IMPROVING EXTREME LOW-LIGHT IMAGE DENOISING VIA RESIDUAL LEARNING

A THESIS
in
Electrical Engineering

Presented to the Faculty of the University
of Missouri–Kansas City in partial fulfillment of
the requirements for the degree

MASTER OF SCIENCE

by
PARAS MAHARJAN

B. E., Kantipur Engineering College,
affiliated to Tribhuvan University,
Lalitpur, Nepal, 2014

Kansas City, Missouri
2019

IMPROVING EXTREME LOW-LIGHT IMAGE DENOISING VIA RESIDUAL LEARNING

Paras Maharjan, Candidate for the Master of Science Degree

University of Missouri–Kansas City, 2019

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 with the noise 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 computation cost which makes them impractical to use in real-time applications or where the computational resource 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 performance 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.

APPROVAL PAGE

The faculty listed below, appointed by the Dean of the School of Computing and Engineering, have examined a thesis titled "Improving Extreme Low-light Image Denoising via Residual Learning," presented by Paras Maharjan, candidate for the Master of Science degree, and hereby certify that in their opinion it is worthy of acceptance.

Supervisory Committee

Zhu Li, Ph.D., Committee Chair
Department of Computer Science & Electrical Engineering

Yugyung Lee, Ph.D.
Department of Computer Science & Electrical Engineering

Baek-Young Choi, Ph.D.
Department of Computer Science & Electrical Engineering

CONTENTS

ILLUSTRATIONS

TABLES

Tables Page

ACKNOWLEDGEMENTS

I would like to express my great appriciation to my academic advisor Dr. Zhu Li for his continuous support and ideas during the planning and development of this research work. Thanks to all members of Multimedia Computing and Communication Lab at UMKC specially Dr. Li Li and visiting scholar from University of Science and Technology of China, Yue Li, this would not have been possible without your help.

This research was funded by gift grant from Snap Inc. So I would like to offer my special thanks to my supervisor from Snap Inc., Ning Xu and Chongyang Ma for financial support and their patient guidance, enthusiastic encouragement and useful suggestion throughout this research work.

Finally, I wish to thank my parents for their support and encouragement throughout my study.

CHAPTER 1

INTRODUCTION

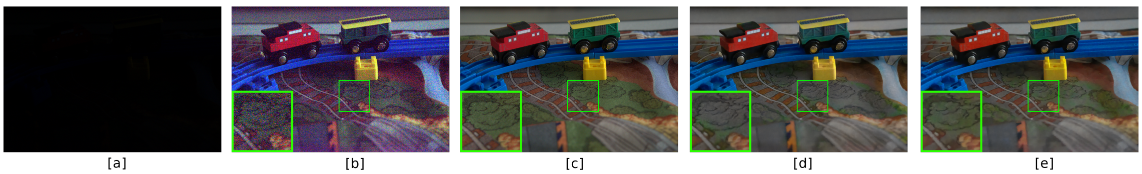

Figure 1: (a) Extreme low-light image from Sony α7S II exposed for 1/10 second . (b) 100x intensity scaling of image in (a). (c) Ground truth image captured with 10 second exposure time. (d) Output from [1]. (e) Output from our method.

Low light imaging [1–4] 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 the natural scene with correct exposure, accurate color and detailed texture from noisy short exposure extreme low-light images. Traditional image denoising approaches, for instance BM3D [5], 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 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 reference images. They propose to use U-Net [6] as the network architecture and reported 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 result in smoother output with blurry edges, although the skip connections could 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 novel residual learning based end-to-end network to deal with the extreme low-light image enhancement problem. 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 [7] for feature re-calibration.

Comparing with the U-Net architecture in [1], the use of residual learning in our proposed network help in better learning of the color and texture information of the 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 proposed method not only leads to much faster inference time, but also results in better objective and subjective qualities. A typical example of the comparison between our proposed method and the work in [1] is shown 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.

Figure 2 shows the traditional Image Signal Processor (ISP) pipeline. These block of ISP are tuned differently for different ISP vendors. Based on the vendors these block might interchanged or some extra blocks may be added to further enhance the performance. However, this architecture of ISP works for normal lighting condition and fails when used under under or over exposed environment. Hence in this work, we propose an end-to-end solution for joint denoising and demosaicing approach for the enhancement of low light images.

