MonoClothCap: Towards Temporally Coherent Clothing Capture from Monocular RGB Video (Supplementary Material)

1. Further Ablation Studies

In this section, we conduct more ablation studies on various loss terms we use in the energy optimization for clothing capture and body shape estimation.

1.1. Loss Terms for Clothing Capture

We first study the loss terms used for clothing capture in Section 5.2 (Eq. 12). In the experiments below, we compare the results of the batch optimization stage with different loss terms, initialized from the same body capture and sequential tracking results.

Clothing segmentation term (Eq. 13). In order to study the effect of the clothing segmentation term, we run an ablative experiment where the weight for the segmentation term is set to 0, while all other terms remain the same. To better visualize the effect, we render the output meshes in three colors: grey for skin, yellow for upper clothing and green for lower clothing. We consider a vertex j as a skin vertex if the length of the clothing offset for this vertex is below a certain threshold ε, or

$$\|D_j\| < \varepsilon,$$

where D_j is defined in Eq. 4 in the main paper. We consider a vertex as belonging to the upper clothing if

$$\|D_j^u\| \geq \|D_j^l\| \quad \text{and} \quad \|D_j\| \geq \varepsilon,$$

or, similarly, as belong to the lower clothing if

$$\|D_j^l\| > \|D_j^u\| \quad \text{and} \quad \|D_j\| \geq \varepsilon.$$

The result of this experiment is shown in Fig. 1. In each frame, we observe that the boundary between the upper and lower clothing is *more consistent* with the original image in the result with segmentation term than the result without segmentation term. Our method adopts a combination of upper clothing and lower clothing models, which might both have non-zero offsets around the body waist. It is important for our method to produce both offsets with correct relative length to realistically reconstruct the spatial arrangement of the T-shirt and trousers in the original images. This result proves the effectiveness and necessity of the clothing segmentation term.

Photometric tracking term (Eq. 14). Similarly, we run an ablative experiment where the weight for the photometric tracking term is set to 0 and other terms remain the same. To visualize its effect, we render the output tracked mesh with the final texture extracted in the sequential tracking stage (see Section 5.2 of the main paper for detail), and compare the results with and without the photometric tracking term with the original images.

The result of this experiment is shown in Fig. 2. Notice that the same final texture image is used to render all the results. In order to assist visual comparison of the rendered pattern, we draw several auxiliary horizontal dashed lines in red. We can observe that the results with photometric tracking term is more consistent with the original image than the result without photometric tracking term, in terms of the location of the white strip on the T-shirt and the boundary between the T-shirt and trousers. This demonstrates that our photometric tracking loss can help to obtain better temporal correspondence across different frames in the video.

Silhouette matching term. We now compare the results with and without the silhouette matching term. We render both results and align them with the original images to visualize how well the silhouette matches.

The result of this experiment is shown in Fig. 3. We observe that the result with silhouette matching term achieves a better alignment of silhouette with the original image. This suggests that the silhouette matching term can help to reconstruct the accurate shape of the clothing in the video.

1.2. Losses Terms for Body Shape Estimation

Although body pose and shape estimation is not a focus of this paper, we conduct ablative studies on the loss terms used in body shape estimation in Section 5.1. (Eq. 9). In each of the experiment in this section, the weight for the loss term under study is set to 0, and all other terms stay the same as the full results. We render the estimated body shapes and compare them with the full results.

Silhouette term. The result of this experiment is shown in Fig. 4. We can observe in the result that silhouette provides critical information for the estimation of body shape and pose in the following two ways. First, the projection of human body should always lie in the interior of the overall silhouette in the image, which includes the region of body and clothes. Second, in the top-right and bottom-left examples, an arm of the subject is occluded by the torso. There is no available information to reason about the location of the arm from the 2D keypoints or DensePose results. In this sit-

Ours	Temporal HMR [2]	SPIN [4]	VIBE [3]
77.3	94.7	89.5	87.2

Table 1: Quantitative comparison with recent SMPL-based 3D body pose estimation approaches on the *Pablo* sequence. All numbers are in mm.

uation, only the silhouette can constrain the position of the arm to be behind the torso in the camera view. This proves the importance of the silhouette term for accurate estimation of human body and shape.

DensePose term. The result of this experiment is shown in Fig. 5. The use of DensePose together with SMPL model for accurate body estimation was first proposed in [1]. In our work, we find that the DensePose term helps to estimate the hand orientation more accurately, as fingers are usually not included in the hierarchy of 2D body pose output.

