Supplementary Material
Pyramidal Signed Distance Learning for Spatio-Temporal Human Shape Completion

Boyao Zhou[1], Jean-Sébastien Franco[1], Martin deLaGorce[2], and Edmond Boyer[1]

[1] Inria-Univ. Grenoble Alpes-CNRS-Grenoble INP-LJK, France
{boyao.zhou, jean-sebastien.franco, edmond.boyer}@inria.fr
[2] Microsoft. Cambridge, United-Kingdom
{martin.delagorce}@microsoft.com

1 Network Details

As illustrated in Fig. 1, the Spatio-Temporal Feature Encoding consists of 4 down blocks to the global information, and 4 up blocks. The global information goes to the bidirectional GRU [1, 2] with the objective to learn the interframe information. In this architecture, the 3 last up blocks are associated with 1×1 convolution to yield the features $\mathcal{Z}^l, l = \{0, 1, 2\}$ at different levels. Fig. 2(a) shows how the features $\mathcal{Z}^l$ are fed into 3 3D convolution blocks. The weights of these 3D convolutions are shared for $l = 0$ and 1. Fig. 2(b) illustrates the Implicit Surface Decoding. The layer dimensions of the MLP are 64, 64 and 1 respectively. The final output is activated with *tanh*.

2 Training and Inference Details

We render the depth images in resolution 512^2 and train 336 4-frame subsequences from scratch for 796 epochs. We use Adam [3] to optimize the network with learning rate 1×10^{-4}. During the inference, query points are chosen in a grid of resolution 256^3. We bi-linearly interpolate the feature $\mathcal{F}^0$. It is then combined with depth information and fed into a MLP to predict the SDF value. It takes 1.04 second per frame, on average, for all SDF values inference. The mesh surface is finally extracted using the marching cubes algorithm [6, 4].

Data	CAPE		DFAUST	
Method	IoU ↑	Chamfer-L1 ↓	IoU ↑	Chamfer-L1 ↓
4-frame pyramidal	**0.839**	**0.161**	**0.898**	**0.103**
6-frame pyramidal	0.817	0.176	0.880	0.111

Table 1. Impact of the number of frames. Spatial completion with IoU and Chamfer-L1 distances ($\times 10^{-1}$) in the real 3D space.

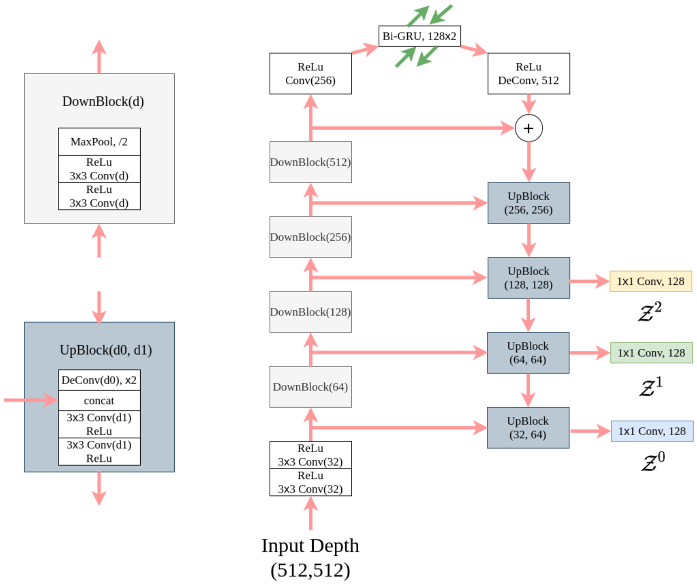

Fig. 1. Spatio-Temporal Feature Encoding.

duration	100ms	200ms	300ms	400ms	500ms
STIF [10]	0.857/0.107	0.852/0.108	0.855/0.109	0.852/0.111	0.850/0.111
ours	**0.900/0.095**	**0.900/0.095**	**0.900/0.096**	**0.899/0.097**	**0.898/0.097**

Table 2. Impact of the sequence duration (ms) in the test. Metrics are IoU/Chamfer-L1 distances ($\times 10^{-1}$).

3 Local Pattern Reasoning

We render depth images from the raw scans, *e.g.* Fig. 3(a), and preserve therefore measurement noise. The signed distances are pre-computed from watertight meshes, as obtained from the full scans and SMPL [5], see Fig. 3(c). Although the SMPL fitting can be locally imperfect, as a result of the global fitting, our network learns locally and naturally tends to reproduce the input depth information as optimal predictions on average over all parts in all the training models. This can be observed in Fig. 3 where the network better predicts the foot than the SMPL fitting on the full scan.

