Human from Blur: Human Pose Tracking from Blurry Images

Yiming Zhao[1] Denys Rozumnyi[1] Jie Song[1]
Otmar Hilliges[1] Marc Pollefeys[1,2] Martin R. Oswald[1,3]
[1]ETH Zürich [2]Microsoft [3]University of Amsterdam

Abstract

We propose a method to estimate 3D human poses from substantially blurred images. The key idea is to tackle the inverse problem of image deblurring by modeling the forward problem with a 3D human model, a texture map, and a sequence of poses to describe human motion. The blurring process is then modeled by a temporal image aggregation step. Using a differentiable renderer, we can solve the inverse problem by backpropagating the pixel-wise reprojection error to recover the best human motion representation that explains a single or multiple input images. Since the image reconstruction loss alone is insufficient, we present additional regularization terms. To the best of our knowledge, we present the first method to tackle this problem. Our method consistently outperforms other methods on significantly blurry inputs since they lack one or multiple key functionalities that our method unifies, i.e. image deblurring with sub-frame accuracy and explicit 3D modeling of non-rigid human motion.

1. Introduction

Accurate tracking of human motion is often crucial for understanding dynamic scenes from images. Human motion estimation has a wide field of applications such as improving human-robot collaboration [2], human-machine interaction in general [12], better safety for autonomous driving [20], markerless human motion capture [36, 35, 27], sports analysis, and the movie and entertainment industry. A particular difficulty occurs when the human motion is fast, or low light conditions demand longer camera exposure times, which can both lead to blurry images from which it is significantly harder to estimate the human pose.

The main goal of our method is accurate 3D human pose tracking from substantially blurred images or videos. Hence, it is related to both human pose estimation and image deblurring methods. On the one hand, while there is a variety of methods that address 3D human pose estimation from RGB or RGB-D images, there is no method that is designed to handle substantially blurred images. Moreover, none of the human pose estimation methods is able to estimate human pose at sub-frame accuracy. On the other hand, there is a large amount of methods that aim at deblurring images and videos, but they mostly only assume simplified scenarios, *e.g.* without out-of-image-plane object rotations, or only for rigidly moving objects [45, 46]. So far, human pose estimation and image deblurring has not been studied jointly. Also, there is no public dataset to evaluate such task since none of standard datasets for human pose estimation include significant amounts of motion blur.

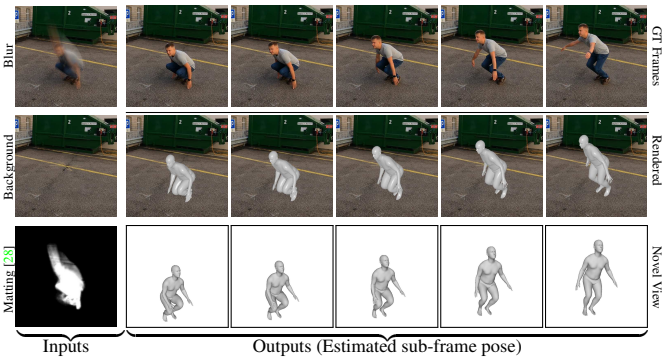

Figure 1. **Human from Blur (HfB) on a real-world sequence.** Given a blurry image with human motion and the corresponding background, HfB recovers the human shape and sub-frame motion. We visualize sub-frame human pose and show the reconstructed mesh from a novel view.

We propose the first method that recovers human pose at sub-frame accuracy from blurry inputs, even from a single blurry image (Fig. 1). We make the following contributions:

(1) We present the first method for human pose estimation from substantially blurred images that recovers sub-frame accurate poses as well as texture and body shape.

(2) We generate a synthetic dataset and collected real-world motion-blurred data of humans for evaluation purposes. We further propose corresponding evaluation metrics to assess and compare to future methods.

(3) The proposed method only relies on test-time optimization and is learning-free, apart from the initialization

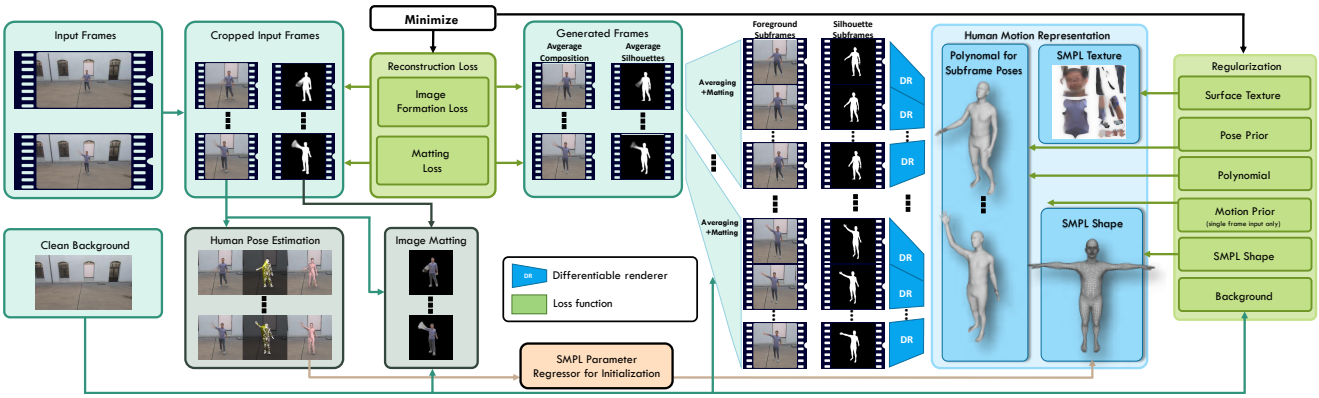

Figure 2. **Method overview.** The input to our method are a single or multiple blurry frames of a human (left), and the output is a 3D representation of a human and its sub-frame motion over time (right). **From Right to Left:** Starting from the human motion representation, our model can be seen as generative model. For a desired set of frames and sub-frames, we can render sub-frame appearances and corresponding silhouettes. Then, the sub-frames are averaged to generate blurry frames and blurry silhouettes (alpha channel), which are composed with the known background to generate the input image according to (2). The central part of our method is the image reconstruction loss which compares the generated images with the actual input images. In order to solve for the human motion estimation, the reconstruction loss is backpropagated through the entire differentiable pipeline. The human pose estimation uses a traditional method [27] to initialize the optimization, and the image matting is precomputated [28] for the matting loss.

and the motion prior, which is only needed for the single-frame case. Hence, our method does not require large amounts of annotated training data.

