Down-Scaling with Learned Kernels in Multi-Scale Deep Neural Networks for Non-Uniform Single Image Deblurring

Dongwon Park1  Jisoo Kim1  Se Young Chun
Ulsan National Institute of Science and Technology (UNIST)
UNIST-gil 50, Ulsan, Republic of Korea
{dong1,rlawltntn1053,sychun}@unist.ac.kr, 1equal contribution.

Abstract

Multi-scale approach has been used for blind image / video deblurring problems to yield excellent performance for both conventional and recent deep-learning-based state-of-the-art methods. Bicubic down-sampling is a typical choice for multi-scale approach to reduce spatial dimension after filtering with a fixed kernel. However, this fixed kernel may be sub-optimal since it may destroy important information for reliable deblurring such as strong edges. We propose convolutional neural network (CNN)-based down-scale methods for multi-scale deep-learning-based non-uniform single image deblurring. We argue that our CNN-based down-scaling effectively reduces the spatial dimension of the original image, while learned kernels with multiple channels may well-preserve necessary details for deblurring tasks. For each scale, we adopt to use RCAN (Residual Channel Attention Networks) as a backbone network to further improve performance. Our proposed method yielded state-of-the-art performance on GoPro dataset by large margin. Our proposed method was able to achieve 2.59dB higher PSNR than the current state-of-the-art method by Tao. Our proposed CNN-based down-scaling was the key factor for this excellent performance since the performance of our network without it was decreased by 1.98dB. The same networks trained with GoPro set were also evaluated on large-scale Su dataset and our proposed method yielded 1.15dB better PSNR than the Tao’s method. Qualitative comparisons on Lai dataset also confirmed the superior performance of our proposed method over other state-of-the-art methods.

1. Introduction

Images and videos are often blurred by various reasons: unwanted camera shake, moving objects, depth variation in the scene. A typical forward model to generate a blurred image $y \in \mathbb{R}^N$ with $N$ number of pixels is

$$y = Gx + n$$

where $G \in \mathbb{R}^{N \times N}$ is a unknown large sparse matrix whose rows contain local blur kernels, $x \in \mathbb{R}^N$ is a unknown latent ground truth image that one would like to estimate, and $n \in \mathbb{R}^N$ is noise. The goal of blind single image deblurring is to recover the values of $G$ and $x$ for the given $y$. Thus, it is a challenging inverse problem that is severely ill-posed.

There have been much effort to tackle blind single image / video deblurring. One is to simplify (1) by assuming uniform blur and to recover both the latent ground truth image $x$ and the blur kernel matrix $G$ by solving a tractable opti-
mization problem based on \([1] [8] [32] [3] [38]\). However, uniform blur is often not accurate enough to approximate the actual blur due to, for example, camera rotation and movements outside the sensor plane or moving objects in the scene with different distances from the camera. Thus, there have been much research on removing non-uniform blur by extending the degree of freedom of the blur model from uniform blur to non-uniform blur in a limited way compared to the dense matrix \(G\) in \([1] [11] [10] [26] [13] [40] [28]\). Other non-uniform blur models have been investigated such as additional segmentations within which simple blur models were used \([4] [14]\) or motion estimation based deblurs \([15] [16]\).

Recently, deep-learning-based approaches for single image / video blind deblurring have been proposed with excellent quantitative results and with fast computation time. There are largely two different ways of using deep neural networks for deblurring. One is to use neural networks to explicitly estimate non-uniform blur kernels \([41]\) or motion estimation based deblurs \([15] [16]\). Our proposed method used our down-scaling with learned kernels and RCAN, employed sub-optimal modular training approach instead of end-to-end training for deeper network with limited hardware resource, and removed the deblur network to process the original blurred image in the multi-scale network without compromising performance.

2. Related Works

Conventional approaches to blind single image / video deblurring usually require to explicitly estimate blur kernels. There have been several works on estimating uniform blurs using optimization algorithm with coarse-to-fine multi-scale approach \([8]\), using a model of the spatial randomness of noise and a local smoothness prior \([32]\), exploiting blurred strong edges to reliably estimate blur kernel \([8]\), and developing a metric to measure the usefulness of image edges for blur kernel estimation \([38]\).