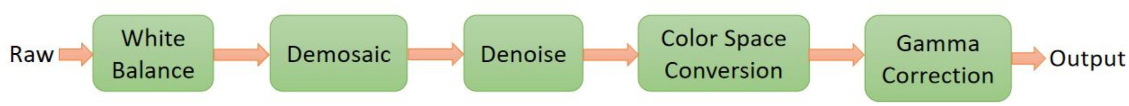

Figure 2: Overview of traditional imaging pipeline

CHAPTER 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 [8] proposed a benchmark dataset of real noisy images to compare traditional image denoising methods and found that the sparse 3D transform-domain collaborative filtering (BM3D) [5] outperforms other methods such as Weighted Nuclear Norm Minimization (WNNM) [9], K-SVD [10], Expected Patch Log Likelihood (EPLL) [11], Field of Experts (FoE) [12] and Nonlocally Centralized Sparse Representations (NCSR) [13]. More recently deep learning based image denoising methods have gained popularity. DnCNN [14] uses Batch Normalization (BN) and ResNet [15] 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 the eyes. However both of these network suffers to produce good quality images when processed with 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 that use multi-exposure fusion like [16] uses a burst of images taken with different exposure time and use deep network to fuse them and produce 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 proposed 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 image, and is also slow in inferencing.

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

CHAPTER 3

METHOD

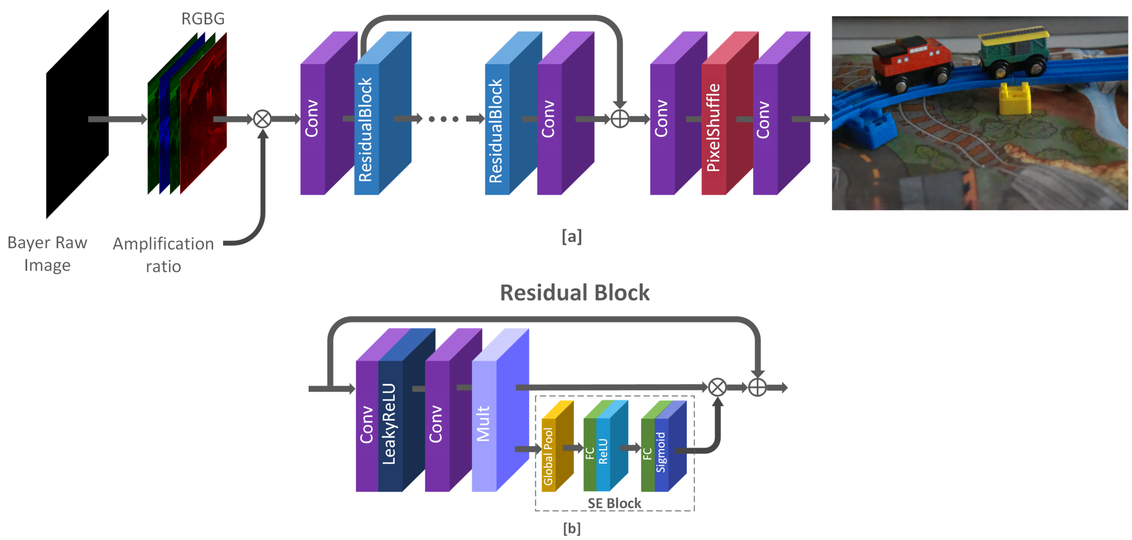

Figure 3: (a) Raw sensor image is separated into different color planes on which an amplification ratio is multiplied. After residual learning, the output is upsampled x2 using convolution layers with pixel shuffling. The network for residual learning contains a number of residue blocks. (b) Residual block details. Each residual block contains LeakyReLU layer and an SE block.

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 3. 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 [17].

3.1 Network Architecture

The main network contains 32 residue blocks [15], and the structure of each residue block is shown in Figure 3[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 of the residual block is re-calibrated by an Squeeze-and-Excitation block [7].

Compared with the network in SID, we replaced 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 the feature. Therefore, on contrary to the U-Net architecture which has the contracting and expanding structure, in this paper we propose to use the network architecture without the downscaling structure. In our proposed network, we uses a constant feature size throughout residual part of the network.

