IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 13, 2019, accepted December 2, 2019, date of publication December 30, 2019,
date of current version January 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2962809

On Combining CNN With Non-Local Self-Similarity Based Image Denoising Methods

ZIFEI YAN [1], SHI GUO [1,2], GANG XIAO [3], AND HONGZHI ZHANG [1]

[1] Harbin Institute of Technology, Harbin 150001, China
[2] Department of Computing, The Hong Kong Polytechnic University, Hong Kong
[3] No. 962 Hospital of PLA, Harbin 150000, China

Corresponding author: Zifei Yan (yanzifei@hit.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61871381, Grant 61671182, and Grant 61872118, in part by the Youth Projects in Humanities and Social Sciences of the Ministry of Education of China under Grant 18YJCZH211, and in part by the Key Laboratory Funding Project of the Ministry of Culture and Tourism of The People's Republic of China.

ABSTRACT Despite the significant advances in convolutional neural network (CNN) based image denoising, the existing methods still cannot consistently outperform non-local self-similarity (NSS) based methods, especially on images with many repetitive structures. Although several studies have been given to incorporate NSS priors with CNN-based denoising,their improvement is generally insignificant when compared with the state-of-the-art CNN-based denoisers. In this paper, we suggest to combine CNN and NSS based methods for improved image denoising, resulting in an NSS-UNet architecture. Motivated by gradient descent inference of TNRD, both the current estimate and noisy observation are considered as the inputs to the CNN. To take the NSS prior into account, the result by NSS (e.g., BM3D or WNNM), is adopted as the initial estimate. And a modified UNet is presented for exploiting the multi-scale information. We evaluate the proposed method on three common testing datasets. The results clearly show that NSS-UNet outperforms the existing CNN and NSS based methods in terms of both PSNR index and visual quality.

INDEX TERMS Non-local self-similarity, convolutional neural network, residual learning, image denoising.

I. INTRODUCTION

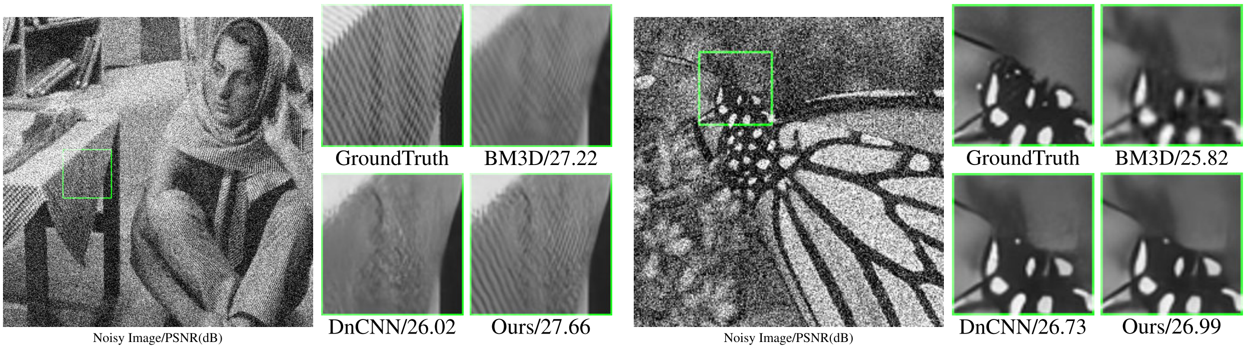
Image denoising is a fundamental and classical topic in image processing and low level vision. Given a noisy image $y = x + n$, image denoising aims to estimate the corresponding clean image x. And n is usually assumed to be an additive white Gaussian noise (AWGN) with the mean zero and standard deviation σ. For decades, a wide range of models have been suggested for image denoising, to name a few, variational models [1]–[3], non-local self-similarity (NSS) based methods [4]–[6], Markov random fields [7]–[9], sparse representation methods [4]–[6], [10]–[12] and discriminative learning based methods [13]–[15]. Among these methods, NSS-based methods and discriminative learning based ones are two representative categories of approaches with state-of-the-art denoising performance.

The NSS-based methods utilize the repetitive local patterns in an image for effective denoising. Generally, for each noisy image patch $\mathbf{y}_i$, one can find a group of similar patches $\mathbf{Y}$ (including $\mathbf{y}_i$) from a larger neighborhood of $\mathbf{y}_i$. The non-local mean (NLM) approach [16] simply adopts the weighted mean of the patches in $\mathbf{Y}$ as the estimate of the clean patch. In BM3D [17], $\mathbf{Y}$ is reshaped as 3D data array, and transform-domain collaborative filtering is then deployed to estimate the clean patch group $\mathbf{X}$. Inspired by the success of BM3D, group sparsity and centralized sparsity are respectively adopted in LSSC [4] and NCSR [6] to model the prior of $\mathbf{X}$. Recently, low rank representation is exploited by WNNM [18] for modeling $\mathbf{X}$, and achieves favorable denoising performance.

Another representative category of state-of-the-art denoising approaches are discriminative learning based methods. Given a training set of noisy observation and ground truth clean image pairs, discriminative learning based methods directly learn a mapping to estimate the clean image from its noisy observation. Based on the truncated inference procedure, the cascade shrinkage fields (CSF) model [13] and the trainable non-linear reaction diffusion (TNRD) model [14], [15] have been proposed to learn stage-wise inference models. However, the performance of such methods may be

The associate editor coordinating the review of this manuscript and approving it for publication was Chao Tong.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/



(a) Barbara

(b) Monar

FIGURE 1. Denoising results ($\sigma = 50$) by discriminative learning based methods: DnCNN [21], NSS-based methods: BM3D [17] and the proposed BM3D-UNet. (a) *Barbara*, (b) *Monar*.

limited by the forms of image priors and inference algorithms. With the progress in deep learning, convolutional neural networks (CNNs) are also utilized for image denoising [19]–[23]. Nowadays, CNN-based solutions have achieved better denoising performance than those based on NSS and truncated inference.