POF term. The result of this experiment is shown in Fig. 6. The use of POF together with deformable human body model was first proposed in [6]. We find that the POF term can help to eliminate the ambiguity of 3D body pose given only 2D keypoints in the front view, and therefore help to estimate more accurate body pose in 3D.

2. Quantitative Comparison with Monocular 3D Pose Estimation Methods

In the first stage of our pipeline, we use a standard model-fitting method to estimate 3D body pose from the video. Although we do not claim any contribution or novelty in this aspect, we still provide a quantitative comparison with recent state-of-the-art approaches that estimate 3D body pose with SMPL model from a monocular view. In particular, we evaluate all methods on the *Pablo* sequence using the same protocol as Section 6.1 in the main paper. The evaluation results are shown in Table 1. As a part of our pipeline, our estimation of 3D body pose is highly accurate even when compared with recent state-of-the-art approaches that focus on 3D body pose only. This lays a solid foundation for the following clothing capture stages.

3. Runtime Analysis

In this section, we present the runtime information of our approach. Our method runs on a Linux server with 40 CPU cores and 4 GTX TITAN X GPUs. Our approach requires the memory of 4 GPUs in order to run the batch optimization on a video of around 250 frames together. For optimization, we use the L-BFGS solver implemented in PyTorch [5]. We measure the average time consumed for each frame in every stage, and the results are shown in Table 2.

Stage	Runtime (s)
Body Estimation (Sec. 5.1)	6
Sequential Tracking (Sec. 5.2)	62
Batch Optimization (Sec. 5.2)	27
Wrinkle Extraction (Sec. 5.3)	232
Total	327

Table 2: Average per-frame runtime of each stage in our pipeline. The numbers are in seconds.

4. Complete Quantitative Evaluation Results

In this section, we present the figures for complete per-frame results of the quantitative experiments conducted in Section 6 of the main paper.

4.1. Evaluation on MonoPerfCap Dataset

Evaluation of Clothing Surface Reconstruction. The complete per-frame results corresponding to the surface error in Table 1 in the main paper are shown in Fig. 7.

Evaluation of 3D Pose Estimation. The complete per-frame results corresponding to the joint error in Table 1 in the main paper are shown in Fig. 8.

4.2. Evaluation on BUFF Dataset

The complete per-frame results corresponding to Table 2 in the main paper are shown in Fig. 9.

References

[1] R. A. Guler and I. 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

[2] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5614–5623, 2019. 2

[3] M. Kocabas, N. Athanasiou, and M. J. Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5253–5263, 2020. 2

[4] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2252–2261, 2019. 2

[5] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019. 2

[6] D. Xiang, H. Joo, and Y. Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10965–10974, 2019. 2

