

Bringing Alive Blurred Moments

Kuldeep Purohit¹ Anshul Shah^{2*} A. N. Rajagopalan¹

¹ Indian Institute of Technology Madras, India ² University of Maryland, College Park

kuldeeph.purohit3@gmail.com, anshulb@cs.umd.edu, raju@ee.iitm.ac.in

Abstract

We present a solution for the goal of extracting a video from a single motion blurred image to sequentially reconstruct the clear views of a scene as beheld by the camera during the time of exposure. We first learn motion representation from sharp videos in an unsupervised manner through training of a convolutional recurrent video autoencoder network that performs a surrogate task of video reconstruction. Once trained, it is employed for guided training of a motion encoder for blurred images. This network extracts embedded motion information from the blurred image to generate a sharp video in conjunction with the trained recurrent video decoder. As an intermediate step, we also design an efficient architecture that enables real-time single image deblurring and outperforms competing methods across all factors: accuracy, speed, and compactness. Experiments on real scenes and standard datasets demonstrate the superiority of our framework over the state-of-the-art and its ability to generate a plausible sequence of temporally consistent sharp frames.

1. Introduction

When shown a motion blurred image, humans can mentally reconstruct (sometimes ambiguously perhaps) a temporally coherent account of the scene that represents what transpired during exposure time. However, in computer vision, natural video modeling and extraction has remained a challenging problem due to the complexity and ambiguity inherent in video data. With the success of deep neural networks in solving complex vision tasks, end-to-end deep networks have emerged as incredibly powerful tools.

Recent works on future frame prediction reveal that direct intensity estimation leads to blurred predictions. Instead, if a frame is reconstructed based on the original image and corresponding transformations, both scene dynamics and invariant appearance can be preserved well. Based on this premise, [6, 45] and [21] model the task as a flow

of image pixels. The methods [38, 43] generate a video from a single sharp image, but have a severe limitation in that they work only on the specific scene for which they are trained. All of these approaches work only on sharp images and videos. However, motion during exposure is known to cause severe degradation in the captured image quality due to the blur it induces. This is usually the case in low-light situations where the exposure time of each frame is high and in scenes where significant motion happens within the exposure time. In [35], it has been shown that standard network models used for vision tasks and trained only on high-quality images suffer a significant degradation in performance when applied to images degraded by blur.

Motion deblurring is a challenging problem in computer vision due to its ill-posed nature. Recent years have witnessed significant advances in deblurring [36, 27, 25]. Several methods [41, 26, 5, 31, 3, 14, 16, 17, 41] have been proposed to address this problem using hand-designed priors as well as Convolutional Neural Networks (CNN) [2, 29, 30] for recovering the latent image. A few methods [33, 7] have been proposed to remove heterogeneous blur but they are limited in their capability to handle general dynamic scenes. Most of these methods strongly rely on the accuracy of the assumed image degradation model and include intensive, sometimes heuristic, parameter-tuning and expensive computations, factors which severely restrict their accuracy and applicability in real-world scenarios. The recent works of [23, 24, 18, 34] overcome these limitations to some extent by learning to directly generate the latent sharp image, without the need for blur kernel estimation.

We wish to highlight here that until recently, all existing methods were limited to the task of generating only ‘a’ deblurred image. In this paper, we address the task of reviving and reliving all the sharp views of a scene as seen by the camera during its flight within the exposure time. Recovering sharp content and motion from motion blurred images can be valuable for uncovering the underlying dynamics of a scene (e.g., in sports, traffic surveillance monitoring, entertainment etc.). The problem of extracting a video from a single blurred observation is challenging due to the fact that a blurred image can only reveal aggregate information

*Work done while at Indian Institute of Technology Madras, India.

about the scene during exposure. The task requires recovery of sharp frames which are temporally and scene-wise consistent in the sense that they emulate recording coming from a high frame-rate camera. State-of-the-art deblurring methods such as [36] [27] estimate at best a group of poses which constitute the camera motion, but with total disregard to their ordering. For example, one would get the same blurred image even if the temporal order is reversed (temporal ambiguity). As a post-processing step, synthesizing a sequence from this group of poses is a non-trivial task. Although the camera motion can be partially detected through gyroscope sensors attached to modern cameras, the obtained data is too sparse to completely describe trajectories within the time interval of a single lens exposure. More importantly, sensor information is seldom available for most internet images. Further, these methods can only handle blur induced by a camera imaging a static planar scene which is not representative of a typical real-world scenario and hence not very interesting.

