RESIDUAL DILATED NETWORK WITH ATTENTION FOR IMAGE BLIND DENOISING

Guanqun Hou, Yujiu Yang

Tsinghua University, China

Jing-Hao Xue

University College London, UK

ABSTRACT

Image denoising has recently witnessed substantial progress. However, many existing methods remain suboptimal for texture restoration due to treating different image regions and channels indiscriminately. Also they need to specify the noise level in advance, which largely hinders their use in blind denoising. Therefore, we introduce both attention mechanism and automatic noise level estimation into image denoising. Specifically, we propose a new, effective end-to-end attention-embedded neural network for image denoising, named as Residual Dilated Attention Network(RDAN). Our RDAN is composed of a series of tailored blocks including Residual Dilated Attention Blocks(RDAB) and Residual Conv Attention Blocks(RCAB). The RDAB and RCAB incorporates both non-local and local operations, which enable a comprehensive capture of structural information. In addition, we incorporate a Gaussian-based noise level estimation into RDAN to accomplish blind denoising. Experimental results have demonstrated that our RDAN can substantially outperforms the state-of-the-art denoising methods as well as promisingly preserve image texture.

***Index Terms*—** Attention mechanism, image blind denoising, non-local operation

1. INTRODUCTION

As a basic low-level vision task, image denoising plays a vital role in a larger number of various high-level vision tasks, and has attracted wide attention in academic research and industry. As the process of image contamination by noise is irreversible, image denoising is a typical ill-posed problem [1].

The current image denoising methods can be mainly divided into two categories, model-based optimization methods and discriminative learning-based methods. The former includes image non-local self-similarity (NSS) model [2–4], sparse model [5] and gradient-based model [6], etc. However, the performance these methods is subject to the choice of model. Because of the powerful representation of image structure by deep learning, discriminant-based methods have gradually become the mainstream method for image denoising in recent years. Zhang et al. proposed DnCNN [7] using residual learning to solve the training difficulty of deeper networks. Mao et al. proposed a very deep full convolutional encoding-decoding framework [8]. Zhang et al. introduced a denoiser prior in IRCNN [9] for fast image restoration. Zhang et al. recently proposed a fast and flexible denoising network, namely FFDNet [10], capable of handling different levels of noise by a single FFDNet. Also, reference learning, residual learning, non-local and other ideas have been applied in the field of image denoising [11–13].

However, these denoising algorithms still have some issues for further improvement. First, they treat different image regions and channels equally, hence there is no focus in the network processing. Secondly, the size of the network receptive field is limited, hence only local information can be used for image denoising. Finally, most of these algorithms still have to know the noise level in advance, which hinders their use in blind denoising. To alleviate the above problems, we propose a new attention-embedded network called RDAN for blind denoising.

The proposed network framework is shown in Fig. 1 and will be presented in detail in Section 3. The main contributions are summarized as follows.

- We introduce the attention mechanism into image denoising, allowing the network to adaptively set different weights for different regions and channels of the image, hence to improve the restoration of texture details.
- We introduce the idea of fusing both local and non-local operations, dense layer, convolution layer, and dilated convolution operation into the attention structure, hence to increase the network's receptive field and capture comprehensively the structural information for the attention mechanism.
- We incorporate Gaussian-based noise level estimation into the proposed network, hence no need to know the noise level in advance.

2. RELATED WORK

Recently, the attention mechanism has achieved good results in high-level vision tasks [14, 15]. It can selectively focus on salient parts in images to capture effective visual structure better [16], which is consistent with the perception of the human visual system. However, attention remains to fully exploit in image denoising.

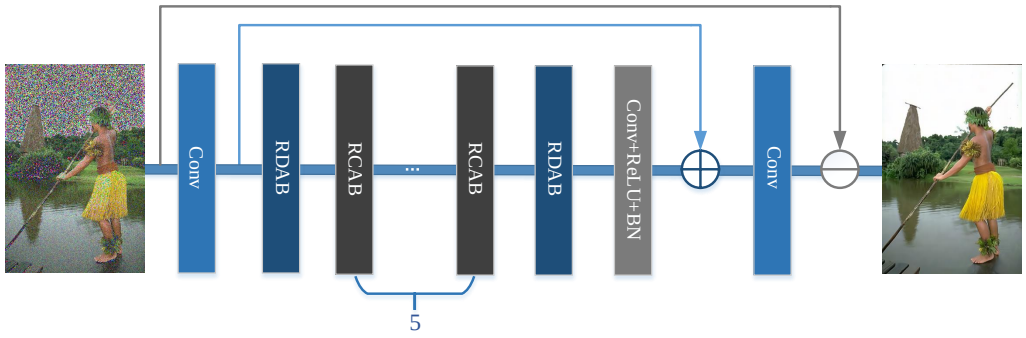


Fig. 1. The framework of our residual dilated attention network (RDAN). RDAB and RCAB denote residual dilated attention block and residual conv attention block, respectively.