There have also been many works on predicting non-uniform blurs assuming spatially linear blur \([11]\), simplified camera motion (from 6D to 3D) \([10]\), parametrized geometric model in terms of camera rotation velocity during exposure \([36]\), filter flow framework based blur model \([13]\), L0 sparse expression for blurs \([40]\), and dark channel prior \([28]\). There was also an attempt to exploit multiple images from videos assuming spatially varying blur \([24]\).

There have also been some works to utilize segmentation information by assuming uniform blur on each segmentation area \([4]\) and to segment motion blur using convex optimization \([14]\), to simplify motion model as local linear without segmentation using coarse-to-fine approach \([15]\), and to use bidirectional optical flows for video deblurring \([16]\).

Since the advent of deep learning \([22]\), many blind single image / video deblurring works employed deep neural networks for estimating blur kernels and/or original sharp images from given blurred input images. There are several works to predict non-uniform blur kernels explicitly: predicting the probabilistic distribution of motion blur at the patch level \([34]\), estimating the complex Fourier coefficients of a deconvolution filter \([2]\), performing blur kernel estimation by division in Fourier space from extracted deep features \([31]\), and analyzing the spectral content of blurry image patches by reblurring them \([1]\).

There are also many works to directly estimate the original sharp image from the given blurred input image without explicitly estimating non-uniform blur kernels. For
video blind deblurring, there have been some works to exploit temporal information: blending temporal information in spatio-temporal recurrent network for online video deblurring [17], taking temporal information into account with recurrent deblur network consisting of several deblur blocks [37], and developing an encoder-decoder network with the input of multiple video frames to accumulate information across frames [33]. There are a few works for blind single image deblurring without temporal information. Xu et al. proposed a direct estimation method of the original sharp image based on conventional optimization to approximate deconvolution by a series of convolution steps using deep neural networks [39]. Later, Nah et al. proposed a multi-scale network architecture with Gaussian pyramid and multi-scale loss functions [27] and Tao et al. proposed convolution long short-term memory (LSTM)-based multi-scale deep neural network for single image deblurring [35].

Multi-scale approaches for single image / video deblurring have two different types of down-scaling. One is a simple filtering & down-sampling operation so that local information will be encoded with reduced spatial dimension [8, 3, 27, 35]. The other is a deep neural network based global information encoding of multiple video frames (e.g., 5) with much further reduced spatial dimension (up to \(\times 8\)) and increased channels (up to 512) [33]. There has been no work on learning based down-scaling to encode local information as an extension of Gaussian / bicubic downsampling in multi-scale single image deblurring.

3. Method

3.1. Proposed Network Architecture

Figure 2 illustrates the overall architecture of our proposed deep neural network model for single image deblurring. While conventional multi-scale approaches [27, 35] are using Gaussian / bicubic downsampling with a fixed kernel to reduce the spatial dimension of the input blurred image, we propose to use a CNN-based downsampling with learned kernels to preserve necessary high-frequency information for deblurring. Our network firstly down-scales the input color image with the size of \(W \times H \times 3\) to the feature maps with the size of \(W/4 \times H/4 \times 64\) using the “Down-Scaling 1” module in Figure 2(b). Then, these feature maps are fed into the backbone network as well as residual in residual (RIR) skip connection to yield initial deblurred feature maps in the spatial dimension of \(W/4 \times H/4\). Then, upsampling and CNN will result in the intermediate deblurred image estimate with the size of \(W/2 \times H/2 \times 3\). This result is combined with the down-scaled blurred image in the “Down-Scaling 2” module in Figure 2(b) for the deblurring at the scale of \(W/2 \times H/2\) using another backbone network, RIR, and up-scaling process to yield the final deblurred output image with the size of \(W \times H\). Unlike other multi-scale approaches for single image deblurring [27, 35], our proposed network does not process the original blurred image at the finest scale so that less parameters are required to be used and computation complexity is significantly reduced. We empirically confirmed that adding the module to directly process the original blurred image did not help to improve the overall performance.