2. Related work

The proposed method is at the intersection of human pose tracking and image/video deblurring.

3D Human pose estimation. The 3D pose of a human is usually represented as a skeleton of 3D joints [34, 36, 35, 65, 51]. In order to obtain more fine-grained representations of the human body, parametric body models such as SCAPE [3] or the SMPL family [41, 42, 38] have been introduced to capture the 3D body pose. Iterative optimization-based approaches have been leveraged for model-based human pose estimation. [9, 48, 11, 4, 42] proposed to estimate the parameters of the human model by leveraging silhouettes or 2D keypoints. On the other hand, direct parameter regression via neural networks has been explored [15, 53, 56, 37, 10, 54, 58, 61, 24, 52, 27, 18]. Given a single RGB image, a deep network is used to regress the human model parameters. There is another line of work that combines the advantages of both optimization and regression to fit the SMPL body [19, 49]. Although there have been significant advances of human pose estimation from monocular images or videos, a method which is able to deal with blurry input is still missing.

Image and video deblurring. A large amount of methods have studied generic image [21, 22] and video deblurring, *e.g.* [14, 39, 62, 47, 64, 7, 23, 60, 50, 16, 43, 44]. Some attempts to specialize on deblurring depicted humans have already been made. For instance, [8] focuses only on deblurring human faces. Closely related to our problem setting, [31] addresses deblurring of human motion using an adversarial approach, which focuses on image deblurring rather than pose estimation, and it does not recover at sub-frame accuracy. The follow-up method [30] generalizes to joint human motion and scene deblurring with a similar methodology, but the sub-frame poses are never recovered.

The proposed method is partially inspired by Shape from Blur (SfB) [45], which uses a similar test-time optimization to recover 3D shape and sub-frame motion of simple rigid objects with spherical topology from a single blurry image with a given background. Motion from Blur (MfB) [46] extends SfB to multiple video frames. There is also a related Animation from Blur method [63], but it assumes a motion guidance is provided as an additional input.

3. Method

The inputs to our method are an image I with the blurred human and the corresponding clean background image B. The desired output is a human shape parameter β, texture image $\mathcal{T}$, and three functions representing sub-frame human motion that depend on a timestamp i. This timestamp represents the sub-frame time interval and is defined between 1 and N, where N is the desired number of sub-frames. Effectively, it means that we generate a temporal super-resolution or a short video with N frames out of each single input frame. Those three functions are human body translation T_i, rotation R_i, and sub-frame human pose θ_i that represents joint rotation. They are all represented by a set of low-degree polynomials, where translations and rotations have each four degrees of freedom (direction with

distance and axis with angle). This polynomial representation generates poses in a strict chronological order and is continuous, differentiable, and can be easily initialized with a given initial pose (Sec. 3.3). Human pose and shape representations follow the SMPL human model [41]. The texture image $\mathcal{T}$ is mapped using a fixed UV mapping from SMPL.

As the first step, we generate the human SMPL mesh Θ_i at timestamp i with a given pose, shape, and texture parameters (Fig. 2). Then, we move the whole mesh Θ_i by translation T_i and rotation R_i given by motion function $\mathcal{M}$:

$$\Theta_i = \mathcal{M}\big(\text{SMPL}(\theta_i, \beta, \mathcal{T}), T_i, R_i\big) \ . \tag{1}$$

To render the sub-frame silhouette and appearance of the mesh, we use Differentiable Interpolation-based Renderer (DIB-R) [6]. This differentiable rendering provides two outputs. The first one is appearance rendering $\mathcal{R}_F(\Theta_i)$ that outputs projected human appearance. The second one is silhouette rendering $\mathcal{R}_S(\Theta_i)$ that outputs projected human silhouette. In this work, we assume a static camera.

Image formation model. Given all previously defined parameters, we can finally define the image formation model. It follows a standard alpha matting approach:

$$\hat{I} = \underbrace{\Big(1 - \frac{1}{N}\sum_{i=1}^{N} \mathcal{R}_S(\Theta_i)\Big)}_{\text{Inverse alpha channel}} \cdot \underbrace{B}_{\text{Background}} + \underbrace{\frac{1}{N}\sum_{i=1}^{N} \mathcal{R}_S(\Theta_i) \cdot \mathcal{R}_F(\Theta_i)}_{\text{Blurred foreground (human)}} \ . \tag{2}$$

The generated image $\hat{I}$ consists of the background image, scaled down by the inverse alpha channel, and the blurred foreground human body. The alpha channel is modeled by averaging all projected sub-frame silhouettes.

3.1. Loss terms

The key components of our method are the image formation loss and the matting loss. The image formation loss forces the reconstructed image to be as close as possible to the input image. The matting loss favors silhouettes that are consistent with the initially estimated alpha channel. The other losses are auxiliary and regularization terms that make the optimization easier and refine the final results.

Image formation loss. This loss measures the input image reconstruction according to the image formation model (2). We compute the mean squared error between the observed input image and our reconstruction as:

$$\mathcal{L}_I = |I - \hat{I}|_2 \ . \tag{3}$$

Matting loss. If the image formation loss (3) is the sole loss to be minimized, the optimization becomes extremely difficult. Experimentally, such optimization is ambiguous and mostly results in an undesired local minimum. Therefore, we further impose a loss on our approximated rendered alpha channel, which is computed as the average of sub-frame silhouettes, $\alpha_{\text{target}} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{R}_S(\Theta_i)$, according to the image formation model (2). The initial alpha channel α_{in} is estimated using a pre-trained Background-Matting-V2 [28] model, based on the input blurry image and the corresponding background. Finally, the matting loss computes the intersection over union between our rendered alpha channel from averaging and the one from [28]:

$$\mathcal{L}_\alpha = 1 - \frac{|\min(\alpha_{\text{in}}, \alpha_{\text{target}})|_1}{|\max(\alpha_{\text{in}}, \alpha_{\text{target}})|_1} \ , \tag{4}$$

where the intersection over union for non-binary inputs is a ratio between the sum of pixel-wise $\min$ and $\max$ operators.

