Human from Blur: Human Pose Tracking from Blurry Images — Supplementary Material —

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

This supplementary material provides further details on our method as well as additional experimental results including an ablation study and more baseline comparisons.

A. More results

We show two extra results on B-AIST++ [19] in Figs. E.7 and E.8 and one on the synthetic BC-CAPE dataset in Fig. E.5. In Fig. E.9, we compare our method with BIN [14] and AfB [19] on a real example with 4 averaged frames. An example of the BC-CAPE dataset construction is shown in Fig. E.1. We also included a supplementary video that shows all examples from the main paper.

B. Ablation study

In Table D.1, we provide an ablation study on part of the BT-AMASS dataset. To evaluate the performance of different loss terms, we test our model by excluding each of loss terms individually while keeping all other experimental conditions the same. Specifically, we test our method each time without Background regularization $\mathcal{L}_B$, Polynomial regularization $\mathcal{L}_C$, SMPL shape regularization $\mathcal{L}_\beta$, and Surface texture smoothness $\mathcal{L}_S$. Then, we evaluate their performance using the IoU and MPJPE losses.

C. Baselines

We clarify that we use Animation-from-Blur (AfB) [19] that was pre-trained on B-AIST++ [19]. Similarly, we use the pre-trained version from Jin *et al*. [4] and the Blurry Video Frame Interpolation (BIN) [14]. We crop input images with a tight bounding box for AfB [19] as suggested by the authors.

Version	MPJPE↓	IoU↑
HfB $\mathcal{L}$ (11)	**79.1**	**0.79**
HfB w/o $\mathcal{L}_B$ (9)	81.6	0.79
HfB w/o $\mathcal{L}_C$ (8)	89.1	0.76
HfB w/o $\mathcal{L}_\beta$ (7)	81.4	0.75
HfB w/o $\mathcal{L}_P$ (6)	100.1	0.73
HfB w/o $\mathcal{L}_S$ (5)	80.5	0.77
HfB w/o $\mathcal{L}_\alpha$ (4)	112.4	0.53
HfB w/o $\mathcal{L}_I$ (3)	84.5	0.77

Table D.1. **Ablation study** with 311 samples generated from our synthetic dataset BT-AMASS with blur rates $\in (0.2, 0.5)$.

Dataset	# of polynomials	4-order		5-order	
		MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
CMU [2]	87787	5.10	3.68	3.56	2.36
ACCAD [1]	13250	5.77	4.85	4.60	3.18

Table D.2. **Reconstruction error for different polynomial orders.** More parameters and a higher order lead to a better model fit and lower errors.

D. Motion prior details

First, we uniformly sample the short motion sequence of the body joint rotations $\boldsymbol{\theta}^J[n]$ from the AMASS dataset [10] (mainly, CMU [2] and ACCAD [1]) with 10 to 90 frames ($n \in (10, 90)$), which translates to blur rates between 0.01 and 0.75. Then, we use the least squares method to fit polynomials to obtain ground truth polynomial coefficients C. The reconstruction error of the polynomials is shown in Table D.2. We denote the temporally normalized coefficients as C_n, which are converted from the original time interval $[0, t_o]$ with $t_o = n/fps$ to $[0, 1]$.

The sampled coefficient C is concatenated with the temporally normalized coefficient as $\mathbf{C} = [C, C_n]$, which has corresponding indicator function map $I_c = \mathbf{0}^{J\times 3}$. Coefficients $\mathbf{C}$ are randomly augmented with noise and reversed joint motion to get coefficient matrix $\tilde{\mathbf{C}} = [\tilde{C},\tilde{C_n}]$. The

entry with value 1 in indicator function $\tilde{I}_c$ indicates which joints' coefficients are unrealistic. The generator G tries to output corrected coefficients $\hat{\mathbf{C}}$ close to true coefficients $\mathbf{C}$. The discriminator D tries to discriminate $\tilde{\mathbf{C}}$, $\hat{\mathbf{C}}$, and $\mathbf{C}$. Inspired by [11], the training is supervised jointly by two loss terms. The adversarial loss $\mathcal{L}_{adv}$ contains adversarial and binary cross-entropy terms. The generator loss $\mathcal{L}_G$ is a reconstruction loss that includes three terms. The first is L1 loss $\mathcal{L}_C$ between the coefficient matrix predicted by the correction generator and the ground truth. The second loss is L2 loss $\mathcal{L}_{Pose}$ between the reconstructed pose and the ground truth pose. The last one $\mathcal{L}_{JP}$ is the mean per joint position error (MPJPE) [15] between the reconstructed joint positions and the ground truth SMPL joint positions. All loss terms are summarized here:

$$\begin{aligned}
\mathcal{L}_{adv} &= -\mathbb{E}[\log(1-|\tilde{I}_c - D(\tilde{\mathbf{C}})|) + \log(1-|D(\mathbf{C})|)] \\
&\quad -\mathbb{E}[\log D(G(\tilde{\mathbf{C}}, D'(\tilde{\mathbf{C}}))) + \log D(G(\mathbf{C}, D'(\mathbf{C})))] && (2)\\
\mathcal{L}_{rec} &= \mathcal{L}_C + \mathcal{L}_{Pose} + \mathcal{L}_{JP} && (3)\\
\mathcal{L}_C &= |G(\tilde{\mathbf{C}}, D'(\tilde{\mathbf{C}})) - \mathbf{C}| + |G(\mathbf{C}, D'(\mathbf{C})) - \mathbf{C}| && (4)\\
\mathcal{L}_{Pose} &= \|\text{Pose}(\tilde{G}) - \text{Pose}(\mathbf{C})\|_2 \\
&\quad + \|\text{Pose}(G) - \text{Pose}(\mathbf{C})\|_2 && (5)\\
\mathcal{L}_{JP} &= \text{MPJPE}(\text{Joint}(\text{Pose}(\tilde{G})), \text{Joint}(\text{Pose}(\mathbf{C}))) \\
&\quad + \text{MPJPE}(\text{Joint}(\text{Pose}(G)), \text{Joint}(\text{Pose}(\mathbf{C}))) && (6)\\
\mathcal{L}_{total} &= \lambda_{rec}\mathcal{L}_{rec} + \lambda_{adv}\mathcal{L}_{adv} && (7)
\end{aligned}$$

Instead of ground truth indicator $\tilde{I}_c$ and $\mathbf{0}$, we feed the detached predicted indicator function $D'(\tilde{\mathbf{C}})$ and $D'(\mathbf{C})$ to the generator. The generator and discriminator also take the sequences as input, which is decomposed from the coefficients $\tilde{\mathbf{C}}$ and $\mathbf{C}$ with exposure time t. The generator and discriminator have a similar structure that contain convolutions and self-attention layers. The output features of the above layers are passed to a 1×1 convolution, followed by an MLP.

Original Clothed Blurred

Figure E.1. **BC-CAPE dataset construction.** The original CAPE [9] human body dataset contains no texture or clothes. Since their human body representation is also an SMPL [12] model, we use textures from the SURREAL [16] dataset. Then, blurry inputs are generated by temporal integration and interpolation between neighboring recorded meshes.

Texture from Fig. E.5 Texture from Fig. E.7

Figure E.2. **Examples of texture estimation.** The top row shows our deblurred sub-frame estimation, and the bottom row shows the corresponding optimized texture maps in the SMPL format.

E. Limitations

Non-tight clothing. Our model optimizes the basic SMPL body model without clothes, which limits the performance on real data when the human is dressed in loose-fitting clothing. As shown in Fig. E.8, in such case the optimization results in a more obese body shape since silhouette and photometric consistency are the major losses that dominate the optimization. The proposed model may fail on humans with more complex or wider clothing. A potential solution is to introduce clothed SMPL models [3, 9].

Coarse texture. As shown in Fig. 3 in the main paper, our model also optimizes the human texture. Some examples are shown in Fig. E.2. However, the optimized texture is not on par with the ground truth sharp texture, especially when the human wears non-tight clothing. Since the core of our method is a gradient descent with differentiable rendering, the lighting is not considered, which makes the textures even less accurate. Thus, when comparing the deblurring quality with usual deblurring metrics, *e.g.* PSNR and SSIM, our method has poor performance. However, the primary purpose of our method is not deblurring and sharp texture estimation from blurry images. Our main goal is sub-frame accurate human pose estimation from blurry images. Accurate sharp texture estimation is a direction for future work.

