THUNDR: Transformer-based 3D HUmaN Reconstruction with Markers
Supplementary Material

In this supplementary material we provide more details on the Marker Poser and mesh fitting process and on the influence of adding noise when training the Marker Poser. We also include videos of our reconstructions from THUNDR on images with various backgrounds, illumination conditions, poses and clothing. We also provide a visual comparison with another state-of-the-art method, SPIN [1]. For some sample comparisons see fig. 2.

1. Additional Details

In THUNDR, we set λ, the step size used in equation 6., equal to 0.1 and the various weights for our losses as follows: $\lambda_{ps} = 2.5$, $\lambda_m = 50$, $\lambda_k = 1$, $\lambda_b = 25$, $\lambda_v = 0.75$ and $\lambda_j = 0.25$.

2. Additional Ablations

Marker Poser - Mesh fitting We also investigate the mesh reconstruction errors of our direct and parametric meshes recovered from ground-truth markers in the Human3.6M dataset. We consider the GHUM mesh obtained through energy minimization as the ground-truth. We get an MPVPE of 26mm for the direct mesh (*i.e.* $\mathbf{V}_d$) and an MPVPE of 30.7mm for the parametric mesh (*i.e.* $\mathbf{V}_p$). These numbers are equal or better than our reported training errors, as the training was done with noise injected on sampled markers. Please see accompanying videos for examples of our reconstructions and also an example in fig. 1, bottom. We show, from left to right: the image with superimposed 3d marker projections; the ground-truth reconstruction from optimization; the direct mesh reconstruction $\mathbf{V}_d$; the parametric mesh reconstruction $\mathbf{V}_p$.

Marker Poser - Noise during training We show in fig. 1, top, the results of reconstructions for the marker poser if no noise is injected during training. Although the training errors are very low on samples drawn from the normalizing flow prior, the marker poser fails on real data which contains noise. This behaviour also applies when trying to regress marker configurations from real images, as in our proposed THUNDR architecture.

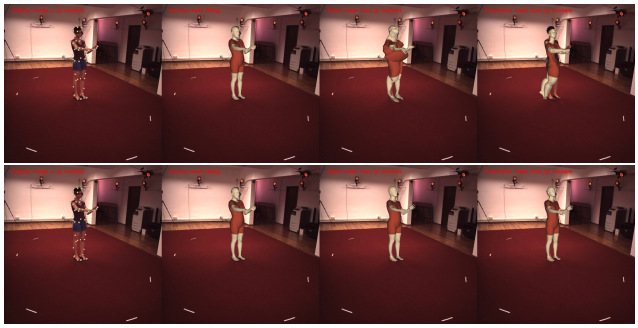

Figure 1: *(Top)* The reconstructions from ground-truth markers if the marker poser is trained **without noise**. From left to right: the image with superimposed 3d marker projections; the ground-truth reconstruction from optimization; direct mesh reconstruction $\mathbf{V}_d$; parametric mesh reconstruction $\mathbf{V}_p$. Notice the failure in reconstruction of our Marker Poser. Even ground-truth markers have noise either due to slightly incorrect manual placement, or because their position slightly changes as the person moves (especially if attached to loose clothing). The Marker Poser needs to be robust to noise. *(Bottom)* Reconstructions when we inject noise during the training of the Marker Poser. Notice much better reconstruction quality.

Figure 2: Comparison between THUNDR and SPIN[1]. From left to right we show the original image, overlaid THUNDR direct mesh reconstructions $\mathbf{V}_d$, overlaid THUNDR parametric mesh reconstructions $\mathbf{V}_p$ and overlaid SPIN mesh reconstructions. While both methods are capable of reconstructing complex poses, notice that THUNDR aligns better the reconstructed meshes to the image evidence and recovers better 3d poses (see e.g. tables 4 and 5 in the paper). For example: i) in the first row, notice the difference in aligning the right leg/foot; ii) in the third row, SPIN misses the correct global rotation and the position of the legs; iii) in the fourth row, the right arm of the tennis player is not correctly reconstructed by SPIN; iv) in the last row, the right leg/foot of the person is poorly reconstructed by SPIN. In the second row, both methods fail in reconstructing the right elbow of the baseball player. The elbow joint is oriented to the back of the player. See our accompanying video for more results.

References

[1] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 2252–2261, 2019. 1, 2