Surface texture smoothness. The UV texture map from SMPL contains many non-overlapping regions (see Fig. 3, HfB row), and the correct neighborhoods are not properly defined. Therefore, the commonly used total variation loss for texture smoothness [46] cannot be directly applied in this case since it will propagate the color of the void area. To address this issue, we propose a surface texture smoothness term that accounts for the mesh faces neighborhood. For a given texture pixel p_k and its 8 surrounding neighboring pixels $p_j \in \mathcal{N}(p_k)$, we pick those ones that are neighbors in the mesh ($\mathbf{c}_{k,j} = 1$), *i.e.* they belong to adjacent triangular faces, and that are visible in at least one of the sub-frames ($\mathbf{v}_j = 1$). Then, we compute the cosine value between the face normal n_k of the current pixel and the face normal n_j of its chosen neighbors. The introduction of the cosine of face normals takes into account the mesh geometry, *i.e.* the texture should be smoother on flat surfaces. Then, the surface texture smoothness is expressed as a weighted sum of absolute differences in RGB pixels:

$$\mathcal{L}_S = \frac{1}{8|\mathcal{T}|} \sum_{p_k \in \mathcal{T}} \sum_{p_j \in \mathcal{N}(p_k)} \mathbf{c}_{k,j} \mathbf{v}_j \cos \angle(n_k, n_j) |p_k - p_j|_1 \ . \tag{5}$$

Pose prior loss. We import the pose prior loss from SMPLify-X [42]. This prior scores how feasible are the estimated pose parameters θ_i:

$$\mathcal{L}_P = \frac{1}{N}\sum_{i=1}^{N} \text{prior}(\theta_i) \ . \tag{6}$$

SMPL shape regularization. We add norm regularization on the SMPL shape parameter β to avoid irregular human body shape as used in SMPL [41]:

$$\mathcal{L}_\beta = |\beta|_2^2 \ . \tag{7}$$

Polynomial regularization. The polynomial coefficients of the pose, translation, and rotation could be serialized into a matrix $\boldsymbol{C} \in \mathbb{R}^{4d \times (J+2)}$, where d is the degree of the polynomial, and J is the number of joints in the SMPL model.

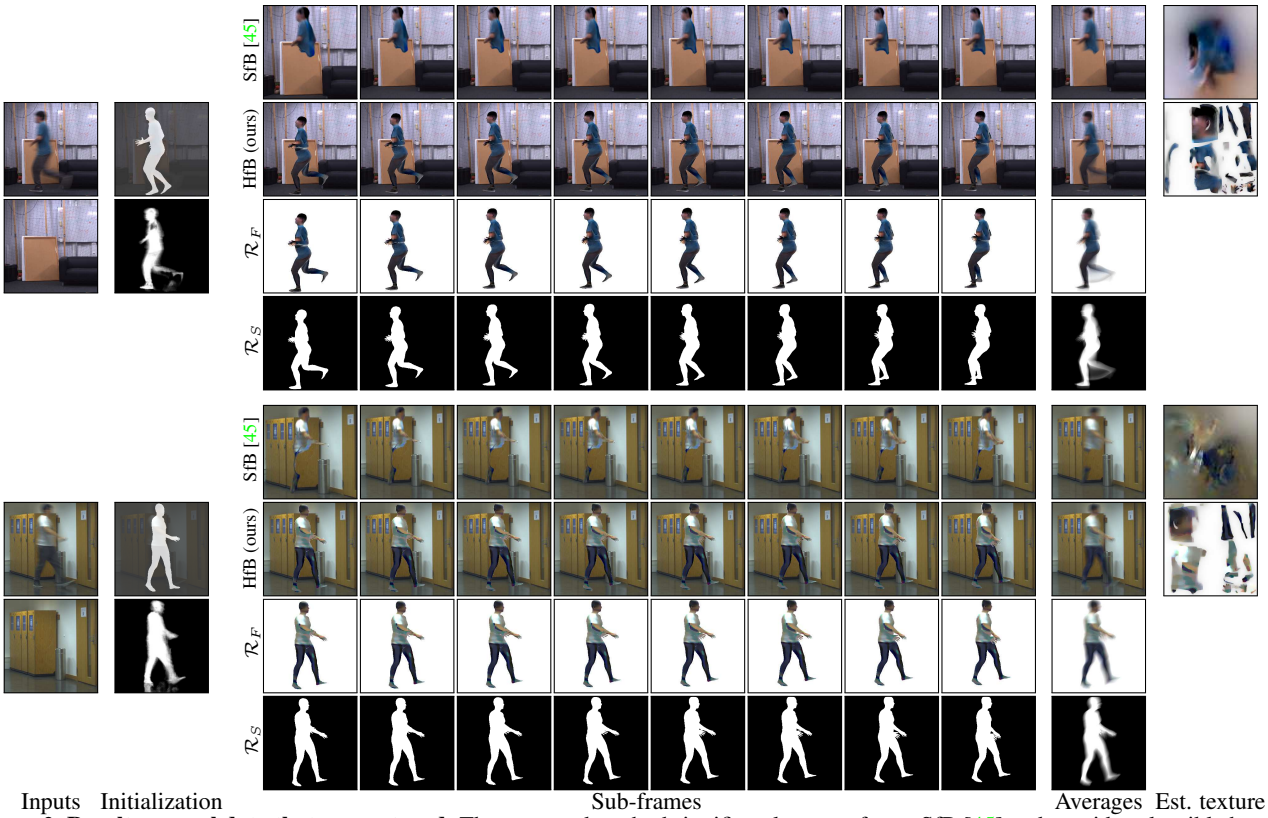

Figure 3. **Results on real data that we captured.** The proposed method significantly outperforms SfB [45] and provides plausible human shape and pose reconstructions. **Left**: initialization of human pose from METRO [27] (top) and alpha channel from [28] (bottom).

The whole body translation and rotation are already incorporated into matrix $\boldsymbol{C}$, thus we have $4(J+2)$ polynomials of d degree. Since rotations, translations, and joint poses have 4 degrees of freedom each, we have a separate polynomial for each degree of freedom. We apply both the L1-norm and the Frobenius norm on the polynomial coefficients:

$$\mathcal{L}_C = |\boldsymbol{C}|_1 + |\boldsymbol{C}|_F \ . \tag{8}$$

The intention of adding this regularization is to avoid extreme joint movement.

