IMPROVING EXTREME LOW-LIGHT IMAGE DENOISING VIA RESIDUAL LEARNING

Paras Maharjan[1], Li Li[1], Zhu Li[1,2], Ning Xu[3], Chongyang Ma[4], Yue Li[5]

[1]University of Missouri-Kansas City, USA,
[2]Peng Cheng Lab, Shenzhen, China,
[3]Amazon Go, USA
[4]Kwai Inc., USA
[5]University of Science and Technology of China, China

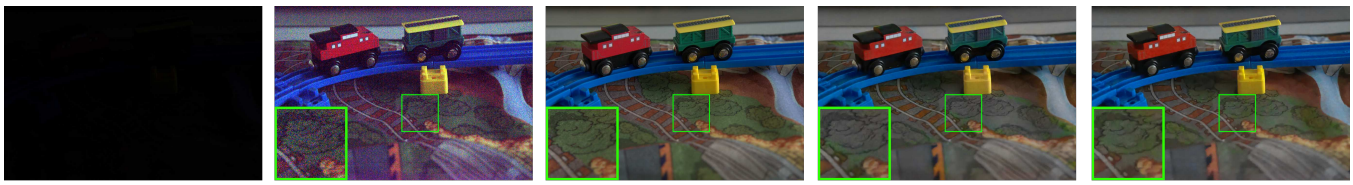

Fig. 1: [a] Extreme low-light image from Sony α7S II exposed for 1/10 second . [b] 100x intensity scaling of image in [a]. [c] Ground truth image captured with 10 second exposure time. [d] Output from [1]. [e] Output from our method.

ABSTRACT

Taking a satisfactory picture in a low-light environment remains a challenging problem. Low-light imaging mainly suffers from noise due to the low signal-to-noise ratio. Many methods have been proposed for the task of image denoising, but they fail to work under extremely low-light conditions. Recently, deep learning based approaches have been presented that have higher objective quality than traditional methods, but they usually have high computational cost which makes them impractical to use in real-time applications or where the processing power is limited. In this paper, we propose a new residual learning based deep neural network for end-to-end extreme low-light image denoising that can not only significantly reduce the computational cost but also improve the quality over existing methods in both objective and subjective metrics. Specifically, in one setting we achieved 29x speedup with higher PSNR. Subjectively, our method provides better color reproduction and preserves more detailed texture information compared to state-of-the-art methods.

Index Terms— Deep Residual Learning; Image Denoising; Low Light Image Enhancement.

1. INTRODUCTION

Low-light imaging is one of the most challenging tasks in image processing and computer vision, especially when the environment is extremely dark. Current image sensors are still suffering from low signal-to-noise ratio (SNR) in extremely low-light environment and will produce very noisy images if there are not enough photons reaching the sensors. Enlarging the aperture will reduce the depth of field and lead to blurry images in most cases, while extending exposure time will cause motion blur and is not feasible when capturing videos. There are extensive studies on how to reproduce natural scenes with correct exposure, accurate color and detailed texture from noisy short exposure low-light images. Traditional image denoising approaches, for instance BM3D [2], work reasonably well for moderate amount of noise in normal lighting conditions. However, they perform poorly in extreme low-light condition.

Recently, a deep learning based method [1] was proposed to deal with the extreme low-light image denoising problem, using a raw image captured from the sensor as input. The authors introduce a dataset of raw short-exposure low-light images, with the corresponding long-exposure photos as reference. They propose to use U-Net [3] as the network architecture and present promising results on this dataset. However, the U-Net architecture used in this work causes two problems. First, the autoencoder based network with the use of max pooling layer for feature downsampling will lose image details and generate output with blurry edges, even with skip connections to mitigate the degradation. Second, the U-Net architecture is slow at inference time, which makes it difficult to be used for fast imaging and video applications under low-light conditions.

To solve the problem of the previous work, we propose a

This work has been supported by the gift grant from Snap Inc.

novel residual learning based end-to-end network to enhance extreme low-light images. In our proposed residual blocks, we replace ReLU layer with LeakyReLU as the nonlinear activation function, remove the batch normalization layers, and add Squeeze-and-Excitation (SE) block [4] for feature recalibration. Comparing with the U-Net architecture in [1], the use of residual learning in our proposed network helps extract and represent the color and texture information in low-light images. Furthermore, using LeakyReLU as activation function in the residual block introduces slope in the negative region of the feature, thus preserves the information of the features with negative values. Finally, the SE block in residual block improves the representation quality by re-calibrating the convolutional features, and also helps converge faster to a stable network. We have found that the integration of above modifications is effective in speeding up the training process and improving the denoising performance.

