Article

Document Image Restore via SPADE-based Super-Resolution Network

Jaehun Kim and Yoonsik Choe *

Posted Date: 9 January 2023

doi: 10.20944/preprints202301.0031.v1

Keywords: Spatially Adaptive De-normalization (SPADE); Super-Resolution; Convolutional Neural Network; Generative Adversarial Network)

Article

Document Image Restore via SPADE-based Super-Resolution Network

Jaehun Kim and Yoonsik Choe *

Department of Electrical and Electronic Engineering, Yonsei University, Seoul 03722, Korea; dbice@yonsei.ac.kr
* Correspondence: yschoe@yonsei.ac.kr

Abstract: With the development of deep learning technology, various structures and research methods for super-resolution restoration of natural images and document images have been introduced. In particular, a number of recent studies have been conducted and developed in image restoration using generative adversarial network. Super-resolution restoration is ill-posed problem because of some complex restraints such as a lot of high-resolution images being restored for the same low-resolution image and also difficulty in restoring noises like edges, light smudging, and blurring. In this study, we utilized the spatially adaptive de-normalization (SPADE) structure for document image restoration to solve previous problems such as edge unclearness, hardness to catch features of texts, and the image color transition. Consequently, it can be confirmed that the edge of the character and the ambiguous stroke are restored more clearly when contrasting with the other previously suggested methods. Also, the proposed method's PSNR and SSIM scores are geting 8% and 15% higher, respectively, compared to the previous methods.

Keywords: Spatially Adaptive De-normalization (SPADE); Super-Resolution; Convolutional Neural Network; Generative Adversarial Network)

1. Introduction

Ultra-resolution, which restores and extracts low-resolution images as high-resolution images, is a field where many studies are conducted due to its high utilization in society as a whole. Only simple photo images were restored in the past, but the area gradually expanded to be used in fields such as text and video. Meanwhile, the digitization of many documents covered by banks, securities, insurance, and public affairs includes characters in the video to be restored and is sensitive to distortion, loss, and noise caused by transmission tasks such as faxes and scans. Therefore, super-resolution restoration technology can help restore the noise lost in the progress to be almost the same as the original document.

The super-resolution restoration problem is an ill-posed problem that is prematurely difficult to contrast with the original image because it shows various results as shown in Figure 1 depending on the degree and method of restoration for the low-resolution image. However, with the development of deep learning structures such as convolutional neural networks (CNN) and the recent use of generative adversarial networks (GAN), we have tried to solve them by declaring mapping functions by combining low-resolution and high-resolution images in pairs.[1,2,4] Furthermore, we conducted research on strings.[15,16] The above study used a function to replace the mean square error loss, and the concept of perceptual loss was also introduced to optimize the model by identifying features such as the overall texture of photographic images rather than pixels.[2] Through this development process, the quality of the image has been significantly improved.[5]

However, although this theoretical development significantly improved the quality of the image, restoration of edges and detailed features still tended to be difficult. In particular, unlike photographs, the distinction between strokes may be ambiguous, and the distinction between edges and main parts is also a problem that needs to be more careful in characters that are difficult to distinguish.[4] This is because there is a possibility that the phrase in the restored high-resolution image for the low-resolution image of the same single character is different from the existing original. Therefore, in the case of restoring the character, detailed restoration is required compared to the edge problem of the single image described above.[16]

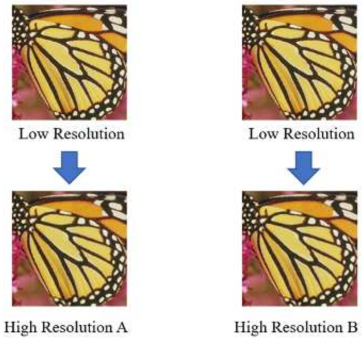

Figure 1. Restored Images: (**a**) High-Resolution A restored from same low-resolution image used to make high-resolution B (**b**) High-Resolution B restored from same low-resolution image used to make high-resolution A. Both of A and B restored from same low-resolution image, however the high-resolution A and B are not the exactly same image

