Reconstructing the Noise Manifold for Image Denoising - Supplementary Material

Anonymous ECCV submission

Paper ID 910

1 Architecture details

In our '*encoder-decoder*' architecture (same for both subnets) 11 layers are used with an latent space of dimensions $MB \times 2 \times 2 \times 1024$, where MB stands for mini-batch size. The architecture is given in Table 1. As it can be seen in this table, an extra layer with linear activation (colored in yellow) is added to the original decoder in order to get the final image output. In the case of RGB image denoising the output image has 3 channels, whereas RAW image denoising outputs 4 channels (R-G-G-B).

Layer	Filter Size	Stride
Conv. 1	$3 \times 3 \times 32$	1
Conv. 2	$3 \times 3 \times 64$	2
Conv. 3	$3 \times 3 \times 128$	2
Conv. 4	$3 \times 3 \times 256$	2
Conv. 5	$3 \times 3 \times 512$	2
Conv. 6	$3 \times 3 \times 1024$	2

(a) Encoder

Layer	Filter Size	Stride
F-Conv. 1	$3 \times 3 \times 512$	2
F-Conv. 2	$3 \times 3 \times 256$	2
F-Conv. 3	$3 \times 3 \times 128$	2
F-Conv. 4	$3 \times 3 \times 64$	2
F-Conv. 5	$3 \times 3 \times 32$	2
Extra last Conv. layer	$3 \times 3 \times \{3, 4\}$	1

(b) Decoder

Table 1: Details of the '*encoder-decoder*' architecture employed in our work. 'Filter size' denotes the size of the filters for the convolutions; the last number denotes the number of output filters. Conv denotes a convolutional layer. F-Conv denotes a transposed convolutional layer with fractional-stride. An extra Conv layer with linear activation (colored in yellow) is added to the original decoder in order to get the final image output.

2 Motivation

Based on the bibliography [10], the primary sources of noise are shot noise, a Poisson process with variance equal to the signal level, and read noise, an approximately Gaussian process caused by a variety of sensor readout effects. The noise is spatially variant (non-Gaussian); hence, the assumption that noise is spatially invariant, employed by many algorithms does not hold for real image

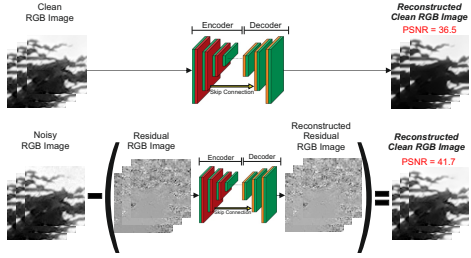


Fig. 1: **Motivation of our method** (picture taken from the main paper): By characterizing directly the image spatially variant noise, the reconstruction of the clean image is much more accurate. Instead of constraining the output of a generator to span the target space, is better to constrain it to remove from the noisy image only that information which spans the manifold of the residual image.

noise. These effects are well-modeled by a signal-dependent Gaussian distribution [10]. Based on that, the variance of noise is proportional to image intensity which means that the noise in real images is structured. Thus, a low dimensional manifold of the image noise can be created. This makes sense because if a low dimensional manifold of real image noise does not exist this means that the same holds for the real image as well, something not true based on bibliography ((cGAN) [13]).

As it is mentioned in the main paper, by characterizing directly the image spatially variant noise the reconstruction of the clean image is much more accurate. In this way of thinking, instead of constraining the output of a generator to span the target space, is better to constrain the generator to remove from the noisy image only that information which spans the manifold of the residual image. To justify our motivation (Fig. 1), we trained a standard '*encoder-decoder*' type architecture with skip connections, same like *Rec subnet*, to:

- directly reconstruct the clean images.
- indirectly reconstruct the clean images by removing from the noisy images the corresponding reconstructed image noise signal.

