SINGLE IMAGE SUPER-RESOLUTION WITH DILATED CONVOLUTION BASED MULTI-SCALE INFORMATION LEARNING INCEPTION MODULE

Wuzhen Shi, Feng Jiang, Debin Zhao

School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China

ABSTRACT

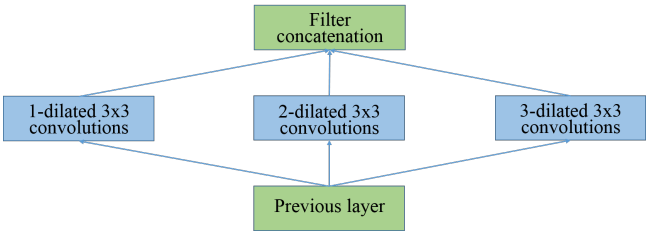
Traditional works have shown that patches in a natural image tend to redundantly recur many times inside the image, both within the same scale, as well as across different scales. Make full use of these multi-scale information can improve the image restoration performance. However, the current proposed deep learning based restoration methods do not take the multi-scale information into account. In this paper, we propose a dilated convolution based inception module to learn multi-scale information and design a deep network for single image super-resolution. Different dilated convolution learns different scale feature, then the inception module concatenates all these features to fuse multi-scale information. In order to increase the reception field of our network to catch more contextual information, we cascade multiple inception modules to constitute a deep network to conduct single image super-resolution. With the novel dilated convolution based inception module, the proposed end-to-end single image super-resolution network can take advantage of multi-scale information to improve image super-resolution performance. Experimental results show that our proposed method outperforms many state-of-the-art single image super-resolution methods.

Index Terms— Image super-resolution, convolutional neural network, multi-scale information, dilated convolution, inception module

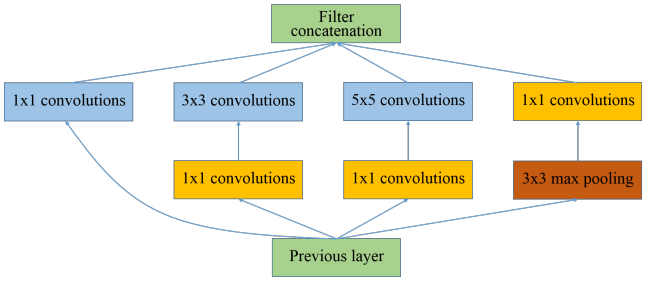
1. INTRODUCTION

In this paper, we focus on single image super-resolution (SR), which aims at recovering a high-resolution (HR) image from a given low-resolution (LR) one. It is always the research emphasis because of the requirement of higher definition video displaying, such as the new generation of Ultra High Definition (UHD) TVs. However, most video content would not at sufficiently high resolution for UHD TVs. As a result, we have to develop efficient SR algorithms to generate UHD content from lower resolutions [2].

Traditional single image SR methods always try to find a new image prior or propose a new way to use the existing image priors. A lot of image prior information have been explored in the image restoration literature, such as local smooth, non-local self-similarity and image sparsity. Based on the assumption that low and high resolution images have the same sparse representation, Yang et al. [3] use two coupled dictionaries to learn a nonlinear mapping between the LR and the HR images. In [4], Glasner et al. begin to use the image multi-scale information for single image SR and obtain state-of-the-art results.



(a) Dilated convolution based inception module



(b) Original GoogLeNet inception module [1]

Fig. 1. Comparison between the proposed dilated convolution based inception module and the original GoogLeNet inception module. Our proposed new inception module contains multiple different scale dilated convolution that makes it can learn multi-scale image information.

Recently, due to the availability of much larger training dataset and the powerful GPU implementation, deep learning based methods achieve great success in many fields, including both high level and low level computer vision problems. Look through the literature, most state-of-the-art single image SR methods are based on deep learning. The pioneering SR method is SRCNN proposed by Dong et al. [5, 6]. They establish the relationship between the traditional sparse-coding based SR methods and their network structure and demonstrate that a convolutional neural network (CNN) can learn a mapping from low resolution image to high resolution one in an end-to-end manner. Dong et al. successfully expand SRCNN for compression artifacts reduction by introducing a

feature enhancement layer [7]. Soon after, they proposed a fast SRCNN method, which directly maps the original low-resolution image to the high-resolution one [8]. Different from [5, 6, 7, 8], some works try to learn image high frequency details instead of the undegraded image. In [9], Kim et al. cascade many convolutional layers to form a very deep network to learn image residual.

