Multi-Temporal Recurrent Neural Networks For Progressive Non-Uniform Single Image Deblurring With Incremental Temporal Training

Dongwon Park*[0000−0001−6060−9705], Dong Un Kang*[0000−0003−2486−2783], Jisoo Kim[0000−0002−6984−2850], and Se Young Chun[0000−0001−8739−8960]

Department of Electrical Engineering, UNIST, Republic of Korea
{dong1,qkrtnskfk23,rlawltn1053,sychun}@unist.ac.kr

Abstract. Blind non-uniform image deblurring for severe blurs induced by large motions is still challenging. Multi-scale (MS) approach has been widely used for deblurring that sequentially recovers the downsampled original image in low spatial scale first and then further restores in high spatial scale using the result(s) from lower spatial scale(s). Here, we investigate a novel alternative approach to MS, called multi-temporal (MT), for non-uniform single image deblurring by exploiting time-resolved deblurring dataset from high-speed cameras. MT approach models severe blurs as a series of small blurs so that it deblurs small amount of blurs in the original spatial scale progressively instead of restoring the images in different spatial scales. To realize MT approach, we propose progressive deblurring over iterations and incremental temporal training with temporally augmented training data. Our MT approach, that can be seen as a form of curriculum learning in a wide sense, allows a number of state-of-the-art MS based deblurring methods to yield improved performances without using MS approach. We also proposed a MT recurrent neural network with recurrent feature maps that outperformed state-of-the-art deblurring methods with the smallest number of parameters.

1 Introduction

Non-uniform single image deblurring is still a challenging *ill-posed* inverse problem to recover the original sharp image from a blurred image with or without estimating unknown non-uniform blur kernels. One approach to tackle this problem is to simplify the given problem by assuming uniform blur and to recover both image and blur kernel [11, 37, 7, 45]. However, uniform blur is not accurate enough to approximate real blur, and thus there has been much research to extend the blur model from uniform to non-uniform in a limited way compared to the full dense matrix [15, 14, 42, 16, 44, 33]. Other non-uniform blur models have been investigated such as additional segmentations within which simple blur models were used [8, 18] or motion estimation based deblurs [19, 20]. Recently, deep-learning-based approaches have been proposed with excellent quantitative

* Equal contribution. Code is available at https://github.com/Dong1P/MTRNN

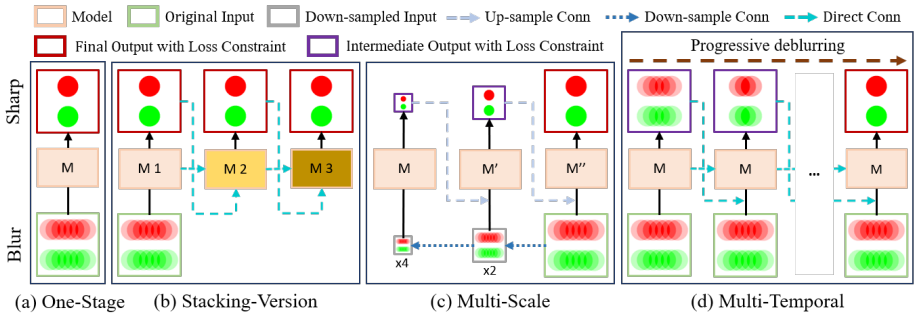

Fig. 1. Pipelines of four approaches for deblurring: (a) one-stage (OS) [24, 23, 2], (b) stacking version (SV) [47, 32], (c) multi-scale (MS) [31, 41, 12] and (d) our proposed multi-temporal (MT). In SV, the models M1, M2, M3 are independent. In MS, the models M, M', M" used to be independent, but recent works used strongly dependent models with parameter sharing. Our MT uses the identical model M over all iterations.

results and fast computation time. There are largely two different ways of using deep neural networks (DNNs) for deblurring. One is to use DNNs to explicitly estimate non-uniform blurs [40, 6, 36, 4] and the other is to use DNNs to directly estimate the sharp image without estimating blurs [46, 21, 43, 39, 31, 41].

Focusing on DNN based non-uniform single image deblurring, there are three different approaches as illustrated in Fig. 1: (a) one-stage (OS) attempts to recover the original image from blurred image in the original spatial scale [24, 23, 2] (b) stacking-version (SV) uses independent models multiple times and each model attempts to restore the original image from blurred or intermediate deblurred image in the original scale iteratively [47, 32] and (c) multi-scale (MS) (or coarse-to-fine) exploits multiple downsampled images in different spatial scales and recovers the downsampled original images in the lowest scale first and then to restore the original images in the original scale at the end [31, 41, 12]. This approach has been the most popular among state-of-the-art methods [12, 41].

OS approach in Fig. 1 (a) is straightforward and the model M is supervised to yield the original sharp image in the original high spatial scale at once. SV approach in Fig. 1 (b) uses multiple independent models M1, M2, M3 and possibly more. Each model is supervised to yield the original sharp image in the original high spatial scale. However, each model has different input, either a given blurred image or an intermediate deblurring result of the previous model. Later models refine the deblurring results for improved performance, but with the price of increased network parameters.