Compared with previous work, our method not only leads to much faster inference, but also results in better objective and subjective qualities. We compare our proposed method with the work in [1] in Figure 1. Our proposed network is able to reconstruct the image from the extreme low-light image with better color accuracy and higher image quality.

2. RELATED WORK

Extensive research has been conducted on low light image denoising and enhancement. Here we provide a brief literature review of existing research work.

2.1. Image Denoising

Many conventional methods have been developed for image denoising. Plotz and Roth [5] propose a benchmark dataset of real noisy images to compare traditional image denoising methods and find that the sparse 3D transform-domain collaborative filtering (BM3D) [2] outperforms other methods such as Weighted Nuclear Norm Minimization (WNNM) [6], K-SVD [7], Expected Patch Log Likelihood (EPLL) [8], Field of Experts (FoE) [9], and Nonlocally Centralized Sparse Representations (NCSR) [10].

More recently, deep learning based image denoising methods have gained popularity. DnCNN [11] uses Batch Normalization (BN) and ResNet [12] to perform image denoising and has shown significant performance gain over traditional methods including BM3D. This network not only performs image denoising, but also achieves super-resolution to the denoised images and makes the image looks more satisfying to human eyes. However, all of these methods cannot produce good quality images when processing extremely low-light images.

2.2. Low-Light Image Enhancement

Histogram equalization and gamma correction are the most common traditional methods for image enhancement. Although these methods work well on normal dark images, they fail on extremely low-light condition because of introduction of quantization errors. Deep learning based methods such as [13] use a burst of images taken with different exposure times and fuse them to produce a single denoised image. These methods are not very practical because of the complex network behind image fusion and time inefficiency for capturing and processing. In addition, this type of methods are not possible for video application.

More recent work in low-light image processing is *Learning to See in the Dark* (SID) [1] that proposes to use an end-to-end fully convolutional network on raw sensor data to replace the whole traditional image processing pipeline. They also introduced a dataset of raw short-exposure low-light images, with the corresponding long-exposure reference images. Their work uses U-Net as the main network architecture which causes some quality issue in resulting images and is also slow at inference time.

Inspired by the residual learning (DnCNN) and See-in-Dark (SID), we propose a new network architecture to address the issues with these methods.

3. OUR METHOD

In this section, we will describe our proposed method for extreme low-light image denoising and enhancement. The overall network architecture of our proposed method is shown in Figure 2. Raw sensor image is separated into RGBG color planes with half size, before an amplification ratio is multiplied. The main structure of our network is a residual learning framework. The residual learning assumes that the residue can be more easily learned by the network rather than the whole image itself. After residual learning, the output is upsampled x2 using convolution layers with pixel shuffling [14].

Our main network contains 32 residue blocks [12]. The structure of each residue block is shown in Figure 2[b]. For this task we design a residue block that contains a first 3x3 convolution layer, followed by a Leaky ReLU layer, a second 3x3 convolution layer, a constant linear scaling unit, and finally the output layer which is re-calibrated by an Squeeze-and-Excitation block [4].

Compared with the network in SID [1], we replace the U-Net architecture with residual learning. We argue that the use of the maxpool layer and reduction of feature size in U-Net architecture will remove the important information from image features. Therefore, unlike the U-Net architecture which has the contracting and expanding structure, we propose to use a network architecture without a downscaling structure. In our network, we use a constant feature size throughout the residual part of the network.

