Deep Combined Image Denoising with Cloud Images

Sifeng Xia[1], Jiaying Liu[1⋆], Wenhan Yang[1], Mading Li[1] and Zongming Guo[1,2]

[1]Institute of Computer Science and Technology, Peking University, Beijing, China
[2]Cooperative Medianet Innovation Center, Shanghai, China

Abstract. Image denoising methods essentially lose some high-frequency (HF) information in denoising. To address this issue, we propose an end-to-end trainable deep network to additionally utilize online retrieved cloud images to compensate for the HF information loss based on the the internal inferred results. In particular, the noise inference network first infers a noise map from the noisy image and derives an intermediate image by removing the noise map from the noisy image. Then the external online compensation is performed based on the intermediate image. The final results are obtained by fusing the intermediate image with external HF maps extracted by the external HF compensation network. Extensive experimental results demonstrate that our method achieves notably better performance than state-of-the-art denoising methods.

Keywords: Image denoising, high-frequency information loss, online compensation, external high-frequency map

1 Introduction

Image denoising aims to obtain a clean image from a noisy one and it is widely applied in practical issues. Due to the information loss caused by the noise contamination, image denoising is an ill-posed problem. Early image denoising methods estimate and remove the noise based on local statistics of the noisy image. For example, Li *et al.* [9] proposed a method that removes the noise by studying both signal and noise characteristics under overcomplete expansion. An adaptive denoising method for images decomposed in overcomplete oriented pyramids is proposed in [7]. The sparse representation based method [4] groups similar 2D image fragments in the noisy image into 3D data arrays to enhance the sparse representation. In these methods, only limited local signal of the noisy image is utilized for removing the noise.

For the purpose of better denoising performance, many methods are successively proposed utilizing additional external data to bring more information for image denoising. The K-SVD algorithm is used in [6] to obtain a dictionary

[⋆] Corresponding author. This work was supported by National Natural Science Foundation of China under contract No.U1636206.

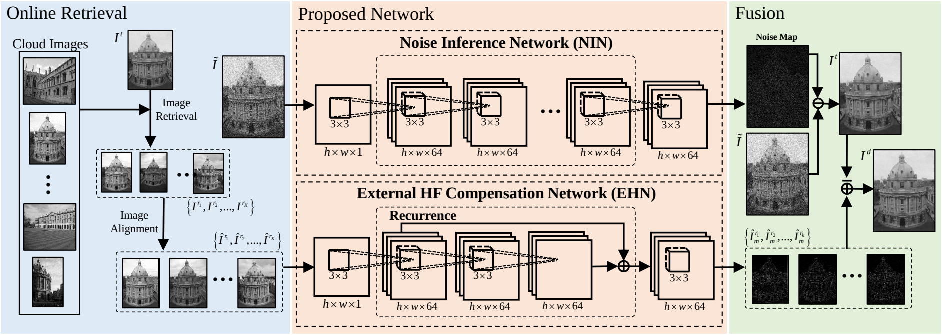

Online Retrieval
Cloud Images
I^t
Image Retrieval
$\{I^{r_1}, I^{r_2}, ..., I^{r_K}\}$
Image Alignment
$\{\tilde{I}^{r_1}, \tilde{I}^{r_2}, ..., \tilde{I}^{r_K}\}$
$\tilde{I}$
Proposed Network
Noise Inference Network (NIN)
3×3
$h \times w \times 1$
$h \times w \times 64$
External HF Compensation Network (EHN)
Recurrence
Fusion
Noise Map
I^t
$\tilde{I}$
I^d
$\{\tilde{I}^{r_1}_m, \tilde{I}^{r_2}_m, ..., \tilde{I}^{r_K}_m\}$

Fig. 1. The framework of the proposed deep denoising method with online compensation.

that describes the image content effectively based on the noisy image and a high-quality image dataset. Mairal *et al.* [12] proposed to combine a learned dictionary from external data with self-similarities of natural images for image denoising. Context-aware sparsity prior was proposed in [13] to facilitate the sparsity-based regularization for image denoising. Dong *et al.* [5] proposed a method that introduces the concept of sparse coding noise and try suppressing the noise over a learned dictionary. Besides, Yue *et al.* [18] combined spatial and frequency filtering with the assistance of retrieved web images to preserve details.