At present, most image denoising algorithms are based on the known magnitude of the noise variance. To make the algorithm truly adaptive, noise estimation is very important. Noise estimation methods can be broadly classified into three categories: block-based methods, filter-based methods, and their fusion methods [2, 17]. However, the current noise estimation algorithms only use the internal information of a single image, and the estimation is not accurate enough.

Deep neural networks recently have been prevalent for image denoising. Chen et al. proposed a trainable nonlinear reaction diffusion (TNRD) model [18]. The boosting strategy has also been introduced to improve denoising performance [19]. Similarly, CNN-based methods have been developed for image denoising [7–11]. Besides CNNs, RNNs have also been applied for image restoration [12, 20].

3. RESIDUAL DILATED ATTENTION NETWORK (RDAN)

We proposed a residual dilated attention network (RDAN) for image blind denoising. The overall framework of RDAN is illustrated in Fig. 1. The framework contains convolutional layes and two types of blocks, RDAB (Residual Dilated Attention Block) and RCAB (Residual Conv Attention Block). Each of these blocks is an attention block consisting of a trunk branch and a mask branch. The difference between RDAB and RCAB is the choice of the first part of the mask branch. In the following sections, we introduce each component of the proposed framework in more details.

3.1. Network Framework

As shown in Fig. 1, the proposed RDAN is constructed by convolutional layers, RDAB and RCAB. The RDAN takes noisy image as input and outputs its noise version, and after the subtraction operation, RDAN attains the clean image.

The first convolutional layer is used to extract the shallow features in the noisy image, and the last convolutional layer is used for image restoration as the decoding process. As shown in Fig. 1, by using the long-range skip-connections which bypass intermediate layers, we link the shallow features with the layers that are close to the output of the whole network. The benefits of using such skip-connections here are twofold: providing the long-range information compensation and facilitating the gradient back-propagation and the pixel-wise prediction. The main part of the network is made up of two RDABs and five RCABs, which will be elaborated in the following section.

3.2. RDAB & RCAB

The structures of RDAB and RCAB are shown in Fig. 2. Each block consists of two branches, a trunk branch and a mask branch. The trunk branch performs feature processing, which consists of two Res-blocks and outputs $T(x)$ for input x. The mask branch uses two Res-blocks in RCAB or the Dilated-block + Res-block structure in RDAB to learn mask $M(x)$, which softly weights output features $T(x)$. The output of Attention Module H is then defined as

$$H_{s,c}(x) = T_{s,c}(x) * M_{s,c}(x), \tag{1}$$

where s ranges over all spatial positions and $c \in \{1, 2, \cdots, C\}$ is the index of a channel. The whole structure can be trained in an end-to-end way.

Wang et al. [15] considered that naive stacking Attention Modules leads to the obvious performance drop, so they proposed attention residual learning to ease this problems, by modifying output H of the Attention Module as:

$$H_{s,c}(x) = T_{s,c}(x) * (1 + M_{s,c}(x)). \tag{2}$$

Inspired by this idea, we link the input x to the output $H(x)$ via skip connection, and get a new output of the block as

$$H_{s,c}(x) = T_{s,c}(x) * M_{s,c}(x) + x. \tag{3}$$

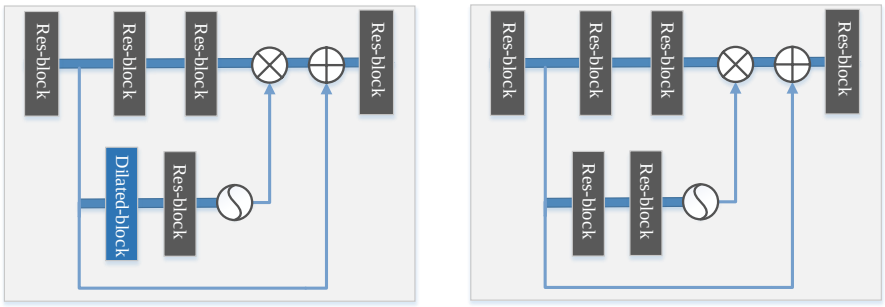


Fig. 2. Left: RDAB (Residual Dilated Attention Block). Right: RCAB (Residual Conv Attention Block).