Despite their success on image denoising, both the two categories of approaches have their respective merits and drawbacks. The discriminative learning methods take advantage of large scale training data, and accumulate image local information to generate high quality denoising estimation. In contrast, non-local based approaches are able to take benefit from image NSS prior, and are superior in handling images with repetitive patterns. Fig. 1 presents the denoising results of two noisy images (i.e., *Barbara* and *Monar.*) with $\sigma = 50$ by a representative NSS-based method (i.e., BM3D [17]) and a state-of-the-art CNN-based method (i.e., DnCNN [21]). Due to the rich repetitive patterns of *Barbara*, BM3D [17] significantly outperforms DnCNN [21] in terms of PSNR index, and recovers more detailed textures from noisy observation. As for images dominated by salient structures (e.g., *Monar.* in Fig. 1), DnCNN [21] generally surpasses BM3D [17], and is effective in recovering sharp edges and suppressing artifacts in the smooth region.

Considering their relative merits and complementarity, it is encouraging to incorporate NSS-based and discriminative learning methods for enhanced denoising performance. Recently, some attempts have already been made for improving discriminatively learned denosing models (e.g., TNRD [15]) with NSS prior [24], [25]. These methods, however, perform marginally better than the NSS-based methods and generally are inferior to the state-of-the-art CNN-based denoisers. Therefore, more studies are still required to take the full advantages of both NSS-based and CNN-based methods.

In this paper, we suggest to combine CNN and NSS based methods for improved image denoising. Concretely, we consider two representative NSS-based methods, i.e., BM3D [17] and WNNM [18], and adopt the UNet architecture [26] due to its optimal tradeoff between efficiency and receptive field size. In order to design appropriate combination scheme, we refer to [21] which shows that DnCNN can be treated as an extension of TNRD with the residual learning formulation and initialized by the noisy observation $\mathbf{y}$. Particularly, DnCNN aims to learn a residual mapping $\mathcal{R}(\mathbf{y}) \approx \mathbf{n}$, and the clean image is then predicted by $\mathbf{x} = \mathbf{y} - \mathcal{R}(\mathbf{y})$. Here, we move one step forward by taking the estimate by some NSS-based method $\mathbf{x}_t$ as the initialization of $\mathbf{x}$. By analyzing the connection with TNRD, we argue that both the initial estimate $\mathbf{x}_t$ and the noisy observation $\mathbf{y}$ should be taken as the inputs to the CNN model. And the modified UNet is deployed for learning the residual mapping $\mathcal{R}(\mathbf{y}, \mathbf{x}_t) \approx \mathbf{x} - \mathbf{x}_t$, resulting in our NSS-UNet model. Finally, the results on three widely used datasets show that BM3D-UNet and WNNM-UNet outperform existing popular CNN and NSS based methods in terms of PSNR index and visual quality. On Set12, given the noise level $\sigma = 50$, the average PSNR by our WNNM-UNet can be about 0.5 dB higher than the state-of-the-art DnCNN and WNNM, and 0.2 dB higher than the baseline UNet.

To sum up, the main contribution of this work is three-fold:

- We suggest to combine the NSS-based denoiser with the CNN-based methods for enhanced image denoising. While the combination of WNNM and CNN greatly benefits denoising performance, the combination of BM3D and CNN can achieve better tradeoff between efficiency and effectiveness.
- A simple but effective combination scheme is presented by dissecting the connection between CNN and TNRD inference. Both the estimate by the NSS-based method and the noisy observation are taken as inputs to the modified UNet for predicting the residual between the ground truth and the estimate.
- Extensive experiments show that our NSS-UNet outperforms the state-of-the-art model-based denoising methods, and even achieve denoising results comparable with current best performing NSS and CNN based methods in terms of quantitative and qualitative evaluation.

The remainder of this paper is organized as follows. Section II presents a brief survey on the related works on image denoising. Section III describes the proposed NSS-UNet method, while Section IV reports the experimental results. Finally, Section V ends this paper by providing several concluding remarks.

II. RELATED WORKS

A. IMAGE DENOISING BASED ON DEEP NEURAL NETWORK

Due to that great successes had been achieved on high level computer vision applications, the application of deep neural networks (DNN) methods on denoising tasks have been attracting great research interests in recent years. In [20], Jain and Seung propose to use convolutional neural networks (CNNs) for image denoising and discuss their relationship with Markov random field (MRF). In [19], Burger et al. show that the multi-layer perceptron (MLP) can achieve state-of-the-art performance for image denoising. In [27], Xie et al. combine sparse coding and deep networks pre-trained with denoising auto-encoder (DA) together for Gaussian noise removal. In [15], Chen et al. propose a simple but effective framework to learn stage-wise inference models. In [21], Zhang et al. design a deep denoising CNN (DnCNN) by incorporating residual learning [28] and batch normalization [29]. Although CNN-based methods (e.g., DnCNN [21]) have achieved state-of-the-art denoising performance, they cannot consistently outperform the NSS-based methods, especially on images with many repetitive structures.

B. IMAGE DENOISING BASED ON NONLOCAL SELF-SIMILARITY

The nonlocal self-similarity (NSS) prior refers to the fact that natural images contain many repetitive structures at different locations. On the basis of above considerations, Buades *et al.* [16] firstly adopt the NSS prior and proposed a non-local mean (NLM) filtering method for image denoising. Inspired by NLM, a lot of NSS-based methods [4], [6], [17], [18], [30] have been suggested in the last decade, and achieved promising denoising performance. In [17], Dabov et al. propose a very efficient and highly engineered approach known as BM3D, which consists of three steps: nonlocal patch grouping, 3D wavelet shrinkage, and patch group aggregation. In [6], Dong et al. present a nonlocally centralized sparse representation (NCSR) model for image restoration, which uses the NSS prior to obtain good estimate of the sparse coding coefficients of the original clean image. Gu *et al.* [18] apply the weighted nuclear norm minimization (WNNM) algorithm to image denoising by exploiting the low-rank property of nonlocal self-similar image patches.

C. INCORPORATION OF NSS AND CNN BASED METHODS

Despite their rapid progress, the CNN-based denoising methods [14], [15], [21], [22] remain local models and rarely take into account the inherent NSS property. In contrast, the NSS-based methods, such as BM3D [17] and WNNM [18], are promising in recovering repetitive structures but may introduce artifacts in smooth region. Therefore, it is natural to expect that the denoising performance can be further boosted by incorporating the NSS and CNN based methods [24], [25], [31]. In [24], [25], the NSS regularization terms are designed for TNRD [15]. In [31], by extending BM3D [17], the algorithm consists of three parts: blocking matching, denoising CNN, and aggregation. For better location of similar patches, the block matching is operated on the DnCNN [21] results. N^{3}Net [32] introduces the continuous deterministic relaxation of the KNN rule to neural network architectures by proposing a non-local processing layer. For the purpose of capturing self-smilar information, GCDN [33] employs graph convolution to create layers with hidden neurons having non-local receptive fields. By computing reliable feature correlations within a confined neighborbood and passing feature correlation messages between adjacent recurrent stages, NLRN [34] incorporates non-local operations into a recurrent neural network for image denosing. Different from [24], [25], [31]–[34], we present a simple yet effective scheme to combine the NSS and CNN-based methods. Concretely, the NSS-based method is first deployed to obtain an initial estimate of the clean image. By taking both the initial estimate and noisy observation as inputs, a modified UNet is then presented to produce the final denoising result. Extensive experiments clearly demonstrate the superiority of our NSS-UNet in comparison with the state-of-the-art denoising methods.