Background regularization. We assume that the human texture is sufficiently distinct from the background. This is enforced by the difference between the projected object appearance and the background:

$$\mathcal{L}_B = \frac{1}{N} \sum_{i=1} \frac{1}{|B - \mathcal{R}_F(\Theta_i)| + \epsilon} \quad \text{with } \epsilon = 10^{-6} \ . \tag{9}$$

Adversarial short motion prior. Since the human body consists of multiple joints, there exists a significant amount of ambiguity in case of a single input blurry image. The ambiguity comes mainly from the unknown motion direction. In fact, both the forward and the backward directions provides the same blurry image according to the image formation model (2). Potentially, there are exponentially many motion directions for each joint that lead to the same input. And it is infeasible to estimate the correct direction directly from a single image without any additional priors. Otherwise, the choice of motion direction will be arbitrary. Many prior studies [42, 18] offer motion priors, but they are not suitable for our setting.

To address this problem we propose the adversarial motion prior to recognize wrong (reversed) motion of joints. Based on our polynomial motion representation, we propose an adversarial model that could supervise on the polynomial coefficients $\boldsymbol{C}$. The model is inspired by the image in-painting methods [40, 29].

Our adversarial model consists of two components: a discriminator D that generates a binary indicator function to identify unrealistic entries in the coefficients $\boldsymbol{C}$, and a correction-generator G that predicts realistic polynomial coefficients from the given polynomial coefficients $\boldsymbol{C}$ and the indicator function $\boldsymbol{I_c}$.

The training data are sampled from the AMASS dataset [33] (CMU [5] and ACCAD [1]). The training is supervised jointly by four loss terms. The discriminator loss is the binary cross entropy loss, which is applied to the indicator function predicted by the discriminator and compared to the ground truth. The generator loss, specifically

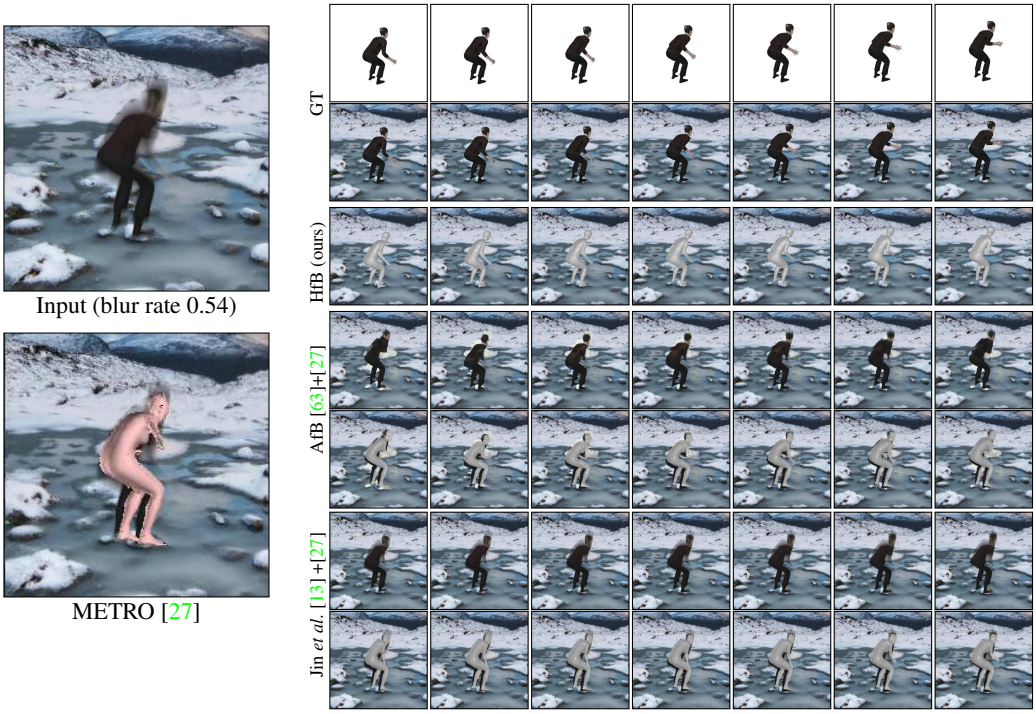

Figure 4. **Comparison on synthetic data.** Given a blurry image with human motion, sub-frame human poses generated by HfB are consistent, whereas for learning-based temporal super-resolution methods [13, 63] (with METRO [27] human poses on sub-frames) the poses are not consistent, *e.g.* motion of the right arm. We also visualize the raw METRO [27] pose prediction on the input blurry image.

the reconstruction loss, comprises three terms. The first one is L1 loss between the coefficient matrix predicted by the correction- generator and the ground truth. The second is L2 loss between the reconstructed pose and the ground truth pose. The last one is the mean per joint position error (MPJPE) [49] between the reconstructed joint positions and the ground truth SMPL joint positions.

This adversarial model is pre-trained as mentioned above and is fixed during optimization. In case of a single input blurry image, the adversarial motion prior is incorporated into the optimization as the L1 loss between the generator output and the polynomial coefficients $\boldsymbol{C}$:

$$\mathcal{L}_A = |G(D(\boldsymbol{C}), \boldsymbol{C}) - \boldsymbol{C}|_1 \ . \tag{10}$$

Joint loss. The final loss is a weighted sum of all previously defined losses:

$$\begin{aligned} \mathcal{L} = & w_I \mathcal{L}_I + w_\alpha \mathcal{L}_\alpha + w_S \mathcal{L}_S + w_P \mathcal{L}_P + \\ & w_\beta \mathcal{L}_\beta + w_C \mathcal{L}_C + w_B \mathcal{L}_B + w_A \mathcal{L}_A \ . \end{aligned} \tag{11}$$

3.2. Multiple blurry images

Our approach can be extended to multiple consecutive blurry images in a video. In this case, the human body shape β and texture $\mathcal{T}$ are assumed to be the same for all input images, while the other parameters, *e.g.* poses, are separate for each frame. In general, this setting is simpler since there are more constraints from more images. Also, there is no more ambiguity in the motion direction of each joint. Therefore, the adversarial motion prior $\mathcal{L}_A$ is not needed anymore.

For smooth joint motion in consecutive frames, we add a boundary restriction on the joints rotation and position. For instance, in case of two input blurry images, we add a boundary restriction at the end timestamp $i = N$ of the first image and the start timestamp $i = 1$ of the second image. The boundary restriction forces the joints rotation and position and their first order derivatives to be equal at the boundary to preserve the motion continuity, and it is implemented by the L1 loss with a unit weight. For images with exposure gap, we extend the end timestamp of first image with exposure time τ (measured in sub-frames) and then apply boundary restriction at $N + \tau$.

