Human Motion Deblurring using Localized Body Prior

Jonathan Samuel Lumentut, Joshua Santoso, and In Kyu Park[1]

Dept. of Electrical and Computer Engineering, Inha University, Incheon 22212, Korea
{jlumentut,22192280}@inha.edu, pik@inha.ac.kr

Abstract. In recent decades, the skinned multi-person linear model (SMPL) is widely exploited in the image-based 3D body reconstruction. This model, however, depends fully on the quality of the input image. Degraded image case, such as the motion-blurred issue, downgrades the quality of the reconstructed 3D body. This issue becomes severe as recent motion deblurring methods mainly focused on solving the camera motion case while ignoring the blur caused by human-articulated motion. In this work, we construct a localized adversarial framework that solves both human-articulated and camera motion blurs. To achieve this, we utilize the result of the restored image in a 3D body reconstruction module and produces a localized map. The map is employed to guide the adversarial modules on learning both the human body and scene regions. Nevertheless, training these modules straight-away is impractical since the recent blurry dataset is not supported by the 3D body predictor module. To settle this issue, we generate a novel dataset that simulates realistic blurry human motion while maintaining the presence of camera motion. By engaging this dataset and the proposed framework, we show that our deblurring results are superior among the state-of-the-art algorithms in both quantitative and qualitative performances.

1 Introduction

The task of restoring sharp and blur-free imaging becomes fundamental in computer vision works for the sake of ameliorating recent high-level tasks such as recognition or detection. The restoration of scene and object motion-blurred images are widely known as deblurring. This motion is represented by discrete representations of point-spread-function (PSF). The PSF is stored in a certain spatial window, known as blur kernel, for deconvolving the blurry image to its sharper version. The simplest way to do deblurring is to treat the scene motion as uniform. Krishnan et al. [1] introduce automatic regularization to estimate the correct blur kernel for optimization. However, in the real-world case, motion blurs variate along spatial region on the image. This happens due to the presence of objects that are located at different depths on a scene. To solve this issue, Kim et al. [2] utilizes optical flow to obtain specific blur location to be restored

[1]Corresponding Author

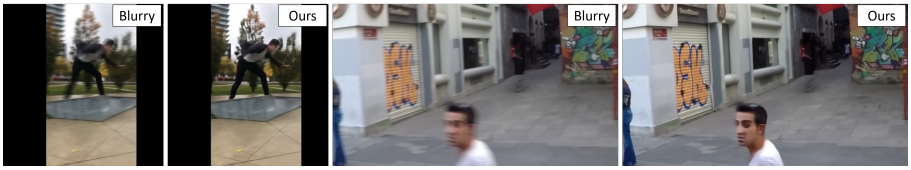

Fig. 1. Blurry images caused by both human and camera motions are restored using our approach.

accordingly. Since the rise of deep learning, the works of non-uniform deblurring are widely exploited. State-of-the-art work of Nah et al. [3] restored deblurred image directly without blur kernel estimation. Recently, the work of deblurring pursed inward with the goals of specific object restoration. The work of Shen et al. [4] introduces a semantic face mask to locally train the discriminators. The masks are generated from each specific human face part, such as eyes, nose, and mouth. Ren et al. [5] utilize human face prior for providing sharper deblurred face results. The human face prior is utilized in the deblurring layers by plugging the face identity and projected 3D rendered face directly in the network. On the other hand, Shen et al. [6] initiate the work of human deblurring. Their approach relies on the network's ability to recognize the rectangular human region that is provided from a sharp ground truth dataset.

In this work, we opt to focus on solving motion blur caused by human articulated and camera motions. Our work is amplified with the adversarial network that takes humans prior to focus on both human and scene regions themselves. To our best knowledge, this work is the first to apply such prior when it comes to deblurring. Training the network with recent human blur datasets [3, 6] seems ineffective with our goal since they are mostly affected by camera motions. Therefore, we take the challenge of providing a novel blurry human dataset that considers both human and camera motions constraints. Human motion blur is mostly non-uniform since the articulated human motion is different between human joints. Thus, we provide a newly blurry human dataset that is generated from both camera and human motions. Then, we crafted the deblurring network that is joined with the 3D body reconstruction in an end-to-end fashion. The output of this generator, namely: the deblurred image and 3D reconstructed body, are utilized in the localized adversarial modules. This is learned via our generated attentional map that is part of our contribution.

In the experiments, we demonstrate our framework's performance using the blurry human dataset and achieve significant improvement in both quantitative and qualitative aspects. Visual results of our approach are demonstrated in Figure 1. With the success of our work, we show that using the additional dataset and utilization of human prior in the deblurring modules, are learnable with generative adversarial network (GAN) approach. In summary, we describe our contributions as follows:

– We provide a learning based deblurring algorithm that utilizes human prior information from the remarkable body statistical model of SMPL [7].