III. THE PROPOSED NSS-UNET MODEL

This section presents the proposed NSS-UNet model for combining NSS and CNN based methods. First, the NSS-based method is employed to generate an initial estimate x_t from the noisy input y. Then, the key issue of NSS-UNet is to design proper CNN model to predict the final denoising result based on the initial estimate x_t.

To this end, in Sec. III-A we resort to the connection between TNRD inference and CNN, and provide two tips for the CNN model: (i) taking both x_t and y as inputs, and (ii) predicting the residual $x - x_t$ instead of $x - y$. Sec. III-B disccusses the choice of x_t based on three principles, i.e., learnability, complementarity, and efficiency. Finally, Sec. III-C introduces the modified UNet architecture to produce the final denoising result.

A. TNRD AND GRADIENT DESCENT INFERENCE

In this work, we treat the denoising result by the NSS-based method as an initial estimate x_t of the clean image. Our combination scheme can then be modeled as the learning of a mapping to the clean image by using x_t as the starting point. Thus, we resort to the gradient descent inference adopted in TNRD [14], which can be operated at any starting point and may shed some light on the design of the CNN model. The objective function of the TNRD model [14] is defined as,

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|^2 + \lambda \sum_{k=1}^{K}\sum_{p=1}^{N} \rho_k((\mathbf{f}_k * \mathbf{x})_p), \tag{1}$$

where N denotes the image size, λ denotes the regularization parameter, $*$ denotes the convolution operator. $\mathbf{f}_k$ and ρ_k are

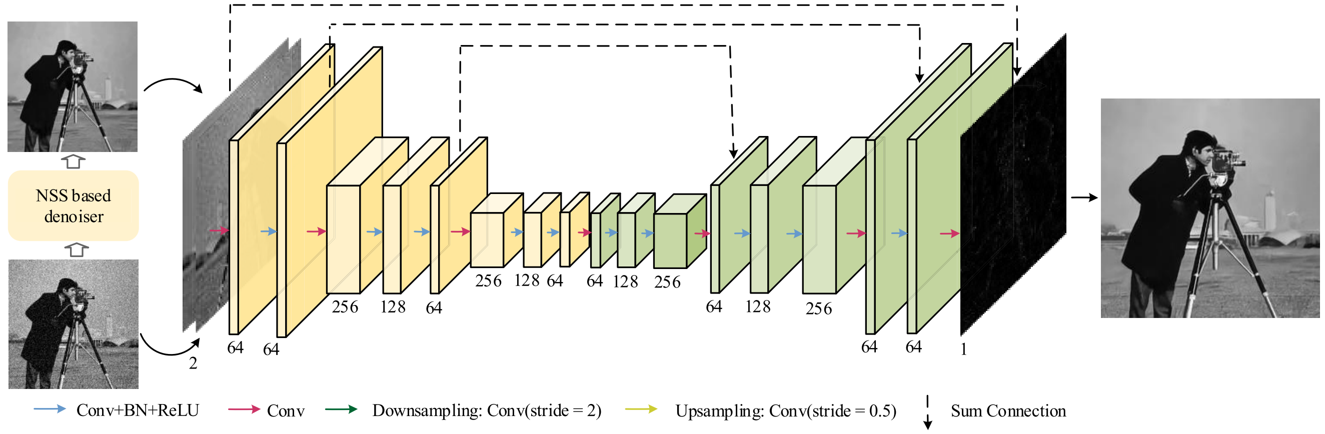


FIGURE 2. Illustration of the proposed NSS-UNet model.

the k-th filter kernel and penalty function, respectively. $(\cdot)_p$ denotes the pixel value at the location p. At a general starting point $\mathbf{x}_t$, the diffusion of TNRD can be written as one gradient descent step,

$$\mathbf{x} = \mathbf{x}_t - \alpha \left((\mathbf{x}_t - \mathbf{y}) + \lambda \sum_{k=1}^{K} \tilde{\mathbf{f}}_k * \rho'_k(\mathbf{f}_k * \mathbf{x}_t) \right), \qquad (2)$$

where $\tilde{\mathbf{f}}_k$ denotes the adjoint filter of $\mathbf{f}_k$, α is the step size, and $\rho'_k(\cdot)$ denotes the derivative of $\rho_k(\cdot)$. Furthermore, we define the residual mapping $\mathcal{R}(\mathbf{y}, \mathbf{x}_t) = \mathbf{x} - \mathbf{x}_t$ as,

$$\mathcal{R}(\mathbf{y}, \mathbf{x}_t) = -\alpha \left((\mathbf{x}_t - \mathbf{y}) + \lambda \sum_{k=1}^{K} \tilde{\mathbf{f}}_k * \rho'_k(\mathbf{f}_k * \mathbf{x}_t) \right). \qquad (3)$$

We can obtain the following two observations from Eqn. (3): (i) the residual mapping is a function of both $\mathbf{y}$ and $\mathbf{x}_t$, and thus we should take both $\mathbf{y}$ and $\mathbf{x}_t$ as inputs to CNN; (ii) instead of predicting $\mathbf{x}-\mathbf{y}$ in DnCNN [21], we should learn the mapping $\mathcal{R}(\mathbf{y}, \mathbf{x}_t)$ to predict $\mathbf{x} - \mathbf{x}_t$.

