CONTENT PRESERVING SCALE SPACE NETWORK FOR FAST IMAGE RESTORATION FROM NOISY-BLURRY PAIRS

Green Rosh K S *Nikhil Krishnan* *B H Pawan Prasad* *Sachin Lomte*

Samsung R&D Institute India, Bangalore

ABSTRACT

Hand-held photography in low-light conditions presents a number of challenges to capture high quality images. Capturing using a high ISO results in noisy images, while capturing using longer exposure results in blurry images. This necessitates post-processing techniques to restore the latent image. Most existing methods try to estimate the latent image either by denoising or by deblurring a single image. Both these approaches are ill-posed and often result in unsatisfactory results. A few methods try to alleviate this ill-posedness using a pair of noisy-blurry images as inputs. However, most of the methods using this approach are computationally very expensive. In this paper, we propose a fast method to estimate a latent image given a pair of noisy-blurry images. To accomplish this, we propose a deep-learning based approach that uses scale space representation of the images. To improve computational efficiency, we process higher scale spaces using shallower networks and the lowest scale using a deeper network. Also, unlike existing scale-space methods that use bi-cubic interpolation, we propose a content preserving scale space transformation for decimation and interpolation. The proposed method generates state-of-the-art results at reduced computational complexity compared to state-of-the-art method. Finally, we also show that computational efficiency can be improved by 90% compared to baseline with only a marginal drop in PSNR.

Index Terms— deblurring, denoising, deep learning

1. INTRODUCTION

Low-light photography is extremely challenging due to the presence of noise. To reduce the noise, the images can be captured using a longer exposure time. However, this leads to motion blur in the event of handshake or object motion. There has been a plethora of work focused on recovery of latent image either by denoising [1] [2] [3] or by deblurring [4] [5] [6] the observed image. However, these methods use only a single image for image restoration thereby limiting the output quality .

Smartphone photography has made it possible to capture multiple images of the same scene in quick succession. This enables us to gather more information about the latent scene by capturing a pair of images consisting of complementary information: a noisy image captured using high ISO and a blurry image captured using a high exposure. Even though there have been methods such as [7] and [8], the research on image restoration using noisy-blur pairs is still limited. The method proposed by [7] alternates between blur kernel estimation and deconvolution. However, this method uses computationally expensive iterative optimization thus limiting its practical deployability. Recent methods such as LSD2 [8] uses deep learning to recover the latent scene from noisy-blurry pairs. However this method is also computationally very expensive and requires ˜820,000 MAC operations to compute one pixel of output data.

In this paper, we propose a computationally efficient scale-space network architecture for image restoration from noisy-blurry image pairs. The input images are converted into scale-space representations with reduced spatial dimensions, and each of the scales are processed using separate CNNs. Contrary to conventional scale-space methods such as [5] and [9] which uses downsampling operation for scale transformation, we use a *learnable content preserving* scale space transform to efficiently learn a representation of input data at different scale-spaces. To improve computational efficiency, we propose to use lightweight networks to process the higher scales thus shifting the computational burden to the scale with the lowest spatial dimensions. We also propose a lightweight postprocessing network to learn an alpha map that can selectively add details from the noisy image to the output image. The proposed method can produce state-of-the-art results compared to that of baseline LSD2 [8] while having 36% less computations. We also show that the proposed method can further reduce the computational complexity compared to baseline by 90% with a marginal drop in PSNR (1.65%).

2. PROPOSED METHOD

An overview of the proposed method is shown in Fig. 1. Given a noisy (I_n) and a blurry (I_b) observation, we propose to recover the underlying scene in two stages. In the first stage, we make an initial estimate of the latent scene using a novel content preserving scale-space neural network architecture. In the second stage, we further enhance the output using a novel light-weight, input guided neural network.