3.2. Proposed Down-Scaling with Learned Kernels

Gaussian or Bicubic down-sampling is a simple, effective method to reduce the spatial dimension of an image with a fixed kernel. For the input image with the size of \(W \times H \times 3\), bicubic \(\times 2\) down-sampling yields \(W/2 \times H/2 \times 3\) image. However, due to the fixed kernel, some important high-frequency information for deblurring may be removed.
to avoid aliasing. For example, it is well-known that strong edges help to reliably estimate blur kernels \([3, 38]\), thus it may be beneficial to preserve some of these high-frequency details even in down-scaled information for deblurring.

We propose CNN-based down-scaling modules instead of using conventional down-sampling. For \(\times 4\) down-scaling, our proposed module consists of convolution layer - ReLU - convolution layer where each convolution layer has 64 channels and stride 2 as illustrated in Figure 2 (b). For \(\times 2\) down-scaling in multi-scale architecture, our proposed architecture consists of convolution layer with stride 2 - concatenation with the output from the lower scale deblurring network - ReLU - convolution later with stride 1 as illustrated in Figure 2 (b). All convolution layers have \(3 \times 3\) filters that have similar sizes as bicubic kernel.

Unlike conventional down-sampling after anti-aliasing filtering with a deterministic kernel, our proposed CNN-based down-scaling has learned kernels with 64 channels and spatial dimension reduction using stride 2. Thus, our proposed method seems to keep the effect of spatial dimension reduction for multi-scale approach while to be more optimal than anti-aliasing filtering (essentially low-pass filtering). Figure 3 illustrates one example of our CNN-based down-scaling. Our proposed down-scaling with learned kernels is also different from the work of \([33]\) and others that are using a deep neural network based global information encoding with further reduced spatial dimension and increased channels. Note that the work of \([33]\) and others do not yield any intermediate deblurred image estimate at low scales, have much larger receptive field size than bicubic kernels so that less local information is preserved, and have increased channels for global information encoding so that it can not be seen as an extension of conventional bicubic down-sampling.

3.3. Residual Channel Attention Network

A backbone network architecture is often shared among single image deblurring and single image super resolution. For example, the work of Nah \([27]\) for deblurring and the work of Lim (EDSR) \([25]\) for super resolution utilized the same residual block as their backbone networks. Recently, residual channel attention network (RCAN) was proposed for super resolution and yielded state-of-the-art results \([41]\). We adopt RCAN as our backbone network for single image deblurring. RCAN contains a small number of network channels, but had a large receptive field due to deep architecture. RCAN consists of a number of residual groups with long skip connections (LSC) and each residual group consists of multiple residual channel attention blocks with short skip connections (SSC). RCAN with large receptive field was able to be well-trained due to SSC as well as LSC.

Figure 4 illustrates the differences among the original residual blocks of ResNet \([12]\) for high-level computer vision task such as image classification, a modified residual blocks in EDSR \([25]\) and the work of Nah \([27]\) for low-level computer vision tasks such as super resolution and image deblurring, and the residual channel attention blocks for super resolution \([41]\). Note that there is no batch normalization layer for super resolution and image deblurring. RCAN additionally combined residual blocks with channel attention mechanism that generated global context information through global average pooling. This information combined with the output of residual block through element-wise product to yield the final output as in Figure 4 (c). Recent image super resolution methods \([25, 23]\) including RCAN are using pixel shuffle for up-scaling due to low computation complexity and high performance. We also adopt pixel shuffle for our up-scaling modules.