B. PROPER CHOICE OF $\mathbf{x}_t$

From Eqn. (3), one can see that $\mathbf{x}_t$ may affect both the form and the result of the residual mapping $\mathcal{R}(\mathbf{y}, \mathbf{x}_t)$. Thus, proper choice of $\mathbf{x}_t$ is also a crucial issue in the proposed method. In general, we consider three principles for choosing proper $\mathbf{x}_t$, i.e., learnability, complementarity, and efficiency. As to the learnability, we require to use some $\mathbf{x}_t$ which makes it easier to learn the residual mapping $\mathcal{R}(\mathbf{y}, \mathbf{x}_t)$. Here $\|\mathcal{R}(\mathbf{y}, \mathbf{x}_t)\|^2$ is simply adopted as a metric on the easiness of learning, where lower $\|\mathcal{R}(\mathbf{y}, \mathbf{x}_t)\|^2$ indicates that it is easier to learn $\mathcal{R}(\mathbf{y}, \mathbf{x}_t)$. Note that $\mathcal{R}(\mathbf{y}, \mathbf{x}_t) \approx \mathbf{x} - \mathbf{x}_t$. Thus, to decrease $\|\mathcal{R}(\mathbf{y}, \mathbf{x}_t)\|^2$, we require $\mathbf{x}_t$ to be close to the ground truth clean image $\mathbf{x}$, and one feasible choice is to use another high-performance denoiser to obtain $\mathbf{x}_t$.

As to the complementarity, $\mathbf{x}_t$ is required to contain some complementary information to the result by CNN-based denoiser. From Fig. 1, one can see that DnCNN only considers the generic image priors, and performs poor on images with rich repetitive structures. In contrast, NSS-based methods usually achieve favorable results on this kind of images. Therefore, we suggest to use the NSS-based methods (e.g., BM3D, WNNM) to produce $\mathbf{x}_t$. Moreover, our empirical experiments show that the combination of CNN with patch based methods (e.g., EPLL) contributes little in improving denoising result, which indicates that complementarity play more important role in improving denoising performance. Through experiments, we find that utilizing WNNM to generate $\mathbf{x}_t$ can achieve the best denoising performance. When taking the efficiency into account, BM3D will be more preferred to combine with the CNN-based method.

C. NETWORK ARCHITECTURE

Based on the discussion in Sec. III-A, the network should take both $\mathbf{y}$ and $\mathbf{x}_t$ as inputs, and learn the residual mapping $\mathcal{R}(\mathbf{y}, \mathbf{x}_t) = \mathbf{x} - \mathbf{x}_t$. Here we adopt a modified UNet architecture to learn $\mathcal{R}(\mathbf{y}, \mathbf{x}_t)$ for better tradeoff between efficiency and denoising performance. With the introduction of skip connections and pooling operations, the UNet [26] provides an efficient way to exploit multi-scale information, and usually exhibits higher computational efficiency and larger receptive field at the moderate increase of memory cost.

As illustrated in Fig. 2, we further modify UNet from two aspects. First, instead of the deployment of pooling layer, we use convolutional layers with stride 2 and 1/2 for in-network downsampling and upsampling. Specifically, we utilize k $(= 2)$ downsampling and upsampling steps in the modified UNet, leading to $k + 1$ spatial scales of feature maps. For each scale, two convolutional blocks (3×3 convolution + Batch Normalization + ReLU) are employed in the encoder and decoder subnetworks, respectively. Thus, our NSS-UNet can have a very large reception field of 85×85. Second, we adopt a simple element-wise summation operation to combine the feature maps from the encoder and decoder subnetworks. In contrast, concatenation has been utilized in [26]. We empirically find that element-wise summation is effective in reducing the network parameters, and can lead to comparable denoising results. For more details

TABLE 1. The PSNR (dB) results by different methods on Set12.

Images	C.man	House	Peppers	Starfish	Monar.	Airpl.	Parrot	Lena	Barbara	Boat	Man	Couple	Average
Noise Level	$\sigma = 15$												
BM3D [17]	31.91	34.93	32.69	31.14	31.85	31.07	31.37	34.26	33.10	32.13	31.92	32.10	32.37
WNNM [18]	32.17	35.13	32.99	31.82	32.71	31.39	31.62	34.27	33.60	32.27	32.11	32.17	32.70
EPLL [35]	31.85	34.17	32.64	31.13	32.10	31.19	31.42	33.92	31.38	31.93	32.00	31.93	32.14
CSF [13]	31.95	34.39	32.85	31.55	32.33	31.33	31.37	34.06	31.92	32.01	32.08	31.98	32.32
TNRD [15]	32.19	34.53	33.04	31.75	32.56	31.46	31.63	34.24	32.13	32.14	32.23	32.11	32.50
DnCNN [21]	32.55	34.93	33.22	32.16	33.03	31.67	31.82	34.57	32.59	32.38	32.45	32.44	32.82
UNet	32.61	35.07	33.28	32.41	33.19	31.71	31.87	34.65	32.67	32.43	32.49	32.50	32.91
BM3D-UNet	32.67	35.31	33.29	32.38	33.16	31.73	31.87	34.67	33.32	32.46	32.48	32.53	32.99
WNNM-UNet	32.74	35.43	33.32	32.43	33.30	31.78	31.92	34.72	33.77	32.50	32.52	32.56	33.08
Noise Level	$\sigma = 25$												
BM3D [17]	29.44	32.85	30.16	28.56	29.25	28.42	28.93	32.07	30.71	29.90	29.61	29.71	29.97
WNNM [18]	29.64	33.22	30.42	29.03	29.84	28.69	29.15	32.24	31.24	30.03	29.76	29.82	30.26
EPLL [35]	29.26	32.17	30.17	28.51	29.39	28.61	28.95	31.73	28.61	29.74	29.66	29.53	29.69
MLP [19]	29.61	32.56	30.30	28.82	29.61	28.82	29.25	32.25	29.54	29.97	29.88	29.73	30.03
CSF [13]	29.48	32.39	30.32	28.80	29.62	28.72	28.90	31.79	29.03	29.76	29.71	29.53	29.84
TNRD [15]	29.72	32.53	30.57	29.02	29.85	28.88	29.18	32.00	29.41	29.91	29.87	29.71	30.06
DnCNN [21]	30.13	33.05	30.82	29.36	30.21	29.10	29.42	32.42	29.90	30.17	30.10	30.07	30.40
UNet	30.19	33.13	30.86	29.85	30.38	29.14	29.48	32.56	30.17	30.28	30.14	30.18	30.53
BM3D-UNet	30.25	33.38	30.81	29.88	30.40	29.17	29.49	32.61	31.01	30.31	30.15	30.23	30.64
WNNM-UNet	30.35	33.51	30.92	29.89	30.45	29.21	29.50	32.66	31.43	30.35	30.20	30.28	30.73
Noise Level	$\sigma = 50$												
BM3D [17]	26.13	29.69	26.68	25.04	25.82	25.10	25.90	29.05	27.22	26.78	26.81	26.46	26.72
WNNM [18]	26.45	30.33	26.95	25.44	26.32	25.42	26.14	29.25	27.79	26.97	26.94	26.64	27.05
EPLL [35]	26.10	29.12	26.80	25.12	25.94	25.31	25.95	28.68	24.83	26.74	26.79	26.30	26.47
MLP [19]	26.37	29.64	26.68	25.43	26.26	25.56	26.12	29.32	25.24	27.03	27.06	26.67	26.78
TNRD [15]	26.62	29.48	27.10	25.42	26.31	25.59	26.16	28.93	25.70	26.94	26.98	26.50	26.81
DnCNN [21]	26.96	29.91	27.30	25.66	26.73	25.84	26.49	29.31	26.02	27.16	27.21	26.85	27.12
UNet	27.03	30.42	27.38	26.27	26.95	25.98	26.57	29.63	26.63	27.36	27.33	27.11	27.39
BM3D-UNet	27.16	30.52	27.43	26.45	26.99	25.99	26.58	29.67	27.66	27.38	27.35	27.17	27.53
WNNM-UNet	27.19	30.79	27.44	26.58	27.05	26.02	26.61	29.74	27.98	27.42	27.39	27.19	27.61

