Supplementary Materials

Anonymous ECCV submission

Paper ID 1848

A More details and discussions of our proposed model

A.1 Overview

As mentioned in Sec. 1 in the original paper, for any burst denoising solution to be practical, it needs to address a few major challenges. First, it needs to be efficient, especially when considering resource-constrained devices. Second, it needs to be flexible and scalable, being able to handle arbitrary length of burst frame input, which is not the case in KPN [1]. Third, it needs to not only pursue objective quality, but also enhance perceptual quality and balance the trade-off between them as indicated in the Perception-Distortion Tradeoff work [2].

In the following sections, we first review more recent related work on single image denoising in Sec. A.2 to put our burst denoising model in a better context. Then we introduce more background and detail of the wavelet transforms operation in our proposed model in Sec. A.3. In Sec. A.4 we provide more details on the temporal max-pooling based feature fusion for better comparing the two pseudo-3D mechanisms which we investigated for the burst denoising task. More details of the camera simulation pipeline is introduced in Sec. A.5 and failure cases are demonstrated in Sec. A.5.

A.2 More related work on single-image denoising.

Image denoising is a long-established low-level computer vision task. Many traditional methods have been proposed to take advantage of the specific statistics of natural images for reducing noise. Anisotropic diffusion [3] and bilateral filters [4] exploit local similarities to selectively smooth pixels locally. Total variation methods [5] use analytical priors for natural images. Domain transform methods, notably using the Wavelet transform [6,7] reduce the spatial dependencies to simplify filtering. Non-local patch methods [8] exploit short-range and long-range similarities. Another approach is to model the image patches as sparse linear combinations of elements of a learned dictionary [9]. Among these traditional methods, BM3D [10], which both selects pixels with block matching and uses transform domain filtering, is generally considered the current state of the art in image denoising.

Meanwhile, due to the popularity of convolutional neural networks (CNNs), image denoising algorithms [11,12,13,14,15] have achieved a significant boost in performance. Notable denoising neural networks, DnCNN [11], and IrCNN [13]

predict the residual noise present in the image instead of the denoised image, while the ground truth noise is utilized for loss computation instead of the original clean image. Recently, many algorithms have focused on blind denoising for images with real noise. The algorithms [13,11] benefited from the modeling capacity of CNNs and have demonstrated the ability to learn a single-blind denoising model. However, the denoising performance is limited, and the results are not as good when the network is trained for a known noise [11,16]. They are even less satisfactory on real photographs [17].

Very recently, CBDNet [18] was proposed as a blind denoising model for real photographs. CBDNet is composed of two sub-networks: noise estimation and non-blind denoising, and also incorporates multiple losses, Furthermore, [18,16] may require manual tuning to improve results. Also, our model was originally designed to facilitate higher resolution feature maps for preserving more details in final denoised output.

A.3 Wavelet transforms for feature decomposition

As discussed in Sec. 3.2 in the original paper, we explicitly decompose the convolutional features to the high-frequency sub-bands and low-frequency subband features with wavelet pooling, and fuse the features later with the inverse wavelet unpooling process. Wavelets has been widely used in signal processing tasks to compactly represent signals while maintaining important information such as edges.

As a result, the reconstruction performance of encoder-decoder type of networks can be improved with minimal noise amplification by utilizing proper wavelet transformation mechanisms. For the Haar wavelet we adopted, the low-pass subband filter is equivalent to the average pooling to capture the structural information of convolutional features, and the high-pass counterpart will capture the representation corresponding to the local details and perceptual quality.

According to our ablation study in Table 1 of the paper, even though the wavelet transform does not contribute as significantly as high-resolution features towards PSNR improvements, we consider it important to the restoration of high frequency contents. Wavelets allow us to explicitly decompose the noisy data into frequency-specific feature channels and then temporally process different frequency channels at various scales adaptively. This reduces over-smoothing of high frequency content, which is not be reflected well in PSNR measurements. Moreover, wavelet, with their reversible transform, have been very successful in denoising before the deep learning era and have unambiguous decoding capability w.r.t. max-pooling + transposed convolutions.