This paper attempts to apply the super-resolution restoration technique using deep neural networks to characters. Existing methods include converting the original high-resolution images into low-resolution images to utilize adversarial generative networks that have recently made significant progress and restoring them to high resolution using deep convolutional neural networks.[6,7] However, for these two methods, we were able to confirm the loss of character feature information, especially during the restoration of edges and characterization of characters are important parts.[8,11] Therefore, it is necessary to restore to high resolution while preserving this information. In this paper, spatial adaptive on normalization is applied to text images.[1] Unlike other methods, this combines the input images into a hierarchical pyramid structure in each layer of the decoder, providing spatial information so that various characteristics can be learned in the zoom generator. Therefore, in the conventional method, it is pretty simple to restore the flat texture of the edge that is easily missed.[1]

2. Related Works

In previous studies, super-resolution restoration usually tends to focus on simple photographs. Several prior studies on documents have also been conducted, but shortcomings have been presented in each method. In addition, unlike the super-resolution restoration of simple photos, where various prior studies have been conducted and sufficient dataset exists, the super-resolution restoration of characters is difficult to prepare dataset first. Therefore, previous studies had no choice but to use their own dataset [11, 15, 16] and as a result, it is difficult to say that their results are enough to say that they achieved learning various languages. All previous studies had to learn only language and fonts limited to self-produced dataset, and the number of them is relatively small. [11,15,16]

2.1. Super-Resolution Convolutional Neural Network

Among the studies that initially attempted to restore documents to ultra-resolution, studies using CNN showed advanced values in Peak Signal-to-noise Ratio (PSNR) compared to the previously used bicubic interpolation.[15] SRCNN applied a simple CNN structure to the field of super-resolution, showing performance beyond all traditional techniques and becoming the beginning of the subsequent super-resolution techniques.[3] As shown in Figure 1, the super-resolution restoration problem is an ill-posed problem in which the answer is not determined. Consequently, conventional state-of-the-art (SOTA) techniques have sought to determine the correct answer to some degree of high-resolution images through prior knowledge. First, there is an example-based method that builds a pre-learning function that maps low-resolution image and high-resolution image patch pairs. Second is a method of preprocessing a low-resolution image in advance in a sparse-coding-based method, performing an encoding process with a sparse coefficient and restoring it through a high-resolution dictionary. In contrast, by relating the super-resolution problem to CNN, we directly designed a CNN structure that learns low-resolution and high-resolution images by passing end-to-end mapping, as shown in Figure 2.[15]

This method has the advantage of processing all prior knowledge within the convolutional layer, as illustrated in Figure 3, and the weight of the model itself is light, so there is no significant difference in time consumption. In addition, it was found that the performance of PSNR indicators was higher than that of the limited dataset.[15]

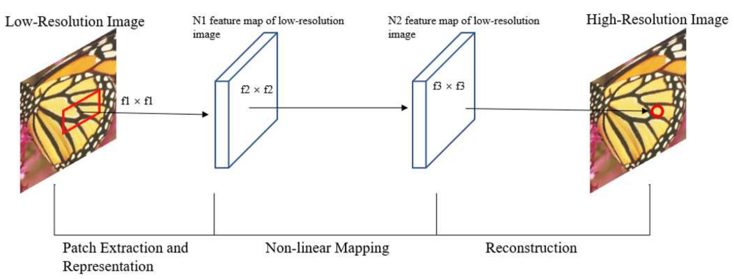
Low-Resolution Image
N1 feature map of low-resolution image
N2 feature map of low-resolution image
High-Resolution Image
f1 × f1
f2 × f2
f3 × f3
Patch Extraction and Representation
Non-linear Mapping
Reconstruction

Figure 2. Structure of Super-Resolution Convolutional Neural Network: the first convolutional layer of the SRCNN extracts a set of feature maps. The second layer maps these feature maps nonlinearly to high-resolution patch representations. The last layer combines the predictions within a spatial neighbourhood to produce the final high-resolution image[3]

