Learning to Reconstruct 3D Human Pose and Shape via Model-fitting in the Loop **Supplementary Material**

Nikos Kolotouros*[1], Georgios Pavlakos*[1], Michael J. Black[2], Kostas Daniilidis[1]
[1] University of Pennsylvania [2] Max Planck Institute for Intelligent Systems

The goal of this Supplementary Material is to provide additional details that were not included in the main manuscript due to space constraints. In Section 1 we present additional quantitative results. Section 2 aims to provide qualitative results for a wide range of settings, including: visualization from novel viewpoints, comparison with the approach of Kanazawa *et al*. [6], comparison of the “unpaired” version and the version that has access to 3D ground truth, etc. Then, in Section 3, we provide more details about the training procedure. Finally, in Section 4, we discuss the evaluation metrics used to report results.

1. Further quantitative evaluation

In this Section we provide more quantitative results of our approach that complement and extend the results found in the main manuscript.

Different dictionary: As we mentioned in Section 3.5 of the main manuscript, we can follow different strategies to initially populate with SMPLify fits the dictionary our approach uses to keep track of the best fits during training. For the results used in the main empirical evaluation, we use a network similar to Martinez *et al*. [8] trained on CMU MoCap data [2, 12] and we regress an initial pose from the given 2D keypoints. SMPLify is initialized with this pose and provides the fit we add in the dictionary. Alternatively, we can run SMPLify from the mean pose, without requiring an external pose regressor. Here, we provide results for this setting as well. In general, as we can see in Table 1, performance remains similar even if we start with this dictionary. More importantly, our in the loop optimization is responsible for consistent improvement in the results, compared to using only the initial dictionary of static fits. This experiment offers additional evidence, that SPIN can improve the performance of the network, regardless of the quality of the initial (non-perfect) dictionary we use in our approach.

	3DPW	LSP (masks)	MPI-INF
static fits (from mean pose init)	66.2	90.99%	71.4
in the loop (from mean pose init)	62.3	**91.85%**	68.1
static fits (from [8] init)	66.3	91.07%	70.1
in the loop (from [8] init)	**59.2**	91.83%	**67.5**

Table 1: Comparison of the proposed in the loop optimization, with vanilla training using the static dictionary of fits we use to initialize SPIN. The static fits are the result of SMPLify initialized with the mean pose, or the regressed pose value of a lifting network similar to the one proposed by Martinez *et al*. [8]. The numbers for 3DPW and MPI-INF are reconstruction errors in mm, while the numbers for LSP are mask segmentation accuracies.

	3DPW	LSP (masks)
regressor	59.2	91.83%
regressor + fit	66.3	89.93%

Table 2: Comparison between the mesh regressed by our network (“regressor”) with the reconstruction we get after applying a post-processing to this mesh by fitting it to *predicted* 2D joints (“regressor+fit”). Since the predicted joints can be noisy, we see that the fit accuracy can potentially decrease by using this post-processing. The numbers for 3DPW are reconstruction errors in mm, while the numbers for LSP are mask segmentation accuracies.

Fitting at test time: SPIN leverages a tight collaboration between the optimization-based and the regression-based approach at training time, to improve the performance of a deep regressor. Other recent approaches, e.g., [3, 10, 11] have also investigated the collaboration of the paradigms, focusing on test time inference. This is an option we also investigated. However, we observed that a potential problem of this strategy is that often the *predicted* 2D joints can be inaccurate at test time and eventually reduce the prediction quality. We observed that this is the case for different datasets and we report our findings in Table 2. In contrast to that, at training time, we have access to ground truth 2D joints, so we can consistently improve accuracy compared to the regressed shape.

Dictionary improvement: To more explicitly discover the level of improvement of the fits we use for training, we compare our initial set of fits (dictionary at the beginning of

	Human3.6M	MPI-INF-3DHP
initial dictionary	72.5	83.4
final dictionary	60.5	74.5

Table 3: Comparison between the accuracy of the fits in the initial and the final dictionary, for the training set of Human3.6M and MPI-INF-3DHP. The results are mean reconstruction errors in mm. As expected, the fits we use for training improve over time, so it is natural to lead to a more accurately trained regressor network.