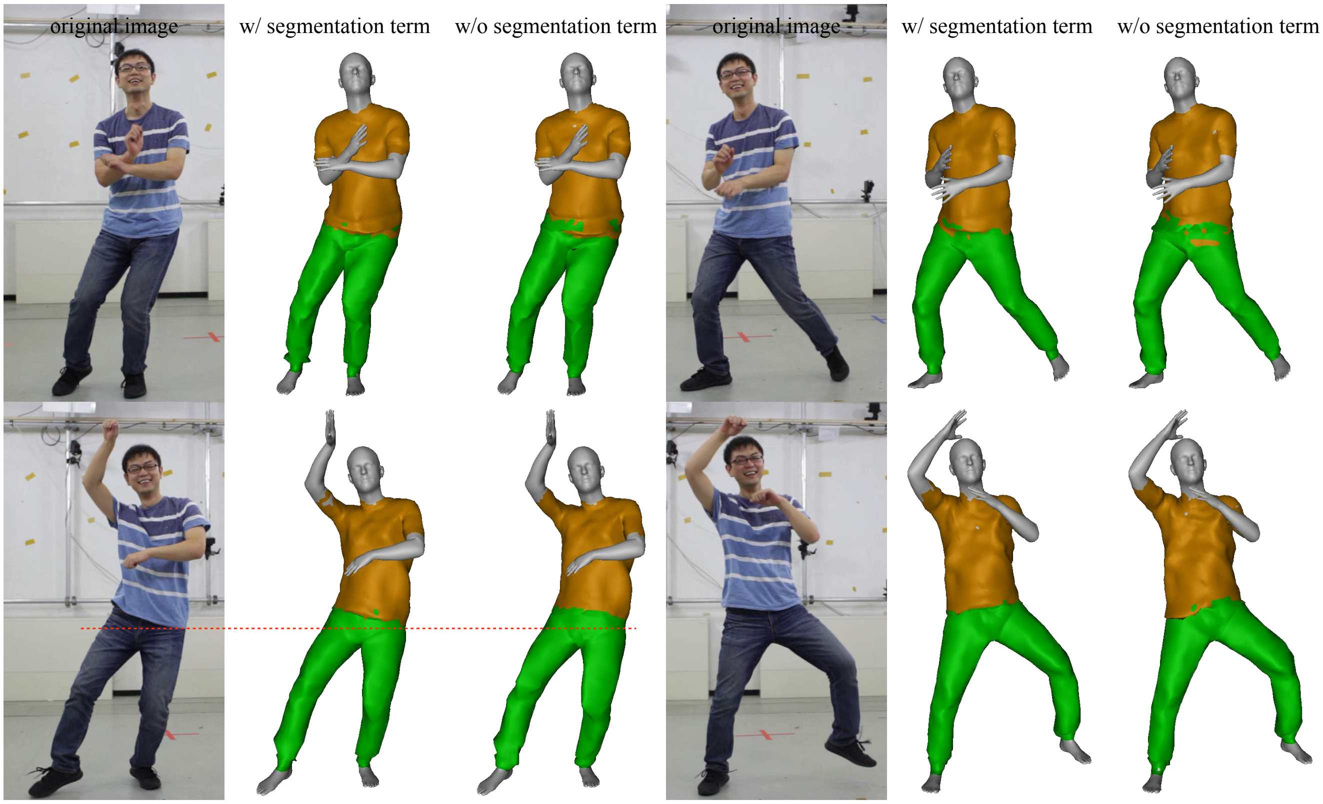

Figure 1: Comparison between results with and without clothing segmentation loss. The vertices for skin, upper clothing and lower clothing are rendered in grey, yellow and green respectively. A horizontal red dashed line is drawn in the bottom left example to help visually check the location of the boundary between upper and lower clothing.

