Multi-scale Information Distillation Network for Image Super Resolution in NSCT Domain

Yu Sang(✉), Jinguang Sun, Simiao Wang, Yanfei Peng, Xinjun Zhang, and Zhiyang Yang

School of Electronic and Information Engineering, Liaoning Technical University, Huludao 125105, China
sangyu2008bj@sina.com

Abstract. Deep learning based methods have dominated super-resolution (SR) field due to their remarkable performance in terms of effectiveness and efficiency. In this paper, we propose a new multi-scale information distillation network (MSID-N) in the non-subsampled contourlet transform (NSCT) domain for single image super resolution (SISR). MSID-N mainly consists of a series of stacked multi-scale information distillation (MSID) blocks to fully exploit features from images and effectively restore the low resolution (LR) images to high-resolution (HR) images. In addition, most previous methods predict the HR images in the spatial domain, producing over-smoothed outputs while losing texture details. Thus, we integrate NSCT and demonstrate the superiority of NSCT over wavelet transform (WT), and formulate the SISR problem as the prediction of NSCT coefficients, which is able to further make MSID-N preserve richer structure details than that in spatial domain. The experimental results on three standard image datasets show that our proposed method is capable of obtaining higher PSNR/SSIM values and preserving complex edges and curves better than other state-of-the-art methods.

Keywords: Single Image Super Resolution (SISR) · Multi-scale information distillation network · Non-subsampled Contourlet Transform (NSCT) · Convolutional Neural Networks (CNNs)

1 Introduction

SISR is an important low-level vision task which has high practical value in many application fields such as remote sensing, medical imaging and object detecting. It aims at reconstructing a HR image from a single LR image, which is an ill-posed inverse problem.

In recent years, CNNs-based models [1–20] significantly improve the super resolution (SR) quality from the first SRCNN [1] to the latest RCAN [17], which are remarkably better than conventional SR methods. The performance of SRCNN was limited by its shallow structure. To achieve higher performance the networks are tend to be deeper and deeper, Kim et al. proposed the VDSR [3] model with a deeper structure. Recently, some very deep models have been proposed such as EDSR [4] and RCAN, which achieves very pleasing performance on super-resolution tasks. Moreover,

T. Gedeon et al. (Eds.): ICONIP 2019, LNCS 11954, pp. 50–59, 2019.
https://doi.org/10.1007/978-3-030-36711-4_5

super-resolution models integrated with dense connections have been proposed, such as SRDenseNet [8] and MemNet [9], which boosts the performance further more. In addition, some more effective CNN-based SR methods construct the entire network by connecting a series of identical feature extraction modules such as MSRN [14], RDN [15], IDN [16], indicating the capability of each block plays a crucial role.

The above SR methods are conducted in the spatial domain. By contrast, SR in the transform domain can preserve the image's context and texture information in different layers to produce better SR results. With that in mind, Guo et al. [18] designed a deep wavelet super-resolution (DWSR) network to acquire HR image by predicting "missing details" of wavelet coefficients of the LR image. Later, the same team [19] integrated discrete cosine transformation (DCT) into CNN and put forward an orthogonally regularized deep network. In addition, Huang et al. [20] applied WT to CNN-based face SR to validate that this method can accurately capture global topology information and local textural details of faces. The existing models received excellent performance in terms of peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) in the SISR problem.

In this paper, we present a novel CNN architecture in the NSCT domain for SISR. The main contributions are as follows:

(1) We propose multi-scale information distillation (MSID) block to fully exploit features from images; and MSID-N is mainly formed by multiple MSID blocks to effectively restore the LR images.
(2) We integrate NSCT and demonstrate the superiority of NSCT over WT, and formulate the SISR problem as the prediction of NSCT coefficients, which is able to make MSID-N preserve richer detail information than that in spatial domain.
(3) We evaluate the proposed method with three standard image datasets. The qualitative and quantitative results confirm that our method is capable of obtaining data with higher PSNR/SSIM values and preserving complex edges and curves better than other state-of-the-art methods.

2 Proposed Method

In this section, we will first describe the architecture of our proposed MSID-N. After that, we will provide a brief introduction to the proposed MSID block, followed by the description of NSCT domain.