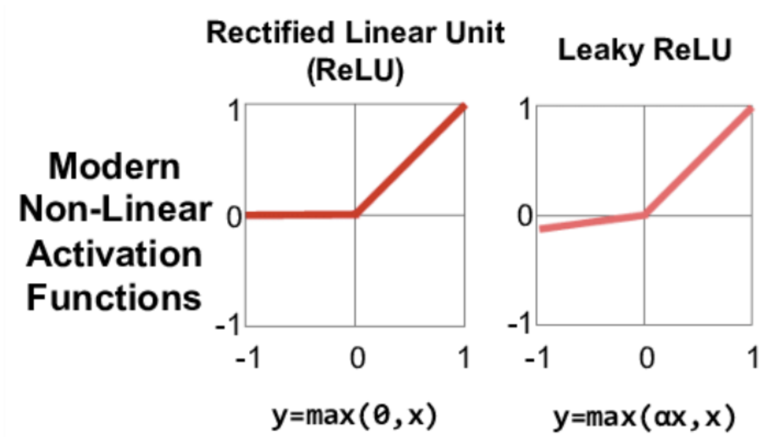

Figure 4: LeakyReLU as activation function

Several modifications are introduced in our network architecture compared to recent residual network [14, 15, 18] which are successfully applied for image super resolution task. In these network, rectified linear unit(ReLU) was adopted 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. LeakyReLU with the negative slope of 0.2 showed better model convergence with no extra computational cost. Figure 4 shows the non linearity curve of LeakyReLU activation function compared to ReLU.

Within each residual block, we also add a Squeeze-and-Excitation block, which has shown improvement in network performance of ResNet and Inception module [7]. It is observed that integration of SE block within the residual block is effective in speeding up the training and boosting the denoising performance. SE block improves the feature representation of network by using the channel wise feature scaling.

In training, we set our input size to 256x256 pixel and used 4 channel RGBG image 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 in learning better features. The input raw sensor image is first linearly scaled by the amplification ratio which is the difference of the exposure time between short and long exposure images.

3.2 Loss function

We use L1 loss as the loss function for our network. L1 Loss function minimizes the absolute differences between the predicted value and the ground truth value. The L1 loss is implements as follows,

$$\mathcal{L}_1 = |\hat{x} - x| \tag{3.1}$$

where, $\hat{x}$ is predicted image and x is corresponding long exposure ground truth image.

CHAPTER 4

EXPERIMENTS

4.1 Dataset and Experimental Setup

We use See-in-the-Dark dataset [1] that contains the real world extreme low-light images with its corresponding noise-free ground truth images. The dataset contains 5094 raw images from Sony a7S II and Fujifilm X-T2 sensors with dark short exposed images and its respective bright long exposed images. 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. There were some misalignment found in the test set of the SID. So, we removed such images from the dataset for performance evaluation.

The input to the network is raw sensor image with short exposure and the output is sRGB image. The ground truth is the corresponding standard RGB long exposure image produced from the raw sensor image with the *libraw* library. In training, the input size is 256x256, randomly cropped from input image set with flipping and rotation for data augmentation and the output is 3 channel 512x512 sRGB image. We experimented with both 16 and 32 residual blocks. The negative slope parameter of LeakyReLU is set as 0.2. L1 loss is used as our loss function and Adam is used as optimizer. The network is trained

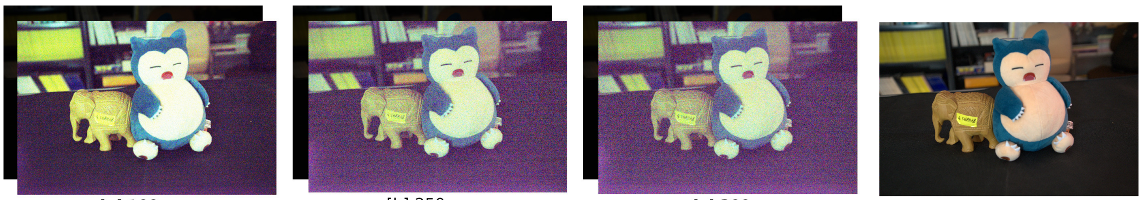

Figure 5: Example of SID Dataset. For each scene there are at most three different images with three exposure time

for 6000 epochs, with learning rate set to 10^{-4} initially and reduced by factor of 10 after every 2000 epochs. Our training process are carried out on a PC with Intel i5-8400 CPU, 16GB memory and Nvidia GTX 1080 GPU.