To do so, we have collected a dataset that consists of four classes of 12MP images:

- Buildings - 1010 images
- Foliage - 841 images
- Text - 838 images
- Misc - 815 images

In purpose, these data were collected under very low ISO conditions by using a smartphone camera. The reason for choosing very low ISO values is to collect clean images which could be used as clean ground truth images. On top of that, we have removed some (very) light residual image noise to further enhance the image quality of the clean ground truth data. The noise model parameters for

the used camera sensor were available (like the noise model described in Section 3.1 in the main paper), thus we were able to generate a very big amount of synthetic noisy data by adding camera sensor-based signal-dependent noise to the clean images (one clean image could provide more than one noisy versions because more than one different ISO values could be used to add noise to the same clean image). To do so, we follow the same pipeline used in [8]. As a result, each class in the dataset had ISO values in the range [40, 64000]. The ISO values were random but uniformly sampled across each class. Some visual examples of the real clean images are depicted in Fig. 2. Afterwards, the synthetic created paired images were combined with real paired images to verify the generalization ability of this experiment. To get real paired images we employed the SSID [1] and RENOIR [2] datasets. In total, 5.5 million training patches of size 128×128 were extracted, while 40K patches (10K images from each class) with random ISO values but uniformly sampled across each class were used as a validation set.

In general, the skip connections enable deeper layers to capture more abstract representations without the need of memorizing all the information. In our '*encoder-decoder*' type architecture, like in *Rec subnet*, only the lower-level representations are propagated directly to the decoder through a (Unet style) shortcut.

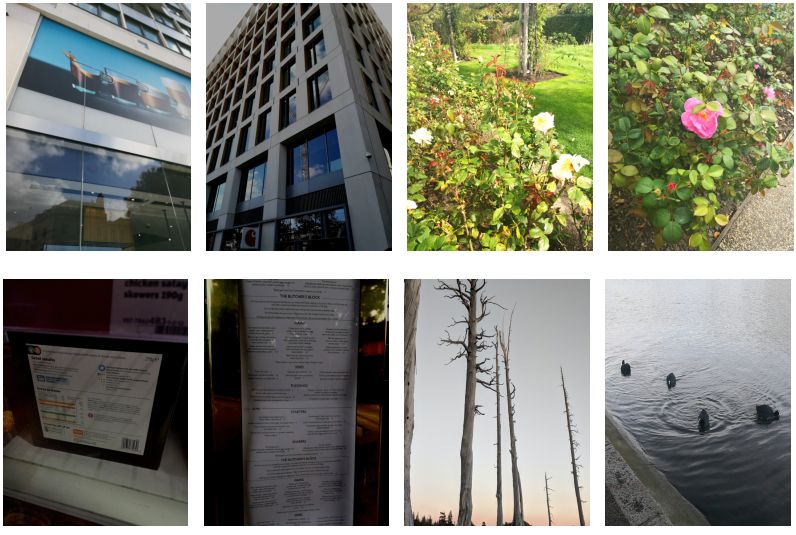


Fig. 2: Examples of images from the dataset used to justify the motivation behind the proposed idea.

The total loss function consists of the content loss and the Deconv loss [5]. The ℓ_1 loss between the ground-truth image and the output of the generator was used in the case of direct image reconstruction as content loss, while the

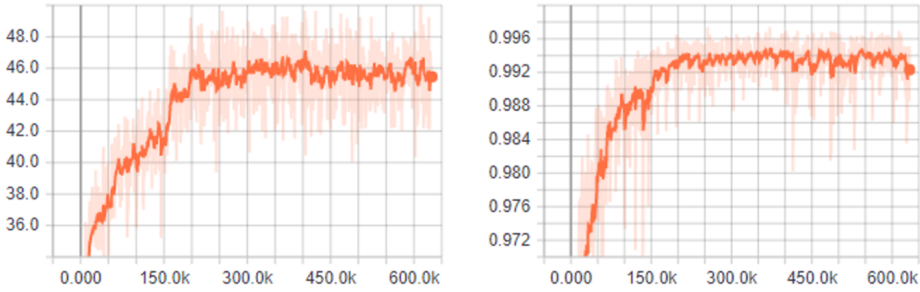


Fig. 3: **The reconstruction performance in terms of PSNR and SSIM metrics for the '*direct*' way of clean image reconstruction during the training procedure. Left:** PSNR values and **Right:** SSIM values.

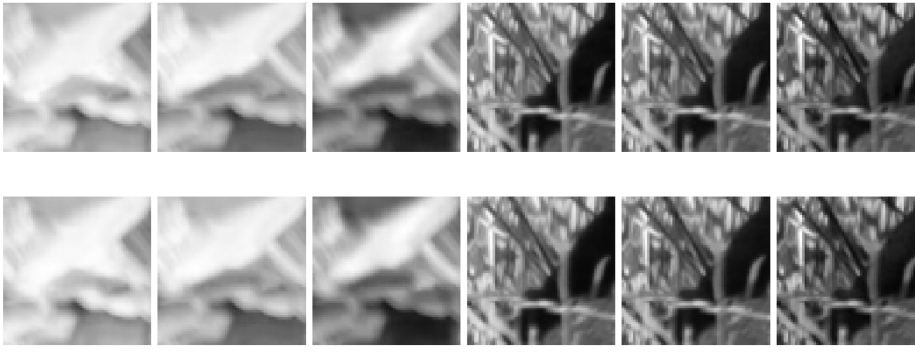


Fig. 4: **Visual example for the '*direct*' reconstruction of a clean validation image (example in RGB domain). First row:** Ground truth clean image patch channels, **Second row:** Corresponding reconstructed clean image patch channels.

