Supplementary Material for 3D Human Mesh Estimation from Virtual Markers

Xiaoxuan Ma[1] Jiajun Su[1] Chunyu Wang[3 *] Wentao Zhu[1] Yizhou Wang[1, 2, 4]

[1] School of Computer Science, Center on Frontiers of Computing Studies, Peking University
[2] Inst. for Artificial Intelligence, Peking University
[3] Microsoft Research Asia
[4] Nat'l Eng. Research Center of Visual Technology

{maxiaoxuan, sujiajun, wtzhu, yizhou.wang}@pku.edu.cn, chnuwa@microsoft.com

We elaborate on the post-processing implementation of the virtual markers and provide additional experimental details and results. At last, we discuss data from human subjects and the potential societal impact.

1. Post-processing on Virtual Markers

As described in Section 3.1, considering the left-right symmetric human body structure, we slightly adjust the learned virtual markers $\mathbf{Z}$ to be symmetric. In fact, after the first step that updates each $\mathbf{z}_i$ by its nearest vertex to get $\widetilde{\mathbf{Z}} \in \mathbb{R}^{3\times K}$. $\widetilde{\mathbf{Z}}$ are almost symmetric with few exceptions. To get the final symmetric virtual markers $\widetilde{\mathbf{Z}}^{sym} \in \mathbb{R}^{3\times K}$, for each virtual marker located in the left body part, we take its symmetric vertex in the right body to be its symmetric counterpart.

Since the human mesh (*i.e.* SMPL [17]) itself is not strictly symmetric, we clarify the *symmetric vertex pair* (*e.g.* left elbow and right elbow) on a human mesh template $\mathbf{X}^t \in \mathbb{R}^{3\times M}$ in Figure 1. We place $\mathbf{X}^t$ at the origin of the 3D coordinate system. Formally, we define the cost of matching i^{th} vertex to j^{th} vertex to be $\boldsymbol{C}_{i,j} = |x_i + x_j| + |y_i - y_j| + |z_i - z_j|$. A symmetric vertex pair $(\mathbf{X}_i^t, \mathbf{X}_j^t)$ is defined to have the minimal cost $\boldsymbol{C}_{i,j}$. In this way, for each virtual marker in the left body, we take its symmetric vertex counterpart to be its symmetric virtual marker and finally get $\widetilde{\mathbf{Z}}^{sym}$.

2. Experiments

In this section, we first add detailed descriptions for datasets and then provide more experimental results of our approach.

*Corresponding author

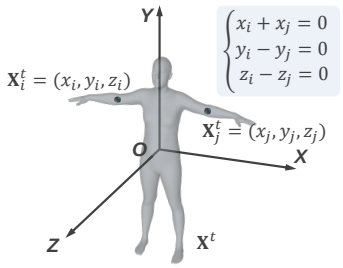

Figure 1. Illustration of the human mesh template $\mathbf{X}^t$ at the 3D coordinate system and a symmetric vertex pair $(\mathbf{X}_i^t, \mathbf{X}_j^t)$.

2.1. Datasets

H3.6M [4]. Following previous works [6, 10, 11, 22], we use the SMPL parameters generated from MoSh [16], which are fitted to the 3D physical marker locations, to get the GT 3D mesh supervision. Following standard practice [6], we evaluate the quality of 3D pose of 14 joints derived from the estimated mesh, *i.e.* $\hat{\mathbf{M}}\mathcal{J}$. We report Mean Per Joint Position Error (MPJPE) and PA-MPJPE in millimeters (mm). The latter uses Procrustes algorithm [3] to align the estimates to GT poses before computing MPJPE. To evaluate mesh estimation results, we also report Mean Per Vertex Error (MPVE) which can be interpreted as MPJPE computed over the whole mesh.

3DPW [25]. The 3D GT SMPL parameters are obtained by using the data from IMUs when collected. Following the previous works [9, 13, 14, 26], we use the train set of 3DPW to learn the model and evaluate on the test set.

MPI-INF-3DHP [19] is a 3D pose dataset with 3D GT pose annotations. Since this dataset does not provide 3D

mesh annotations, following [6, 10], we only enforce supervision on the 3D skeletons (Eq. (9)) in mesh losses.

UP-3D [12] is a wild 2D pose dataset with natural images. The 3D poses and meshes are obtained by SMPLify [1]. Due to the lack of GT 3D poses, the fitted meshes are not accurate. Therefore we only use the 2D annotations to train the 3D virtual marker estimation network as in [23].

