accurate image superresolution using very deep convolutional networks,” Conference on Computer Vision and Pattern Recognition (CVPR), pp.1646-1654, 2016

[22] J. Kim; J. Kwon Lee; K. M. Lee; “Deeply-recursive convolutional network for image super-resolution,” Conference on Computer Vision and Pattern Recognition (CVPR), pp.1637-1645, 2016

[23] Dmitry Ulyanov; Andrea Vedaldi; Victor Lempitsky; “Instance Normalization: The Missing Ingredient for Fast Stylization,” arXiv:1607.08022v3, 2017 (accessed Dec., 24, 2019).

[24] Sergey Ioffe; Christian Szegedy; “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” Proc. of the 32nd International Conference on Machine Learning, vol.37, pp.448-456, 2015

[25] Yuxin Wu; Kaiming He; “Group Normalization,” *International Journal of Computer Vision*, pp.1-14, 2018

[26] Wenzhe Shi; Jose Caballero; Ferenc Huszar; Johannes Totz; Andrew P. Aitken; “Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network,” Conference on Computer Vision and Pattern Recognition (CVPR), 2016

[27] QuangNhat Vo; GueeSang Lee; SooHyung Kim; HyungJeong Yang; “Recognition of Music Scores with Non-Linear Distortions in Mobile Devices,” *Multimedia Tools and Applications*, vol.77, pp.15951-15969, 2018

[28] Jorge Calvo-Zaragoza; Jose J. Valero-Mas; Antonio Pertusa; “End-to-End Optical Music Recognition Using Neural Networks,” Proc, of the 18th International Society for Music Information Retrieval Conference (ISMIR), pp.472-477, 2017

[29] Luu-Ngoc Do; Hyung-Jeong Yang; Soo-Hyung Kim; Guee-Sang Lee; Cong Minh Dinh; “A Covariance-matching-based Model for Musical Symbol Recognition,” *Smart Media Journal*, vol.7, no.2, pp.23-33, 2018

[30] Son Tung Trieu; Guee-Sang Lee; “Machine Printed and Handwritten Text Discrimination in Korean Document Images,” *Smart Media Journal*, vol.5, no.3, pp.30-34, 2016

[31] Van Khien Pham; Soo-Hyung Kim; Hyung-Jeong Yang; Guee-Sang Lee; “Text Detection based on Edge Enhanced Contrast Extremal Region and Tensor Voting in Natural Scene Images,” *Smart Media Journal*, vol.6, no. 4, pp.32-40, 2017

Authors

Minh–Trieu Tran

He received a B.S. degree in Electronics and Telecommunications from Vietnam Aviation Academy, Vietnam in 2016. He received a M.S. degree in Electronics from Hochimin University of Technology and Education in 2018. He is currently a Ph.D student in the dept. of Electronics and Computer Engineering, Chonnam National University, Korea. His interests include deep learning based image processing, computer vision and pattern recognition.

Guee– Sang Lee

He received a B.S. degree in Electrical Engineering and a M.S. degree in Computer Engineering from Seoul National University, Korea in 1980 and 1982, respectively. He received a Ph.D. degree in Computer Science from Pennsylvania State University in 1991. He is currently a professor of the Department of Electronics and Computer Engineering in Chonnam National University, Korea. His research interests are mainly in the field of image processing, computer vision and video technology.