	Rec. Error
Training in the loop (from mean pose init)	62.3
Training with final dict (from mean pose init)	62.2
Training in the loop (from [8] init)	59.2
Training with final dict (from [8] init)	59.9

Table 4: Comparison of our in the loop training with vanilla training using the final dictionary recovered by our in the loop approach. The numbers are mean reconstruction errors in mm on the 3DPW dataset. Performance is almost identical in both cases. This implies that the benefit from SPIN comes explicitly from recovering more accurate fits, and not because the gradual improvement of supervision can interfere with the training procedure.

training) with the final fits recovered when we finish our in the loop approach (dictionary at the end of training). The results for Human3.6M and MPI-INF-3DHP are presented in Table 3. These results correspond to the "unpaired" setting, where we do not have explicit 3D ground truth and we can only start with inaccurate fits. As expected, we observe an improvement of the fits quality in the final dictionary, compared to the initial fits. These improved fits help our network to be more accurate at the end of our in the loop training.

Effect of improved fits: To demonstrate the benefit of using the improved ground truth shapes that SPIN provides, we also consider training a network using only the updated shapes that are included in the dictionary after the end of training. This way, we can demonstrate that the network performance improves explicitly because the ground truth shapes are more accurate, and not because the labels improve in a gradual manner during training (which could interplay with the idiosyncrasies of neural network training). The results for the 3DPW dataset are provided in Table 4. Considering the small difference between the results of the two models, we deduce, that the improved quality of the fits we recover during training is what primarily contributes to the improved performance of our in the loop models.

2. Further qualitative evaluation

In this Section we provide more qualitative results of our approach, that were not included in the main manuscript due to space constraints.

Side views: Figure 5 of the main manuscript provides a variety of qualitative results of our approach for all the datasets involved in our quantitative evaluation. Here we provide even more examples, with the addition of visualizations from novel viewpoints, which is a typical way to evaluate qualitatively 3D human pose estimation methods. These additional visualizations including novel viewpoints have been collected in Figure 1.

SMPLify failures: In the main manuscript (Subsection 3.5), we discuss the typical failure modes of SMPLify. Here we provide more visual results of these failures in Figure 2. These errors motivate our decision to avoid training with some very bad fits that SMPLify can provide to our network. From inspection, these failures include wrong orientation of the body and/or extreme shape parameters. In the second case of extreme shape parameters, we observed that the camera translation is typically off (estimated to be too close or too far), because the assumptions of [1] are violated (i.e., the person is not standing parallel to the image plane). It is important to clarify though, that these failures happen when the optimization starts from the mean pose, and results are typically improved over the course of training, when the SMPLify routine is initialized with a reasonable pose estimate from the network.

Comparison with HMR [6]: Based on the results of the main manuscript, our closest competitor is the HMR approach of Kanazawa *et al.* [6]. To provide additional intuition over the benefits of our approach with respect to [6], (beyond the quantitative results), here we include further qualitative comparison with our approach, by applying HMR and our network on the same images. Example reconstruction from both approaches are presented in Figure 3. Based on this comparison, we identify that although HMR is quite robust, it has more issues with estimating the global orientation correctly, while it is less accurate for the body extremities. In contrast, our network is trained with successful SMPLify fits, which tend to get these cases correctly, so we observe more successful reconstructions also from a qualitative point of view.

"Paired" vs "Unpaired" supervision: Although for our best models we do use examples where 3D ground truth is available for training (e.g., Human3.6M and MPI-INF-3DHP), our approach is applicable even when we have access to no image with corresponding 3D ground truth. Here, we provide a qualitative comparison between these two training settings. The corresponding results are presented in Figure 4. Interestingly, our "unpaired" network produces very similar results to the network that has been trained with limited access to 3D ground truth. Significant differences can be observed only in cases with very challenging poses, or in cases with ordinal depth ambiguities, where SMPLify itself is also prone to failure.

