Real Photographs Denoising with Noise Domain Adaptation and Attentive Generative Adversarial Network

Kai Lin[1,2] Thomas H. Li[1,3] Shan Liu[4] Ge Li ✉[1,2]
[1]School of Electronic and Computer Engineering, Peking University [2]Peng Cheng Laboratory
[3]Advanced Institute of Information Technology, Peking University [4]Tencent America
geli@ece.pku.edu.cn

Abstract

Nowadays, deep convolutional neural networks(CNNs) based methods have achieved favorable performance in synthetic noisy image denoising, but they are very limited in real photographs denoising since it's hard to obtain ground truth clean image to generate paired training data. Besides, the existing training datasets for real photographs denoising are too small. To solve this problem, we construct a new dataset and obtain corresponding ground truth by averaging, and then extend them through noise domain adaptation. Furthermore, we propose a attentive generative network by injecting visual attention into the generative network. During the training, visual attention map learns noise regions. The generative network will pay more attention to noise regions, which contributes to balancing between noise removal and texture preservation. Extensive experiments show that our method outperforms several state-of-the-art methods quantitatively and qualitatively.

1. Introduction

Image denoising is one of the most basic tasks in image processing. Owing to images are easily corrupted by noise in the process of acquisition and transmission, they cannot be used directly for high-level tasks such as object recognition, remote sensing imaging and medical imaging ..., etc. In most cases, the distribution of noise in real-world photographs is uncertain and is much different from Gaussian distribution due to factors such as in-camera processing pipeline and environment. In this paper, we focus on blind denoising of real photographs.

Many image prior based methods have been proposed for removing image noise of a certain type and promising results have been achieved(e.g., BM3D [4], LSSC [9] and NCSR [5]). However, on one hand, the image priors are designed mostly by human knowledge, which would fail to capture full-extent features of images. On the other hand, real-world noisy images often exhibit noises of multiple types, making the situation more complicated.

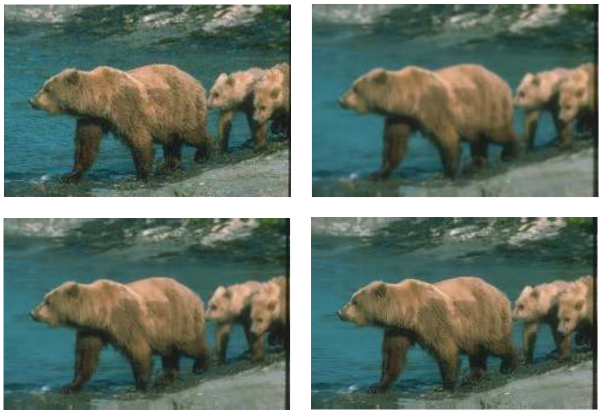

Figure 1. Restoration results of one real noisy image from DND [11]. From left to right and top to bottom: the reconstructed images by BM3D, GCBD, Noise2Noise and our proposed method.

Owing to the great success of deep CNNs, discriminative learning based methods [13, 15, 7, 16] have achieved state-of-the-art performance toward known noise like Gaussian noise. Under the assumption that noise type is known, they can generate synthetic noisy image to form paired training data and train a deep network to obtain remarkable denoising results. However, such a pair of clean and real noisy images are hard to get for real photographs. Actually, what we can get is noise images which are deviated from synthetic Gaussian noise images. These days, some datasets are proposed for real photographs denoising such as DND [11] and Nam [10], but they are proposed for testing and are too small for training. Therefore, such learned based methods perform poorly on real photographs due to lack of paired training data.

The contributions of this paper are as follows: Firstly, we construct a new benchmark dataset which contains real-world noisy images. And corresponding ground truth clean

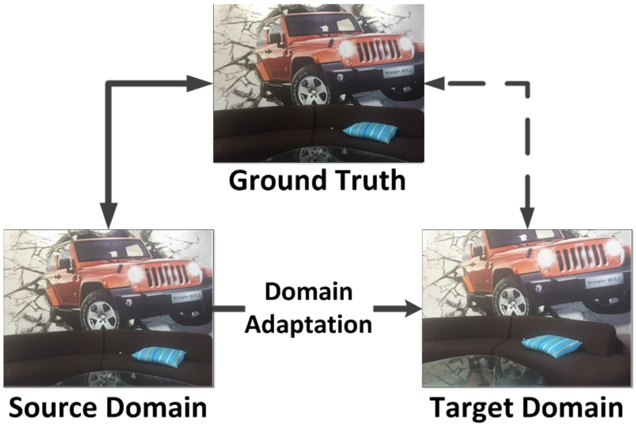

Figure 2. The process of expanding the benchmark with noise domain adaptation. The source domain and target domain represent noise image.