MS approach in Fig. 1 (c) also uses multiple models like SV approach, but the models are supervised to yield the original or down-scaled images in the different spatial scales. It is well-known that blurs become relatively smaller as image scale decreases and recovering image from intermediate result of deblurring is easier than restoring image from given blurred image. Thus, MS approach breaks a challenging deblurring problem for severe blur into multiple easy prob-

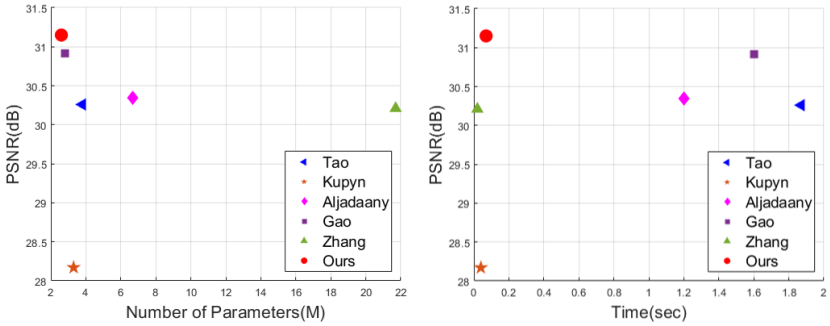

Fig. 2. Number of parameters (Million) and Time (Sec) vs. PSNR (dB) evaluated on the GoPro dataset. Our proposed MT-RNN method (Ours) yielded the best PSNR with the smallest parameters, real-time computation among state-of-the-art image deblurring methods such as Tao [41], Kupyn [24], Aljadaany [2], Gao [12] and Zhang [47].

lems (dealing with small blur in low spatial scale or deblurring from intermediate result of deblurring in high spatial scale) that can be seen as a form of curriculum learning [5] in a wide sense. However, since edge information is important for reliable deblurring [7, 45], performing deblurring in low spatial scales using MS approach could be a potential drawback. Note that MS approach requires incremental spatial training with spatially augmented training data (*i.e.*, downsampled sharp and blurred images). MS approach used to require large number of network parameters for different spatial scales [31], but recently many state-of-the-art MS based methods are using shared network parameters over spatial scales [41, 12]. The models at different spatial scales are strongly dependent.

Here, we investigate a novel alternative approach to MS, called multi-temporal (MT), for non-uniform single image deblurring by exploiting time-resolved deblurring dataset from high-speed cameras like the popular GoPro dataset [31]. We model severe blurs as a series of small blurs so that MT approach deblurs small amount of blurs in the original spatial scale progressively instead of restoring the images in different spatial scales as illustrated in Fig. 1 (d). Our MT approach, that can be seen as another form of curriculum learning [5] in a wide sense, also breaks down a challenging deblurring problem into a series of easy deblurring problems with small blurs. Note that unlike MS approach, each deblurring sub-problem in MT approach is still in the original spatial scale so that high-frequency information can be used for reliable deblurring [7, 45].

To realize MT approach, we propose progressive deblurring over iterations and incremental temporal training. Our scheme does not require special parameter sharing across spatial scales like [12], but allows natural parameter sharing in the same spatial scale over iterations, yielding better performance than MS approach on the GoPro [31] and its variant, Su [39] datasets. We also proposed a MT recurrent neural network (MT-RNN) with recurrent feature maps that outperformed state-of-the-art methods on the GoPro [31], Lai [25] datasets with the smallest number of parameters and real-time computation as in Fig. 2.

2 Related Works

Non-DNN Deblurring: There have been works on predicting non-uniform blurs assuming spatially linear blur [15], simplified camera motion [14], parameterized model [42], filter flow [16], l_0 sparsity [44], and dark channel prior [33]. There have also been some works to exploit multiple images from videos [26], to utilize segmentation by assuming uniform blur on each segmentation area [8], to segment motion blur using optimization [18], to simplify motion model as local linear using MS approach [19], and to use bidirectional optical flows [20].
DNN Image Deblurring: Blind image / video deblurring employed DNNs for original sharp images from blurred input images. Xu *et al.* proposed a direct estimation of the sharp image with optimization to approximate deconvolution by a series of convolutions using DNNs [46]. Aljadaany *et al.* proposed a learning of both image prior and data fidelity for deblurring [2]. Kupyn *et al.* [24] proposes generative adversarial network based on feature pyramid and relativistic discriminator [29] with a least-square loss [17]. Zhang *et al.* proposed a multi-patch hierarchical network for different feature levels on the same spatial resolution [47]. They also proposed a stacked multi-patch network without parameter sharing. Nah *et al.* proposed a MS network with Gaussian pyramid [31] and Tao *et al.* proposed convolution long short-term memory (LSTM)-based MS DNN [41]. Gao *et al.* proposed MS parameter sharing and nested skip connections [12].
Curriculum Learning: MS approach for deblurring [31, 41, 12, 38] can be seen as a form of curriculum learning [5], tackling a challenging deblurring problem with less challenging sub-problems in lower spatial scales. At each scale, DNN is trained more effectively so that it helped to achieve state-of-the-art performances. Li [27] trained the model to generate the intermediate goals using Gaussian blurs and to progressively perform image super-resolution. Our MT approach is another form of curriculum learning, but breaks the deblurring problem in a different way. We exploit temporal information to generate intermediate goals with *non-uniform* blurs in the original spatial scale, while MS is generating intermediate goals with *uniform* blurs in lower scales or in the original scale.
RNN Video Deblurring: There have been video deblurring works to exploit temporal information: blending temporal information in spatio-temporal RNN [21], taking temporal information into account with RNN of several deblur blocks [43] and accumulating video information across frames [39]. Zhou [49] proposed spatio-temporal variant RNN. RNN utilizes previous frames effectively such as convolutional LSTM [41]. Similar to SV, Nah [32] proposed RNN with intra-frame iterations by reusing RNN cell parameters. RNN based video deblurring and our MT-RNN share similar architectures. However, the former has inputs across frames while our MT-RNN has inputs over deblurring sub-problems.
Deblurring Dataset: The importance of image deblurring dataset has been raised with remarkable progress of image deblurring. Several existing popular uniform deblurring dataset [40, 22, 13] are sythesized by blur kernel. In [40, 22, 13], single sharp image is convolved with a set of motion kernels for blurred image. Recently, several works [31, 41, 12, 30] generated dynamic motion blurred image by averaging consecutive video frames captured by high frame rate camera.