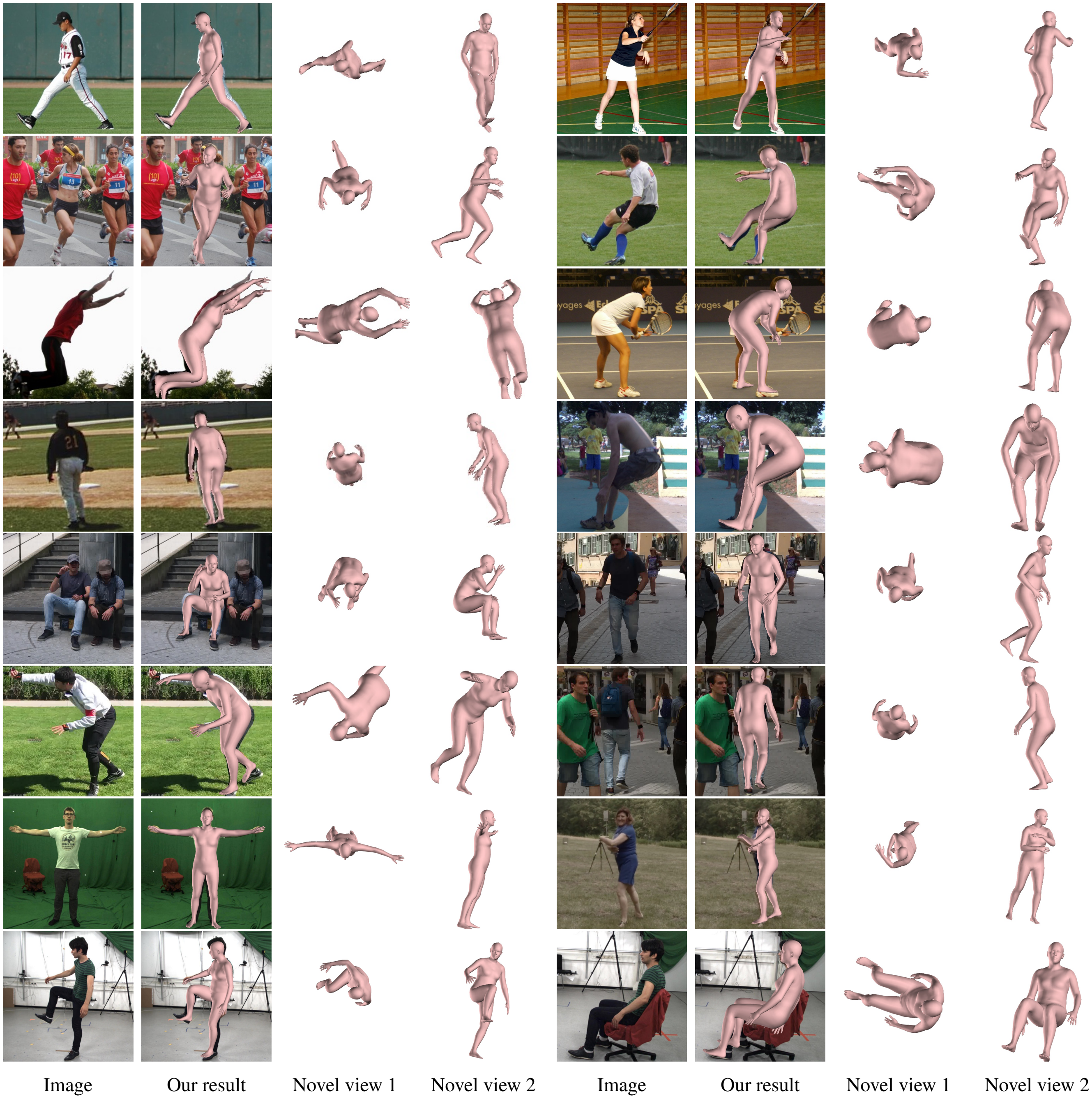

Figure 1: Successful results of SPIN. For each example from left to right: Image, Our reconstruction result in the camera frame, Our reconstruction result from a novel view (top view), Our reconstruction result from a novel view (side view).

