Garment4D: Garment Reconstruction from Point Cloud Sequences — *Supplementary Material*

Fangzhou Hong[1], **Liang Pan**[1], **Zhongang Cai**[1,2,3], **Ziwei Liu**[1] ✉
[1]S-Lab, Nanyang Technological University [2]SenseTime Research [3]Shanghai AI Laboratory
{fangzhou001, liang.pan, ziwei.liu}@ntu.edu.sg caizhongang@sensetime.com

In this supplementary material, we provide the following sections along with a demo video for a better understanding of the main paper. The Implementation details are elaborated in Section 1. More qualitative and quantitative results are further compared and analyzed in Section 2 and the attached demo video.

1 Implementation Details

Network Details. The Hierarchical Garment Feature Pooling uses query balls with radii of $r = \{0.1m, 0.2m, 0.4m\}$ and sample numbers of $n = \{32, 16, 8\}$. Because the point number of point clouds for sampling decreases, the number of sampling points decreases. The Hierarchical Body Surface Encoder uses query balls with radii of $r = \{0.1m, 0.2m, 0.4m\}$ and sample numbers of $n = \{8, 16, 32\}$. With larger sampling radius, each vertex of the proposal garment mesh gains wider perception field. The encoding dimension of each module of the Proposal-Guided Hierarchical Feature Network is set to 32. Therefore, for each vertex of the proposal garment mesh, the dimension of the features input into the Iterative GCN is $32 \times 6 + 3 = 195$. The additional three dimension is set to the coordinates of the vertex itself. The Iterative GCN is applied for 3 times. Inside each iteration, 4 layers of GCN layers is stacked to predict displacements. The hidden dimension of GCN layers is set to 128. The output feature dimension of the Temporal Transformer is also set to 128. The sequence length T is set to 10. At test time, a sliding window with step size of 3 is used to ensemble the reconstructed meshes. The same operation are used on both the adapted MGN [1] and our Garment4D.

Training Details. The training process consists of two stages. We train the canonical garment estimation first. Then the posed garment reconstruction is trained with the parameters of the first part fixed. The learning rates of two stages of the network are set to 1×10^{-3}. The first stage is trained with 2 NVIDIA V100 GPUs (internal cluster) and the batch size of $20T$ frames. The second stage is trained with 2 NVIDIA V100 GPUs and the batch size of $4T$ frames. We train each step up to 100 epochs. The training time is approximately 24 hours.

2 More Qualitative and Quantitative Results

Quanlitative Results on CAPE. A visualization of the reconstructed T-shirts and trousers on CAPE [2] is shown in Fig. 1. Considering the fact that we directly inference the network on CAPE that is pretrained on Cloth3D [3], the reasonable reconstruction results shows the generalizability of our Garment4D.

MGN *vs.* Ours. We provide several qualitative and quantitative comparisons between the adapted MGN [1] and our Garment4D. As shown in Fig. 2, 3, 4, 5, we plot the per-vertex L2 error of each frame in one sequence. In these sequences, our Garment4D outperforms the adapted MGN by large gaps in most frames. Several representative frames are visualized for comparison. Due to the limitation of the SMPL+D model, the adapted MGN inevitably produces ripped and jagged meshes

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

Figure 1: Visualization of reconstruction results on CAPE.

when the legs are stretched to the sides. While our Garment4D avoids generating low-quality meshes by using interpolated LBS. In addition, our Garment4D is able to capture better garment dynamics with the help of the Proposal-Guided Hierarchical Feature Network.

Demo Video. To further show the advantages of our method, we provide a demo video in the attached materials. More qualitative results and comparisons of different garment types and body movements are available in the video.

References

[1] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5420–5430, 2019. 1

[2] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J Black. Learning to dress 3d people in generative clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6469–6478, 2020. 1

[3] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. Cloth3d: Clothed 3d humans. In *European Conference on Computer Vision*, pages 344–359. Springer, 2020. 1

