ISBN: 978-981-18-5852-9
WCSE 2022 Spring Event: 2022 9th International Conference on Industrial Engineering and Applications
doi: 10.18178/wcse.2022.04.063

Multi-channel Image Super-resolution Reconstruction based on ESRGAN

Pingan Qiao[1], Jing Li[1]

[1] School of Computer Science and Technology, Xi'an University of Posts and Telecommunications

Abstract. Image super-resolution reconstruction, the task of producing high quality images from existing low quality images, is to solve the problems of poor perception effect and unclear texture of image super-resolution results. Then a method of multi-channel image super-resolution under ESRGAN is proposed. Firstly, 2D-3D CNN is used to reconstruct the multi-channel image data and is set to collect more image texture information. Secondly, combined with DPN, the generation network is improved to reduce the memory usage of parameters and alleviate the problem of gradient disappearance. Finally, the texture loss function is introduced to ensure the high frequency details of the generated image while ensuring the integrity of the contents. The experiment indicates that compared with the contrast algorithm, the average value of PSNR index increased by 2.31dB，and the SSIM index of structural similarity increased by 0.071. The visual result is also significantly enhanced by the method of multi-channel image super-resolution.

Keywords: ESRGAN, Image super-resolution reconstruction, DPN, Multichannel image

1. Introduction

The higher the image resolution, the more information it contains. At present, the common way to obtain high-resolution images is to upgrade the hardware equipment, but it costs too much. To lower the cost, Image super-resolution reconstruction technology came into being. In recent years, the introduction of deep learning has brought rapid progress to image super-resolution reconstruction tasks. In 2015, Dong et al. proposed the SRCNN [1], which only uses a three-layer Convolutional Neural Network to train the interrelation between low-resolution and high-resolution images. The results obtained are superior to the traditional mehods [2,3,4]. VDSR [5] deepens the network structure on the basis of SRCNN. To speed up deep network convergence, the idea of global residual is introduced to simplify the training difficulty and further improve the reconstruction effect. However, there remains a problem — a lack of high-frequency details. SRGAN [6] proposed to use the adverse objective function to promote super-resolution image output to solve it, which is close to the manifold of the natural image, but the reconstruction was too smooth. ESRGAN [7] is based on SRGAN. By deleting the batch normalization and merging the artificial artifacts in the dense block SRGAN results, the visual quality of the reconstruction results of specific data sets is enhanced to some extent, but the number of parameters is large and the denoising effect is not obvious.

To improve the above problems, this paper proposes a method to reconstruct high-resolution images through multi-channel image feature extraction based on ESRGAN. (MC-ESRGAN).Firstly, 2D-3D CNN[8] method was used to reconstruct multi-channel image data set to extract the unique image information of different bands of real world images. Then, combined with DPN[9], the generation network of ESRGAN is improved for feature value extraction to make more reasonable use of image features and reduce the number of network parameters. Finally, the texture loss function is added to make sure to generate high frequency details of the image.

2. Method

ESRGAN has a relatively good reconstruction result in the current reconstruction field. However, the number of parameters is big and the noise processing is not good enough. The generator part is improved on the basis of SRGAN. Firstly, ESRGAN does not use batch standardization. The BN layer normalizes test data by using mean and variance estimated over the entire training set, which results in artifacts in the reconstructed image. Secondly, in the Basic Block part,the residual Block of SRGAN sequential connection

is changed, and RRDB residual cover residual structure is used for feature extraction. The shallow feature and deep feature are combined, and the information of each layer is fully utilized so that the network can learn more fine details. Although relatively excellent results have been obtained, the number of calculations increases. Finally, through the study of the loss area, it is established that the activated features are very sparse, especially in the deep network. This sparse activation provides poor supervision, resulting in poor performance, and the use of activated features can cause the reconstructed image to be inconsistent with the brightness of GroundTruthy's.image.Therefore,pre-activation characteristics are chosen.In this paper, on the basis of ESRGAN, the DPN module is used to improve the generator structure, choosing the residual network as the backbone, and adding a dense linpath to build a dual path. It helps to slow down the width of intensive link path increment and the cost of the GPU memory, by adding a slice on the existing network layer, and the link layer can be simple to implement. Compared with the RRDB network, DPN has higher parameter efficiency and lower memory consumption. More texture features are extracted from multi-channel images to enrich high-frequency information, and texture loss function is combined to train the model and improve network performance better.

