Supplementary Document for Function4D: Real-time Human Volumetric Capture from Very Sparse Consumer RGBD Sensors

A. Overview

This supplementary document provides more details about the comparisons (Sec.B) and evaluations (Sec.C) in the paper. Please also refer to the supplementary video for more clear demonstrations.

B. Comparison Details

To better evaluate the effectiveness of our method, we compare with both the state-of-the-art volumetric fusion methods (Motion2Fusion and UnstructuredFusion) and the state-of-the-art learning-based 3D human reconstruction methods (PIFu, IFNet and IPNet).

B.1. Comparison with Volumetric Fusion Methods

We first compare with the state-of-the-art volumetric fusion method of Motion2Fusion [5] qualitatively. The implementation used to generate the results of Motion2Fusion [5] is an improved version made by the original author. Specifically, the final SDF values in Motion2Fusion are solved by a Poisson equation with the normal constraints instead of weighted averaging. This causes blob artifacts for depth off the surface. Apart from the artifacts, we notice that Motion2Fusion has limited capacity to track the challenging non-rigid motions and topological changes using very sparse consumer sensors, thus producing severely erroneous surfaces when tracking fails.

To further demonstrate the effectiveness of our method, we also compare with UnstructuredFusion [11] qualitatively in the supplementary video, which also uses sparse consumer depth sensors for volumetric performance capture. Note that UnstructuredFusion does not support multi-person reconstruction, so we compare with UnstructuredFusion using a single-person sequence. Based on a parametric model of human body (SMPL [8]), they eliminate the requirement for explicit multi-camera calibration and generate plausible dynamic 3D reconstruction results for tight-clothed humans. However, the incorporation of the parametric model restrict UnstructuredFusion from handling topological changes and more general clothes. More importantly, when severe tracking failure happens, it is hard for UnstructuredFusion to recover from it efficiently, which finally leads to unstable reconstruction results under fast motions. Please refer to the supplementary video for the qualitative comparison with Motion2Fusion and UnstructuredFusion.

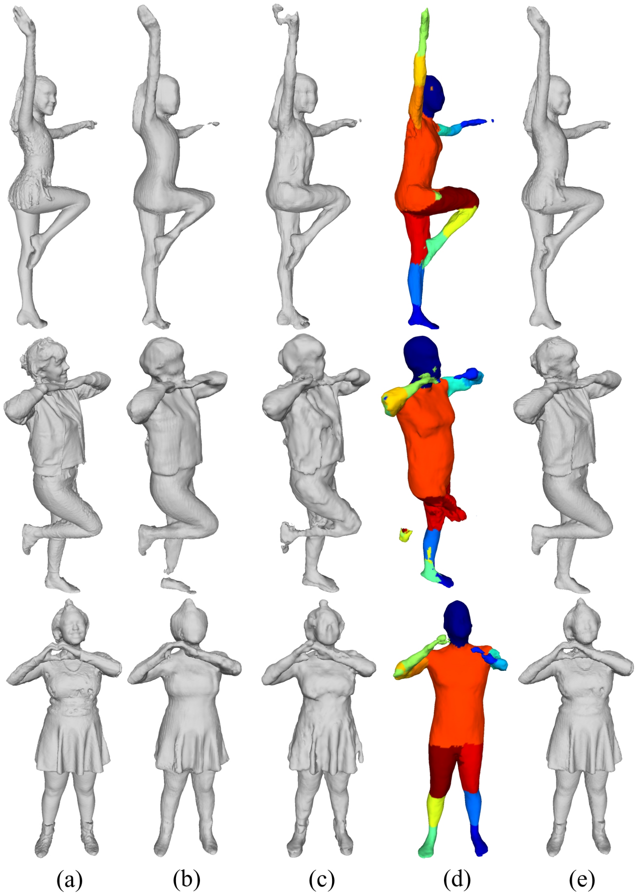

Figure A: Qualitative comparison with Multi-view PIFu and IPNet. From (a) to (e) are the ground truth model, the results of Multi-view PIFu, the outter layer and inner reconstruction results of IPNet, and our results, respectively.

