Super-Resolution using Dense Attention Network

Jiangping Gao u6987832 | **Shidong Pan** u6342277 | **Weifan Jiang** u6683698 | **Yu Wen** u7024019

Abstract

Super-resolution based on deep convolutional networks has been rapidly developed recently, and various networks using different methods are emerging in an endless stream. In this article, we propose a dense attention network aiming to deal with the super-resolution task, and achieve outstanding performance. The major improvement of our model lies in two aspects, one is the adoption of the Dense to Dense model and the other is the application of the attention mechanism. Compared with the latest methods on the benchmark dataset, our proposed method shows excellent performance.

1 Introduction

Before the super-resolution was widely researched, to gain images with better visual quality, interpolation techniques that base on sampling theory were generally used on many tasks [1, 14, 21]. Specifically, there are multiple interpolation operations such as bilinear, bicubic, nearest-neighbour, etc. However, none of them can bring more information about the image, which inevitably affects the quality of the image. Thus, a better solution must be non-manual, and have a higher resolution with refined details. With the development of deep learning, it naturally takes over in the super-resolution field.

Generating an image with higher resolution can be performed by using either a single image or multiple images in the same scenario, however, in practice, researchers mainly focus on single image super-resolution (SISR). Specifically, SISR refers to the reconstruction of the corresponding high-resolution image from the observed low-resolution image, which has significant application value in the fields of monitoring equipment, satellite images, medical images and historical images restoration. In this report, when we discuss SR, it indicates the SISR scenario.

SR-related methods usually gain extra visual information from following two aspects, one is pixel density, and another is high-frequency information. Pixel density is well understood, referring to the number of pixels per inch of the image; The SR method is definitely to increase the pixel density. On the other hand, high-frequency information is based on increasing pixel density. When adding pixels, the values of these pixels must reflect the correct information of the picture as much as possible. Furthermore, the SR method is to find the most accurate pixel value method.

Although methods such as SRCNN [7], VDSR[12] and EDSR[15] successfully solved SISR with a decent performance, they did not completely exploit the potential of ResNet structure. As mentioned by Huang et al. [11], to ensure maximum information flow between layers in the network, directly connecting all layers can preserve its feed-forward nature. Meanwhile, each layer can obtain additional inputs from all preceding layers.

To overcome these disadvantages, based on the EDSR architecture, firstly, we introduce the idea of dense connections in Huang et al. [11] to ensure maximum information flow between layers. Furthermore, many previous works [4, 10, 16] showed that Attention mechanism exhibits a remarkable performance on natural language process (NLP) field, and the core concept of attention mechanism in neural network is capturing information from features map with different weights. Intuitively, in an image, the pattern frequencies at different areas are highly different, including colors and shapes,

thus we are naturally inspired to apply attention mechanism on this computer vision task. Above all, we combine attention mechanisms with dense blocks to explore better performance.

We train our models on DIV2K[20] dataset and evaluate our models on the Set5, Set14 and BSD100 datasets and PSNR and SSIM are used as our evaluation metrics.

2 Related Work

2.1 SRCNN

Super-Resolution Convolutional Neural Network(SRCNN) [7] is an essential milestone in the history of super-resolution. It is the first method to apply a neural network for super-resolution. SRCNN needs to go through a pre-processing bicubic interpolation before putting the image into the neural network which turns the original image into the desired high-resolution image.

Based on the relationship between deep learning and traditional sparse coding, SRCNN divides the three-layer network into Patch extraction and representation (9 x 9 x 64 convolution kernel), Non-linear mapping (1 x 1 x 32 convolution kernel), and finally reconstruction (5 x 5 x 1 convolution kernel). The loss function used in SRCNN is a simple mean square error (MSE), which is defined as the difference between each pixel of the reconstructed photo and each pixel of the real picture. However, as mentioned by Lim et al. [15], ResNet was proposed for higher-level computer vision tasks such as image classification, which means it is not appropriate by being directly adapted in super-resolution which is always considered as a kind of low-level computer vision task. We remove the batch normalization and ReLU activation parts to improve the performance.

SRCNN also has some shortcomings. When it is trained for a single scale factor, if a new scale is required, the new model must be trained. Feature extraction only uses one layer of a convolutional layer, which has the problem of relatively small receptive fields, so the extracted features are very local features, which will cause the details to be unable to be recovered.

