MULTI-SCALE DENSE NETWORK FOR SINGLE-IMAGE SUPER-RESOLUTION

Chia-Yang Chang, Shao-Yi Chien

Media IC and System Lab
Graduate Institute of Electronics Engineering and Department of Electrical Engineering
National Taiwan University
BL-421, No. 1, Sec. 4, Roosevelt Rd., Taipei 10617, Taiwan

ABSTRACT

Recently, deep neural networks have led to tremendous advances in image super-resolution. As a well-known one-to-many inverse problem, the deep learning based methods tackle this issue via large receptive field. By that, the deep network could infer each output pixel from sufficient context information. However, most existing studies use larger kernel size or design a very deep network model to attain sufficient receptive field. The computational cost dramatically increments along with the training difficulty. Concerning this problem, the goal of this paper is to design an effective and trainable convolutional neural network. We proposed a multi-scale dense network (MSDN) which is composed of deep concatenation and basic blocks, namely multi-scale dense block (MSDB). The proposed MSDB use different dilated convolutions to gather multi-scale information; meanwhile concatenating the different dilated convolution results magnify the receptive field of a single layer. To facilitate the training difficulty, there are the dense skip connections in the proposed MSDB. Moreover, the deep concatenation and global skip connection are also adopted for improving training furthermore. Consequently, we achieve a large receptive field network without deeper structure. The experiments indicate that the quality of the proposed MSDN yields the state-of-the-art result.

Index Terms— Single-image super-resolution, learning-based super-resolution, dilation convolution, image restoration

1. INTRODUCTION

The goal of single image super-resolution (SISR) is enhancing resolution of a low-resolution image. This technology is widely used in surveillance, satellite, and editing the old pictures. As a classical problem, lots of SISR methods were investigated over the last decades. Those include interpolation-based, self-similarity-based [1], and dictionary-based methods [2].

With the huge success in computer vision and signal processing, there are several deep learning based SISR methods proposed. Among them, Dong *et al.* first proposed a simple three-layer structure called SRCNN [3]. Despite that the quality does not take the lead of traditional methods, the simple end-to-end learning convolutional neural network (CNN) method shows the great ability of mapping low-resolution image to high-resolution image. After that, Kim *et al.* address that the insufficient receptive field of the SRCNN causes quality degradation. Thus, they proposed very deep neural network(VDSR) [4] which has 20 layers with 3×3 convolutional kernels. Compared to the SRCNN(13×13), the receptive field of VDSR is 41×41. Also, with the assistance of some training tricks, such as global skip connection and adjustable gradient clipping, the VDSR outperform the SRCNN and traditional methods. Subsequently, Tai *et al.* proposed DRRN [5] and MemNet [6]. The DRRN contains multi-path local skip connections and recursive units, and the depth of network reach to 52 layers. Based on the recursive manner, the MemNet use the long-/short-term memory as a deep concatenation strategy and the depth of network expands to 80 layers. However, both the DRRN and the MemNet take bicubic image as network input, the exact receptive field to the original low-resolution image grows slower than the network depth.

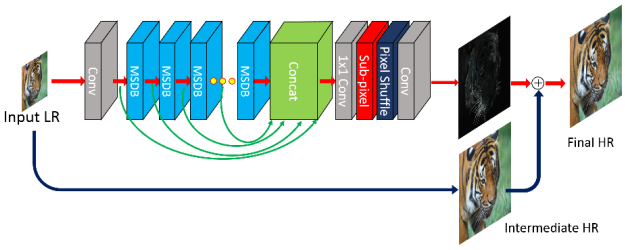


Fig. 1. Overview of the proposed multi-scale dense network (MSDN). The MSDN mainly consists of three stages: 1) initial feature extraction stage, 2) multi-scale dense block (MSDB) feature extraction stage, and 3) reconstruction stage.