Recently, many learning based methods have been proposed and they posses impressive performance. A random field-based architecture is constructed in [14] combining the image model and the optimization algorithm in a single unit for image denoising. Chen and Pock [3] described a flexible learning framework that simultaneously learns all the parameters including the filters and the influence functions from the training data through a loss based approach. Besides, a feed-forward convolutional network (CNN) is constructed by Zhang *et al.* to infer a map from the noisy image for denoising [19], which obtains state-of-the-art performance. However, despite the impressive result Zhang has achieved, some high-frequency (HF) information is still lost due to the ambiguity nature of image denoising and the problem that mean squared error leads to "regression to mean" [15].

To address the above issues, we propose a unified deep network that utilizes additional online retrieved cloud images to facilitate image denoising. Specifically, our work first initially removes an inferred noise map from the noisy image to obtain an intermediate image. Then the intermediate image is utilized to retrieve and align multiple reference images. The final result with well recovered HF detail is derived by extracting external HF (EHF) maps with the external HF compensation network and fusing the maps with the intermediate image.

The rest of the paper is organized as follows. Sec. 2 introduces the proposed denoising method. The unified network is first illustrated and then we describe the utilization of reference images. Experimental results are shown in Sec. 3 and concluding remarks are given in Sec. 4.

2 Deep Online Compensation for Image Denoising

In this section, the proposed denoising method is presented. As shown in Fig. 1, given a noisy image $\tilde{I}$, we first derive an intermediate image I^t by removing the noise map inferred with the noise inference network (NiNet) from $\tilde{I}$. And then I^t is utilized for retrieving and aligning the reference images. There are K reference cloud images defined as $\{I^{r_1}, I^{r_2}, ..., I^{r_K}\}$. Their corresponding aligned images are represented by $\{\hat{I}^{r_1}, \hat{I}^{r_2}, ..., \hat{I}^{r_K}\}$. For each $\hat{I}^r$ of $\{\hat{I}^{r_1}, \hat{I}^{r_2}, ..., \hat{I}^{r_K}\}$, it is taken as an input of the external HF compensation network (EhNet), which extracts the EHF map $\hat{I}^r_m$. Finally, the extracted EHF maps $\{\hat{I}^{r_1}_m, \hat{I}^{r_2}_m, ..., \hat{I}^{r_K}_m\}$ are fused with the intermediate image I^t based on the patch matching results between each $\hat{I}^r$ and I^t.

2.1 The Proposed Network

As can be observed in Fig. 1, the proposed network consists of two components: the noise inference network (NiNet) and the external high-frequency compensation network (EhNet). The first component NiNet proposed by [19] is utilized to infer a noise map from the noisy image $\tilde{I}$. As shown in Fig. 1, $\tilde{I}$ is directly used as the input of NiNet. The size of convolutional layers is $h \times w \times *$, where h and w are the height and the width of the input image, respectively. Except for the last layer, there are 64 filters with the size $3 \times 3 \times c$ to generate feature maps in each layer. c is 1 in the first layer and 64 for the rest. The rectified linear units (ReLU) are utilized for nonlinearity and batch normalization is utilized between convolution and Relu in the intermediate layers. One $3 \times 3 \times 64$ filter is used in the last layer to estimate the noise map.

With the inferred noise map, the intermediate image I^t is then generated as follows:

$$I^t = \tilde{I} \ominus \varphi(\tilde{I}), \tag{1}$$

where $\ominus$ is the direct minus operation between $\tilde{I}$ and $\varphi(\tilde{I})$. $\varphi(\tilde{I})$ represents the process that infers the noise map from $\tilde{I}$ using NiNet. The training loss function of NiNet is defined by the MSE between I^t and the ground truth signal.