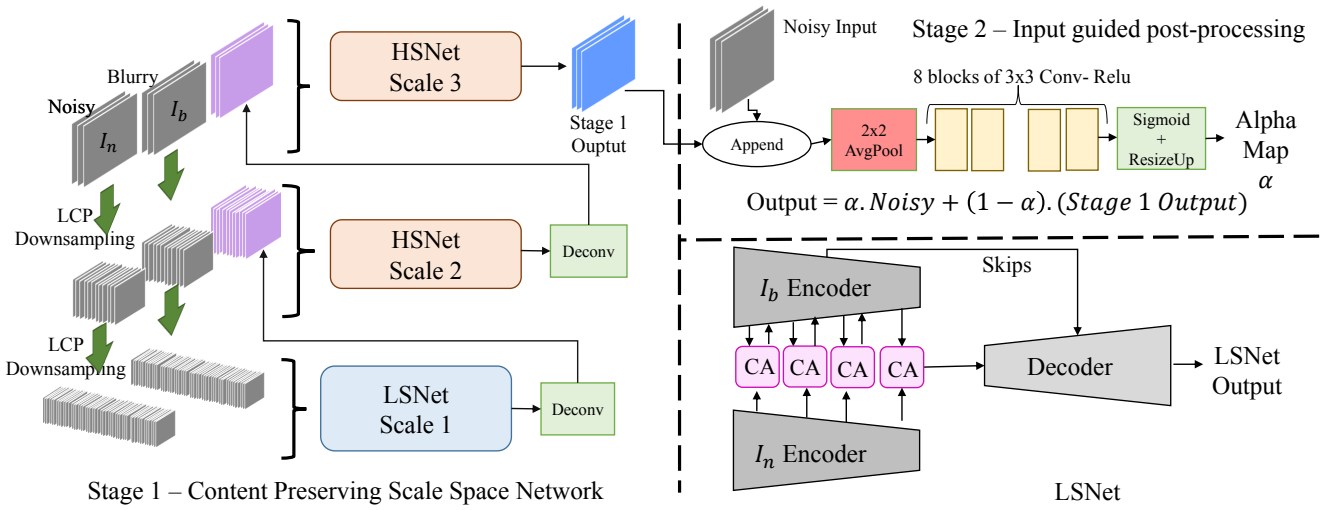

Fig. 1. Network Architecture. A learnable content preserving scale space architecture used to generate initial output followed by a lightweight post-processing stage to generate final output

2.1. Content Preserving Scale Space Network

In this stage, we propose a computationally efficient method to estimate the latent image using a novel scale space network architecture. We convert the input noisy and blurry images to scales of lower spatial resolution using learnable transformations. Each of these scales is then processed by separate CNNs. The output from each scale is transformed to higher spatial dimensions and is provided as input to the next scale. Specifically, we use three scales to process the input images. The salient features of our method is explained in the following subsections.

2.1.1. Learnable Content Preserving (LCP) Scale Transformation

Scale-space methods such as [5] [9] uses downsampling operations to reduce spatial dimensions at each scale. However, in this approach, every downsampling operation removes three-fourths of the data available in the higher scale. This severely diminishes the amount of data available at the lowest scale thus impacting the restoration quality. To address this issue, we propose to use *Learnable Content Preserving* (LCP) transformation to change scales of the data. Specifically, given a $(H \times W \times N)$ image, we use depth separable filters to transform to a $(\frac{H}{2} \times \frac{W}{2} \times 4N)$ dimensional space for lower scale transformation. Similarly, deconvolutional filters are used to learn a higher scale transformation, where a $(H \times W \times N)$ dimensional input data gets transformed to a $(2H \times 2W \times \frac{N}{4})$ dimensional data. By using LCP transformations, the network is able to learn efficient data representations during scale transformation, which improves the restoration quality at each scale. This allows us to use shallower networks at higher scales to improve computational efficiency. Further, since this transformation affect only the input and output number of channels, the computational overhead at each scale is marginal (Table 3).

2.1.2. Variable Complexity Network Architectures

We reduce the spatial resolution by half at every scale resulting in a complexity reduction of 4 times at every scale. Hence the network complexity can significantly be reduced by using shallower networks at higher scales. To this end, we propose to use two different architectures of different complexities: *Low Scale Network (LSNet)* and *High Scale Network (HSNet)*.

LSNet learns to recover the latent scene at the lowest sale using LCP scale transformed data. We follow a multi-branch encoder-decoder architecture similar to methods such as [10] and [11] as shown in Fig. 1. The scale transformed noisy and blur images are processed using separate encoders. The features extracted from the two encoders after every convolutional block are appended together and then passed through a channel attention module [12]. The merged features after the encoding is passed through a decoder to generate the output. We also use spatial attention modules [12] after every convolutional block in the blur encoder to address spatially varying blur.