Conversely, Lai *et al.* developed LapSRN [7] from the Laplacian pyramid, and with the progressive reconstruction LapSRN shows that the neural network could reconstruct image well without the pre-upscaling low-resolution image. Meanwhile, Ledig *et al.* proposed SRResNet [8] based on residual block [9]. And several methods are proposed

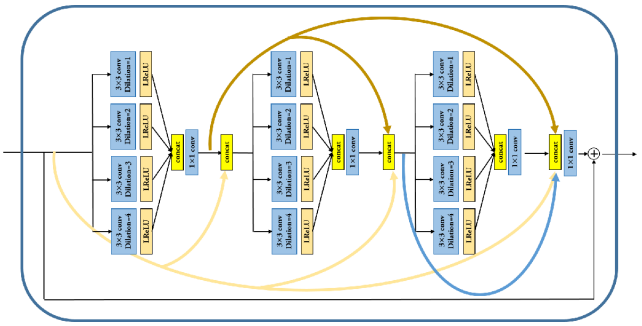


Fig. 2. The overview of the proposed MSDB. There are two major parts in MSDB : 1)The dilation inception-like block. 2) The dense skip connection.

based on SRResNet, such as EDSR [10] ,CARN [11], and IDN [12]. There are also several super-resolution methods [13][14] which are inspired by DenseNet [15].

According to the recent challenges with respect to super-resolution [16][17], the performance improvements of super-resolution deep networks have led to increases in their depth. That means the quality improvements come from either receptive field or parameter number. However, there are some analytic studies in [7][5][11] which show that the parameter number is not the key factor to performance gain. Those studies illustrate the importance of the receptive field.

To the mid-level vision tasks, the receptive field is also a critical problem, such as [18] [19]. Among them, Chen *et al.* design the large receptive field network with the help of the atrous/dilation convolution [20], and achieve remarkable performance in semantic segmentation.

Inspired by the dilation convolution, we proposed a multi-scale dense block to achieve a large receptive field network. As shown in Figure 2. The proposed MSDB consists different dilation convolution similar to inception block [21]. To handle the sparse sampling of dilation convolution, we adopt dense connections to attain dense sampling to the input hidden features. The proposed network contains eight MSDB, and output of each MSDB would be concatenated to the final reconstruction stage. Overview of the proposed multi-scale dense network is shown in Figure 1. Li *et al.* share the same idea [22] to the proposed MSDN. However, Li *et al.* adopt 5×5 filters as a larger-scale feature extractor, which is more inefficient than dilation convolution and increases parameter number. The dilation convolution is also adapted to handle image denoise problem in the IRCNN [23] proposed by Zhang *et al.* and DDRN [24] proposed by Wang *et al.* . Nevertheless, they both simply replace the 3×3 with dilation convolution. Shi *et al.* [25] and Lin *et al.* [26] both proposed super-resolution network with dilation convolution. However, the quality of their proposed network do not attain the state-of-the-art result

The main contributions of the proposed method are summarized as follows: **1)** we proposed a large receptive field network for single-image super-resolution. The proposed MSDN gather large-scale information via multi-scale dilation convolutions. The large receptive field assists MSDN in attaining the state-of-the-art result. **2)** To handle the sparse sampling issue of large dilation convolution, the proposed MSDB contains branches as same as inception block and concatenates the different scale features together. Furthermore, the dense connections confirm that the output features of MSDB densely sample the input features. **3)** With the MSDB and the deep concatenation, the proposed MSDN achieve state-of-the-art results with comparative fewer layers. It means that the proposed technologies in MSDN could improve the representational ability of the network.

2. PROPOSED METHOD

2.1. Network Architecture

As shown in Figure 1, the proposed MSDN mainly consists of three stages: initial feature extraction, multi-scale dense block feature extration, and reconstruction stage.