3. Training details

Our model follows the architecture of Kanazawa *et al*. [6]. The only difference is that instead of using an axis-angle representation for the 3D rotations (as done by [6]), we instead change the output to regress the representation of Zhou *et al*. [14]. Our models were trained using the Adam optimizer with a batch size of 64, and the learning rate set to $3e-5$. We did not use learning rate decay. Training with SPIN lasts for $300k$ iterations. The model without access to any form of 3D ground truth ("unpaired") was initialized from a model pretrained on ImageNet. The model with limited access to 3D ground truth ("paired") was initialized with a model pretrained on Human3.6M [4] using full 3D

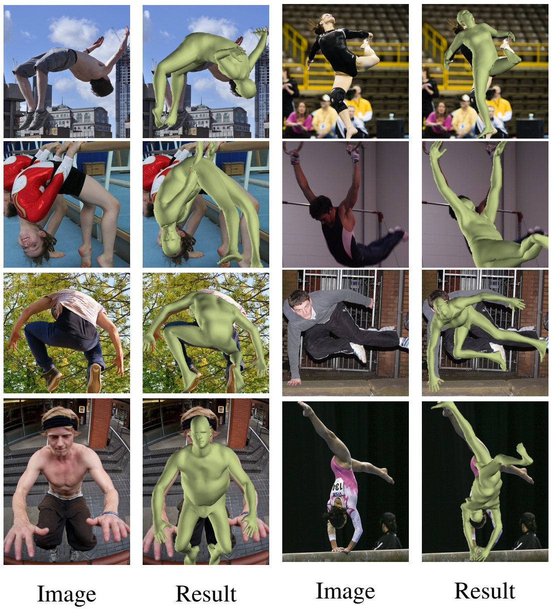

Figure 2: Erroneous reconstructions of SMPLify. Failures typically occur because of errors in the orientation of the body or specific parts (first and second row), or in the estimated shape parameters (third and fourth row). In the second case, the distance from the camera has been heavily over- or under-estimated, which can produce extreme values for the shape parameters.

pose and shape ground truth. Pretraining in this case was useful, such that the model provides better initial 3D shape estimates for the iterative fitting. The weights for the losses on the SMPL pose and beta parameters were set to 1 and 0.001 respectively. The loss weights on the 2D and 3D keypoints were set to 5. We did not use a loss at the mesh level. All training losses were then scaled by a factor of 60.

The SMPLify optimization in the loop is done using the Adam optimizer in batch mode. The step size was set to $1e-2$ and and the maximum number of iterations to 50. In these conditions, for a batch size of 64 images, the optimization takes about 3 seconds on a GeForce 1080Ti GPU, allowing us to include it within the training loop.

4. Evaluation metrics

In the main manuscript, we report results using a variety of different metrics, always following the literature and computing the metrics the same way that competing approaches do. In this Section, we provide more details about the relevant metrics and give pointers to previous works that use or define them.

Rec. Error: In Tables 1 and 3 of the main manuscript we report results on 3DPW and Human3.6M respectively using the Reconstruction error. This error computes the mean Euclidean error over all the joints after aligning the prediction with the ground truth 3D pose through Procrustes alignment. A definition of this error is given formally in [13]