– We propose, to our best knowledge, the first adversarial-based framework that is trained using localized regions that are extracted from humans and the scene's blurry locations.
– We present a novel way to synthesize blurry human motion dataset.

2 Related Works

Motion Deblurring Early deblurring algorithms utilize the traditional way of firstly estimating the blur kernel. The kernel is used to deconvolve the blurry input in order to restore it back to a sharp image. Various regularization priors have been utilized for improving this approach [1, 8]. These approaches further targeted multi-view imagings, such as the works of [9] in stereo and [10, 11] in light field imaging. With the advancement of deep learning, various approaches are introduced. Nah et al. [3] proposes a multi-scale deep framework that shows deblurring robustness under various scales. Recently, the generative adversarial network (GAN) approach by [12] takes a particular interest in the deblurring field. State-of-the-art GAN deblurring is introduced by [13] that utilizes conditional GAN [14] with similar architecture to domain translation work [15]. This work is improved with the addition of multi-scale discriminators with more compact layers [16]. Recently, the work of human deblurring is introduced by [6], where a convolutional neural network is learned with the capability to find the human region in an image. However, they train the network by providing a non-precise human region since it is defined under a rectangular box.
Attention Modelling The work of attention modeling is pioneered by [17] that provides a spatial transformer network to spatially warps specific feature regions during classification tasks. This study shows that localized attention is beneficial for a specific learning task. Pioneer work of [18] utilizes global and local regions in a GAN based framework to solve the traditional image inpainting method. Recently, the work of face de-occlusion that has the task of removing specific objects with inpainted pixels is done with the utilization of a local region of the human face in the discriminator [19]. Furthermore, face motion deblurring is shown to be improved when local face regions, such as forehead, eyes, nose, and mouth, are learned in the discriminators [4, 5]. Following these trends, we introduced the utilization of the local human region to restore blurry images.
3D Body Reconstruction In order to generate the human body region, a sophisticated model is needed. In recent years, a statistical model is introduced to represent a 3D virtual human body that is extracted from a single image. This state-of-the-art model is designed by [7] and widely known as SMPL. This model is constructed by mainly 2 parameters, namely: body pose (β) and shape (θ). The predicted outputs of SMPL are body vertices and joints. Kanazawa et al. [20] utilized deep learning approach that regresses image features for predicting those outputs. An additional discriminative network is added for distinguishing real human and non-human body during learning. Their improved version of this work is done in the multi-frames (video) domain [21]. Recent work of [22] utilizes 2 images of the same person with different poses and views to generate the human

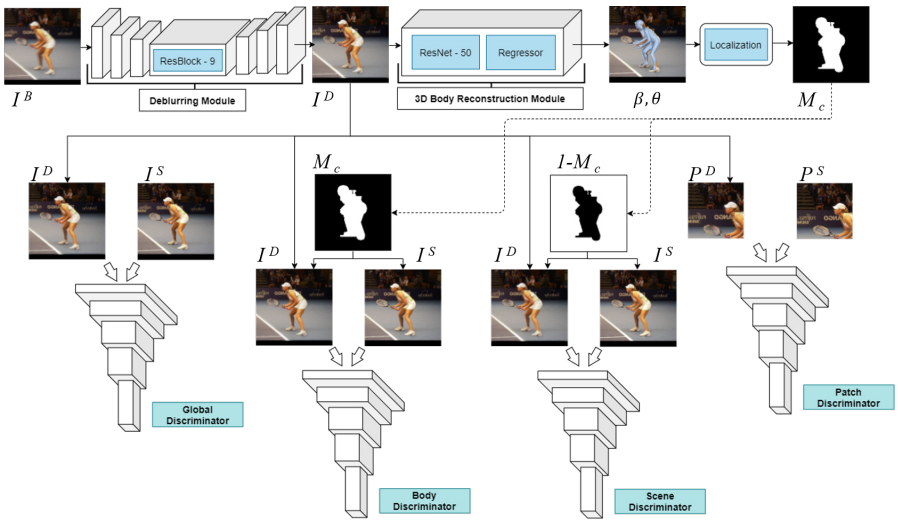

Fig. 2. The main architecture of our end-to-end framework. Blurry RGB image is fed through deblurring module where its output is processed through the 3D body reconstruction module to produce human body. Multiple discriminators improved the deblurring and predicted 3D body location results.

body with colored textures. Both images are passed through the model of [20], and the textures from each view is used to complement each other.

Based on these works, we are motivated to utilize the body reconstruction model to generate our localized attention map. More precisely, this map defines the region of humans and its nearby blurry pixels. This map is regarded as prior information for the localization procedure in our adversarial framework.