2.2 VDSR

Very Deep Super-Resolution (VDSR)[12] inspired by the VGG-Net[19], which mostly is applied in image recognition uses deep networks to improve the accuracy of results, while it avoids slow convergence in the deep network. To achieve these, VDSR makes the number of network layers to 20 with the 3*3 convolution kernel and using a larger receptive field to take more image context. As most information in both low resolution images and high resolution images is the same, learning differences (residual) between two images will be more efficient. By employing the residual image, VDSR provides an adjustable gradient clipping to allow a high learning rate, which can accelerate the training speed, and get away from the gradient explosion at the same time. With multi-scale input images, VDSR has a flexible ability to handle multi-scale super-resolution problems with a single model. However, similar as SRCNN, VDSR also needs to pre-process bicubic interpolation of input images, which needs amount of memory and time. We use transposed convolution up-sampling to overcome this limitation.

2.3 EDSR

Enhanced Deep Residual Networks (EDSR)[15] changed ResNet's architecture and trains architecture with L1 loss to improve the model performance. What is more, by applying trained model arguments of small super-resolution to initialize arguments in larger super-resolution models, EDSR reduces training time and achieves better performance. Similar to VDSR, EDSR also handles the multi-scale super-resolution problem with proposing a method called multi-scale super resolution network (MDSR)[15]. MDSR model adds scale-related layers only on input and output parts to parallelly process scale-dependent information. Instead of using residual blocks, we use dense blocks to ensure the maximum information flow between layers for exploring better performance.

2.4 RCAN

After the introduction of residual blocks in ResNet, many major improvements in image super-resolution are based on the depth of network representation. However, most CNNs treat each channel

of features equally, which undoubtedly lacks the flexibility to process different types of information. This time, the network lacks the ability to discriminate and learn. Therefore, the Residual Channel Attention Network (RCAN)[23] proposes a residual in residual (RIR) architecture, in which residual group (RG) is used as the basic module, and long skip connection (LSC) is used for rough residual learning. In each RG ,it superimposes a few simple residual blocks and short jump connections (SSC). After that, RCAN introduced the channel attention mechanism (CAM) into the super-resolution task, and adaptively rescaled the features of each channel by modeling the interdependence between feature channels. Inspired from RCAN, we apply attention mechanisms in our dense blocks to ensure the interdependence between each blocks.

3 Proposed Methods

In this section, we mainly introduce the proposed model architecture. Firstly, we generally review and analysis recently published super-resolution networks.Then we suggest dense blocks with attention (DBA) structure as basic model which reconstructs super-resolution images in a single model.

3.1 Dense Blocks

Many previous works have showed that residual networks[9] can successfully solve both low-level and high-level tasks on computer vision field. According to Ledig et al. [13], they proposed SRResNet architecture which is based on ResNet to set a new state-of-the-art of super-resolution problem. In addition, Lim et al. [15] made some modifications compared with SRResNet, specifically batch normalization (BN) layers are removed. To fully exploit the potential of network, inspired by Huang et al. [11], we apply dense connection to replace residual connection to explore a possible better performance on super resolution task.

In Figure 1, original ResNet [11], EDSR [15], and our proposed densely-connect architecture are compared. Previous works [15, 18] show that removing BN layers can not only increase the range flexibility from networks, but also remarkably reduce GPU memory usage by approximately 40%. Hence, a more efficient computational performance can be expected under limited resources.

