

Efficient-VDSR Network for Super-Resolution

Xi Chen

School of Information and Control Engineering, China University of Mining and Technology,
Xuzhou, China.

chenxi2019@cumt.edu.cn

Abstract. Super-resolution image processing is an important research topic in the field of image processing. The exploration in this paper comes from the use of VDSR network, which can improve the learning rate by only detecting and learning the residuals since the difference between low-resolution and high-resolution images lies more in the high-frequency detail part. Increasing the depth of the network can significantly improve the learning rate and further affect the final rendering of the image, but the increase of weight layers also increases the number of iterations and training time. This paper intends to reduce the number of iterations by decreasing the number of weight layers and correspondingly shorten the training time of the model. Then, we create low-resolution images and improve the resolution of the images by using bicubic interpolation method and the trained network respectively, and then obtain the evaluation data of the image quality by final visualization and quantitative comparison. The results show that reducing the number of weight layers still leads to a high accuracy of the images and will reduce the number of iterations and time accordingly.

Keywords: Super-resolution; Deep Learning; Residual Learning; VDSR Network; Bicubic interpolation; Quantitative Comparison.

1. Introduction

Super-resolution is the process of reconstructing a low-resolution image into a high-resolution one by using redundant information of the image. Single-image super-resolution (SISR) [1] [2] is able to accomplish the low-resolution (LR) image obtained by scaling a given original high-resolution (HR) image to estimate the high-resolution (HR) image, that is, it solves the problem of recovering high-resolution images to some extent. Super-resolution can be directly applied in many fields, such as target detection of aerial images, medical imaging [3], satellite imaging [4], and face recognition [5]. There are many ways to optimize single image super-resolution, such as statistical methods, adaptive learning, deep learning and convolutional neural networks. However, the limited quality of high-resolution images is a highly discomforting problem due to the lack of high-frequency information in low-resolution images, and there are great limitations, including strong dependence on images and slow training speed due to the large amount of data.

Therefore, Jiwon Kim [6] et al. proposed to solve the problem by using residual learning and gradient cropping. VDSR network is a convolutional neural network architecture used to perform single image super-resolution. VDSR network has more advantages in that it has deeper network structure, which makes the deeper network layers have a larger field of perception. If the convolutional layer has D-layer network, then it has a receptive field of $(2D+1) \times (2D+1)$. Also, the VDSR network uses residual learning and most of the values are relatively small, so the convergence is fast. It has better training effect and high applicability, and one trained model can solve multiple problems. Jiwon Kim et al. used a 20-layer network for training, and the more layers of the network, the deeper the network will be, so each training consumes a lot of time and resources. In this paper, it aims to optimize the VDSR network by reducing the number of network layers with small training quantity and high accuracy.

Jiwon Kim et al. first introduced residuals to reduce the computational effort to speed up convergence, and also used adaptive gradient cropping to limit the gradient range and speed up convergence. If the gradient is not processed and the learning rate is only increased, it may bring about the disappearance or explosion of gradient [13]. Therefore, this paper aims to continue

improving the VDSR network by using the residuals as the basis. We believe that the high number of 20 convolutional layers does not substantially change the quality of the picture, but brings about a waste of resources. Therefore, it is important to find the appropriate number of layers, i.e., to find the minimum number of convolutional layers that will not affect the image quality. We take 20 layers as the upper bound and reduce the number of layers one by one to observe the obtained evaluation data and find the number of layers with fewer iterations but without affecting the image quality. Also, we have to compare with the traditional method, so we choose the bicubic interpolation method, which does not need to be trained by the neural network but also gives a better image quality. We need to compare the results obtained by the VDSR network with the network obtained by the bicubic interpolation method to draw conclusions.