COCO [15] is a large wild 2D pose dataset with natural images. Previous work [20] used SMPLify-X [21] to obtain pseudo SMPL mesh annotations but they are not accurate. However, we find that if we project the 3D mesh to 2D image, the resulting 2D mesh vertices align well with the image. So we leverage the 2D annotations to train the virtual marker estimation network as in [23].

SURREAL [24] is a large-scale synthetic dataset containing 6 million frames of synthetic humans. The images are photo-realistic renderings of people under large variations in shape, texture, viewpoint, and body pose. To ensure realism, the synthetic bodies are created using the SMPL body model, whose parameters are fit by the MoSh [16] given raw 3D physical marker data. All the images have a resolution of 320×240. We use the same training split to train the model and evaluate the test split following [2].

2.2. Implementation Details and Computation Resource

Following common practice [2, 6, 8, 11, 13, 14, 20, 26], we conduct mix-training by using the above 2D and 3D datasets for experiments on the H3.6M and 3DPW datasets. To leverage the 3D pose estimation dataset, *i.e.* MPI-INF-3DHP [19], we extend the 64 virtual markers with the 17 landmark joints (*i.e.* skeleton) from the MPI-INF-3DHP dataset. For experiments on the SURREAL dataset, we use its training set alone as in [2, 18]. We implement the proposed method with PyTorch. All the experiments are conducted on a Linux machine with 4 NVIDIA 16GB V100 GPUs. The whole network is trained for 40 epochs with batch size 32 using Adam [7] optimizer.

We evaluate the model complexity in terms of FLOPs (G) and the number of model parameters in Table 1. Compared to the most recent state-of-the-art methods that directly regress *all mesh vertices*, such as I2L-MeshNet [20], METRO [13], and Mesh Graphormer [14], our approach with virtual marker representation reduces the computation overhead by a large margin while getting better estimation quality. The last column shows the MPVE errors on 3DPW test set for performance reference.

Methods	FLOPs (G) ↓	Params (M)	MPVE↓
I2L-MeshNet [20] ECCV'20	28.7	141.2	110.1
METRO [13] CVPR'21	153.0	397.5	88.2
Mesh Graphormer [14] ICCV'21	48.8	180.6	87.7
Ours	**22.1**	**109.6**	**77.9**

Table 1. Computation overhead comparison with the recent state-of-the-art methods that directly regress *all 3D vertices*. The rightmost column shows the MPVE errors on the 3DPW test set for performance reference.

	Ours	*w/o* $\mathcal{L}_{conf}$	*w/o* $\mathcal{L}_{pose}$	*w/o* $\mathcal{L}_{normal}$	*w/o* $\mathcal{L}_{edge}$
MPVE↓	**58.0**	59.2	58.3	60.6	60.4

Table 2. MPVE errors on H3.6M [4] test set when ablating different loss terms.

Occ. VM Parts	MPVE↓	MPJPE↓	PA-MPJPE↓
None (Ours)	**77.9**	**67.5**	**41.3**
2 Arms	79.2 ↑1.3	68.2 ↑0.7	42.2 ↑0.9
2 Legs	78.3 ↑0.4	67.9 ↑0.4	41.7 ↑0.4
Body	78.6 ↑0.7	68.0 ↑0.5	41.8 ↑0.5
Random	78.7 ↑0.8	68.1 ↑0.6	41.9 ↑0.6

Table 3. Results on 3DPW [25] test set when different parts of virtual markers (VM) are occluded.

2.3. Additional Quantitative Results

Different loss terms. Table 2 reports the MPVE error on H3.6M [4] test set when ablating different loss terms. The confidence loss [5] is used to encourage the interpretability of the heatmaps to have a maxima response at the GT position. Without the confidence loss, the error increases slightly. If ablating the surface losses, MPVE increases a lot, which demonstrates the smoothing effect of these two terms.

Robustness to occlusion. We report results when different virtual markers are occluded by a synthetic mask in Table 3. The errors are slightly larger than the original image (None), which validates the effectiveness of the *locality* of the virtual marker representation. Occluding arm regions results in a larger error increase. This may be because the arm has larger variations in the dataset.

Comparison to fitting. In order to disentangle the ability of mesh regression from markers using $\hat{\mathbf{A}}$ with the ability to detect the virtual markers accurately from images, we first compute the estimation errors of the virtual markers. The MPJPE over all the virtual markers is 35.5mm, which demonstrates that these virtual markers can be accurately detected from the images. We then fit the mesh model parameters to these virtual markers. Table 4 shows the metrics of the fitted mesh on the SURREAL [24] test set. As we can

Figure 2. Meshes estimated by our approach on Internet images with challenging cases (complex poses or extreme body shapes).