We present a two-stage deep convolutional architecture to carve out a video from a motion blurred image that is applicable to non-uniform motion caused by individual or combined effects of camera motion, object motion and arbitrary depth variations in the scene. We avoid overly simplified models to represent motion and hence refrain from creating synthetic datasets for supervised training. The first stage consists of training a video auto-encoder wherein the encoder accepts a sequence of video frames to extract a latent motion representation while the decoder estimates the same video by applying estimated motion trajectories to a single sharp frame in a recurrent fashion. We use this trained video decoder to guide the training of a CNN (which we refer to as Blurred Image Encoder (BIE)) to extract the same motion information from a blurred image as the video encoder would from the image sequence corresponding to that blurred image. For testing, we propose an efficient deblurring network to first estimate a sharp frame from the given blurred image. The BIE is responsible for extracting motion features from the blurred image. The video decoder uses the outputs of the BIE and the deblurred sharp frame to generate the video underlying the motion blurred image.

As the only other work of this kind, [13] very recently proposed a method to estimate a video from a single blurred image by training multiple neural networks to estimate the underlying frames. In contrast, our architecture utilizes a single recurrent neural network to generate the entire sequence. Our recurrent design implicitly addresses temporal ambiguity to a large extent, since generation of any frame in the sequence is naturally preconditioned on all the previous frames. The approach of [13] is limited to small motion, owing to its architecture and training procedure. We estimate pixel level motion instead of intensities which proves to be an advantage for the task at hand, especially

in cases with large blur (which is an issue with [13]). Our deblurring architecture not only outperforms all existing deblurring methods but is also smaller and significantly faster. In fact, separating the processes of content and motion estimation allows our architecture to be used with any off-the-shelf deblurring approach.

Our work advances the state-of-the-art in many ways. The main contributions are:

- A novel solution for extracting a sharp video from a single motion blurred image. In comparison to the state-of-the-art [13], our network is faster, more accurate (especially for large blur) and contains fewer parameters.
- A two-stage training strategy with a recurrent architecture for learning to extract an ordered spatio-temporal motion representation from a blurred image in an unsupervised manner. Unlike [13], our network is independent of the number of frames in the sequence.
- An efficient architecture to perform real-time single image deblurring that also delivers superior performance over the state-of-the-art in deblurring [34] across all factors: accuracy, speed (20 times faster) and compactness.
- Qualitative and quantitative analysis using benchmark datasets to demonstrate the superiority of our framework over competing methods in deblurring as well as video generation from a single blurred image.

2. The Proposed Architecture

Convolutional neural networks (CNNs) have been successfully applied for various vision tasks on images but translating these capabilities to video is non-trivial due to their inefficiency in exploiting temporal redundancies present in videos. Recent developments in recurrent neural networks provide powerful tools for sequence modeling as demonstrated in speech recognition [8] and caption generation for images [37]. Long short term memory networks (LSTMs) can be used to generate outputs that are highly correlated along the temporal dimension and hence form an integral part of our video generation framework. Though Conv3Ds have been used for video classification approaches, we found that for our application, recurrent networks were more efficient. Considering that we are working with images, the spatial information across the image is equally important. Hence we use Convolutional LSTM units [40] as our building blocks, which are capable of capturing both spatial and temporal dependencies.

The task of generating an image sequence requires the network to understand and efficiently encode static as well as dynamic information for a certain period of time. Although such an encoding is not clearly defined and hence

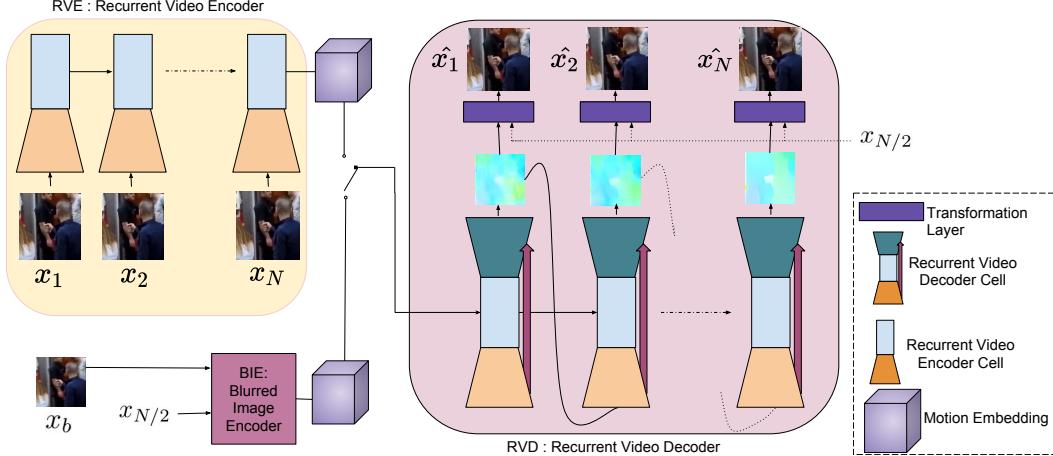


Figure 1. An overview of our video generation architecture during training. The first step involves training the RVE-RVD for the task of video reconstruction. This is followed by guided training of BIE through the trained RVD.

unavailable in labeled datasets, we overcome this challenge by unsupervised learning of motion representation. We propose to use video reconstruction as a surrogate task for training our BIE. Our hypothesis is that a successful solution to the video reconstruction task will allow a video autoencoder to learn a strong and meaningful motion representation which will enable it to impart spatio-temporal coherence to the generated moving scene content.