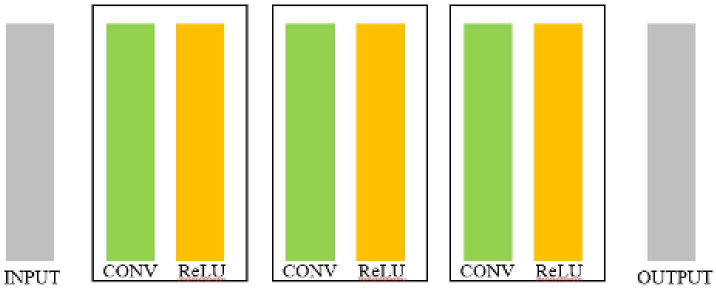

Figure 3. The Structure of SRCNN: As Figure 2 mentioned about the works of each layer do in SRCNN structure, more specifically this structure use ReLU as there activation function.[3]

2.2. Super-Resolution Generative Adversarial Network

It is clear that the aforementioned SRCNN has made progress in the field of super resolution, but the problem of whether it is possible to restore more detailed texture is still presented.[2] Although loss function is important in solving the super-resolution problem, it is limited to complement such parts as high textured detail because the previously used mean squared error (MSE) method lacks visual satisfaction with high-frequency detail and both MSE and PSNR are defined based on the differences in pixel-wise image.[2] Figure 4 shows the structure of GAN that supplemented this. The significance of this previous study is to propose a permanent loss function including adversarial loss and content loss, and adversarial loss is related to the training of the discriminator network with the structure below in Figure 4. Therefore, the super-resolution GAN with this structure can recover the heavily down-sampled image.[2]

It is true that SRGAN exhibits superior super-resolution performance than SRCNN, but there is

doi:10.20944/preprints202301.0031.v1

still a challenging constraints in capturing detailed textures. In addition, there is a problem in that the color of the image changes relatively due to the processing of the edge portion, or the overall image quality appears to be impaired.[1] In the process of restoring documents to super-resolution, not just photographs, it is important to restore detailed edges and find features.

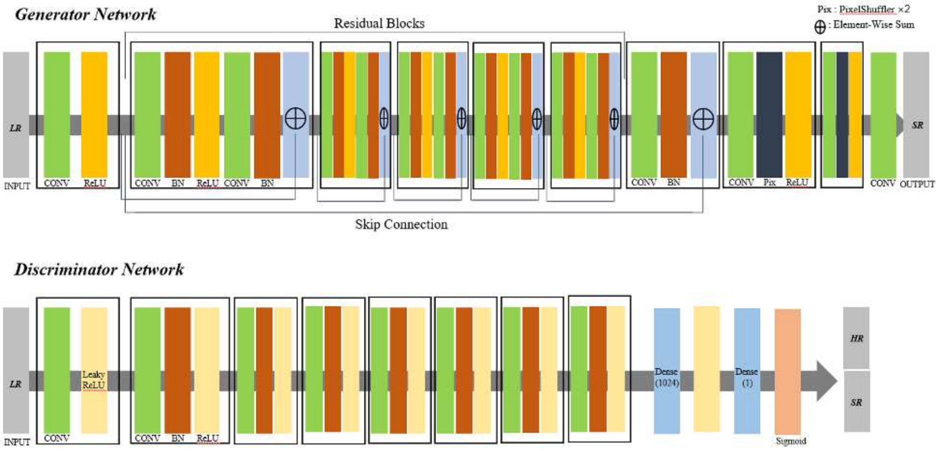

Figure 4. Architecture of SRGAN: The SRGAN structure consists with Discriminator model and Generator model. Each of them have own kernel size, number of feature maps and strides.[2]

3. The Proposed Method

3.1. Single Image Super Resolution

Super-resolution restoration of a single image restores a high-resolution image with one image, and the result may not be unique. Therefore, the original image is defined in Figure 5, and it is made into a low-resolution image through down sampling and noise addition. Thereafter, a model recovered it to a high-resolution original image. In this paper, we used the bicubic interpolation, which is representatively adapted in the low-resolution generation.

