Personalized 3D Human Pose and Shape Refinement
—–Supplementary Material—–

Tom Wehrbein[1,2†] Bodo Rosenhahn[1] Iain Matthews[2] Carsten Stoll[2]
[1]Leibniz University Hannover, [2]Epic Games
wehrbein@tnt.uni-hannover.de

1. Implementation Details

Training details. We train our model for 600K steps with a batch size of 8 using Adam [10]. The learning rate is set to 5×10^{-4} and we sample training examples from H36M, 3DPW and SURREAL with probabilities 0.4, 0.3 and 0.3. The images are cropped and resized to 224×224 while maintaining the aspect ratio. Additionally, with a probability of 0.5, we add Gaussian noise to the pose, shape and camera parameters. We select the PARE prediction with probability of 0.3 and take a sample from ProHMR with probability of 0.5. To also focus on fine-grained displacements, we use ground-truth pose with PARE predicted shape and camera parameters with probability of 0.2. Following [12], image data augmentation includes random rotations, scaling and channel-wise pixel noise [12]. Besides, we adopt photometric distortion [3] and for H36M and SURREAL self-mixing [5]. The channel-wise pixel noise is also applied on the texture map.

Preprocessing details. To benchmark our approach, we generate predictions using the latest OpenPose [4] version (v1.7.0) and a state-of-the-art DensePose [8] model[1]. For fair comparison, we feed both models with the images cropped around the target subject using the ground-truth bounding boxes. By transforming the DensePose predictions to points on the SMPL body, they can be used for the reprojection loss [7]. For 3DPW, we use the OpenPose detections included in the dataset. Because RICH only provides SMPL-X bodies, we convert the provided model parameters to SMPL using the official implementation [1].

Runtime. The PyTorch implementation of the displacement field prediction network takes on average 26.4 ms to process one frame on a RTX4090. Running our slightly modified SMPLify [2, 12] implementation for 100 iterations with the reconstructed 2D vertices brings no overhead compared to sparse 2D keypoints and takes around 614 ms and 769 ms with the GMM [2] and VPoser [13] prior respectively. Rendering and transforming the per-pixel 2D displacements to per-vertex displacements is in total done in 1 ms. For faster evaluation, we run SMPLify in batch mode. SMPLify with a batch size of 32 takes around 644 ms and 815 ms with GMM and VPoser pose prior respectively. Note that we did not spend any effort optimizing the runtime of our approach. A highly optimized custom implementation can reduce the fitting time to a few milliseconds [6], which would enable our approach to run in real-time. Additionally, by using the refined estimate of the last frame as initialization for the next frame, the 3D pose regressor would only need to be evaluated once.

2. Additional Results

We provide more qualitative refinement results on images from 3DPW [14] and RICH [9] in Fig. 1 and Fig. 2. We use PARE [11] predictions and SMPLify with VPoser. Our approach generalizes well to different scenes and subjects with varied body shapes, can handle poor lighting and challenging poses, and can even improve fine details such as head rotation.

OpenPose comparison. We show additional visual comparisons with refinements using OpenPose keypoints in Fig. 3. Our approach better refines the reconstruction of the back (row 4, 6), better detects barely visible body parts (row 2, 5, 7) and leads to more accurate depth estimates (row 3). Additionally, body parts that are visible in the initial SMPL prediction but not in the image can be correctly pushed to be occluded in the refinements (row 1) using our approach.

DensePose comparison. We visually compare our refined 3D human models with refinements using DensePose predictions in Fig. 4. Each person pixel detected by DensePose is colorized in the image and shown on the ground-truth human body. While DensePose is good at detecting pixels belonging to a person, the predicted correspondences between

[†] Work primarily done during an internship at Epic Games.

[1] https://github.com/facebookresearch/detectron2/blob/main/projects/DensePose/configs/densepose_rcnn_R_101_FPN_DL_s1x.yaml

the pixels and the 3D SMPL surface lack in accuracy. This is especially noticeable at the boundary between body parts, where no pixels are assigned to even though the regions are visible in the image. Our approach computes more accurate dense correspondences, leading to significantly better refined 3D bodies.

Failure cases. In Fig. 5, we show a few examples where our network fails to estimate reasonable 2D displacement vectors. The scenarios range from (a) extreme occlusion, (b) very poor initial body estimates and (c) close interactions and overlap with other subjects. To improve the performance for large occlusions, it could be helpful to learn visibility masks [15,16] or per-pixel confidence scores. The problem of wrongly associating limbs could be mitigated by integrating more examples of closely interacting persons in the training set. Finally, in some cases our refinement leads to improved image-model alignment but degrades the 3D pose (see Fig. 6). This is due to the depth ambiguity inherent in monocular 3D motion capture and could be alleviated by regarding multiple images or integrating scene constraints.