In our proposed video autoencoder, the encoder utilizes all the video frames to extract a latent representation, which is then fed to decoder which estimates the frame sequence in a recurrent fashion. The Recurrent Video Encoder (RVE) reads N sharp frames $x_{1..N}$, one at each time-step. It returns a tensor at the last time-step, which is utilized as the motion representation of the image sequence. This tensor is used to initialize the first hidden state of another ConvLSTM based network called Recurrent Video Decoder (RVD) whose task is to recurrently estimate N optical flows. Since the RVE-RVD pair is trained using reconstruction loss between the estimated frames $\hat{x}_{1..N}$ and ground-truth frames $x_{1..N}$, the RVD must return the predicted video. To enable this, the (known) central frame of the video is acted upon by the flows predicted by the RVD. Specifically, the estimated flows are individually fed to a differentiable transformation layer to transform the central frame $x_{\lfloor \frac{N}{2} \rfloor}$ to obtain the frames $\hat{x}_{1..N}$. Once trained, we have an RVD which can estimate sequential motion flows, given a particular motion representation.

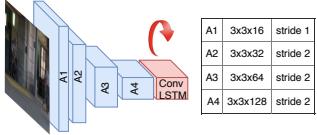
In addition, we introduce another network called Blurred Image Encoder (BIE) whose task is to accept blurred image x_B corresponding to the spatio-temporal average of the input frames $x_{1..N}$ and return a motion encoding, which too can be used to generate a sharp video. To achieve this task, we employ the already trained RVD to guide the training

of BIE so as to extract the same motion information from the blurred image as the RVE would from that image sequence. In other words, the weights are to be learnt such that $BIE(x_B) \approx RVE(x_{1..N})$. We refrain from using the encoding returned by RVE for training due to lack of ground truth for the encoded representation. Instead, the BIE is trained such that the predicted video at the output of RVD for the given x_B matches as closely as possible to the ground truth frames $x_{1..N}$. This ensures that the BIE learns to capture ordered motion information for the RVD to return a realistic video. Directly training the BIE-RVD pair poses a challenge since it requires learning to perform two tasks jointly: “video generation from motion representation” and “ambiguity-invariant motion extraction from a blurred image”. Such training delivers below-par performance (see supplementary material).

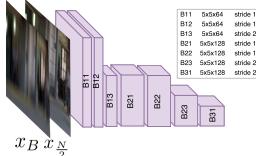
The overall architecture of the proposed methodology is given in Fig. 1. It is fully convolutional, end-to-end differentiable and can be trained using unlabeled high frame-rate videos, without the need for optical flow supervision, which is challenging to produce at large scale. During *testing*, the central sharp frame is not available and is estimated using an independently trained deblurring module (DM). We now describe the design aspects of the different modules.

2.1. Recurrent Video Encoder (RVE)

At each time-step, a frame is fed to a convolutional encoder, which generates a feature-map to be fed as input to the ConvLSTM cell. Interpreting ConvLSTM’s hidden-states as a representation of motion, the kernel-size of a ConvLSTM is correlated with the speed of the motion which it can capture. Since we need to extract motion taking place within a single exposure at fine resolution, we choose a kernel-size of 3×3 . As can be seen in Fig. 2(a), the en-



(a) RVE architecture.



(b) BIE architecture.

Figure 2. Architectures of BIE and RVE. The RVE is trained to extract a motion representation from a sequence of frames while the BIE is trained to extract a motion representation from a blurred image and a sharp image.

coder block is made of 4 convolutional blocks with 3×3 filters. The first block is a conv layer with stride of 1 and the rest contain a conv layer with stride of 2, followed by a Resblock. The number of feature maps in the outputs of these blocks are 16, 32, 64 and 128, respectively. A ConvLSTM cell operates on the features returned by the last block and augments it with memory from previous time-steps.

Overall, each module can be represented as $h_n^{enc} = enc(h_{n-1}^{enc}, x_n)$, where h_n^{enc} is encoder ConvLSTM state at time step n and x_n is the n^{th} sharp frame of the video.

2.2. Recurrent Video Decoder (RVD)

The task of RVD is to construct a sequence of frames using the motion representation provided by RVE and the (known) central frame ($x_{\lfloor \frac{N}{2} \rfloor}$) of the sequence. The RVD contains a flow encoder which utilizes a structure similar to the RVE. Instead of accepting images, it accepts optical flows. The flow encoding is fed to a ConvLSTM cell whose first hidden state is initialized with the last hidden state $h_{e,N}$ of the RVE. To estimate optical flows for a time-step, the output of the ConvLSTM cell is passed to a Flow decoder network (F_D). The flow estimated by F_D at each time-step is fed to a transformer module (T) which returns the estimated frame \hat{x}_n . The descriptions of F_D and T are provided below.