A.4 Temporal max-pooling for pseudo-3D feature fusion

As described in Sec. 3.3 in the original paper, temporal max-pooling we adopted as one of the two pseudo-3D feature fusion mechanisms is straightforward and order-invariant to all the input frames. The corresponding network architecture for burst denoising is illustrated in Fig. 1. Specifically, each of the input frame is

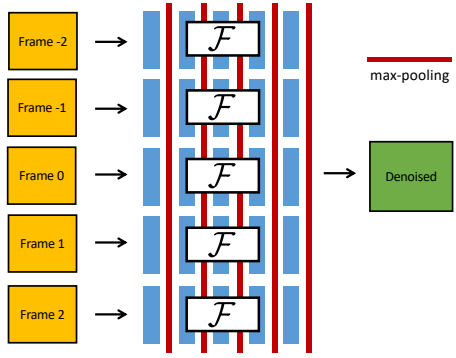

Fig. 1: Overview of the temporal max-pooling architecture in our experiments. Each input frame is processed by a copy of the same 2D backbone network with tied weights, but the information is repeatedly exchanged between the copies. Specifically, the maximum value of each activation between all the tracks are computed, then these *global representations* are concatenated back with the per-frame local features for further convolution. At the last layer, all the tracks are collapsed by a final max-pooling layer and jointly predict one clean image. Note that this mechanism is order-invariant in contrast with the temporal feature shifting operation.

processed by a tied copy of the same 2D backbone network (either the baseline or our proposed 2D wavelet transform model), and the outputs are max-pooled across all frames to generate the global representation, which are then concatenated back to each track with the original local features. This would enable repeated back-and-forth information exchanges between the members of the set in an order-invariant fashion. However, the potential drawback is that this design cannot preserve precious scene motion information across consecutive frames, which might be critical for the dynamic burst denoising scenarios with not only camera motion, but also scene motion.

A.5 Camera simulation pipeline

Following the discussion in Sec. 4.1. The simplest and most commonly used noise model is the homoscedastic Gaussian assumption, also known as the additive white Gaussian noise (AWGN). Despite its prevalence, the Gaussian model does not represent the fact that photon noise is signal-dependent. To better model noises, the Poisson-Gaussian model or heteroscedastic Gaussian model (i.e. novel level function) are widely used. However, in real images there may still exist other noise sources that may not be accurately represented by such models, e.g. fixed-pattern noise, defective pixels, clipped intensities, spatially correlated noise, amplification, and quantization noise thus a realistic camera simulation pipeline which could add noises in the RAW domain would be helpful. [19,20,21].

Although there not exist a perfect camera simulation pipeline yet which could simulate all these kind of noises as from the physical camera, we have tested

Ground truth *Noisy:* 12.52*dB* *Denoised:* 21.41*dB*

Fig. 2: For Sec. 5.6: Failure cases of our model under extremely amplified noises levels far beyond the training range. The trained model outputs less than ideal denoised result, which shows explicit artifacts including color shift and distortion.

and finally choose `Cam_Sim` [22] to synthesize all the synthetic noisy bursts for both static and dynamic scenes. The proposed pipeline in [22] contains over 40 individual modules, and covers a good range of typical camera processes such as tone-mapping, demosaicking, and denoising. The main stages for a forward process contain artifact generation, demosaicking, cellphone denoising (with bilateral filters), and tone-mapping and post-processing. For all available sRGB images in DIV2K and Vimeo90K datasets, we apply the inverse process based on the same pipeline, add more significant Gaussian and Poisson noises (as described in the paper), and finally process them back to the sRGB domain. We also tested the pipelines proposed in [23] and [24], but found `Cam_Sim` provides the best simulation quality and flexibility. We do find that there still exist deficiencies in Cam_Sim, including some unprocessed chroma noises and potentially mishandled artifacts which could hamper the generalization of the proposed model. Thus one future direction is to further improve the utilized camera simulation pipelines.

A.6 Failure case of the proposed model

As discussed in Sec. 5.6 in the paper, the main limitation of our proposed model is that it is still trained in an non-blind denoising fashion without taking a noise estimation map as input from a individual noise estimator such as the ones proposed in [13,51], thus lacking the capability of adaptive noise-aware burst denoising which is recently popular in single-image denoising methods. Fig. 2 here demonstrates a failure of our model on a severely corrupted burst, which is considerably noisier than examples seen during training. The model struggles to recover the corrupted detail, but instead produces unsatisfactory artifacts. Integrating the proposed model with a noise level estimation mechanism is a promising future research direction to mitigate this problem.

B More interactive qualitative results.