Investigating an effective way to use multi-scale information is also important. The degraded image can be successful recovered is mainly based on the assumption that patches in a natural image tend to redundantly recur many times inside the image. However, it is not only exist in the same scale but also across different scales. It has been demonstrated that make full use of multi-scale information can improve the restoration result in traditional methods [4]. However, the multi-scale information has been little investigated in deep learning methods. In [9], Kim et al. try to train a multi-scale model for different magnification SR. It is a very rough tactics to exploit the scale information since they just put different scale image as input for training. Its success can be attribute to the powerful learning ability of CNN instead of the multi-scale information being considered in the network structure.

In this paper, we propose a dilated convolution based inception module to learn multi-scale information and design a deep network for single image SR. Fig. 1 makes a comparison between the proposed dilated convolution based inception module and the original GoogLeNet inception module. Our proposed new inception module contains multiple different scale dilated convolution that makes it can learn multi-scale image information. Furthermore, we cascade multiple dilated convolution based inception modules to constitute a deep network for single image SR. In short, the contributions of this work are mainly in three aspects: 1) we proposed a dilated convolution based inception module, which can learn multi-scale information with only single scale image input; 2) we design a novel deep network with the proposed dilated convolution based inception module for single image SR. 3) experimental results show that our proposed new method outperforms many state-of-the-art methods.

2. DILATED CONVOLUTION BASED INCEPTION MODULE

Dilated Convolution: It has been referred to in the past as convolution with a dilated filter [10, 11]. In [12], Yu et al. call it dilated convolution instead of convolution with a dilated filter to clarify that no dilated filter is constructed or represented. It can be formulated as

$$(F *_l k)(p) = \sum_{s+lt=p} F(s)\, k(t) \qquad (1)$$

where $*_l$ is called as a dilated convolution or an l-dilated convolution, F and k is a discrete function and a discrete filter of size $(2r+1)^2$, respectively. Three dilated convolutions have

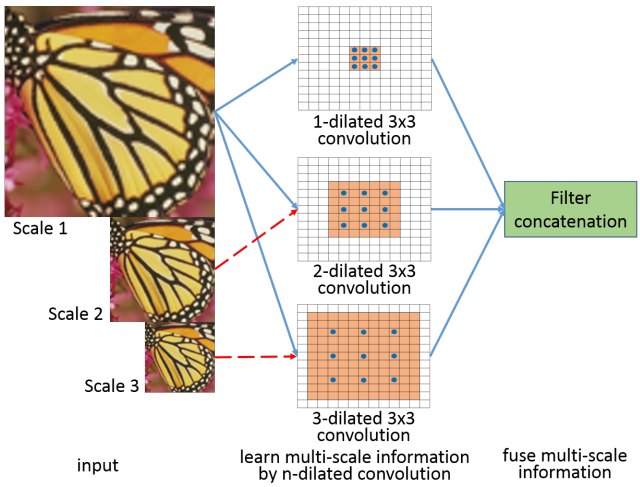


Fig. 2. Illustration of the proposed inception how to learn multi-scale information.

been shown in the middle part of Fig. 2, where k is a 3×3 filter and the kernel dilation factors are 1, 2 and 3, respectively. It shows that filters are dilated by inserting $dilated - 1$ zeros between filter elements for a given dilation factor. We refer the reader to [12] for more information about the dilated convolution.

Dilated Convolution based Inception Module: To make full use of the image multi-scale information, Szegedy et al. [1] proposed an inception module for image classification. As shown in Fig. 1b, the GoogLeNet inception module contains multiple convolution with different kernel size. It concatenates the outputs of these different size filter to fuse different scale information. With this multi-scale information learning structure, GoogLeNet achieves the new state-of-the-art performance for classification and detection in the ImageNet Large-Scale Visual Recognition Challenge 2014 (ILSVRC14). Inspired by this successful work, we propose a dilated convolution based inception module to learn multi-scale information for improving single image SR performance.

As shown in Fig. 1a, our new inception module first use three dilated convolution with kernel size 3×3 and dilated factor 1, 2 and 3, respectively, to operate on the previous layers. Then, we concatenate all these convolution outputs to fuse different scale information.