2.1. Image Preprocessing

Hyperspectral images provide rich information in multi-frequency bands [10], including color, texture, light, and shade contrast, which is conducive to image super-resolution reconstruction. Based on the performance of deep learning in various computer vision applications, RGB image restoration algorithms for hyperspectral images have made great progress. In this paper, the 2D-3D CNN method is used to rebuild multi-channel image data set. There are 31 feature maps in the last layer of the model, and the wavelength difference between each feature map is 10nm [7]. The 2D-CNN model uses 2D to convolute single channels and calculates the average value generated by each pixel in these channels, to extract the available hyperspectral information in the spatial domain of specific channels, effectively integrating spatial data. The 3D-CNN model is used to add and convolute the responses of all corresponding pixels in 3D space and channels. Then the relevant information between the channels is merged. The 2D-CNN model mainly uses the spatial correspondence of channels in the image to obtain spectral data, while the 3D-CNN model uses the correspondence between channels to refine the extraction of spectral data.

Fig. 1 shows the transformed multi-channel image. As there are too many bands in the 31 channels image, which will affect the effectiveness of the experiment, a genetic algorithm is adopted to select the optimal band. When extracting feature bands, the band shape features in the data set are first arranged into a feature matrix of 256×256, and then 12 features of each band are optimized for band selection. The final six bands are 430nm, 480nm, 540nm,600nm,650nm,and 700nm respectively. Experimental results show that these six bands can contain the texture details of 31 channels to the maximum extent and provide sufficient texture information for subsequent image reconstruction.

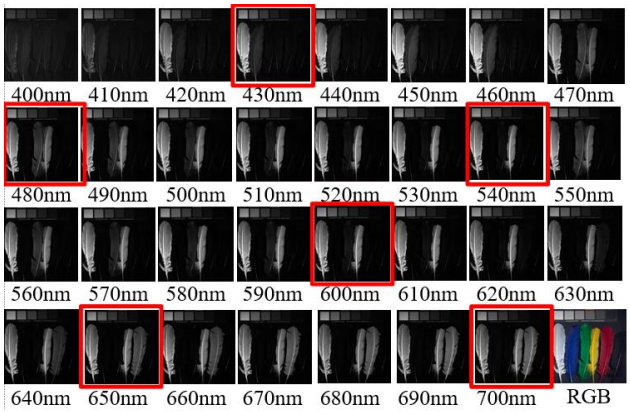

Fig. 1:Multichannel image band selection

2.2. Network Structure

The network structure diagram generated in this paper is indicated in Fig. 2. Firstly, each channel of the 6 input multi-channel images is convoluted with a 3*3 convolution kernel and 64 feature maps respectively,

then the feature enhancement is carried out through dual-path connection of 8 DPN blocks, and finally, the up-sampling layer and convolution layer are carried out. The final output generates an image.

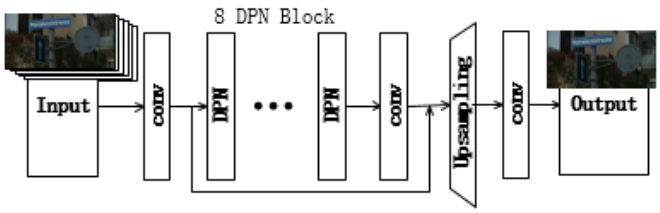

Fig. 2: Generate network model diagrams.

DPN model structure is shown in Fig. 3. The DenseNet and ResNet branches share the first 1x1 convolution in DPN Block. In practical application, 3x3 convolution uses the GROUP mode in ResNeXt to improve performance. The number of features in ResNet branches will also increase, which will reduce the problem of DenseNet feature width increasing as the hierarchy increases. The output of the last 1x1 convolution in the structure is split into two, one half of which has the same number of features as the input of the ResNet branch, so that it adds exactly to the input of the ResNet branch. DPN Block improves the accuracy without increasing the parameter complexity, reduces the number of hyperparameters, and ultimately fuses the extraction features of its mapping and residual mapping.