3.2 Attention Mechanism

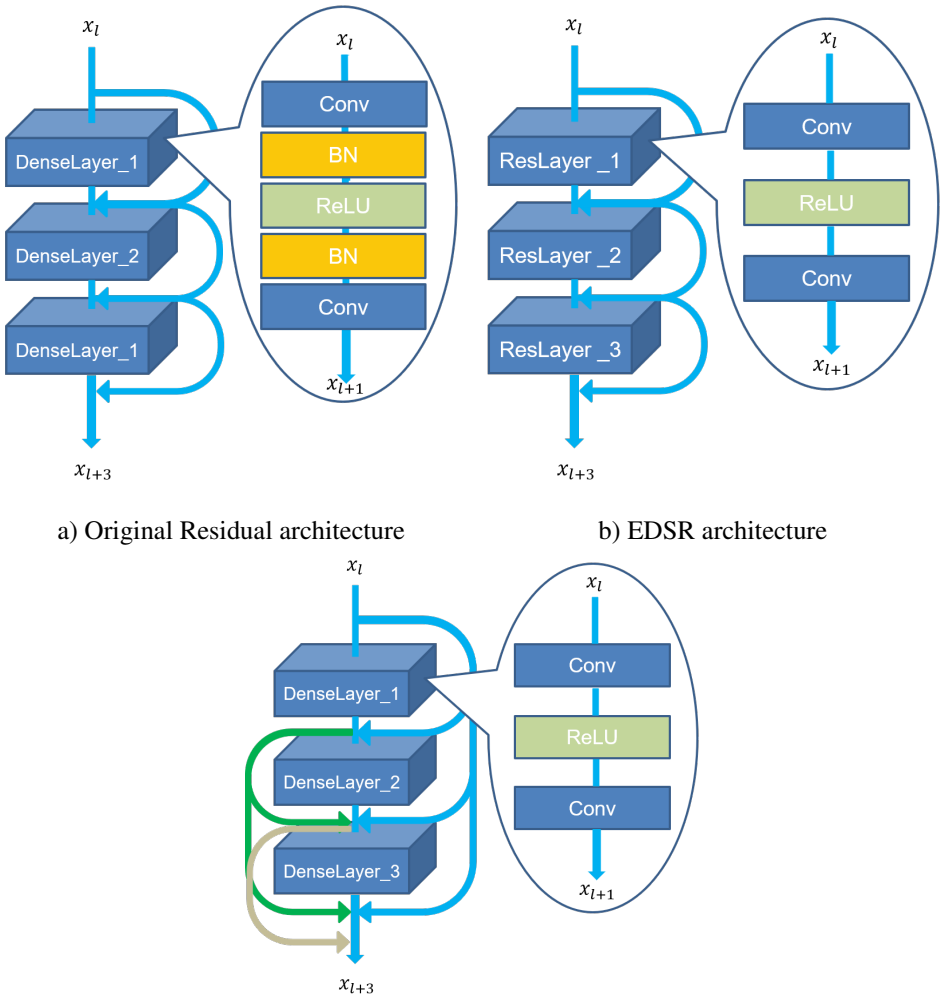
Before the attention mechanism was proposed, the network design to solve the super-resolution problem believe that all spatial positions and channels are of uniform importance for super-resolution.

But considering that not all features of a image are essential for super resolution, and they share different importance. Therefore, the attention mechanism is a very critical technology for the super-resolution problem, allowing the deep learning model to pay more attention to certain local information. Recently, channel attention networks have shown excellent performance in super-resolution tasks, such as SelNet[6], RCAN[23], DRLN[2] and SRRAM[3].

In Figure 2, it illustrates our attention block. First of all, we use the convolutional layer to replace the full convolutional layer, which is more flexible, and don't need to limit the resolution of the input image. Secondly, for our model, we use the concatenate operator to replace element-wise sum, which just fits the idea of dense to dense.

3.3 Dense Blocks with Attentions (DBA)

As we discussed in previous two sections, the dense block ensures the network to take maximum information flows and keeps as many as possible features of input images which can leave more opportunities to generate a more reliable high-resolution image. With the attention of channel to adaptively rescaled of the features of each channel, our model can process various types of information to achieve a higher flexibility and robustness. Meanwhile, the deletion of batch normalization and ReLU have a common purpose of obtaining more adaptability with the improvement of efficiency. We encapsulate and call this dense blocks with attentions as DBA, and its architecture is shown in Figure 2.



a) Original Residual architecture b) EDSR architecture

c) Proposed

Figure 1: Comparison of architecture of original ResNet, EDSR and ours

4 Experiments

In this section, we briefly describe the dataset we used for training and evaluation. Details for data pre-processing, training and testing are also covered.

4.1 Datasets

DIV2K[20] is a newly collected large-scale RGB images dataset with a wide variety of scenarios, and only train data and validation data are used. Specifically, there are 800 high-resolution images with corresponding low-resolution images of different levels in the training set, and 100 images in the validation set. Since the ground truth of the test dataset of DIV2K is not released, we use three well-known super-resolution image reconstruction datasets: Set5[5], Set14[22] and BSD100[17] as our test datasets.

4.2 Evaluation criteria

We used two criteria to evaluate the performance of our proposed model.