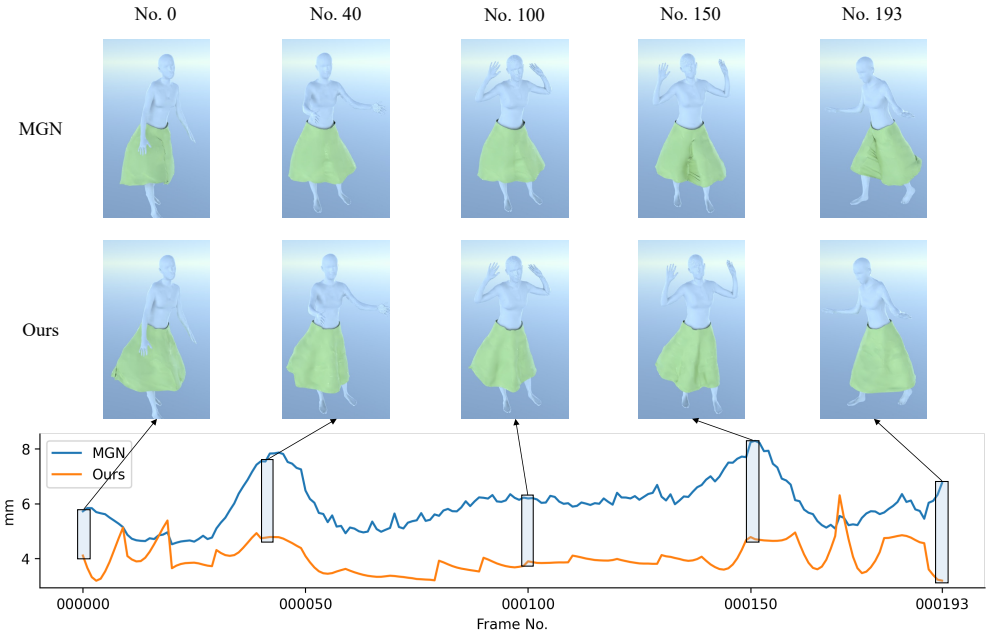

Figure 2: Detailed Comparison on Sequence No. 00550. The x-axis of the line chart represents the frame number. The y-axis is the per-vertex L2 error of the posed-garment reconstruction. The lower the better.

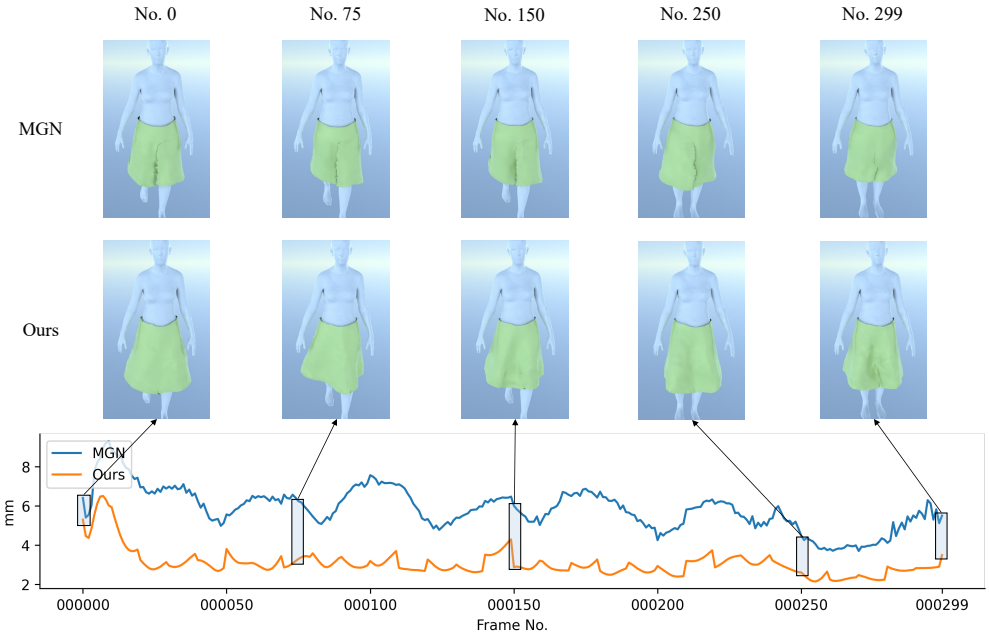

Figure 3: Detailed Comparison on Sequence No. 01000. The x-axis of the line chart represents the frame number. The y-axis is the per-vertex L2 error of the posed-garment reconstruction. The lower the better.