There is only one convolutional layer to extract the initial hidden feature H_{init} from the low-resolution input image I_{LR}. The initial feature extraction layer is formulated as:

$$H_{init} = \phi\left(\Theta_{3\times3}\left(I_{LR}, w_{init}\right)\right) \tag{1}$$

where w_{init} denotes the weights of the initial feature extraction layer. $\Theta_{3\times3}$ and ϕ are the 3×3convolution and activation function operator. After that, H_{init} is the input of multi-scale feature extraction with the proposed MSDB. The successive hidden feature from the first MSDB would be

$$H_{ms_1} = \beta_1\left(H_{init}, w_{ms_1}\right) \tag{2}$$

where the β_1 denotes the operation of first multi-scale dense block which will be described in next subsection. And w_{ms_1} is the weights of the $MSDB_1$. In the MSDB feature extration stage, there will be N multi-scale dense block to extract sufficient feature from the large receptive field. The formulation of MSDB feature extration stage is as following:

$$H_{ms_N} = \beta_N\left(..\beta_n\left(\beta_1\left(H_{init}, w_{ms_1}\right)..., w_{ms_n}\right)..., w_{ms_N}\right) \tag{3}$$

where β_n and w_{ms_n} are the operation function and weights of n-th MSDB. And the multi-scale hidden features extract from that is denoted by H_{ms_n}.

After the initial feature extraction and multi-scale dense block feature extraction stages, all of the hidden features $(H_{init}, H_{ms_1}, ..., H_{ms_N})$ would be concatenated and pass to a depth-wise convolutional layer. The bottleneck layer reduces the dimension of the concatenation hidden features, and send to sub-pixel convolutional layer and pixel shuffler to enlarge the hidden features to the desired resolution; the detail could be found in ESPCN [27]. Refer to VDSR [4]; we adopt the global skip connection to add the bicubic high-resolution image to the network output. Therefore, the high-resolution hidden features would generate a one-channel

Table 1. The experiment on different dilation inception block of the proposed MSDB.

dilation = 1	√				√	√	√				√
dilation = 2		√			√			√	√		√
dilation = 3			√			√		√		√	√
dilation = 4				√			√		√	√	√
PSNR	29.01	28.90	28.82	28.75	29.03	29.02	29.08	29.01	29.05	29.06	29.22

Table 2. Different structure experiments.

DenseBlock	√		√	√	√
ResBlock		√			
Deep concat	√	√	√		
Residual	√	√		√	
PSNR	29.22	29.10	29.03	29.15	28.92

gray-scale residual image or a three-channel color residual image. This strategy is to reduce the mapping difficulty from dense high-resolution image to sparse high-resolution residual.

2.2. Multi-Scale Dense Block

As described in Sec.1, in order to increase receptive field, we proposed the multi-scale dense block. The concept of MSDB is shown in Figure 2. There are two major parts of MSDB. One is the dilation inception-like block; which gathers multi-scale information with dilation convolutions. The dilation inception block consists of four different dilation settings convolution. Another part is dense skip connection; which compensates the sparse sampling of large dilation convolution.

The first dilation inception block would extract input hidden feature as follows:

$$H^n_{d_1} = \Theta_{1\times1}\left([H^n_{d_{11}}, H^n_{d_{12}}, H^n_{d_{13}}, H^n_{d_{14}}], w^n_{b_1}\right) \quad (4)$$

$$H^n_{d_{1k}} = \sigma\left(\Theta_{3\times3_{dk}}\left(H_{ms_{n-1}}, w^n_{d_{1k}}\right)\right) \quad (5)$$

where the $\Theta_{3\times3_{dk}}$ denotes the k dilation convolution, and $w^n_{d_{1k}}$ is the corresponding weights. The [] is concatenating operation, and the output of first dilation inception block is denoted by $H^n_{d_1}$. After that, the following m-th dilation inception block take input as:

$$\Theta_{1\times1}\left([H_{ms_{n-1}}, H^n_{d_1}, ..., H^n_{d_{m-1}}], w^n_{c_m}\right) \quad (6)$$

The dense connection would concatenate all the output features of previous dilation inception block including the input features of this MSDB. Then a 1×1 convolution reduce the feature dimension to prevent the growth of computing requirements.

2.3. Implementation Details

In the proposed multi-scale dense net, the filter number of the initial feature extraction layer is 32, and each dilation convolution in MSDB also has 32 filters. After the concatenating of dilation inception blocks, the 1×1 convolution will reduce the feature dimension back to 32. Also, the bottleneck convolution in the reconstruction stage has 32 filters, either. The number of filters in the sub-pixel layer depends on the scaling factor. The negative slope is 0.2 to all the LeakyReLUs. The final network contains 8 MSDB, and each MSDB has four dilation inception block. It means that the depth of the proposed MSDN is about 34 layers. The receptive field of each output pixel is 261×261 to the input low-resolution image.