3 Temporal Data Augmentation

Unlike MS approach [31, 41, 12] to augment training data with down-sampling that could be sub-optimal for reliable deblurring [7, 45], we propose temporal training data augmentation for deblurring. Most deblurring training datasets were obtained from high-speed cameras [31, 38, 39], thus our MT augmentation scheme for intermediate goals and inputs can be widely applicable.

3.1 Motion Blur Dataset

Recent non-uniform deblurring datasets were generated by the integration of the sharp images [31, 38, 39]. The blurred image $y \in R^{M \times N}$ from a sequence of images $x \in R^{M \times N}$ can be constructed as follows:

$$y = g\left(\frac{1}{T}\int_{t=0}^{T} x(t)dt\right) \approx g\left(\frac{1}{n}\sum_{i=0}^{n} x[i]\right) \tag{1}$$

where T and $x(t)$ denote an exposure time and a sharp image at time t in continuous domain, n and $x[i]$ denote the number of images and the ith sharp image in discrete domain, and g is a camera response function (CRF). We denote the dataset of blurred images y from n frames as Temporal Level n (TLn).

For example, motion blur datasets in [31, 38, 39] were captured by GoPro Hero camera (240 frame per sec) and 7-13 frames were averaged to yield a blurred image where a mid-frame image was selected as a ground truth image. Thus, the training / test datasets of [31] (called the GoPro dataset) consist of TL7, TL9, TL11 and TL13 with the ground truth of TL1.

3.2 Temporal Data Augmentation For MT Approach

Our MT approach requires more intermediate goals and inputs. Our temporal data augmentation further generates more blurred images to complete the whole training set with TLn where n is an odd number. For the GoPro dataset [31], we temporally augmented the data to generate TL1 (ground truth), TL3, ..., TL13 (we denote them Temporal GoPro or T-GoPro dataset). Unlike previous works using TL7-13 for the inputs of training, our MT exploits TL3-13 for both inputs and intermediate goals of training as proposed in the next section.

4 Multi-Temporal (MT) Approach

Fig. 1(d) illustrates the concept of our MT approach that progressively predicts intermediate deblurred image (*e.g.*, predicting TL($n-2$) from TLn) to finally yield the desired sharp image that is close to the ground truth (TL1). As illustrated in Fig. 1, our proposed MT approach is different from others such as OS (*e.g.*, predicting TL1 from TLn), SV (*e.g.*, predicting TL1 from TLn or intermediate results from previous network), and MS (*e.g.*, predicting downsampled TL1

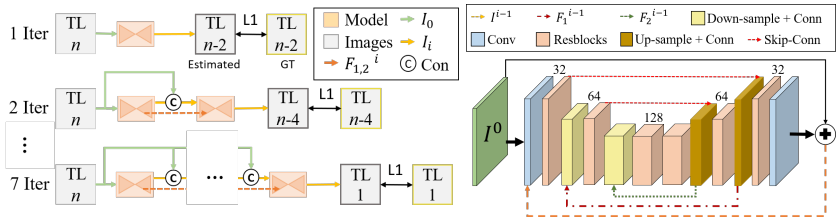

Fig. 3. (Left) Pipeline of incremental temporal training with our proposed MT-RNN, (Right) proposed neural network architecture of MT-RNN.

from downsampled TLn or intermediate results from previous scale). Here, we present incremental temporal training for our MT approach to use intermediate goals (*e.g.*, TL$(n-2)$). Then, we propose the MT-RNN with recurrent feature maps as a representative implementation of our MT approach for progressive deblurring. Lastly, we briefly discuss empirical convergence of our MT-RNN.

4.1 Incremental Temporal Training

Our MT approach conjectures that it is easier to predict TL5 from TL7 than to directly estimate TL1 from TL7, which seems reasonable (see supplementary material for further details). Curriculum learning approach can be used and incremental temporal training uses various temporally augmented dataset as intermediate goals as illustrated in Fig. 3 (left).

At the first iteration, a network is trained with randomly selected blurred images TLn (*e.g.*, 7, 9, 11, 13) as inputs and with corresponding less blurred images TL$(n-2)$ as intermediate goals using $L1$ loss. At the next iteration, the estimated image from the previous iteration is taken as input and corresponding less blurred images TL$(n-4)$ as intermediate goals. This process is continued if intermediate goals become the final goals with TL1. Finally, 1-3 more iterations are done with the same final goals TL1. The max iteration for training was set to be 7 to reduce the overall training time. Temporal step (TS) is defined to be the difference between the input TL and the output TL over 1 iteration for training. Unless specified, we set TS=2 based on the ablation studies in Table 1.

Our model uses identical parameters and training was performed independently for all iterations. This allows us to train the DNN with limited memory and to reduce the size of network without special parameter sharing.