3.4. Loss Function

Recently, a mix loss was proposed to combine a conventional L1 loss and a multi-scale SSIM (structural similarity) loss to yield shaper images than L1 loss only \([42]\). We used this mix loss for training both our coarse-scale sub-network
as well as our fine-scale sub-network as follows:

$$L_{tot}(\theta_c, \theta_f) = \lambda_{L1}L_{L1}(\theta_c, \theta_f) + \lambda_{SS}L_{SS}(\theta_c, \theta_f)$$  \hspace{1em} (2)$$

where $\theta_c, \theta_f$ are deep neural network parameters for the coarse-scale sub-network, the fine-scale sub-network, respectively. The L1 loss is defined as

$$L_{L1}(\theta_c, \theta_f) = \|I^{GT} - I^{out}\|_1$$

where $I^{GT}$ is a ground truth image and $I^{out}$ is a predicted blurred image, $h^c, h^f$ are coarse-scale, fine-scale deep neural networks, respectively. The multi-scale SSIM loss is defined as

$$L_{SS}(\theta_c, \theta_f) = 1 - \sum_{m=0}^{2} w_m S(D^m_{ap}(I^{out}, D^m_{ap}(I^{GT}))$$

Then, we finally trained the top network $h^f$ in Figure 2 (a) with the original ground truth image as well as the output of the bottom sub-network $h^c$ in Figure 2 (a) by minimizing $2$ with the fixed $\theta_c$ or $L_{tot}(\theta_c, \theta_f)$. Thus, our final trained network is the optimal fine-scale sub-network (the top network in Figure 2 (a)) for the fixed, sub-optimal coarse-scale sub-network (the bottom network in Figure 2 (a)). Due to our modular approach, we were able to train the network that is effectively 2 times deeper than the original RCAN [41].

4. Experimental Setup

4.1. Datasets

GoPro dataset [27] for deblurring was used to train our proposed model. This dataset consists of 3,214 blurred images with the size of 1,280×720 that are divided into 2,103 training images and 1,111 test images. We also used Köhler dataset [19] that contains 48 blurred images with the size of 800×800 for testing. We further evaluated our proposed methods and other state-of-the-art methods on Su dataset [33]. Su dataset consists of 71 videos (6,708 images) with the size of 1,920×1,080 or 1,280×720 from multiple devices including iPhone 6s, GoPro Hero 4, and Canon 7D. We used all videos for test. Lastly, we performed qualitative performance comparisons on Lai dataset [21] whose image sizes are varying within 351-1,024×502-1,024.

4.2. Training Details

We used PyTorch [29] for all our implementations and MATLAB for all PSNR/SSIM evaluation. NVIDIA Titan V GPU was used for training and testing. Adam optimizer [18] was used with learning rate= $0.5 \cdot 10^{-5}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. Total epoch number was 1,000 with reducing learning rate by half every 300 epochs. No pre-training was performed for RCAN backbone network. Patch based training was performed with the sizes of 192×192 at ×4 scale and 96×96 at ×2 scale with data augmentation using random crop, horizontal flip, and 90 rotation.

5. Experimental Results

5.1. Down-Scaling in Single-Scale Architecture

First of all, we investigated the effect of our down-scaling approach for different scale levels in single-scale architecture. We took out the sub-network $h^c(I^{in}; \theta_c)$ or

| Measure | PSNR | SSIM | Time |
|---------|------|------|------|
| $\times 1 \to \times 1 \to \times 1$ | 27.44 | 0.8984 | 21.2 s |
| $\times 1 \to \times 2 \to \times 1$ | 29.37 | 0.9275 | 4.0 s |
| $\times 1 \to \times 4 \to \times 1$ | 31.40 | 0.9501 | 1.0 s |

Table 1: Effects of down-scaling at different scales.
Figure 5: An example for the effects of down-scaling at different scales. (a) input blurred image. (b) \( \times 1 \rightarrow \times 1 \rightarrow \times 1 \). (c) \( \times 1 \rightarrow \times 2 \rightarrow \times 1 \). (d) \( \times 1 \rightarrow \times 4 \rightarrow \times 1 \).