2.1 Network Architecture

As shown in Fig. 1, our MSID-N consists of two parts, the shallow feature extraction (SFE) module and the deep feature extraction (DFE) module. Let's denote the I^{LR} and I^{HR} as the LR images and HR images respectively. Our ultimate goal is to learn an end-to-end mapping function F between I^{LR} and I^{HR}. So, we solve the following problem:

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} L^{SR}\left(F_{\theta}\left(I_i^{LR}\right), I_i^{HR}\right), \tag{1}$$

where $\theta = \left\{w^1, w^2, \cdots, w^p, b^1, b^2, \cdots, b^p\right\}$ denotes the weights and bias of our p convolutional layers. N is the number of training samples. L^{SR} is the loss function for minimizing the difference between the I^{LR} and I^{HR}.

The mean square error (MSE) function is the most widely-used objective optimization function in image super-resolution [2, 7, 9]. However, Lim et al. [14] have experimentally demonstrated that training with MSE loss is not a good choice. In order to avoid introducing unnecessary training tricks and reduce computations, we use the mean absolute error (MAE) function L^{SR} as a better alternative, as defined below

$$L^{SR}\left(F_{\theta}\left(I_i^{LR}\right), I_i^{HR}\right) = \frac{1}{N} \sum_{i=1}^{N} \left\| I_i^{LR} - I_i^{HR} \right\|_1. \tag{2}$$

Specially, we use two convolution layers to extract the shallow feature M_0 from the noisy seismic signals. So we can have

$$M_0 = H_{SFE1}\left(H_{SFE2}\left(I^{LR}\right)\right), \tag{3}$$

where H_{SFE1} and H_{SFE2} denote the convolution operation of the two layers in SFE module respectively. After shallow feature module, the shallow feature M_0 is used for DFE module, which contains a set of cascaded MSID blocks. Each MSID block can gather more information as much as possible and distill more useful information. After that we use a 1×1 convolutional layer to adaptively control the output information. We name this operation as feature fusion formulated as

$$M_{GF} = H_{GFF}([M_1, M_2, \cdots, M_D]), \tag{4}$$

where $[M_1, M_2, \cdots, M_D]$ denotes the concatenation of feature maps produced by MSID blocks 1, 2, …, D. H_{GFF} is a composite function of 1×1 convolutional layer. Global residual learning is utilized after feature fusion to obtain the feature maps I_{Output}, which can be formulated as

$$I_{Output} = M_{GF} + M_0. \tag{5}$$

In our MSID-N, all convolutional layers have 64 filters, except that in feature fusion, whose has 128 filters.

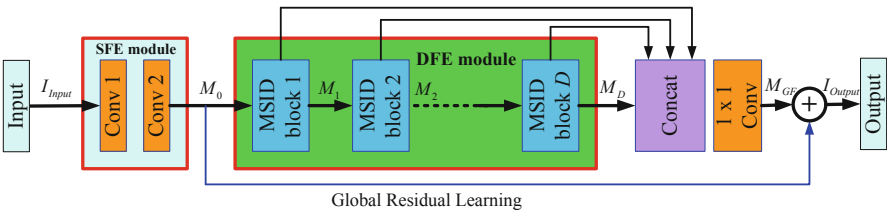


Fig. 1. The architecture of our proposed MSID-N.

2.2 MSID Block

The proposed MSID block is shown in Fig. 2. Each MSID block can be divided into two parts, which are used for exploiting short and long-path features. Different from the IDN model [19], we construct three-bypass network in each part and different bypass use different convolutional kernels. In this way, our model can adaptively detect the short and long-path features at different scales.

Supposing the input and output of the first part are M_{d-1} and O_{P1}, we have

$$O_{P1} = \sigma\left(Y^1_{1\times1}\left(\left[\sigma\left(Y^2_{3\times3}(M_{d-1})\right) + \sigma\left(Y^3_{5\times5}(M_{d-1})\right) + \sigma\left(Y^4_{7\times7}(M_{d-1})\right)\right]\right)\right), \tag{6}$$

where $Y^1_{1\times1}$, $Y^2_{3\times3}$, $Y^3_{5\times5}$ and $Y^4_{7\times7}$ refers to the function of 1×1, 3×3, 5×5 and 7×7 convolutional layers in the first part respectively. $[\cdot]$ indicates the concatenation of feature maps by different convolutional kernels. σ denotes the ReLU function [21]. After that, the feature maps with 64 dimensions of O_{P1} and the input M_{d-1} are concatenated in the channel dimension,