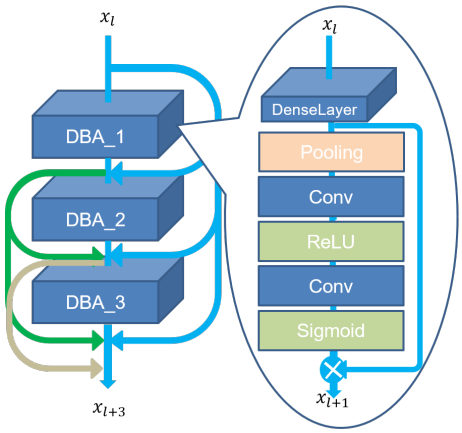


Figure 2: The architecture of attention block in our model

4.2.1 Peak Signal to Noise Ratio

Peak signal to noise ratio (PSNR), an objective standard for evaluating images, is used to assess the quality estimation of our model. To calculate PSNR, we need to obtain the mean square error (MSE) first:

$$MSE = \frac{1}{m \cdot n} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} \left[SR(i,j) - HR(i,j) \right]^2, \tag{1}$$

where SR is our model output, HR is the ground truth image with size of m rows and n columns. Then, PSNR can be expressed as :

$$PSNR = 10 \cdot log_{10}(\frac{MAX_{HR}^2}{MSE}), \tag{2}$$

where MAX_* is the range of the pixel values. For example, if HR is an 8-bit RGB image, then the value of MAX_{HR} is 255.

4.2.2 Structural Similarity Index Measure

Structural similarity index measure (SSIM) is used to consider the similarity between two images from three aspects-luminance, contrast and structure:

$$\begin{aligned} luminance(x,y) &= \tfrac{2\mu_x\mu_y+c_1}{\mu_x^2+\mu_y^2+c_1} \\ contrast(x,y) &= \tfrac{2\sigma_x\sigma_y+c_2}{\sigma_x^2+\sigma_y^2+c_2} \\ structure(x,y) &= \tfrac{\sigma_{xy}+c_3}{\sigma_x\sigma_y+c_3} \end{aligned} \tag{3}$$

$$c_1 = (0.01MAX)^2, c_2 = (0.03MAX)^2, c_3 = 0.5 * c_2,$$

where x is the output of our model; y is the ground truth image; σ_* and μ_* are standard deviation and mean value of corresponding data respectively; σ_{xy} is the convariance values of x and y; The definition of MAX is same with Equation 2; and c_1, c_2, c_3 are constants to avoid dividing 0. With the three parameters, the SSIM can be defined as

$$SSIM(x,y) = \left[luminance(x,y)^\alpha \cdot contrast(x,y)^\beta \cdot structure(x,y)^\gamma \right], \tag{4}$$

where α, β, γ are three constants that are usually set to 1. Above all, we have:

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{5}$$

Options	SRResNet	EDSR	ours
Layers	16	32	72
# Filters	64	256	256
# Parameters	1.5M	43M	96.7M
Use BN	Yes	No	No
Loss function	L2	L1	L1

Table 1: Model specification.

Parts		Settings
Software	System	Windows 10 professor 64-bit
	Python	3.7.3
	Torch	1.4.0
	CUDA	10.1
Hardware	CPU	Intel(R)Core(TM) CPU i5-9600K @ 3.70GHZ Core 6
	GPU	Nvidia GeForce RTX 2070 (8GB)
	Memory	16GB DDR4 2666MHz

Table 2: Experiment environment settings.

4.3 Training Details

We used a mini-patch for training instead of using a whole image to avoid excessive model parameters. In addition, we performed the same image augment processing on the input image and ground truth, specifically random rotation, random horizontal flipping, and random Gaussian noise.

We train our model using Adam optimizer with an initial learning rate 0.0001 and reduce 0.95 for every 100 epochs. L2 loss is usually used in super-resolution tasks, as PSNR includes L2, i.e., PSNR can be maximized in minimizing L2 to obtain better performance. However, according to Lim et al. [15], compared with L2, the L1 loss has better convergence. Hence, we use the L1 loss as the loss function. After each epoch, there is a test session to calculate PSNR and SSIM. For convenience, we use mean square error loss in the test session. We trained 2 networks with 2 and 4 up-sampling scales.