Insights: there are a lot of ways to learn the multi-scale information by deep network. For example, the GoogLeNet inception module uses different kernel size of convolution with different receptive field to operate on the previous layer. However, it is still operated on the same image scale space. The other choice is to operate on different scale image input. Farabet et al. [13] use this way to learn multi-scale feature for scene parsing. For single image SR, the lower solution image always has much more sharp detail information than the interpolation result. Therefore, it can improve the SR result by taking these small scale images into account. Fig. 2 shows the process of our proposed dilated convolution based inception module how to learn multi-scale information. The blue solid

lines indicate that all these dilated convolutions actually operate on the same scale image, while the red dash lines mean these convolutions with different dilated factors learn the corresponding scale image information. Furthermore, the output of the dilated convolution can keep the same size with its input, so we can fuse these different scale information through concatenation operator easily.

3. PROPOSED MULTI-SCALE INFORMATION LEARNING NETWORK STRUCTURE

The configuration of our proposed single image SR network structure is outlined in Fig. 3, which cascades $m+1$ dilated convolution based inception modules and $m+2$ common discrete convolutions. It can be explained as three phases, i.e. feature extraction, feature enhancement and image reconstruction. Since residual learning have been proven to be an effective way to accelerate the network convergence in recent works [9], we follow them to make our network predict image high frequency details instead of the HR image itself. The predicted image frequency details will be added to the input LR image to get the final desired HR one.

As most single image SR works done, we first up-sample the LR image to the desired size using bicubic interpolation. For the ease of presentation, we still call the interpolation result as a "low-resolution" image, although it has the same size as the HR image. For the feature extraction phase, n dilated convolution based inception modules operate on the LR input. The filter kernel size is $3 \times 3 \times c$, where c is the number of image channel, for the first inception module layer. The inception module combines different scale feature information through concatenation operator. To further fuse these multi-scale information, in the next we add a nonlinear transformation layer after the inception module layer, which contains $3n$ common convolution with filter kernel size $3 \times 3 \times 3n$.

Both the high level and low level vision works have proven the deeper the network the better the performance. Furthermore, the larger reception field of the network can catch more contextual information that gives more clues for predicting high frequency details. Therefore, we iterate m times the process of dilated convolution based inception modules following with a common convolution operator for the feature enhancement phase. In this phase, the filter kernel size is $3 \times 3 \times 3n$ as show in Fig. 3. In the image reconstruction phase, we use a single common convolution of size $3 \times 3 \times 3n$ to predict the high frequency details. Finally, the predicted high frequency details will add to the interpolation output to get the desired HR image.

3.1. Training

There are a lot of perceptually relevant characteristics based loss functions have been proposed in the literature. But for a fair comparison with SRCNN and FSRCNN, we adopt the mean square error (MSE) as the cost function of our network. Our goal is to train an end-to-end mapping F to predict high frequency detail $\hat{y} = F(x)$, where x is an interpolation result of the LR image and $\hat{y}$ is the estimated high frequency detail image. Given a training dataset $\{x_i, y_i\}_{i=1}^{N}$, the optimization objective is represented as

$$\min_{\theta} \frac{1}{2N} \sum_{i=1}^{N} \|F(x_i;\theta) - y_i\|_F^2 \tag{2}$$

where θ is the network parameters needed to be trained, $F(x_i;\theta)$ is the estimated high frequency image with respect to the interpolation result of a LR image. In the literature, people suggest to use to recently proposed Parametric Rectified Linear Unit (PReLU) as the activation function instead of the commonly-used Rectified Linear Unit (ReLU). But, in order to reduce parameters, ReLU is used after each convolution layer in our very deep network that has got a satisfactory result for comparison. We use the adaptive moment estimation (Adam) [14] to optimize all network parameters instead of the commonly used stochastic gradient descent one.

4. EXPERIMENTAL RESULTS

A lot of experiments have been done to show dramatic improvement in performance can be obtained by our proposed method. We give the experimental details and report the quantitative results on three popular dataset in this section. We name the proposed method as Multiple Scale Super-Resolution Network (MSSRNet).

4.1. Datasets for Training and Testing

It is well known that training dataset is very important for the performance of learning based image restoration methods. A lot of training dataset can be found in the literature. For example, SRCNN [5, 6] uses a 91 images dataset and VDSR [9] uses 291 images dataset. However, the 91 images dataset is too small to push a deep model to the best performance. Some image in the 291 image dataset are JPEG format, which are not optimal for the SR task. We follow FSRCNN [8] to use the General-100 dataset, which contains 100 bmp-format images (with no compression). We set the patch size as 48×48, and use data augmentation (rotation or flip) to prepare training data. Following FSRCNN, SRCNN and SCN, we use three dataset, i.e. Set5 [15] (5 images), Set14 [16] (14 images) and BSD200 [17] (200 images) for testing, which are widely used for benchmark in other works.