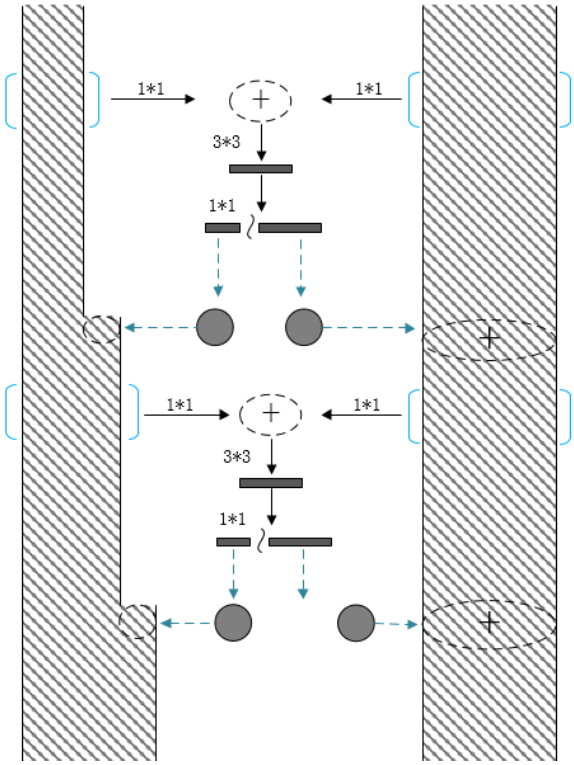

Fig 3. Dual Path Architecture

In the design part of the discriminant network, the Patch Disiciminator proposed by Isola et al is used in this paper. [11] It was completely composed of the convolution layer, and the mean value of the output matrix was taken as the discriminant criterion to determine whether the image was a reconstructed image or a real image. As the discriminant criterion is matrix mean, it is more likely to take into account the influence between image regions. It has been proved that Markov discriminant has a significant effect on maintaining the sharpness and texture details of high-resolution images. Finally, the reconstructed image is discriminated against the HR image, and it is shown in Fig. 4

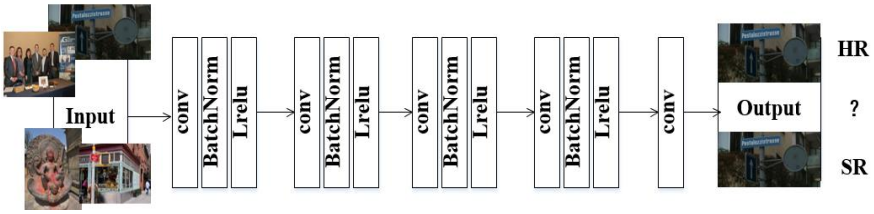

Fig 4. Discriminant network model diagram

2.3. Loss Functions

The loss functions used for training in MC_ESRGAN include pixel loss $L1$, antagonism loss L_{adv}, and texture loss L_{tex}. As shown in (1).

$$L = \alpha L_{adv} + \beta L_{tex} + \delta L1 \tag{1}$$

Pixel loss uses L1 loss, as shown in (2).

$$L1 = \sum_{i=0}^{n} \left| y_i - f(x_i) \right| \tag{2}$$

Where y_i represents high resolution image and $f(x_i)$ stands for the reconstructed image.

Adversarial loss is a commonly used loss function for adversarial network generation, which is used to enhance the details of the generated image and make its visual effect more real, as shown in (3).

$$L_{adv} = E_{x\sim P_G}[D(x)] - E_{x\sim P_{data}}[D(x)] + E_{x\sim P_z(x)}[1\text{-}D(G(x)] \tag{3}$$

D represents the generation network and represents the signal distribution of random noise based on the data taken from the real distribution.

Since the high-frequency information of HR image obtained through double cubic down-sampling is missing, and the image characteristic parameters that can be obtained from the multi-channel image are much larger than RGB image [10],the generator is not able to learn the high-frequency information, so texture loss Ltex is added, as shown in (4).

$$L_{tex} = \left\| G(x_i) - y \right\|_1 \tag{4}$$

Among them, G is the Gram matrix, and Ltex is an improvement on perceived loss. Feature mapping of the output of the convolution intermediate layer of the generation network and discriminant network is obtained and the corresponding Gram matrix is calculated. The style and texture of the picture can be extracted by viewing the values in the feature graph and computing the correlation of the feature graph.

In this paper, the values of equilibrium coefficients are α =5×10-3, β =1, δ =1×10-2。

3. Experimental Analysis

3.1. Experimental Environment

The experiment in this paper is carried out on a machine with 16GB(Gigabyte) memory, Inter (R) Core (TM) I7-4710HQ CPU, 64-bit Ubuntu operating system, and NVIDIA GTX 1080TI graphics card. The running software is JetBrains PyCharm 2019.2.4x64, and the experimental framework is Tensorflow.