3.3. Optimization

The joint loss (11) is minimized using the ADAM optimizer [17] for 200 iterations with learning rate 0.01 on a single 12 GB RTX 2080 Ti graphics card.

Initialization. To initialize our method, we use the METRO [27] human pose estimation method, which reconstructs a single human pose from a blurry image reasonably well, albeit without sub-frame accuracy. We fit the initial body translation, rotation, pose, and shape parameters to

blur rate int.	[0.05, 0.1]		[0.1, 0.2]		[0.2, 0.3]		[0.3, 0.4]		[0.4, 0.5]		[0.5, 0.6]		[0.6, 0.7]		[0.7, 0.8]		[0.8, 0.9]		[0.9, 1.1]	
#images	123		206		154		139		106		108		81		79		76		52	
Method	MPJPE↓	IoU↑	MPJPE↓	IoU↑	MPJPE↓	IoU↑	MPJPE↓	IoU↑	MPJPE↓	IoU↑	MPJPE↓	IoU↑	MPJPE↓	IoU↑	MPJPE↓	IoU↑	MPJPE↓	IoU↑	MPJPE↓	IoU↑
SfB[45]	N.A.	0.498	N.A.	0.487	N.A.	0.493	N.A.	0.455	N.A.	0.428	N.A.	0.408	N.A.	0.409	N.A.	0.397	N.A.	0.378	N.A.	0.363
METRO [27]	70.0	0.632	71.9	0.637	84.1	0.615	101.2	0.590	121.3	0.540	121.4	0.500	132.8	0.489	143.4	0.428	147.1	0.421	146.4	0.423
HfB w/o AMP	65.1	0.829	65.7	0.813	68.7	0.803	80.6	0.785	98.5	0.763	107.2	0.739	115.1	0.731	129.0	**0.685**	**121.7**	**0.670**	**138.9**	**0.645**
HfB (ours)	**56.3**	**0.859**	**59.4**	**0.837**	**66.4**	**0.820**	**78.3**	**0.805**	**89.0**	**0.775**	**101.1**	**0.743**	**110.8**	**0.734**	**128.7**	0.682	124.6	0.659	140.5	0.633

Table 1. **Single-frame evaluation for different blur rates on BT-AMASS dataset.** The proposed method outperforms SfB [45] (no human pose output) and METRO [27], which we also use for initialization. Our method improves significantly over the initialization. The proposed AMP prior (Sec. 3.1) improves results only slightly, and even becomes harmful for higher blur rates due to more ambiguity.

blur rate int.	[0.05, 0.1]		[0.1, 0.2]		[0.2, 0.3]		[0.3, 0.4]		[0.4, 0.5]		[0.5, 0.6]		[0.6, 0.7]		[0.7, 0.8]		[0.8, 0.9]		[0.9, 1.1]	
#videos	35		39		47		42		38		32		25		19		13		15	
Method	MPJPE↓	IoU↑	MPJPE↓	IoU↑	MPJPE↓	IoU↑	MPJPE↓	IoU↑	MPJPE↓	IoU↑	MPJPE↓	IoU↑	MPJPE↓	IoU↑	MPJPE↓	IoU↑	MPJPE↓	IoU↑	MPJPE↓	IoU↑
MfB [46]	N.A.	0.513	N.A.	0.547	N.A.	0.528	N.A.	0.509	N.A.	0.384	N.A.	0.356	N.A.	0.342	N.A.	0.281	N.A.	0.268	N.A.	0.229
METRO [27]	68.9	0.645	69.6	0.689	71.6	0.653	67.0	0.601	98.6	0.598	99.6	0.508	108.5	0.477	111.3	0.399	116.8	0.408	121.1	0.388
HfB (ours)	**65.4**	**0.819**	**64.4**	**0.828**	**69.4**	**0.823**	**56.2**	**0.787**	**77.2**	**0.779**	**83.4**	**0.738**	**102.4**	**0.766**	**101.4**	**0.695**	**109.4**	**0.671**	**112.0**	**0.656**

Table 2. **Two-frame evaluation for different blur rates on BT-AMASS dataset.** Similarly to Table 1, HfB outperforms other methods even when there are two input blurry frames. We compare to MfB [46] (multi-frame method) and interpolated poses from METRO [27].

the mesh generated from METRO using a SMPL registration model [57]. In case of sub-frame translation, rotation, and poses, we initialize the polynomial coefficients at timestamp $i = 1$ and all other coefficients to zero.

4. Experiments

Among the chosen baselines, we selected SfB [45] and MfB [46], designed for simple objects sub-frame deblurring and 3D reconstruction. Then, we compare our method to general temporal super-resolution methods, *i.e*. Jin *et al*. [13] and Animation-from-Blur (AfB) [63] for single frame experiments and Blurry Video Frame Interpolation (BIN) [47] for multi-frame experiments. To make them competitive, we also apply 3D human pose estimation METRO [27] on top of their deblurred sub-frames (temporal super-resolution), except for SfB and MfB, where the output sub-frames are of low quality, and human pose estimation methods do not detect anything.

Blur rate. In general, motion blur is determined by many factors. However, the main factors are the camera exposure time and the speed of the object motion. Even with those two factors, it is still a challenging task to quantify the exact amount of motion blur. In order to measure the approximate blur level, we define blur rate as:

$$\text{blur rate} = \frac{|\bigcup_{i=1}^{N} \mathcal{R}_S(\Theta_i)|_1}{|\mathcal{R}_S(\Theta_1)|_1} - 1 \ . \tag{12}$$

Here, we compute the union of all projected sub-frame silhouettes and divide it by the first silhouette. When the human stays still, the blur rate value is zero. When the human moves over a distance larger than its size within one blurry frame, *i.e*. there is no overlap between the rendered silhouettes at the first and last timestamps, the blur rate is larger than one. We use this blur rate to classify the experiments.

	original	2 avg. frames	3 avg. frames
avg. blur rate	0.27	0.36	0.45
Method	PA-MPJPE(mm)↓	PA-MPJPE(mm)↓	PA-MPJPE(mm)↓
HfB (ours)	69.1	77.3	**81.4**
AfB [63]	**52.3**	**63.3**	87.3
Jin *et al*. [13]	55.3	81.6	96.1

Table 3. **Results on B-AIST++ [63] dataset.** We average 2 and 3 original blurry frames to increase the blur amount. We apply METRO [27] on top of the output of two baselines [14, 63].