2. Related Work

Super-resolution is an important technique in computer vision and image processing and now has a wide range of applications. At the early stage, the ways of processing images by deep learning methods are based on convolutional neural networks (CNN), such as SRCNN [7]. With its development, currently, SR [8] methods are seen which use generative adversarial networks (GAN). The main differences between the different algorithms lie in the network architectures, loss functions, principles and strategies, etc. Existing deep learning-based super-sampling methods are mainly classified into supervised SR, unsupervised SR and domain-specific SR [9]. Four structures are common in supervised super-sampling methods: Pre-upsampling SR, Post-upsampling SR, Progressive upsampling SR and Iterative up-and-down Sampling SR. In pre-upsampling SR, the low-resolution image is first interpolated to roughly obtain the high-resolution image, and the commonly used interpolation methods are nearest neighbor interpolation, bilinear interpolation and bicubic interpolation. These methods only improve the image resolution based on its own signal, so it is faster and the obtained image results are better. In our experiments, we will also take the bicubic interpolation method to process the images for control and observe the effectiveness of the VSDR network model. However, this method will amplify the noise and is more serious for the waste of resources such as time. In the post-upsampling SR, the lifted-resolution images are passed to the CNNs and operated at the end of the network with lower spatial dimensionality and less computational complexity. The progressive upsampling SR decomposes the task and reduces the learning difficulty while maintaining a high performance, but the mode type is more complicated. Iterative up-and-down Sampling SR can better reflect the relationship between high-resolution images and low-resolution images, and thus it is easier to obtain reconstruction results. Deep learning-based on super-resolution plays a very important role in practical applications, so its study is also very critical.

The VDSR network is an improved model for the SRCNN. The SRCNN approach has been shown to map LR to HR in an end-to-end manner via CNN [10]. Its structure is simple and consists of three convolutional layers that can divide the network into feature extraction, mapping, and reconstruction layers. However, it is highly dependent on the information around the image and consumes a lot of resources during training, so its drawbacks can be optimized by using VDSR networks. The application of VDSR networks has certain advantages. First of all, it is the use of residuals for operations. In order to solve the problem of gradient disappearance that easily occurs in deep networks and to improve the iteration speed, He et al. proposed a residual neural network [11] in 2015, which learns residual features by using multiple residual blocks connected at the beginning and end, and the structure schematic is shown in Figure 1. In Figure 1, weight layer is the convolutional layer, x is the input, $F(x)$ is the mapping function, and $H(x)$ is the output. It can be seen that the residual neural network learns the difference between the input and output, while He et al. proved that the model can increase the residual gradient and has stronger learning ability. By using the residual neural network in the VDSR model and making the difference operation, most numerical results are obtained as 0 or very small, which simplifies the processing of the data.

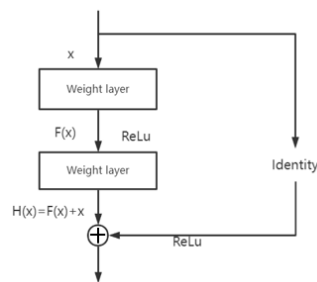


Figure1: Schematic Diagram of Residual Learning Structure

Secondly, the VDSR network is able to ensure that the image size does not change. For the conventional convolution operation, the image size decreases as the number of convolutions increases, and this operation is obviously not applicable to networks with very deep layers. The conventional convolution operation may lead to poor results due to the change of the receptive field, so the surrounding area is usually cropped, which in turn leads to a smaller image. In contrast, the VDSR network uses a complementary 0 operation, which not only restores the original image size one and solves the problem caused by the depth of the convolutional network, but also preserves the boundary details, so that the image quality can be improved.

Jiwon Kim et al. used a convolutional network structure with 20 weight layers with a receptive field of 41×41 . Ideally, the receptive field is the same as the image size. Of the 20 layers, it can be divided into an input layer, a two-dimensional convolutional layer, an intermediate layer including 18 convolutional layers and a linear unit layer, and finally a convolutional layer and a regression layer. In this paper, it is believed that the running time can be reduced by changing the number of intermediate layers and it has little impact on the image quality.