3 Human and Scene Deblurring

Human articulated motion blur is a challenging task since the non-uniform blur differs much in certain parts of the human body in an image. This problem becomes severe with the addition of a camera motion blur. To tackle these issues, we define a framework that is constructed by generator and discriminator networks, as shown in Figure 2. The generator is built with a concatenation of our deblurring model and the state-of-the-art 3D body prediction model. The key model of our framework is located in the discriminators, where they directly target the alleviation process of human and scene motion blur. Blur caused by human motion leaves distinguished trails on the blurry image. These trails are mostly located nearby the human region itself. Inversely, blurry scene regions can be captured outside the human region. This information is utilized in our discriminator modules. We elaborate on the details of our proposed generator and discriminator modules as follows.

3.1 Generator

The generator module of our framework has 2 main tasks, namely: predicting the deblurred image and estimating the 3D body vertices and joints. These tasks are described in the following modules: *deblurring* and *3D body reconstruction*.

Deblurring Module In the first scheme, an input of RGB blurry image, I^B, is fed into the deblurring model. As shown in Figure 2, the network receives an input of RGB blurry image, I^B. The image is initially passed through 3 convolutional blocks. The convolutional block is presented by a sequence of convolutional layer followed by instance normalization (IN) and $ReLU$ activation ($CONV \rightarrow IN \rightarrow ReLU$). The stride of the first 2 and last 2 convolutional layers are 2. Thus the intermediate feature is downscaled to 4 times. This feature is propagated through 9 modified Residual Net blocks (ResBlocks) of [23]. Each block contains $CONV \rightarrow IN \rightarrow ReLU \rightarrow CONV \rightarrow IN$ with the final addition with input at the last layer without any dropout function. More details about the deblurring module architecture is informed in the supplementary material. The output of this deblurring module is a residual image and it is added with input image, I^B, to generate final deblurred output, I^D.

3D Body Reconstruction Module 3D body prediction module is performed as an intermediator between the deblurring and the discriminative networks. This module benefits the performance of state-of-the-art work of [20] that uses a statistical model to predict the 3D human body. The statistic model is dubbed as SMPL [7] and is known for its capability to provide a high anatomic representation of the human body. With this model, a human body is defined by the shape (β) and pose (θ) parameters. Plugging these parameters into an SMPL function produces a renderable human body that contains 6890 vertices and 24 body joints. We utilize this model without its discriminator as in the original version [20] to balance our framework that has multiple adversarial modules. This module is utilized directly using the recent sophisticated weight of [20]. In our implementation, I^D is passed through a ResNet-50 feature extraction model and the regressor will generate features for predicting the SMPL output.

3.2 Discriminator

Besides the deblurring, our key contribution in the framework is the utilization of humans prior in the adversarial (discriminator) module. Our discriminator is trained to focus on learning a localized region. This is done by providing a specific attentional map that has the constraint of finding the local human region and its neighboring blurry region. We describe our approach in the following.

Human-based Attention Map An attention map is an approach of finding the local region of an image that provides useful information for the algorithm to learn. In this work, we show that omitting un-resourceful information such as the non-blurry region is beneficial. This information is penalized by a simple binary map with a value of 0. Instead of a generator, we utilized the map in the discriminator module as opposed to the approach of [6]. Our idea is motivated

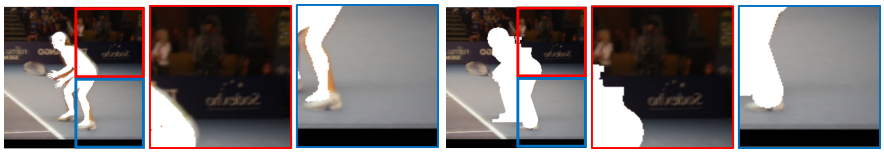

Fig. 3. First-3: attention map from rendered shape. **Last-3**: localized attention map that is obtained from connected body joints and blurry-sharp edge differences. The application of localized map fully covers human's and its nearby blurry region. Subsequently, its reversed version covers the non-human region.

by several prevailing works such as as [4, 19] that utilizes face region from a 3D statistical face model [24]. Applying a similar idea by using the output of the 3D body reconstruction module seems like a direct application. However, it is worthy to wisely choose which information should be propagated. As known above, 2 main information is provided from the 3D body module, namely: Body vertices and joints. Utilizing body vertices have a major disadvantage as predicted shape might not fully cover the blurry human region as depicted in the first 3 columns of Figure 3. This condition leaves us to the body joint prior option. First, the joint information from the predicted model contains only 24 coordinates that are far less than the vertices (6890), which is faster for training. Secondly, the body joint map can be connected to indicates favorable regions. Thus, we opt to connect the 24 body joint coordinates to visualize the human region. The result of this approach is dubbed as a line-joint (M_j) map and shown in the fourth column of Figure 4. While connected body joint implies good exploration, the blurry region on the human body may not fully be covered. This is expressed in the first 3 columns of Figure 3. We elaborate on the improvement of M_j to fully cover the blurry human region in the next region.