4.2. Implementing Details and Parameters

For each dilated convolution based inception module layer, we set $n = 8$. That is, there are 8 inception module per layer. For feature enhancement process, we iterate 5 inception modules to make the network deeper, i.e. $m = 5$. Filters are

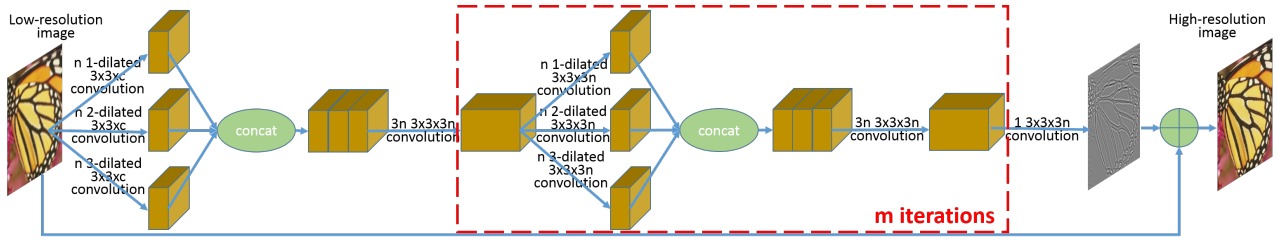


Fig. 3. The proposed single image super-resolution network structure with dilated convolution based multi-scale information learning inception module. It uses inception module, which contains different scale dilated convolution, to learn multi-scale information. Multiple inception modules cascade to constitute a deep network to predict high frequency detail information.

Table 1. The results of average PSNR (dB) and SSIM on the Set5 [15], Set14 [16] and BSD200 [17] dataset

Dataset	Scale	Bicubic	A+	SRF	SRCNN	SCN	FSRCNN	MSSRNet
		PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
Set5	×2	33.66/0.9299	36.55/0.9544	36.87/0.9556	36.34/0.9521	36.76/0.9545	37.00/0.9558	**37.33/0.9581**
	×3	30.39/0.9299	32.59/0.9088	32.71/0.9098	32.39/0.9033	33.04/0.9136	33.16/0.9140	**33.38/0.9178**
	×4	28.42/0.8104	30.28/0.8603	30.35/0.8600	30.09/0.8503	30.82/0.8728	30.71/0.8657	**31.10/0.8777**
Set14	×2	30.23/0.8687	32.28/0.9056	32.51/0.9074	32.18/0.9039	32.48/0.9067	32.63/0.9088	**32.89/0.9117**
	×3	27.54/0.7736	29.13/0.8188	29.23/0.8206	29.00/0.8145	29.37/0.8226	29.43/0.8242	**29.57/0.8282**
	×4	26.00/0.7019	27.32/0.7471	27.41/0.7497	27.20/0.7413	27.62/0.7571	27.59/0.7535	**27.83/0.7631**
BSD200	×2	29.70/0.8625	31.44/0.9031	31.65/0.9053	31.38/0.9287	31.63/0.9048	31.80/0.9074	**32.08/0.9118**
	×3	27.26/0.7638	28.36/0.8078	28.45/0.8095	28.28/0.8038	28.54/0.8119	28.60/0.8137	**28.78/0.8188**
	×4	25.97/0.6949	26.83/0.7359	26.89/0.7368	26.73/0.7291	27.02/0.7434	26.98/0.7398	**27.17/0.7489**
Avg.		28.80/0.8151	30.53/0.8491	30.67/0.8505	30.39/0.8474	30.81/0.8542	30.88/0.8537	**31.13/0.8596**

initialized using the initializer proposed by He et al. [18] with values sampled from the Uniform distribution. For the other hyper-parameters of Adam, we follow [19] to set the exponential decay rates for the first and second moment estimate to 0.9 and 0.999, respectively. Each model was trained only 100 epochs and each epoch iterates 2000 times with batch size of 64. We set a lager learning rate in the initial training phase to accelerate convergence, then decrease it gradually to make the model more stable. Therefore, the learning rates are 0.001, 0.0001 and 0.00001 for the first 50 epochs, the 51 to 80 epochs and the last 20 epochs, respectively. Our model is implemented by the MatConvNet package [17].