Failure of HPE model. Our method employs a Human Pose Estimation (HPE) model for the initial pose. We show the comparison of several different HPE models on real blur examples in Fig. E.4. We observed that METRO [7] and HybrIK [6] are usually robust against the blurry input and produce a single pose, which is usually a random sub-frame pose Examples are shown in Fig. E.3.

Background as input. Our model requires a background image without the blurry human as additional input, which could limit in-the-wild applications. As shown in Fig. E.6, the background matting model BGMv2 [8] can produce ar-

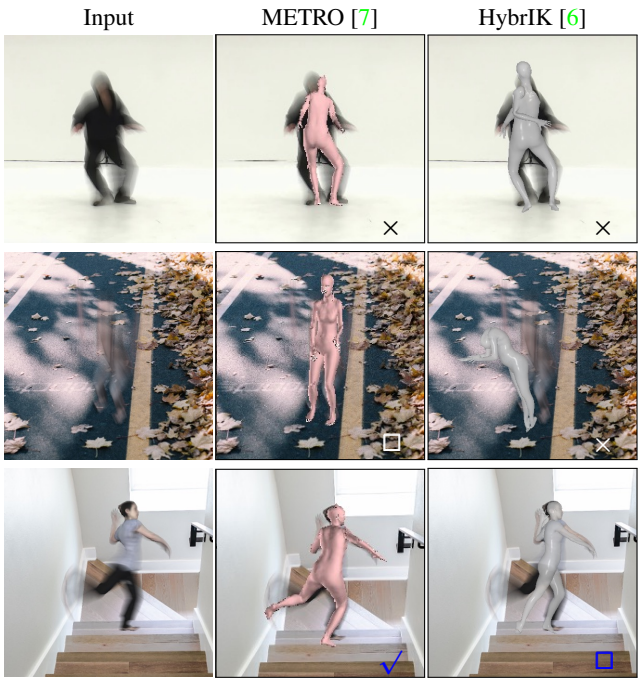

Figure E.3. **Failure of initial human pose estimation.** In some cases, when the initialization is far from the real pose, our method cannot converge to the right solution. We have notations for three cases: (✓) converges and initial pose is one of the sub-frame poses, (□) still converges but initial pose is not one of the sub-frame poses, and (×) does not converge. In the top row, both models predict a flipped pose, which optimization cannot rotate back. The middle row shows a fast squatting pose with stretched arms, for which only the METRO [7] initialization allows for convergence by predicting a standing pose. In the bottom row, only the METRO [7] detects the moving leg.

tifacts for large amounts of blur. Since BGMv2 is designed and trained for a general matting problem, fine-tuning it on highly motion-blurry inputs may improve the performance of our method.

References

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

[2] Carnegie Mellon University. CMU MoCap Dataset. 1

[3] Enric Corona, Albert Pumarola, Guillem Alenyà, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *CVPR*, 2021. 2

[4] M Jin et al. Learning to extract a video sequence from a single motion-blurred image. In *CVPR 2018*, 2018. 1, 4, 5, 6

[5] 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. 4

[6] 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, 3, 4

[7] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021. 2, 3, 4, 5, 6, 7

[8] 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. 2, 4, 5, 6

[9] Q Ma et al. Learning to Dress 3D People in Generative Clothing. In *CVPR*, 2020. 2

[10] 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. 1

[11] 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. 2

[12] 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

[13] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. 4

[14] 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. 1, 7

[15] 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

[16] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 2

[17] Xiangyu Xu, Hao Chen, Francesc Moreno-Noguer, Laszlo A Jeni, and Fernando De la Torre. 3d human shape and pose from a single low-resolution image with self-supervised learning. In *ECCV*, 2020. 4

[18] 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. 4, 6

[19] 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. 1, 4, 5, 6, 7

Figure E.4. **Robustness w.r.t. blur.** We test several models on real data. Notice the misalignment for larger blur rates. PyMaF [18] and VIBE [5] depend on YOLO [13] tracking method to get consistent tracking of bounding boxes between frames. YOLO [13] shows less robustness against blur. We show an example of PyMaF [18] reconstruction in Fig. E.8.