3. Method

In [j], Dong et al. tried to train the images by using a deeper network model, but the results obtained were not satisfactory, so they concluded that a deeper network would not lead to better performance. In [6], Kim et al. used a 20-layer network for training and converged faster and obtained images with high accuracy. However, we believe that although Kim et al. demonstrated that deeper networks can achieve higher accuracy, we can still reduce the number of intermediate layers appropriately to achieve faster convergence without affecting the accuracy.

First, the convolutional network layers are defined: the first and last layers are the image input layer and the regression output layer respectively, and each of the remaining layers contains 64 filters of size $3 \times 3 \times 64$, in which each layer is operated on the input image. After the convolution operation in the intermediate layer, the image size will be reduced accordingly, and the inference of the central pixel is more dependent on the edge pixel. As the number of convolutions increases, the image size will obviously change, causing visual discomfort, and the image quality will also degrade gradually. Therefore, we adopt the method of zero-filling before convolution to keep the image size consistent before and after convolution, and the operation process is shown in Figure 2. The applicability of this method was also demonstrated by Kim et al. Also, since the weights are initialized to random values, which introduces asymmetry, it is necessary to add a ReLU layer to introduce nonlinearity after each layer of convolution is completed. The final layer no longer uses the ReLU layer, but a regression layer for computing the mean square error between the residuals and the network predictions.

For the calculation of image residuals, we define x to denote the low-resolution image data, y to denote the high-resolution image data, and $f(x)$ to be the network prediction. Then the residuals are obtained as $r=y-x$, where the value of r is mostly zero or very small. Therefore, the loss function is defined as $\frac{1}{2} \|r - f(x)\|^2$. After defining for each layer, the VDSR network setup is completed. The structure of the defined network is shown in Figure 2.

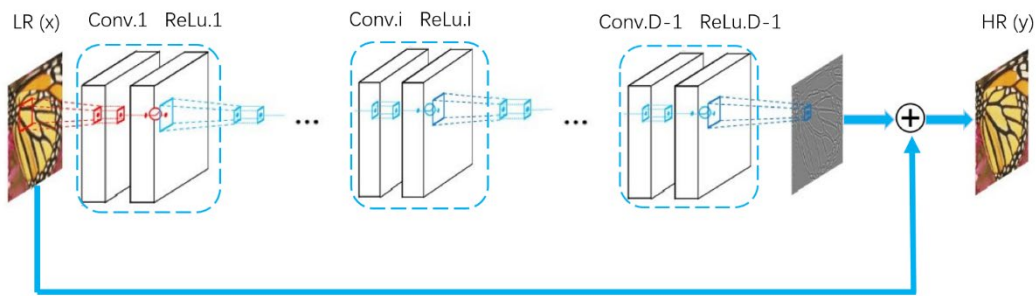


Figure2: Schematic Diagram of VDSR Network Flow

In order to investigate the effect of changing the number of layers of the convolutional network on the super-resolution effect, we use the total number of layers of the network to be changed. Since the first and final layers are already defined, and a ReLU layer is added after each convolutional layer to introduce nonlinearity, so changing the total number of layers actually changes the number of intermediate convolutional layers. In order to obtain a comparison with the traditional method, we simultaneously use bicubic interpolation and make the resulting high-resolution image the same size as the reference image. We change the total number of layers to 12, 14, 16, 18, and 20 respectively, and train them by the VDSR network and the bicubic interpolation method respectively, to obtain the effect of the number of network layers on the image quality and the comparison of the effect of VDSR and bicubic interpolation by testing the graphical effect and comparing the obtained data.

After processing the image, we measure the image quality by using the peak signal-to-noise ratio (PSNR), the structural similarity index of image (SSIM) and the natural image quality evaluation method (NIQE) respectively. PSNR is an image quality evaluation based on the error of the corresponding pixel point, and does not consider the visual characteristics of human eyes, and the larger, the better. SSIM measures the image similarity in terms of brightness, contrast and structure, which is also a full-reference image quality evaluation index. Its value range is [0,1], and the larger the value, the lower the image distortion. In PSNR or SSIM, since the image texture details do not necessarily match the human eye's habit, the image is also evaluated by using NIQE, and a smaller value indicates better perceptual quality.