B.2. Comparison with Learning-based Methods

We compare with the state-of-the-art learning-based 3D human reconstruction methods (PIFu [9], IFNet [2] and IPNet [1]) both quantitatively and qualitatively to demonstrate the effectiveness of our networks. All of the 3 methods are specifically designed for learning-based 3D human reconstruction. Specifically, PIFu uses multi-view RGB images as input while IFNet and IPNet uses point cloud as input. Note that IPNet further improves IFNet by incorporating the parametric model for constraining a multi-layer surface reconstruction process. The quantitative comparison can be found in the Table.1 of the main paper, and the qualitative comparison is shown in Fig. A. We can see from the comparisons that our method outperforms all of the 3 methods mentioned above both qualitatively and quantitatively. To conclude, our method is not only more accurate (for multi-view RGB-D input) and robust (generalize well to different large poses and cloth styles), but also much more efficient (orders of magnitude faster than existing methods, real-time run-time efficiency) than existing methods. Next, we will introduce the metrics and implementations corresponding to this comparison in detail.

Metrics For quantitative evaluation on geometry reconstruction, we firstly randomly sample 100K points from the reconstructed surface and the ground truth surface, and then construct a KD-tree to estimate the corresponding distances. We calculate three different metrics: average point-to-surface Euclidean distance(P2S) from the sampling vertices on the reconstructed surface to the ground truth measuring the accuracy of the surface (lower is better), the Chamfer-L_1 distance defined as the mean of an accuracy and a completeness metric (lower is better), and a normal consistency score as the mean absolute dot product of the normals in one mesh and that at the corresponding nearest neighbors in the other mesh, measuring the accuracy of the shape normal (higher is better).

Implementation Details Tab.1 shows the quantitative comparison on geometry reconstruction between the proposed *GeoNet*, Multi-view PIFu[9] (with RGB images as input) and IPNet[1] (with voxelized point clouds as input). The two baseline methods are also based on implicit functions and multi-view information to reconstruct complete meshes. We retrain Multi-view PIFu (using perspective projection model) and IPNet using our training dataset and perform evaluation on a testing dataset which contains 116 high-quality scans with various poses, clothes and human-object interactions.

For point sampling during the training stage, we use the gradient-based point sampling strategy for recovering more surface details. Specifically, for each scan, we first calculate its Discrete Gaussian Curvature using the method in [4] with radius 0.002. And in each training pass, we sample 5000 points which contains 2500 high-curvature-points and 2500 free-sampled points (in which the curvature threshold for filtering high-curvature nodes is set as 0.004).

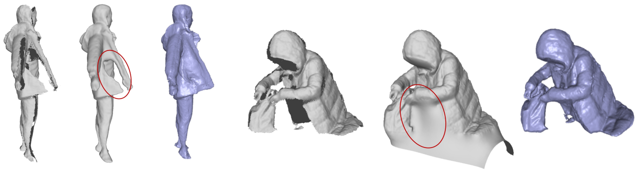

Figure B: For each subject, we demonstrate the results of dynamic sliding fusion(DSF)(left), the results of the DSF with Screened-Possion-Surface-Reconstruction(SPSR) [6](middle), and the results produced by our method(right).

For the comparison with multi-view PIFu, we also use HRNetV2-W18-Small-v2[10] (with the same feature map resolution 64 × 64) as the backbone. Different from our implementation (in which the output channel number is 32 for the *GeoNet*), we use 256 channel feature map for multi-view PIFU to keep consistent with [9]. In Table 1, we can see that even with much less feature dimension, our *GeoNet* still achieves much better results. It fully demonstrates that our model can learn more depth-relevant and fine-grained information with the help of the truncated PSDF values in the encoding stage.

The IPNet is derived from IFNet[3] with point clouds in 3D volume as input and the same multi-scale 3D CNN as encoder to extract volume features and implicit function to learn the SDF volume. For training IPNet, we strictly follow [1] to make training dataset containing SMPL registered scans with our dataset. From the results in Table 1, we can find that IPNet produces better results than multi-view PIFU benefiting from the strong capacity of 3D CNN and multi-view depth inputs. However, due to the heavy dependency on exact SMPL inner bodies as the ground truth (which is very hard to acquire especially under large poses and human-object interactions), the reconstruction accuracy of the IPNet is still lower than ours.