4.2 MT-RNN for Progressive Deblurring

Baseline MS deblurring: Among MS based deblurrings [41, 38, 31], the DNN of Tao [41] shares parameters over scales that can be modeled as follows:

$$\{\hat{I}^j, h^j\} = \mathrm{DNN}_{\mathrm{Tao}}(U(I^j), U(\hat{I}^{j+1}), U(h^{j+1}); \theta_{\mathrm{Tao}}) \tag{2}$$

where j refers to a spatial scale where $j = 1$ represents the original high spatial scale, I^j and $\hat{I}^j$ are blurred and estimated images at the jth scale, respectively,

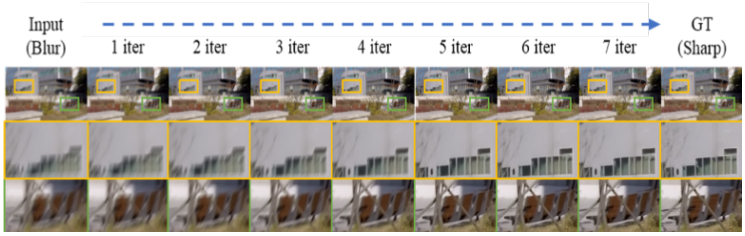

Fig. 4. Progressively deblurred images over iterations using our proposed MT-RNN.

DNN_{Tao} is the MS based DNN and θ_{Tao} is a set of parameters in the network, I^j is a down-sampled image from I^1 if $j > 1$, h is an intermediate feature map of convolutional LSTM, and U is a up-sampling operation by bilinear interpolation. Due to the encoder-decoder structure of U-Net [35], a base network of Tao [41], the receptive field of the DNN of Tao was relatively large, which is desirable for good deblurring performance. Thus, the DNN of Tao [41] was chosen as the base model for our proposed MT-RNN as illustrated in Fig. 3 (right).

Proposed MT-RNN: We propose MT-RNN with recurrent feature maps that can be modeled as follows:

$$\{\hat{I}^i, F_1^i, F_2^i\} = \text{DNN}_{\text{Ours}}(\hat{I}^{i-1}, I^0, F_1^{i-1}, F_2^{i-1}; \theta_{\text{Ours}}) \tag{3}$$

where i refers to an iteration number, F_1^{i-1} and F_2^{i-1} are recurrent feature maps from the $(i-1)$th decoder, I^0 is an input blurred image(TLn), $\hat{I}^{i-1}$ and $\hat{I}^i$ are predicted images at the ith iteration, respectively. Since the network utilizes previous feature maps, the output recurrent feature maps F_1^i and F_2^i are fed into the feature skip connection layer in the next iteration. DNN_{Ours} is our MT-RNN and θ_{Ours} is a set of network parameters to be trained as shown in Fig. 3 (right) with feature extraction layers and residual blocks of 32, 64, 128 channels at the top, middle and bottom encoder-decoders, respectively [31, 41].

For our proposed MT-RNN, we made a number of modifications on the DNN of Tao [41]. Firstly, changing kernel size from 5×5 to 3×3 was responsible for 0.13dB improvement in PSNR and substantially decreased number of parameters by 26%. Secondly, residual skip connection for input was responsible for 0.15dB improvement in PSNR. Fig. 4 illustrates progressive deblurring of our proposed MT-RNN over iterations. Fig. 5 quantitatively shows that our proposed MT approach recovers frequency components over iterations unlike SV approach.

Recurrent feature maps: Recurrent features F^{i-1} are from the last residual block of each decoder and are concatenated with the feature maps of previous encoder at feature extraction layer as illustrated in Fig. 3 (right):

$$F_{enc}^i = Cat(F^{i-1}, f^i) \tag{4}$$

where f^i is the feature map of previous encoder at the ith iteration. Estimated image $\hat{I}^{i-1}$ is concatenated with I^0:

$$I_{cat}^i = Cat(\hat{I}^{i-1}, I^0) \tag{5}$$

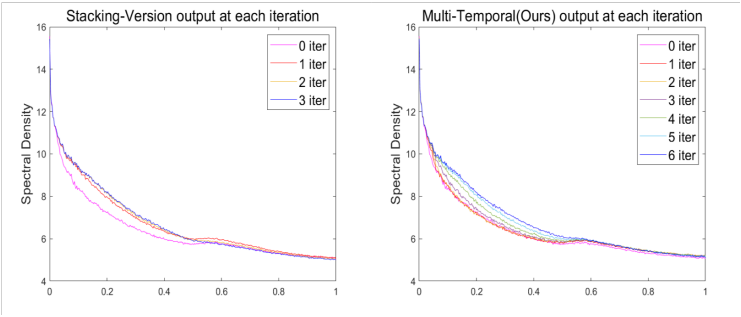

Fig. 5. Output spectral densities at each iteration for SV and our MT approaches. MT approach progressively recovers frequency components while SV approach does not.

and then the encoder takes the I^i_{cat} and $F^i_{encoder}$ as inputs.

Similar to other works of Tao [41] using convolutional LSTM for passing intermediate feature maps to the next spatial scale or of Nah [32] using hidden state h_{t-1} in RNN cell, our MT-RNN uses intermediate feature maps F^{i-1} from decoder that may include information about blur patterns and intermediate results for I^i. Using recurrent feature maps F^{i-1} was responsible for improved performance by 0.31dB.

Residual learning: Kupyn [24], Gao [12] and Zhou [49] utilized residual learning for deblurring. We conducted an ablation study for residual learning. In Fig. 3, our proposed network takes I^0 and $\hat{I}^{i-1}$ as inputs and residual skip connection is linked to I^0. The linked I^0 was responsible for improved performance over $\hat{I}^{i-1}$ as summarized in Table 1.