ℓ_2 loss used in the other case. Different content loss functions used because we reported in this study the best results we got for each case. Except the different content loss functions, during the training the same setup was used in both cases. Adam [11] was used as the optimizer with default parameters; the learning rate is initially set to 10^{-3} and then halved after 10^5 iterations; ReLU activation used; the network ran for 45 epochs. Regarding the training procedure, Fig. 3 shows the reconstruction performance in terms of PSNR and SSIM metrics for the direct way of clean image reconstruction, while Fig. 5 shows the reconstruction performance in terms of PSNR and SSIM metrics for the indirect way of clean image reconstruction. Regarding the validation procedure, Fig. 4 shows a visual example for the direct way of clean image reconstruction, while Fig. 6 shows a visual example for the indirect way of clean image reconstruction. Based on the validation dataset, the averaged difference in PSNR was 7.5db. Based on these results, the indirect way of clean image reconstruction is by far better than the

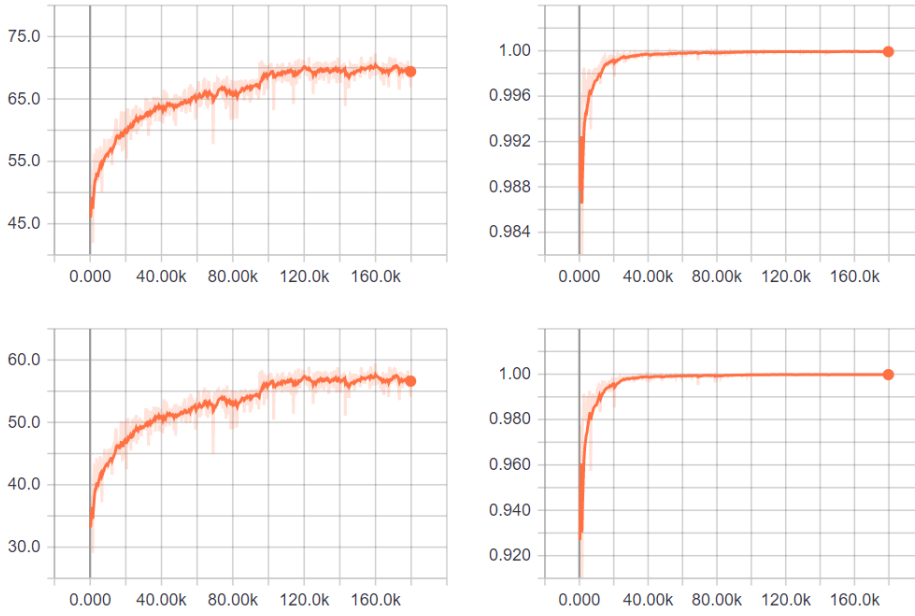


Fig. 5: **The reconstruction performance in terms of PSNR and SSIM metrics for the '*indirect*' way of clean image reconstruction during the training procedure. Top-Left:** PSNR values for the clean image reconstruction, **Top-Right:** SSIM values for the clean image reconstruction, **Bottom-Left:** PSNR values for the image noise reconstruction and **Bottom-Right:** SSIM values for the image noise reconstruction.

direct one (especially when a non very complex '*encoder-decoder*' architecture is employed (Table 1)). Thus, our motivation is experimentally justified. In short, by using the residual learning to directly characterize the image noise makes the task of image denoising much easier.

3 Experiments

3.1 Type of denoising

For all benchmarks, a blind and a non-blind version of our method had been tested based on the info that $\boldsymbol{c}$ represents. The blind version uses no extra conditional information along with the noisy input image (empty $\boldsymbol{c}$). As described in the main paper (Section 3.3), in the non-blind version, $\boldsymbol{c}$ could contain information regarding the camera noise model parameters (signal-dependent noise variance) and/or the camera id.