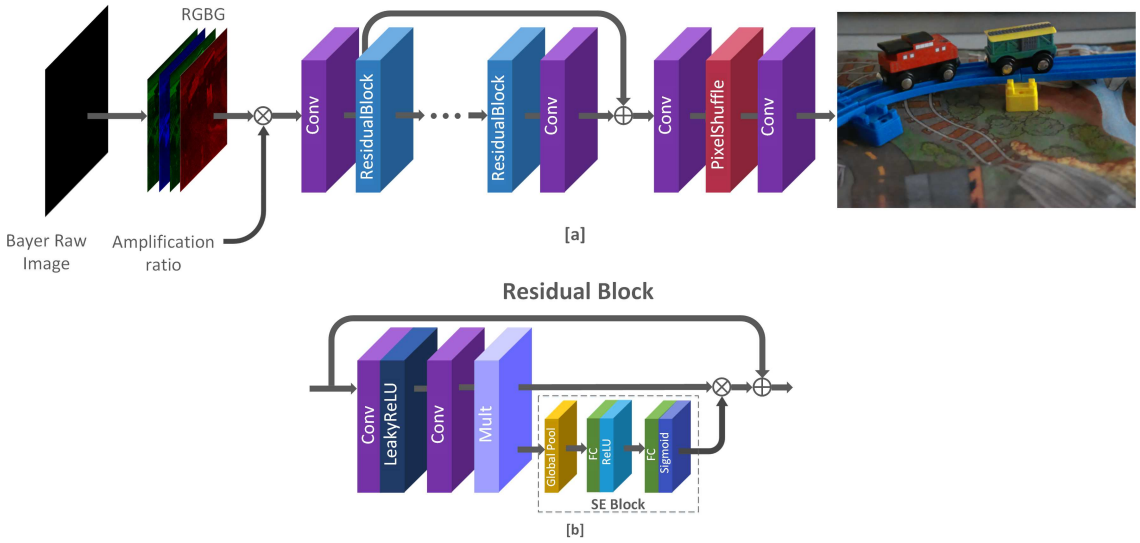

Fig. 2: [a] Overview of our system. Raw sensor image is separated into different color planes on which an amplification ratio is multiplied. After residual learning, the output is upsampled x2 using convolution layers with pixel shuffling. The network for residual learning contains a number of residue blocks. [b] Residual block details. Each residual block contains LeakyReLU layer and an SE block.

We introduce several modifications in our network architecture compared to recent residual network [12, 11, 15] which are successfully applied for image super resolution task. In these methods, rectified linear unit (ReLU) was used as the activation function for each residual block. ReLU zeros out the negative information from the feature, which also carries important information about the local structure and should be preserved for better reconstruction of the output image. In our design, we use LeakyReLU instead of ReLU as the activation function for the residual block.

Within each residual block, we also add a Squeeze-and-Excitation block, which has shown improvement in network performance of ResNet and Inception module [4]. It is observed that integration of SE block within the residual block is effective in speeding up the training process and boosting the denoising performance. SE block improves the feature representation of network by using the channel wise feature scaling.

During the training process, we set our input size to be 256×256 pixel and use four-channel RGBG input extracted from the raw images of SID dataset [1]. Since our proposed network is less complex than its counterpart SID [1], we are able to increase the depth of the network to 32 residual blocks, while keeping the inference speed of 4K resolution image fast enough for realtime processing. Increasing the depth of the residual learning helps learn better visual features. The input raw sensor image is first linearly scaled by the amplification ratio which is the difference of the exposure time between the short exposure images and long exposure ones.

4. EXPERIMENTS

4.1. Dataset and Experimental Setup

We use the SID dataset [1] which contains real-world extreme low-light images with the corresponding noise-free ground truth images. The dataset consists of 5094 raw images from Sony a7S II and Fujifilm X-T2 sensors. Our network is trained with images from Sony sensor that uses the full-frame Bayers filter array. The dataset contains the dark images with three different exposure time of $1/10$, $1/25$ and $1/30$ seconds and the corresponding ground truth images with exposure of 10 seconds. The time difference between the shutter speed is taken as the amplification ratio for dark image and ground truth pair.

The input to the network is a raw image captured with a short exposure time and the output is an sRGB image. The ground truth is the corresponding standard RGB long exposure image produced from the raw sensor image with the *libraw* library. During the training process, the input size is 256×256, randomly cropped from input image set with flipping and rotation for data augmentation and the output is 3 channel 512×512 sRGB image. We have experimented with both 16 and 32 residual blocks. The negative slope parameter of LeakyReLU is set to 0.2. We use L1 loss and Adam optimizer. The network is trained for 6000 epochs with an initial learning rate of 10^{-4} which is reduced by a factor of 10 after every 2000 epochs. Our training process is performed on a PC with Intel i5-8400 CPU, 16GB memory, and NVIDIA GTX 1080 GPU.

Table 1: Quantitative comparison.

Experiments	PSNR	SSIM
SID	28.97	**0.8857**
Ours - No SE Block	28.49	0.8817
Ours - 16 Residual Blocks	29.15	0.8829
Ours - 32 Residual Blocks	**29.16**	0.8856

Table 2: Performance analysis.