Flow Decoder (F_D): Realizing that the flow at current step is related to the previous one, we perform recurrence on optical flows for consecutive frames. The design of F_D is illustrated in Fig. 3. F_D accepts the output of ConvLSTM unit at any time-step and generates a flow-map. For robust estimation, we further perform estimation of flow at multiple scales using deconvolution (deconv) layers which “unpool” the feature maps and increase the spatial dimensions by a factor of 2. Inspired by [28], we make use of skip connections between the layers of flow encoder and F_D . All deconv operations use 4×4 filters and the convolutional operations use 3×3 filters. The output of the ConvLSTM cell is passed through a convolutional layer to estimate the flow $f_{n,1}$. The cell output is also passed through a deconv layer before being concatenated with the upsampled $f_{n,1}$ and the corresponding feature-map coming from the encoder, to ob-

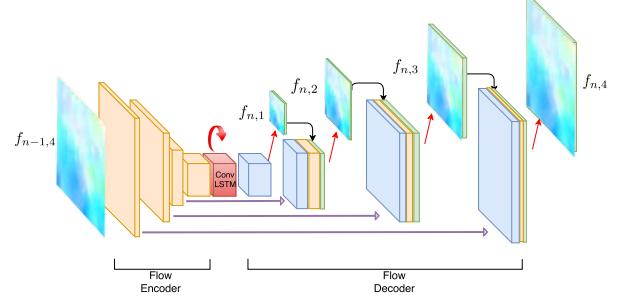


Figure 3. Our Recurrent Video Decoder (RVD). This module recurrently generates optical flows which are warped to transform the sharp frame. Flows are estimated at 4 different scales.

tain a hybrid feature map at that scale. As shown in Fig. 3, this process is repeated 3 more times to obtain the flow maps at subsequently higher scales ($f_{n,2..4}$).

Transformer(T): This generates a new frame by transforming a sharp frame using the output returned by F_D . It is a modified version of the Spatial Transformer Layer [11], which comprises of a grid generator followed by a sampler. Instead of a single transformation for the entire image (as originally proposed in [11]), T accepts one transformation per pixel. Since we focus on learning features for motion prediction, it provides immediate feedback on the flow map predicted by the optical flow generation layers. Effectively, the RVD function can be summarized as follows:

$$h_1^{dec} = h_N^{enc} \quad (1)$$

$$h_n^{dec}, f_{n,1..4} = G(h_{n-1}^{dec}, f_{n-1,4}) \quad (2)$$

$$\hat{x}_{n,1..4} = T(x_{\lfloor \frac{N}{2} \rfloor}, f_{n,1..4}) \quad (3)$$

for $n \in [1, N]$ where h_n^{dec} is decoder hidden state, $f_{n,1..4}$ are flows predicted at n and $\hat{x}_{n,1..4}$ are sharp frames predicted at different scales and G refers to a recurrent cell of RVD.

2.3. Blurred Image Encoder (BIE)

We make use of the trained encoder-decoder couplet to solve the task of extracting video from a blurred image. We advocate a novel strategy of utilizing spatio-temporal embeddings to guide the training of a CNN. The trained decoder has learnt to generate optical flow for all time-steps from the encoder’s hidden state. We employ this proxy network to solve the task of blurred image to video generation.

The use of optical flow recurrence enables our network to prefer temporally consistent sequences, which preempts it from returning arbitrarily ordered frames. However, directional ambiguity stays. For a scene with multiple objects, the ambiguity becomes more pronounced as each object can have its own independent motion. The BIE is connected with the pre-trained RVD and the pair is trained (RVD is fine-tuned) using a combination of ordering-invariant frame

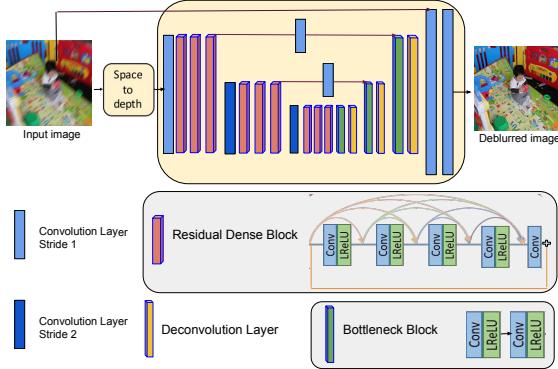


Figure 4. An overview of our dense deblurring architecture which we utilize to estimate the central sharp frame. It follows an encoder-decoder design with residual-dense blocks, bottleneck blocks, and skip connections present at 3 different sub-scales.

reconstruction loss and spatial motion smoothness loss over the RVD outputs (described later). No such ambiguity exists in the video autoencoder since the RVD has to exactly reproduce the video which is fed to RVE.