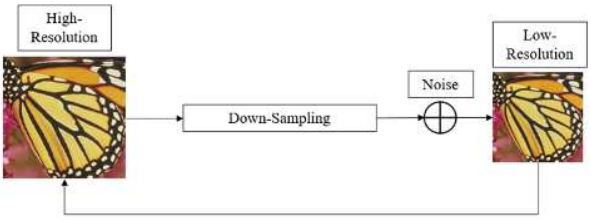

Figure 5. Single image super resolution structure

3.2. Super-Resolution of Low-Resolution Image

To restore high-resolution images from low-resolution images, we used domain mapping of two pair-defined images. By defining the domain and high-resolution domain of low-resolution images, we attempted to restore single-image super resolution through image conversion between low-resolution domains.

$$F^*(\alpha) = N() \tag{1}$$

In Equation 1, $F^*(\alpha)$ is a function of mapping a high-resolution output image from a low-resolution input image, which means image conversion from a low-resolution domain to a high-resolution domain.

Furthermore, $N()$ is a neural network proposed in this paper, consisting of conditional adversarial generation networks.

3.3. Spatially Adaptive De-Normalization Network

As a network for converting high-resolution images from low-resolution images, high-resolution images are estimated from low-resolution input images through a conditional adversarial generation network.

First of all, a typical adversarial generation network consists of a generator G that receives data distribution and generates data close to reality, and a discriminator D that determines whether the input data is real or made from the generative model G, all of which are nonlinearly mapped. Accordingly, the objective function $L_{GAN}(G, D)$ is expressed by the following equation.

$$L_{GAN}(G, D) = min_G max_D[V(D, G) = E_x[\log D(x)] + E_y[\log(1 - D(G(Z)))] \tag{2}$$

In Equation 2, x means data, and z means any noise variable. The purpose of discriminator D is to determine whether the input data is the original or generated by generator G, so it is aims to label the data accurately. The purpose of generator G is to increase both the probability that the determinator D makes a mistake, i.e., the probability that D will be determined to be the data given by the original or the probability of that from G. These two networks G and D are learned with different goals through adversarial learning.

Conditional adversarial generation networks can create conditional generation models with the condition of additional information y at generator G and discriminator D. In this case, y may be a class label or other distribution of data, and y may be additionally added to the input layers of G and D. In generator G, noise z and condition y are combined, and in discriminator D, data x and condition y are entered as the inputs. This is expressed by the formula as follows.

$$L_{CGAN}(G, D) = min_G maxD[V(D, G) = E_x[\log D(x|y)] + E_y[\log(1 - D(G(z|y)))]] \tag{3}$$

Conditional adversarial generation network is constructed from condition y data paired with data x. Unlike the adversarial generation network, noise z and condition y are the inputs of the generator G , and the paired condition y with the data x are the inputs of the discriminator D. This allows generator G of the conditional adversarial generation network to create images that match a pair of data x and condition y, and discriminator D must determine whether the input and output images are correctly made in pairs that satisfy the condition. That is, the conditional adversarial generation network may acquire an image satisfying the corresponding condition.

Based on this, we design a conditional adversarial generation network for super-resolution restoration from a single image. Accordingly, the objective function is expressed by the following equation.

$$[L_{our}(G, D) = E_x[\log D(H|L)] + E_z[\log(1 - D(G(z|L)))] \tag{4}$$

In Equation 4, L means an input low-resolution image and H means a high-resolution image corresponding thereto. Therefore, the proposed objective function allows us to learn a function F^* that maps the input low-resolution image L to the corresponding high-resolution image H via a conditional adversarial generation network.