$$R = C(S(O_{P1}, 64), M_{d-1}), \tag{7}$$

where C and S indicate concatenation operation and slice operation respectively. Therefore, the 64 dimensional features are fetched from S. The purpose is to combine the current multi-scale information with the previous information. It can be regarded as retained short path information. And then, we take the rest of 64 dimensional feature maps as the input of the second part, which mainly further extracts long path information,

$$O_{P2} = \sigma\left(Y^5_{1\times1}\left(\left[\sigma\left(Y^6_{3\times3}(O_{P1}, 64)\right) + \sigma\left(Y^7_{5\times5}(O_{P1}, 64)\right) + \sigma\left(Y^8_{7\times7}(O_{P1}, 64)\right)\right]\right)\right), \tag{8}$$

where $Y^5_{1\times1}$, $Y^6_{3\times3}$, $Y^7_{5\times5}$ and $Y^8_{7\times7}$ refers to the function of 1×1, 3×3, 5×5 and 7×7 convolutional layers in the second part respectively. Finally, the input information, the short path information and the long path information are aggregated, which can be formulated as follows:

$$M_d = R + O_{P2}. \tag{9}$$

where M_d indicates the output of the MSID block.

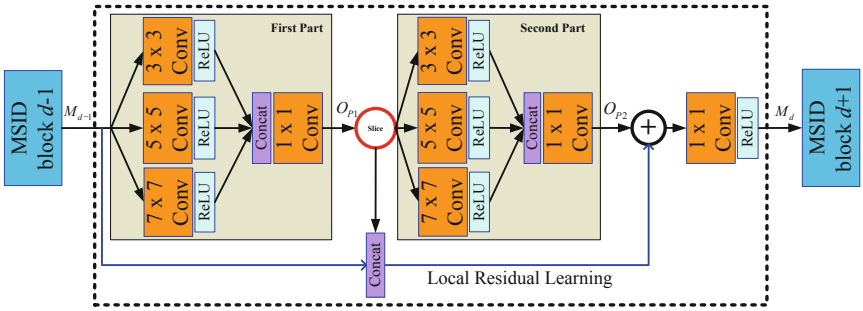


Fig. 2. The architecture of MSID block.

2.3 Non-subsampled Contourlet Transform (NSCT) Prediction

Wavelet analysis [22] cannot "optimally" represent image functions with straight lines and curves. The contourlet transform (CT) [23] improved WT. It is constructed by two filter-bank stages, a Laplacian Pyramid (LP) followed by a Directional Filter Bank (DFB). CT is a shift-variant transform, as it involves sampling at both the LP and the DFB stages. However, shift-variance is not a desirable property for various multimedia processing tasks. To overcome this problem, Cunha et al. [24] proposed NSCT, which is a translation-invariant version of the CT. This transform eliminates all sub-sampling operations, resulting in high redundancy.

NSCT mainly comprises non-subsampled pyramid filter bank (NSPFB) and non-subsampled direction filter bank (NSDFB) in cascade. Firstly, decomposition is made on the image by NSPFB and the resulting sub-bands is taken as input of NSDFB to get decomposition results of the original image in multiple dimensions and directions *K*-level decomposition is made on any image by NSCT to get one low-frequency sub-band and some high-frequency band-pass sub-bands, all of which have the same size as the original image. Both NSPFB and NSDFB are eligible for full reconstruction so NSCT is fully rebuilt as well.

As stated in the introduction, SR in the transform domain can achieve better results than spatial domain. In this paper, we formulate the SISR problem as the prediction of NSCT coefficients, which is able to make MSID-N further preserve richer structure details. In Fig. 3, we compare the high-frequency coefficients of NSCT and WT, where we can clearly see that NSCT represents the curvature more accurately. This demonstrates the superiority of NSCT over WT.

(a) (b) (c)

Fig. 3. Comparison of NSCT and WT coefficients on the Lena image: (a) The original HR image Lena; (b) the fusion of NSCT high-frequency coefficients; (c) the fusion of discrete WT high-frequency coefficients.