The BIE is implemented as a CNN which specializes in extracting motion features from a blurred image (we experimentally found that feeding the central sharp frame along with the blurred image improves its performance). The BIE is tasked to extract the sequential motion in the image by capturing local motion, e.g. at the smeared edges in the image. Moreover, the generated encoding should be such that the RVD can reconstruct motion trajectories. The BIE has 7 convolutional layers with kernel sizes as shown in Fig. 2(b). Each layer (except the last) is followed by batch-normalization and leaky ReLU non-linearity.

2.4. Cost Function

Both our network pairs (RVE-RVD and BIE-RVD) are trained by calculating the cost on the flows and frames estimated by the RVD. Since RVD implicitly estimates optical flows, we utilize a cost function motivated by learning-free variational method [1] which resembles the original formulation of [9] to impose flow smoothness. At each time step, the data loss measures the discrepancy between intensities of target frame and the output of transformation layer (obtained using the predicted optical flow field). The smoothness cost is in the form of total variation-loss on the estimated flow-maps: $TV(s) = \sum |\nabla_x s| + |\nabla_y s|$.

Coarse-to-Fine: Motivated by the approach employed in FlowNet [4], we improve our network's accuracy by estimating the flow-maps and frames in a coarse-to-fine manner. At each time-step, four loss terms are calculated using four optical flows $f_{n,1..4}$ predicted at sizes which are $(\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1)^{th}$ fraction of the original image resolution and applied on the corresponding down-sampled central frame

using the transformation layers. Reconstruction losses are calculated at each scale using suitably down-sampled ground truth videos. Effectively, we use a weighted sum of loss to guide the information flow over the network, which can be expressed as

$$\mathcal{L} = \sum_{j=1}^4 \lambda_j \left[\mathcal{L}_j + \sum_{n=1}^N \mu TV(f_{n,j}) \right] \quad (4)$$

For RVE-RVD training, the data term we use is

$$\mathcal{L}_j = \sum_{n=1}^N \left| \hat{x}_{n,j} - x_{n,j} \right|_1 \quad (5)$$

As mentioned in section 2.3, training of BIE-RVD requires a loss term that preempts the network from penalizing a video which correctly explains the blurred image but does not match the available ground truth. Following [13], we use loss function

$$\mathcal{L}_j = \sum_{n=1}^{\frac{N}{2}} \left| \left| \hat{x}_{n,j} + \hat{x}_{N-n,j} \right| - \left| x_{n,j} + x_{N-n,j} \right| \right|_1 + \left| \left| \hat{x}_{n,j} - \hat{x}_{N-n,j} \right| - \left| x_{n,j} - x_{N-n,j} \right| \right|_1 \quad (6)$$

Here, j represents the scale, n represents time-step, μ is the regularization weight for total-variation loss empirically set to 0.02. The relative weights λ_j s for each scale were adopted according to the loss weight suggested in [22].

2.5. Deblurring Module (DM)

We propose an independent network for deblurring the motion blurred observation. The estimated sharp frame is fed to both BIE and RVD during testing.

Recent works on image restoration have proposed end-to-end trainable networks which require labeled pairs of degraded and sharp images. Among them, [23, 34] have achieved promising results using multi-scale CNN composed of residual connections. We explore a more effective network architecture which is inspired by prior methods that use multi-level and multi-scale features. Our high-level design is similar to that of U-Net [28], which has been used extensively for preserving global context information in various image-to-image tasks [10]. Based on the observation that increase in number of layers and connections across them leads to a boost in feature extraction capability, the encoder structure of our network utilizes a cascade of Residual Dense Blocks (RDB) [44] instead of convolutional layers. An RDB is a cascade of convolutional layers connected through a rich set of residual and concatenation connections which immensely improves feature extraction capability by reusing features across multiple layers. Inclusion of such connections maximizes information flow along the intermediate layers and results in better convergence. These units

efficiently learn deeper and more complex features than a network with residual connections (which have been used extensively in recent deblurring methods[23, 18, 34, 13]), while requiring fewer parameters.

Our proposed deblurring architecture is depicted in Fig. 4. The decoder part of our network contains 3 pairs of up-sampling blocks to gradually enlarge the spatial resolution of feature maps. Each up-sampling block contains a bottleneck layer [12] followed by a deconvolution layer. Each convolution layer (except the last) is followed by a non-linearity. Similar to U-Net, features corresponding to the same dimension in encoder and decoder are merged with the help of projection layers. The output of the final up-sampling block is passed through two additional convolutional layers to reconstruct the output sharp image. Our network uses an asymmetric encoder-decoder architecture, where the network capacity becomes higher benefiting from the dense connections.