4.3 Convergence of Progressive MT-RNN

Determining the number of iterations for MT-RNN is important for performance. We studied iteration vs. PSNR / SSIM for the network that was trained only with one type of TL images (e.g., TL13) for all TL7, 9, 11, 13. Training was performed until the 7th iteration for all cases. As illustrated in Fig. 6, all networks yielded increased PSNR / SSIM over iterations until 5th / 6th iterations, and then decreased performances beyond the trained iteration. We set the number of iterations to be 6 for all experiments of our proposed MT-RNN. In all cases with different TL images, our proposed MT-RNN methods outperform state-of-the-art MS methods (Tao [41]) as in Fig. 6 (solid lines vs. dotted lines).

5 Experiments

Datasets The GoPro dataset [31] consists of 3,214 blurred images with the size of 1280×720 that are divided into 2,103 training images and 1,111 test images. In both training and test sets, TL7, 9, 11, 13 images were evenly distributed. We generated our T-GoPro dataset that includes more intermediate TL images

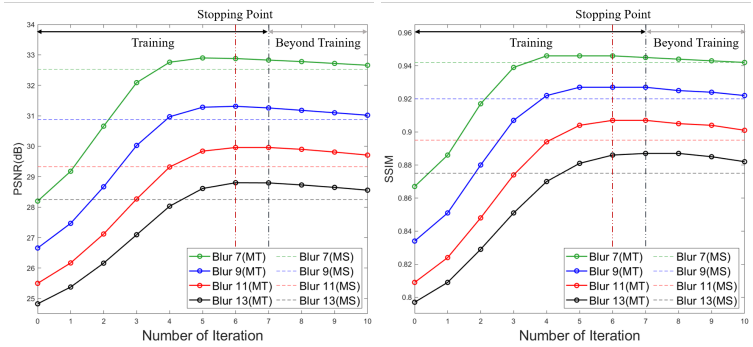

Fig. 6. Iteration vs. PSNR / SSIM for our proposed MT-RNN trained with one of TL7, 9, 11, 13. Corresponding MS models are also trained only with each TL.

(TL3, 5) with 5,500 training, 110 validation and 1,200 test images. Note that this new training dataset does not include any video from the original GoPro test dataset.

Su dataset [39] consists of 71 videos (6,708 images) with the size of 1,920×1080 or 1,280×720 from multiple devices. They are divided into 61 training and 10 test videos. Lastly, qualitative comparison was performed on Lai dataset [25] whose image sizes are varying within 351-1,024×502-1,024.

Implementation Details For fair comparisons, we evaluated our proposed method and state-of-the-art methods on the same machine with NVIDIA Titan V GPU using PyTorch During training, the Adam optimizer was used with learning rate 2×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. Patch size was set to be 256×256 and data augmentations such as random crop, horizontal flip and 90° rotation were used. Note that since the number of channel is changed by concatenation replacing the original add operation of skip connection [41], 1×1 convolution was used. PSNR and SSIM were used for evaluations. Run time was recorded with batch size 1 and data-loading time was not counted.

For Tables 1, 2, 3, the training iteration was 92×10^3 with reduced learning rate by half every 46×10^3 iterations. Both our T-GoPro dataset and the original GoPro dataset [31] were used. For Table 3, Su dataset [39] was used only for test. For Table 4 and Fig. 7, the total iteration was 46×10^4 with reduced learning rate by half every 46×10^3 iterations. The GoPro dataset [31] and the Lai dataset [25] were used for quantitative and qualitative evaluations.

5.1 Ablation Studies For MT-RNN Designs

There are a number of components that affect the performance of our MT-RNN and we performed ablation studies on the T-GoPro and GoPro datasets to select the best possible combinations of the components: temporal step (TS), residual

Table 1. Ablation studies with temporal step (TS), residual learning (ResL) and recurrent feature map (RFM) with training 92×10^3 iterations.

Test dataset				T-GoPro		GoPro [31]	
TS	Iteration	ResL	RFM	PSNR	SSIM	PSNR	SSIM
(a) **2**	**6**	I^0	**o**	**30.74** dB	**0.917**	**29.98** dB	**0.908**
(b) n/a	1 (OS)	I^0	n/a	29.93 dB	0.904	29.26 dB	0.895
(c) 4	4	I^0	o	30.44 dB	0.913	29.85 dB	0.889
(d) 2	6	I^0	x	30.43 dB	0.911	29.70 dB	0.888
(e) 2	6	$\hat{I}^{i-1}$	o	30.05 dB	0.905	29.41 dB	0.896
(f) 2	6	x	o	30.57 dB	0.913	29.90 dB	0.905

learning (ResL) and recurrent feature map (RFM). All results are summarized in Table 1 with our MT-RNN in the first row (a).

Table 1 (a), (b), (c) are corresponding to the ablation study for temporal step (TS). The row (b) is one-step approach, thus there is no TS as well as no recurrent feature map (RFM). The row (c) is the case with TS = 4, yielding improved performance over OS. In (c), iteration was set to be 4 to account for large TS as compared to our proposed MT-RNN with TS = 2. Note that (a), (c) are our MT approaches with different TS parameters and they outperformed one-step (OS) approach on both T-GoPro and GoPro datasets.

Table 1 (a), (d) are corresponding to the ablation study for recurrent feature map (RFM). It turned out that using RFM increased performances in MT-RNN by 0.31dB on the T-GoPro dataset and by 0.28dB on the original GoPro dataset.

Lastly, Table 1 (a), (e), (f) are corresponding to the ablation study for residual learning (ResL). It seems that using the original blurred image in the residual learning is important for improved performances as compared to using the previous output image in the residual learning. ResL in MT-RNN was the least important component among all three components according to the performance results in Table 1 (a), (f), especially for the original GoPro dataset.