Method	MPVE↓	MPJPE↓	PA-MPJPE↓
Fitting	44.6	34.8	29.5
Ours	44.7	36.9	28.9

Table 4. Results on SURREAL [24] test set when the mesh is obtained by fitting to the estimated virtual markers.

see, the fitted mesh has a similar error as our regression ones which uses the interpolation matrix $\hat{\mathbf{A}}$, which validates the accuracy of the estimated virtual markers.

2.4. Additional Qualitative Results

Figure 4 shows more qualitative comparisons with Pose2Mesh [2] on the SURREAL test set in which has diverse body shapes. The skeleton representation used in Pose2Mesh loses the body shape information so the method [2] can only recover mean shapes. For example, in Figure 4 (d) (e), the estimated meshes of Pose2Mesh tend to have the average body shape and fail to estimate the real body shape, regardless of whether the person is slim or stout. This is caused by the limited skeleton representation bottleneck so that the model learns a mean shape for the whole training dataset implicitly. In contrast, our approach with virtual marker representation generates more accurate mesh estimation results.

Figure 5 shows more qualitative comparisons with Pose2Mesh [2] and METRO [13] on the 3DPW test set. Pose2Mesh and METRO use the skeleton or all 3D vertices as intermediate representations, respectively. The estimated meshes are overlaid on the images according to the camera parameters. Pose2Mesh [2] has difficulty in estimating correct body pose and shapes when truncation occurs (a) or in complex postures (c). The results of METRO [13] have

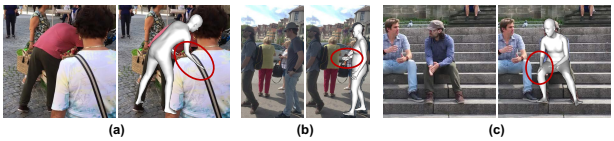

Figure 3. Typical failure cases. **(a)** The right arm has inaccurate shape estimation due to the inaccurate virtual marker estimation around the arm when occluded. **(b)** Our method treats the lower arm of another person as its own due to occlusion. **(c)** Interpenetration around the right hand.

many artifacts where the estimated mesh is not smooth, and they also fail to align the image well. In contrast, our method estimates more accurate human poses and shapes and has smooth human mesh results. In addition, it is more robust to truncation and occlusion and aligns the image better.

Figure 6 shows more quality results of our approach on the 3DPW [25], H3.6M [4], MPI-INF-3DHP [19], and COCO [15] datasets. Figure 2 shows more qualitative results on Internet images with challenging cases, such as extreme body shapes or complex poses. Our method generalizes well on the natural scenes. Figure 3 shows typical failure cases, including inaccurate shape estimation and interpenetration, which are mainly caused by inaccurate 3D virtual marker estimation when occlusion occurs. But as expected, the rest body parts are barely affected due to the *local and sparse* property of the virtual marker.

3. Human Subject Data

We use existing public datasets of human subjects in our experiments following their official licensing requirements. With proper usage, the proposed method could be beneficial to society (*e.g.* elderly fall detection).