3.4. Document Single Image Super-Resolution via Spatially Adaptive De-normalization

Through the conditional adversarial generation network proposed in Section 2.2, the orientation of the output can be adjusted by the input. In the structure of the conditional adversarial generation network, the convergence rate of learning was accelerated by using a batch normalization layer. In deep neural networks, batch normalization layers are generally important and the presence or absence of this batch normalization can cause a large difference in performance. In this case, the batch normalization layer is a process of calculating the average and variance of the characteristic vectors of the previous hidden layer and then normalizing them. However, when generator G must contain complex structural features and high-frequency components, such as super-resolution restoration, this batch normalization rather adversely affects high-resolution image generation. In this paper, we rely on a single input low-resolution image for learning. This results in spatial loss and loss of important high-frequency components contained in the input low-resolution image through the batch normalization process, which eventually fails to draw correct inference. To solve this problem, spatially adaptive de-normalization [8] is utilized. Based on this, important components of low-resolution images are preserved and spatial loss is minimized.

As shown in Figure 6, if the characteristic space from generator G is H, the spatial adaptive de-normalization technique S can be expressed in the form of element- wise product or element agreement of γ and β. This is expressed as a formula as follows.

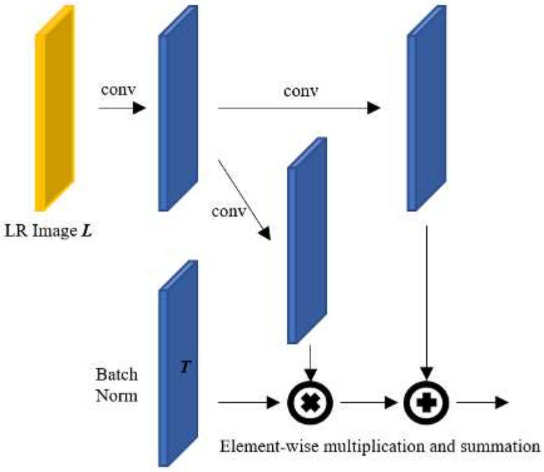

Figure 6. Spatially adaptive de-normalization structure for single image super resolution

$$[S(L,T) = \gamma(L) \times_{element} \frac{(T-\mu_T)}{\sigma_T} +_{element} \beta(L)] \tag{5}$$

μ_T and σ_T are the mean and standard deviation considering all the elements of each characteristic space channel, and $\times_{element}\frac{T-\mu_T}{\sigma_T}$ and $+_{element}\beta(L)$ represent the product of each element and the sum of each element, respectively. At this time, the product of each element and the sum of each element are calculated for each element of each channel. Since T, γ, and β are tensors with the same channel depth, the product of each element and the sum of each element are possible. In addition, as shown in the figure, the spatial adaptive non-normalization technique normalizes the characteristic space T and then combines the information of the characteristic space of the input low-resolution image into the product of each element and the sum of each element, respectively. As previously explained, spatial feature information and high-frequency information are lost due to the general batch normalization process, and through this technique, features with various spatial characteristics can be reflected in the decoder. This is similar to analyzing images over different sizes, such as image pyramid shapes, and identifying features of each size. Through this, it is possible to take the form of a hierarchical structure, which means that both structural features in a wide range of images and detailed features in a narrow range of images can be restored in super-resolution. Figure 7 presents the overall network structure of the proposed method. In the entire network, low-resolution images are converted into input, the encoder into original size input, and the decoder into hierarchical image size as mentioned above and used as input. After that, the high-resolution image generated by the discriminator and the high-resolution image of the original is used to determine whether it is real or fake.