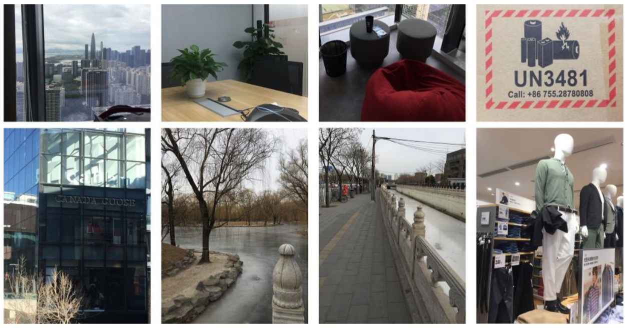

Figure 3. Some example images in our constructed dataset.

images are obtained by averaging and expanding them through noise domain adaptation. Secondly, an attention map is used to guide the generator focusing on noise region, which reduces over-smoothing on noiseless region. Experiments on two real noisy photographs dataset and one synthetic datasets show that our method outperforms several state-of-the-art methods in terms of both quantitative metrics and visual quality.

2. Benchmark

In this section, we introduce the details of capturing a small part of noisy images on different scenes and how to obtain corresponding ground truth images. Then, how the limited benchmark are expanded with noise domain adaptation is discussed.

2.1. The Based Benchmark Construction Process

In the process of constructing basic datasets, we use four cameras from three different camera brands, including Sony (A7R, RX100 IV), Canon (600D), and Nikon (D800). Each scene is captured with 5 different ISO settings, i.e., 800, 1,600, 3,200, 6,400, 12,800. For each ISO setting, we set the shutter speed, the aperture, and the ISO value to ensure that the scenes are in a natural lighting condition. In summary, we capture totally 40 different scenes by using 4 cameras in different camera settings. All images are captured in RAW format. For generating ground truth image, to avoid the problem of slightly different illuminations and human bias as illustrated in [1, 11], we obtain the ground truth by capturing images of the same static scenes for 500 times and average the captured images. (The first image is selected as noise image.) The camera is fixed by a tripod. The data collection is automatically done with shutter release after the button is pressed by a person. Hence, the misalignment problem can be nearly avoided in the acquisition process of 500 images for one scene. Besides, to remove outliers, we remove the images with misalignment and inconsistent luminance. After generating ground truth, we construct training datasets based on pairs of noise image and its ground truth noise-free image. The detailed description of cameras and camera settings is shown in Table 1.

Table 1. Cameras and camera setting used in our new dataset.

Camera	#scenes	Sensor size(mm)	ISO
Sony A7R	10	36 × 24	128-6.4k
Sony RX100 IV	10	13.2 × 8.8	128-6.4k
Canon 600D	12	22.3 × 14.9	128-12.8k
Nikon D800	12	35.9 × 24	128-12.8k

2.2. Expanding the Benchmark with Noise Domain Adaptation

Owing to we having to shot the scene for 500 times to get one noisy image's ground truth, thus creating such sufficient datasets is prohibitively expensive and difficult. The purpose of noise domain adaptation is to expand the training datasets. By mapping a source noisy image to a target noisy image and the two images share the same ground truth noise-free image, the training datasets will be expanded. These two noisy images are shot either at the same scene with different viewpoint or at different scene with similar environment style. The mapping process is shown in Fig. 2. The noise images referenced in section 2.1 are named source domains. More formally, let $X^s = \{x_i^s, y_i^s\}_{i=0}^{N^s}$ represent a dataset including noisy images (source domain) x and their ground truth images y. Firstly, for source domain x_i^s, we shot some noisy images with some shifts and different ISO in the same scene, the dataset consist of these noisy images is named target domain. Let $X^t = \{x_i^t\}_{i=0}^{N^t}$ represent a dateset without ground truth samples from the target domain. Secondly, the generator function $G(x^s) \rightarrow x^a$ in noise domain adaptation model maps a source domain image x^s to an adapted image x^a. Given the generator function G, it is possible to create a dataset $X^a = \{G(x_i^s), y_i^s\}$ of any size. For the process of domain adaptation (DA), we employ the architecture as proposed in [2]. Some example images in the dataset are shown in Fig. 3. The dataset will