Localized Attention Map To solve the previous issue, M_j is augmented with the difference of I^S and I^B maps. Since this approach is run on discriminator, I^S can be utilized. Following the basic computer vision technic, we firstly obtain the edges difference between I^S and I^B in both horizontal and vertical directions by a 3×3 Sobel filter, as shown in first and second columns of Figure 4. The difference between horizontal and vertical edge maps is combined to produce the final map, M_b, of blurry edges, as shown in the third column of Figure 4. However, M_b might appear in all regions on the image, although it is not nearby the human body. Thus, M_b is spatially limited with the position that corresponds to most-top, -bottom, -left, and -right of body joint. After cropped, the map is convolved with Maxpool function to do the hole-filling. Finally, this map is added with M_j to produce the combined map, M_c, which is shown in the fifth column of Figure 4. As shown in the fourth column of Figure 3, this map covers blur inside and the nearby human body. M_c is termed as a localized attention map as it utilizes both human prior and blurry scene information. Scene's blurry region located outside M_c is also provided by reversing its value as $1 - M_c$.

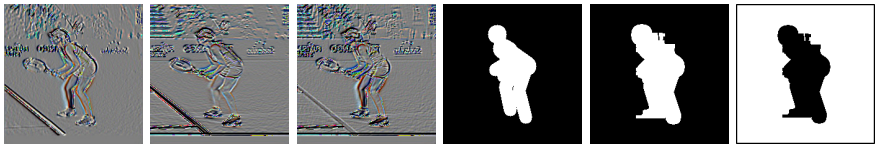

Fig. 4. First-3: Edge difference between blurry input and sharp ground truth in horizontal-vertical directions, and combined edge map from both directions. **Last-3:** Line joint map (M_j), localized map (M_c) for the body region, and reverse localized map ($1 - M_c$) for the scene region.

Multiple Discriminator Networks The discriminator networks act as the counter-learning process to the deblurring network in the generator scheme. Therefore, additional input, such as sharp (un-blurred) ground truth image, is utilized. In this module, deblurred and sharp ground truth images act as fake and real data distribution. Our adversarial networks are constructed by 4 discriminator modules, namely: global, body, scene, and patch discriminators, where each of them is assigned to solve specific tasks. Global and patch discriminators are utilized as a base framework that indicates no human prior. The utilization of M_c and $1 - M_c$ maps in the body and scene discriminators are regarded as our contribution to human prior. Discriminator networks' settings are provided in the supplementary material.

4 Optimization

In the initial stage, only the deblurring network is being trained without the human prior. Thus, only global and patch discriminators are included. During learning, the energy function is utilized to find the difference between the deblur output (I^D) and sharp ground truth image (I^S). At the first stage, we utilize perceptual loss that is calculated from L2 error between features of (I^D) and (I^S). These features are obtained from the convolution process of pre-trained VGG-19 network [25] until *conv3.3* layer. This approach is written in the following manner

$$L_{\mathbf{Deb}} = \frac{1}{b}\sum_{i}^{b} \gamma \|\Phi(I_i^S) - \Phi(I_i^D)\|^2 + (1-\gamma)\|I_i^S - I_i^D\| \tag{1}$$

where Φ is the *conv3.3* function of VGG network and i represents each pixel in a set of batch multiplied by images spatial and channel size together (b). The γ value is a binary value. γ is set to 0 for final refinement using the L1 loss. For adversarial loss, we opt to utilize Least-squares GAN (LSGAN) [26]. The real and fake data probability losses are calculated by the average of discriminator scores upon real and fake results from the global and patch data, as:

$$L_{\mathbf{Real}} = \frac{0.5}{b}\sum_{i}^{b}(\Pi_{\mathbf{Glob}}(I_i^S) + \Pi_{\mathbf{Patch}}(P_i^S)); \tag{2}$$

$$L_{\mathbf{Fake}} = \frac{0.5}{b}\sum_{i}^{b}(\Pi_{\mathbf{Glob}}(I_i^D) + \Pi_{\mathbf{Patch}}(P_i^D)), \tag{3}$$

where $\Pi(\cdot)$ represents the discriminator function. Combining the equations above, the base LSGAN function of the discriminators and generator are defined by:

$$L_{\mathbf{Disc}} = 0.5 \times (\|L_{\mathbf{Real}} - 1\|^2 + \|L_{\mathbf{Fake}}\|^2); \tag{4}$$

$$L_{\mathbf{Gen}} = 100 \times L_{\mathbf{Deb}} + 0.5 \times (\|L_{\mathbf{Fake}} - 1\|^2). \tag{5}$$