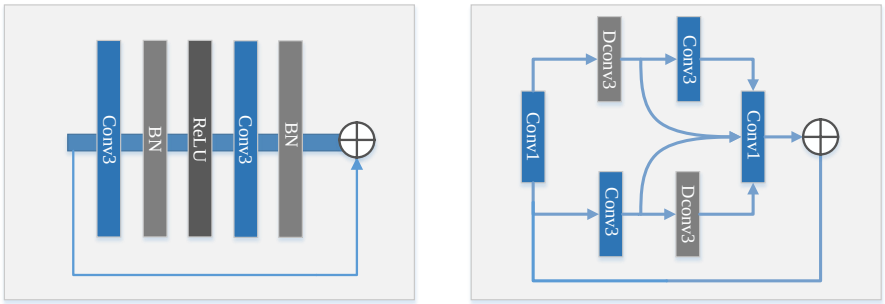


Fig. 3. Left: Res-block. Right: Dilated-block.

In RDAB and RCAB, the mask branch can not only serve as a feature selector during forward inference, but also as a gradient update filter during back propagation. This attention residual learning can retain more low-level image features and is conducive to subsequent denoising.

The difference between RDAB and RCAB is the choice of the first block in the mask branch (i.e. using Dilated-block or Res-block). The structures of Dilated-block and Res-block are shown in Fig. 3. We use two convolutional layers with small 3×3 kernels and 64 feature maps followed by batch-normalization layers and ReLU as the activation function to form the Res-block. The Dilated-block concatenates Conv1 (convolution 1×1), Conv3 (convolution 3×3) and Dconv3 (dilated convolution 3×3 (dilated=2)) by dense connection. Conv1 can perform non-local processing on images. The Conv3-Dconv3 and Dconv3-Conv3 branches can learn different feature representations. The Dilated-block can widen the receptive field while extracting local and non-local features. Experimental results demonstrate that the denoising performance can be improved in this way (see Section 4.3).

3.3. Noise Level Estimation

Most of the existing methods have to know the noise level in advance for testing, but in practices the noise level is unknown, so we design a simple method to firstly estimate the noise level of the test image and then incorporate it into the network when testing.

Noise and texture are both high-frequency information in the image, but noise is much more sensitive to the Laplacian operators, so the Laplacian operator can be used for noise level estimation. A reasonable way to achieve this is to filter the noise image with two different Laplacian filters, and then calculate the difference between two result images obtained by the two filters. Since the image texture information is not sensitive to the Laplacian operators, only the noise is left. Based on Gaussian assumption, the noise magnitude can be estimated as

$$\sigma = \sqrt{\frac{\pi}{2}} \frac{1}{6(W-2)(H-1)} \sum_{imageI} |I * N|, \qquad (4)$$

where W and H are the width and height of the noise image, I is the noisy image, and N is the difference of the two Laplacian filterings. Due to interference by other information in the image, this noise estimation is not precise and thus cannot be directly used for denoising. Hence we roughly discretize the noise estimate into three levels, consistent with [7, 12, 19], as shown in Table 3. In this way, we can obtain the noise level automatically to realize image blind denoising.

Table 1. Comparison for the Gaussian denoising (grey-level). Training and testing protocols are as in [7]. We evaluate the mean PSNR (dB) on the "Set12" and "BSD68" datasets; the best performance is shown in bold.

Dataset	Noise level	BM3D	WNNM	TNRD	DnCNN	RED	FFDNet	DDFN	RDAN
Set12	15	32.37	32.70	32.50	32.86	32.90	32.75	32.98	**33.04**
	25	29.97	30.26	30.06	30.44	30.48	30.43	30.60	**30.70**
	50	26.72	27.05	26.81	27.18	27.33	27.32	27.46	**27.56**
BSD68	15	31.07	31.37	31.42	31.73	31.76	31.63	31.83	**31.85**
	25	28.57	28.83	28.92	29.23	29.27	29.19	29.35	**29.41**
	50	25.62	25.87	25.97	26.23	26.32	26.29	26.42	**26.46**

Table 2. Comparison for the Gaussian denoising (color-level). We evaluate the mean PSNR(dB) on the "BSD68C" dataset; the best performance is shown in bold.