We take `DIV2K` as the training dataset. In each training batch, 8 low-resolution patches with the size of 128128 would be cropped from dataset randomly. The rotation, horizontal, and vertical flipping are also adopted as data augmentation. The optimizer we used is Adam by setting β_1 = 0.9, β_2 = 0.999, and $\epsilon = 10^{-8}$. The learning rate begins at 10^{-5} and is halved every 10000 iterations. The weights of all the filters are initialized by [28], and biases are set to zeros.

3. EXPERIMENTAL RESULT

3.1. Network Analysis

To understand the properties of the proposed multi-scale dense block, we first analyze the dilation settings. We maintain the total number of filters in the dilation inception block. It means that if dilation inception block contains T different scales, there would be $128/T$ in each dilation convolution. The experiments are summarized in Table 1, and the PSNR results are all evaluated in `BSD100` benchmark with scaling factor 3×. We found that if there is only one dilation convolution in MSDB, the simple 3×3 and dilation $= 1$ would have the better result. Conversely, when the dilation inception block contains more than one dilation setting, the large dilation convolution would be more favorable for the network performance.

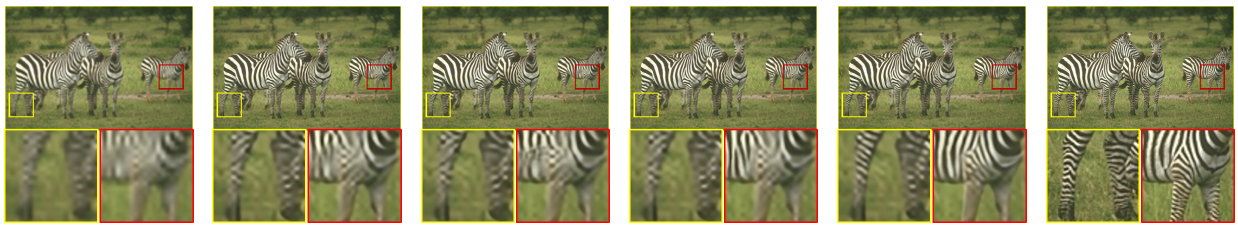
We also analyze the network performance of the different structures. The basis multi-scale block could be DenseBlcok or ResBlcok. Both of them are the popular basic unit in the deep learning literature. However, in the experiment of our multi-scale super-resolution network, the DenseBlock is slightly better than ResBlock. The performance increments with deep concatenation and global skip connection are also summarized in Table 2.

3.2. Comparison with the-state-of-the-arts

Table 3 summarizes the quantitative comparison of the proposed MSDN to the-state-of-the-art. There are VDSR [4], LapSRN [7], DRRN [5], IDN [12], and MSRN [22]. We calculate the PSNR and SSIM on the well-know benchmarks; which are `Set5`, `Set14`, `BSD100`, and `Urban100`. As de-

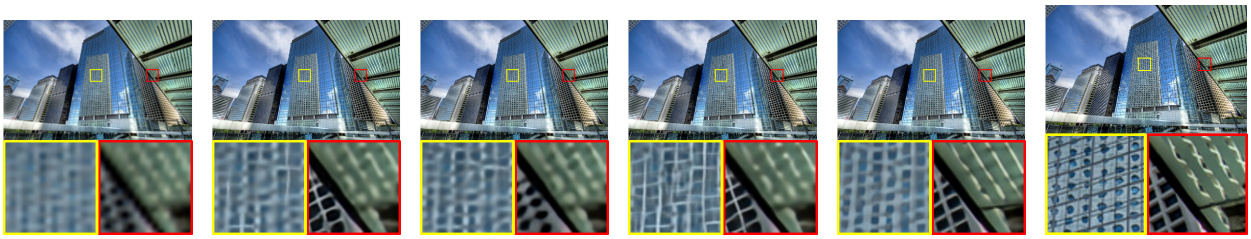
Table 3. Quantitative comparisons of state-of-the-art methods: PSNR/SSIMs for $\times 2$, $\times 3$ and $\times 4$ in benchmark.