In the next stage, we include the set of our contributions, namely: the body and scene discriminators. These discriminators are being trained with input data that is penalized by the generated mask from the 3D body predictor model. Thus, we define additional adversarial losses for the new discriminators. The real and fake data losses for the body and scene discriminators are calculated as follows:

$$L^*_{\mathbf{Real}} = \frac{0.5}{b}\sum_{i}^{b}(\Pi_{\mathbf{Body}}(M_c \odot I_i^S) + \Pi_{\mathbf{Scene}}((1 - M_c) \odot I_i^S)); \tag{6}$$

$$L^*_{\mathbf{Fake}} = \frac{0.5}{b}\sum_{i}^{b}(\Pi_{\mathbf{Body}}(M_c \odot I_i^D) + \Pi_{\mathbf{Scene}}((1 - M_c) \odot I_i^D)), \tag{7}$$

where M_c and $(1 - M_c)$ represent the body and scene masks that are used to piece-wisely penalized the images, respectively. This extension means that the total loss between real and fake data are calculated together, and written as:

$$L^{Tot}_{\mathbf{Real}} = 0.5 \times (L_{\mathbf{Real}} + L^*_{\mathbf{Real}}); L^{Tot}_{\mathbf{Fake}} = 0.5 \times (L_{\mathbf{Fake}} + L^*_{\mathbf{Fake}}). \tag{8}$$

Thus, the final discriminator and generator losses using 4 adversarial networks are simplified as:

$$L^{Tot}_{\mathbf{Disc}} = 0.5 \times (\|L^{Tot}_{\mathbf{Real}} - 1\|^2 + \|L^{Tot}_{\mathbf{Fake}}\|^2), \tag{9}$$

$$L^{Tot}_{\mathbf{Gen}} = 100 \times L_{\mathbf{Deb}} + 0.5 \times (\|L^{Tot}_{\mathbf{Fake}} - 1\|^2). \tag{10}$$

The constant parameters of 100 in $L_{\mathbf{Deb}}$ is used to balance the error score in generator while the value of 0.5 is an average constant for each LSGAN loss.

5 Experiment Procedure

Another key factor of robust deep learning based algorithm is the quality of its datasets. As mentioned before, to handle the lack of a blurry human dataset, we propose our method. Our method generates a new human image with a new pose (I_k^S) at time-stamp k from the given initial image (I_0^S). I_k^S is extracted by employing the algorithm of [27]. We provide additional foreground discriminator, as shown in the left column of Figure 5, to solve the unrealistic I_k^S result. This foreground prior information is obtained by the gaussian-based segmentation map to

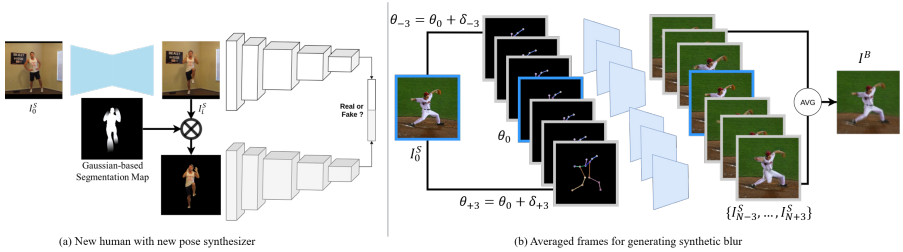

Fig. 5. Synthesizing the same human with a new pose is obtained by a neural network based algorithm. We provide additional discriminator to mask the human region for obtaining the sharp output of the synthesized pose. This process is done sequentially to produce multiple frames to be averaged for final blurry output.

distinguish the human body and background pixels. This map is different from our approach as it is extracted from the sharp image I_0^S. For producing a blurry image, I^B, N frames of I_k^S are extracted and averaged. Each I_k^S has different pre-defined human pose θ_k. This pose is varied differently from θ_0 according to the change parameters, δ_k. The scene background is also translated to simulates camera motion. General visualization of our dataset generation approach is shown on the right column of Figure 5. Details about the change parameters, synthesizer network, and its discriminators are provided in the supplementary material. Note that, our synthesized dataset is only used for training purpose and the images are collected from Leeds Sport Pose dataset [28]. The testing case of blurry human dataset are attained by averaging real blurry videos of InstaVariety dataset [21].

Before running the full experiments, an ablation study is performed to obtain the finest weight of our deblurring module. We divide the ablation procedure by 3 schemes: partial, full, and refined schemes. The partial scheme only includes Eqs. (1) for generator with $\gamma = 1$, (4), and (5) for the discriminators. The full scheme includes whole equations with 4 discriminators in the discriminative modules. Both partial and full-schemes utilize the learning rate of 10^{-4}. The refined scheme has equal structures with the full scheme; however, the learning rate is reduced to 10^{-5} and γ is set to 0 to train on the L1 loss in Eq. (1). Each scheme is initially trained using GoPro dataset until 20K iterations and then our blurry human dataset is included until 65K iterations. Spatial augmentation is done during training with a size of 224 $\times$ 224. The Patch discriminator cropped the fake and real images into 80 $\times$ 80. Batch is set to 8 and the network is backpropagated using ADAM optimizer. The training is done in a TITAN RTX GPU for around 2.5 days for each scheme. Whole implementations are scripted using TensorFlow [29] framework.

