




SkelFormer: Markerless 3D Pose and Shape Estimation using Skeletal Transformers

Vandad Davoodnia^{1,2} , Saeed Ghorbani² , Alexandre Messier², and Ali Etemad¹ 

¹ Queen’s University, Canada

² Ubisoft LaForge, Canada

{vandad.davoodnia, ali.etemad}@queensu.ca

{saeed.ghorbani, alexandre.messier}@ubisoft.com

1 Data Augmentation

We apply the following data corruption techniques when training our skeletal transformer to increase its robustness to different noise sources that are often seen in 3D keypoints estimation algorithms:

Masking. We apply random masking on each joint with a 20% chance to model partially occluded inputs, simulating scenarios where joints are not visible or considered outliers. For this task, we exploit the inner mechanics of transformers by masking the attention of the occluded joints in the encoder and between the encoder and decoders.

Rotation. We randomly rotate the 3D keypoints around the mid-hip point to simulate various orientations of the human body. Specifically, we rotate the points by $\pm 180^\circ$ along the vertical axis to enhance the model’s ability to predict the root rotation accurately. Additionally, we occasionally rotate the body by 90° to simulate lying and sleeping postures, which are often considered challenging in markerless motion capture applications.

Noise Addition. To model the faulty predictions in the 3D keypoint estimation module, we add random Gaussian noise with a standard deviation proportional to 5% of the joint annotation confidences provided for COCO WholeBody [1].

Left-Right Mirroring. We increase the robustness of our model to mediolateral flips by mirroring the keypoints along the YZ plane by 50% chance. Since the SMPL model is not symmetrical, we apply the mirroring directly on the inputs and the output of the model before the SMPL Forward Kinematics (FK) layer.

Shape Augmentation. Although the AMASS [2] training set has a large variety of common and exotic poses, it lacks body shape variety and includes only 300 different body shapes. To mitigate this issue, we randomly augment the body shape parameters before keypoint corruption by an additive Gaussian noise with a standard deviation equal to the standard deviation of all available body shapes.

Outliers. A common noise source in 3D triangulation pipelines is heavy 2D keypoint shifts caused by faulty detection in one or more views. Therefore, a straightforward approach is to consider such heavy shifts as outliers and mask those inputs in the network. However, the outlier detection algorithms might fail

to detect such instances. To increase the robustness of our model to such noises, we apply a large additive Gaussian noise with a standard deviation of 1 meter to each keypoint with a chance of 1% during training to increase the robustness of our model to outliers.

2 Additional Qualitative Assessment

We present a qualitative assessment of our model to Out-of-Distribution (OoD) data to highlight our model’s robustness and generalization. In Fig. 1, we showcase the visual fidelity of SkelFormer by presenting more images on public datasets alongside a proprietary dataset that uses 6 GoPro cameras. These videos include a running sequence in a large volume and a dying and sitting sequence in a small volume.

Figure 2 visualizes SkelFormer’s fitting capabilities in noisy and occlusion experiments on the sequences with its highest error from AMASS [2] testing set. Interestingly, in very noisy circumstances, the model tries to predict a plausible pose while adhering to the input as much as possible. It also generates viable and relaxed poses in the end-point experiments. Consistent with our results, we see almost no changes to the prediction in the presence of occlusions. Please refer to our video demo for more animations containing noise and occlusion experiments and a comparison to VPoser-t.

3 Supplementary Video

We provide a supplementary video describing our solution with visual examples. Our video contains several clips on Human3.6m [3], RICH [4], and our proprietary dataset. We also compare our work with the pseudo ground truth [5] and show better image alignment on the Human3.6m dataset.

References

1. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 740–755.
2. N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, “Amass: Archive of motion capture as surface shapes,” in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 5442–5451.
3. C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, “Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.
4. C.-H. P. Huang, H. Yi, M. Höschle, M. Safroshkin, T. Alexiadis, S. Polikovskiy, D. Scharstein, and M. J. Black, “Capturing and inferring dense full-body human-scene contact,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13 274–13 285.

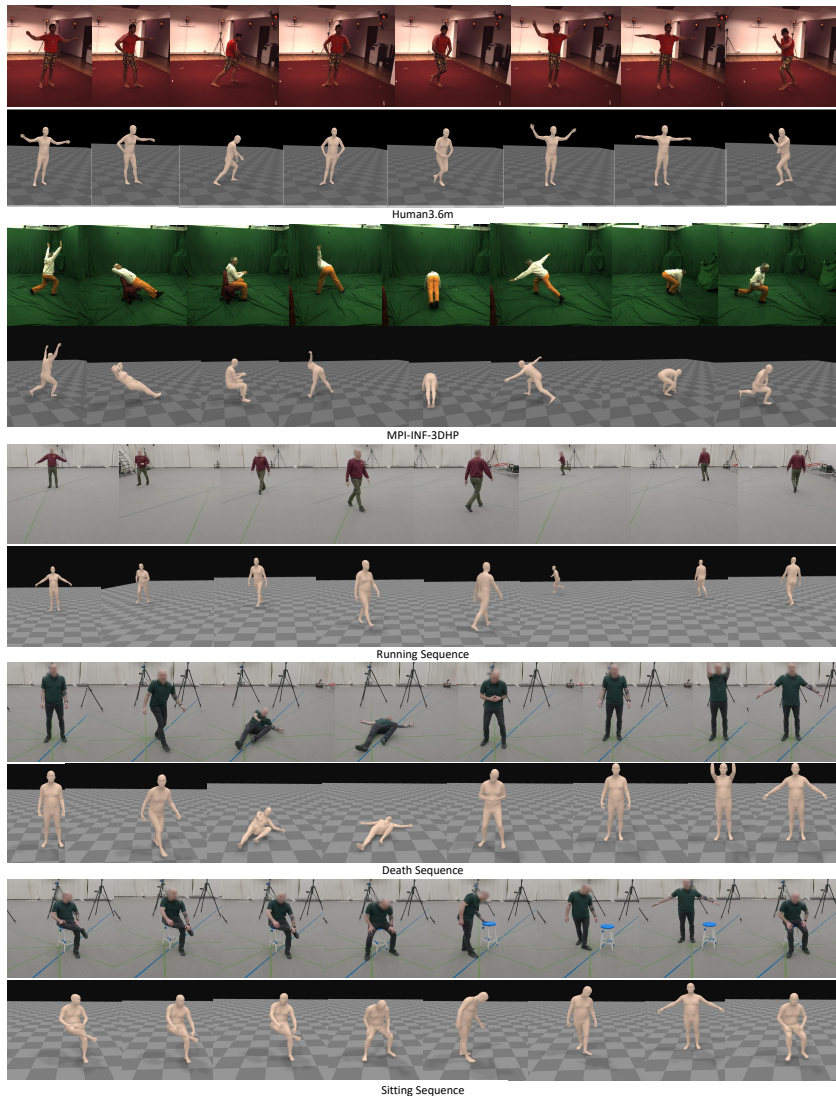


Fig. 1: Sample results on Human3.6m, MPI-INF-3DHP, and our collected videos are presented.

5. G. Moon, H. Choi, and K. M. Lee, "Neuralannot: Neural annotator for 3d human mesh training sets," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 2299–2307.



Fig. 2: We demonstrate the predictions of SkelFormer on its worst-performing sequences during the occlusion and noise experiments. The red points represent the model's inputs.