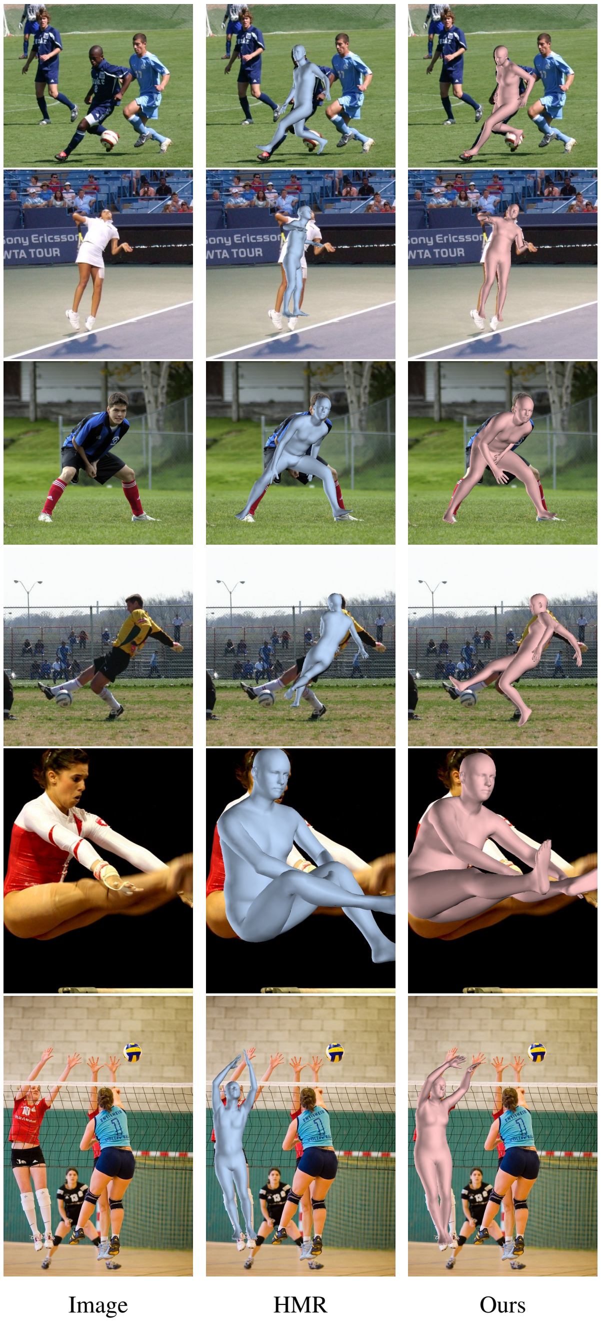

Figure 3: Comparison of SPIN with HMR [6] on the LSP dataset [5]. From left to right: Input image, HMR result, Our result. HMR failures include errors in the estimation of the global orientation and the pose of the extremities (arms and legs). In contrast, SPIN is more robust in these cases, because adding the optimization in the training loop, provides more accurate supervision to the network.

Segmentation: In Table 2 of the main manuscript we evaluate 3D shape implicitly through mesh reprojection us-