the bottom network in Figure 2 (a) and modified to yield an image with the original size instead of the \( \times 2 \) down-scaled size. Three different configurations are as follows: 1) \( \times 1 \rightarrow \times 1 \rightarrow \times 1 \) changed stride 2 to 1 in CNNs with no up-scaling, 2) \( \times 1 \rightarrow \times 2 \rightarrow \times 1 \) changed stride 2 to 1 in the second CNN, and 3) \( \times 1 \rightarrow \times 4 \rightarrow \times 1 \) changed up-scaling to be \( \times 4 \). Table 2 summarizes the quantitative results on GoPro test set. Surprisingly, more down-scaling yielded better PSNR / SSIM and faster computation time. There are a few explanations for these results such as well-preserved high-frequency details even up to \( \times 4 \) scale, much larger receptive field for \( \times 1 \rightarrow \times 4 \rightarrow \times 1 \), and reduced blur effect in lower spatial dimensions. Figure 5 also shows qualitative results for this investigation, clearly visualizing superior results of \( \times 1 \rightarrow \times 4 \rightarrow \times 1 \) over others.

5.2. Ablation Study

Compared to Nah [27], our proposed method contained our CNN-based down-scaling with learned kernels, new backbone network using RCAN [41] and no \( \times 1 \rightarrow \times 1 \rightarrow \times 1 \) path in the multi-scale architecture. In order to show the contribution of each component, we performed ablation study: RCAN vs. EDSR [25] for backbone network, CNN-based down-scaling vs. bicubic down-scaling, and removing \( \times 1 \rightarrow \times 1 \rightarrow \times 1 \) vs. keeping \( \times 1 \rightarrow \times 1 \rightarrow \times 1 \).

Table 2 presents PSNR, SSIM, computation time (Time), and parameters size (Param) for different components such as backbone network (Network), CNN-based down-scaling (denoted by C), and \( \times 1 \)-scale network (denoted by 1).

Firstly, RCAN (b) yielded substantially better PSNR and SSIM over a modified EDSR (a) that was also used in the work of Nah et al. [27]. Note that EDSR requires 80 million (M) parameters while RCAN does substantially smaller parameters (32M), but EDSR had faster computation time than RCAN possibly due to shallower depth that can utilize parallel GPU computing more efficiently.

Secondly, our proposed CNN-based down-scaling yielded significantly better PSNR (1.98dB up) and better SSIM (0.017 up) than bicubic down-sampling as in Table 2 (d) and (b). Thus, our CNN-based down-scaling is the key factor to improve the overall performance of our proposed single image deblurring among all new components.

Lastly, we investigated the role of \( \times 1 \) path in our proposed method. Table 2 (e) shows that this original scale path did not improve quantitative results with substantially increased computation time and parameter size when using our CNN-based down-scaling. However, as shown in Table 2 (c), it was important to use \( \times 1 \) path to improve the quality of deblurred images substantially when using conventional down-sampling. Thus, we can argue that our CNN-based down-scaling well-preserved necessary high-frequency details at lower scales compared to bicubic down-sampling so that the original input image at the fine scale did not help much to improve the quality of deblurred images.

5.3. Benchmark Results

We performed comparison studies on a few open datasets such as GoPro, Köhler, and Su. Tables 3 and 4 present quantitative results of our proposed method and other state-

Table 2: Ablation study results on GoPro dataset

| Network  | C | 1  | PSNR  | SSIM  | Time  | Param |
|----------|---|----|-------|-------|-------|-------|
| (a) EDSR | x | x  | 29.94 | 0.934 | 2.8s  | 80M   |
| (b) RCAN | x | x  | 30.87 | 0.945 | 8.2s  | 32M   |
| (c) RCAN | x | o  | 32.88 | 0.962 | 14.6s | 48M   |
| (d) RCAN | o | x  | 31.23 | 0.948 | 28.6s | 48M   |
| (e) RCAN | o | o  | 29.94 | 0.934 | 2.8s  | 80M   |

Table 3: Quantitative results on GoPro and Köhler datasets.