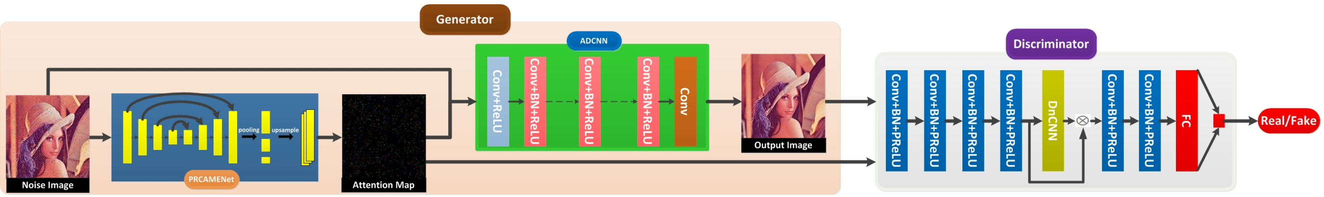

Figure 4. The architecture of our proposed attentive denoising GAN. The generator consists of a pyramid residual connected attention map estimation network and an attentive denoising CNN. The discriminator consists of a series of convolution layers and a DnCNN, and guided by the attention map.

be publicly available after the paper be published.

3. Attentive denoising GAN

A pixel in a noise region is not only influenced by one point but by its whole context. A binary attentive mask M is obtained as follows: if the pixel belongs to noise pixel we set the value to 1 while if the pixel belongs to noise-free pixel we set the value to 0. For synthetic data, we record the noise pixel indexes during generating noise images. For real-world data, we subtract the noise image with its corresponding clean image, and then use a threshold to determine whether a pixel belongs to noise region.

The overall architecture of attentive denoising generative adversarial networks(ADGAN) are shown in Fig. 4. Following the idea of generative adversarial networks [6], The ADGAN includes two parts: the generative and discriminative networks. Given a noisy image, the generative network attempt to produce an image as real as possible to fake the discriminator, while the discriminative network tries to distinguish whether the image is fake or real. Beyond the standard GAN architecture, visual attention computation is incorporated in the generator to accurately localize noisy regions.

3.1. Generative Network

As shown in Fig. 4, the generator consists of two subnetworks: an pyramid residual connected attention map estimation network(PRCAMENet) and an attentive denoising CNN(ADCNN). The purpose of PRCAMENet is to find noise regions that the ADCNN needs to get attention, so that it can avoid over-smoothing over the premise of denoising performance.

Pyramid Residual Connected Attention Map Estimation Network. In order to make use of the features from multiple layers of a CNN, the network is constructed based on a residual connected encoder-decoder structure, where the residual block is used as the basic structure. The reason to use residual block lies in that when the original mapping is more like an identity mapping result, the residual mapping will be much easier to calculate. Besides, owing to the features from different scales are not used to directly estimate the final attention map, the result will lack global information with different scales. To solve this problem, inspired by usefulness of nonlocal context information in segmentation and classification tasks [14, 17], a three-level pyramid pooling module with pooling sizes 1/4, 1/8 and 1/16 is adopted to refine the learned features by considering the global information into the optimization. Then, all three-level features are up-sampling to the original feature size and are concatenated with the original feature before estimating. The feature size at the end of encoding part is 1/16 of the input size. To reconstruct the attention map into the original size, we stack four residual blocks with the refined up-sampling blocks as the decoding module. Besides, the features corresponding to the same dimension are concatenated. The loss function is defined as the mean squared error(MSE) between the output attention map A and the ground truth binary mask M.