All the higher scales are processed using HSNet, which is shallower than LSNet. The major task of HSNet is to take the LCP up-scaled data from the previous scale and to add details from the inputs at its corresponding scale. The scale transformed noisy and blurry data along with the output from previous scale are appended together and passed to HSNet. We use a modified version of UNet [13] with significantly reduced number of filters for our HSNet. The proposed HSNet

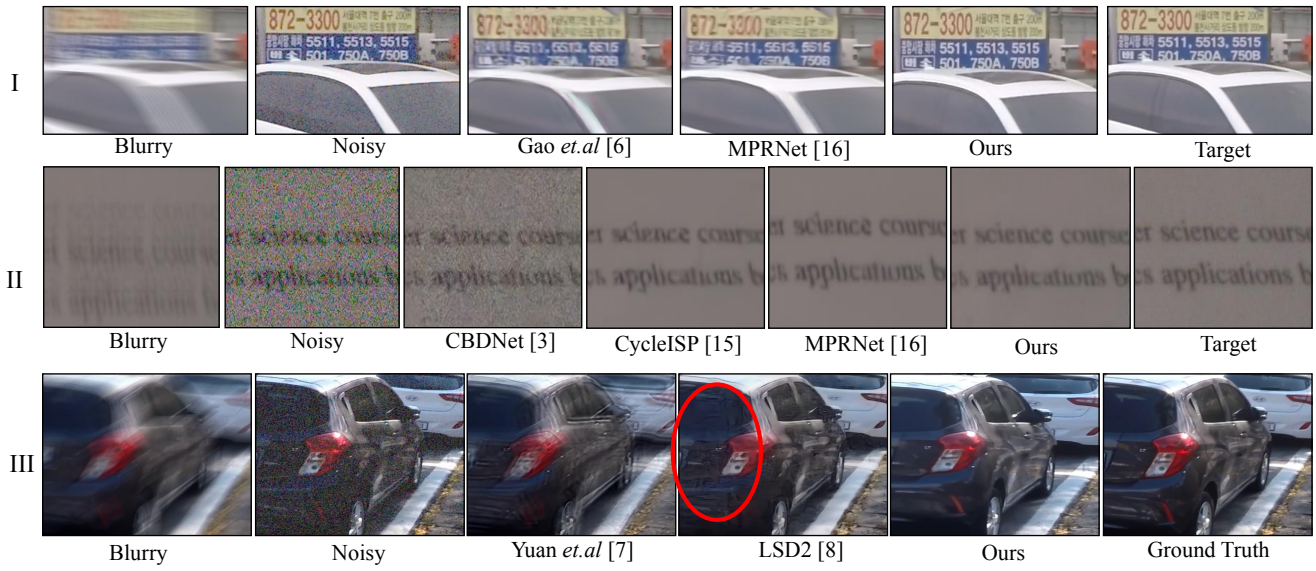

Fig. 2. Comparisons against state of the art methods. I: Single Image Deblurring methods (GOPRO); II: Single Image Denoising methods (SIDD); III: Joint Denoise-Deblur methods. Note the distortions in text regions and red circle

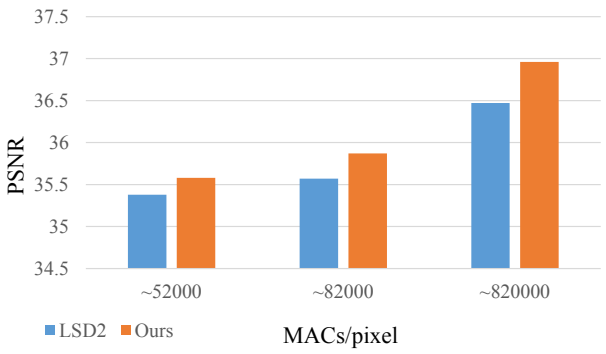

Fig. 3. Complexity Analysis. Our method outperforms LSD2 [8] at all complexity levels.

needs only 3.5% MACs per pixel compared to original UNet.