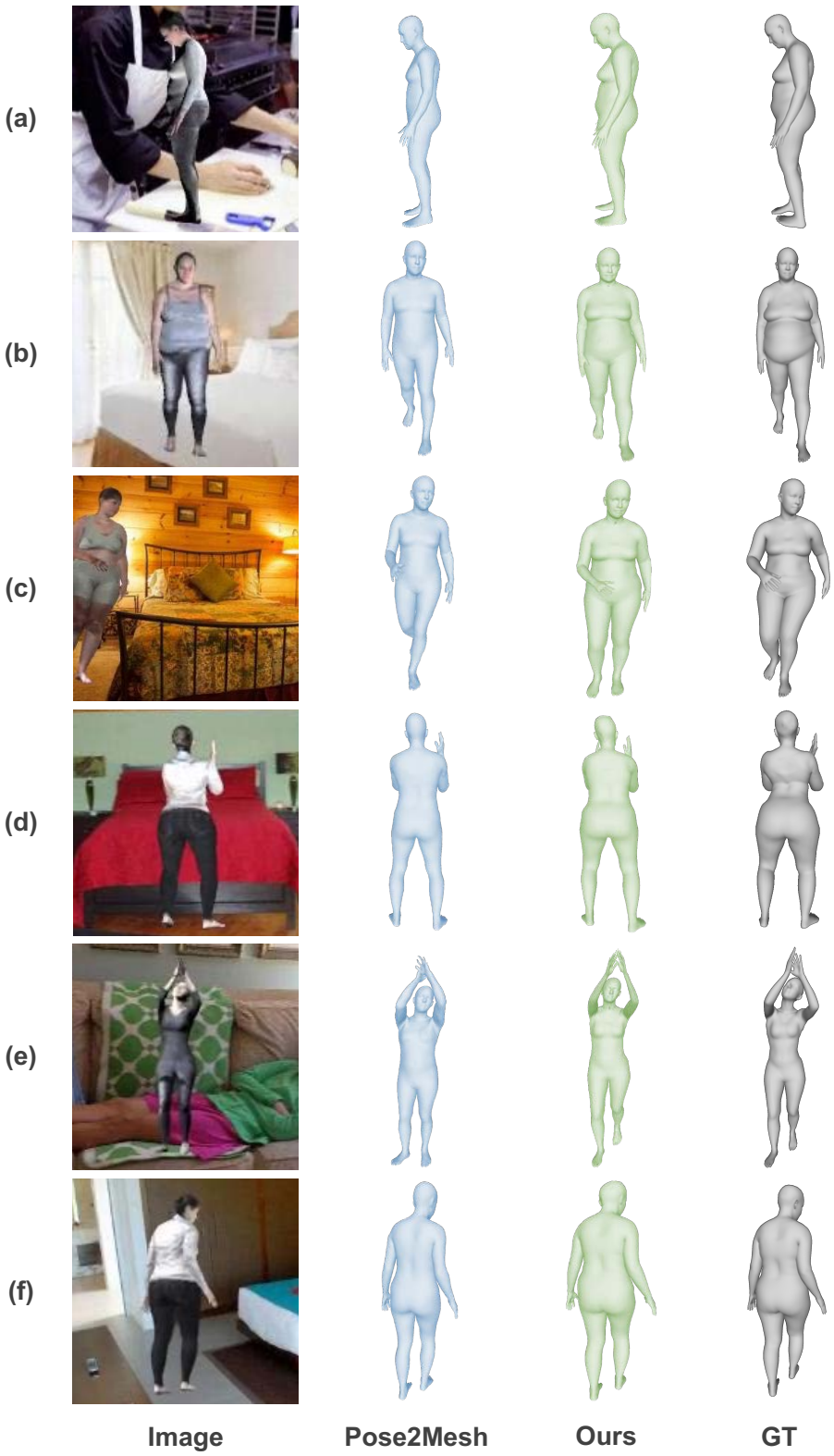

Figure 4. Qualitative comparison between our method and Pose2Mesh [2] on SURREAL test set [24]. Our approach generates more accurate mesh estimation results on images of diverse body shapes.