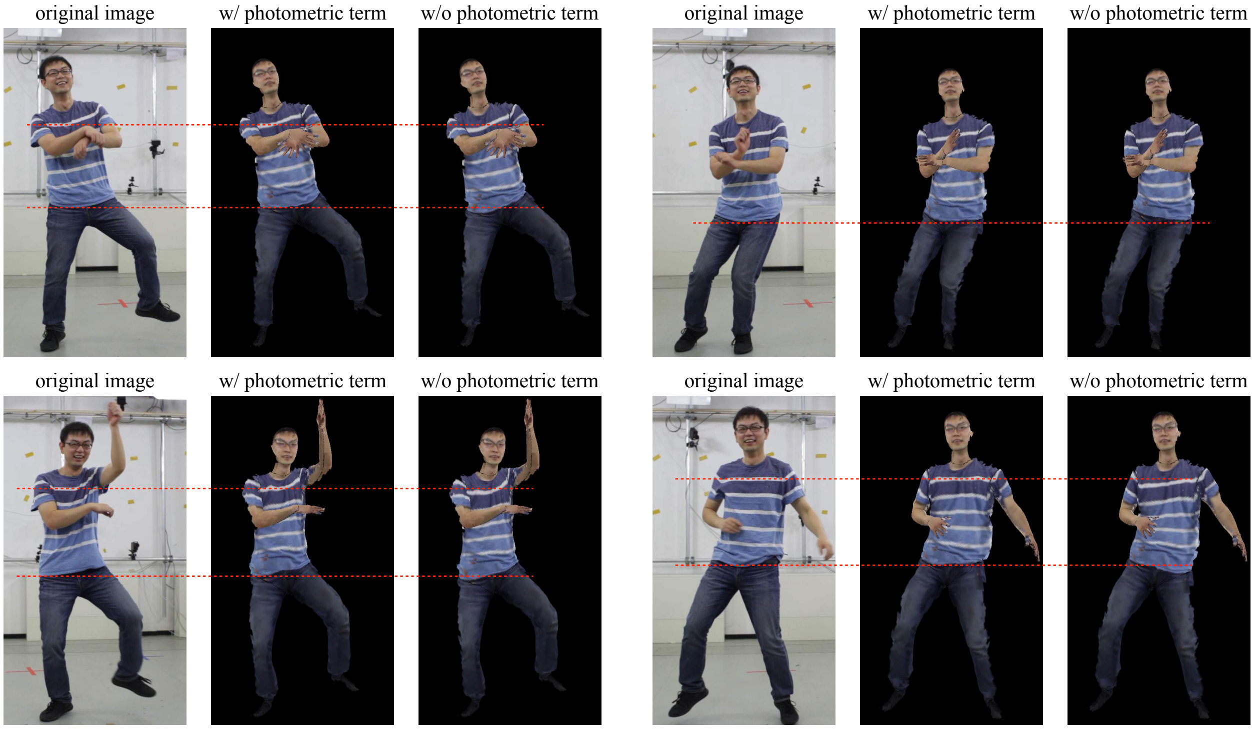

Figure 2: Comparison between results with and without photometric tracking loss. Horizontal dashed lines are drawn in red to help visually compare the location of rendered texture pattern.

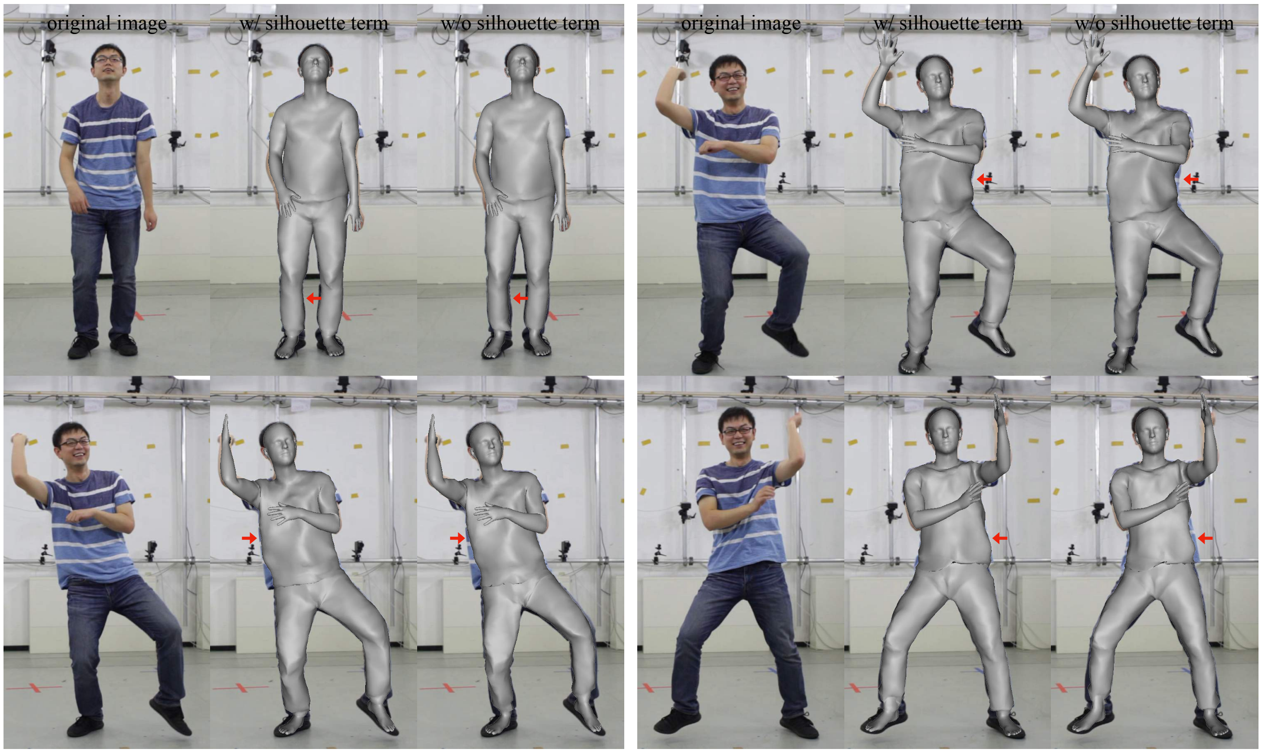

Figure 3: Comparison between results with and without silhouette matching loss. In each example, we show the original image, the result with and without silhouette matching loss from left to right.