Garments with complex texture. When evaluating on the 3DPW test subject with the most complex texture pattern using VPoser and PARE as base model, we achieve an MPJPE of 68.1 and a PVE of 81.2, compared to 75.9 and 90.2 when using OP joints and 75.6 and 88.8 with the base model. Note that most real world cases allow for a cooperative setting where the person is first turning around in front of a camera, which would allow accurate texture estimation even for complex patterns.

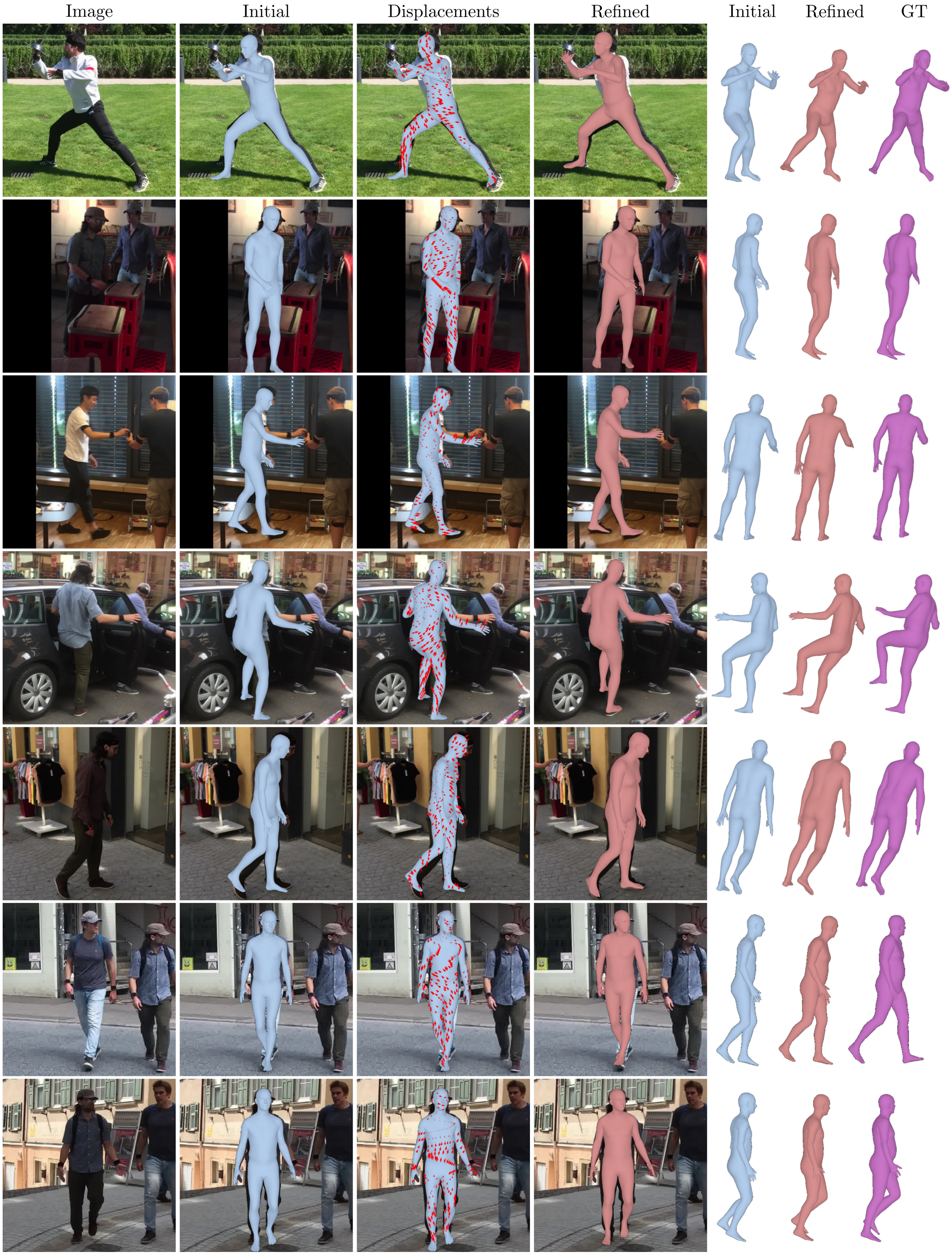

Figure 1. Additional results from the 3DPW [14] dataset. From left to right: input images, initial body estimates, our predicted displacement fields, our refined 3D human models and side views of initial, refined and ground-truth bodies.