NiNet works well in predicting the noise map and removing the noise from the input image. However, during the removal process, it also essentially leads to losing some HF information, as shown in Fig. 2 (b). This inspires us to construct EhNet to extract a significant EHF map $\hat{I}^r_m$ from each aligned reference image $\hat{I}^r$ for further compensation.

During the training process of EhNet, $\hat{I}^r$ is utilized as the input. In particular, $\hat{I}^r$ is generated from the ground truth image in the training process of NiNet. The settings of convolutional layers and filters are the same as NiNet, and ReLU is also used for nonlinearity. Besides, the recurrent network is utilized here to accelerate training. With the EHF maps extracted by EhNet, the intermediate image can be enhanced as follows:

$$I^d = I^t \bar{\oplus} \psi(\hat{I}^r), \tag{2}$$

where ψ is the formulation of the process that extracts the HF map $\hat{I}^r_m$. I^d represents the final result. The operation $\bar{\oplus}$ represents the combination of I^t and $\hat{I}^r_m$. In the training process, operation $\bar{\oplus}$ directly adds $\hat{I}^r_m$ to I_t. In the practical testing process, the EHF maps will be utilized based on patch matching results, which will be elaborated in Sec. 2.3. The training loss function of EhNet is defined as MSE between I^d and the ground truth image.

2.2 References Retrieval and Registration

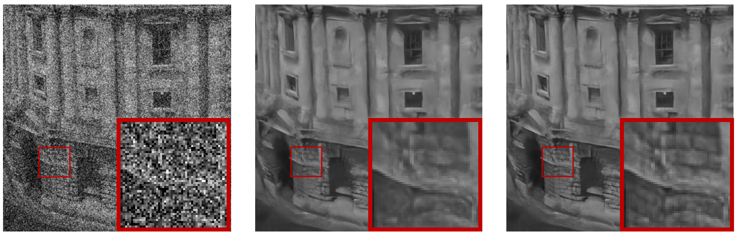
In the practical testing process, we search for K reference images to extract EHF maps for compensation. The method mentioned in [10] is utilized for retrieving and the immediate image I^t is used as the query image. During the searching process, the SURF detector [2] is first used to detect key points. Then a 144-dimension vector that contains discriminative information is extracted for each patch centered at the key point. Finally, the BOW model [8] is used for indexing and retrieving the reference images with extracted feature vectors.

However, considering different scales and viewpoints factors between I^t and reference images, we cannot directly extract EHF maps from the references and add them to I^t for compensation. As a result, each reference image I^r is aligned to I^t before we extract EHF maps. In order to align these reference images, we first detect SIFT features [11] of I^t and I^r and match their feature points. Then the RANSAC algorithm is performed over the matched points to find the best homography transformation matrix. Finally, the aligned reference images $\{\hat{I}^{r_1}, \hat{I}^{r_2}, ..., \hat{I}^{r_K}\}$ are derived from $\{I^{r_1}, I^{r_2}, ..., I^{r_K}\}$ based on the transformation matrix.

2.3 External High-Frequency Maps Fusion

After obtaining the aligned reference images, the EHF maps are then extracted from each $\hat{I}^r$ by EhNet for compensation. As pixels in the aligned references are still not exactly corresponding to those pixels at the same position in I^t, patch matching is further performed to find corresponding pixels between I^t and each aligned reference image $\hat{I}^r$ to guide the combination between I^t and the extracted EHF maps.

Specifically, I^t is first split into $\sqrt{n} \times \sqrt{n}$ query patches with 4-pixel overlapping. And then for each aligned reference image $\hat{I}^r$, we search for its best matching patch of each query patch within a search window. Since small patches contain little structural information of raw images, patch matching results with small patch sizes are not accurate. Thus we perform the patch matching with large patch sizes. Considering that it is impossible for each large patch in I^t



(a) Input $\tilde{I}$ (b) Intermediate I^t (c) Result I^d

Fig. 2. An example of our denoising method at noise level 50.

to have an exact corresponding large patch in $\hat{I}^r$, a method that adaptively adjusts patch sizes according to patch difference [17] is adopted for more accurate patch matching.