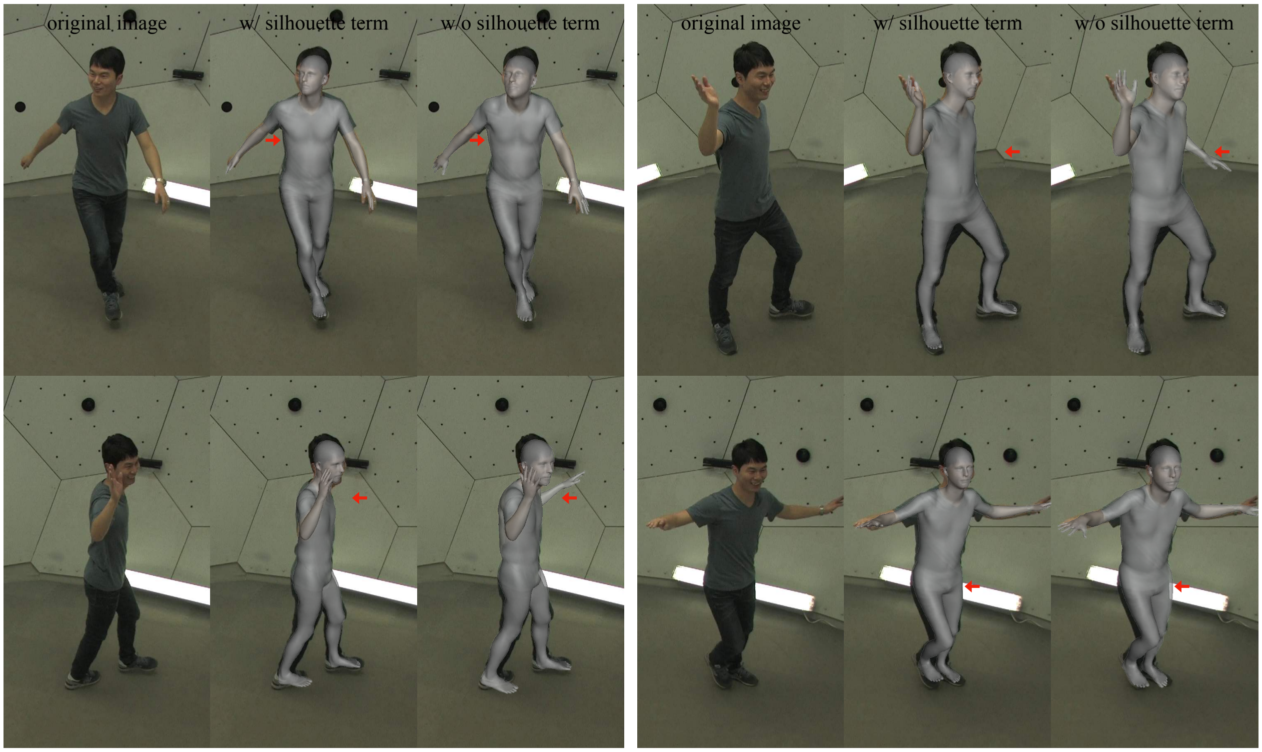

Figure 4: Comparison between results with and without silhouette loss. In each example, we show the original image, the result with and without silhouette loss from left to right.