4.2 Performance metrics

We use peak signal to noise ratio (PSNR) and SSIM [19] as our performance metrics. The PSNR is calculated as,

$$MSE = \frac{1}{N}\sum_{n=1}^{N}(\hat{x}(n) - x(n))^2 \tag{4.1}$$

$$PSNR = 10\log_{10}(\frac{1}{MSE}) \tag{4.2}$$

where N is the number of pixels and MSE is the mean square error between the predicted and the ground truth image. The higher value of PSNR is expected. Since the PSNR alone cannot determine the perceptual quality of the image, we also used SSIM as performance metric. The SSIM value closer to 1 represents predicted image is perceptually similar to ground truth.

Figure 6: (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

4.3 Subjective Quality

4.3.1 Denoising

The proposed network reduces the noise of low-light images while preserving the color and texture information. Figure 11 shows the results of our method compared with SID [1] and BM3D [5].

BM3D is applied after linear scaling up of the original images with corresponding amplification scaling. 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 and 300 for the 250x and 300x amplification scaling of the image. Even with the optimal sigma level setting, the BM3D results are still poor as compared to our method for these extreme low light image cases. This is expected as explained in [1]. SID results are obtained using the source code provided in [1].

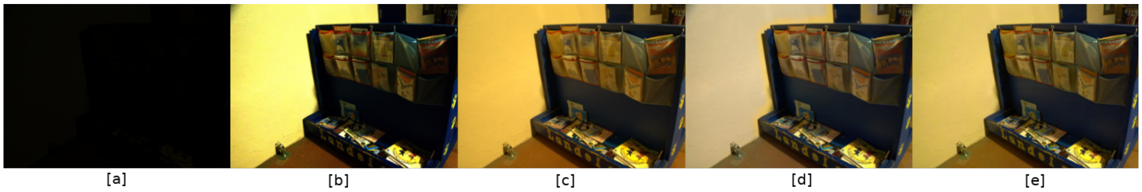

Figure 7: (a) Input dark image (b) 100x scaled version of dark images (c) Ground truth with exposure time of 10 seconds (d) SID output with missing color information, PSNR: 20.48dB (e) Output from our network with close approximation to ground truth image, PSNR: 27.17dB.

Figure 8: (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) Output from our network with close approximation to ground truth image.

4.3.2 Color Accuracy

The output image color is more accurately recovered in the our proposed network than in SID, when compared to the ground truth image. Most of the images reproduced by SID are either discolored or have no color information, while our proposed method produce the color closer to the ground truth.

Figure 7 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.

4.3.3 Color Spreading

We also notice a common green and yellow color spread in the output of SID results. As we can see in Figure 8 the grass is replaced by the barren land like structure in the SID output. However, our proposed method is able to generate the close approximation to the ground truth.