In the case of non-blind denoising for the all benchmarks (Darmstadt Noise Dataset (DnD) [15], Nam Dataset [14] and Smartphone Image Denoising Dataset (SIDD) [1]) the camera id was provided (4 different standard consumer cameras

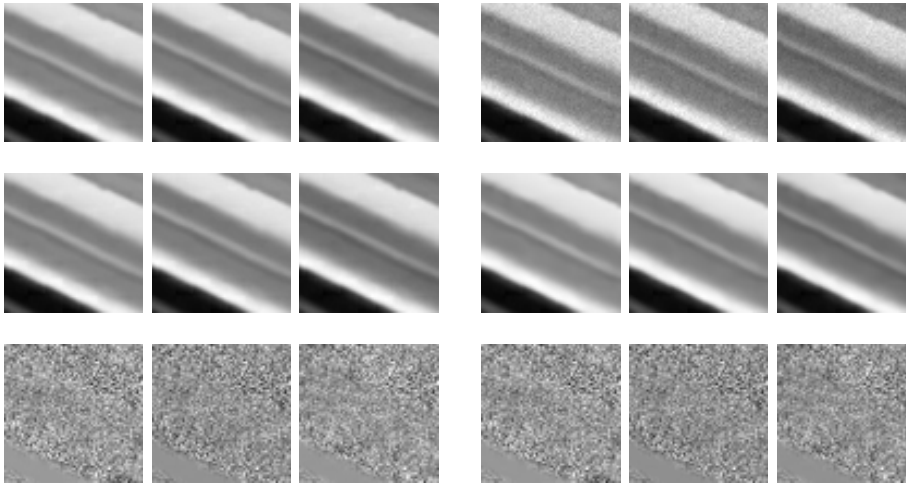
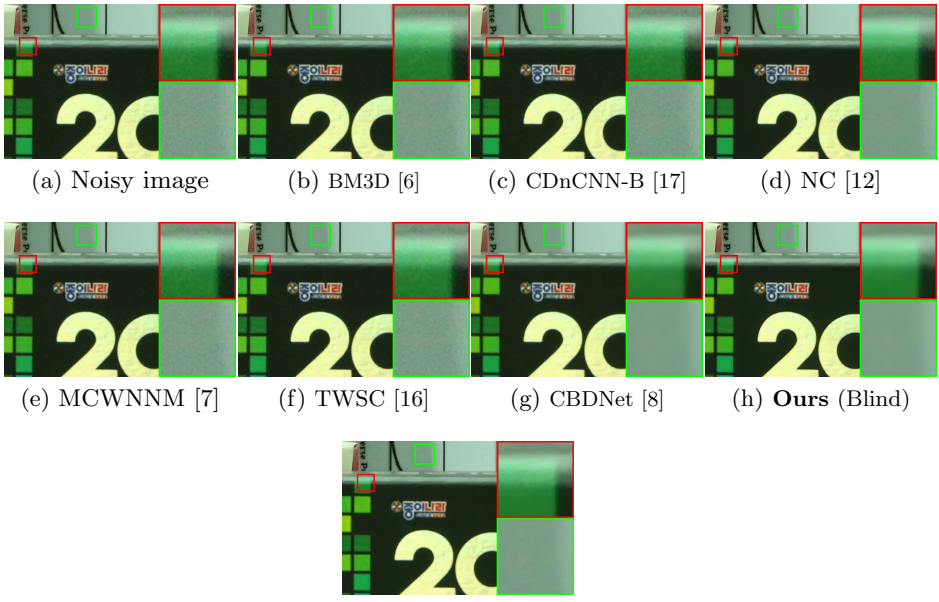


Fig. 6: **Visual example for the '*indirect*' reconstruction of a clean validation image (example in RGB domain). First row:** The three most left images depict the clean image channels, while the three most right images depict the corresponding noisy image channels, **Second row:** The three most left images depict the reconstructed clean image channels, while the three most right images depict the corresponding denoised image channels, **Third row:** The three most left images depict the reconstructed image noise channels, while the three most right images depict the corresponding ground truth image noise channels.



(a) Noisy image (b) BM3D [6] (c) CDnCNN-B [17] (d) NC [12]

(e) MCWNNM [7] (f) TWSC [16] (g) CBDNet [8] (h) **Ours** (Blind)

(g) **Ours** (Non Blind)

Fig. 7: A real noisy example from Nam dataset [14] for comparison of our method against the state-of-the-art algorithms. Results of the proposed method shown when ResNet [9] used as backbone network.

used for DnD, 3 for Nam and 5 for SIDD) while the noise model parameters for each camera sensor were provided as well.

