Human Body Shape Completion with Implicit Shape and Flow Learning Supplementary Materials

Boyao Zhou Di Meng Jean-Sébastien Franco Edmond Boyer
Inria, Université Grenoble Alpes, CNRS, Grenoble INP, LJK, Grenoble France
{boyao.zhou, di.meng, jean-sebastien.franco, edmond.boyer}@inria.fr

This supplementary materials provides additional information on the method and its performances. In particular, more insights on the respective contributions of the flow, attention mechanism and hierarchical learning are given. In addition, supplementary results on another dataset, THUman3.0 [6], are shown.

1. Implementation Details

Depth images are rendered in resolution 256^2 with a fixed monocular camera system [5]. Our network is trained on 510 pairs of depth images with batch size 2 for 496 epochs. During the inference, query points are uniformly distributed in a regular grid of resolution 380^3. The final reconstruction meshes are extracted with the marching cubes [2, 3].

2. Flow Contribution

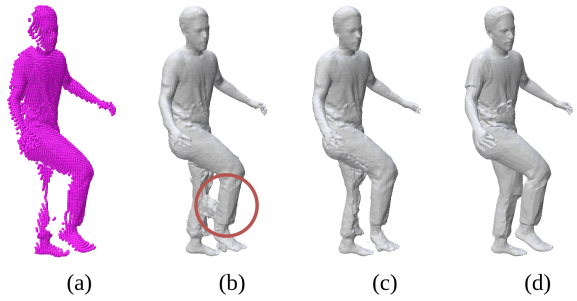

Figure 1. From left to right: (a) Input depth map; (b) Shape completion without the flow; (c) Full model; (d) Ground truth.

We trained one model without the flow: all cross-frame temporal paths are removed in the training and in the inference. It contains therefore the single frame 2D feature map $\mathcal{F}^f$, the coarse-dense feature map $\mathcal{Q}^c$, the sparse-fine feature Q^f and the self-attention $\mathcal{A}^{self}$. The model was trained with the same data as full model and evaluated with data used in Tab. 4 in the main paper. Table 1 provides quantitative results while Figure 1 shows one example. Despite showing good performances with the Chamfer distance (Tab. 1) this model is more prone to large volumetric errors, as in Fig. 1 in occluded parts, which impact less a surface based metric.

Method	IoU(%) ↑	Chamfer-L1 ↓($\times 10^{-2}$)
w/o flow	85.5	**1.044**
Full model	**86.1**	1.047

Table 1. Ablation studies for the flow. Chamfer $\times 10^{-2}$.

3. Attention Mechanism

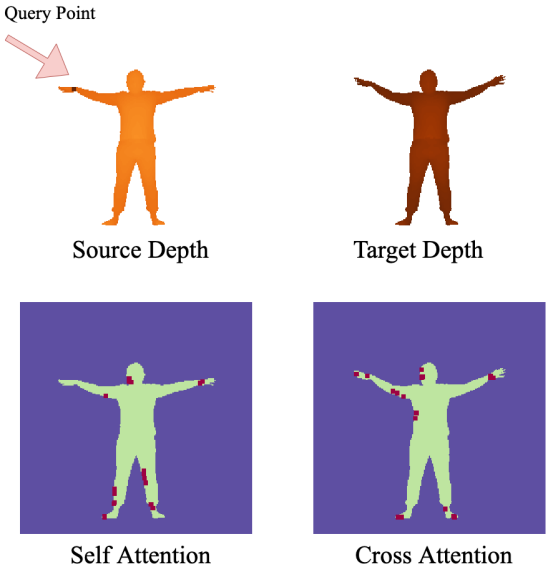

Figure 2. Source/target depth images and the 20 highest self/cross attention weights.

Figure 2 illustrates the attention module weights. Given a query point on the right hand in the source frame, the bottom images show the 20 pixels with the highest values of

q_{att} in equation 9 of main paper in both the source and target frames. We note that the self-attention mechanism focuses on the left hand for symmetry, while the cross-attention mechanism focuses on points on the extremities for the flow estimation, this in particular with the corresponding points on the right hand.