Ours is our network trained with GoPro training dataset.

|                  | GoPro  | Köhler |      |      |
|------------------|--------|--------|------|------|
|                  | PSNR   | SSIM   | PSNR | MSSIM|
|                  |        |        |      |      |
| Xu [40]          | 25.10  | 0.890  | 27.47| 0.811|
| Kim [15]         | 23.64  | 0.824  | 24.68| 0.794|
| Sun [34]         | 24.64  | 0.843  | 25.22| 0.774|
| Gong [9]         | 27.19  | 0.908  | 26.59| 0.808|
| Ram. [30]        | 28.94  | 0.922  | 27.08| 0.812|
| Nah [27]         | 29.08  | 0.914  | 26.48| 0.808|
| Kupyn [20]       | 28.70  | 0.958  | 26.10| 0.816|
| Tao [35]         | 30.26  | 0.934  | 26.75| 0.837|
| Ours             | 32.85  | 0.962  | 26.08| 0.810|

Table 4: MSSIM metric over a modified EDSR (a) that was also used in the work of Nah et al. [27]. Note that EDSR requires 80 million (M) parameters while RCAN does substantially smaller parameters (32M), but EDSR had faster computation time than RCAN possibly due to shallower depth that can utilize parallel GPU computing more efficiently.

Table 2 (c), it was important to use \( \times 1 \) path to improve the quality of deblurred images substantially when using conventional down-sampling. Thus, we can argue that our CNN-based down-scaling well-preserved necessary high-frequency details at lower scales compared to bicubic down-sampling so that the original input image at the fine scale did not help much to improve the quality of deblurred images.

5.3. Benchmark Results

We performed comparison studies on a few open datasets such as GoPro, Köhler, and Su. Tables 3 and 4 present quantitative results of our proposed method and other state-
Figure 6: (a) Input blurred images. (b) Results of Tao et al. [35]. (c) Results of Nah et al. [27]. (d) Results of our proposed method. The results on the 1st-2nd are from GoPro dataset [27], those on the 3rd-4th rows are from Su dataset [33], and those on the 5th row are from Lai dataset [21]. Our proposed method deblurred fine details clearly compared to other state-of-the-art methods on all three datasets.
of-the-art methods. Our proposed method trained with GoPro training dataset yielded state-of-the-art quantitative results over previous state-of-the-art deblurring methods such as Tao [35] in PSNR and Kupyn [20] in SSIM on GoPro test dataset (1,111 images). Our proposed method yielded 2.59dB better PSNR and 0.028 better SSIM than Tao and also yielded 4.15dB better PSNR and 0.004 better SSIM than Kupyn on GoPro test set.

Our method trained with GoPro dataset was also evaluated with Köhler dataset. As shown in Table 3, our method was able to achieve comparable results to the work of Kupyn [20], but lower than Tao [35]. In fact, recent state-of-the-art methods were not able to perform well with Köhler dataset in terms of PSNR compared to early work of Xu [40]. Note that even though Köhler dataset contains 48 blurred images, they were actually generated by using 12 different blurs for 4 different images. Thus, the results with this dataset could be severely biased for these 4 images and it may explain rather inconsistent improvements of previous methods between GoPro and Köhler datasets.

For fair evaluations, we further evaluated our method as well as two other state-of-the-art methods by Nah and Tao with all images in Su dataset. The codes that were provided by the authors were used. As shown in Table 5, our proposed method yielded significantly better PSNR / SSIM than the work of Nah and the work of Tao. These results are consistent with the results for GoPro test dataset even though Su dataset contains data from other devices.

Figure 6 shows qualitative comparisons for five different examples from GoPro dataset (1-2 rows), Su dataset (3-4 rows) and Lai dataset (5 row) using our proposed method (d column) as well as the work of Nah (c column) and the work of Tao (b column) for given input blurred images (a column). Visual comparisons clearly show the superior performance of our proposed method over previous state-of-the-art methods, especially for fine details such as texts, numbers, walking people, and moving cars.

6. Discussion

We propose a CNN-based down-scaling method with learned kernels for multi-scale deep neural networks in single image deblurring. Unlike other multi-scale deblurring methods with Gaussian / bicubic down-sampling, our proposed down-scaling seems to preserve necessary high-frequency details such as strong edges for reliable and high performance deblurring even at reduced spatial dimension.