Let $\mathbf{P}_i$ denote the query patch of size $\sqrt{n} \times \sqrt{n}$ in I^t centered at position i and $\mathbf{Q}^i_j$ denote the matching candidate in $\hat{I}^r$ centered at j. For a query patch $\mathbf{P}_i$, we search for its best matching candidate patch with the minimum patch difference within the search window of size $3\sqrt{n} \times 3\sqrt{n}$ centered at i in $\hat{I}^r$. The patch difference between $\mathbf{P}_i$ and $\mathbf{Q}^i_j$ is defined as:

$$d(\mathbf{P}_i, \mathbf{Q}^i_j) = ||\mathbf{P}_i - \mathbf{Q}^i_j||^2_2 + \rho||\nabla(\mathbf{P}_i) - \nabla(\mathbf{Q}^i_j)||^2_2, \tag{3}$$

where ∇ is the operation that calculates the gradient of the patches and ρ is the weighting parameter that controls the relative importance of pixel value differences and their gradient differences, which is empirically set to be 10. Then the value of $d(\mathbf{P}_i, \mathbf{Q}^i_j)/(\sqrt{n} \times \sqrt{n})$ is defined as the gradient mean square error (GMSE) and G^{min}_i is set as the minimum GMSE value between the query patch $\mathbf{P}_i$ and the best candidate patch $\mathbf{Q}^i_j$. In particular, the value of G^{min}_i is consistent with the quality of patch matching. In order to improve the quality of patch matching, patch sizes are then adaptively adjusted according to G^{min}_i as follows:

$$\sqrt{n} = \begin{cases} 21, & G^{min}_i <= 400, \\ 17, & 400 < G^{min}_i <= 600, \\ 13, & 600 < G^{min}_i <= 800, \\ 9, & G^{min}_i > 800. \end{cases} \tag{4}$$

Patch matching is performed at initial size 21×21 and will be adjusted if the value of G^{min}_i is too large according to Eq. 4. Finally the best candidate patch $\mathbf{Q}^i_{j_0}$ will be found.

After patch matching, pixels at the same position in the matched patches between I^t and each $\hat{I}^r$ are matched. Note that since the EHF maps are directly derived from the aligned reference images, the pixel-wise matching correlation between I^t and each $\hat{I}^r_m$ is the same as that of I^t and each $\hat{I}^r$. For pixel $\mathbf{p}$ in I^t, we define the set of its matching pixels in K EHF maps as $\Omega_{\mathbf{p}}$. The final result image I^d is then obtained by combining I^t with the EHF maps as follows:

$$I^d_{\mathbf{p}} = I^t_{\mathbf{p}} + \begin{cases} \dfrac{\sum\limits_{\mathbf{q}\in\Omega_{\mathbf{p}}} \hat{I}^r_{m,\mathbf{q}} \cdot e^{\frac{-d(\mathbf{p},\mathbf{q})}{100}}}{\sum\limits_{\mathbf{q}\in\Omega_{\mathbf{p}}} e^{\frac{-d(\mathbf{p},\mathbf{q})}{100}}}, & |\Omega_{\mathbf{p}}| \neq 0, \\ 0, & |\Omega_{\mathbf{p}}| = 0, \end{cases} \tag{5}$$

where $I^d_{\mathbf{p}}$ and $I^t_{\mathbf{p}}$ are the values of the pixel $\mathbf{p}$ in image I^d and I^t, respectively. Similarly, $\hat{I}^r_{m,\mathbf{q}}$ is the value of pixel $\mathbf{q}$ in map $\hat{I}^r_m$. $|\Omega_{\mathbf{p}}|$ represents the number of elements in set $\Omega_{\mathbf{p}}$. $d(\mathbf{p},\mathbf{q})$ is the GMSE value between the patches that p and q belong to.