5.2 Empirical comparisons of OS, SV, MS and MT Approaches

We performed empirical comparison studies for different deblurring approaches as illustrated in Fig. 1: one-stage (OS), stacking version (SV), multi-scale (MS) and our multi-temporal (MT). Table 2 summarizes the performances of different approaches in PSNR (dB), SSIM and the number of parameters (Million).

Firstly, Table 2 (g), (h), (i), (j) are comparing the performances of OS, SV, MS and MT approaches where OS, MS and MT contain the same amount of network parameters. Note that the original MT contains 0.041 M more parameters (1.58% increase) than MS does due to recurrent feature map (RFM), thus we removed RFM in MT (called MT w/o RFM) to have the same number of parameters. Even though RFM was an important component for improved performance as shown in Table 1, our proposed MT approach without RFM still outperformed OS and MS approaches with the same number of parameters in

Table 2. Empirical comparisons among different deblurring approaches with training 92×10^3 iterations: one-stage (OS), stacking version (SV), multi-scale (MS) and our proposed multi-temporal (MT). MT has 0.041 M more parameters (1.58% increase) than MS does due to recurrent feature map (RFM). MS w/ TL3-5 was trained with the training set of MT for inputs (*i.e.*, more training data) to yield ground truth images.

Test dataset	T-GoPro		GoPro [31]		Param
Approach	PSNR	SSIM	PSNR	SSIM	
(g) OS	29.93 dB	0.904	29.26 dB	0.895	2.594 M
(h) SV	30.38 dB	0.912	29.71 dB	0.903	7.890 M
(i) MS	30.25 dB	0.908	29.50 dB	0.898	2.594 M
(j) MT w/o RFM	30.43 dB	0.911	29.70 dB	0.888	2.594 M
(k) **MT**	**30.82** dB	**0.917**	**30.04** dB	**0.908**	2.635 M
(l) MS + MT	30.58 dB	0.915	29.87 dB	0.905	2.637 M
(m) MS w/ TL3-5	30.04 dB	0.906	29.27 dB	0.893	2.594 M

most cases and yielded comparable performances to SV with 3 times less parameters. With RFM, MT approach yielded state-of-the-art performances on both T-GoPro and GoPro datasets in all metrics over all the other approaches as shown in Table 2 (k).

Table 2 (l) are the results of the case to combine MS and MT (with RFM). MS + MT still outperformed all other approaches, but was not able to achieve better performance than the original MT approach. Lastly, Table 2 (m) are the results of MS approach to be trained with the original dataset along with additional dataset (TL3-5) that was used for training MT approaches (called MS w/ TL3-5). It turned out that using more data for training in MS degraded the performance of the original MS approach trained without TL3-5. In other words, using more training data only seems to help appropriate approaches such as MT, not any approaches such as MS.

We further investigated on different deblurring approaches by converting the original approach into our MT: for Kupyn [24] and Zhang [47], OS approach was

Table 3. Performance comparisons for state-of-the-art methods: Kupyn [24], Zhang [47] and Gao [12] before / after converting the original approach (OS / MS) into our MT approach with training 92×10^3 iterations. Evaluations were performed on T-GoPro, GoPro [31] and Su [39] datasets. Converting into MT approach does not change the number of parameters much since RFM was not used. PSNR in dB.

Test dataset		T-GoPro		GoPro [31]		Su [39]		Param
Method	Approach	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	(M)
Kupyn [24]	OS	28.27	0.870	27.58	0.858	28.32	0.865	3.28
Kupyn*	MT	**28.36**	**0.872**	**27.70**	**0.860**	**28.53**	**0.868**	3.28
Zhang [47]	OS	30.25	0.908	29.591	0.900	28.59	0.866	5.42
Zhang*	MT	**30.91**	**0.918**	**30.21**	**0.910**	**29.56**	**0.892**	5.43
Gao [12]	MS	30.70	0.916	29.930	0.907	29.73	0.897	3.87
Gao*	MT	**31.01**	**0.921**	**30.32**	**0.915**	**29.79**	**0.898**	3.40

converted into MT approach (called Kupyn* and Zhang*) and for Gao [12], MS approach was converted into MT (called Gao*). Note that the number of parameters in Gao was decreased by converting into MT since independent feature extraction modules at different scales was removed except for the original scale. As shown in Table 3, in all cases on all datasets, MT conversions of state-of-the-art methods yielded improved performances in PSNR and SSIM over original deblurring approaches. These results demonstrated the superior performances of our proposed MT approaches over other deblurring approaches.

5.3 Benchmark Results

We performed benckmark studies on the popular GoPro dataset [31]. Our proposed MT-RNN method was trained with our T-GoPro dataset that is generated using temporal data augmentation of the original GoPro dataset and then it was evaluated on the GoPro test dataset (1,111 images) that other previous methods were also evaluated on. The total training iteration was 4×10^4. Table 4 summarized the reported performances of state-of-the-art methods in the literature in PSNR (dB), SSIM as well as other information such as the number of parameters, run time and used training sets. Our MT-RNN yielded the highest PSNR (31.15 dB) with the smallest number of parameters (2.6 M) thanks to our MT approach to use the identical network over all iterations and effective curriculum learning approach to break challenging problem into easy sub-problems. Moreover, our MT-RNN is real-time - its run time is only 0.07 second, which is advantageous over other state-of-the-art methods such as Kupyn [23], Aljadaany [2] or Gao [12]. Fig. 2 summarized the results of Table 4 in graphs.