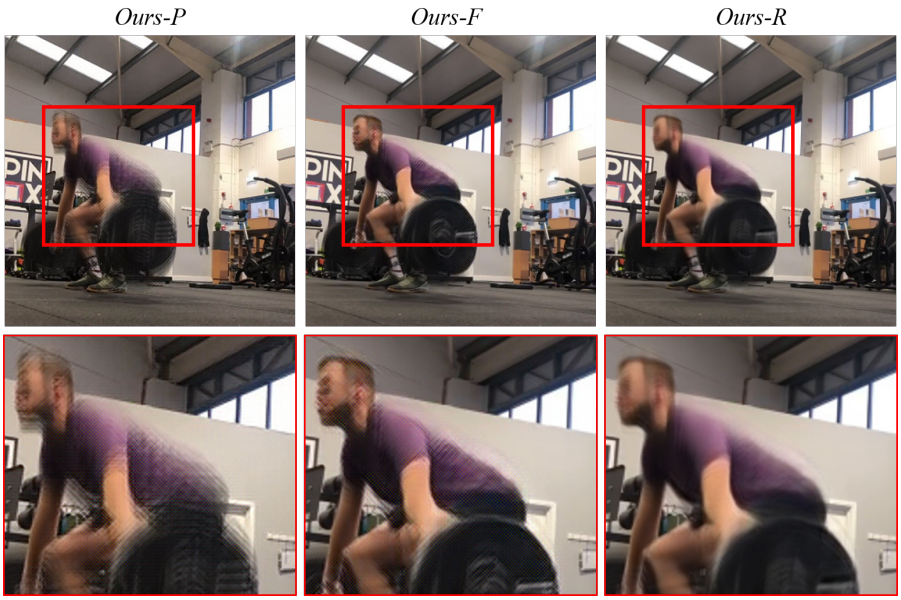

Fig. 6. Visual ablation results between our methods: *Ours-P*, *Ours-F*, and *Ours-R*.

Table 1. Quantitative comparisons on our blurry human test set between the deblurring algorithms. Ablation study is included to show the improvement of our methods. Last row indicates the performance of *Ours-R* approach.

Methods	[13]	[30]	[16]	*Ours-P*	*Ours-F*	*Ours-R*
SSIM	0.869	0.901	0.899	0.803	0.840	0.891
PSNR	33.29	34.93	35.25	33.07	34.56	36.04

6 Experimental Results

In the experiment section, we provide a comparison using our blurry human test set and general deblurring datasets. The general datasets are obtained from the test collections of GoPro [3] and recent HIDE [6] dataset. For comparison, we utilize recent state-of-the-art deblurring algorithms, precisely: DeblurGAN-V1 [13], Deep Hierarchical Multi Patch (DHMP) [30], and DeblurGAN-V2 [16], that are publicly available. The metric of peak signal to noise ratio (PSNR) and structural similarity index (SSIM) are used in the calculation.

In the initial step, we provide ablation study on our approaches: partial (*Ours-P*), full (*Ours-F*), and refined (*Ours-R*) schemes. These schemes are tested in our blurry human test cases. The quantitative scores are shown in the last 3 columns of Table 1. Without the body prior, our method suffers from restoring the blurry region caused by human motion, as shown in the first column of Figure 6. By providing full-scheme approach, that includes body prior, blurry

Fig. 7. Qualitative results between deblurring methods on our human blurry test set. Magnified images signify large motion blur on human region. Last row represents our refined (*Ours-R*) approach.

human motion on the second column of Figure 6 can be restored. However, this approach leaves artifact on the deblurred region caused by the mismatched size of deconvolution filter size when VGG loss is used. Thus, we refine this approach using L1 loss as described in Sec. 4. This strategy successfully restores the blurry human motion without artifact as shown in the third column of Figure 6.

In the second step, we perform a comparison using the state-of-the-art methods using our blurry human dataset. The first 3 columns of Table 1 show the result of other methods. For fairness' sake, those algorithms are fine-tuned using our blurry human training set. It is clearly seen from Figure 7 that our approach provides better visual results compared to others. Significant human motion blur, such in the case of the sixth column of Figure 7, is hardly restored by other state-of-the-art methods. The main reason is that their methods are trained for deblurring camera motion cases only. However, in our case, we em-