Finally, we vary the scaling factors of the images so that they are 2, 3, and 4 respectively, and use them to calculate the mean of PSNR and SSIM to evaluate the overall effect of the images.

4. Experiment

4.1 Data Set

During the experiment, we adopted the publicly available dataset IAPR TC-12 Benchmark [12], which contains 20,000 static images for training. And 14 intact images are used for testing, including animals, plants, and scenery, etc. Here, food is selected as the test image to analyze its metric data.

4.2 Data Processing

The high-resolution image was the first one obtained by using the VDSR network. After that the images are visualized and quantitatively compared, and we use roi vectors to specify part of the area. Here we choose $roi=[350 \ 200 \ 180 \ 220]$, four metrics are $[x, y, width, height]$, the left side is the image obtained by bicubic interpolation and the right side is the image obtained by VDSR network.

The peak signal-to-noise ratio (PSNR) is defined by the maximum possible pixel value L and the mean square error MSE. After obtaining L and MSE, the expression for PSNR can be obtained as:

$$PSNR = 10 \log_{10} \left(\frac{L^2}{MSE} \right)$$

The structural similarity (SSIM) highlights the changing relationship between brightness, contrast and structure. Therefore, it can be marked as the brightness, contrast and structure, which are calculated separately and generally takes the weights of them as 1. The expression of SSIM is thus obtained as follows:

$$SSIM = \frac{(2\mu_I\mu_{\hat{I}} + C_1)(\sigma_{I\hat{I}} + C_2)}{(\mu_I^2 + \mu_{\hat{I}}^2 + C_1)(\sigma_I^2 + \sigma_{\hat{I}}^2 + C_2)}$$

In this expression, C_1 , C_2 , C_3 is the constant and σ is the standard deviation. The HR image is denoted by I and the reconstructed image is denoted by $\hat{I}$. NIQE is the distance between the MVG model extracted from the test image and the MVG model extracted from the natural image. The NSS feature model is formed by fitting the model density function. The density function is as follows:

$$f_X(x_1, \dots, x_k) = \frac{1}{(2\pi)^{\frac{k}{2}} |\Sigma|^{-\frac{1}{2}}} e^{(-\frac{1}{2}(x-v)^T \Sigma^{-1}(x-v))}$$

where $x_1, \dots, x_k$ are the features of NSS, v and Σ denote the mean and covariance matrices.

$$D(v_1, v_2, \Sigma_1, \Sigma_2) = \sqrt{(v_1 - v_2)^T \left(\frac{\Sigma_1 + \Sigma_2}{2} \right)^{-1} (v_1 - v_2)}$$

where $v_1, v_2, \Sigma_1, \Sigma_2$ denote the mean and covariance matrices of the natural model and the distortion model respectively.

Meanwhile, the scaling factor is also changed to observe whether the change in scaling factor would have an effect on the quality of the images. Here the scaling factors are selected as 2, 3 and 4 for training and the evaluation parameters PSNR and SSIM are selected. The trend of data change is observed. Firstly, feature extraction from the null domain NSS is required, followed by image block selection and characterization of image blocks.

4.3 Conclusions

Through the above analysis, the first conclusion the experiment got is that the training effect of deep-learning VDSR network is better than that of bicubic interpolation.

The following figure shows the comparison between the bicubic interpolation method and the VDSR network with different number of layers. The left side shows the results obtained by the bicubic interpolation method and the right side shows the results obtained by the VDSR network.



Figure 3: The number of VDSR network layers is 12.



Figure 4: The number of VDSR network layers is 14.



Figure 5: The number of VDSR network layers is 16.



Figure 6: The number of VDSR network layers is 18.