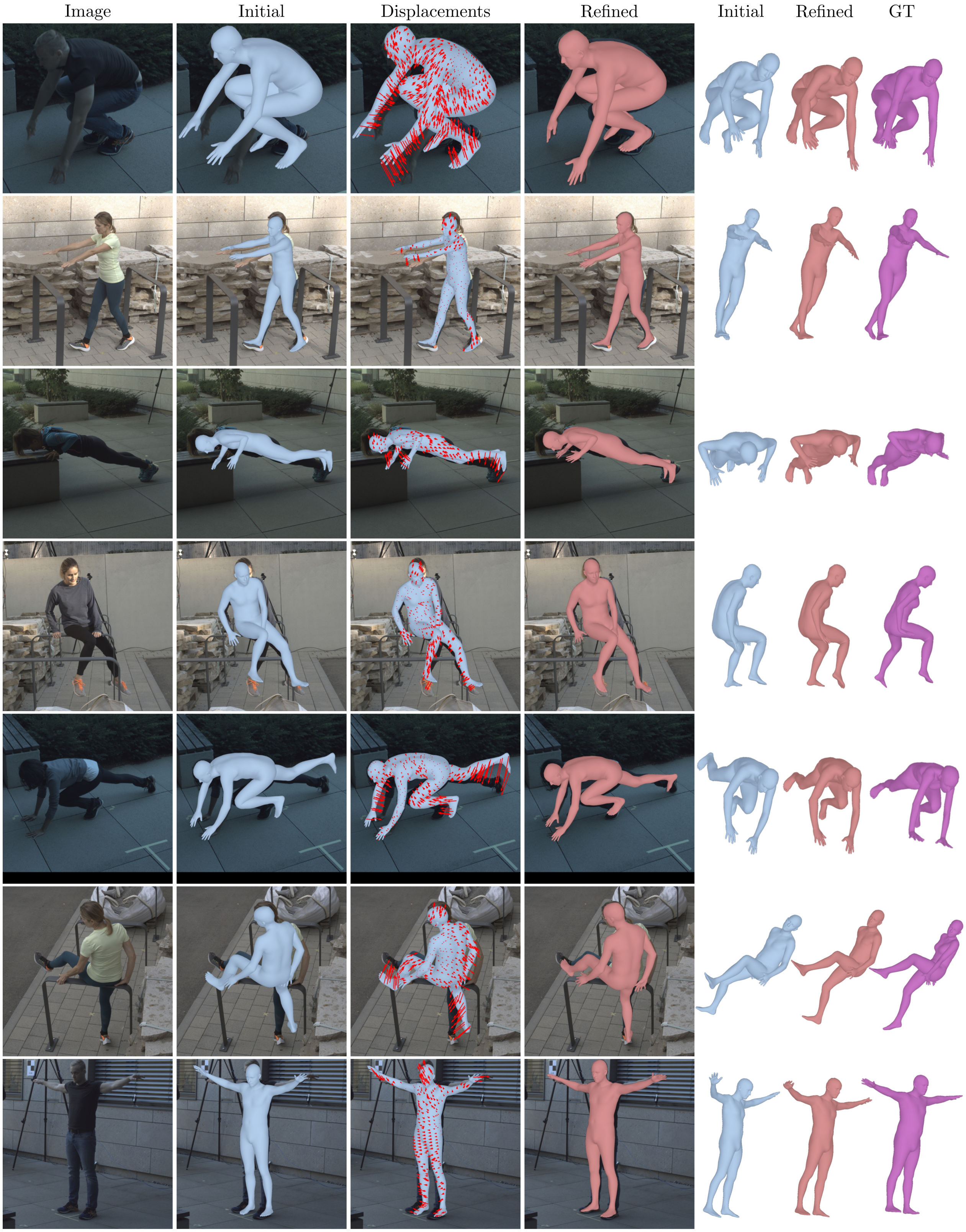

Figure 2. Additional results from the RICH [9] dataset. From left to right: input images, initial body estimates, our predicted displacement fields, our refined 3D human models and side views of initial, refined and ground-truth bodies.