In this paper, we formulate the SR problem as the prediction of NSCT coefficients as show in Fig. 4, which is able to make MSID-N further preserve richer structure details than that in spatial domain. It is worth mentioning that NSCT can be used in different SR networks, which is a simple and effective way to improve the performance. Speaking of the role of NSCT, it is to take further experiment in Sect. 3.3. The detailed process of NSCT implementation can be found in [24].

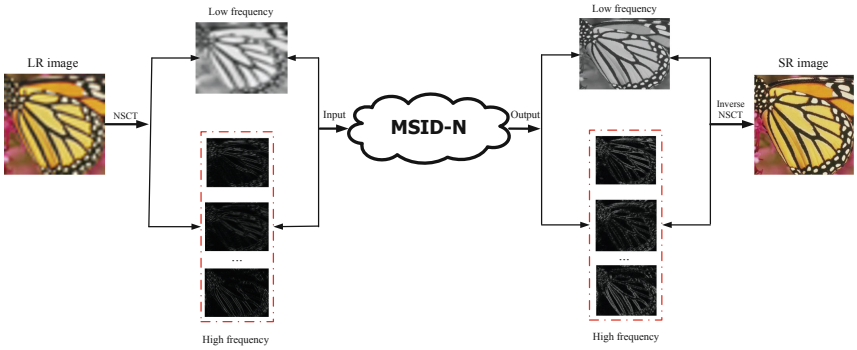


Fig. 4. NSCT domain coefficients prediction.

3 Experiments

In the experiments, the performance of the proposed method is evaluated on both qualitative and quantitative aspects. PSNR and SSIM [25] are used for quantitative evaluation.

3.1 Implementation Details

Recently, Timofte et al. [26] have released a high-quality dataset DIV2 K for image restoration applications. We train our model with 800 training images and use 5

validation images in the training process. For testing, we use the standard benchmark datasets: Set5, Set14, and BSD100. We use the RGB input patches of size 48 × 48 from the LR input for training. We sample the LR patches randomly and augment them by flipping horizontally or vertically and rotating 90. We implement our method with the Torch7 framework and update it with the ADAM optimizer. The mini-batch size is 64, and the learning rate begins with 0.0001 and decreases half for every 100 epochs. Training our model roughly takes 12 h with a NVIDIA Tesla P100 for 200 epochs.

3.2 Evaluation of Results

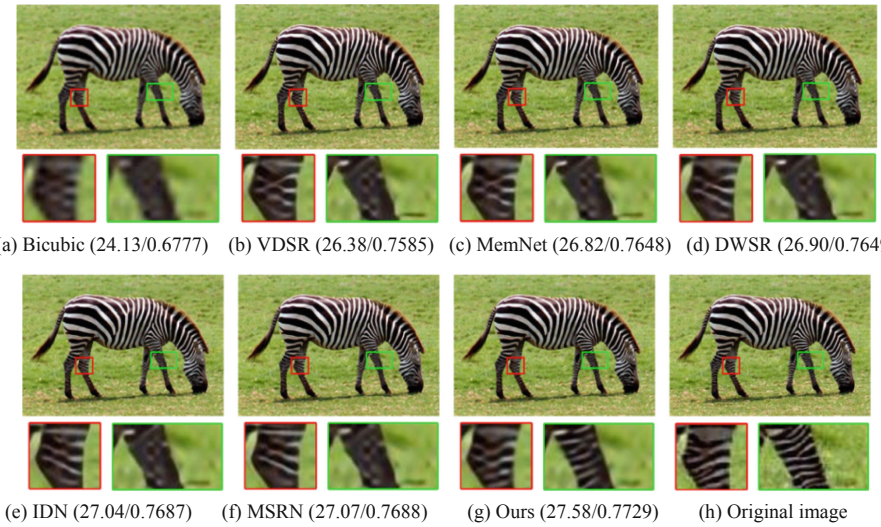
In this section, we evaluate the performance of our method on three standard image datasets. In order to evaluate the SISR performance, we use PSNR and SSIM as a quantitative evaluation metric to justify the reconstruction results. For fair comparison, we use the released codes of the above models and train all models with the same training set. The PSNR (dB) and SSIM values for comparison are shown in Tables 1 and 2. The tables show that our proposed method obtains higher PSNR/SSIM values than other methods; it is that our model constructs multi-bypass network to adaptively detect the short and long-path features and distill more useful information at different scales in transform domain (Fig. 5).