3.2. Experimental Data Set

In the experiment, the high-quality DIV2K[12] data set was selected for training, with a total of 1000 2K resolution images, including 800 training images, 100 verification images, and 100 test images. The test data adopted Set14, BSD100, and DPED[13,14,15] data sets to test the performance of the model. The low resolution images required for training are obtained by four times degradation of high resolution images in DIV2K dataset by bicubic interpolation.

3.3. Experimental Result And Evaluation

To verify the effectiveness of the proposed method, Bicubic, SRGAN, ESRGAN, and the proposed method were tested on public datasets Set14, BSD100, and DPED. The Peak signal-to-noise Ratio (PSNR) and Structural Similarity Index (SSIM) values at 4 times of upsampling by different algorithms are

calculated. PSNR is an image quality evaluation based on error sensitivity, the formula is shown in (5). SSIM is an image quality evaluation standard that conforms to human vision. It is calculated based on the luminance and contrast of the image. The closer the value is to 1, the better the quality of generated images has, the formula is shown in (6).

$$PSNR = 10 \times \log_{10} \frac{255^2 \times W \times H \times C}{\sum_{i=1}^{W} \sum_{j=1}^{H} \sum_{z=1}^{C} \left[\bar{X}(i,j) - x(i,j) \right]^2 + 1 \times 10^{-9}} \tag{5}$$

$$SSIM(X,Y) = \frac{(2\mu_x \mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_{X^2} + \mu_{Y^2} + C_1)(\sigma_{X^2} + \sigma_{Y^2} + C_2)} \tag{6}$$

The results are shown in TableI. It can be found that the MC-ESRGAN model proposed in this paper has improved its PSNR compared with Bicubic, SRCNN,SRGAN,andESRGAN, the four mainstream image reconstruction models used for testing. The mean value of the PSNR index increased by 2.31dB, and the SSIM index was 0.008 lower than the ESRGAN algorithm on the Set14 dataset, but higher than the comparison algorithm on the other two datasets.

TABLE I. IMAGE RECONSTRUCTION RESULTS OF FIVE ALGORITHMS ON SET14，BSD100 AND DPED DATASETS

Data set	evaluation index	Bicubic	SRCNN	SRGAN	ESRGAN	Ours
Set 14	PSNR	28.43	28.93	29.97	30.77	32.33
	SSIM	0.820	0.839	0.837	0.897	0.889
BSD100	PSNR	25.93	26.08	27.16	27.13	27.43
	SSIM	0.678	0.690	0.722	0.728	0.737
DP ED	PSNR	27.21	29.36	29.28	30.11	33.15
	SSIM	0.731	0.763	0.811	0.789	0.879

The running time of the five algorithms on Set14 data set, BSD100 data set and DPED data set is shown in TableII. It can be seen that the Bicubic algorithm has the shortest time due to its simple operation, while other algorithms are slow in network training due to deep learning. Because SRCNN has only three simple convolution layers, the computational amount is minimal in the deep learning method.Among them, the SRGAN algorithm has the slowest reconstruction speed because the BN layer is not removed from the network structure. MC-ESRGAN algorithm is slightly faster than SRGAN algorithm and ESRGAN algorithm because the DPN algorithm is introduced. Based on TableI and TableII, it can be seen that the MC-ESRGAN algorithm has higher PSNR and SSIM values than other algorithms when the running time is reduced, which indicates that the proposed algorithm has a better image reconstruction effect.

TABLE II. RUNNING TIME OF FIVE ALGORITHMS ON SET14 AND BSD100 AND DPED DATASETS

arithmetic	Set14	BSD100	DPED
Bicubic	1.894	13.568	136.484
SRCNN	3.352	25.228	229.562
SRGAN	4.788	28.482	298.563
ESRGAN	4.381	27.332	275.635
Ours	3.865	25.423	254.269

In this paper, there images are selected from the DPED data set for high resolution image reconstruction test, which is used for visual comparison with four classic super-resolution reconstruction methods: Bicubic, SRCNN， SRGAN, and ESRGAN. Since the images in the DPED data set are unprocessed real images, containing noise, blur, dark light, and other low-quality problems, there is no corresponding ground truth, which can only provide an intuitive comparison. To visualize the advantages of this model, local high-resolution images of tree branches ,small tiger whiskers and the license plate number were selected for magnification and comparison, as shown in Fig 5.