Attentive Denoising CNN. The purpose of ADCNN is to balance denoising performance and texture preservation under the guidance of attention map. The input to ADCNN is the concatenation of the input noisy image and the attention map from PRCAENet. Given ADCNN with depth D, the layers are divided into three types as shown in Fig. 4. For the first layer, 32 filters of size $3 \times 3 \times 2c$ Conv are applied to generate 32 feature maps and followed by ReLU. Here c indicates the number of image channels. Since the noisy image and the attention map are concatenated in the $3rd$ channel, we denote $2c$ here. From the $2nd$ layer to the $(D-1)th$ layer, 32 filters of size $3\times3\times32$ Conv are applied, which is followed by batch normalization(BN) and ReLU. For the last layer, c filters of size $3 \times 3 \times 32$ are used to reconstruct the output. D is set to 15. For ADCNN, we use L_2 loss. Overall, the loss of our generator can be written as:

$$\begin{aligned} L_G(F,A,M,T) = &log(1-D(F)) + L_{MSE}(A,M) \\ &+ L_{MSE}(F,T) \end{aligned} \tag{1}$$

where F is the output of the whole generator, A is the output of the PRCAMENet, M is corresponding ground truth binary mask, T is the ground truth image and D is the discriminator.

3.2. Attentive Discriminative Network

To differentiate fake images from real ones more accurately, we adopt global and local image-content consistency by constructing an attentive discriminator. Specifically, we use $(Conv + BN + ReLU)$ to extract features from the interior layers of the discriminator, and feed them into a simple CNN-based network (DnCNN [15]). The loss function L_{map} is calculated between DnCNN's output and the attention map generated by our PRCAMENet. Moreover, before feeding the output of DnCNN into the next layers of the discriminator, they are multiply with the input of DnCNN to guide the discriminator to focus more on the regions signed by the attention map. Finally, a full connected layer is used to decide whether the image is real or fake. The whole loss function of the discriminator is expressed as follows:

$$\begin{aligned} L_D(F, R, A) = & -log(D(R)) - log(1 - D(F)) \\ & + \alpha L_{map}(F, R, A) \end{aligned} \tag{2}$$

where

$$L_{map} = L_{MSE}(D_{att}(F), A) + L_{MSE}(D_{att}(R), \mathbf{0}) \tag{3}$$

where D_{att} represents the process of producing a 2D attention map by the discriminator. R is image drawn from real images and F is the output of the generator. $\mathbf{0}$ represents a 2D map containing only 0 values. α is set to 0.01.

4. Experimental Results

In this section, we demonstrate the denoising performance by conducting experiments on two real-world datasets and two synthetic datasets.

4.1. Experimental Data

Training Data. Our constructed dataset involves 40 different scenes captured by 4 different cameras of different settings. 5k paired training images are generated using noise domain adaptation. Since the images are of large size (3000×3000), we crop some regions from these images of size 512×512.

Testing Data. In real-world data experiments, DND [11] and Nam [10] are used. In the experiments on synthetic data, the evaluations are conducted on BSD68 [12] and Set12 [5].

4.2. Parameter Setting and Network Training

During training, ADAM is used as the optimization algorithm with learning rate of 10^{-3} for both generator and discriminator and batch size of 64. We train the network for 100 epochs.

4.3. Evaluation with Synthetic and Real-world Noise

We evaluate the performance of our methods against several state-of-the-art noise removal methods: BM3D [4], GCBD [3] and Noise2Noise [8].

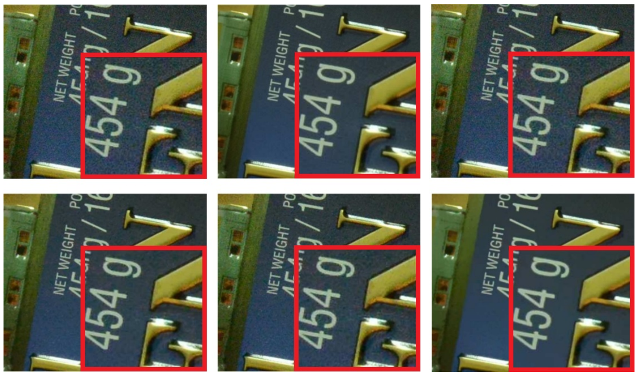

Figure 5. Restoration results of one real noisy image from Nam[10]. From left to right and top to bottom: the noise image, the ground truth, the reconstructed images by BM3D, GCBD, Noise2Noise and our proposed method.

Table 2. The quantitative results from the DND benchmark [11].

Method	BM3D	GCBD	Noise2Noise	Ours
PSNR	34.51	37.72	37.57	**38.13**
SSIM	0.8507	0.9408	0.9360	**0.9580**

Table 3. The quantitative results from the Nam benchmark [10].