Fig. 7 shows deblurred results on GoPro and Lai datasets for visual comparisons. The images on the first row are input blurred images and the results

Table 4. Benchmarks on the GoPro test dataset [31] for PSNR, SSIM, parameter size, run time and training datasets. The 1st, 2nd and 3rd best performances are highlighted with red, blue and green. The run times of [44], [19], [40] and [2] are from their papers.

Method	PSNR	SSIM	Param	Run time	Training sets
Xu [44]	25.10 dB	0.890	-	13.41 sec	-
Kim [19]	23.64 dB	0.824	-	1 hour	-
Sun [40]	24.64 dB	0.843	-	20 min	[10]
Gong [13]	27.19 dB	0.908	-	-	[28, 3]
Ram [34]	28.94 dB	0.922	-	-	[31, 28, 9]
Nah [31]	29.08 dB	0.914	21.0 M	0.91 sec	[31]
Kupyn [23]	28.70 dB	0.958	-	0.15 sec	[31, 28]
Tao [41]	30.26 dB	0.934	3.8 M	0.34 sec	[31]
Kupyn [24]	28.17 dB	0.925	3.3 M	0.03 sec	[31, 28]
Zhang [47]	30.21 dB	0.934	21.7 M	0.02 sec	[31, 39]
Aljadaany [2]	30.35 dB	0.961	6.7 M	1.20 sec	[31, 1, 22]
Gao [12]	30.92 dB	0.942	2.84 M	1.01 sec	[31]
Ours	31.15 dB	0.945	2.6 M	0.07 sec	Temporal [31]

Fig. 7. Visual comparisons among state-of-the-art methods and our proposed MT-RNN on GoPro dataset [31] for (a), (b), (c) and on Lai dataset [25] for (d). Four input blurred images are on the 1st row, deblurred images of Nah [31] on the 2nd row, deblurring results of Tao [41] on the 3rd row, results of Zhang [48] on the 4th row. Our results using MT-RNN are on the 5th row (bottom row). Our proposed method yielded deblurred images that are visually better than the results of state-of-the-art methods for all 4 image cases for fine details.

of Nah [31], Tao [41], Zhang [47], and our MT-RNN are on the 2nd, 3rd, 4th, 5th rows of Fig. 7, respectively, showing that our MT-RNN outperforms other state-of-the-art methods visually on both GoPro and Lai test datasets.

6 Discussion

The GoPro training dataset has been the most popular dataset for single image deblurring works as shown in Table 4, but most works also used additional dataset such as Microsoft COCO dataset [28] for improved performance. Thus, it seems disadvantageous to use the GoPro dataset [31] only for training. However, our proposed MT-RNN was able to achieve better performance than other state-of-the-art methods without using additional dataset. Even though we increased the training set size by temporal data augmentation, as shown in Table 2 (m), this increased training dataset is not always helpful for performance boost.

In Fig. 6, MT-RNN yielded increased PSNR over early iterations (usually, before 6th or 7th iterations) and then yielded decreased PSNR for later iterations. This seems to be related to the generalization of deep learning and thus this issue is beyond the scope of this work. Deep learning beyond training scenarios often fails to yield expected, reliable results. Active stopping criterion (*e.g.*, gating unit in [32]) can potentially improve the performance of our MT-RNN.

Many state-of-the-art MS based single image deblurring methods exploit network weights across different spatial scales by parameter sharing [41] or partial networks weight sharing [12]. Weight sharing allows to reduce the number of network parameters significantly while performance is increased. However, weight sharing across scales seems to require special techniques and they are usually slow in computation. Our MT approach can be seen as natural weight sharing across temporal iterations without special methods. Thus, our MT approach seems to yield fast computation and high performance. We observed that the performance of MT was substantially decreased without weight sharing over iterations.

7 Conclusion

We investigate a novel alternative approach to MS, called MT, for non-uniform image deblurring by exploiting time-resolved deblurring dataset from high-speed cameras. Our proposed MT approach with progressive deblurring, incremental temporal training and MT-RNN yielded improved performance over previous deblurring approaches (OS, SV, MS) and outperformed state-of-the-art deblurring methods with the smallest number of parameters and real-time computation.

Acknowledgement 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 through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI18C0316).