Experiments	x100	x250	x300
BM3D	21.23	19.97	19.01
SID	30.08	28.42	**28.52**
Ours - 32 Residual Blocks	**30.53**	**28.78**	28.38

Table 3: Complexity analysis.

Experiments	# of parameters	Time(sec)
BM3D	-	385.90
SID	7.76M	0.235
Ours - 16 Residual Blocks	1.38M	0.008
Ours - 32 Residual Blocks	2.5M	0.011

4.2. Subjective Quality

4.2.1. Denoising

Our proposed network reduces the noise of low-light images while preserving the color and texture information. Figure 3 shows the results of our method compared with SID [1] and BM3D [2]. BM3D is applied after linear scaling up of the original images with an amplification ratio. For each scaling factor, multiple sigma values are tried and the best one is used to obtain the results. Specifically, the sigma value is set to 200 for the 100x scaling while 300 is set for the 250x and 300x amplifications of the input images. Even with the optimal sigma level setting, our method achieves better results than BM3D for these extreme low light image cases. SID results are obtained using the source code provided in [1].

4.2.2. Color Accuracy

The image color is more accurately recovered in our proposed network than in SID, when taking the ground truth image as the reference. Most of the images produced by SID are either less colorful than the ground truth or have no color information, while our proposed method produces colorful results closer to the ground truth. Figure 4 shows an example where the output of the SID has completely different color on the wall. It only produces some color at the edge of the wall. The floor in the image is slightly discolored. Our proposed method is able to reproduce the wall color and the floor color more accurately.

4.2.3. Color Spreading

We also notice a common green and yellow color spreading issue in the output of SID results. As we can see in Figure 5 the grass is replaced by the barren land like structure in the SID output. However, our proposed method is able to generate results which are closer to the ground truth.

4.2.4. Image Details

Since we do not reduce the feature size, we find our approach can better preserve the texture and edge details in the output images. On the contrary, SID produces output with smoother texture and may lose details due to contracting and symmetric expanding structure of the U-Net architecture. Figure 6 shows that the output image in the zoom-in area is much clearer in the results by our proposed network than those from SID.

4.3. Objective Quality

Figure 7 shows comparison in loss curve for our proposed method vs SID. The loss in our proposed approach is converging faster as compared to SID. The use of the Squeeze-and-Excitation (SE) [4] block in the our network is effective in speeding up the training and boosting the denoising performance. As we can see in the figure, our proposed method converges much faster at the beginning and keep a big margin along the way for the entire training process.

We uses peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) as performance metrics for objective image quality comparison, and the results are shown in Table 1. As we can see in the table, our methods outperforms SID in terms of PSNR. At the same time, in terms of average SSIM, our results are comparable to SID. In Table 1, we can also see that the performance of our methods with SE block is much better than the one without SE block.

We further break the input images into three categories based on the aplification ratio, and find that our methods has better results for the amplification ratios of x100 and x250. Table 2 shows the performance for each of the scaling factor in comparison to SID and BM3D.

4.4. Complexity Analysis

Our proposed network architecture has much less model parameters compared to the U-Net architecture used in SID [1]. Table 3 shows the complexity analysis of our proposed network compared with SID and BM3D. There are two configurations of our proposed network, one has 32 residual blocks

Fig. 3: Image denoising results. [a] Ground truth image. [b] Output from SID. Noise is still present in few parts of the image. [c] Output from BM3D. Denoised image is darker than the ground truth. [d] Denoised output from our network.

Fig. 4: Comparison of color Accuracy. [a] Input dark image. [b] 100x scaled version of dark images. [c] Ground truth with exposure time of 10 seconds. [d] SID output with missing color information, PSNR: 20.48dB. [e] Output from our network with close approximation to ground truth image, PSNR: 27.17dB.

and the other has 16 residual blocks. With our network with 32 residual blocks we get around 21x faster processing time, while in another setting with 16 residual blocks we get 29x faster processing speed with higher PSNR than the SID.

5. CONCLUSIONS

In this paper we propose a new deep residual learning network with Squeeze-and-Excitation block for denoising and enhancement of extremely low-light image. The experimental results show that our network not only has better PSNR gain over the SID counterpart but also has reduced computational cost. With our residual network we are able to denoise the image under extremely low light condition while preserving most of the color and texture information. This advantage makes our network suitable for fast processing of low light images and videos on resource constrained devices. In the future we plan to design low-light image understanding solution via end-to-end learning for various vision tasks.