4.3. Comparisons with State-of-the-Art Methods

We compare our method with five state-of-the-art learning based SR algorithms that rely on external databases, namely the A+ [20], SRF [21], SRCNN [5, 6], FSRCNN [8] and SCN [22]. A+ and SRF are two state-of-the-art traditional methods, while SRCNN, FSRCNN and SCN are three newest popular deep learning based single image super-resolution image methods. In Table 1, we provide a summary of quantitative evaluation on several datasets. The results of other five methods are the same as reported at FSRCNN. Our method outperforms all previous methods in these datasets. Compare with the newest FSRCNN, our method can improve roughly 0.33 dB, 0.22Db and 0.37 dB on average with respect to up-sample factor 2, 3 and 4 on Set5 dataset, respectively. Over the three dataset and three up-sample factor, our MSSRNet can improve roughly 2.33 dB, 0.6 dB, 0.46 dB, 0.74 dB, 0.32 dB and 0.25 dB on average, in comparison with Bicubic, A+, SRF, SRCNN, SCN and FSRCNN, respectively. To get better performance, we can increase the network depth (larger m), which is called deeper is better in the literature, and the network width with larger n. In our experiments, we have implemented the fatter network with $n = 16$ and $n = 32$, and the deeper network with $m = 10$ and $m = 15$. Both the deeper and the fatter networks show PSNR and SSIM gain. The reader can download our test code[1] to get more quantitative and qualitative results.

5. CONCLUSION

In this paper, we use deep learning technology to solve the single image super-resolution problem. We first propose a dilated convolution based inception module, which can learn multi-scale information from the single scale input image. We design a deep network, which cascades multiple dilated convolution based inception modules, for single image super-resolution. Experimental results show that the proposed method outperforms many state-of-the-art ones. As future work we plan to explore MSSRNet for video processing.

6. ACKNOWLEDGEMENTS

This work has been supported in part by the Major State Basic Research Development Program of China (973 Program 2015CB351804), the National Science Foundation of China under Grant No. 61572155.

[1] https://github.com/wzhshi/MSSRNet

7. REFERENCES

[1] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[2] Armin Kappeler, Seunghwan Yoo, Qiqin Dai, and Aggelos K Katsaggelos, "Video super-resolution with convolutional neural networks," *IEEE Transactions on Computational Imaging*, vol. 2, no. 2, pp. 109–122, 2016.

[3] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma, "Image super-resolution via sparse representation," *IEEE transactions on image processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[4] Daniel Glasner, Shai Bagon, and Michal Irani, "Super-resolution from a single image," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 349–356.

[5] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Learning a deep convolutional network for image super-resolution," in *European Conference on Computer Vision*. Springer, 2014, pp. 184–199.

[6] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[7] Chao Dong, Yubin Deng, Chen Change Loy, and Xiaoou Tang, "Compression artifacts reduction by a deep convolutional network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 576–584.

[8] Chao Dong, Chen Change Loy, and Xiaoou Tang, "Accelerating the super-resolution convolutional neural network," in *European Conference on Computer Vision*. Springer, 2016, pp. 391–407.

[9] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1646–1654.

[10] Matthias Holschneider, Richard Kronland-Martinet, Jean Morlet, and Ph Tchamitchian, "A real-time algorithm for signal analysis with the help of the wavelet transform," in *Wavelets*, pp. 286–297. Springer, 1990.

[11] Mark J Shensa, "The discrete wavelet transform: wedding the a trous and mallat algorithms," *IEEE Transactions on signal processing*, vol. 40, no. 10, pp. 2464–2482, 1992.

[12] Fisher Yu and Vladlen Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[13] Clément Farabet, Camille Couprie, Laurent Najman, and Yann LeCun, "Scene parsing with multiscale feature learning, purity trees, and optimal covers," *arXiv preprint arXiv:1202.2160*, 2012.

[14] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[15] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *Proceedings of the British Machine Vision Conference*. 2012, pp. 135.1–135.10, BMVA Press.

[16] Roman Zeyde, Michael Elad, and Matan Protter, "On single image scale-up using sparse-representations," in *International conference on curves and surfaces*. Springer, 2010, pp. 711–730.

[17] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*. IEEE, 2001, vol. 2, pp. 416–423.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[19] Jun Guo and Hongyang Chao, "Building dual-domain representations for compression artifacts reduction," in *European Conference on Computer Vision*. Springer, 2016, pp. 628–644.

[20] Radu Timofte, Vincent De Smet, and Luc Van Gool, "A+: Adjusted anchored neighborhood regression for fast super-resolution," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 111–126.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[22] Z Wang, D Liu, J Yang, W Han, and T Huang, "Deeply improved sparse coding for image super-resolution. arxiv preprint," *arXiv preprint arXiv:1507.08905*, vol. 2, no. 3, pp. 4, 2015.