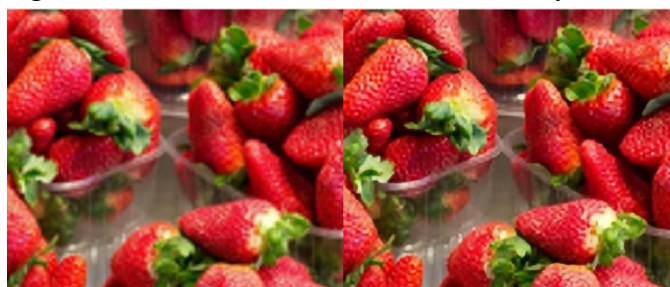


Figure 7: The number of VDSR network layers is 20.

Table 1 Data obtained by bicubic interpolation and VDSR network with different number of layers

Layers	bPSNR	vPSNR	bSSIM	vSSIM	bNIOE	vNIOE
12	23.7935	24.2571	0.9427	0.9500	6.2859	6.1002
14	23.7935	24.2606	0.9427	0.9499	6.2859	6.0036
16	23.7935	24.2606	0.9427	0.9499	6.2859	6.0036
18	23.7935	24.2606	0.9427	0.9499	6.2859	6.0036
20	23.7935	24.2606	0.9427	0.9499	6.2859	6.0036

Of the above images, the left side are the images obtained by image processing using the bicubic interpolation method, while on the right side are the results obtained by image processing of the VDSR network. From the comparison of the images, it is obvious that the image processing effect of VDSR network is significantly better than that of the bicubic interpolation method. It has higher accuracy, clearer details and sharper edges.

The data presented in the table are the evaluation data derived from a single image. b is the data obtained by bicubic interpolation and v is the result obtained by VDSR network training. For the image obtained by the VDSR network, it has a greater signal-to-noise ratio, indicating better image quality. The SSIM value is closer to 1, indicating better brightness, contrast and structure, and more consistent with the reference image, while the NIQE is smaller, indicating better human perception of the image.

Therefore, we can conclude that the image results obtained from the VDSR network trained by using the neural network are better than the bicubic interpolation method that processes the image directly.

The second significant finding is that there is no impact on image quality when the number of convolution layers is 14, but fewer layers are more efficient. It is also concluded from the data that the larger the scaling factor is, the worse the image quality is.

The following figure shows the comparison between the image obtained after training by VDSR network and the partial magnified image.

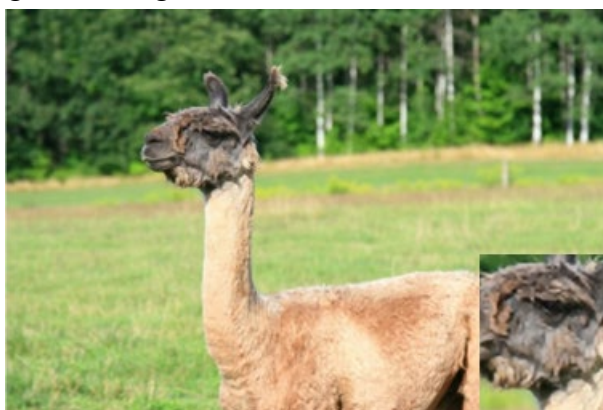


Figure 8: The number of VDSR network layers is 12.



Figure 9: The number of VDSR network layers is 14.

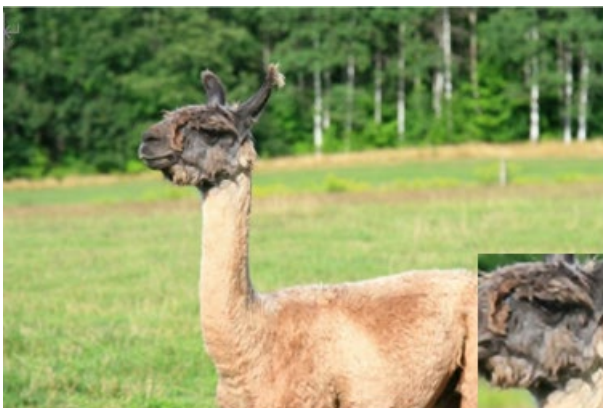


Figure 10: The number of layers of VDSR network is 16.