Dataset	scale	Bicubic	VDSR [4]	LapSRN [7]	DRRN [5]	IDN [12]	MSRN [22]	Ours
Set5	$\times 2$	33.66 / 0.9299	37.53 / 0.9587	37.52 / 0.9591	37.74 / 0.9591	37.83 / 0.9600	38.08 / **0.9605**	**38.13 / 0.9605**
	$\times 3$	30.39 / 0.8682	33.66 / 0.9213	33.81 / 0.9220	34.03 / 0.9244	34.11 / 0.9253	34.38 / 0.9262	**34.46 / 0.9272**
	$\times 4$	28.42 / 0.8104	31.35 / 0.8838	31.54 / 0.8852	31.68 / 0.8888	31.82 / 0.8903	32.07 / 0.8903	**32.32 / 0.8943**
Set14	$\times 2$	30.24 / 0.8688	33.03 / 0.9124	32.99 / 0.9124	33.23 / 0.9136	33.30 / 0.9148	33.74 / 0.9170	**33.87 / 0.9183**
	$\times 3$	27.55 / 0.7742	29.77 / 0.8314	29.79 / 0.8325	29.96 / 0.8349	29.99 / 0.8354	30.34 / 0.8395	**30.51 / 0.8403**
	$\times 4$	26.00 / 0.7027	28.01 / 0.7674	28.09 / 0.7700	28.21 / 0.7721	28.25 / 0.7730	28.60 / 0.7751	**28.82 / 0.7853**
BSD100	$\times 2$	29.56 / 0.8431	31.90 / 0.8960	31.80 / 0.8952	32.05 / 0.8973	32.08 / 0.8985	32.23 / **0.9013**	**32.34** / 0.9011
	$\times 3$	27.21 / 0.7385	28.82 / 0.7976	28.82 / 0.7980	28.95 / 0.8004	28.95 / 0.8013	29.08 / 0.8041	**29.22 / 0.8062**
	$\times 4$	25.96 / 0.6675	27.29 / 0.7251	27.32 / 0.7275	27.38 / 0.7284	27.41 / 0.7297	27.52 / 0.7273	**27.70 / 0.7314**
Urban100	$\times 2$	26.88 / 0.8403	30.76 / 0.9140	30.41 / 0.9103	31.23 / 0.9188	31.27 / 0.9196	32.22 / 0.9326	**32.51 / 0.9342**
	$\times 3$	24.46 / 0.7349	27.14 / 0.8279	27.07 / 0.8275	27.53 / 0.8378	27.42 / 0.8359	28.08 / 0.8554	**28.63 / 0.8593**
	$\times 4$	23.14 / 0.6577	25.18 / 0.7524	25.21 / 0.7562	25.44 / 0.7638	25.41 / 0.7632	26.04 / 0.7896	**26.25 / 0.7931**



(a) Bicubic. (b) VDSR[4]. (c) LapSRN[7]. (d) DRRN[5]. (e) Our MSDN. (f) Ground Truth.

Fig. 3. Visual comparison in scaling factor 3. "253027.jpg" image from `BSD100` benchmark.



(a) Bicubic. (b) VDSR[4]. (c) LapSRN[7]. (d) DRRN[5]. (e) Our MSDN. (f) Ground Truth.

Fig. 4. Visual comparison in scaling factor 4. "img_061.png" image from `BSD100` benchmark.

scribed in Sec 1, VDSR [4] and DRRN [5] use bicubic interpolation result as input image. Therefore, the exact receptive fields are 41×41 and 103×103 divided by scaling factor. And the LapSRN [7] use the special structure, the receptive fields are about 15×15, 20×20, and 22×22 in scaling $2\times$, $3\times$, and $4\times$. And IDN [12] and MSRN [22] have 53×53 and 69×69 respectively. The proposed MSDN have the 261×261 receptive filed. Note that the parameter number of ours is half than that of MSRN [22]. The receptive field is an essential factor to the hard cases; especially the scaling $4\times$ in the `BSD100` and `Urban100` benchmark. There are visual comparisons in Figure 3 and Figure 4. The crop regions are high-frequency sub-images; which usually cause aliasing in the super-resolution problem. However, with the help of large receptive field, the proposed MSDN could refer to the high-resolution output pixels based on sufficient context information. It could be found at Figure 3(e) and Figure 4(e).