Fig 5. Four algorithms are used to reconstruct visual contrast images

The MC-ESRGAN reconstructed images of the whiskers and branches of the tiger cub are clearly textured and show details closest to the real texture seen by human eyes in the real world. In contrast, Bicubic images are very blurry, and even some texture details are different from real images. Due to the simple network structure of SRCNN, more texture details can be learned.SRGAN images can recover some high-frequency information, but they are too smooth and not effective in sharpening. ESRGAN method is better than the first there methods in overall image reconstruction, but it will introduce noise information. In the third picture, the license plate is under the quadruple magnification factor. Our method is obviously superior to other algorithms in terms of digital clarity. Compared with SRGAN, it reduces the generation of artifacts and has clearer texture details than ESRGAN.

4. Conclusion

In this paper, 2D-3D CNN is firstly used to reconstruct the multi-channel image data set and enrich the high-frequency details of the image by taking advantage of the characteristics that different bands of the multi-channel image contain different texture information. In addition, a genetic algorithm is used to select six bands from the 31-channel image as the input of the generation network to reduce the number of parameters. Secondly, the RRDB module in the ESRGAN generation network structure is improved by combining the DPN idea to reduce the memory usage of parameters and alleviate the problem of gradient disappearance. Finally, the texture loss function is introduced to reduce the loss of high-frequency details. After the discussion before, the MC-ESRGAN algorithm proposed in this paper has improved the evaluation index

results and subjective visual effects obtained on Set14 data set, BSD100 data set, and DPED data set compared with other reconstruction methods. However, this experiment was only reconstructed at a factor of 4, and further studies should be carried out on reconstruction at 8 magnification and greater magnification.

5. Acknowledgment

This work was supported by the National Natural Science Foundation of China(Item NO:61105064),Scientific Research Project of Shaanxi Provincial Department of Education(Item NO:16JK1689),and the Key Laboratory of Network Data Analysis of Shaanxi Province.

6. Rererences

[1] Dong, C., Loy, C. C., He, K., & Tang, X.. " Image super-resolution using deep convolutional networks. " IEEE transactions on pattern analysis and machine intelligence, vol.38,no.2, pp.295-307, 2015.

[2] Zhu, S., Zeng, B., Zeng, L., & Gabbouj, M. Image interpolation based on non-local geometric similarities and directional gradients. IEEE transactions on Multimedia, vol.18,,no.9,pp.1707-1719, 2016

[3] Suhail Hamdan, Yohei Fukumizu, Tomonori Izumi, and Hironori Yamauchi, "Face Image Super-Resolution with Adaptive Patch Size to Scaling Factor," Journal of Image and Graphics, vol. 6, no. 2, pp. 167-173, December 2018.

[4] PENG, Y., GAO, Y., DU, T., SANG, Y., & ZI, L. Single Image Super-Resolution Reconstruction Method for Generative Adversarial Network. Journal of Frontiers of Computer Science and Technology, vol.14,no.9, pp.1612-1620, 2020.

[5] Kim, J., Lee, J. K., & Lee, K. M. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition ,pp. 1646-1654, 2016.

[6] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., ... & Shi, W. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE conference on computer vision and pattern recognition ,pp. 4681-4690.2017.

[7] Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., ... & Change Loy, C. Esrgan: Enhanced super-resolution generative adversarial networks. In Proceedings of the European conference on computer vision (ECCV) workshops ,pp. 0-0,2018.

[8] Koundinya, S., Sharma, H., Sharma, M., Upadhyay, A., Manekar, R., Mukhopadhyay, R., ... & Chaudhury, S. 2d-3d cnn based architectures for spectral reconstruction from rgb images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops ,pp. 844-851, 2018.

[9] Chen, Y., Li, J., Xiao, H., Jin, X., Yan, S., & Feng, J. Dual path networks. arXiv preprint arXiv,pp.1707.01629,2017.

[10] Li, J., Cui, R., Li, B., Song, R., Li, Y., Dai, Y., & Du, Q. Hyperspectral image super-resolution by band attention through adversarial learning. IEEE Transactions on Geoscience and Remote Sensing, ,vol.58,no.6, 4304-4318, 2020.

[11] Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition ,pp.1125-1134,2017.

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

[13] Zeyde, R., Elad, M., & Protter, M. On single image scale-up using sparse-representations. In International conference on curves and surfaces,pp. 711-730, Springer, Berlin, Heidelberg,2010.

[14] Martin, D., Fowlkes, C., Tal, D., & Malik, J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001,Vol. 2, pp. 416-423. IEEE,2001.

[15] Ignatov, A., Kobyshev, N., Timofte, R., Vanhoey, K., & Van Gool, L. Dslr-quality photos on mobile devices with deep convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision ,pp. 3277-3285,2017