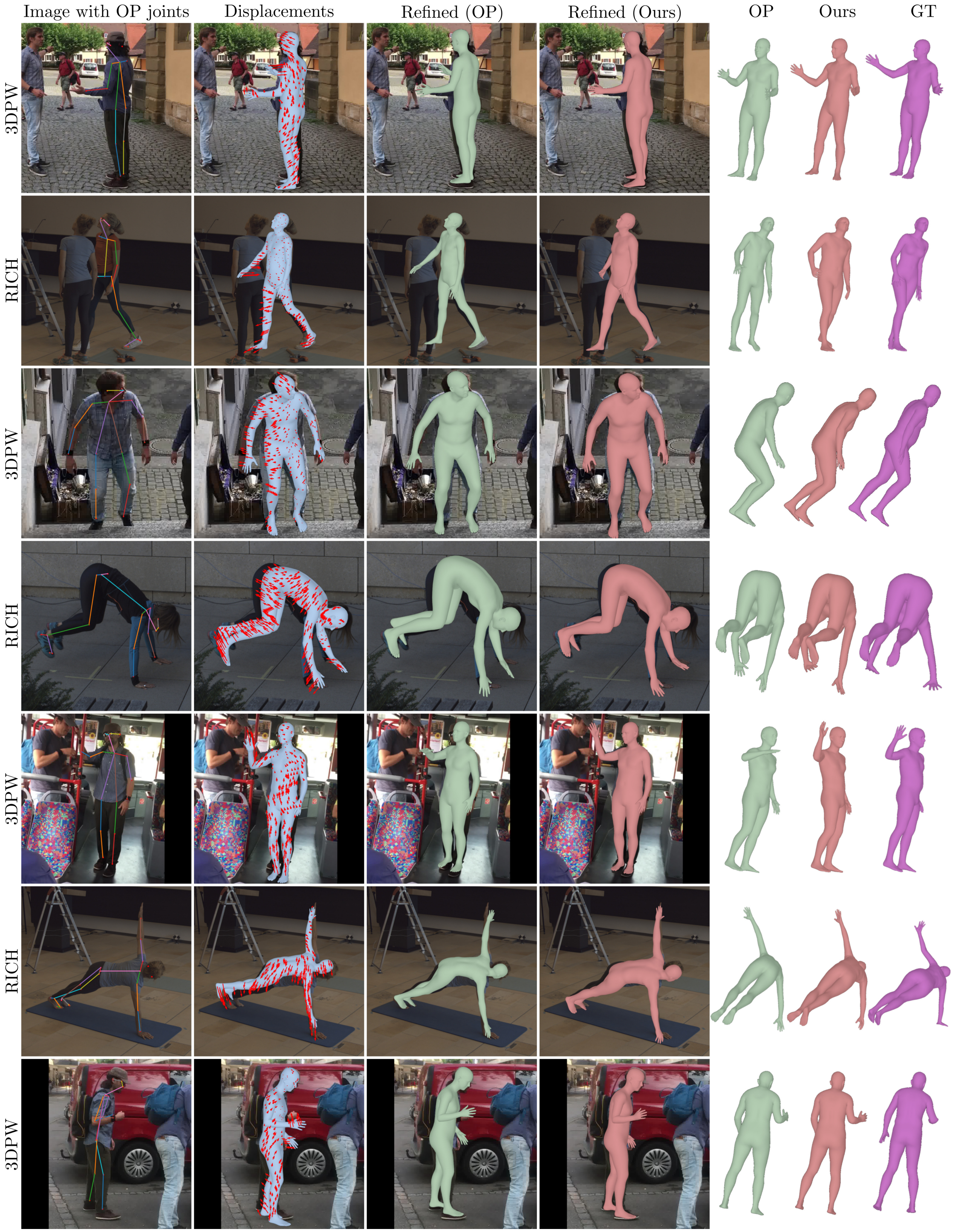

Figure 3. Comparison with refinements using OpenPose [4] keypoints on images from 3DPW [14] and RICH [9].