4.1. Synthetic datasets

We generated two datasets: BC-CAPE (Blur-Clothed CAPE [32]) and BT-AMASS (Blur-Textured AMASS [33]). The BT-AMASS is sampled on real-world human poses θ_i, rotations R_i, and translations T_i from the ACCAD [1] and CMU [5] dataset of the AMASS [33] database with 120 fps. The UV textures $\mathcal{T}$ are sampled from the SURREAL [55] dataset. Finally, the background images are randomly selected from a set of random images from the BG-20K [25] database, capturing both indoors and outdoors scenes. We take random motion captures with length of 5 to 60 frames. This covers blur rates in the range between 0.05 and 1.1. We utilize the SMPL-X plugin [42] in Blender to generate dataset images.

The BC-CAPE is based on the CAPE dataset [32], which contains SMPL human models with poses for each frame with 60 fps. For BC-CAPE, we interpolate human poses to render higher speed footage. The camera position is randomly selected facing the human model. Then, we render sub-frame silhouettes $\mathcal{R}_S(\Theta_i)$ and appearances $\mathcal{R}_F(\Theta_i)$. In the end, we average these rendered silhouettes and appearances to acquire blurry images according to the image formation model (2). The jittering effect is eliminated

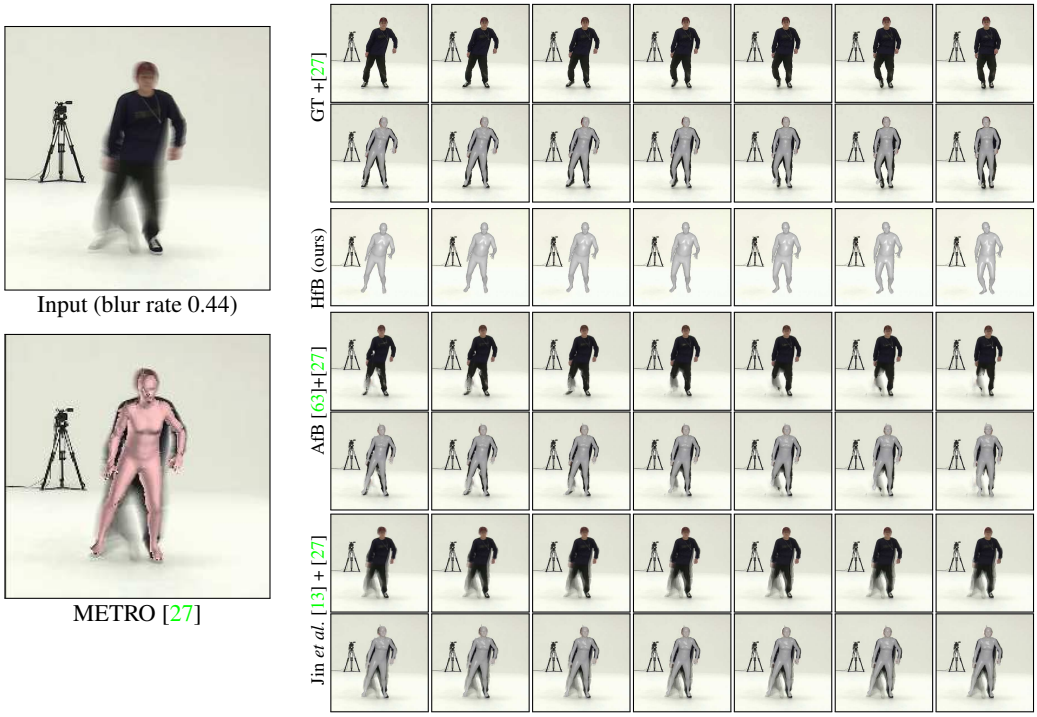

Figure 5. **Comparison on real data.** We evaluate our method on the real B-AIST++ [63] dataset. This example shows an average of 3 frames (see Table 3). Our method produces more consistent and accurate sub-frame human poses compared to carefully selected baselines.

by up-sampling and interpolation at a high frame rate of 600 fps. In total, we generated 1861 blurry images for a single frame experiment and 305 short videos with two video frames to evaluate our multi-frame setting.

Evaluation metrics. We evaluate HfB on the joint position error in millimeters: Mean Per Joint Position Error (MPJPE) and Procrustes Analysis MPJPE (PA-MPJPE) as in [49]. For MPJPE, we initially align the coordinate axis orientation of the predicted motion sequence with the ground truth. For comparison to SfB and MfB, we also measure the intersection-over-union (IoU) between the generated silhouette and the ground truth one.

4.2. Results on BT-AMASS

First, we compare our method on the generated BT-AMASS dataset to the following baselines: SfB [45], MfB [46], and static interpolated METRO [27] (1124 single frames and 305 two-frames). The single-frame results with 1124 images are shown in Table 1, whereas multi-frame results in Table 2. Our method outperforms all baselines by a wide margin, especially for higher blur rate intervals. As excepted, the performance steadily decreases with the increased blur rate. Additionally, we evaluate the influence of the Adversarial Motion Prior (AMP), which is only used for single-frame experiment. This prior improves results only for higher blur raters, whereas for lower blur raters, where there is less ambiguity, it is harmful.

4.3. Results on BC-CAPE

Next, we evaluate the proposed method on the generated BC-CAPE dataset, which contains 609 single frames and 205 short sequences with 4-frames. In this case, we compare to three temporal super-resolution methods: AfB [63], Jin *et al*. [13], and BIN [47]. For fair comparison, we augmented their sub-frame output with human pose estimation metods, either METRO [27] or HybrIK [24]. As shown in Tables 5 and 4, the proposed Human from Blur (HfB) method outperforms these baselines by a large margin, especially on larger blur rates. The performance gain is even higher for the multi-frame experiment (Table 5).