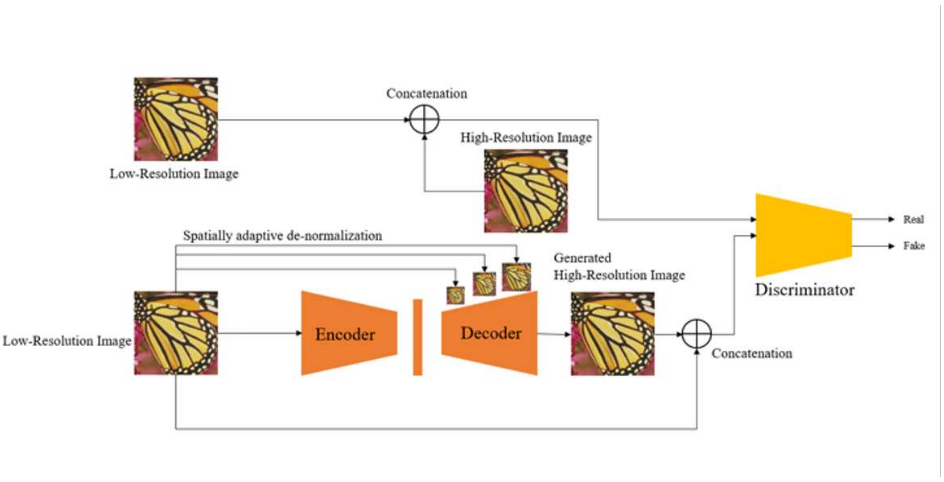

Figure 7. The overall structure of single image super-resolution via spatially adaptive de-normalization: (**a**) This proposed structure with SPADE-GAN has differences compare to ordinary super-resolution methods in Decoder by using hierarchical image inputs to get feature maps (**b**) With new method the discriminator D compare the fake image and real image.

4. Experiments and Results

The PC environment used in this paper consisted of an Intel Core i9 CPU and a GTX 1080Ti 16GB GPU. Both the proposed method and the method for comparison were implemented through the Pytorch environment. Before conducting the experiment with text images, we used DIV2K dataset [17] as a data set for pre-training, released in 2017 CVPR's "New Trends in Image Restoration and Enhancement Workshop", and selected and compared SRNN [15] and SRGAN [2]. For the comparison between the existing method and the method used in this paper, 800 pairs of data were used as a training set, 100 pairs were used as a validation set, and 100 pairs were used as a test set. A total of 300 epochs were conducted for training, and the learning rate was adjusted by the Adam method as the optimization method. However, in the case of this dataset, there were no photos of documents or characters, so additional 500 pairs of self-produced datasets were attached to conduct learning. Therefore, 300 pairs of training set and 150 pairs of validation and training set were pre-trained on a total of 500 pairs of dataset. In addition, in the case of OCR, which is the main purpose of text super-resolution restoration, there is no significant difference in efficiency for documents with image quality between 300 and 600 dpi. However, for documents between 75 and 100 dpi, its efficiency drops sharply. [11, 15, 16] Therefore, the main target of super-resolution restoration of document images that have undergone a process such as scanning is documents of 75 dpi or less. In this paper, after setting the original image of the same document image, the resolution was lowered to 72 dpi to generate a low-resolution image, and then a comparative analysis was performed. The Peak Signal-to-noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) were used as objective indicators for performance identification. These two indicators are typically used to determine the similarity of images, PSNR comparing the size between image pixels and SSIM determining the structural similarity of images.

Figure 8 shows the results of the ultra-high resolution restoration technique of the document image through each method. In the case of high-resolution images restored with SRCNN, it was difficult to accurately identify the characters after relatively high-resolution restoration and showed a spreading appearance. The results of high-resolution restoration through SRGAN allow natural edge restoration and stroke separation compared to SRCNN. But not only does the color of the image change due to the formation of a grid on the edge [1], but the overall image quality also appears to be

doi:10.20944/preprints202301.0031.v1

degraded.[2] On the other hand, we show that the method presented in this paper uses a conditional adversarial generation network to show superior image quality problems and fine edge planning restoration, to preserve characteristic space features in both images and characters. In addition to the results in Figure 8, Table 1 summarizes the quantitatively judged PSNR and SSIM indicators, showing that the quantitative indicators are high in the order presented in this paper, i.e. SRGAN to SRCNN. In particular, the commonly used quantitative indicators of SRCNN and SRGAN show similar patterns, but it can be seen that the method presented in this paper shows the highest figure. This means that the method presented in this paper represents the best quality super-resolution image even as a quantitative indicator outside the visible area. Through Table 2, the learning time for each structure can also be confirmed. As previously described, the advantage of SRCNN was that learning was possible in a relatively short time [11]. However, it can be confirmed that the method presented in this paper is slightly faster in the same dataset and PC environment. This paper shows that our method is more efficient with less weight in the learning environment.