Noise level	CBM3D	CDnCNN	RED	IRCNN	FDDNet	CRDAN
15	33.52	33.89	33.74	33.86	33.87	**33.92**
25	30.71	31.23	31.11	31.16	31.21	**31.29**
50	27.38	27.92	27.89	27.86	27.96	**28.11**

Table 3. Noise level discretizing

Estimated Noise	Noise level
$\sigma \leq 20$	15
$20 < \sigma \leq 35$	25
$\sigma > 35$	50

4. EXPERIMENTAL STUDIES

4.1. Experiment Settings

We train the network on a popular dataset [21], which includes 400 images of size 180×180. We set the patch size as 64×64 and the stride as 16. The test images were taken from two widely-used datasets: one consists of 68 natural images from the Berkeley Split Data Set (BSD68) [22] and the other is the Set12 dataset. All of these images are not included in the training dataset. For color image denoising, we use the color version of the BSD68 dataset for testing and the remaining 432 color images from [22] are adopted as the training images. About 32×4739 patches of size 64×64 are cropped to train the model.

We select the average squared error as the loss function:

$$l(\Theta) = \frac{1}{2N} \sum_{i=1}^{N} \parallel f(y_i; \Theta) - (y_i - x_i) \parallel_F^2 \qquad (5)$$

To optimize the network parameters Θ, the Adam solver is adopted. The learning rate is initially set to 0.001, and the batch-size is set to 32. We use the Keras platform to implement our models with a Titan XP GPU.1 trained for 50 epochs. We evaluate the performance of the synthesized data in terms of Peak Signal-to-Noise Ratio (PSNR).

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

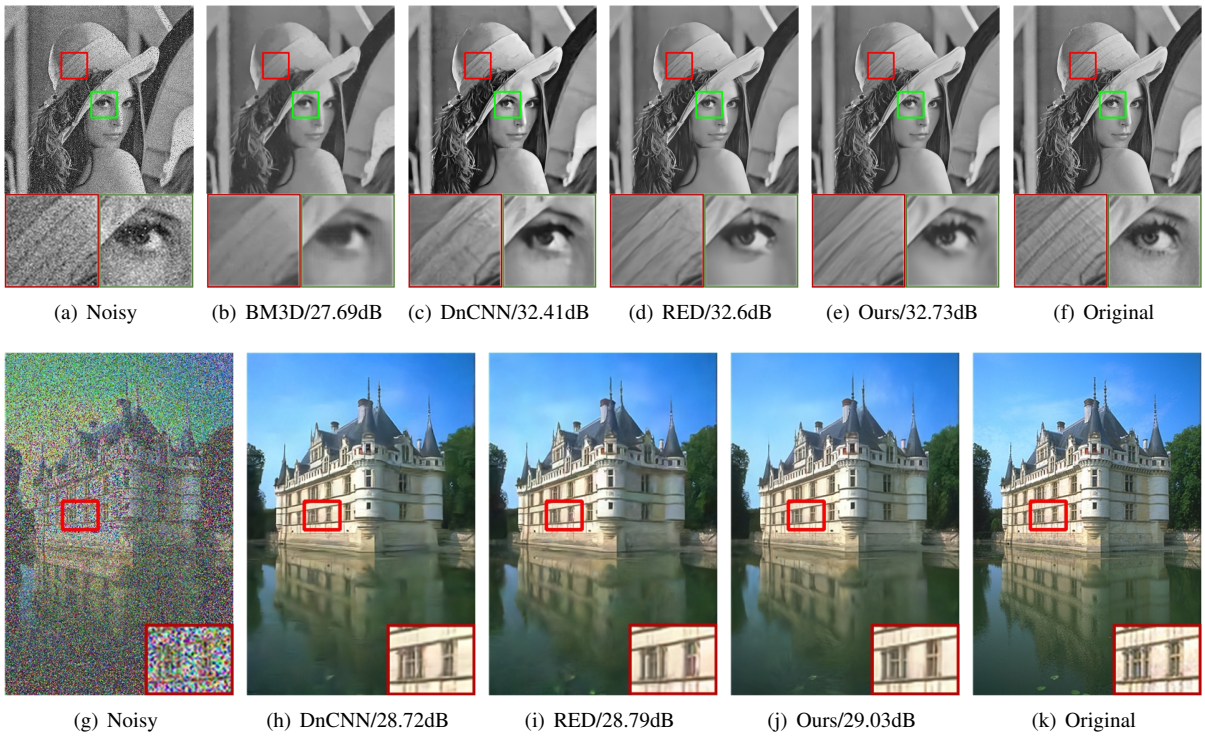
We compared our proposed RDAN with several denoising results of different methods on the "Set12" and "BSD68" datasets; the results are shown in Table 1. The proposed RDAN can produce the highest PSNR: it achieves a gain of about 0.2dB over DnCNN. Table 2 lists the color image denoising results: it can be seen that the proposed RDAN also outperforms the state-of-the-art methods. An illustration of visual comparison on two images is provided in Fig. 4, from which we can observe the proposed RDAN is more effective for restoration of texture.