Further, we optimize the inference time of the network by performing computationally intensive operations on features at lower spatial resolution. This also reduces memory footprint while increasing the receptive field. Specifically, before feeding the input blurred image to encoder, we map the image to a lower resolution space using space-to-depth transformation. Following [20, 23], we omit normalization layers for stable training, better generalization and reduced computational complexity and memory usage. To further improve performance, we also exploit residual scaling [20].

3. Experiments

In this section, we carry out quantitative and qualitative comparisons of our approach with state-of-the-art methods for deblurring as well as video extraction tasks.

3.1. Implementation Details

We prepared our training data from GoPro dataset [23], following standard train-test split, wherein 22 full videos were used for creating training sets and 11 full videos were reserved for validation and testing. Each blurred image is produced by averaging 9 successive latent frames. Such an averaging simulates a photo taken at approximately 26 fps, while the corresponding sharp image shutter speed is 1/240. We extract 256×256 patches from these image sequences for training. Finally, our dataset is composed of 10^5 sets, each containing $N = 9$ sharp frames and the corresponding blurred image x_B . We perform data augmentation by random horizontal flipping and zooming by a factor in the range $[0.2, 2]$. The network is trained using Adam optimizer with learning rate 1×10^{-4} . The batch size was set to 10 and the training of our video-autoencoder took 5×10^4 iterations to converge. We then train the BIE-RVD pair with the same training configuration and reduce the learning rate for RVD parameters to 2×10^{-5} , for stable training.

Method	[42]	[39]	[33]	[7]	[23]	[18]	[34]	Ours
PSNR(dB)	21	24.6	24.5	26.4	28.9	27.2	30.10	30.58
SSIM	0.740	0.845	0.851	0.863	0.911	0.905	0.933	0.941
Time (s)	3800	700	1500	1200	6	0.8	0.4	0.02
Size(MB)	-	-	54.1	41.2	300	45.6	27.5	17.9
Hardware	CPU	CPU	CPU	CPU	GPU	GPU	GPU	GPU

Table 1. Performance comparison of our deblurring network with existing methods on the benchmark dataset [23].

For training and evaluating our single image deblurring network, we utilized the same train-test split of the GoPro dataset [23] as recent deblurring methods [23][34]. The batch size was set to 16 and the entire training took 4.5×10^5 iterations to converge.

3.2. Results for Single Image Deblurring

We evaluated the efficacy of our network (DM shown in Fig. 4) for the intermediate task of deblurring, both quantitatively and qualitatively on 1100 test images (resolution 1280×704) from the GoPro dataset [23]. The method of [39] is selected as representative traditional method for non-uniform blur. We also compare our performance with deep networks [23, 18, 34]. All the codes were downloaded from the respective authors' websites. Quantitative and qualitative comparisons are presented in Table 1 and Fig. 5, respectively. Since traditional method of [39] cannot model combined effects of general camera shake and object motion, it fails to faithfully restore most of the images in the test-set. On the other hand, the method of [18] trains a residual network containing instance-normalization layers using a mixture of deep-feature losses and adversarial losses, but leads to suboptimal performance on images containing large blur. The methods [23, 34] use a multi-scale strategy to improve capability to handle large blur, but fail in challenging situations. Fig. 5 shows that results of prior works suffer from incomplete deblurring or ringing artifacts. In contrast, our network is able to restore scene details more faithfully, while being 20 times faster than the nearest competitor [34]. These improvements are also reflected in the quantitative values presented in the table.

3.3. Results and Comparisons for Video Extraction

In Fig 6, we give results on standard test blurred images from the dataset of [23]. Note that some of them suffer from significant blur. Fig. 6(a) shows an image of a planar scene which is blurred due to dominant camera motion. Fig. 6(b) shows a 3D scene blurred due to camera motion. Figs. 6(c-f) show results on blurred images with dynamic object motion. Observe that the videos generated by our approach are realistic and qualitatively consistent with the blur and depth of the scene, even when the foreground incurs large motion. Our network is able to reconstruct videos from blurred images with diverse motion and scene content.



Figure 5. Visual comparisons of deblurring results on test dataset [23] (best viewed in high resolution).



Figure 6. Comparisons of our video extraction results with [13] on motion blurred images obtained from the test dataset of [23]. The first row shows the blurred images while the second and third rows show deblurred frames generated by our method and [13], respectively. Videos extracted using our method and [13] are provided in the supplementary document.

In comparison, the results of [13] suffer from local errors in deblurring, inconsistent motion estimation, as well as color distortions. We have observed that in general the method of [13] fails in cases involving high blur as direct image regression becomes difficult for large motion. In contrast, we divide the overall problem into two sub-tasks of deblurring and motion extraction. This simplifies learning and yields improvement in deblurring quality as well as motion estimation. The color issue in [13] can be attributed to the design of their networks, wherein feature extraction and reconstruction branches are different for different color channels. Our method applies the same motion to each color channel. By having a single recurrent network to generate the video, our network can be directly trained to extract even higher number of frames (> 9) without any design change or additional parameters. In contrast, [13] requires training of an additional network for each new pair of frames. Our overall architecture is more compact (45 MB vs 70 MB) and much faster (0.02s vs 0.45s for deblurring and 0.39s vs 1.10s for video generation) as compared to [13].