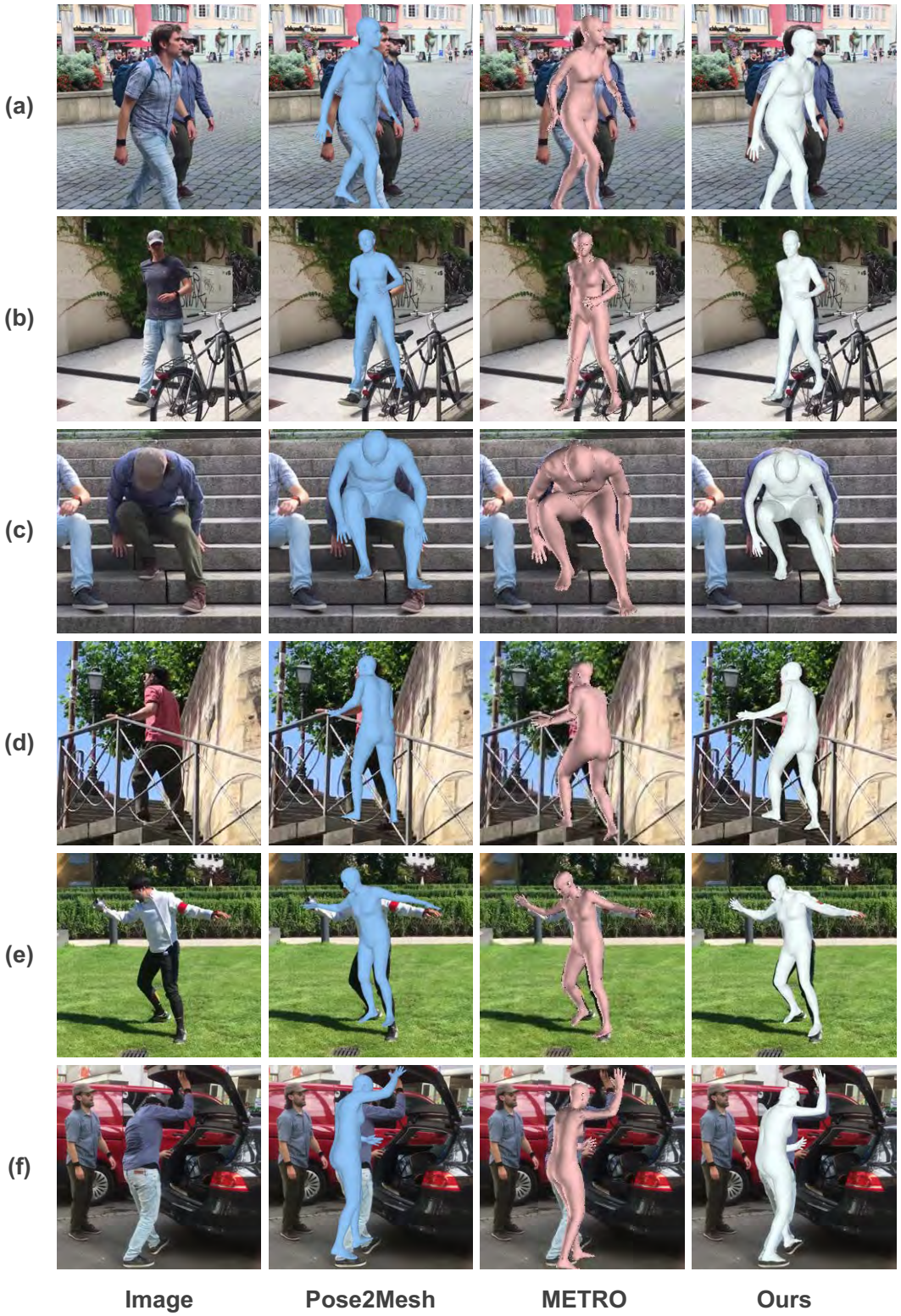

Figure 5. Qualitative comparison between our method and Pose2Mesh [2], METRO [13] on 3DPW test set [25]. Our approach is more robust to occlusion and truncation and generates more accurate mesh estimation results that align images well.