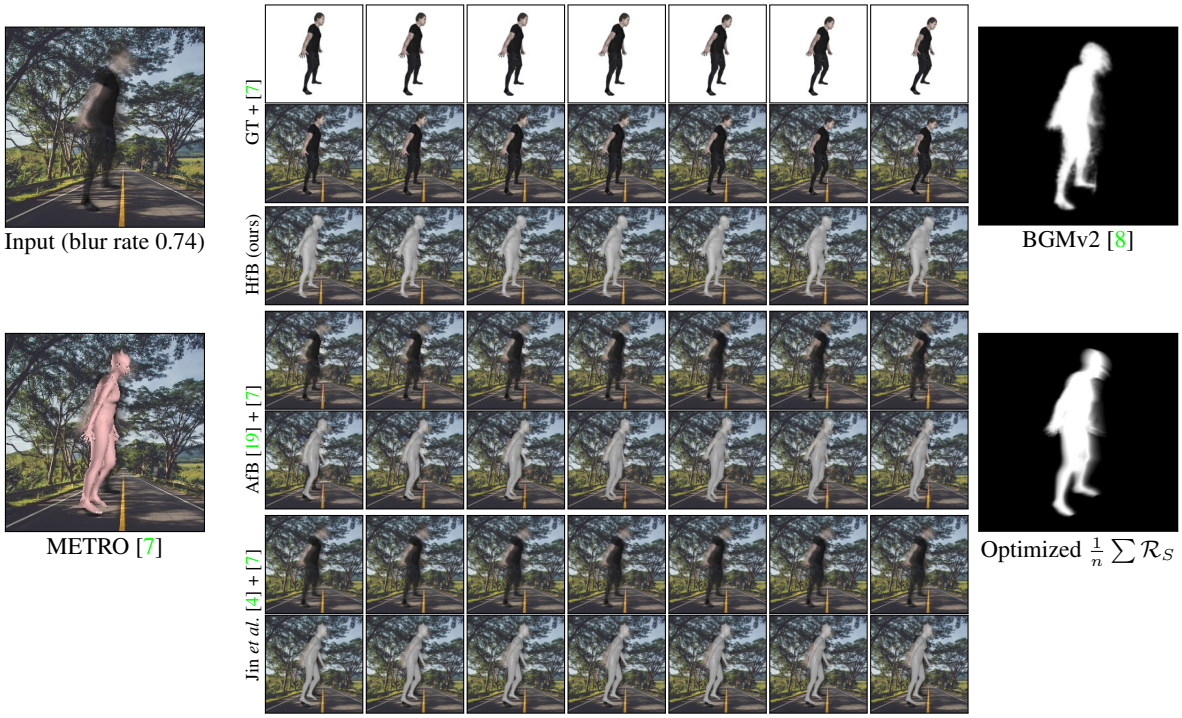

Figure E.5. **Results on the BC-CAPE dataset.**

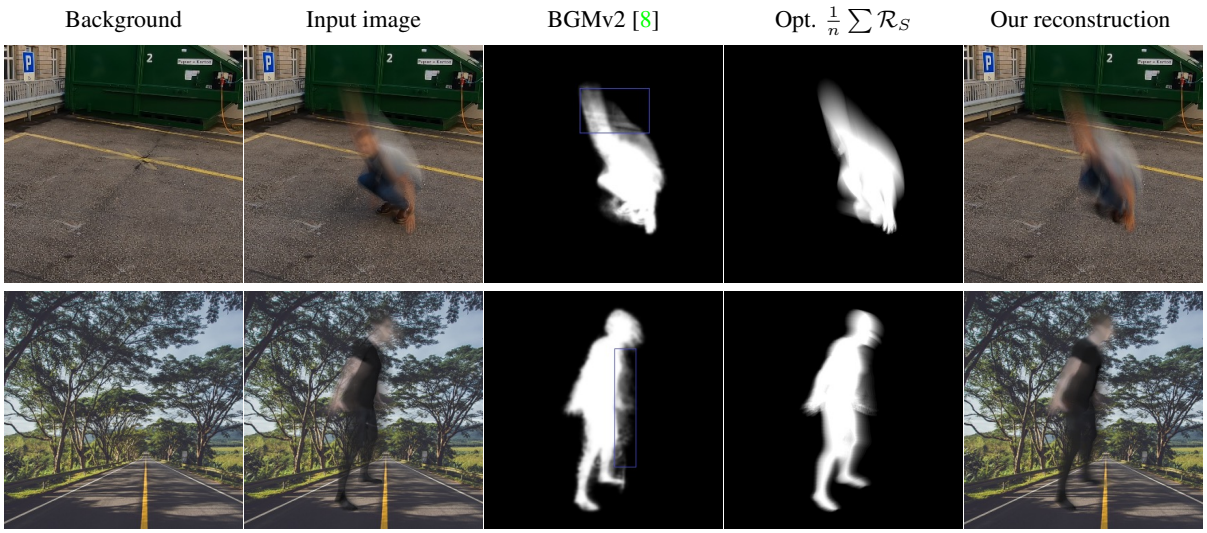

Figure E.6. **Examples of inconsistent background matting.** The top row is the real image from Fig. 1 in the main paper, while the bottom row is a synthetic image from Fig. E.5 in this supplementary file. The estimated background matting by BGMv2 [8] shows slight errors, but HfB still estimates the correct motion.

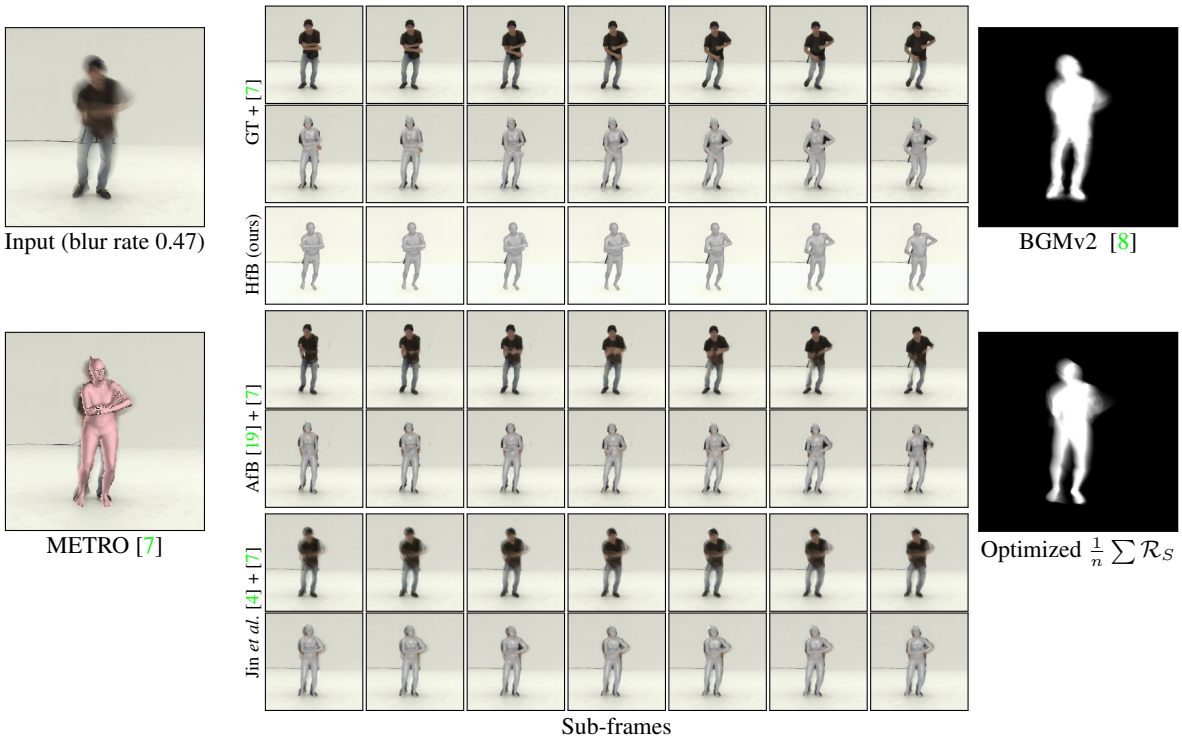

Figure E.7. **Comparison on real data.** We show an input with three averaged frames from B-AIST++ [19]. In the right-most column, we show the initial matting BGMv2 [8] (top) and the optimized result from our method (bottom).

Figure E.8. **Comparison on real data:** three averaged frames from the B-AIST++ [19] dataset. Since the human subject is dressed in loose-fitting clothing, HfB incorrectly estimates an obese body shape because the clothing is not modeled explicitly. We also show an additional result of PyMaF [18], which is applied to the prediction from AfB [19]. Even though PyMaF [18] is video-based, it shows discontinuity on coarse inputs, *e.g*. the third frame.