Table 4: Quantitative results on Su dataset for the networks trained with GoPro training dataset.

|          | Nah [27] | Tao [35] | Ours  |
|----------|----------|----------|-------|
| PSNR     | 30.34    | 30.93    | 32.08 |
| SSIM     | 0.917    | 0.931    | 0.940 |

Figure 7: Decomposition study. (a) Input blurred image. (b) Output deblurred image. (c) Output without using backbone networks. (d) Output without RIR skip connections.

Our scheme also allows us to use no network processing the original scale data with negligible performance degradation so that computation complexity is not increased. Our proposed method yielded state-of-the-art performance on large-scale datasets such as GoPro and Su. Our method also yielded visually pleasing deblurring results on Lai dataset compared to other state-of-the-art methods.

Even though our proposed method yielded similar results to other state-of-the-art methods on Köhler dataset. Most previous methods and our proposed method did not yield consistent performance gain. Methods with excellent performance on GoPro dataset often failed to get similar state-of-the-art results on Köhler dataset. Even though it seems true that the results on Köhler dataset could be biased due to small number of images in it, training data must be carefully investigated. For example, GoPro and Köhler datasets have different image sizes, used different ways of generating blurs, and different scenes and environments. Thus, we argue that GoPro dataset may not be the best training set for Köhler test set. Our observation can also be seen in other literature: Kupyn et al. investigated two training datasets (ImageNet based synthetic blur dataset and GoPro dataset) and obtained quite different performance results [20]. Thus, there have been much research on blur data generation [31, 17, 27, 9, 20]. As an example, we fine-tuned our GoPro-trained network with Su training dataset with several epochs only and it yielded 27.26dB on Köhler dataset. This shows the importance of training dataset for performance. Further investigation is needed to address this.
Lastly, Figure 7 shows one example of the input blurred image, the output deblurred image by our proposed method, the output image without using backbone networks, and the output image without RIR skip connections. This decomposition study illustrates the roles of each component in our method. Even though our training process does not enforce to separate the role of each component, backbone networks estimate blurring information that is corresponding to the blur kernel estimation in conventional deblurring methods.

Acknowledgments

This work was supported partly by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2017R1D1A1B05035810), the Technology Innovation Program or Industrial Strategic Technology Development Program (10077533, Development of robotic manipulation algorithm for grasping/assembling with the machine learning using visual and tactile sensing information) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea), and a grant of the Korea Health Technology R&D Project of Korea(NRF) funded by the Ministry of Education(NRF-2017R1D1A1B05035810), the Technology Innovation Program through the National Research Foundation (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI18C0316).

References

[1] Y. Bahat, N. Efrat, and M. Irani. Non-uniform Blind Deblurring by Reblurring. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3306–3314, 2017.

[2] A. Chakrabarti. A Neural Approach to Blind Motion Deblurring. In *European Conference on Computer Vision (ECCV)*, pages 221–235, 2016.

[3] S. Cho and S. Lee. Fast Motion Deblurring. *ACM Transactions on Graphics*, 28(5):145:1–8, Dec. 2009.

[4] F. Couzinie-Devy, J. Sun, K. Alahari, and J. Ponce. Learning to estimate and remove non-uniform image blur. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1075–1082, 2013.

[5] E. Denton, S. Chintala, a. szlam, and R. Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1486–1494, 2015.

[6] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2650–2658, 2015.

[7] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2366–2374, 2014.

[8] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman. Removing camera shake from a single photograph. In *ACM Transactions on Graphics*, pages 787–794, July 2006.

[9] D. Gong, J. Yang, L. Liu, Y. Zhang, I. Reid, C. Shen, A. Van Den Hengel, and Q. Shi. From motion blur to motion flow: a deep learning solution for removing heterogeneous motion blur. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2319–2328, 2017.

[10] A. Gupta, N. Joshi, C. Lawrence Zitnick, M. Cohen, and B. Curless. Single image deblurring using motion density functions. In *European Conference on Computer Vision (ECCV)*, pages 171–184, 2010.