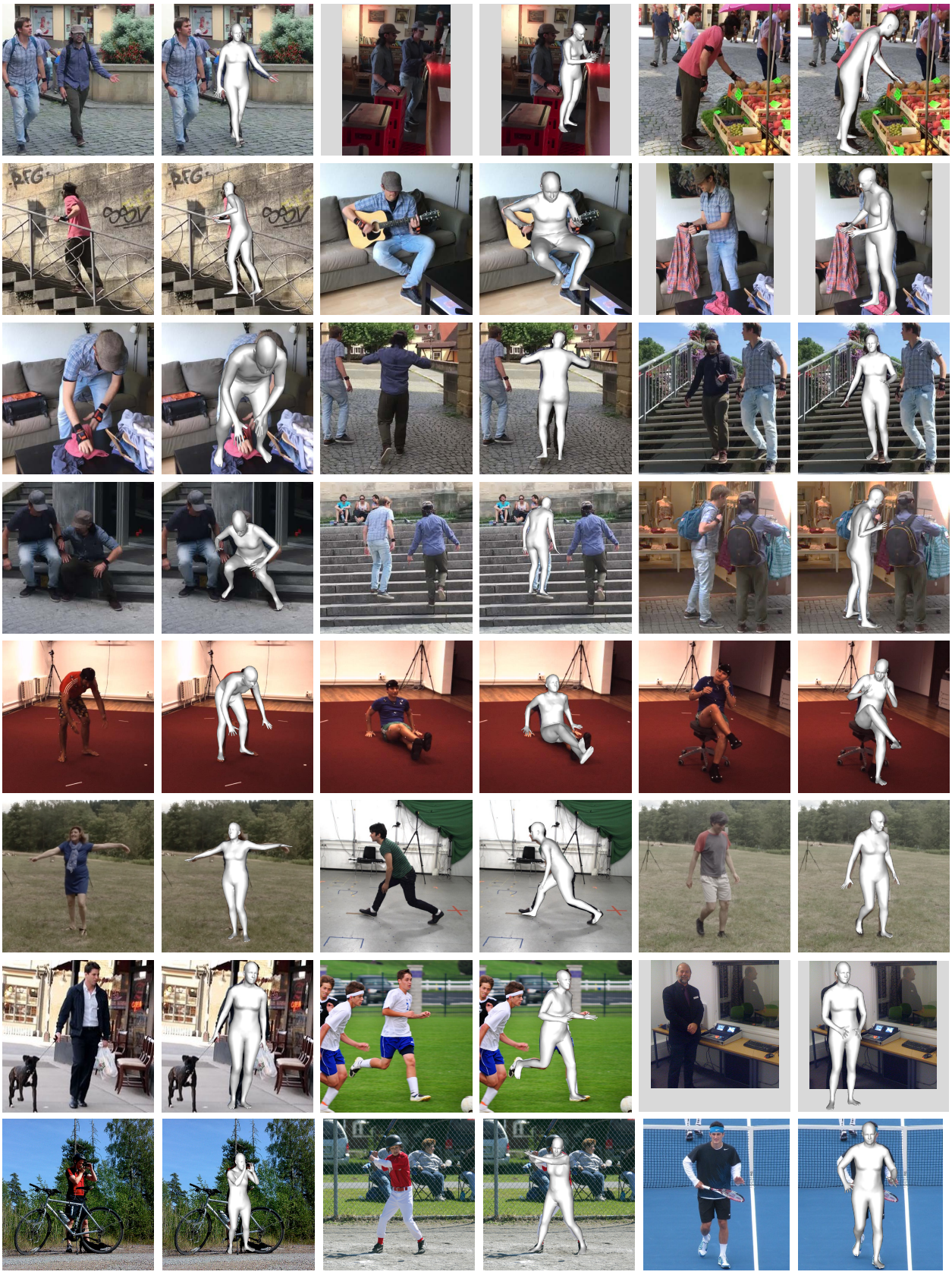

Figure 6. Meshes estimated by our approach on images from the 3DPW [25] dataset (row 1-4), H3.6M [4] dataset (row 5), MPI-INF-3DHP [19] dataset (row 6), and COCO dataset (last 2 rows) [15].

References

[1] 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*, pages 561–578, 2016. 2

[2] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *ECCV*, pages 769–787, 2020. 2, 3, 4, 5

[3] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. 1

[4] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE TPAMI*, 36(7):1325–1339, 2013. 1, 2, 3, 6

[5] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *ICCV*, pages 7718–7727, 2019. 2

[6] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131, 2018. 1, 2

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 2

[8] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, pages 5253–5263, 2020. 2

[9] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. Pare: Part attention regressor for 3d human body estimation. In *ICCV*, pages 11127–11137, October 2021. 1

[10] 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*, pages 2252–2261, 2019. 1, 2

[11] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, pages 4501–4510, 2019. 1, 2

[12] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *CVPR*, pages 6050–6059, 2017. 2

[13] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, pages 1954–1963, 2021. 1, 2, 3, 5

[14] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *ICCV*, pages 12939–12948, 2021. 1, 2

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 2, 3, 6

[16] Matthew Loper, Naureen Mahmood, and Michael J Black. Mosh: Motion and shape capture from sparse markers. *TOG*, 33(6):1–13, 2014. 1, 2

[17] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *TOG*, 34(6):1–16, 2015. 1

[18] Tianyu Luan, Yali Wang, Junhao Zhang, Zhe Wang, Zhipeng Zhou, and Yu Qiao. Pc-hmr: Pose calibration for 3d human mesh recovery from 2d images/videos. In *AAAI*, pages 2269–2276, 2021. 2

[19] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3DV*, pages 506–516, 2017. 1, 2, 3, 6

[20] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *ECCV*, pages 752–768, 2020. 2

[21] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, pages 10975–10985, 2019. 2

[22] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *CVPR*, pages 459–468, 2018. 1

[23] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, pages 529–545, 2018. 2

[24] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, pages 109–117, 2017. 2, 3, 4

[25] 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*, pages 601–617, 2018. 1, 2, 3, 5, 6

[26] Mihai Zanfir, Andrei Zanfir, Eduard Gabriel Bazavan, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Thundr: Transformer-based 3d human reconstruction with markers. In *ICCV*, pages 12971–12980, 2021. 1, 2