4.4. Real dataset B-AIST++

Finally, we evaluate on the real-world dataset with various human motion and garments: B-AIST++ [63]. They use frame interpolation to generate high-speed frames from original dancing dataset AIST++ [26]. B-AIST++ provides significantly blurred images with human motion. We generate ground-truth sub-frame human pose by running METRO [27] on top of the ground-truth sub-frames. Table 3 shows that our method outperforms other baselines when 3 consecutive frames are averaged, which translates to blur rate 0.45. Note that AfB [63] is trained on this dataset,

Figure 6. **Multi-frame evaluation.** We compared to BIN [47], with METRO [27] human poses on top of their sub-frames. The visual results show that BIN fails, however METRO is still robust to some amount of blur and detects human poses, which are not consistent over time. The proposed method generates motion which is more consistent with the ground truth.

blur rate	<0.2		[0.2,0.3]		[0.3,0.4]		[0.4,0.5]		[0.5,0.6]		[0.6,0.8]	
Method	PA-MPJPE ↓	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓	MPJPE↓
HfB (ours)	75.2	81.2	76.4	83.0	84.6	94.2	**89.7**	**100.5**	**96.2**	**110.6**	**98.5**	**114.6**
AfB [63] + METRO [27]	82.4	89.1	84.1	89.2	90.6	100.8	105.3	119.9	107.5	133.3	112.8	135.2
Jin *et al*. [14] + METRO [27]	**74.4**	**77.2**	**76.2**	**82.0**	**82.6**	**90.8**	100.1	112.8	99.5	119.2	105.6	124.6
Jin *et al*. [14]+PyMaF [59]	79.9	83.6	80.3	85.5	93.4	106.1	115.6	123.3	119.5	137.9	125.2	142.5
AfB [63]+PyMaF [59]	83.5	87.0	86.7	91.42	93.7	109.8	113.1	124.6	117.1	141.8	127.4	147.4

Table 4. **Results on Blurred-Clothed CAPE dataset.** Our method outpeforms competitive baselines on larger blur rates.

blur rate	<0.2		[0.2,0.3]		[0.3,0.4]		[0.4,0.5]		[0.5,0.7]		[0.7,0.9]	
Method	PA-MPJPE ↓	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓	MPJPE↓
Hfb (ours)	83.7	85.3	**86.1**	**91.3**	**90.5**	**96.4**	**92.6**	**99.3**	**104.2**	**114.4**	**110.5**	**117.4**
BIN [47] + HybrIK [24]	**76.6**	**78.8**	86.7	96.5	93.5	106.9	107.0	122.9	116.3	134.5	120.2	150.0
AfB [63] + HybrIK [24]	82.4	84.6	96.8	105.3	100.8	111.1	103.6	120.1	119.1	137.9	120.9	156.2
BIN [47] + METRO [27]	84.0	85.9	88.5	97.5	94.0	106.8	100.1	118.4	112.8	133.7	117.4	146.4
AfB [63] + METRO [27]	87.7	90.2	90.8	98.1	97.7	108.9	109.0	126.2	113.0	133.3	121.8	151.2
PyMaF [59]	97.7	119.5	118.7	152.5	128.8	172.0	138.3	202.0	152.7	231.4	157.9	252.9
BIN [47]+PyMaF [59]	78.29	98.04	95.05	126.2	105.2	151.2	110.7	147.6	121.7	168.4	124.0	187.7

Table 5. **Results on Blurred-Clothed CAPE dataset with 4 consecutive frames.** We also combine both METRO [27] and HybrIK [24] with two baselines (BIN [47] and AfB [63]) to show the impact of different human pose estimation methods. We also show results of BIN [47] with PyMaF [59], which include the interpolation of the joint positions by directly applying on blur frames. With multiple input frames, HfB outpeforms other baselines for almost all blur rates (0.2 and higher).

whereas our method is purely optimization based.

4.5. Captured data

We captured 21 real-world sequences with significant amounts of motion blur, including four male and one female subjects. The used cameras are the IDS camera and a GoPro 7, which were deliberately set at a low frame rate of 30 fps with exposure time of 30 ms to 50 ms. The recorded humans were asked to move fast. Background images are captured as well. As shown in Fig. 3 and Fig. 1, the final reconstructions are plausible. When compared to SfB [45], our method achieves significantly better results.

5. Conclusion

We proposed the first method to reconstruct subframe human motion and textured shape from substantially blurred images. The key idea is to approach the problem from a generative viewpoint and describe a fully differentiable forward process to generate blurry images from a given 3D human motion model. The core of our method is an image reconstruction loss that allows to solve the inverse problem with standard gradient descent methods. Experiments showed that the proposed method achieves the best results on both synthetic and real blurry data.

Acknowledgements. This research was supported by a Google Focused Research Award, Innosuisse grant No. 34475.1 IP-ICT, and a research grant by FIFA.

References

[1] Advanced Computing Center for the Arts and Design. ACCAD MoCap Dataset. 4, 6

[2] Saad D Al-Sheekh and Majid Dherar Younus. Real-time pose estimation for human-robot interaction. In *2020 2nd Annual International Conference on Information and Sciences (AiCIS)*, pages 86–90. IEEE, 2020. 1

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM transactions on graphics (TOG)*, volume 24, pages 408–416. ACM, 2005. 2

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision*, pages 561–578. Springer, 2016. 2

[5] Carnegie Mellon University. CMU MoCap Dataset. 4, 6

[6] Wenzheng Chen, Jun Gao, Huan Ling, Edward Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3d objects with an interpolation-based differentiable renderer. In *NeurIPS*, 2019. 3

[7] Zhixiang Chi, Yang Wang, Yuanhao Yu, and Jin Tang. Test-time fast adaptation for dynamic scene deblurring via meta-auxiliary learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9137–9146, June 2021. 2

[8] Grigorios G Chrysos, Paolo Favaro, and Stefanos Zafeiriou. Motion deblurring of faces. *International journal of computer vision*, 127(6):801–823, 2019. 2

[9] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. Estimating human shape and pose from a single image. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1381–1388. IEEE, 2009. 2

[10] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10884–10894, 2019. 2

[11] Nils Hasler, Hanno Ackermann, Bodo Rosenhahn, Thorsten Thormählen, and Hans-Peter Seidel. Multilinear pose and body shape estimation of dressed subjects from image sets. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1823–1830. IEEE, 2010. 2

[12] Christoph Heindl, Markus Ikeda, Gernot Stübl, Andreas Pichler, and Josef Scharinger. Metric pose estimation for human-machine interaction using monocular vision. In *IROS Factory of the Future Workshop*, 2019. arXiv preprint arXiv:1910.03239. 1

[13] M Jin et al. Learning to extract a video sequence from a single motion-blurred image. In *CVPR 2018*, 2018. 5, 6, 7