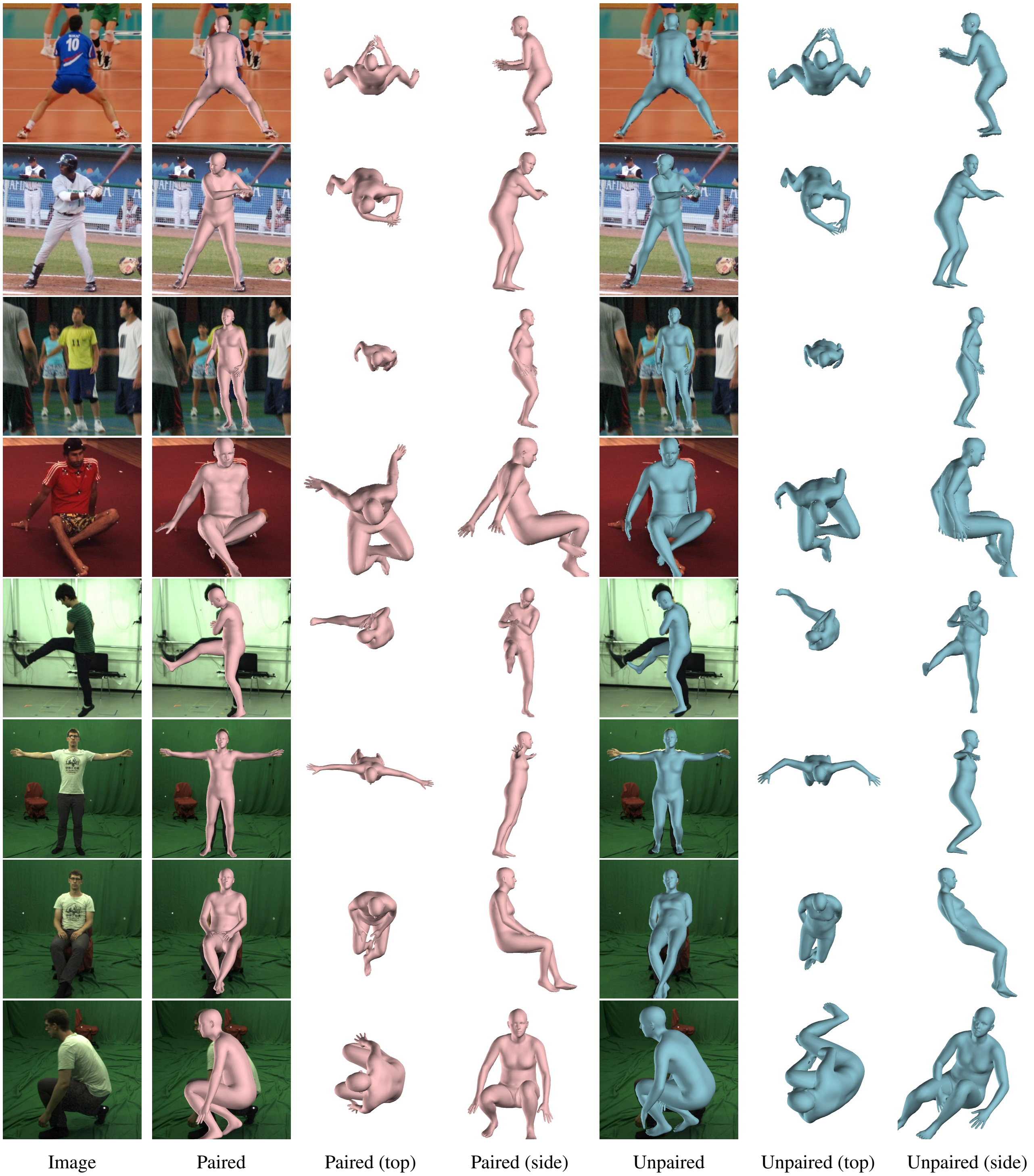

Figure 4: Comparison of “unpaired” model (no access to images with 3D ground truth) with the “paired” version (limited access to images with 3D ground truth). For each row from left to right: Image, Reconstruction result of “paired” model in the camera frame, Reconstruction result of “paired” model from top view, Reconstruction result of “paired” model from side view, Reconstruction result of “unpaired” model in the camera frame, Reconstruction result of “unpaired” model from top view, Reconstruction result of “unpaired” model from side view. Interestingly, in most cases the two versions recover similar human shapes. Important differences can only be observed in the presence of very challenging poses.

ing segmentation accuracy metrics. We report accuracy scores and f1 scores when considering only the silhouette (FB - Foreground/Background case), and also considering Part segmentation. The evaluation on LSP using these segmentation metrics is originally done by Lassner *et al*. [7].

MPI-INF-3DHP evaluation: The evaluation on MPI-INF-3DHP [9] in Table 4 of the main manuscript includes a variety of metrics reported with or without rigid alignment, i.e., with or without aligning our prediction with the ground truth using Procrustes alignment. In this case, MPJPE stands for the mean Euclidean error over all the joints. PCK is the percentage of correctly localized keypoints, where a keypoint is considered to be correctly localized if its Euclidean error is below a specific threshold (here $150mm$). Finally AUC stands for Area Under the Curve and is computed as in [9], by estimating the PCK for a variety of thresholds, from 0 to 150, with a step equal to 5.

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*, 2016. 2

[2] CMU. Graphics lab motion capture database. http://mocap.cs.cmu.edu, 2000. 1

[3] Riza Alp Guler and Iasonas Kokkinos. HoloPose: Holistic 3D human reconstruction in-the-wild. In *CVPR*, 2019. 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. *PAMI*, 36(7):1325–1339, 2014. 3

[5] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. 4

[6] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 1, 2, 3, 4

[7] 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*, 2017. 6

[8] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3D human pose estimation. In *ICCV*, 2017. 1, 2

[9] 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*, 2017. 6

[10] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *CVPR*, 2018. 1

[11] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *ECCV*, 2018. 1

[12] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 1

[13] Xiaowei Zhou, Menglong Zhu, Georgios Pavlakos, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. MonoCap: Monocular human motion capture using a CNN coupled with a geometric prior. *PAMI*, 41(4):901–914, 2019. 4

[14] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 3