[11] S. Harmeling, M. Hirsch, and B. Schölkopf. Space-variant single-image blind deconvolution for removing camera shake. In *Advances in Neural Information Processing Systems (NIPS)*, pages 829–837, 2010.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[13] M. Hirsch, C. J. Schuler, S. Harmeling, and B. Schölkopf. Fast removal of non-uniform camera shake. In *IEEE International Conference on Computer Vision (ICCV)*, pages 463–470, 2011.

[14] T. H. Kim, B. Ahn, and K. M. Lee. Dynamic scene deblurring. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3160–3167, 2013.

[15] T. H. Kim and K. M. Lee. Segmentation-free dynamic scene deblurring. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2766–2773, 2014.

[16] T. H. Kim and K. M. Lee. Generalized video deblurring for dynamic scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5426–5434, 2015.

[17] T. H. Kim, K. M. Lee, B. Schölkopf, and M. Hirsch. Online Video Deblurring via Dynamic Temporal Blending Network. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4058–4067, 2017.

[18] D. P. Kingma and J. L. Ba. ADAM: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[19] R. Köhler, M. Hirsch, B. Mohler, B. Schölkopf, and S. Harmeling. Recording and playback of camera shake: Benchmarking blind deconvolution with a real-world database. In *European Conference on Computer Vision (ECCV)*, pages 27–40, 2012.

[20] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8183–8192, 2018.

[21] W. S. Lai, J. B. Huang, Z. Hu, N. Ahuja, and M. H. Yang. A comparative study for single image blind deblurring. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1701–1709, 2016.

[22] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[23] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4681–4690, 2017.
[24] Y. Li, S. B. Kang, N. Joshi, S. M. Seitz, and D. P. Huttenlocher. Generating sharp panoramas from motion-blurred videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2424–2431, 2010.

[25] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee. Enhanced deep residual networks for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 136–144, 2017.

[26] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. In *International Conference on Learning Representations (ICLR)*, 2016.

[27] S. Nah, T. H. Kim, and K. M. Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 257–265, 2017.

[28] J. Pan, D. Sun, H. Pfister, and M. H. Yang. Blind image deblurring using dark channel prior. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1628–1636, 2016.

[29] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic Differentiation in PyTorch. In *Neural Information Processing Systems (NIPS) Workshops*, 2017.

[30] C. J. Schuler, M. Hirsch, S. Harmeling, and B. Schölkopf. Learning to Deblur. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1439–1451, July 2016.

[31] Q. Shan, J. Jia, and A. Agarwala. High-quality motion deblurring from a single image. *ACM Transactions on Graphics*, 27(3):73:1–10, Aug. 2008.

[32] S. Su, M. Delbracio, J. Wang, G. Sapiro, W. Heidrich, and O. Wang. Deep video deblurring for hand-held cameras. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 237–246, 2017.

[33] J. Sun, W. Cao, Z. Xu, and J. Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 769–777, 2015.

[34] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia. Scale-recurrent network for deep image deblurring. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8174–8182, 2018.

[35] O. Whyte, J. Sivic, A. Zisserman, and J. Ponce. Non-uniform deblurring for shaken images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 491–498, 2010.

[36] P. Wieschollek, M. Hirsch, B. Schölkopf, and H. P. A. Lensch. Learning Blind Motion Deblurring. In *IEEE International Conference on Computer Vision (ICCV)*, pages 231–240, 2017.

[37] L. Xu, J. S. Ren, C. Liu, and J. Jia. Deep Convolutional Neural Network for Image Deconvolution. In *Advances in neural information processing systems (NIPS)*, pages 1790–1798, 2014.

[38] L. Xu, S. Zheng, and J. Jia. Unnatural l0 sparse representation for natural image deblurring. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1107–1114, 2013.

[39] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In *European Conference on Computer Vision (ECCV)*, pages 294–310, 2018.

[40] H. Zhao, O. Gallo, I. Frosio, and J. Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3(1):47–57, 2017.