6. REFERENCES

[1] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun, "Learning to see in the dark," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3291–3300.

[2] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," in *IEEE Transactions on Image Processing*, Aug 2007, pp. 2080–2095.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, vol. 9351 of *LNCS*, pp. 234–241.

[4] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[5] Tobias Plotz and Stefan Roth, "Benchmarking denoising algorithms with real photographs," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[6] Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng, "Weighted nuclear norm minimization with application to im-

Fig. 5: Comparison of color spreading issue. [a] Input dark image form Sony 300x subset. [b] 300x amplification of dark image. [c] Ground truth image with exposure time of 10 seconds. [d] U-Net output with unnecessary color spread at the ground. [e] Our output with close approximation to ground truth image.

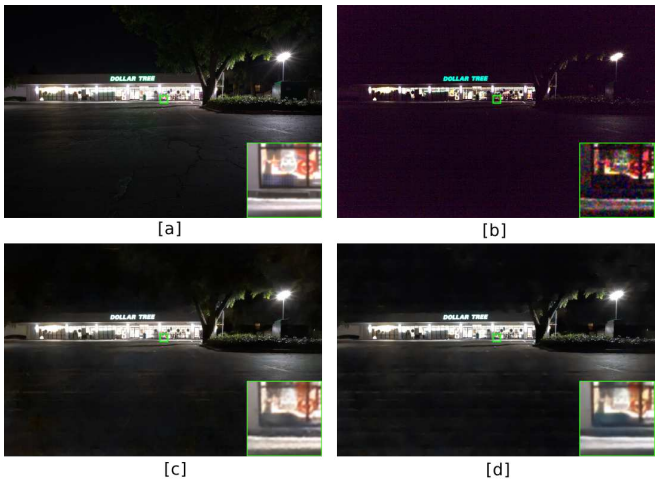

Fig. 6: Comparison of image details. [a] Ground truth. [b] Input image amplified by 300x. [c] U-Net output. [d] Our result with higher image quality.

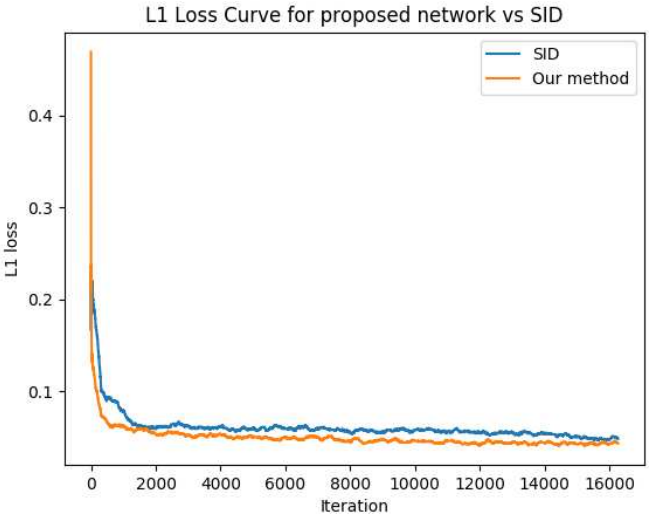

Fig. 7: L1 loss curves for our proposed method vs SID for 100 epochs.

age denoising," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 2862–2869.

[7] Michal Aharon, Michael Elad, and Alfred Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *Signal Processing, IEEE Transactions on*, vol. 54, pp. 4311–4322, Dec 2006.

[8] Daniel Zoran and Yair Weiss, "From learning models of natural image patches to whole image restoration," in *Proceedings of the IEEE International Conference on Computer Vision*, Nov 2011, pp. 479–486.

[9] Stefan Roth and Michael Black, "Fields of experts: A framework for learning image priors.," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jan 2005, vol. 2, pp. 860–867.

[10] Weisheng Dong, Lei Zhang, Guangming Shi, and Xin li, "Nonlocally centralized sparse representation for image restoration," *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, vol. 22, Dec 2012.

[11] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *The IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[13] K. Ram Prabhakar, V Sai Srikar, and R. Venkatesh Babu, "Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[14] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1874–1883, 2016.

[15] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Jul 2017, pp. 136–144.