Figure 11: The number of VDSR network layers is 18.



Figure 12: The number of VDSR network layers is 20.

Table 2: Average values of indicators at different scaling factors

Layers	PSNR2	PSNR3	PSNR4	SSIM2	SSIM3	SSIM4
12	20.6265	18.3122	17.7931	0.6231	0.5792	0.5650
14	20.5267	18.2874	17.7817	0.6228	0.5788	0.5646
16	20.5267	18.2874	17.7817	0.6228	0.5788	0.5646
18	20.5267	18.2874	17.7817	0.6228	0.5788	0.5646
20	20.5267	18.2874	17.7817	0.6228	0.5788	0.5646

As can be seen from the picture, the effect when the number of convolutional layers is 14 is slightly better than that of 12, but when the number of convolutional layers continues to increase, the evaluation index for the image no longer changes and the image quality does not improve significantly. When the VDSR network is used to perform the hyper-segmentation operation on the image, the quality of the image will not be enhanced all the time with the deepening of the number of layers of the network. But when training is performed, the increase of the number of layers of the network will make the iteration at this time and time increase significantly. Therefore, when we perform the hyper-segmentation operation, we can directly use the 14-layer convolutional network, which can guarantee the image quality and also improve the training efficiency and reduce the waste of resources.

Also, according to the above table, it can be found that when the scaling factor changes, the image quality also changes. As the scaling factor increases, the value of PSNR decreases and the value of SSIM gradually moves away from 1. Therefore, it can be concluded that the smaller the scaling factor is, the higher the image quality is.

References

- [1] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael. Learning low-level vision. *International journal of computer vision*, 40(1):25–50, 2000. 1
- [2] M. Irani and S. Peleg. Improving resolution by image registration. *CVGIP: Graphical models and image processing*, 53(3):231–239, 1991. 1
- [3] J. S. Isaac and R. Kulkarni, “Super-resolution techniques for medical image processing,” in *Proc. Int. Conf. Technologies Sustain. Develop.*, 2015, pp. 1–6.
- [4] M. W. Thornton, P. M. Atkinson, and D. a. Holland. Sub-pixel mapping of rural land cover objects from fine spatial resolution satellite sensor imagery using super-resolution pixel-swapping. *International Journal of Remote Sensing*, 27(3):473–490, 2006. 1
- [5] B. K. Gunturk, A. U. Batur, Y. Altunbasak, M. H. Hayes, and R. M. Mersereau. Eigenface-domain super-resolution for face recognition. *IEEE Transactions on Image Processing*, 12(5):597–610, 2003. 1
- [6] Kim, J., J. K. Lee, and K. M. Lee. "Accurate Image Super-resolution Using Very Deep Convolutional Networks." *Proceedings of the IEEE®Conference on Computer Vision and Pattern Recognition*. 2016, pp. 1646-1660.
- [7] C. Dong, C. C. Loy, K. He, and X. Tang, “Learning a deep convolutional network for image super-resolution,” in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 184–199.
- [8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
- [9] Zhihao Wang , Jian Chen , and Steven C. H. Hoi ,” Deep Learning for Image Super-resolution: A Survey”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.43, No.10, October 2021, pp.3369-3377
- [10] C. Dong, C. C. Loy, K. He, and X. Tang. Image super-resolution using deep convolutional networks. *TPAMI*, 2015. 1-4
- [11] He K, Zhang X, Ren s, et al. Deep residual learning for image recognition[C]. *IEEE Conference on Computer Vision*, 2015, 115(3):211-252.
- [12] Grubinger, M., P. Clough, H. Müller, and T. Deselaers. "The IAPR TC-12 Benchmark: A New Evaluation Resource for Visual Information Systems." *Proceedings of the OntoImage 2006 Language Resources For Content-Based Image Retrieval*. Genoa, Italy. Vol. 5, May 2006, p. 10.
- [13] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *Neural Networks*, *IEEE Transactions on*, 5(2):157–166, 1994. 3, 4