4. CONCLUSION

In this paper, we presented a large receptive field network for single-image super-resolution; which is based on the multi-scale dense block. The proposed MSDB indicated that gathering sufficient context information is very useful for network performance. With the help of dilation inception block, the receptive field of MSDN reaches 261×261 via only 34 layers. The computational cost and number of parameters are fewer than the-state-of-the-art relatively, and the results beyond them at the same time. In the future, the recursive manner could be adopted into the proposed MSDN. The weights of each MSDB could be shared; that will reduce lots number of parameters. Furthermore, we think that the different dilation convolution in one dilation inception block could also share weights to each other. Lots of traditional vision methods have already demonstrated that the scale-invariant is a good property for feature extraction.

5. REFERENCES

[1] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja, "Single image super-resolution from transformed self-exemplars," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5197–5206.

[2] Radu Timofte, Vincent De Smet, and Luc Van Gool, "A+:

Adjusted anchored neighborhood regression for fast super-resolution," in *Asian Conference on Computer Vision*, pp. 111–126. 2014.

[3] 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.

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

[5] Ying Tai, Jian Yang, and Xiaoming Liu, "Image super-resolution via deep recursive residual network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, vol. 1.

[6] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu, "Memnet: A persistent memory network for image restoration," in *IEEE International Conference on Computer Vision*, 2017, pp. 4539–4547.

[7] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 624–632.

[8] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[10] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, vol. 1, p. 3.

[11] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn, "Fast, accurate, and, lightweight super-resolution with cascading residual network," *arXiv preprint arXiv:1803.08664*, 2018.

[12] Zheng Hui, Xiumei Wang, and Xinbo Gao, "Fast and accurate single image super-resolution via information distillation network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 723–731.

[13] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao, "Image super-resolution using dense skip connections," in *IEEE International Conference on Computer Vision*, 2017, pp. 4809–4817.

[14] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu, "Residual dense network for image super-resolution," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[15] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten, "Densely connected convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, vol. 1, p. 3.

[16] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Yang Ming-Hsuan, and et al., "Ntire 2017 challenge on single image super-resolution: Methods and results," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

[17] Radu Timofte, Gu Shuhang, Wu Jiqing, Luc Van Gool, and et al., "Ntire 2018 challenge on single image super-resolution: Methods and results," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018.

[18] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, "Pyramid scene parsing network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2881–2890.

[19] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[20] Qifeng Chen, Jia Xu, and Vladlen Koltun, "Fast image processing with fully-convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, vol. 9, pp. 2516–2525.

[21] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning.," in *National Conference on Artificial Intelligence*, 2017, vol. 4, p. 12.

[22] Juncheng Li, Faming Fang, Kangfu Mei, and Guixu Zhang, "Multi-scale residual network for image super-resolution," in *European Conference on Computer Vision*, 2018, pp. 517–532.

[23] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang, "Learning deep cnn denoiser prior for image restoration," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, vol. 2.

[24] Tianyang Wang, Mingxuan Sun, and Kaoning Hu, "Dilated deep residual network for image denoising," in *IEEE International Conference on Tools with Artificial Intelligence*, 2017, pp. 1272–1279.

[25] Wuzhen Shi, Feng Jiang, and Debin Zhao, "Single image super-resolution with dilated convolution based multi-scale information learning inception module," in *IEEE International Conference on Image Processing*. IEEE, 2017, pp. 977–981.

[26] Guimin Lin, Qingxiang Wu, Lida Qiu, and Xixian Huang, "Image super-resolution using a dilated convolutional neural network," *Neurocomputing*, vol. 275, pp. 1219–1230, 2018.

[27] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1874–1883.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *IEEE International Conference on Computer Vision*, 2015, pp. 1026–1034.