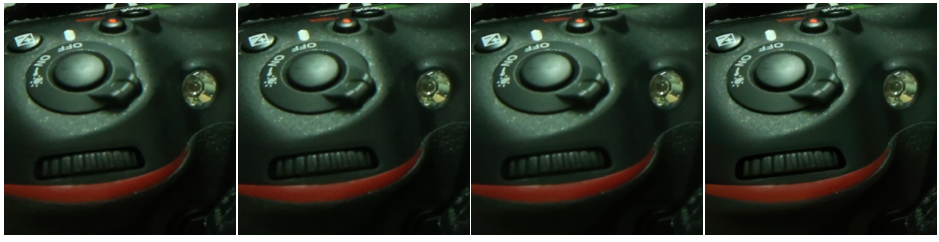
3.2 Visual results

Fig. 7 and 8 show some visual image denoising comparisons based on noisy testing samples from Nam Dataset [14]. Fig. 9 shows a visual image denoising comparison based on a real noisy test sample from DND dataset [15]. Fig. 10, 11, 12, 13, 14 and 15 show some image denoising results by using the proposed method given as input the noisy validation images from our collected dataset used to verified our motivation in Section 2. Based on all our experiments, the proposed idea restores better the true colors which are closer to the original pixel values than the competing methods. Also, by directly characterizing the image noise, our method avoids in great degree the image over-smoothing.

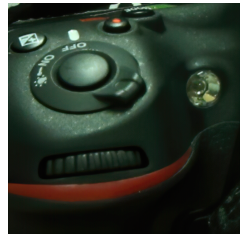
References

1. Abdelhamed, A., Lin, S., Brown, M.S.: A high-quality denoising dataset for smartphone cameras. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)

(a) Noisy image (b) RoCGAN [4] (c) BM3D [6] (d) DnCNN-B [17]



(e) CBDNet-JPEG [8] (f) RIDNet [3] (g) NC [12] (h) **Ours** (Blind)



(g) **Ours** (Non Blind)

Fig. 8: A real noisy example (the non-cropped version of the Nam image reported in the main paper) from Nam dataset [14] for comparison of our method against the state-of-the-art algorithms. Results of the proposed method shown when ResNet [9] used as backbone network.

2. Anaya, J., Barbu, A.: Renoir - a dataset for real low-light noise image reduction. Journal of Visual Communication and Image Representation **51**, 144–154 (2018)
3. Anwar, S., Barnes, N.: Real image denoising with feature attention. In: IEEE International Conference on Computer Vision (ICCV) (2019)
4. Chrysos, G.G., Kossaifi, J., Zafeiriou, S.: Robust conditional generative adversarial networks. In: International Conference on Learning Representations (ICLR) (2019)
5. Cogswell, M., Ahmed, F., Girshick, R., Zitnick, L., Batra, D.: Reducing overfitting in deep networks by decorrelating representations. In: International Conference on Learning Representations (ICLR) (2016)
6. 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 (2007)
7. Gu, S., Zhang, L., Zuo, W., Feng, X.: Weighted nuclear norm minimization with application to image denoising. IEEE Conference on Computer Vision and Pattern