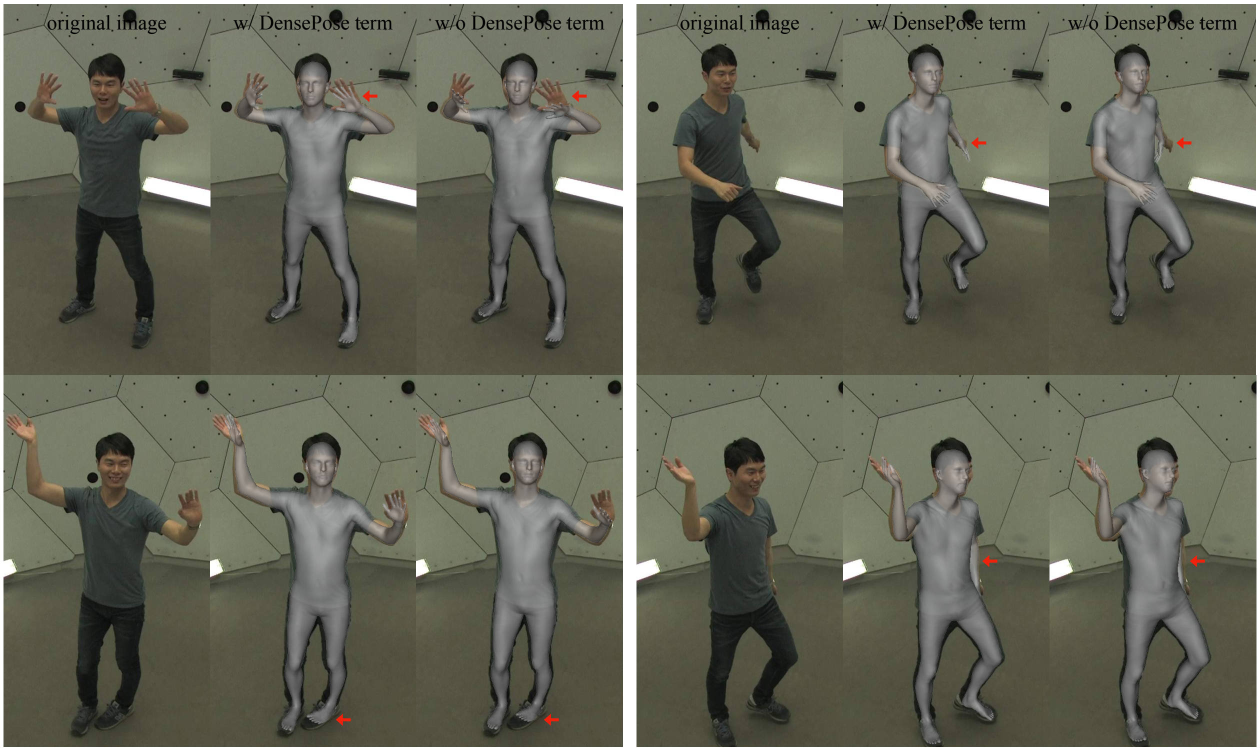

Figure 5: Comparison between results with and without DensePose loss. In each example, we show the original image, the result with and without DensePose loss from left to right.

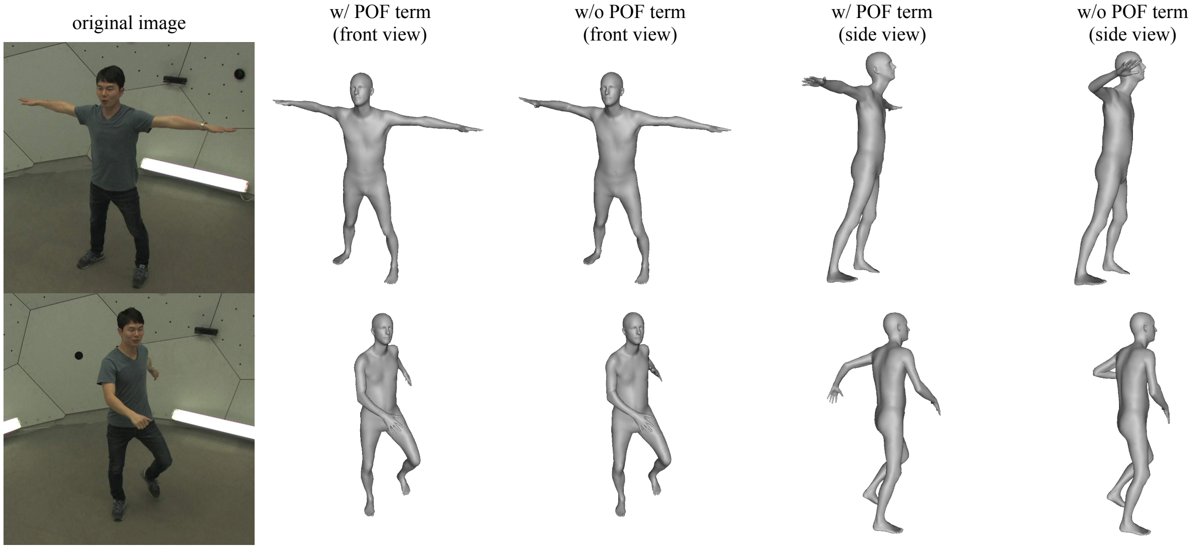

Figure 6: Comparison between results with and without POF loss. In each example, we show the original image, the result with and without POF loss from both the front view and the side view.