4.3. Ablation Experiments

In order to validate the effectiveness and necessity of our proposed attention mechanism in RDAB and RCAB, we design extensive ablation experiments to evaluate them. we compare RDAN with its five variants on the BSD68 dataset at $\sigma = 25$. $\mathcal{R}_a$ is our RDAN; $\mathcal{R}_b$ denotes one network stacking 28 Res-block; $\mathcal{R}_c$ indicates a network include 7 RCAB; $\mathcal{R}_d$ is a network include 7 RDAB; $\mathcal{R}_e$ ($\mathcal{R}_f$) means a network include one RDAB (RCAB). The corresponding performance changes in terms of PSNR are listed in Table 4. The $\mathcal{R}_a$ shows that our RDAN with the attention mechanism is more effective than others; and the RDAB is more valid than RCAB by comparing $\mathcal{R}_e$ and $\mathcal{R}_f$. A visual comparison is provided in Fig. 5. As can be seen, the attention mechanism and the RDAB are effective for texture restoration.

5. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a new deep network called RDAN which integrates the attention mechanism with deep



(a) Noisy (b) BM3D/27.69dB (c) DnCNN/32.41dB (d) RED/32.6dB (e) Ours/32.73dB (f) Original

(g) Noisy (h) DnCNN/28.72dB (i) RED/28.79dB (j) Ours/29.03dB (k) Original

Fig. 4. Visual comparison of 'Lena' denoising from "Set12" at $\sigma = 25$ and one color image denoising from "Set68C " at $\sigma = 50$. Our RDAN ((e) & (j)) performs the best in terms of PSNR/dB; moreover, it is effective for texture restoration.

Table 4. Ablation study on different components of our proposed RDAN framework at "BSD68" dataset. $\mathcal{R}_a$ and $\mathcal{R}_d$ performs the best in terms of PSNR/dB, but $\mathcal{R}_d$ spent more time than $\mathcal{R}_a$. WA: the Net without attention; WRD: the Net without RDAB; WRC: the Net without RCAB; SRD: the Net with a single RDAB; and SRC: the Net with a single RCAB.

Model	$\mathcal{R}_a$	$\mathcal{R}_b$	$\mathcal{R}_c$	$\mathcal{R}_d$	$\mathcal{R}_e$	$\mathcal{R}_f$
WA		$\surd$				
WRD			$\surd$			
WRC				$\surd$		
SRD					$\surd$	
SRC						$\surd$
PSNR	29.41	29.40	29.39	29.41	29.11	29.09

learning for image denoising for the first time. It is composed of a series of tailored RDAB and RCAB. The RDAB is at the beginning and end of the network, which incorporates non-local operations for image denoising. The RCAB is then used to extract finer local features. RDAB and RCAB enable a comprehensive capture of structural information crucial for the attention mechanism. In addition, we combine the noise level estimation with image denoising to achieve the task of blind denoising. Experimental results have demonstrated that our proposed RDAN can effectively denoise while promisingly preserve the image texture, which remarkably outperforms the state-of-the-art methods. In our future research, we plan to extend the proposed algorithm to a wider range of image restoration tasks beyond denoising, including but not limited to image de-raining, image de-hazing and image super-resolution.

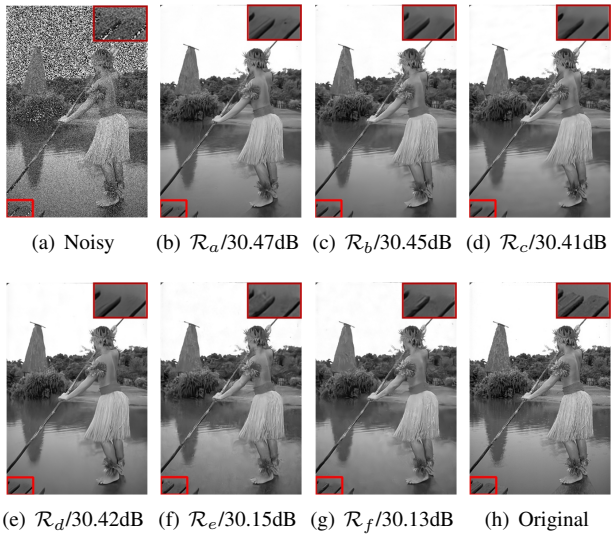
6. REFERENCES

[1] Ryan Wen Liu and Chuansheng Wu, "A predictor-corrector iterated Tikhonov regularization for linear ill-posed inverse problems," *Applied Mathematics and Computation*, vol. 221, pp. 802–818, 2013.