on the setting of NSS-UNet, please refer to Fig. 2. Finally, the average mean squared error (MSE) loss is adopted to train the modified UNet,

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N} \|x_{gt}^{i} - \mathbf{x}_t^i - \mathcal{R}(\mathbf{y}^i, \mathbf{x}_t^i)\|_2^2, \tag{4}$$

where x_{gt}^{i} is the ground truth, $\mathbf{x}_t^i$ denotes the denoising result by NSS-based methods, and N is the number of the training pairs. Given a testing noisy image $\mathbf{y}$, the NSS-based method are first used to produce $\mathbf{x}_t$, and the denoising result is then obtained by $\mathbf{x} = \mathbf{x}_t + \mathcal{R}(\mathbf{y}, \mathbf{x}_t)$.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. EXPERIMENTAL SETTING

1) TRAINING AND TESTING DATA

We choose 200 images from BSD300 [36] to train our model, where the image size is 321 × 481. Data augmentation including flipping, rotation and data refinement by resizing images with different factors, i.e. 1, 0.9, 0.8 and 0.7, have been used for generating more training data. We crop 60,000 image patches with size of 120 × 120 to train our model. To generate $\mathbf{x}_t$, we only use the noisy training patches as the input to NSS-based algorithms. To test NSS-UNet, three datasets are used for evaluation: (*i*) Set12 (The 12 testing images used in [21]), (*ii*) BSD68 (The 68 natural images from Berkeley segmentation dataset [36]) and (*iii*) Urban100 (a dataset contains structured scenes in [37]).

TABLE 2. Run Time (in seconds, *s*) of different methods with noise level $\sigma = 50$, NSS-based method runs on CPU while UNet runs on both CPU (left) and GPU (Right).

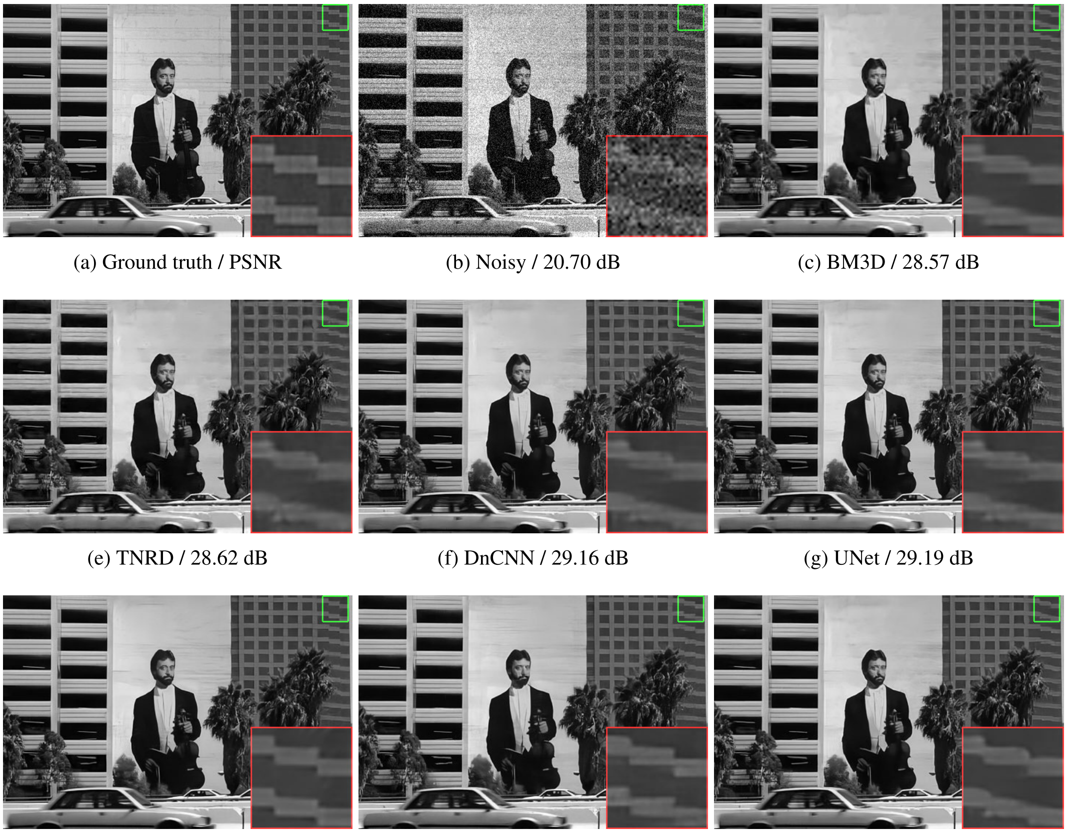
Methods	UNet	BM3D-UNet	WNNM-UNet
256 × 256	0.65/0.015	1.60/0.97	138.55/136.47
512 × 512	2.56/0.022	6.57/4.15	562.52/558.39

2) PARAMETER SETTING

The Adam [38] algorithm with $\beta_1 = 0.9$ is used to train our model, and the size of mini-batch is 32. All the models are trained with 35 epochs, the learning rate for the first 20 epoches is 10^{-3}, and it becomes 10^{-4} for the remaining epoches. The network parameters are initialized based on the method in [39]. It takes about 9.5 hours to train our model with the MatConvNet package [40] on a Nvidia GeForce GTX 1080 GPU. The source code and test results will be released after the publication of this work.