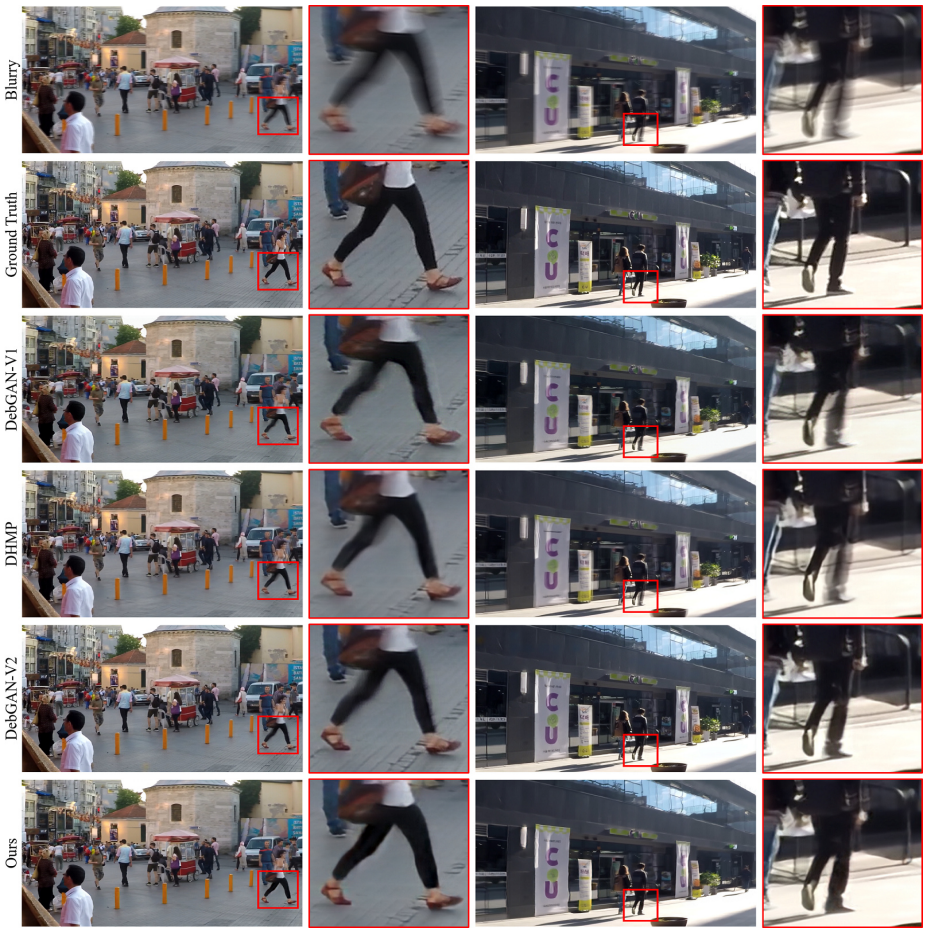

Fig. 8. Qualitative results between deblurring methods on GoPro [3] test set. Last row represents our refined (*Ours-R*) approach.

ploy localized body prior information to train the deblurring network. This prior lets the network learn the blur caused by articulated human motion, which is different from general camera motion. Therefore, our method achieves superior performance in terms of qualitative. These results are also reflected in Table 1 as our approach (*Ours-R*) achieves the highest PSNR. Our method achieves a competitive SSIM score compared to the non-GAN method [30] since the GAN approach produces synthesized pixels during restoration.

In third step, we compare the deblurring algorithms with the GoPro dataset [3] that is widely known for benchmarking. In this case the GoPro dataset contains scene with and without humans and the total test set is 1111 images. The results in Table 2 show that our network able to outperforms previous deblurring methods. Note that our localized approach is done on image with single human in the

Table 2. Quantitative comparisons on GoPro test set [3] between the deblurring algorithms. Our method with refined approach (*Ours-R*) is used for the experiment.

Methods	[13]	[30]	[16]	*Ours-R*
SSIM	0.958	0.940	0.9340	0.805
PSNR	28.70	31.20	29.55	32.51

Table 3. Quantitative comparisons on HIDE test set [6] between the deblurring algorithms. Our method with refined approach (*Ours-R*) is used for the experiment.

Methods	[13]	[30]	[6]	[16]	*Ours-R*
SSIM	0.871	0.924	0.931	0.875	0.778
PSNR	24.51	29.09	28.89	26.61	32.76

middle during training. However, our deblurring network is fully-convolutional. Thus, multiple humans that present in the GoPro dataset, are well-deblurred using our method. Table 2 shows new record as *Ours-R* achieves the best score in terms of PSNR. Qualitative results are shown in Figure 8. Our deblurring method that is trained for both camera and human motion blurs handles the blurry region eloquently by restoring some blurry scene's edges. Other methods show similar performance except in the case of restoring large blurry human motion. Second and fourth columns of Figure 8 show magnified results of the blurry case when people are walking. These results clearly show that the articulated human motion is solved by our method with the more faithful result compared to others. Moreover, the blurry non-human region is also restored similarly compared to other state-of-the-art methods. This is achieved by our network as it is guided by the non-human region $(1 - M_c)$ during the optimization procedure. Best quantitative performance using this dataset is also achieved by our method, as shown in Table 2. For our final step, we also compare using the recent blurry dataset, known as HIDE [6], as they provide whole images with the presence of humans. Note that the HIDE dataset of [6] contains long-shot and close-up blurry human images, and most of the blur is caused by the scene motion. Our approach achieves the highest score in terms of PSNR compared to other methods, as shown in Table 3. Additional visual results are included in the supplementary material.