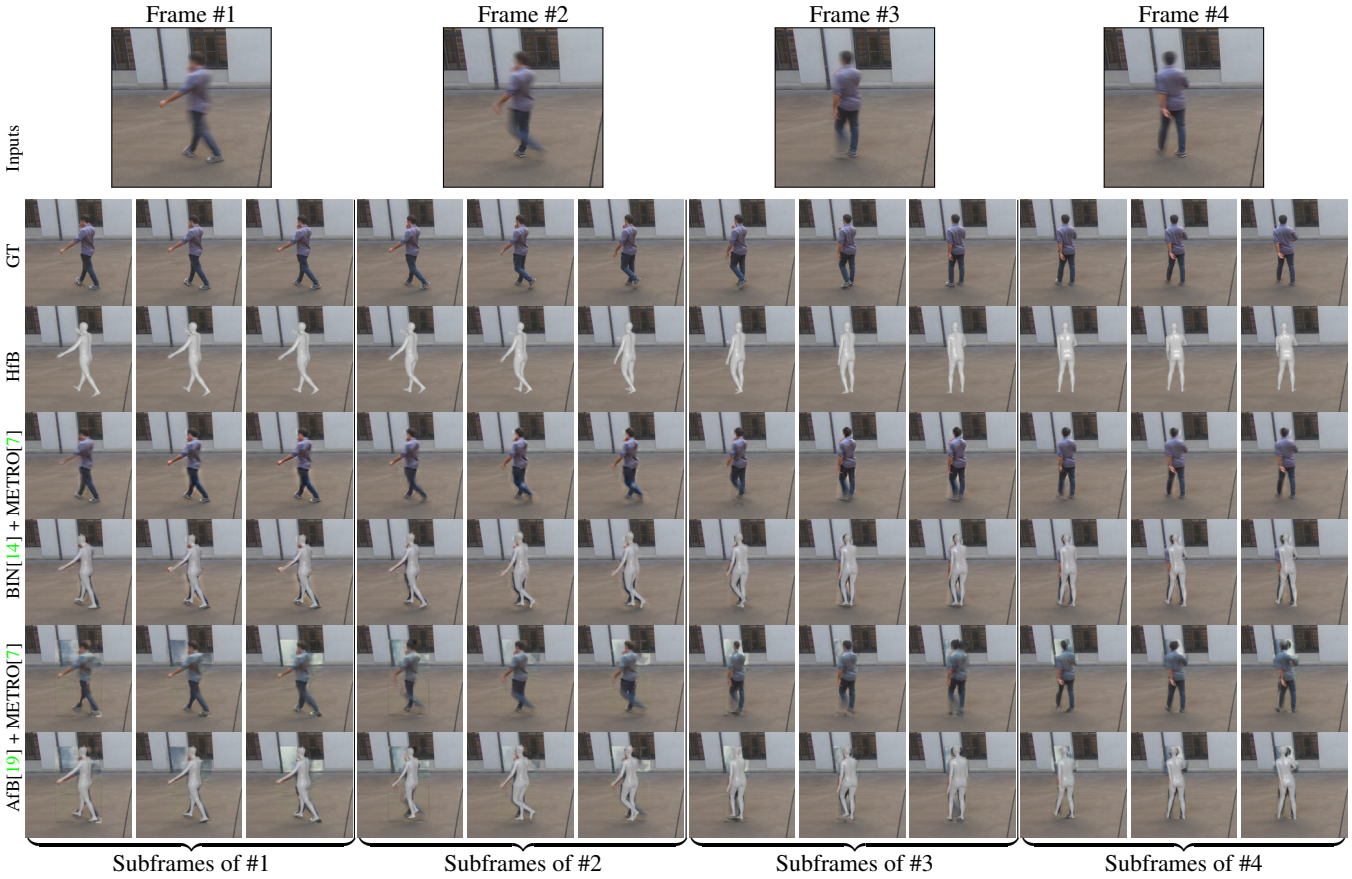

Figure E.9. **Qualitative multi-frame evaluation on a real video.** For small amounts of motion blur, the performance of image-deblurring-based methods is typically good, but for larger amounts of blur (as in Frame #3), their deblurring often fails, and the subsequent human model fitting becomes rather noisy.