B. QUANTITATIVE AND QUALITATIVE EVALUATION

We compare our NSS-UNet with several state-of-the-art denoising algorithms, including BM3D [17], WNNM [18], EPLL [35], CSF [13], TNRD [15], and DnCNN [21]. The results of the baseline UNet are also reported. Three noise levels, i.e., $\sigma = 15$, 25 and 50, are considered in our experiments. Table 1 lists the PSNR results of the competing methods on Set12. The best two PSNR results for each image are highlighted in red and blue, respectively. Our BM3D-UNet and WNNM-UNet achieve the highest



(a) Ground truth / PSNR (b) Noisy / 20.70 dB (c) BM3D / 28.57 dB

(e) TNRD / 28.62 dB (f) DnCNN / 29.16 dB (g) UNet / 29.19 dB

(g) GCDN / 29.92 dB (g) NLRN / 29.50 dB (h) BM3D-UNet / 29.35 dB

FIGURE 3. Denoising results of the image in Set68 with noise level 25.

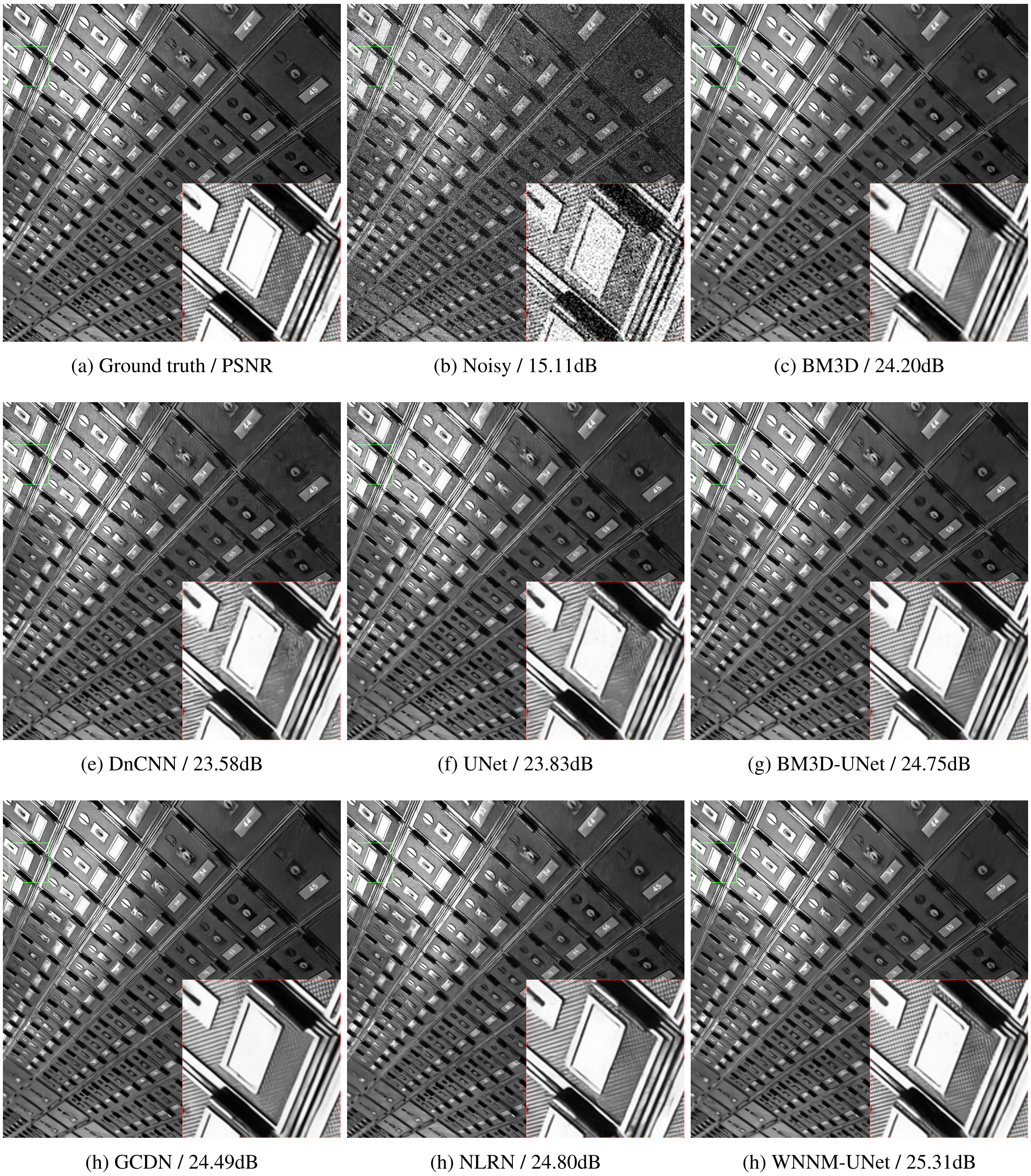
TABLE 3. The PSNR/SSIM results of different methods on BSD68 and Urban100.