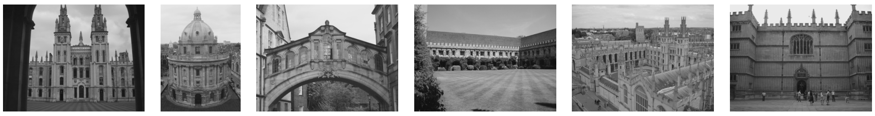


Fig. 3. Testing images from 'a' to 'f'.

3 Experimental Results

We train our EhNet based on 91 images in [16] and 200 training images in *BSD500* [1]. The noisy images are generated by adding Gaussian noise with a random standard deviation of $\sigma \in [0, 55]$. With the mentioned 291 images, we first transfer the images to gray images. Then we generate sub-images at the size of 50×50 from the 291 images with the stride step of 25 pixel. As a result, 45824 sub-images are obtained for training. The learning rate is initially set as 10^{-3}. As for NiNet, weights provided by [19] are directly used.

Table 1. PSNR values of the denoising results of 6 images at noise level 40 by different methods.

Images	BM3D	NCSR	DnCNN	Proposed
a	29.11	29.68	30.45	**30.61**
b	28.26	28.48	29.05	**29.26**
c	27.20	27.57	**28.15**	**28.15**
d	25.90	26.67	27.09	**27.27**
e	26.39	26.75	27.31	**27.43**
f	28.69	28.99	29.52	**29.70**
Avg.	27.67	28.11	28.68	**28.85**

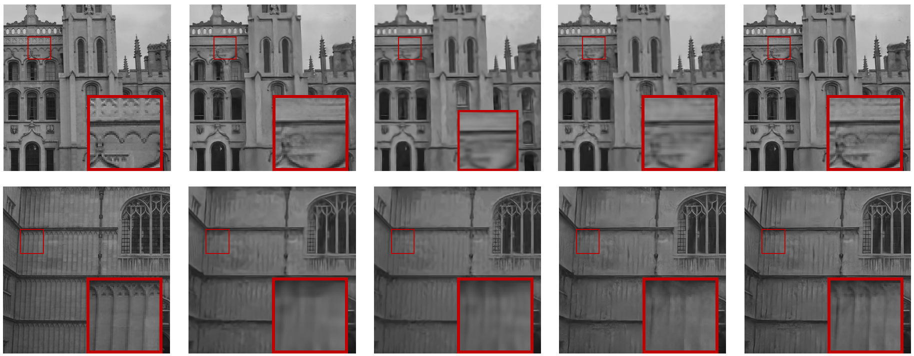

(a) GT (b) BM3D (c) NCSR (d) DnCNN (e) Proposed

Fig. 4. Subjective results of different methods at noise level 40 for images ‘a’ and ‘f’. The regions with HF signals have been highlighted in the red rectangle and enlarged for comparison.

Table 2. Average PSNR and SSIM values of the results at noise level 30, 40 and 50 by different methods.

Noise Level	BM3D	NCSR	DnCNN	Proposed
30	28.07	29.37	29.93	**30.12**
	0.7599	0.8068	0.8232	**0.8392**
40	27.67	28.11	28.68	**28.85**
	0.7490	0.7661	0.7844	**0.7998**
50	27.27	27.21	27.77	**27.89**
	0.7356	0.7353	0.7529	**0.7662**

We compare our algorithm with different denoising methods including a single image based denoising method BM3D [4] and a dictionary based method NCSR [5]. Besides, the intermediate images derived by NiNet [19] are also shown and named as DnCNN. DnCNN is one of the newest deep based denoising methods without using external references. The Oxford Building dataset[1] is utilized to simulate the web images. There are totally 6 testing images named from ‘a’ to ‘f’ for comparison, as shown in Fig. 3. The reference images number K is set to be 4.

Table 1 shows the objective results of the 6 testing images at noise level 40. Our proposed method has obtained the best PSNR values for all of the 6 images. We also show average PSNR and SSIM values for the 6 images at different noise levels in Table 2. Our method outperforms all the other methods at each noise level in PSNR and SSIM values.