From these experiments, our method achieves state-of-the-art performance as it solves both camera and human articulated motion blurs. The human prior works well on guiding the discriminator, which eventually trains the deblurring generator on distinguishing human and non-human blurry regions.

7 Conclusion

While current deblurring methods perform well, most of them only focus on scene motion blur case. Human motion deblurring plays a crucial role in the recent computer vision's 3D body reconstruction task. In this paper, we explore several methods to handle human motion deblurring, specifically: we introduce localized body prior that guide the network to give more attention on the human region; we introduce adversarial framework from human prior that helps network on restoring blurred human and scene regions; and finally, we also introduce synthetic human motion blur dataset to train on the network. From experimental results, we show that our approach is able to reach state-of-the-art performance on both human and scene motion deblurring. We believe this exploration can be applied to various human-based image processing tasks.

Acknowledgement. This work was supported by Samsung Research Funding Center of Samsung Electronics under Project Number SRFCIT1901-06. This work was supported by Inha University Research Grant.

References

1. Krishnan, D., Tay, T., Fergus, R.: Blind deconvolution using a normalized sparsity measure. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. (2011) 233–240
2. Kim, T.H., Lee, K.M.: Segmentation-free dynamic scene deblurring. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. (2014) 2766–2773
3. Nah, S., Kim, T.H., Lee, K.M.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 257–265
4. Shen, Z., Lai, W.S., Xu, T., Kautz, J., Yang, M.H.: Deep semantic face deblurring. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 8260–8269
5. Ren, W., Yang, J., Deng, S., Wipf, D., Cao, X., Tong, X.: Face video deblurring using 3D facial priors. In: Proc. of the IEEE International Conference on Computer Vision. (2019) 9387–9396
6. Shen, Z., Wang, W., Lu, X., Shen, J., Ling, H., Xu, T., Shao, L.: Human-aware motion deblurring. In: Proc. of the IEEE International Conference on Computer Vision. (2019) 5571–5580
7. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: a skinned multi-person linear model. ACM Trans. on Graphics **34** (2015) 248:1–248:16
8. Pan, J., Hu, Z., Su, Z., Yang, M.H.: Deblurring text images via L_0-regularized intensity and gradient prior. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. (2014) 2901–2908
9. Sellent, A., Rother, C., Roth, S.: Stereo video deblurring. In: Proc. of the European Conference on Computer Vision. Volume 9906., Springer (2016) 558–575
10. Srinivasan, P.P., Ng, R., Ramamoorthi, R.: Light field blind motion deblurring. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 2354–2362

11. Lumentut, J.S., Kim, T.H., Ramamoorthi, R., Park, I.K.: Deep recurrent network for fast and full-resolution light field deblurring. IEEE Signal Processing Letters **26** (2019) 1788–1792
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems. (2014) 2672–2680
13. Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., Matas, J.: DeblurGAN: Blind motion deblurring using conditional adversarial networks. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 8183–8192
14. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of Wasserstein GANs. In: Advances in Neural Information Processing Systems. (2017) 5767–5777
15. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 5967–5976
16. Kupyn, O., Martyniuk, T., Wu, J., Wang, Z.: DeblurGAN-v2: Deblurring (orders-of-magnitude) faster and better. In: Proc. of the IEEE International Conference on Computer Vision. (2019) 8877–8886
17. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. In: Advances in neural information processing systems. (2015) 2017–2025
18. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and Locally Consistent Image Completion. ACM Trans. on Graphics **36** (2017) 107:1–107:14
19. Yuan, X., Park, I.K.: Face de-occlusion using 3D morphable model and generative adversarial network. In: Proc. of the IEEE International Conference on Computer Vision. (2019) 10061–10070
20. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 7122–7131
21. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3D human dynamics from video. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 5614–5623
22. Pavlakos, G., Kolotouros, N., Daniilidis, K.: Texturepose: Supervising human mesh estimation with texture consistency. In: Proc. of the IEEE International Conference on Computer Vision. (2019) 803–812
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 770–778
24. Blanz, V., Vetter, T.: Face recognition based on fitting a 3D morphable model. IEEE Trans. on Pattern Analysis and Machine Intelligence **25** (2003) 1063–1074
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
26. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proc. of the IEEE International Conference on Computer Vision. (2017) 2813–2821
27. Balakrishnan, G., Zhao, A., Dalca, A.V., Durand, F., Guttag, J.: Synthesizing images of humans in unseen poses. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 8340–8348
28. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: Proc. of the British Machine Vision Conference, BMVA Press (2010) 1–11

29. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: a system for large-scale machine learning. In: 12th USENIX Conference on Operating Systems Design and Implementation. Volume 16. (2016) 265–283
30. Zhang, H., Dai, Y., Li, H., Koniusz, P.: Deep stacked hierarchical multi-patch network for image deblurring. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 5978–5986