4. Hierarchical Learning

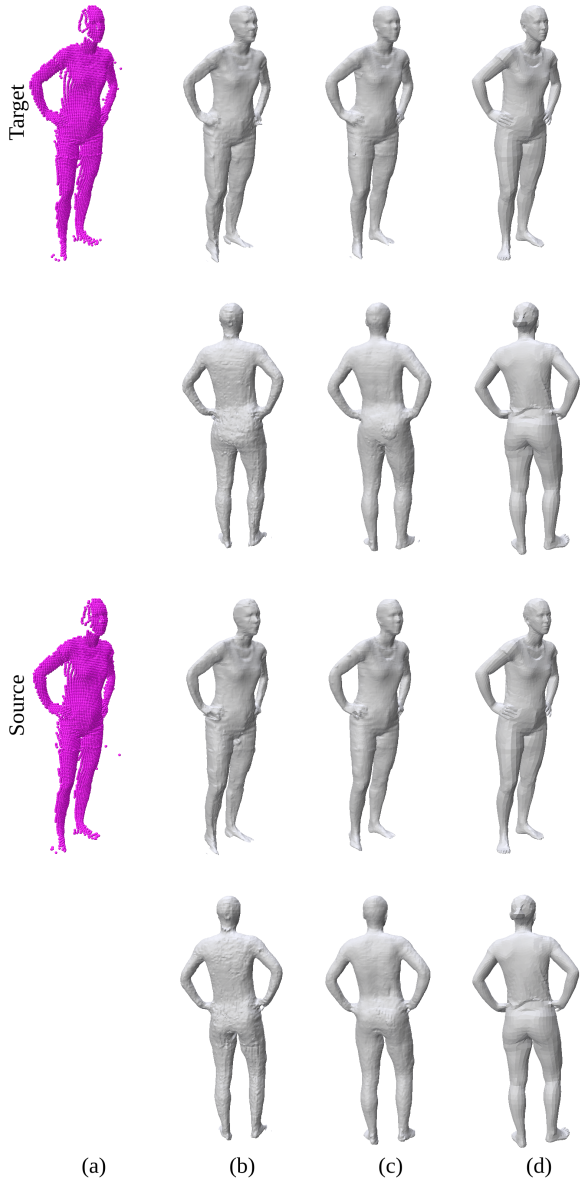

Figure 3. From left to right: (a) input depth maps; (b) Completion without hierarchical module; (c) Full model; (d) Ground truth.

Our approach adopts a coarse-to-fine hierarchical learning strategy to efficiently embed high-frequency 3D features. In order to illustrate its contribution, we trained a model without any hierarchical learning but containing all the other modules, including the attention mechanism, within which only the fine feature $\mathcal{F}^f_{3D}$ is fed into `SConv3D` by $Q^f = \texttt{SConv3D}(\mathcal{F}^f_{3D}, vox^f)$ (see equation 8 in main paper). Figure 3 shows the comparison between the full approach and without the hierarchical module. We observe that without coarse-level information, this model suffers from high-frequency noises in the completions (see Fig. 3(b) and (c)).

5. Other Data

We show here results on more challenging models, from the THUman3.0 [6] dataset, with more complex cloth styles. The model was trained as in the main paper (see Sec. 4.2) without any fine tuning. Unlike CAPE [4] and DFAUST [1], THUman3.0 provides identities in random poses instead of consecutive sequential data. Moreover, such data is generated directly from raw scan without fitting with any template or parametric model. The rendered depth is in general noisy, see Fig. 4(a), and we can not evaluate the flow with this dataset. However, even without any temporal consistency our method can provide high quality results, as illustrated in Fig. 4.

References

[1] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J Black. Dynamic faust: Registering human bodies in motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[2] Thomas Lewiner, Hélio Lopes, Antônio Wilson Vieira, and Geovan Tavares. Efficient implementation of marching cubes' cases with topological guarantees. *Journal of Graphics Tools*, 8(2):1–15, 2003. 1

[3] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM Siggraph Computer Graphics*, 21(4):163–169, 1987. 1

[4] 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 Conference on Computer Vision and Pattern Recognition*, 2020. 2

[5] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 1

[6] Zhaoqi Su, Tao Yu, Yangang Wang, and Yebin Liu. DeepCloth: Neural garment representation for shape and style editing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 2, 3

[7] Xingguang Yan, Liqiang Lin, Niloy J Mitra, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Shapeformer: Transformer-based shape completion via sparse representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 3