2.2. Input Guided Post Processing

Image denoising techniques often lead to loss in texture or details [17], since the network sometimes fail to distinguish between fine texture and noise, both of which contain high frequency information. To recover the lost details, we use an input guided post-processing stage where an alpha map is learned to blend the noisy input image and the output from the first stage (I_o^1) (Fig 1). I_n and I_o^1 are appended together and passed through a 2×2 average pooling operation to reduce the spatial dimensions for faster processing. The resulting tensor is passed through 8 convolutional blocks. The output from the last convolution layer is passed through a sigmoid activation layer to limit the range of values to [0,1]. The output, which is a single channel image is then bilinear upsampled, to get the alpha map (α). The final output can then be obtained as follows:

$$I_o = \alpha.I_n + (1-\alpha).I_o^1 \quad (1)$$

3. EXPERIMENTS AND RESULTS

3.1. Training Details

To train the proposed network, a dataset consisting of blurry, noisy and clean observations of the same scene is required. We use an approach wherein real blur/noise dataset is augmented using synthetic noise/blur to create a *partially synthetic dataset*. Specifically we use two kinds of datasets: a) Real blur and synthetic noise and b) Real noise and synthetic blur.

We use GOPRO [9] dataset to create our training dataset with real blur. We add synthetic noise to the ground-truth images using the method by [3] to get a (noisy, blurry, clean) training pair. To create the second type of dataset with real noise, we use SIDD [18] dataset. We add synthetic blur to the clean groundtruth of this dataset using the method proposed by [19].

We train the network end-to-end using a mix of L1 and perceptual losses. The network is trained using Adam optimizer with an initial learning rate of 10^{-4} for around 2 million iterations.

3.2. Comparison against state-of-the-art

In Fig. 2, We compare our approach against state-of-the-art methods in single image deblurring (Gao et.al [6], MPRNet [16])), single image denoising (CBDNet [3], CycleISP [15],

Table 1. Comparisons on SIDD validation dataset (PSNR (dB))

Blurry	Noisy	BM3D [1]	FFDNet [14]	CBDNet [3]	RIDNet [2]	CycleISP [15]	LSD2 [8]	Ours
26.31	22.8	25.78	29.20	30.78	38.71	39.52	39.61	**40.17**

Table 2. Comparisons on GOPRO test dataset (PSNR (dB))

Blurry	Noisy	SRN [5]	Gao [6]	Suin [4]	MPRNet [16]	Ours (Lightweight)	LSD2 [8]	Ours
25.64	21.96	30.26	30.90	32.02	32.66	35.87	36.47	**36.96**

Table 3. Ablation Studies on Network Architecture

Method	MACs/% increase	PSNR
Bicubic scale-space	77283 / -	35.4
Learnable scale-space	77776 / 0.6%	35.55
LCP scale-space	80002 / 2.8%	35.72
LCP scale-space + postprocess	82024 / 2.5%	35.87

MPRNet [16]) and joint deblurring/denoising (Yuan et.al [7], LSD2 [8]). In regions with high blur, single image deblurring methods (Row I) generates artifacts such as color distortions (Gao et.al) and structural distortions (MPRNet - text regions). The proposed method is able to recover details better in these regions with the help of information from the noisy image. Row II shows an example against single image denoising methods. While CBDNet is unable to remove heavy noise, CycleISP and MPRNet generates distorted output in text regions. It can be observed that our method is able to generate clean output without distortions even with a heavily blurred complementary image. Finally, comparisons against joint denoise-deblur methods are provided in Row III. It can be seen that Yuan et.al [7] fails to deblur images with spatially varying blur and LSD2 [8] generates distortions in the final output. Whereas our method is able to generate results without blur or distortions.

We also provide quantitative comparisons of the proposed method against the state-of-the-art methods. In Table 1, the proposed method is compared against denoising methods on SIDD test dataset and in Table 2, the proposed method is compared against deblurring methods on GOPRO test dataset. It can be observed that the proposed method obtains the best PSNR scores on both the datasets against all the other methods.