To perform quantitative comparisons with [13], we also trained another version of our network on the restricted

case of blurred images produced by averaging 7 successive sharp frames. For testing, 250 blurred images of resolution 1280×704 were created using the 11 test videos from the dataset of [23]. We compared the videos estimated by the two methods using the ambiguity invariant loss function defined in Eq. 6. The average error was found to be 49.06 for [13] and 44.12 for our method. Thus, even for the restricted case of small blur, our method performs favorably. Repeating the same experiment for 9 frames (i.e. for large blur from the same test videos) led to an error of 48.24 for our method, which is still less than the 7-frame error of [13]. We could not compute the 9-frame error for [13] as their network is rigidly designed for 7 frames only.

3.4. Additional Results on Video Extraction

Results on Camera Motion Dataset: For evaluating qualitative performance on videos with camera motion alone, we tested our network's ability to reconstruct videos from blurred images taken from datasets of [7], [15] and [19], which are commonly used for benchmarking deblurring techniques. Fig. 7(a) shows the video obtained on a synthetically blurred image provided in [7]. Fig. 7(b) shows



Figure 7. Video generation from images blurred with global camera motion from datasets of [7, 15] and [19]. First row shows the blurred images and our deblurred frames are shown in second row (generated videos are provided in the supplementary document).



Figure 8. Video generation results on real motion blurred images from dataset of [32]. The first row shows the blurred images. Second row contains the deblurred images estimated with our method (extracted videos are provided in the supplementary document).

result on an image from the dataset of [15]. We can observe that the motion in the generated video conforms with the blur. The dataset [19] consists of both synthetic and real images collected from various conventional prior works on deblurring. Figs. 7(c-d) show our network’s results on synthetically blurred images from this dataset using non-uniform camera motion. The examples in Figs. 7(e-f) are real blurred images obtained from the same dataset. Our method is able to re-enact underlying motion quite well.

Results on Blur Detection Dataset: In Fig. 8, we show videos generated from real blurred images taken from the dataset of [32] which contains dynamic scenes. The results reaffirm that our network can sense direction and magnitude even in severely blurred images.

3.5. More Results and Ablation Studies

Additional results and experiments to highlight the motivation for our design choices are given in the supplementary material. Specifically, for the video autoencoder, we study the effects of motion flow estimation (instead of direct intensity estimation) and the recurrent design. This is followed by an analysis on the influence of different loss functions. Regarding training of BIE, we study the effect of input sharp frame on its performance and also compare our two-stage strategy (BIE trained using pre-trained

RVD) with the case where BIE and RVD are trained directly from scratch. We also include an analysis on variations in growth-rate and residual-dense connection topology on the training performance of our deblurring network.

4. Conclusions

We introduced a new methodology for video generation from a single blurred image. We proposed a spatio-temporal video auto-encoder based on an end-to-end differentiable architecture that learns motion representation from sharp videos in a self-supervised manner. The network predicts a sequence of optical flows and employs them to transform a sharp central frame and return a smooth video. Using the trained video decoder, we trained a blurred image encoder to extract a representation from a single blurred image, that mimics the representation returned by the video encoder. This when fed to the decoder returns a plausible sharp video representing the action within the blurred image. We also proposed an efficient deblurring architecture composed of densely connected layers that yields state-of-the-art results. The potential of our work can be extended in a variety of directions including blur-based segmentation, video deblurring, video interpolation, action recognition etc.

Acknowledgements: The first author gratefully acknowledges travel support from Google Research India.