4.4 Results

The Table 4.4 shows the comparison of our two models with other models on the benchmark dataset. From the table, we can find that the performance of our model is close to that of the EDSR, but still slightly inferior to RCAN.

The Figure 3 shows some demonstration. The first image is many Barbie dolls stacking on each other, and second image is a starfish lying on the seabed. In terms of the visual quality, our model is far better than bicubic interpolation method and achieves a basically same visualized results compared to EDSR. Also, it is obviously that the details of hairs of dolls and spikes of urchins are well restored.

Method	Scale	Set 5		Set 14		B100	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	×2	33.66	0.9299	30.24	0.8688	29.56	0.8431
SRCNN [7]	×2	36.66	0.9542	32.45	0.9067	31.36	0.8879
FSRCNN[8]	×2	37.05	0.9560	32.66	0.9090	31.53	0.8920
VDSR[12]	×2	37.53	0.9590	33.05	0.9130	31.90	0.8960
EDSR[15]	×2	38.11	0.9602	33.92	0.9195	32.32	0.9013
RCAN[23]	×2	**38.27**	**0.9614**	**34.12**	**0.9216**	**32.41**	**0.9027**
Ours	×2	38.14	0.9602	33.84	0.9183	32.14	0.9002
Bicubic	×4	28.42	0.8104	26.00	0.7027	25.96	0.6675
SRCNN [7]	×4	30.48	0.8628	27.50	0.7513	26.90	0.7101
FSRCNN[8]	×4	30.72	0.8660	27.61	0.7550	26.98	0.7150
VDSR[12]	×4	31.35	0.8830	28.02	0.7680	27.29	0.0726
EDSR[15]	×4	32.46	0.8968	28.80	0.7876	27.71	0.7420
RCAN[23]	×4	**32.73**	**0.9013**	**28.87**	**0.7889**	**27.85**	**0.7455**
Ours	×4	32.44	0.8964	28.54	0.7763	27.73	0.7429

Table 3: Quantitative results comparison with different models.