[14] Meiguang Jin, Zhe Hu, and Paolo Favaro. Learning to extract flawless slow motion from blurry videos. In *CVPR*, June 2019. 2, 6, 8

[15] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 2

[16] Adam Kaufman and Raanan Fattal. Deblurring using analysis-synthesis networks pair. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2015. 5

[18] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 4

[19] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 2

[20] Viktor Kress, Janis Jung, Stefan Zernetsch, Konrad Doll, and Bernhard Sick. Human pose estimation in real traffic scenes. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 518–523. IEEE, 2018. 1

[21] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[22] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *ICCV*, Oct 2019. 2

[23] Dongxu Li, Chenchen Xu, Kaihao Zhang, Xin Yu, Yiran Zhong, Wenqi Ren, Hanna Suominen, and Hongdong Li. Arvo: Learning all-range volumetric correspondence for video deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7721–7731, June 2021. 2

[24] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3383–3393, 2021. 2, 7, 8

[25] Jizhizi Li, Jing Zhang, Stephen J. Maybank, and Dacheng Tao. Bridging composite and real: Towards end-to-end deep

image matting. *Int. J. Comput. Vision*, 130(2):246–266, feb 2022. 6

[26] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Learn to dance with aist++: Music conditioned 3d dance generation, 2021. 7

[27] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021. 1, 2, 4, 5, 6, 7, 8

[28] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. *arXiv*, pages arXiv–2012, 2020. 1, 2, 3, 4

[29] Guilin Liu, Fitsum Reda, Kevin Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. *Image Inpainting for Irregular Holes Using Partial Convolutions*, pages 89–105. Springer International Publishing, 09 2018. 4

[30] Jonathan Samuel Lumentut and In Kyu Park. Human and scene motion deblurring using pseudo-blur synthesizer. *IEEE Access*, 9:146366–146377, 2021. 2

[31] Jonathan Samuel Lumentut, Joshua Santoso, and In Kyu Park. Human motion deblurring using localized body prior. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2

[32] Q Ma et al. Learning to Dress 3D People in Generative Clothing. In *CVPR*, 2020. 6

[33] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5441–5450, Oct. 2019. 4, 6

[34] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2640–2649, 2017. 2

[35] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. Xnect: Real-time multi-person 3d motion capture with a single rgb camera. *Acm Transactions On Graphics (TOG)*, 39(4):82–1, 2020. 1, 2

[36] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4):44, 2017. 1, 2

[37] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 International Conference on 3D Vision (3DV)*, pages 484–494. IEEE, 2018. 2

[38] Ahmed AA Osman, Timo Bolkart, and Michael J Black. Star: Sparse trained articulated human body regressor. In *European Conference on Computer Vision*, pages 598–613. Springer, 2020. 2

[39] Jinshan Pan, Haoran Bai, and Jinhui Tang. Cascaded deep video deblurring using temporal sharpness prior. In *CVPR*, June 2020. 2

[40] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei Efros. Context encoders: Feature learning by inpainting. In *Computer Vision and Pattern Recognition (CVPR)*, 2016. 4

[41] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3

[42] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2, 3, 4, 6

[43] D. Rozumnyi, J. Kotera, F. Šroubek, and J. Matas. Sub-frame appearance and 6d pose estimation of fast moving objects. In *CVPR*, pages 6777–6785, 2020. 2

[44] Denys Rozumnyi, Martin R. Oswald, Vittorio Ferrari, Jiri Matas, and Marc Pollefeys. Defmo: Deblurring and shape recovery of fast moving objects. In *CVPR*, Nashville, Tennessee, USA, Jun 2021. 2

[45] Denys Rozumnyi, Martin R. Oswald, Vittorio Ferrari, and Marc Pollefeys. Shape from blur: Recovering textured 3d shape and motion of fast moving objects. In *NeurIPS*, 2021. 1, 2, 4, 6, 7, 8

[46] Denys Rozumnyi, Martin R. Oswald, Vittorio Ferrari, and Marc Pollefeys. Motion-from-blur: 3d shape and motion estimation of motion-blurred objects in videos. In *CVPR*, Jun 2022. 1, 2, 3, 6, 7

[47] Wang Shen, Wenbo Bao, Guangtao Zhai, Li Chen, Xiongkuo Min, and Zhiyong Gao. Blurry video frame interpolation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2, 6, 7, 8

[48] Leonid Sigal, Alexandru Balan, and Michael J Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *Advances in neural information processing systems*, pages 1337–1344, 2008. 2

[49] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. In *European Conference on Computer Vision*, pages 744–760. Springer, 2020. 2, 5, 7

[50] Maitreya Suin, Kuldeep Purohit, and A. N. Rajagopalan. Spatially-attentive patch-hierarchical network for adaptive motion deblurring. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[51] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 529–545, 2018. 2

[52] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, 2021. 2

[53] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *Advances in Neural Information Processing Systems*, pages 5236–5246, 2017. 2

[54] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 109–117, 2017. 2

[55] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 6

[56] Ignas Budvytis Vince Tan and Roberto Cipolla. Indirect deep structured learning for 3d human body shape and pose prediction. In Gabriel Brostow Tae-Kyun Kim, Stefanos Zafeiriou and Krystian Mikolajczyk, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 15.1–15.11. BMVA Press, September 2017. 2

[57] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4116–4125, 2020. 6

[58] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. Denserac: Joint 3d pose and shape estimation by dense render-and-compare. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7760–7770, 2019. 2

[59] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 8

[60] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *CVPR*, June 2020. 2

[61] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7739–7749, 2019. 2

[62] Zhihang Zhong, Ye Gao, Yinqiang Zheng, and Bo Zheng. Efficient spatio-temporal recurrent neural network for video deblurring. In *ECCV*, pages 191–207. Springer, 2020. 2

[63] Zhihang Zhong, Xiao Sun, Zhirong Wu, Yinqiang Zheng, Stephen Lin, and Imari Sato. Animation from blur: Multimodal blur decomposition with motion guidance. *arXiv preprint arXiv:2207.10123*, 2022. 2, 5, 6, 7, 8

[64] Shangchen Zhou, Jiawei Zhang, Jinshan Pan, Haozhe Xie, Wangmeng Zuo, and Jimmy Ren. Spatio-temporal filter adaptive network for video deblurring. In *ICCV*, 2019. 2

[65] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4966–4975, 2016. 2