Dataset	Noise	BM3D	TNRD	DnCNN	UNet	N^{3}Net	GCDN	NLRN	BM3D-UNet	WNNM-UNet
		PSNR / SSIM	PSNR / SSIM	PSNR / SSIM	PSNR / SSIM	PSNR / SSIM	PSNR / SSIM	PSNR / SSIM	PSNR / SSIM	PSNR / SSIM
Set12	$\sigma = 15$	32.37 / 0.8952	32.50 / 0.8967	32.82 / 0.9027	32.91 / 0.9042	- / -	33.14 / 0.9072	33.16 / 0.9070	32.99 / 0.9053	33.08 / 0.9062
	$\sigma = 25$	29.97 / 0.8504	30.06 / 0.8520	30.40 / 0.8618	30.53 / 0.8649	30.55 / -	30.78 / 0.8687	30.80 / 0.8689	30.64 / 0.8670	30.73 / 0.8685
	$\sigma = 50$	26.72 / 0.7676	26.81 / 0.7667	27.12 / 0.7827	27.39 / 0.7919	27.43 / -	27.60 / 0.7957	27.64 / 0.7980	27.53 / 0.7973	27.61 / 0.7997
BSD68	$\sigma = 15$	31.08 / 0.8722	31.42 / 0.8826	31.73 / 0.8906	31.77 / 0.8923	- / -	31.83 / 0.8933	31.88 / 0.8932	31.76 / 0.8922	31.80 / 0.8927
	$\sigma = 25$	28.57 / 0.8012	28.92 / 0.8157	29.23 / 0.8278	29.29 / 0.8320	29.30 / -	29.35 / 0.8332	29.41 / 0.8331	29.31 / 0.8324	29.35 / 0.8335
	$\sigma = 50$	25.62 / 0.6869	25.97 / 0.7029	26.23 / 0.7061	26.36 / 0.7277	26.39 / -	26.38 / 0.7389	26.47 / 0.7298	26.39 / 0.7283	26.43 / 0.7304
Urban100	$\sigma = 15$	32.34 / 0.9220	32.05 / 0.9197	32.67 / 0.9250	32.66 / 0.9274	- / -	33.47 / 0.9358	33.42 / 0.9348	32.96 / 0.9311	33.31 / 0.9322
	$\sigma = 25$	29.70 / 0.8777	29.28 / 0.8733	29.97 / 0.8792	30.01 / 0.8884	30.19 / -	30.95 / 0.9020	30.88 / 0.9003	30.40 / 0.8946	30.85 / 0.8968
	$\sigma = 50$	25.94 / 0.7791	25.64 / 0.7741	26.28 / 0.7869	26.50 / 0.8014	26.82 / -	27.41 / 0.8160	27.40 / 0.8244	26.91 / 0.8122	27.28 / 0.8206

average PSNR result. Moreover, even for the images *House* and *Barbara*, BM3D-UNet also outperforms BM3D, while WNNM-UNet performs consistently better than the competing methods on all the 12 images. In order to investigate the performance of the proposed method comparing with the state-of-the-art non-local-based image denoising methods, Table 3 further reports the average PSNR and SSIM results on Set12, BSD68 and Urban100 with different noise level. All results have been obtained running the pretrained models provided by the authors, except for N^{3}Net at $\sigma = 15$ which is unavailable. Our WNNM-UNet and BM3D-UNet can always achieve denoising results comparable with current best performing non-local-based methods, and work especially well at medium to high levels of noise. As illustrated in Fig. 4, the proposed method provides the best visual quality on high noise level image which contains rich repetitive structures, and outperforms the strong baselines GCDN and NLRN with large margins, namely 1.02 dB and 0.51 dB for WNNM-UNet respectively.

C. THE EFFECT OF $\mathbf{x}_t$ BY DIFFERENT DENOISERS

Using Set12, we evaluate the effect of $\mathbf{x}_t$ produced by different denoisers. We consider one non-NSS-based method, i.e., EPLL, and two NSS-based methods, i.e., BM3D, and WNNM. Three variants of our models are then implemented, i.e., EPLL-UNet, BM3D-UNet, and WNNM-UNet. The denoising results and the run time are provided in Tables 4 and 2, respectively. One can see that NSS-based estimate of $\mathbf{x}_t$ generally results in better performance than both baseline UNet and the corresponding NSS-based method. And verify that a better NSS-based estimate also lead to better denoising result.


(a) Ground truth / PSNR

(b) Noisy / 15.11dB

(c) BM3D / 24.20dB

(e) DnCNN / 23.58dB

(f) UNet / 23.83dB

(g) BM3D-UNet / 24.75dB

(h) GCDN / 24.49dB

(h) NLRN / 24.80dB

(h) WNNM-UNet / 25.31dB

FIGURE 4. Denoising results on a representative image from Urban100 with noise level 50.

TABLE 4. The PSNR (DB) results of different non-NSS-based and NSS-based UNet on Set12 with noise level $\sigma = 50$.

Images	C.man	House	Peppers	Starfish	Monar.	Airpl.	Parrot	Lena	Barbara	Boat	Man	Couple	Average
UNet	27.03	30.42	27.38	26.27	26.95	25.98	26.57	29.63	26.63	27.36	27.33	27.11	27.39
EPLL	26.10	29.12	26.80	25.12	25.94	25.31	25.95	28.68	24.83	26.74	26.79	26.30	26.47
EPLL-UNet.	27.09	30.39	27.39	26.19	26.97	26.01	26.59	29.62	26.46	27.38	27.36	27.13	27.38
BM3D	26.13	29.69	26.68	25.04	25.82	25.10	25.90	29.05	27.22	26.78	26.81	26.46	26.72
BM3D-UNet	27.16	30.52	27.43	26.45	26.99	25.99	26.58	29.67	27.66	27.38	27.35	27.17	27.53
WNNM	26.45	30.33	26.95	25.44	26.32	25.42	26.14	29.25	27.79	26.97	26.94	26.64	27.05
WNNM-UNet	27.19	30.79	27.44	26.58	27.05	26.02	26.61	29.74	27.98	27.42	27.39	27.19	27.61

V. CONCLUSION

In this paper, we present a simple yet effective NSS-UNet architecture to combine the NSS-based method with deep CNN for image denoising. In NSS-UNet, the NSS-based method is first employed to generate an estimate of x_t from the noisy image y. Then, by taking both x_t and y as inputs, a modified UNet is trained to obtain the final denoising result. NSS-UNet outperforms existing popular denoising methods in terms of both quantitative measure and visual perceptual quality. On Set12, at the noise level $\sigma = 50$, the average PSNR by our WNNM-UNet can be about 0.5 dB higher than the state-of-the-art DnCNN, and 0.2 dB higher than the baseline UNet. One limitation of NSS-UNet is that its efficiency is highly dependent on the complexity of the NSS-based method (e.g., WNNM). In future work, we will investigate efficient solution to directly incorporate the NSS prior into appropriate CNN architecture.

REFERENCES

[1] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D, Nonlinear Phenomena*, vol. 60, nos. 1–4, pp. 259–268, Nov. 1992.

[2] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin, "An iterative regularization method for total variation–based image restoration," *Multiscale Model. Simul.*, vol. 4, no. 2, pp. 460–489, Jan. 2005.