Table 1. Average PSNR values for scaling factor ×2, ×3, and ×4.

Datasets	Scale	Bicubic	VDSR	MemNet	DWSR	IDN	MSRN	Ours
Set5	×2	33.66	37.53	37.78	37.43	37.83	38.08	38.51
	×3	30.39	33.66	34.09	33.82	34.11	34.38	34.75
	×4	28.42	31.35	31.74	31.39	31.82	32.07	32.36
Set14	×2	30.24	33.03	33.28	33.07	33.30	33.74	34.21
	×3	27.55	29.77	30.00	29.83	29.99	30.34	30.68
	×4	26.00	28.01	28.26	28.04	28.25	28.60	28.88
BSD100	×2	29.56	31.90	32.08	31.80	32.08	32.23	32.59
	×3	27.21	28.82	28.96	–	28.95	29.08	29.44
	×4	25.96	27.29	27.40	27.25	27.41	27.52	27.84

Table 2. Average SSIM values for scaling factor ×2, ×3, and ×4.

Datasets	Scale	Bicubic	VDSR	MemNet	DWSR	IDN	MSRN	Ours
Set5	×2	0.9299	0.9587	0.9597	0.9568	0.9600	0.9605	0.9664
	×3	0.8682	0.9213	0.9248	0.9215	0.9253	0.9262	0.9289
	×4	0.8104	0.8838	0.8893	0.8833	0.8903	0.8903	0.8937
Set14	×2	0.8688	0.9124	0.9142	0.9106	0.9148	0.9170	0.9203
	×3	0.7742	0.8314	0.8350	0.8308	0.8354	0.8395	0.8412
	×4	0.7024	0.7674	0.7723	0.7669	0.7730	0.7751	0.7774
BSD100	×2	0.8431	0.8960	0.8978	0.8940	0.8985	0.9013	0.9061
	×3	0.7385	0.7976	0.8001	–	0.8013	0.8041	0.8064
	×4	0.6675	0.7251	0.7281	0.7240	0.7297	0.7273	0.7302



(a) Bicubic (24.13/0.6777) (b) VDSR (26.38/0.7585) (c) MemNet (26.82/0.7648) (d) DWSR (26.90/0.7649)

(e) IDN (27.04/0.7687) (f) MSRN (27.07/0.7688) (g) Ours (27.58/0.7729) (h) Original image

Fig. 5. Visual results (PSNR/SSIM) of our method and several state-of-the-art methods.

3.3 Ablation Investigation

Given that in this paper, we introduce to predict NSCT coefficients for SISR. We evaluate the effect of the contribution on scale 2×. We use three methods (MSRN, IDN and MSID-N) and integrate them with NSCT prediction. Figure 6(a) shows the comparison results of MSID-N across different image datasets. Figure 6(b) shows the comparison results of IDN and MSRN. From Fig. 6, we can see that three methods can be improved significantly when integrated with NSCT. Experimental results demonstrate that NSCT prediction is superior to spatial domain; the improvements are consistent across various networks and benchmarks.

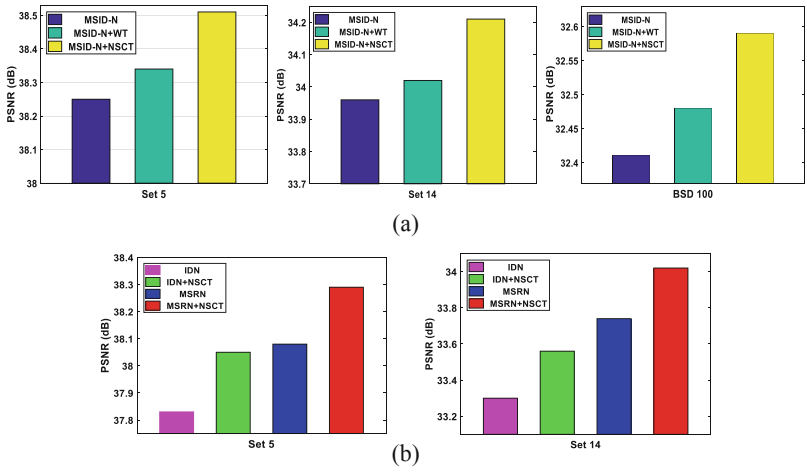


Fig. 6. Effectiveness of NSCT prediction. (a) Comparison for spatial domain, WT domain and NSCT domain. (b) NSCT prediction using different networks.