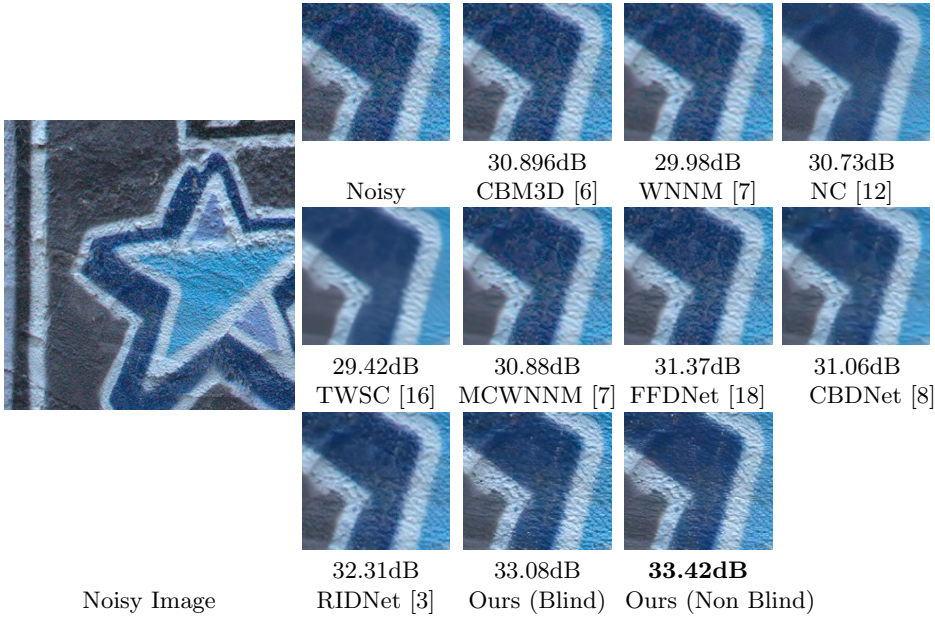
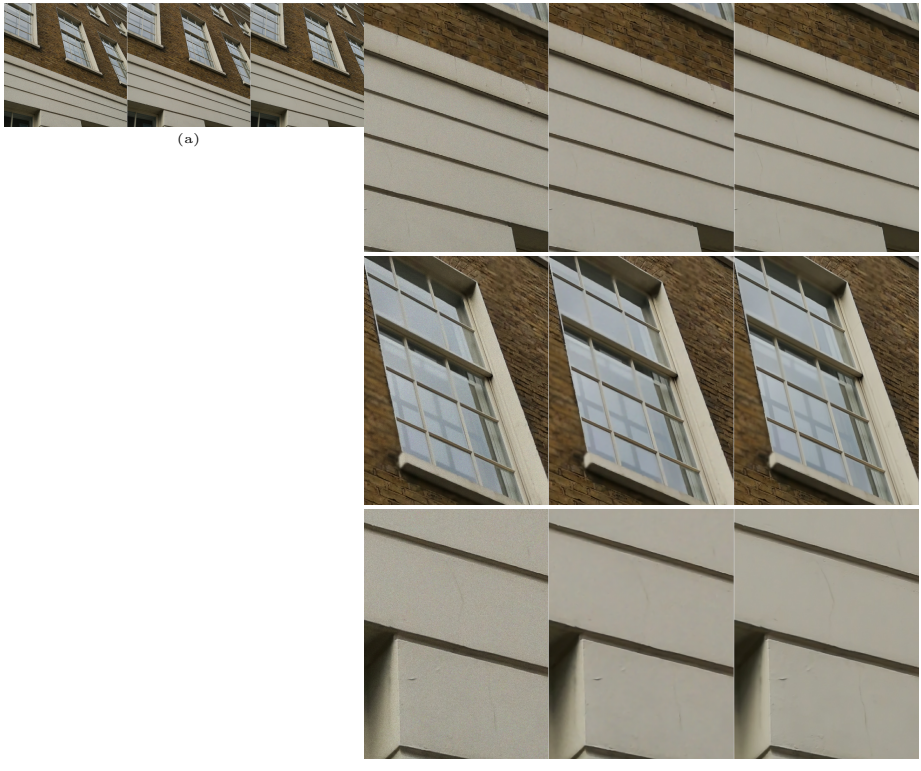


Fig. 9: A real noisy example from DND dataset [15] for comparison of our method against the state-of-the-art algorithms. Results of the proposed method shown when ResNet [9] used as backbone network.

Recognition pp. 2862–2869 (2014)

8. Guo, S., Yan, Z., Zhang, K., Zuo, W., Zhang, L.: Toward convolutional blind denoising of real photographs. 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
10. Healey, G., Kondepudy, R.: Radiometric ccd camera calibration and noise estimation. IEEE Trans. Pattern Anal. Mach. Intell. **16**, 267–276 (1994)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
12. Lebrun, M., Colom, M., Morel, J.M.: The noise clinic: a blind image denoising algorithm. Image Processing On Line (IPOL) **5**, 1–54 (2015)
13. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
14. Nam, S., Hwang, Y., Matsushita, Y., Kim, S.J.: A holistic approach to cross-channel image noise modeling and its application to image denoising. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1683–1691 (2016)
15. Plotz, T., Roth, S.: Benchmarking denoising algorithms with real photographs. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2750–2759 (2017)



(a)