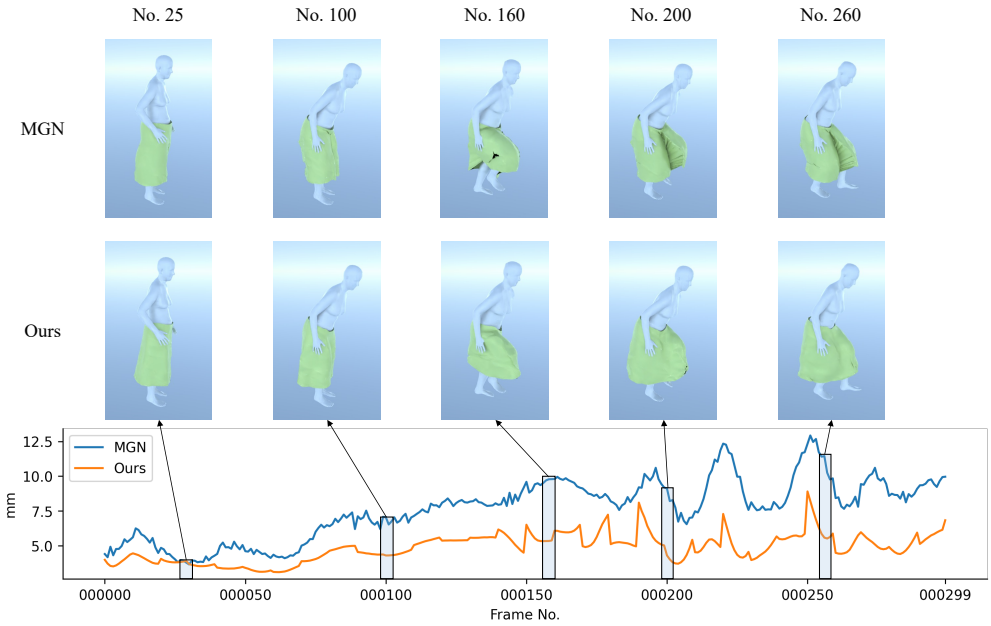

Figure 4: Detailed Comparison on Sequence No. 01232. The x-axis of the line chart represents the frame number. The y-axis is the per-vertex L2 error of the posed-garment reconstruction. The lower the better.

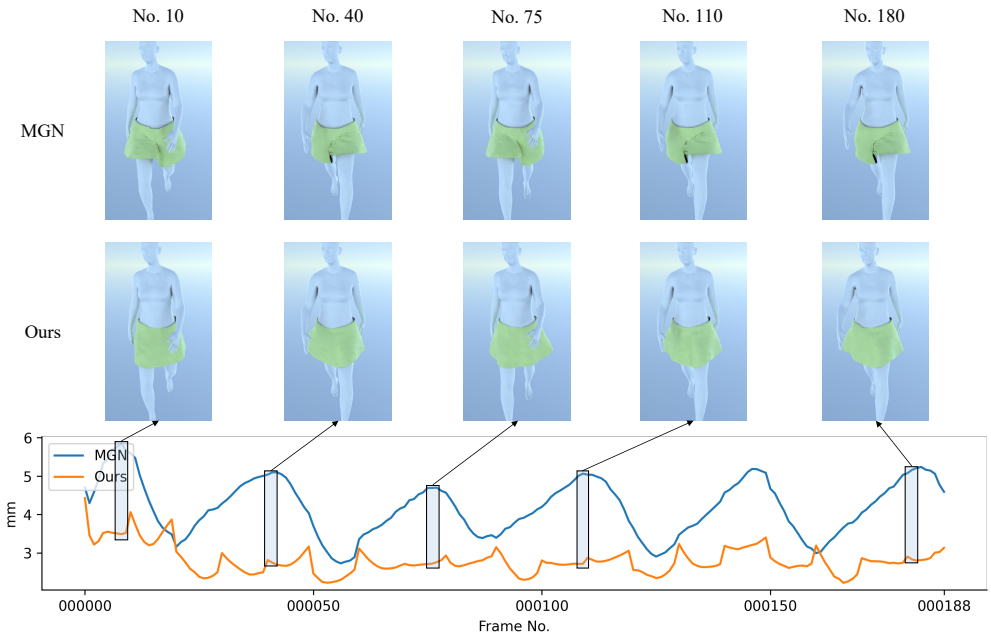

Figure 5: Detailed Comparison on Sequence No. 04176. The x-axis of the line chart represents the frame number. The y-axis is the per-vertex L2 error of the posed-garment reconstruction. The lower the better.