4 Conclusions

In this paper, a CNN-based SISR method is proposed. Our network MSID-N contains a set of cascaded MSID blocks, which effectively exploit features of image to improve the SISR performance. In addition, NSCT is applied to the network structure to effectively preserve richer detail information than spatial domain, which further improves the SISR performance. Qualitative and quantitative results show that the proposed method is much better than other state-of-the-art methods, boosting restoration ability of LR images.

Acknowledgements. This work was supported in part by the National Science Foundation of China under Grant Nos. 61602226; in part by the PhD Startup Foundation of Liaoning Technical University of China under Grant No. 18-1021; in part by the Basic Research Project of Colleges and Universities of Liaoning Provincial Department of Education under Grant No. LJ2017FBL004.

References

1. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 184–199. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_13
2. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 391–407. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_25
3. 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 (CVPR), pp. 1646–1654 (2016)
4. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2017)
5. Kim, J., Lee, J.K., Lee, K.M.: Deeply-recursive convolutional network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1637–1645 (2016)
6. Wang, Z.W., Liu, D., Yang, J.C., Han, W., Huang, T.: Deep networks for image super-resolution with sparse prior. In: International Conference on Computer Vision (ICCV), pp. 370–378 (2016)
7. Tai, Y., Yang, J., Liu, X.M.: Image super-resolution via deep recursive residual network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3147–3155 (2017)
8. Tong, T., Li, G., Liu, X., Gao, Q.Q.: Image super-resolution using dense skip connections. In: International Conference on Computer Vision (ICCV), pp. 4809–4817 (2017)
9. Tai, Y., Yang, J., Liu, X.M., Xu, C.Y.: MemNet: a persistent memory network for image restoration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4539–4547 (2017)

10. Bricman, P.A., Ionescu, R.T.: CocoNet: a deep neural network for mapping pixel coordinates to color values. In: Cheng, L., Leung, A.C.S., Ozawa, S. (eds.) ICONIP 2018. LNCS, vol. 11302, pp. 64–76. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-04179-3_6
11. Ahn, N., Kang, B., Sohn, K.A.: Fast, accurate, and lightweight super-resolution with cascading residual network. In: Proceedings of ECCV (2018)
12. Shocher, A., Cohen, N., Irani, M.: Zero-shot super-resolution using deep internal learning. In: Proceedings of CVPR (2018)
13. Zhang, K., Zuo, W., Zhang, L.: Learning a single convolutional super-resolution network for multiple degradations. In: Proceedings of CVPR (2018)
14. Li, J.C., Fang, F.M., Mei, K.F., Zhang, G.X.: Multi-scale residual network for image super-resolution. In: European Conference on Computer Vision (ECCV), pp. 527–542 (2018)
15. Zhang, Y.L., Tian, Y.P., Kong, Y., Zhong, B.N., Fu, Y.: Residual dense network for image super-resolution. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
16. Hui, Z., Wang, X.M., Gao, X.B.: Fast and accurate single image super-resolution via information distillation network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 723–731 (2018)
17. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 294–310. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_18
18. Guo, T.T., Mousavi, H.S., Vu, T.H., Monga, V.: Deep wavelet prediction for image super-resolution. In: The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 104–113 (2017)
19. Guo, T.T., Mousavi, H.S., Monga, V.: Orthogonally regularized deep networks for image super-resolution. arXiv preprint arXiv:1802.02018 (2018)
20. Huang, H.B., He, R., Sun, Z.N., Tan, T.N.: Wavelet-srnet: a wavelet-based CNN for multi-scale face super resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1689–1697 (2017)
21. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 315–323 (2011)
22. Mallat, S.: A Wavelet Tour of Signal Processing: The Sparse Way. Academic Press, Cambridge (2008)
23. Do, M.N., Vetterli, M.: The contourlet transform: an efficient directional multiresolution image representation. IEEE Trans. Image Process. **14**(12), 2091–2160 (2005)
24. Cunha, A.L., Zhou, J., Do, M.N.: The nonsubsampled contourlet transform: theory, design, and applications. IEEE Trans. Image Process. **15**(10), 3089–3101 (2006)
25. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
26. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: dataset and study. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 126–135 (2017)