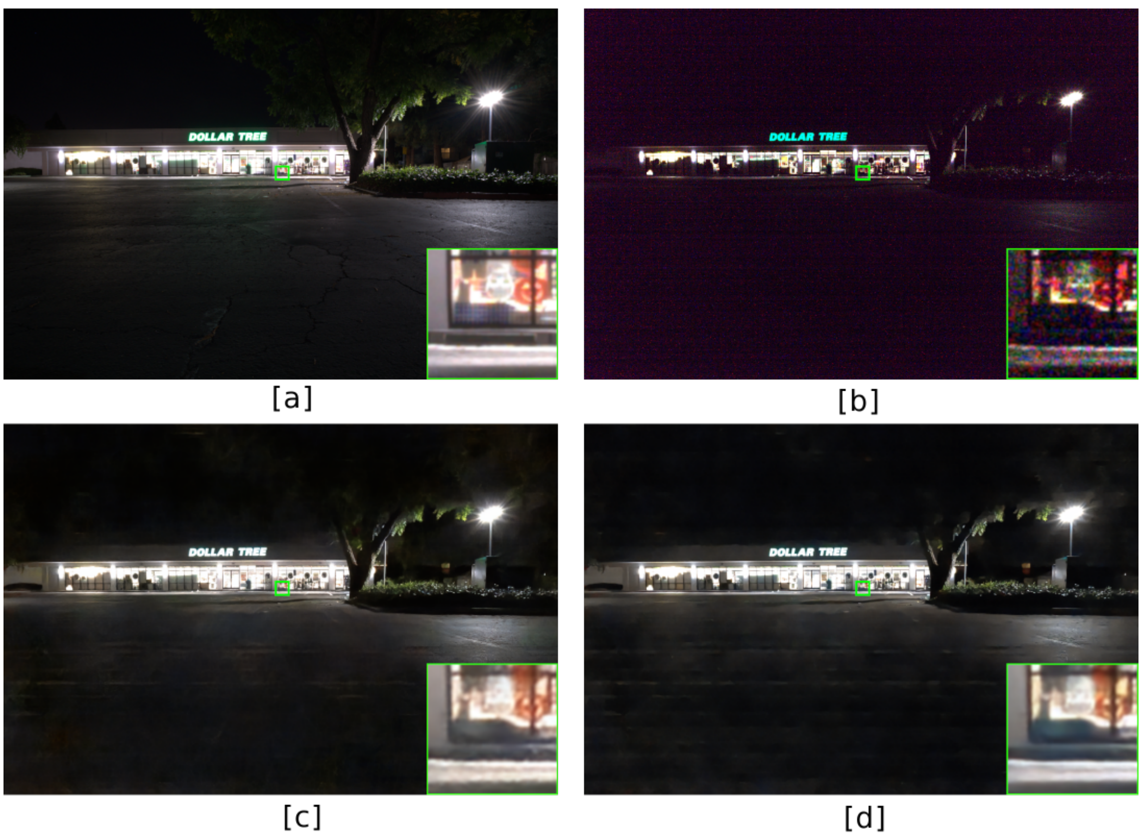

Figure 9: (a) Ground truth (b) 300x amplified dark image (c) U-Net output. Image not clear due to pixelated effect. (d) Output from our network with higher image quality