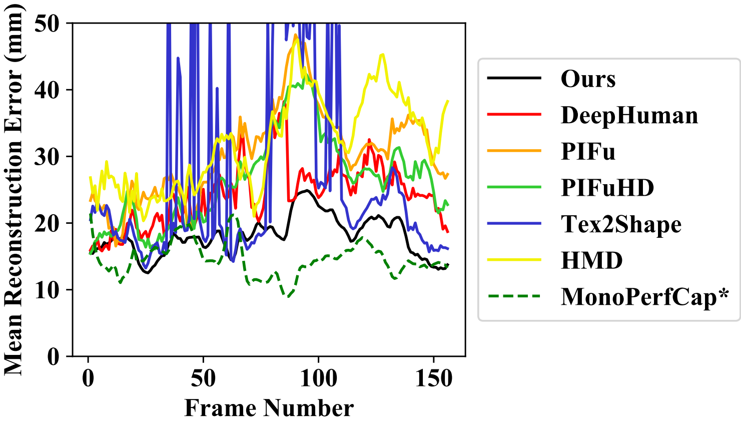

Figure 7: Per-frame results of the quantitative comparison with previous work on *Pablo* sequence using mean point-to-surface error. Notice that the method annotated with '*' uses a pre-scanned personalized template that provides strong shape prior.

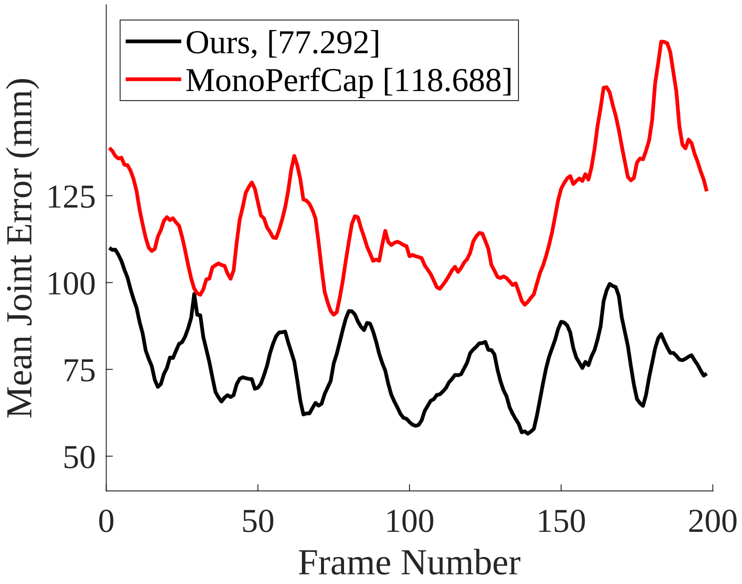

Figure 8: Per-frame results of the quantitative comparison with previous work on *Pablo* sequence using mean joint error.

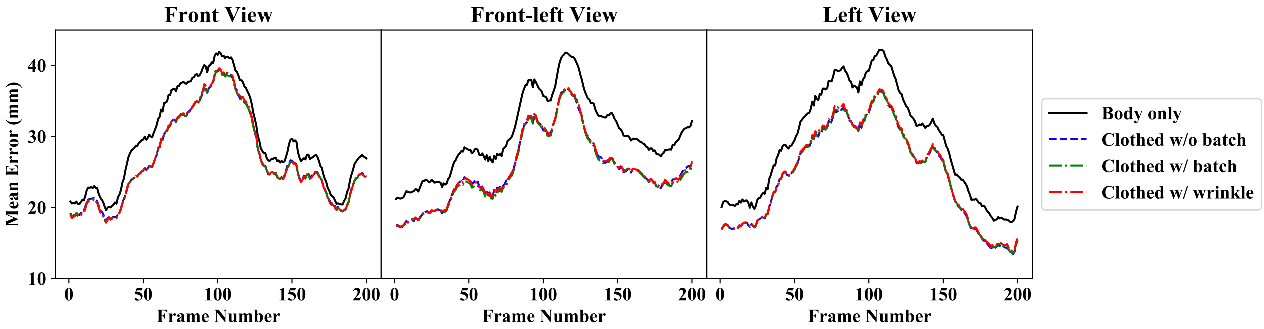

Figure 9: Per-frame results of the quantitative ablation study for different stages of our method on rendered BUFF dataset using mean point-to-surface error.