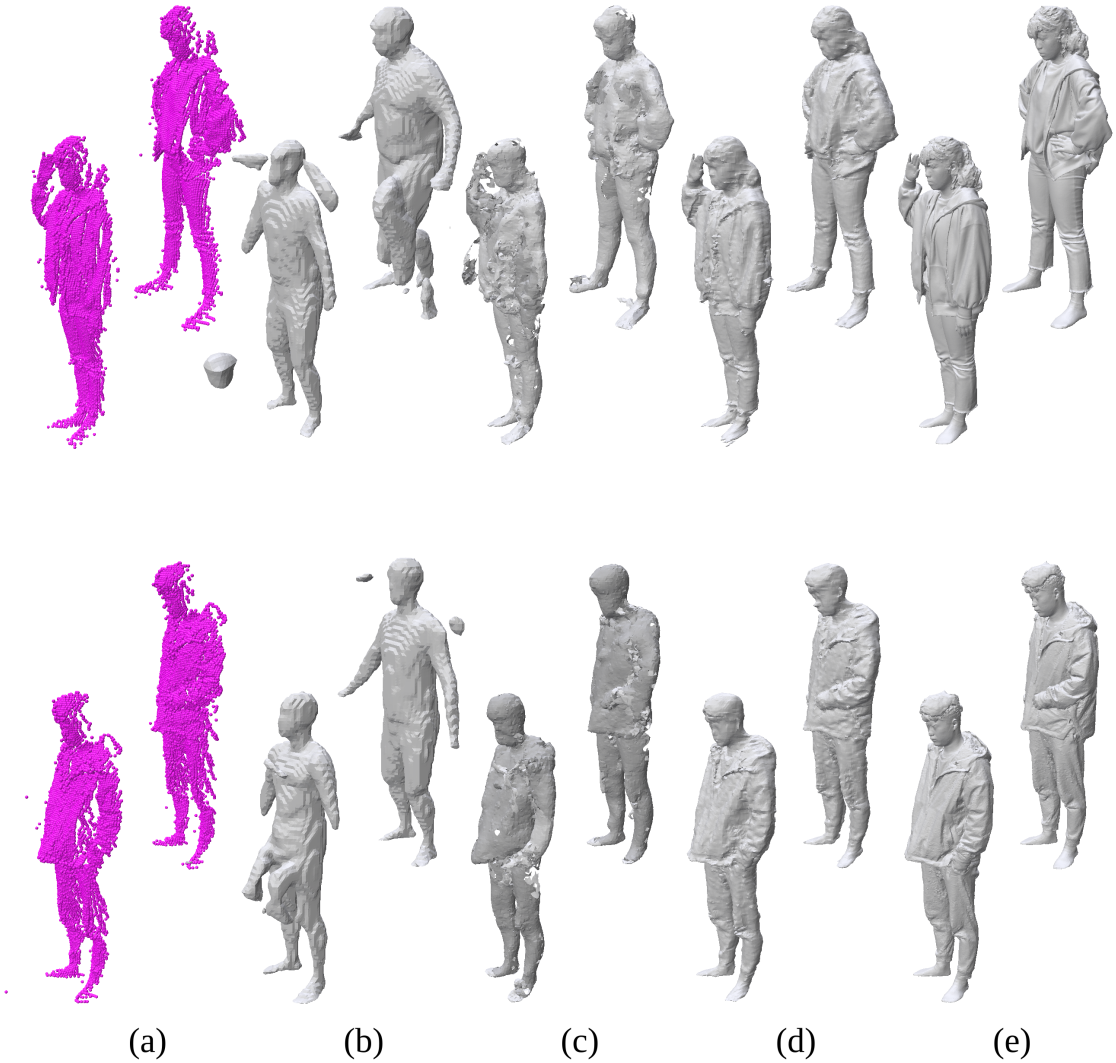

Figure 4. THUman3.0 [6] models. From left to right: (a) Input depth maps; (b) Reconstructions from ShapeFormer [7]; (c) Reconstructions from SeedFormer [8]; (d) our approach; (e) Ground truth.

[8] Haoran Zhou, Yun Cao, Wenqing Chu, Junwei Zhu, Tong Lu, Ying Tai, and Chengjie Wang. Seedformer: Patch seeds based point cloud completion with upsample transformer. In *Proceedings of the European Conference on Computer Vision*, 2022. 3