References

- [1] Thomas Brox and Jitendra Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE transactions on pattern analysis and machine intelligence*, 33(3):500–513, 2011.
- [2] Ayan Chakrabarti. A neural approach to blind motion deblurring. In *European Conference on Computer Vision*, pages 221–235. Springer, 2016.
- [3] Sunghyun Cho and Seungyong Lee. Fast motion deblurring. In *ACM Transactions on Graphics (TOG)*, volume 28, page 145. ACM, 2009.
- [4] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2758–2766, 2015.
- [5] Rob Fergus, Barun Singh, Aaron Hertzmann, Sam T Roweis, and William T Freeman. Removing camera shake from a single photograph. In *ACM transactions on graphics (TOG)*, volume 25, pages 787–794. ACM, 2006.
- [6] John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. Deepstereo: Learning to predict new views from the world’s imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5515–5524, 2016.
- [7] Dong Gong, Jie Yang, Lingqiao Liu, Yanning Zhang, Ian Reid, Chunhua Shen, AVD Hengel, and Qinfeng Shi. From motion blur to motion flow: a deep learning solution for removing heterogeneous motion blur. In *The IEEE conference on computer vision and pattern recognition (CVPR)*, 2017.
- [8] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE, 2013.
- [9] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.
- [10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.
- [11] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025, 2015.
- [12] Simon Jégou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1175–1183. IEEE, 2017.
- [13] Meiguang Jin, Givi Meishvili, and Paolo Favaro. Learning to extract a video sequence from a single motion-blurred image. *arXiv preprint arXiv:1804.04065*, 2018.
- [14] Neel Joshi, Richard Szeliski, and David J Kriegman. Psf estimation using sharp edge prediction. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.
- [15] Rolf Köhler, Michael Hirsch, Betty Mohler, Bernhard Schölkopf, and Stefan Harmeling. Recording and playback of camera shake: Benchmarking blind deconvolution with a real-world database. In *European Conference on Computer Vision*, pages 27–40. Springer, 2012.
- [16] Dilip Krishnan and Rob Fergus. Fast image deconvolution using hyper-laplacian priors. In *Advances in Neural Information Processing Systems*, pages 1033–1041, 2009.
- [17] Dilip Krishnan, Terence Tay, and Rob Fergus. Blind deconvolution using a normalized sparsity measure. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 233–240. IEEE, 2011.
- [18] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiri Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. *arXiv preprint arXiv:1711.07064*, 2017.
- [19] Wei-Sheng Lai, Jia-Bin Huang, Zhe Hu, Narendra Ahuja, and Ming-Hsuan Yang. A comparative study for single image blind deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1709, 2016.
- [20] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *The IEEE conference on computer vision and pattern recognition (CVPR) workshops*, volume 1, page 4, 2017.
- [21] Ziwei Liu, Raymond Yeh, Xiaou Tang, Yiming Liu, and Asseem Agarwala. Video frame synthesis using deep voxel flow. *arXiv preprint arXiv:1702.02463*, 2017.
- [22] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4040–4048, 2016.
- [23] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, volume 1, page 3, 2017.
- [24] TM Nimisha, Akash Kumar Singh, and AN Rajagopalan. Blur-invariant deep learning for blind-deblurring. In *Proceedings of the IEEE E International Conference on Computer Vision (ICCV)*, 2017.
- [25] Jinshan Pan, Zhe Hu, Zhixun Su, and Ming-Hsuan Yang. Deblurring text images via l0-regularized intensity and gradient prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2908, 2014.
- [26] Jinshan Pan, Zhouchen Lin, Zhixun Su, and Ming-Hsuan Yang. Robust kernel estimation with outliers handling for image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2800–2808, 2016.
- [27] Jinshan Pan, Deqing Sun, Hanspeter Pfister, and Ming-Hsuan Yang. Blind image deblurring using dark channel prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1628–1636, 2016.
- [28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmen-

- tation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [29] Christian J Schuler, Harold Christopher Burger, Stefan Harmeling, and Bernhard Scholkopf. A machine learning approach for non-blind image deconvolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1067–1074, 2013.
- [30] Christian J Schuler, Michael Hirsch, Stefan Harmeling, and Bernhard Schölkopf. Learning to deblur. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1439–1451, 2016.
- [31] Qi Shan, Jiaya Jia, and Aseem Agarwala. High-quality motion deblurring from a single image. In *Acm transactions on graphics (tog)*, volume 27, page 73. ACM, 2008.
- [32] Jianping Shi, Li Xu, and Jiaya Jia. Discriminative blur detection features. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2965–2972. IEEE, 2014.
- [33] Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 769–777, 2015.
- [34] 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*, pages 8174–8182, 2018.
- [35] Igor Vasiljevic, Ayan Chakrabarti, and Gregory Shakhnarovich. Examining the impact of blur on recognition by convolutional networks. *arXiv preprint arXiv:1611.05760*, 2016.
- [36] Subeesh Vasu and AN Rajagopalan. From local to global: Edge profiles to camera motion in blurred images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4447–4456, 2017.
- [37] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.
- [38] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Advances In Neural Information Processing Systems*, pages 613–621, 2016.
- [39] Oliver Whyte, Josef Sivic, Andrew Zisserman, and Jean Ponce. Non-uniform deblurring for shaken images. *International journal of computer vision*, 98(2):168–186, 2012.
- [40] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015.
- [41] Li Xu and Jiaya Jia. Two-phase kernel estimation for robust motion deblurring. In *European Conference on Computer Vision*, pages 157–170. Springer, 2010.
- [42] Li Xu, Shicheng Zheng, and Jiaya Jia. Unnatural l0 sparse representation for natural image deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1107–1114, 2013.
- [43] Tianfan Xue, Jiajun Wu, Katherine Bouman, and Bill Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2016.
- [44] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [45] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *European Conference on Computer Vision*, pages 286–301. Springer, 2016.