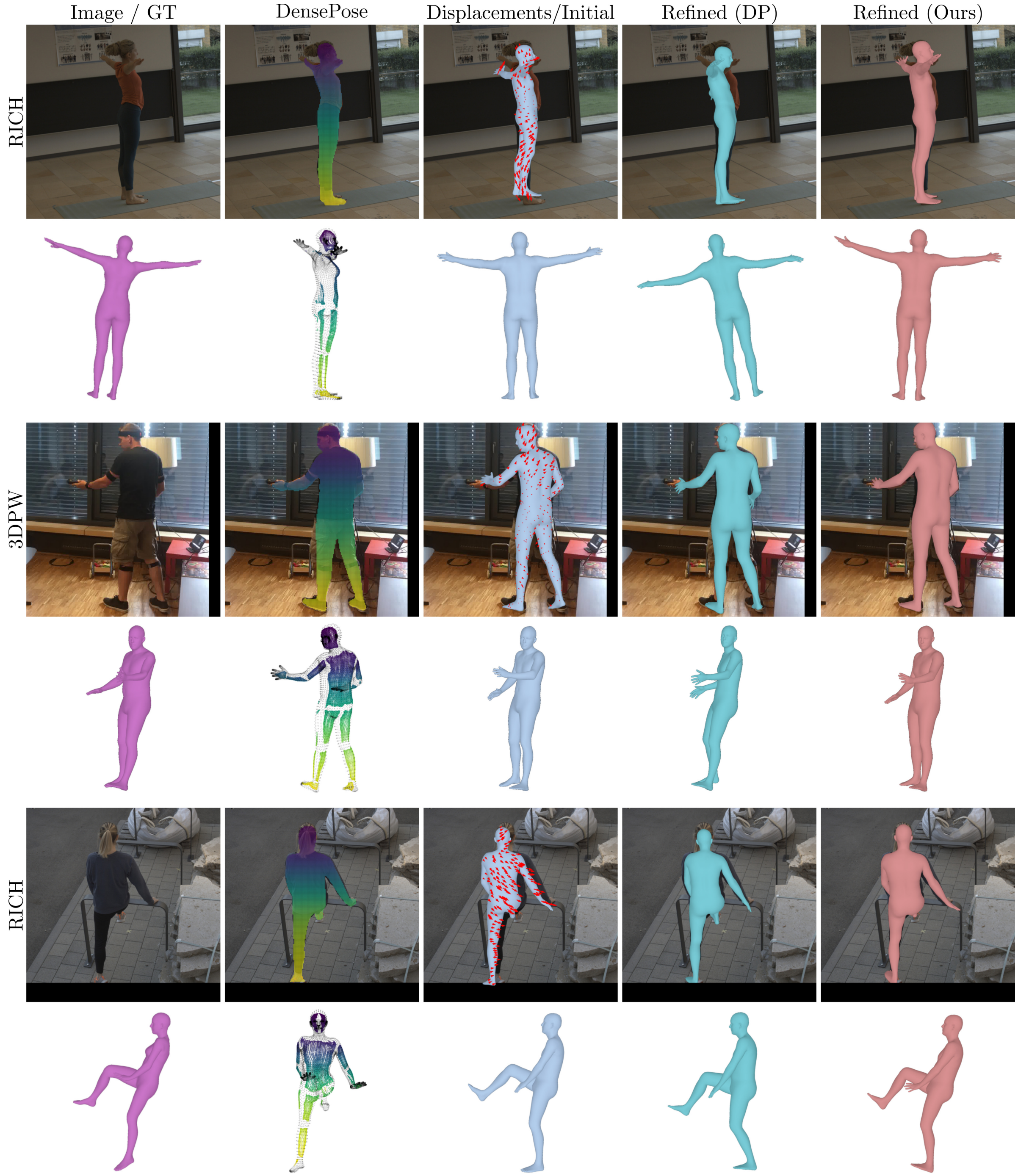

Figure 4. Comparing refinements using DensePose [8] on 3DPW [14] and RICH [9]. Each person pixel detected by DensePose is colorized in the image and shown on the ground-truth human body.