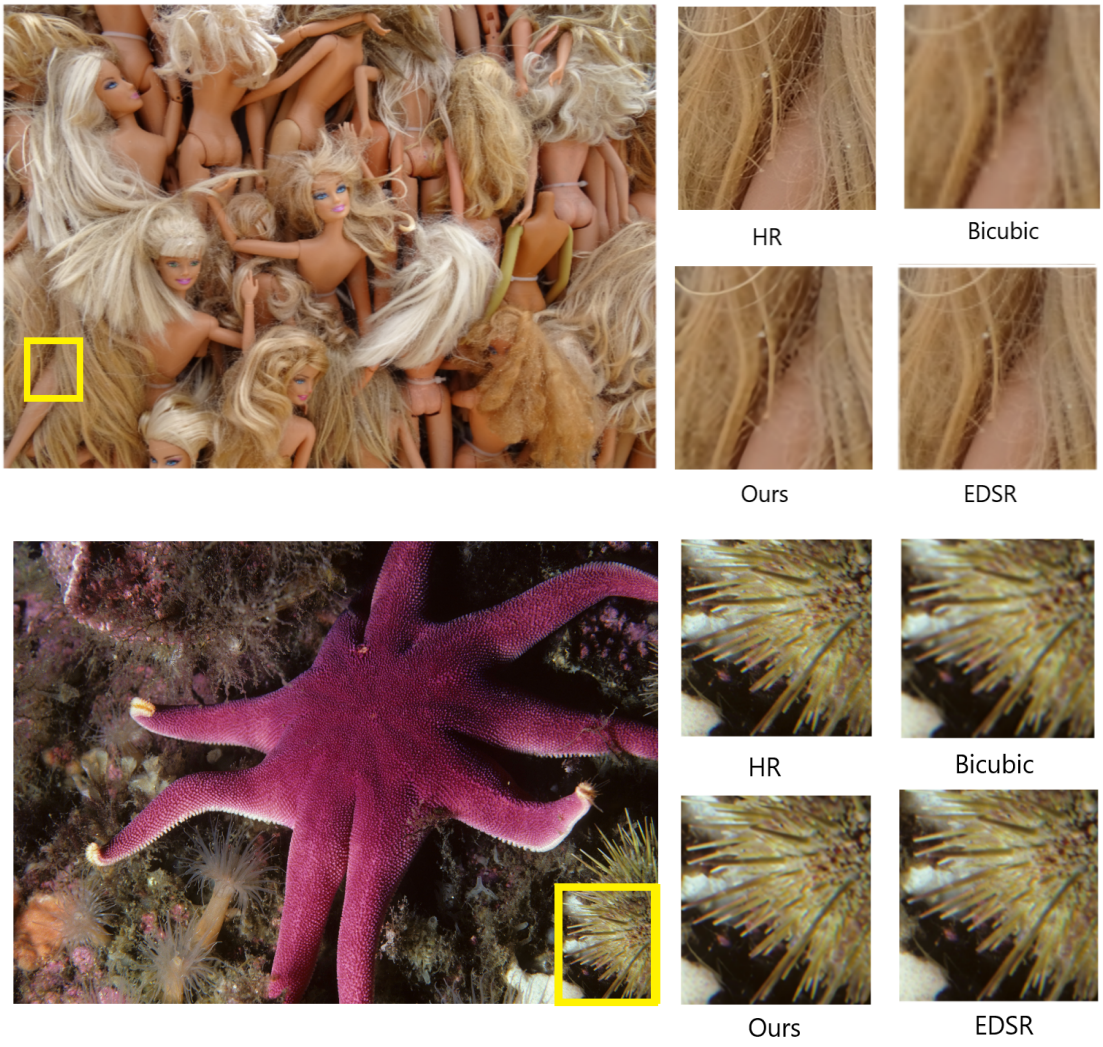


Figure 3: Demonstration of bicubic degradation, our model, EDSR, and ground truth.

5 Discussion and Conclusion

In deep neural network layers, we assume that the dense connection cannot achieve a better performance than residual shortcut. The core idea behind either ResNet or DenseNet is using skip connections (shortcuts) to let some inputs into the layer without any modification, to implement the integration of information flow. In fact, for tradition ResNet, the gradients magnitude is high in middle of depth and approximately zero in the deeper layer. Thus, even more skip connections are added-in, the gradient vanishing actually not completely solved. In future, we believe that more researches on effective paths can make more improvement toward this issue.

In this paper, we propose an enhanced super-resolution algorithm. By using dense to dense, we can obtain improved results while making the model compact. Our attention mechanism also focuses limited attention on key information, thereby saving resources and quickly obtaining the most effective information. The performance of the dense attention network we proposed surpasses most current models, and ranks in the forefront of both the standard benchmark dataset and the DIV2K dataset.

References

[1] J. Allebach and P. W. Wong. Edge-directed interpolation. In *Proceedings of 3rd IEEE International Conference on Image Processing*, volume 3, pages 707–710. IEEE, 1996.

[2] S. Anwar and N. Barnes. Densely residual laplacian super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[3] S. Anwar, S. Khan, and N. Barnes. A deep journey into super-resolution: A survey. *ACM Computing Surveys (CSUR)*, 53(3):1–34, 2020.

[4] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[5] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012.

[6] J. Choi and M. Kim. A deep convolutional neural network with selection units for super-resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1150–1156, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society. doi: 10.1109/CVPRW.2017.153. URL `https://doi.ieeecomputersociety.org/10.1109/CVPRW.2017.153`.

[7] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

[8] C. Dong, C. C. Loy, and X. Tang. Accelerating the super-resolution convolutional neural network. In *European conference on computer vision*, pages 391–407. Springer, 2016.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] D. Hu. An introductory survey on attention mechanisms in nlp problems. In *Proceedings of SAI Intelligent Systems Conference*, pages 432–448. Springer, 2019.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[12] J. Kim, J. Kwon Lee, and K. Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016.

[13] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[14] X. Li and M. T. Orchard. New edge-directed interpolation. *IEEE transactions on image processing*, 10(10):1521–1527, 2001.

[15] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

[16] M.-T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[17] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int'l Conf. Computer Vision*, volume 2, pages 416–423, July 2001.

[18] S. Nah, T. Hyun Kim, and K. Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3883–3891, 2017.

[19] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[20] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, and L. Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 114–125, 2017.

[21] A. Youssef. Image downsampling and upsampling methods. *National Institute of Standards and Technology*, 1999.

[22] R. Zeyde, M. Elad, and M. Protter. On single image scale-up using sparse-representations. In *International conference on curves and surfaces*, pages 711–730. Springer, 2010.

[23] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301, 2018.