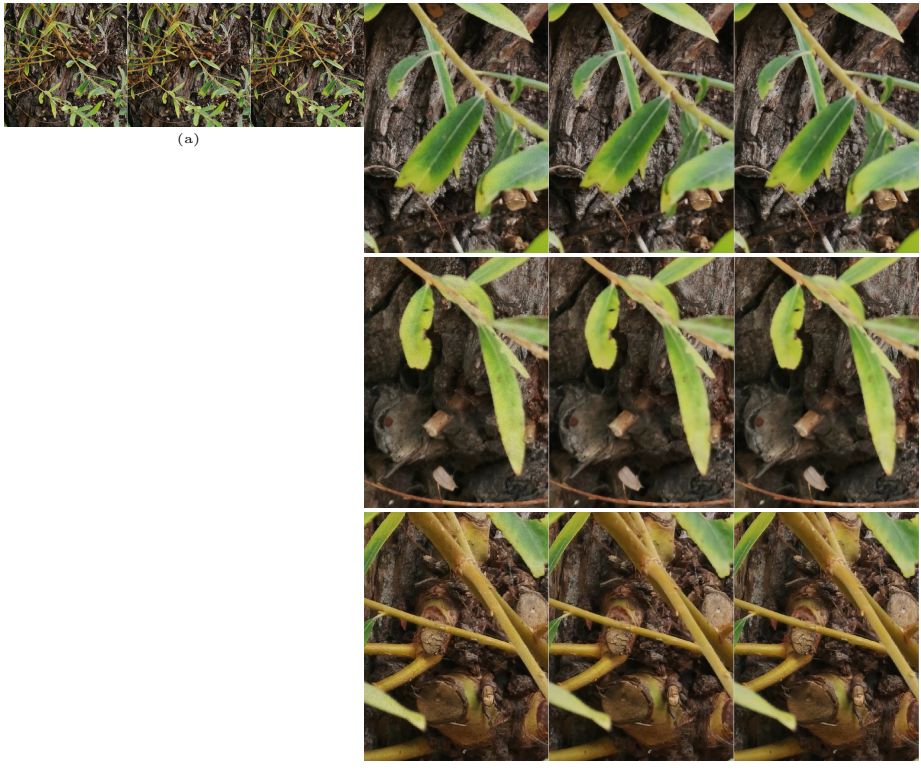
(b)

Fig. 10: **Example of image denoising by using the proposed method and an image (from '*Buildings*' class) from our collected dataset.** (a) **Left:** Noisy image, **Middle:** Denoised image (PSNR: 43.76, SSIM: 0.9869), **Right:** Clean image (Ground truth), (b) **Left:** Noisy image patch, **Middle:** Denoised image patch, **Right:** Clean image patch (Ground truth).

(a)

(b)

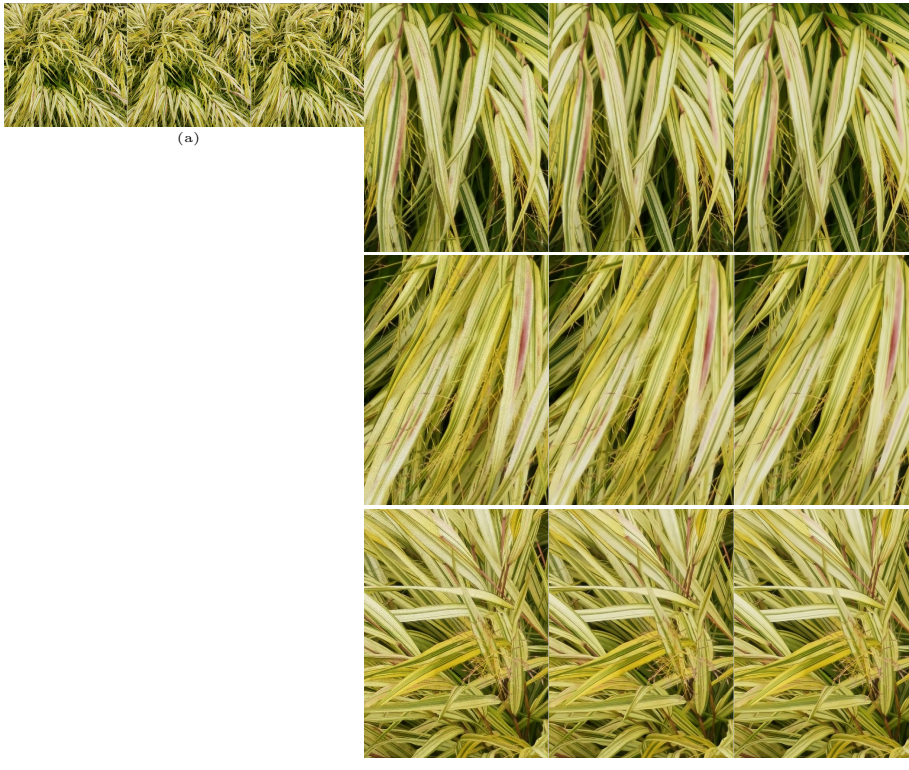
Fig. 11: **Example of image denoising by using the proposed method and an image (from '*Buildings*' class) from our collected dataset.** (a) **Left:** Noisy image, **Middle:** Denoised image (PSNR: 40.82, SSIM: 0.9809), **Right:** Clean image (Ground truth), (b) **Left:** Noisy image patch, **Middle:** Denoised image patch, **Right:** Clean image patch (Ground truth).