3.3. Studies

We performed a study to analyze the computational efficiency of the proposed method against a baseline LSD2 [8] network. We trained both LSD2 and our method at three different complexity levels. The complexity of the networks are altered by changing the number of filters in each convolutional block. All the networks are trained using GOPRO dataset for the same number of iterations. The results are summarized in Fig. 3. It can be seen that the proposed method always outperforms LSD2 for the same complexity level. Further it can be seen that the proposed method is able to generate similar PSNR scores as that of LSD2 at 36% reduced computations. Moreover, the proposed method is able to reduce the computational complexity by 90% compared to that of LSD2, while the PSNR drops only by 1.65%.

We also conducted an ablation study to analyze the impact of various components used in our network. We evaluated four variations of the proposed method: a) Scale space architecture with bicubic interpolation for scale transformation; b) Learnable filters for scale transformation; c) Learnable and content preserving (LCP) filters for scale transformation and d) LCP scale transformation followed by input guided post-processing. These results are summarized in Table 3. It can be seen that the proposed network architecture results in a net PSNR gain of 0.5 dB compared to a multi-scale approach which uses bicubic down/up sampling for scale transformation, at a marginal increase in total computations.

4. CONCLUSIONS

In this paper, we propose a novel deep learning based method to estimate a latent image using a pair of observations corrupted by motion blur and noise respectively. We develop a fast method using computationally efficient scale-space representation of input data. Contrary to other scale-space methods in image restoration, we propose a novel content preserving scale transformation wherein the number of channels of the new scale space is changed relative to change in spatial dimensions. To make the network computationally efficient, we use shallow networks in all the higher scales, thus shifting the computational burden to the lowest scale which has reduced spatial dimensions. Further we also use a novel light-weight post processing network to learn an alpha map for recovering details from noisy input image. We provide qualitative and quantitative results to show the superiority of our approach compared to state-of-the-art methods in image denoising and deblurring. Further, we also show that our method achieves similar psnr scores as that of baseline network used by LSD2, with 36% reduced computation. Our method also improve computational efficiency by 90% with a marginal drop in PSNR compared to baseline.

5. REFERENCES

[1] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Transactions on image processing*, vol. 16, no. 8, pp. 2080–2095, 2007.

[2] Saeed Anwar and Nick Barnes, "Real image denoising with feature attention," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3155–3164.

[3] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang, "Toward convolutional blind denoising of real photographs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1712–1722.

[4] Maitreya Suin, Kuldeep Purohit, and AN Rajagopalan, "Spatially-attentive patch-hierarchical network for adaptive motion deblurring," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3606–3615.

[5] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia, "Scale-recurrent network for deep image deblurring," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8174–8182.

[6] Hongyun Gao, Xin Tao, Xiaoyong Shen, and Jiaya Jia, "Dynamic scene deblurring with parameter selective sharing and nested skip connections," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3848–3856.

[7] Lu Yuan, Jian Sun, Long Quan, and Heung-Yeung Shum, "Image deblurring with blurred/noisy image pairs," in *ACM SIGGRAPH 2007 papers*, pp. 1–es. 2007.

[8] Janne Mustaniemi, Juho Kannala, Jiri Matas, Simo Srkk, and Janne Heikkil, "Lsd2 - joint denoising and deblurring of short and long exposure images with cnns," in *The 31st British Machine Vision Virtual Conference (BMVC)*, September 2020.

[9] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3883–3891.

[10] Green Rosh KS, Anmol Biswas, Mandakinee Singh Patel, and BH Pawan Prasad, "Deep multi-stage learning for hdr with large object motions," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 4714–4718.

[11] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang, "Deep high dynamic range imaging with large foreground motions," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 117–132.

[12] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[14] Kai Zhang, Wangmeng Zuo, and Lei Zhang, "Ffdnet: Toward a fast and flexible solution for cnn-based image denoising," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4608–4622, 2018.

[15] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao, "Cycleisp: Real image restoration via improved data synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2696–2705.

[16] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao, "Multi-stage progressive image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14821–14831.

[17] Matteo Maggioni, Yibin Huang, Cheng Li, Shuai Xiao, Zhongqian Fu, and Fenglong Song, "Efficient multi-stage video denoising with recurrent spatio-temporal fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3466–3475.

[18] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown, "A high-quality denoising dataset for smartphone cameras," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1692–1700.

[19] Shuochen Su and Wolfgang Heidrich, "Rolling shutter motion deblurring," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1529–1537.