References

1. Agustsson, E., Timofte, R.: NTIRE 2017 challenge on single image super-resolution: Dataset and study. In: CVPRW (2017)
2. Aljadaany, R., Pal, D.K., Savvides, M.: Douglas-Rachford networks: Learning both the image prior and data fidelity terms for blind image deconvolution. In: CVPR (2019)
3. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. IEEE T-PAMI (2011)
4. Bahat, Y., Efrat, N., Irani, M.: Non-uniform Blind Deblurring by Reblurring. In: ICCV (2017)
5. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: ICML (2009)
6. Chakrabarti, A.: A Neural Approach to Blind Motion Deblurring. In: ECCV (2016)
7. Cho, S., Lee, S.: Fast Motion Deblurring. ACM Transactions on Graphics (2009)
8. Couzinie-Devy, F., Sun, J., Alahari, K., Ponce, J.: Learning to estimate and remove non-uniform image blur. In: CVPR (2013)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR (2009)
10. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes (VOC) Challenge. IJCV (2010)
11. Fergus, R., Singh, B., Hertzmann, A., Roweis, S.T., Freeman, W.T.: Removing camera shake from a single photograph. In: ACM Transactions on Graphics (2006)
12. Gao, H., Tao, X., Shen, X., Jia, J.: Dynamic scene deblurring with parameter selective sharing and nested skip connections. In: CVPR (2019)
13. Gong, D., Yang, J., Liu, L., Zhang, Y., Reid, I., Shen, C., Van Den Hengel, A., Shi, Q.: From motion blur to motion flow: a deep learning solution for removing heterogeneous motion blur. In: CVPR (2017)
14. Gupta, A., Joshi, N., Lawrence Zitnick, C., Cohen, M., Curless, B.: Single image deblurring using motion density functions. In: ECCV (2010)
15. Harmeling, S., Hirsch, M., Schölkopf, B.: Space-variant single-image blind deconvolution for removing camera shake. In: NIPS (2010)
16. Hirsch, M., Schuler, C.J., Harmeling, S., Schölkopf, B.: Fast removal of non-uniform camera shake. In: ICCV (2011)
17. Jolicoeur-Martineau, A.: The relativistic discriminator: a key element missing from standard GAN. arXiv preprint arXiv:1807.00734 (2018)
18. Kim, T.H., Ahn, B., Lee, K.M.: Dynamic scene deblurring. In: ICCV (2013)
19. Kim, T.H., Lee, K.M.: Segmentation-free dynamic scene deblurring. In: CVPR (2014)
20. Kim, T.H., Lee, K.M.: Generalized video deblurring for dynamic scenes. In: CVPR (2015)
21. Kim, T.H., Lee, K.M., Schölkopf, B., Hirsch, M.: Online Video Deblurring via Dynamic Temporal Blending Network. In: ICCV (2017)
22. Köhler, R., Hirsch, M., Mohler, B., Schölkopf, B., Harmeling, S.: Recording and playback of camera shake: Benchmarking blind deconvolution with a real-world database. In: ECCV (2012)
23. Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., Matas, J.: DeblurGAN: blind motion deblurring using conditional adversarial networks. In: CVPR (2018)
24. Kupyn, O., Martyniuk, T., Wu, J., Wang, Z.: DeblurGAN-v2: Deblurring (Orders-of-Magnitude) Faster and Better. In: ICCV (2019)

25. Lai, W.S., Huang, J.B., Hu, Z., Ahuja, N., Yang, M.H.: A comparative study for single image blind deblurring. In: CVPR (2016)
26. Li, Y., Kang, S.B., Joshi, N., Seitz, S.M., Huttenlocher, D.P.: Generating sharp panoramas from motion-blurred videos. In: CVPR (2010)
27. Li, Z., Yang, J., Liu, Z., Yang, X., Jeon, G., Wu, W.: Feedback network for image super-resolution. In: CVPR (2019)
28. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: ECCV (2014)
29. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Smolley, S.P.: Least squares generative adversarial networks. In: ICCV (2017)
30. Nah, S., Baik, S., Hong, S., Moon, G., Son, S., Timofte, R., Lee, K.M.: NTIRE 2019 challenge on video deblurring and super-resolution: Dataset and study. In: CVPRW (2019)
31. Nah, S., Kim, T.H., Lee, K.M.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: CVPR (2017)
32. Nah, S., Son, S., Lee, K.M.: Recurrent neural networks with intra-frame iterations for video deblurring. In: CVPR (2019)
33. Pan, J., Sun, D., Pfister, H., Yang, M.H.: Blind image deblurring using dark channel prior. In: CVPR (2016)
34. Ramakrishnan, S., Pachori, S., Gangopadhyay, A., Raman, S.: Deep generative filter for motion deblurring. In: ICCVW (2017)
35. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
36. Schuler, C.J., Hirsch, M., Harmeling, S., Schölkopf, B.: Learning to Deblur. IEEE transactions on pattern analysis and machine intelligence (2016)
37. Shan, Q., Jia, J., Agarwala, A.: High-quality motion deblurring from a single image. ACM Transactions on Graphics (2008)
38. Shen, Z., Wang, W., Lu, X., Shen, J., Ling, H., Xu, T., Shao, L.: Human-aware motion deblurring. In: ICCV (2019)
39. Su, S., Delbracio, M., Wang, J., Sapiro, G., Heidrich, W., Wang, O.: Deep video deblurring for hand-held cameras. In: CVPR (2017)
40. Sun, J., Cao, W., Xu, Z., Ponce, J.: Learning a convolutional neural network for non-uniform motion blur removal. In: CVPR (2015)
41. Tao, X., Gao, H., Shen, X., Wang, J., Jia, J.: Scale-recurrent network for deep image deblurring. In: CVPR (2018)
42. Whyte, O., Sivic, J., Zisserman, A., Ponce, J.: Non-uniform deblurring for shaken images. In: CVPR (2010)
43. Wieschollek, P., Hirsch, M., Schölkopf, B., Lensch, H.P.A.: Learning Blind Motion Deblurring. In: ICCV (2017)
44. Xu, L., Zheng, S., Jia, J.: Unnatural l0 sparse representation for natural image deblurring. In: CVPR (2013)
45. Xu, L., Jia, J.: Two-phase kernel estimation for robust motion deblurring. In: ECCV (2010)
46. Xu, L., Ren, J.S., Liu, C., Jia, J.: Deep Convolutional Neural Network for Image Deconvolution. In: NIPS (2014)
47. Zhang, H., Dai, Y., Li, H., Koniusz, P.: Deep stacked hierarchical multi-patch network for image deblurring. In: CVPR (2019)
48. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: ECCV (2018)
49. Zhou, S., Zhang, J., Pan, J., Xie, H., Zuo, W., Ren, J.: Spatio-temporal filter adaptive network for video deblurring. arXiv preprint arXiv:1904.12257 (2019)