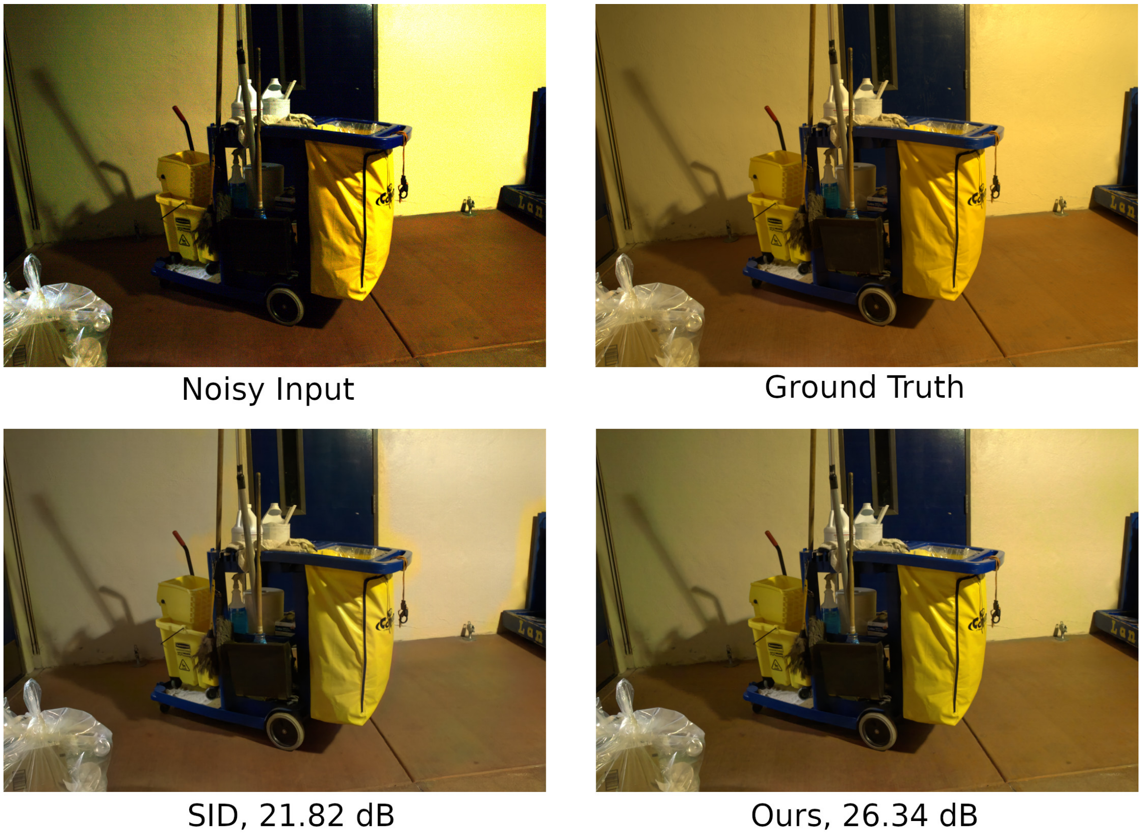

Figure 10: More example on color accuracy

4.3.4 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 lost details due to contracting and symmetric expanding structure of U-Net applied. Figure 9 shows the output image in the zoom-in area is much clearer in the results from our proposed network than from SID.

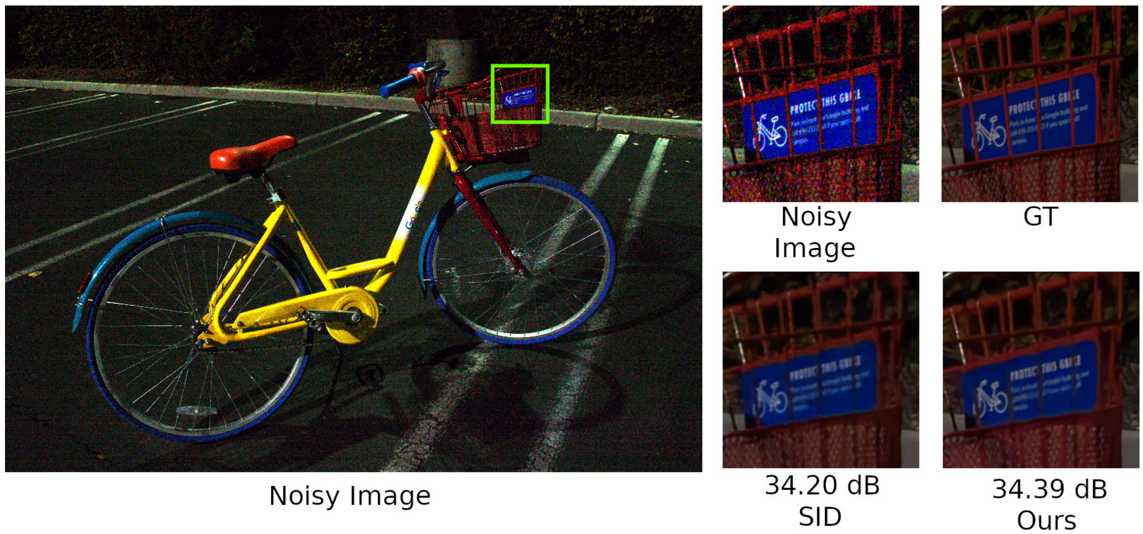

Figure 11: Example showing preservation of textural information

4.4 Objective Quality

Figure 12 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) [7] 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) [19] as performance metrics for objective image quality compared to the ground truth image, and the comparison 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