B.1 Burst denoising on synthetic noisy burst data

`Please check the interactive web-based image gallery for images.`

As shown in the image gallery, while the previous methods are prone to generate over-smoothed results with losing local details, our proposed model is capable of preserving more precious local high-frequency information, and also generating less artifacts. Also, slight simulated camera motion are comparably well tackled by our proposed model, even though the kernel-sampling based methods such as KPN are naturally more robust to slight disturbances.

B.2 Burst denoising on real noisy burst data

`Please check the interactive web-based image gallery for images.`

As shown in the image gallery, though our method visually outperforms KPN on realistic bursts, they are both non-blind denoising models and it is still challenging and difficult for them to adapt to various realistic scenes in contrast to a conditional blind denoising model. We feel the proposed model can be further improved by incorporating a noise estimator that would allow our denoising method to adapt to a wider range of spatially-variant noise levels.

References

1. Mildenhall, B., Barron, J.T., Chen, J., Sharlet, D., Ng, R., Carroll, R.: Burst denoising with kernel prediction networks. In: CVPR. (2018) 2502–2510 1
2. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 6228–6237 1
3. Weickert, J.: Anisotropic diffusion in image processing. Teubner, Stuttgart (1998) 1
4. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: ICCV. (1998) 839–846 1
5. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D: nonlinear phenomena **60**(1-4) (1992) 259–268 1
6. Antonini, M., Barlaud, M., Mathieu, P., Daubechies, I.: Image coding using wavelet transform. IEEE Transactions on image processing (TIP) **1**(2) (1992) 205–220 1
7. Portilla, J., Strela, V., Wainwright, M.J., Simoncelli, E.P.: Image denoising using scale mixtures of gaussians in the wavelet domain. Trans. Img. Proc. **12**(11) (November 2003) 1338–1351 1
8. Buades, A., Coll, B., Morel, J.M.: A non-local algorithm for image denoising. In: CVPR. Volume 2., IEEE (2005) 60–65 1
9. Elad, M., Aharon, M.: Image denoising via learned dictionaries and sparse representation. In: CVPR. Volume 1., IEEE (2006) 895–900 1

10. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3-d transform-domain collaborative filtering. IEEE Transactions on Image Processing (TIP) **16** (2007) 2080–2095 1
11. 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 (TIP) **26**(7) (2017) 3142–3155 1, 2
12. Liu, P., Zhang, H., Zhang, K., Lin, L., Zuo, W.: Multi-level wavelet-cnn for image restoration. In: CVPR Workshop. (2018) 773–782 1
13. Zhang, K., Zuo, W., Gu, S., Zhang, L.: Learning deep cnn denoiser prior for image restoration. In: CVPR. (2017) 3929–3938 1, 2
14. Laine, S., Lehtinen, J., Aila, T.: High-quality self-supervised deep image denoising. arXiv preprint arXiv:1901.10277 (2019) 1
15. Batson, J., Royer, L.: Noise2self: Blind denoising by self-supervision. In: ICML. (2019) 524–533 1
16. 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) (2018) 4608–4622 2
17. Zhou, Y., Jiao, J., Huang, H., Wang, Y., Wang, J., Shi, H., Huang, T.: When awgn-based denoiser meets real noises. arXiv preprint arXiv:1904.03485 (2019) 2
18. Guo, S., Yan, Z., Zhang, K., Zuo, W., Zhang, L.: Toward convolutional blind denoising of real photographs. In: CVPR. (2019) 1712–1722 2
19. Abdelhamed, A., Lin, S., Brown, M.S.: A high-quality denoising dataset for smartphone cameras. In: CVPR. (2018) 1692–1700 3
20. Plotz, T., Roth, S.: Benchmarking denoising algorithms with real photographs. In: CVPR. (July 2017) 3
21. Abdelhamed, A., Brubaker, M., Michael, B.: Noise Flow: Noise Modeling with Conditional Normalizing Flows. In: ICCV. (2019) 3
22. Jaroensri, R., Biscarrat, C., Aittala, M., Durand, F.: Generating training data for denoising real rgb images via camera pipeline simulation. arXiv preprint arXiv:1904.08825 (2019) 4
23. Brooks, T., Barron, J.T.: Learning to synthesize motion blur. In: CVPR. (2019) 6840–6848 4
24. Karaimer, H., Brown, M.: A software platform for manipulating the camera imaging pipeline. In: ECCV. (2016) 4