To conclude, the lack of depth information deteriorates the reconstruction accuracy of Multi-view PIFu. Moreover, even with multi-view depth images as input, the limited feature resolution and heavy dependency on accurate SMPL fitting restricts the IPNet from generating highly accurate results. Finally, by explicitly encoding depth observations using truncated PSDF values, the proposed *GeoNet* can not only achieve accurate reconstruction results, but also orders of magnitude faster than current methods.

B.3. Comparison with Raw Fusion Results and Classical Surface Completion Methods

Without implicit surface reconstruction, the raw fusion results (output of the dynamic sliding fusion step) will not be complete (especially under very sparse views) due to self-occlusions and missing observations as shown in Fig.2

of the main paper (the red-colored model) and Figure. B at here. Moreover, the traditional surface in-painting methods, like Screened Poisson Surface Reconstruction (SPSR), may *fail on hole filling*(Figure. B(the red circle on the left)) and *hallucinates wrong geometries*(Figure. B(the red circle on the right)) even given the raw fusion results from the DSF.

C. Evaluation Details of the Proposed Method

Evaluation of the Truncated PSDF Feature For quantitative evaluation of the truncated PSDF values, all the networks are trained with the same training settings using the same training dataset (containing 500 scans), and finally evaluated in the same testing dataset containing 116 scans. Tab.2 in the main paper shows that without using the truncated PSDF feature, the depth-only model and RGBD model produce similar results, which indicates that explicit PSDF information is more important to discriminate whether the point is inside or outside of the observed surface.

Evaluation of Dynamic Sliding Fusion To better evaluate the proposed dynamic sliding fusion method, we provide a video comparison in the supplementary video. We show the final reconstruction results of 2 kinds of multi-view depth inputs: (a) using the original captured multi-view depth images as input, and (b) using the re-rendered multi-view depth images from dynamic sliding fusion as input.

Note that without dynamic sliding fusion (setup (a)), the depth observations corresponding to different viewpoints may not be consistent with each other due to sensor noise, missing observations and spatial distortions, and this leads to noisy and incomplete reconstruction results. By fusing multi-view depth images in a sliding window into the TSDF volume non-rigidly (setup (b)), we can guarantee that the re-rendered multi-view depth images not only contain much less noise, but are also consistent across different viewpoints. As a result, by using the dynamic sliding fusion, we can generate more complete and noise-suppressed reconstruction results as shown in the supplementary video.

C.1. Network Training Losses

Regarding to the losses for training the *GeoNet* and the *ColorNet*, we follow Monoport [7] to use the Binary Cross Entropy loss and the L1 loss, respectively. The training of the *GeoNet* takes 12 hours in total (approximate 30 epochs in total), and the training of the *ColorNet* takes 36 hours in total (approximate 30 epochs in total).

References

[1] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision (ECCV)*. Springer, August 2020. 2

[2] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020. 2

[3] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[4] David Cohen-Steiner and Jean-Marie Morvan. Restricted delaunay triangulations and normal cycle. In *ACM SYMPOSIUM ON COMPUTATIONAL GEOMETRY*, 2003. 2

[5] Mingsong Dou, Philip L. Davidson, Sean Ryan Fanello, Sameh Khamis, Adarsh Kowdle, Christoph Rhemann, Vladimir Tankovich, and Shahram Izadi. Motion2fusion: real-time volumetric performance capture. *ACM Trans. Graph.*, 36(6):246:1–246:16, 2017. 1

[6] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Trans. Graph.*, 32(3), July 2013. 2

[7] Ruilong Li, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. Monocular real-time volumetric performance capture. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 3

[8] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):1–16, 2015. 1

[9] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2304–2314, 2019. 2

[10] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 2

[11] Lan Xu, Zhuo Su, Lei Han, Tao Yu, Yebin Liu, and Lu Fang. Unstructuredfusion: Realtime 4d geometry and texture reconstruction using commercial RGBD cameras. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(10):2508–2522, 2020. 1