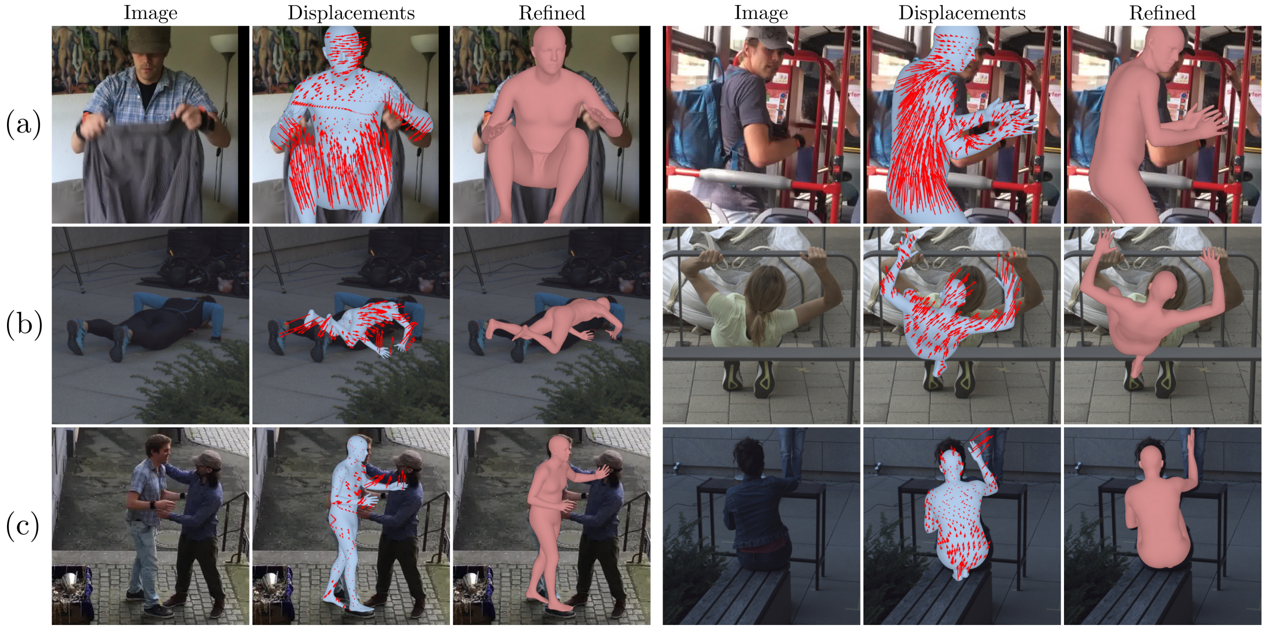

Figure 5. Failure cases of our approach with examples from 3DPW [14] and RICH [9]. (a) Large occlusions may lead to wrong displacement estimates. (b) If the initial estimate is too far away, displacements may not be enough to fit the model. (c) In some cases of close interactions and overlap with other actors the model may wrongly associate limbs.

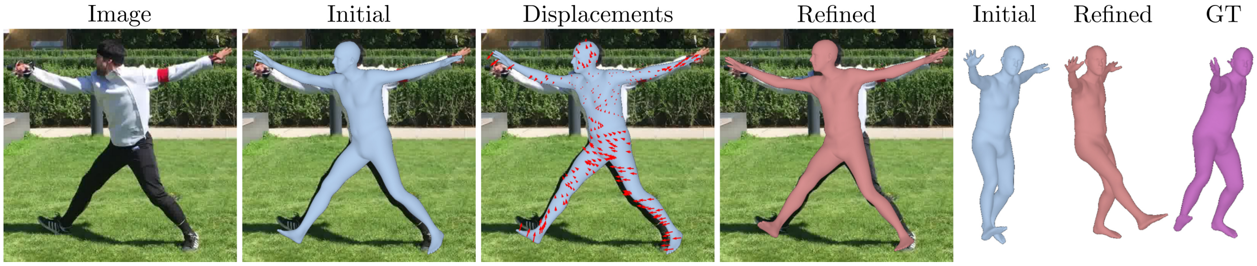

Figure 6. In some cases the 2D alignment may be improved by our approach while leading to a worse 3D pose. Example from 3DPW [14].

References

[1] https://github.com/vchoutas/smplx/tree/master/transfer_model. 1

[2] 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 *ECCV*, 2016. 1

[3] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 11(2), 2020. 1

[4] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1, 5

[5] Minghui Chen, Zhiqiang Wang, and Feng Zheng. Benchmarks for corruption invariant person re-identification. In *NeurIPS*, 2021. 1

[6] Taosha Fan, Kalyan Vasudev Alwala, Donglai Xiang, Weipeng Xu, Todd Murphey, and Mustafa Mukadam. Revitalizing optimization for 3d human pose and shape estimation: A sparse constrained formulation. *ICCV*, 2021. 1

[7] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *CVPR*, 2019. 1

[8] Riza Alp Guler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 1, 6

[9] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *CVPR*, 2022. 1, 4, 5, 6, 7

[10] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 1

[11] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *ICCV*, 2021. 1

[12] 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. 1

[13] 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 *CVPR*, 2019. 1

[14] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *ECCV*, 2018. 1, 3, 5, 6, 7

[15] Chun-Han Yao, Jimei Yang, Duygu Ceylan, Yi Zhou, Yang Zhou, and Ming-Hsuan Yang. Learning visibility for robust dense human body estimation. In *ECCV*, 2022. 2

[16] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *CVPR*, 2020. 2