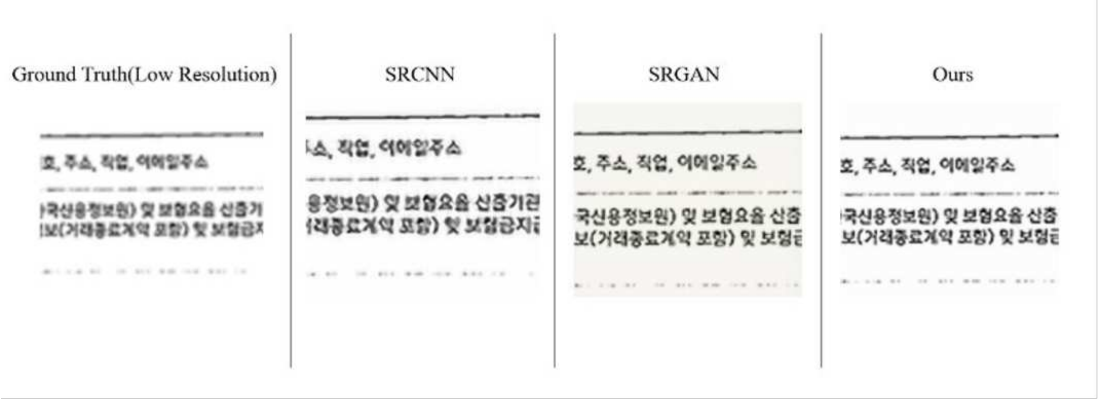
Ground Truth(Low Resolution)
SRCNN
SRGAN
Ours
호, 주소, 직업, 이메일주소
국신용정보원) 및 보험요율 산출
보(거래종료계약 포함) 및 보험금

Figure 8. Spatially adaptive de-normalization structure for single image super resolution

Table 1. Results of each models : Compare to ordinary methods, the proposed method have more higher score both in PSNR and SSIM

	SRCNN	SRGAN	Ours
PSNR[dB]	26.25	26.98	28.32
SSIM	0.7145	0.7319	0.8217

Table 2. Speed of each models : With same environment the proposed method have high-quality speed compare to ordinary methods

	SRCNN	SRGAN	Ours
Time[min]	90	70	78

5. Conclusion

In this paper, we proposed a technique for restoring ultra-high-resolution images through spatial adaptive de-normalization from a single image. We connected the existing ultra-high-resolution problems with low-resolution and high-resolution domain conversions to solved them and used convolutional neural network and adversarial generation networks. In addition, spatial adaptive non-normalization techniques were applied to prevent spatial information loss and image feature loss caused by a structure using a batch normalization layer. As a result, we achieved ultra-high-resolution restoration that preserves detailed textures and edges within the document image, which showed higher performance figures qualitatively and quantitatively than conventional methods. Also, this paper showed that this proposed method can be adapt not only for the natural image, but also to the document images. However, since the detailed texture and edge restoration still do not match the original high-resolution image, we will analyze the hierarchical information of the input image and upgrade it to apply it to further enhance the performance index.