Method	BM3D	GCBD	Noise2Noise	Ours
PSNR	39.84	40.02	40.41	**41.38**
SSIM	0.9657	0.9687	0.9731	**0.9810**

Table 4. The PSNR and SSIM results of all the compared methods on BSD68 and Set12 under noise level $\sigma = 20$ on Gaussian noise.

Method	BM3D	GCBD	Noise2Noise	Ours
BSD68	32.15(0.9013)	32.78(0.9088)	32.85(0.9093)	**33.01(0.9210)**
Set12	31.04(0.8651)	31.87(0.8980)	31.90(0.8989)	**32.03(0.9005)**

Real-World Noise. On DND [11], The competing methods are evaluated in sRGB space. PSNR and SSIM results are shown in Table 2. ADGAN outperforms the other methods with a large margin. Fig. 1 shows the visually comparison results. BM3D and GCBD fail to remove all noise, and Noise2Noise suffers from the problem of over-smoothing. Compared with other methods, ADGAN does well in balancing noise removal and texture preservation. On Nam [10], the quantitative and qualitative comparisons are shown in Table 3 and Fig. 5. ADGAN outperforms the other methods by about 0.5 dB at least. For ablation study, the performance of our network get worse when attention map is removed.

Synthetic Noise. BSD68 and 12 commonly used images are evaluated under a noise level $\sigma = 20$ on Gaussian noise. The comparison results are shown in Table 4. Obviously, our method achieves the best denoising performance on synthetic noise.

5. Conclusion

In this paper, we present an attentive based GAN for real-world photographs denoising. Firstly, we capture pairs of images under wide scene and parameters setting and obtain corresponding ground truth by averaging. Then we extend them by using noise domain adaptation to guarantee the datasets are adequate for the network training. Secondly, attentive denoising GAN is proposed for real photographs denoising, where the generator produces the attention map via a PRCAMENet and applies the map along with the input image to generate a noise-free image through the AD-CNN. Our discriminator then assesses the validity of the generated output locally and globally. Extensive experiments demonstrate that our method outperforms several state-of-the-art methods quantitatively and qualitatively. Furthermore, our network is effective in balancing noise removal and structure preservation.

Acknowledgement. This work was supported by Shenzhen Key Laboratory for Intelligent Multimedia and Virtual Reality (No.ZDSYS201703031405467), Shenzhen Municipal Science and Technology Program (No. JCYJ20170818141146428), National Engineering Laboratory for Video Technology - Shenzhen Division and Peng Cheng Laboratory.

References

[1] Josue Anaya and Adrian Barbu. Renoir a dataset for real low-light image noise reduction . *Journal of Visual Communication Image Representation*, 51:144–154, 2014.

[2] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. pages 95–104, 2016.

[3] Jingwen Chen, Jiawei Chen, Hongyang Chao, and Ming Yang. Image blind denoising with generative adversarial network based noise modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3155–3164, 2018.

[4] K. . Dabov, A. . Foi, V. . Katkovnik, and K. . Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16:2080–2095, 2007.

[5] Weisheng Dong, Lei Zhang, Guangming Shi, and Xin Li. Nonlocally centralized sparse representation for image restoration. *IEEE Transactions on Image Processing*, 22:1618–1628, 2013.

[6] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *International Conference on Neural Information Processing Systems*, pages 2672–2680, 2014.

[7] Stamatios Lefkimmiatis. Non-local color image denoising with convolutional neural networks. pages 5882–5891, 2017.

[8] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2noise: Learning image restoration without clean data. 2018.

[9] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Non-local sparse models for image restoration. In *IEEE International Conference on Computer Vision*, pages 2272–2279, 2010.

[10] Seonghyeon Nam, Youngbae Hwang, Yasuyuki Matsushita, and Seon Joo Kim. A holistic approach to cross-channel image noise modeling and its application to image denoising. In *Computer Vision and Pattern Recognition*, pages 1683–1691, 2016.

[11] Tobias Pltz and Stefan Roth. Benchmarking denoising algorithms with real photographs. In *Computer Vision and Pattern Recognition*, pages 2750–2759, 2017.

[12] Stefan Roth and Michael J. Black. Fields of experts. *International Journal of Computer Vision*, 82:205, 2009.

[13] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11:3371–3408, 2010.

[14] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. 2018.

[15] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26:3142–3155, 2016.

[16] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2808–2817, 2017.

[17] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. pages 6230–6239, 2016.