Subjective results are shown in Fig. 4. The other three methods all fail to preserve some HF signal while removing the noise. On the contrary, our method successfully extracts some HF information from the reference images and com-

[1] `http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/`

pensates for the HF loss in noise removal, which is well identified by the subjective results. Our method indeed achieves a much better success in visual quality recovering the lost HF signal.

4 Conclusion

In this paper, we propose a deep online compensation network for image denoising. With the noise map estimated by NiNet, we initially obtain an intermediate image by removing the noise map. Then the EHF maps of the aligned reference images are further extracted for compensation. The final compensated denoising result is obtained by fusing the EHF maps with the intermediate image. Extensive experimental results demonstrate that the proposed method can well extract external HF maps from the reference images and significantly improve the denoising results via compensating for HF information loss with EHF maps.

References

1. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(5), 898–916 (2011)
2. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Speeded-up robust features (surf). Computer vision and image understanding 110(3), 346–359 (2008)
3. Chen, Y., Pock, T.: Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. IEEE Transactions on Pattern Analysis and Machine Intelligence PP(99), 1–1 (2016)
4. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3-d transform-domain collaborative filtering. IEEE Transactions on Image Processing 16(8), 2080–2095 (Aug 2007)
5. Dong, W., Zhang, L., Shi, G., Li, X.: Nonlocally centralized sparse representation for image restoration. IEEE Transactions on Image Processing 22(4), 1620–1630 (April 2013)
6. Elad, M., Aharon, M.: Image denoising via sparse and redundant representations over learned dictionaries. IEEE Transactions on Image Processing 15(12), 3736–3745 (Dec 2006)
7. Guerrero-Colon, J.A., Portilla, J.: Two-level adaptive denoising using gaussian scale mixtures in overcomplete oriented pyramids. In: Proc. IEEE Int'l Conf. Image Processing. vol. 1, pp. I–105–8 (September 2005)
8. Li, F., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition. pp. 524–531 (2005)
9. Li, X., Orchard, M.T.: Spatially adaptive image denoising under overcomplete expansion. In: Proc. IEEE Int'l Conf. Image Processing. vol. 3, pp. 300–303 vol.3 (2000)
10. Liu, J., Yang, W., Zhang, X., Guo, Z.: Retrieval compensated group structured sparsity for image super-resolution. IEEE Transactions on Multimedia 19(2), 302–316 (February 2017)
11. Lowe, D.: Distinctive image features from scale-invariant keypoints. Int'l Journal of Computer Vision 60(2), 91–110 (November 2004)

12. Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A.: Non-local sparse models for image restoration. In: 2009 IEEE 12th International Conference on Computer Vision. pp. 2272–2279 (Sept 2009)
13. Ren, J., Liu, J., Guo, Z.: Context-aware sparse decomposition for image denoising and super-resolution. IEEE Transactions on Image Processing 22(4), 1456–1469 (Apr 2013)
14. Schmidt, U., Roth, S.: Shrinkage fields for effective image restoration. In: Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition. pp. 2774–2781 (June 2014)
15. Timofte, R., Smet, V., Gool, L.: Semantic super-resolution: When and where is it useful? Computer Vision and Image Understanding 142, 1–12 (2016)
16. Yang, J., Wright, J., Huang, T., Ma, Y.: Image super-resolution via sparse representation. IEEE Transactions on Image Processing 19(11), 2861–2873 (2010)
17. Yue, H., Sun, X., Yang, J., Wu, F.: Landmark image super-resolution by retrieving web images. IEEE Transactions on Image Processing 22(12), 4865–4875 (December 2013)
18. Yue, H., Sun, X., Yang, J., Wu, F.: Cid: Combined image denoising in spatial and frequency domains using web images. In: Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition. pp. 2933–2940 (June 2014)
19. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Deep edge guided recurrent residual learning for image super-resolution. In: Arxiv, 1608.03981 (2016)