[2] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen O. Egiazarian, "Image Denoising by Sparse 3-D Transform-Domain Collaborative Filtering," *IEEE Trans. Image Processing*, vol. 16, no. 8, pp. 2080–2095, 2007.

[3] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel, "A Non-Local Algorithm for Image Denoising," in *CVPR*, 2005, pp. 60–65.

[4] Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng, "Weighted Nuclear Norm Minimization with Application to Image Denoising," in *CVPR*, 2014, pp. 2862–2869.



(a) Noisy (b) $\mathcal{R}_a$/30.47dB (c) $\mathcal{R}_b$/30.45dB (d) $\mathcal{R}_c$/30.41dB

(e) $\mathcal{R}_d$/30.42dB (f) $\mathcal{R}_e$/30.15dB (g) $\mathcal{R}_f$/30.13dB (h) Original

Fig. 5. Visual comparison of denoising an image from "Set68" at $\sigma = 25$. The meaning of $\mathcal{R}_a$–$\mathcal{R}_f$ can be found in Section 4.3.

[5] Julien Mairal, Francis R. Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman, "Non-local sparse models for image restoration," in *ICCV*, 2009, pp. 2272–2279.

[6] Yair Weiss and William T. Freeman, "What makes a good model of natural images?," in *CVPR*, 2007.

[7] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang, "Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising," *IEEE Trans. Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[8] Xiao-Jiao Mao, Chunhua Shen, and Yu-Bin Yang, "Image Restoration Using Very Deep Convolutional Encoder-Decoder Networks with Symmetric Skip Connections," in *NIPS*, 2016, pp. 2802–2810.

[9] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang, "Learning Deep CNN Denoiser Prior for Image Restoration," in *CVPR*, 2017, pp. 2808–2817.

[10] Kai Zhang, Wangmeng Zuo, and Lei Zhang, "Ffdnet: Toward a Fast and Flexible Solution for CNN-Based Image Denoising," *IEEE Trans. Image Processing*, vol. 27, no. 9, pp. 4608–4622, 2018.

[11] Filippos Kokkinos and Stamatios Lefkimmiatis, "Deep Image Demosaicking Using a Cascade of Convolutional Residual Denoising Networks," in *ECCV*, 2018, pp. 317–333.

[12] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S. Huang, "Non-Local Recurrent Network for Image Restoration," *CoRR*, vol. abs/1806.02919, 2018.

[13] Ke Yu, Chao Dong, Liang Lin, and Chen Change Loy, "Crafting a Toolchain for Image Restoration by Deep Reinforcement Learning," in *CVPR*, 2018, pp. 2443–2452.

[14] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua, "SCA-CNN: spatial and channel-wise attention in Convolutional Networks for Image Captioning," in *CVPR*, 2017, pp. 6298–6306.

[15] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang, "Residual Attention Network for Image Classification," in *CVPR*, 2017, pp. 6450–6458.

[16] Hugo Larochelle and Geoffrey E. Hinton, "Learning to combine foveal glimpses with a third-order boltzmann machine," in *NIPS 2010*, 2010, pp. 1243–1251.

[17] Guangyong Chen, Fengyuan Zhu, and Pheng-Ann Heng, "An Efficient Statistical Method for Image Noise Level Estimation," in *ICCV*, 2015, pp. 477–485.

[18] Yunjin Chen and Thomas Pock, "Trainable Nonlinear Reaction Diffusion: A Flexible Framework for Fast and Effective Image Restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1256–1272, 2017.

[19] Chang Chen, Zhiwei Xiong, Xinmei Tian, and Feng Wu, "Deep Boosting for Image Denoising," in *ECCV*, 2018, pp. 3–19.

[20] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu, "MemNet: A Persistent Memory Network for Image Restoration," in *ICCV*, 2017, pp. 4549–4557.

[21] David R. Martin, Charless C. 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 *ICCV*, 2001, pp. 416–425.

[22] Stefan Roth and Michael J. Black, "Fields of Experts," *International Journal of Computer Vision*, vol. 82, no. 2, pp. 205–229, 2009.