Author Contributions: Conceptualization, Y.Choe and J.Kim; methodology, Y.Choe and J.Kim; software, J.Kim; validation, Y.Choe and J.Kim; formal analysis, J.Kim; investigation, J.Kim; resources, Y.Choe; data curation, J.Kim; writing—original draft preparation, J.Kim; writing—review and editing, Y.Choe; visualization, J.Kim; supervision,Y.Choe; project administration, Y.Choe; funding acquisition, Y.Choe All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by Ministry of Culture, Sports and Tourism and Korea Creative Content Agency(Project Number: R2020040238).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yoon, Jongsu, Taehyeon Kim, and Yoonsik Choe. "Gan based single image super-resolution via spatially adaptive de-normalization." *Transactions of the Korean Institute of Electrical Engineers*, **2021**, pp. 402-407.
2. Ledig, Christian, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken et al. "Photo-realistic single image super-resolution using a generative adversarial network." *Proceedings of the IEEE conference on computer vision and pattern recognition*, **2017**, pp. 4681-4690.
3. Dong, Chao, Chen Change Loy, Kaiming He, and Xiaoou Tang. "Learning a deep convolutional network for image super-resolution." *European conference on computer vision*, **2014**, pp. 184-199.
4. Jo, Y. J., K. M. Bae, and J. Y. Park. "Research trends of generative adversarial networks and image generation and translation." *Electronics and Telecommunications Trends 35*, **2020**, no. 4, pp. 91-102.
5. Jeong, Woojin, Bok Gyu Han, Dong Seok Lee, Byung In Choi, and Young Shik Moon. "Study of Efficient Network Structure for Real-time Image Super-Resolution." *Journal of Internet Computing and Services 19*, **2018**, no. 4, pp. 45-52.
6. Dong, Chao, Chen Change Loy, Kaiming He, and Xiaoou Tang. "Image super-resolution using deep convolutional networks." *IEEE transactions on pattern analysis and machine intelligence 38*, **2015**, no. 2, pp. 295-307.
7. Park, Taesung, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. "GauGAN: semantic image synthesis with spatially adaptive normalization." *ACM SIGGRAPH 2019 Real-Time Live*, **2019**, pp. 1-1.
8. Wang, Ting-Chun, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. "High-resolution image synthesis and semantic manipulation with conditional gans." *Proceedings of the IEEE conference on computer vision and pattern recognition*, **2018**, pp. 8798-8807.
9. Yang, Jianchao, John Wright, Thomas Huang, and Yi Ma. "Image super-resolution as sparse representation of raw image patches." *IEEE conference on computer vision and pattern recognition*, **2008**, pp. 1-8.
10. Yang, Jianchao, John Wright, Thomas S. Huang, and Yi Ma. "Image super-resolution via sparse representation." *IEEE transactions on image processing 19*, **2010**, no. 11, pp. 2861-2873.
11. Pandey, Ram Krishna, K. Vignesh, and A. G. Ramakrishnan. "Binary document image super resolution for improved readability and OCR performance." *arXiv preprint arXiv:1812.02475*, **2018**.
12. Farsiu, Sina, M. Dirk Robinson, Michael Elad, and Peyman Milanfar. "Fast and robust multiframe super resolution." *IEEE transactions on image processing 13*, **2004**, no. 10, pp. 1327-1344.
13. Irani, Michal, and Shmuel Peleg. "Improving resolution by image registration." *CVGIP: Graphical models and image processing 53*, **1991**, no. 3, pp. 231-239.
14. Park, Sung Cheol, Min Kyu Park, and Moon Gi Kang. "Super-resolution image reconstruction: a technical overview." *IEEE signal processing magazine 20*, **2003**, no. 3, pp. 21-36.
15. Pandey, Ram Krishna, and A. G. Ramakrishnan. "Efficient document-image super-resolution using convolutional neural network." *Sādhanā 43*, **2018**, no. 2, pp. 1-6.
16. Lat, Ankit, and C. V. Jawahar. "Enhancing OCR accuracy with super resolution." *2018 24th International Conference on Pattern Recognition (ICPR)*, **2018**, pp. 3162-3167.
17. Agustsson, Eirikur, and Radu Timofte. "Ntire 2017 challenge on single image super-resolution: Dataset and study." *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, **2017**, pp. 126-135.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.