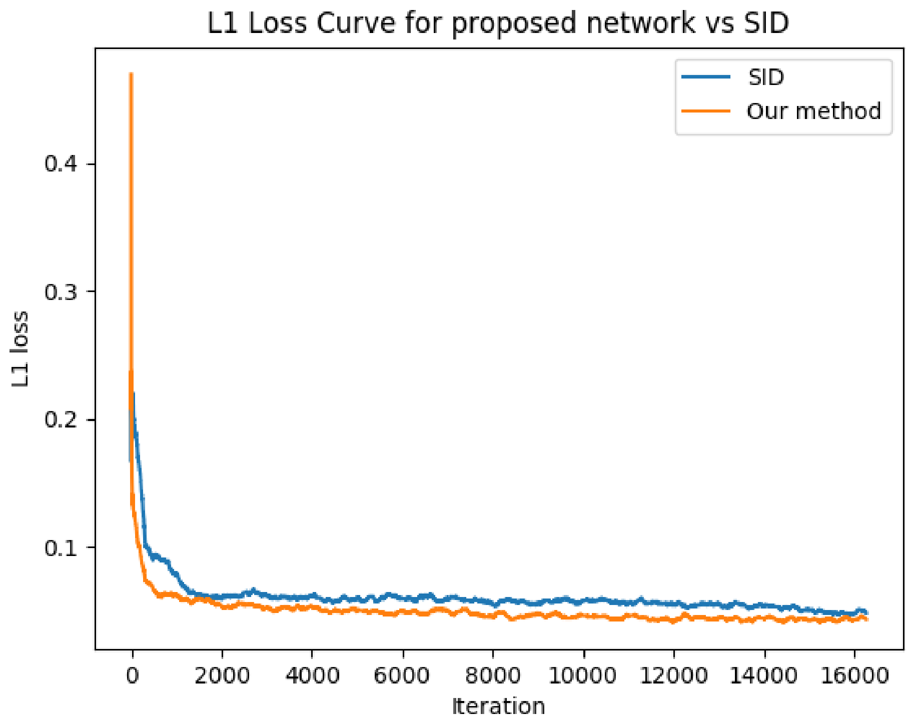

Figure 12: (a) L1 loss curves for our proposed method vs SID for 100 epochs.

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 amplification ratio, and find that our methods has better results for the amplification ratios of x100 and x250. We compared our method with the state-of-the-art SID and traditional approach BM3D. For the fair comparison of BM3D with the deep learning based method we first scaled the dark image by the amplification factor similar to our proposed method and used C-BM3D in sRGB version of low-light image generated by 'Rawpy' similar to the ground truth image. The sigma value of the BM3D is selected such that it generates the highest PSNR value. Table 2 shows the performance for each of the scaling factor in comparison to SID and BM3D.

Table 1: Results 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

4.5 Complexity Analysis

The proposed network architecture in this paper has much less model parameters compared to the neural network architecture of U-Net 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 with 32 residual blocks and the other has 16 residual blocks. On our proposed network with 32 residual block we get around 21x faster processing time, while in another setting with 16 residual block we get 29x faster processing speed with higher PSNR than the SID. In particular, our residual based learning has almost three times less trainable parameters than the U-Net, which allowed to train deeper network in limited amount of GPU resource. Also, the inference time of our network is 0.11 sec for 4K full frame image hence making more practical in real-time

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

application.

CHAPTER 5

CONCLUSION

In this paper we propose a new deep residual learning network with Squeeze-and-Excitation block for denoising and enhancement of extreme low-light images.Our experimental results show that our network has not only 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 deep learning for various vision tasks. Further, we will use a decomposition-based network to divide and conquer the problem. Additionally, we will optimize our work for low-end mobile devices with limited resources and computation power.

REFERENCE LIST

[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)*, pages 3291–3300, 2018.

[2] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *CoRR*, abs/1808.04560, 2018.

[3] Xiaojie Guo. Lime: A method for low-light image enhancement. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, pages 87–91, New York, NY, USA, 2016. ACM.

[4] Shuhang Wang, Jin Zheng, Hai-Miao Hu, and Bo Li. Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, 22, 05 2013.

[5] 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*, pages 2080–2095, Aug 2007.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, 2015.

[7] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

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

[9] Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Weighted nuclear norm minimization with application to image denoising. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2862–2869, 2014.

[10] Michal Aharon, Michael Elad, and Alfred Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. volume 54, pages 4311–4322, Dec 2006.

[11] 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*, pages 479–486, Nov 2011.

[12] 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*, volume 2, pages 860–867, Jan 2005.

[13] Weisheng Dong, Lei Zhang, Guangming Shi, and Xin li. Nonlocally centralized sparse representation for image restoration. volume 22, Dec 2012.

[14] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. volume 26, pages 3142–3155, 2017.

[15] 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.

[16] 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.

[17] 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)*, pages 1874–1883, 2016.

[18] 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*, pages 136–144, Jul 2017.

[19] Zhou Wang, Alan Bovik, Hamid Sheikh, and Eero Simoncelli. Image quality assessment: From error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13:600 – 612, 05 2004.

VITA

Paras Maharjan was born on December 19, 1991 in Lalitpur, Nepal. He completed his Bachelor's Degree in Electronics and Communication Engineering from Kantipur Engineering College (affiliated to Tribhuvan University), Lalitpur, Nepal in 2014. Currently he is working as a Graduate Research Assistant in Multimedia Computing and Communication Lab - UMKC. His research interest involves machine learning, deep learning with its application in computer vision and image processing.