(a)

(b)

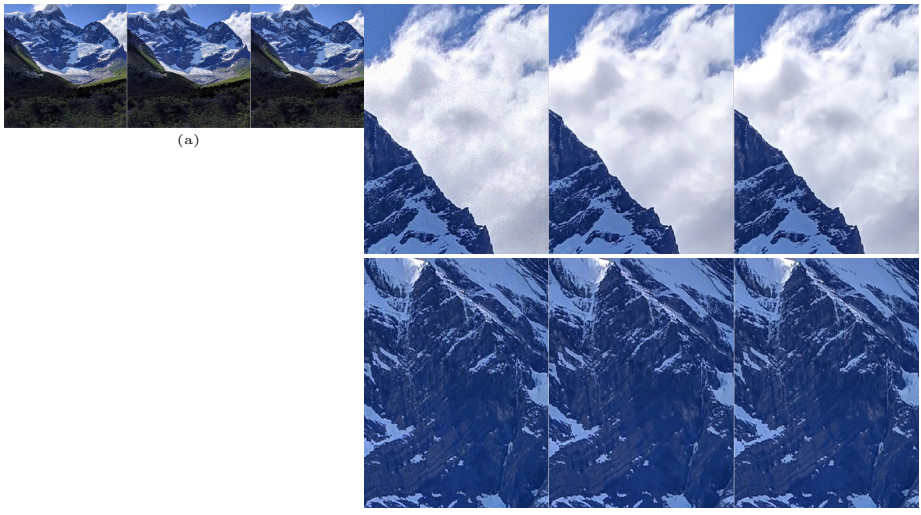
Fig. 12: **Example of image denoising by using the proposed method and an image (from '*Foliage*' class) from our collected dataset.** (a) **Left:** Noisy image, **Middle:** Denoised image (PSNR: 39.52, SSIM: 0.9776), **Right:** Clean image (Ground truth), (b) **Left:** Noisy image patch, **Middle:** Denoised image patch, **Right:** Clean image patch (Ground truth).



(a)

(b)

Fig. 13: **Example of image denoising by using the proposed method and an image (from '*Foliage*' class) from our collected dataset.** (a) **Left:** Noisy image, **Middle:** Denoised image (PSNR: 35.36, SSIM: 0.9769), **Right:** Clean image (Groundtruth), (b) **Left:** Noisy image patch, **Middle:** Denoised image patch, **Right:** Clean image patch (Groundtruth).



(a)

(b)

Fig. 14: **Example of image denoising by using the proposed method and an image (from '*Misc*' class) from our collected dataset.** (a) **Left:** Noisy image, **Middle:** Denoised image (PSNR: 38.71, SSIM: 0.9770), **Right:** Clean image (Groundtruth), (b) **Left:** Noisy image patch, **Middle:** Denoised image patch, **Right:** Clean image patch (Groundtruth).

16. Xu, J., Zhang, L., Zhang, D.: A trilateral weighted sparse coding scheme for real-world image denoising. In: European Conference on Computer Vision (ECCV). pp. 21–38 (2018)
17. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. IEEE Transactions on Image Processing **26**(7), 3142–3155 (2017)
18. Zhang, K., Zuo, W., Zhang, L.: Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. IEEE Transactions on Image Processing (TIP) **27**(9), 4608–4622 (2018)

Fig. 15: **Example of image denoising by using the proposed method and an image (from '*Text*' class) from our collected dataset. Top:** Noisy image, **Middle:** Denoised image (PSNR: 47.27, SSIM: 0.9930), **Bottom:** Clean image (Groundtruth).