[3] J. Zou, M. Shen, Y. Zhang, H. Li, G. Liu, and S. Ding, "Total variation denoising with non–convex regularizers," *IEEE Access*, vol. 7, pp. 4422–4431, 2019.

[4] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 2272–2279.

[5] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.

[6] W. Dong, L. Zhang, G. Shi, and X. Li, "Nonlocally centralized sparse representation for image restoration," *IEEE Trans. Image Process.*, vol. 22, no. 4, pp. 1620–1630, Apr. 2013.

[7] X. Lan, S. Roth, D. Huttenlocher, and M. J. Black, "Efficient belief propagation with learned higher-order Markov random fields," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Berlin, Germany: Springer, May 2006, pp. 269–282.

[8] S. Z. Li, *Markov Random Field Modeling in Image Analysis*. London, U.K.: Springer, 2009.

[9] S. Roth and M. J. Black, "Fields of experts," *Int. J. Comput. Vis.*, vol. 82, no. 2, pp. 205–229, Apr. 2009.

[10] X. Lu, H. Yuan, P. Yan, L. Li, and X. Li, "Image denoising via improved sparse coding," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2011, pp. 1–11.

[11] T. J. Abrahamsen and L. K. Hansen, "Sparse non-linear denoising: Generalization performance and pattern reproducibility in functional MRI," *Pattern Recognit. Lett.*, vol. 32, no. 15, pp. 2080–2085, Nov. 2011.

[12] X. Zhang, X. Feng, W. Wang, and G. Liu, "Image denoising via 2D dictionary learning and adaptive hard thresholding," *Pattern Recognit. Lett.*, vol. 34, no. 16, pp. 2110–2117, Dec. 2013.

[13] U. Schmidt and S. Roth, "Shrinkage fields for effective image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2774–2781.

[14] Y. Chen, W. Yu, and T. Pock, "On learning optimized reaction diffusion processes for effective image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5261–5269.

[15] Y. Chen and T. 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, Jun. 2017.

[16] A. Buades, B. Coll, and J.-M. Morel, "A non–local algorithm for image denoising," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2005, pp. 60–65.

[17] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform–domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007.

[18] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Weighted nuclear norm minimization with application to image denoising," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2862–2869.

[19] H. C. Burger, C. J. Schuler, and S. Harmeling, "Image denoising: Can plain neural networks compete with BM3D?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2392–2399.

[20] V. Jain and S. Seung, "Natural image denoising with convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2009, pp. 769–776.

[21] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[22] D. Wu, K. Kim, G. El Fakhri, and Q. Li, "A cascaded convolutional neural network for X-ray low-dose CT image denoising," 2017, *arXiv:1705.04267*. [Online]. Available: https://arxiv.org/abs/1705.04267

[23] S. Lefkimmiatis, "Universal denoising networks : A novel CNN architecture for image denoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 3204–3213.

[24] P. Qiao, Y. Dou, W. Feng, R. Li, and Y. Chen, "Learning non-local image diffusion for image denoising," in *Proc. ACM Multimedia Conf. (MM)*, 2017, pp. 1847–1855.

[25] S. Lefkimmiatis, "Non-local color image denoising with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3587–3596.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv. (MICCAI)*, 2015, pp. 234–241.

[27] J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 341–349.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[29] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 448–456.

[30] A. Buades, B. Coll, and J.-M. Morel, "Nonlocal image and movie denoising," *Int. J. Comput. Vis.*, vol. 76, no. 2, pp. 123–139, Feb. 2008.

[31] B. Ahn and N. I. Cho, "Block-matching convolutional neural network for image denoising," Apr. 2017, *arXiv:1704.00524*. [Online]. Available: https://arxiv.org/abs/1704.00524

[32] T. Plötz and S. Roth, "Neural nearest neighbors networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 1087–1098.

[33] D. Valsesia, G. Fracastoro, and E. Magli, "Deep graph-convolutional image denoising," 2019, *arXiv:1907.08448*. [Online]. Available: https://arxiv.org/abs/1907.08448

[34] D. Liu, B. Wen, Y. Fan, C. C. Loy, and T. S. Huang, "Non-local recurrent network for image restoration," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 1673–1682.

[35] D. Zoran and Y. Weiss, "From learning models of natural image patches to whole image restoration," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 479–486.

[36] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011.

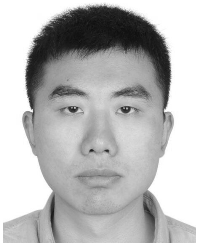
[37] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5197–5206.

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.

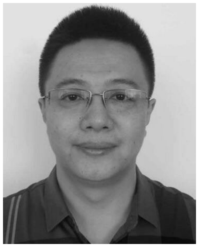
[39] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human–level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[40] A. Vedaldi and K. Lenc, "MatConvNet: Convolutional neural networks for MATLAB," in *Proc. 23rd ACM Int. Conf. Multimedia (MM)*, 2015, pp. 689–692.

ZIFEI YAN received the Ph.D. degree in computer application technology from the Harbin Institute of Technology, Harbin, China, in 2010. From May 2007 to August 2007 and from December 2008 to January 2009, she was a Research Assistant with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong. From September 2014 to September 2015, she was a Visiting Scholar with the University of Pittsburgh. She is currently a Lecturer with the Department of Media Technology and Art, School of Architecture, Harbin Institute of Technology. She has published more than 20 articles in academic journals and conferences. Her current research interests include machine learning and computer vision.



SHI GUO received the B.S. degree from the Harbin Institute of Technology, Harbin, China, in July 2017, where he is currently pursuing the master's degree in computer science. He has published two articles in academic conferences. His research areas focus on image denoising and image superresolution.



GANG XIAO received the M.M. degree in clinical medicine from Harbin Medical Sciences University, Harbin, China, in 2002. He is currently an Associate Chief Physician and also the President of the No. 962 Hospital of PLA, Harbin. He has published more than 20 articles in academic journals. His research interests include medical biometrics and computerized diagnosis.

HONGZHI ZHANG received the M.Sc. degree in material science and engineering and the Ph.D. degree in computer application technology from the Harbin Institute of Technology, Harbin, China, in 2001 and 2007, respectively. He is currently an Associate Professor with the School of Computer Science and Technology, Harbin Institute of Technology. His research interests include image processing, pattern recognition, data mining, and biomedical engineering of Chinese medicine.