4 Impact of Number of Frames

We consider 4 consecutive frames as it experimentally appears to be a good trade-off between the quality and the computational cost. Absent any explicit or estimated long term correspondence information of surface points, adding more frames does not necessarily contribute, increases the GPU load during training

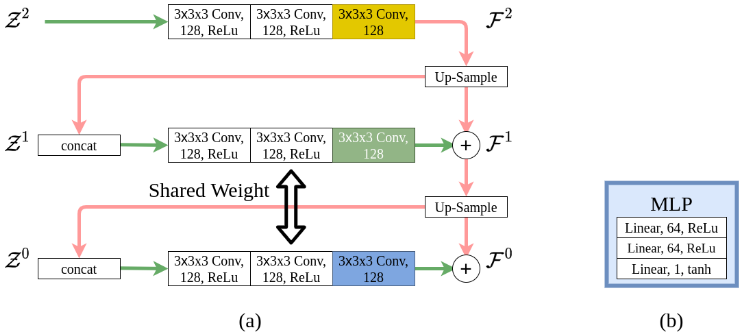

Fig. 2. Detailed Architecture of (a) Pyramidal Feature Decoding and (b) Implicit Surface Decoding.

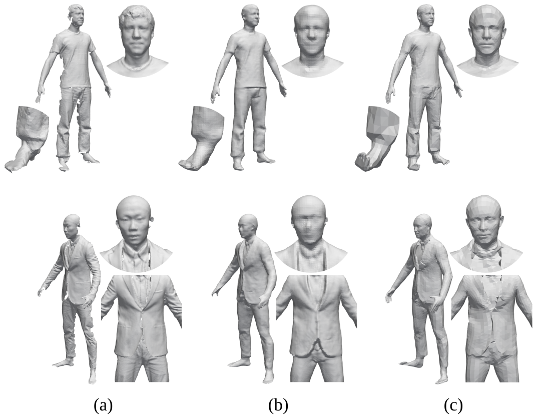

Fig. 3. Local pattern reasoning. From left to right, (a) partial scan, (b) completion with our method and (c) watertight pseudo ground truth with SMPL [5] on full scans.

and even leads to loss of precision. To illustrate this, we train the 6-frame-input model in which there are 1 supervision for level 2, 3 supervisions for level 1 and 6 supervisions for level 0. The result in Tab. 1 is slightly degraded. Here we use the same test data as in Tab. 1 of the paper. We believe that the network would require a more complex additional stage of explicit correspondence estimation to propagate details from more distant frames, which in our mind is a different contribution altogether worthy of exploration in future work. Conversely, our method offers an interesting trade-off already improving the quality of geometric estimation, that doesn't require dealing with correspondences.

5 Impact of Sequence Duration

During the training, we use both the short and long term intervals, resulting in sequences of 200ms and 500ms respectively. To give more insights on the impact of the sequence duration, we provide results with the same trained network but with different frame interval values for testing on the STIF [10] benchmark, in Tab. 2. We give also the comparison with STIF that shows that our architecture improves for all sequence durations and is more robust for long-term sequence completion.

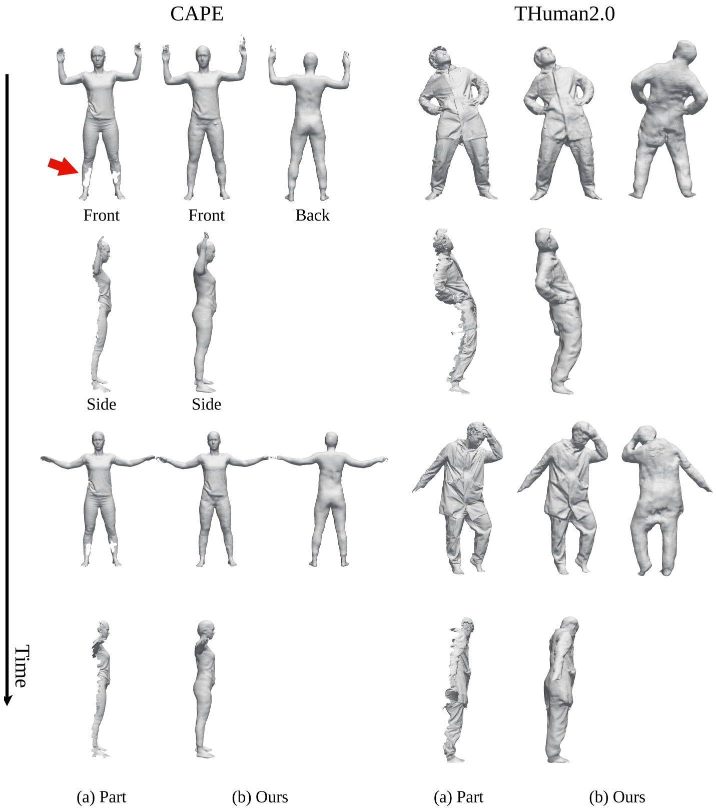

Fig. 4. Some challenging cases. We show (a) the partial real scan data and (b) our completion. For each sub-sequence, we show 2 consecutive frames for CAPE and 2 random poses for THuman2.0.

6 Challenging Cases

In Fig. 4, we show the completion for the scan of CAPE [7] with real missing holes during acquisition, and for THuman2.0 [9] data. We focus on front-view completion and consider missing data in the real scan (red arrow in Fig. 4). We note that THuman2.0 is not a 4D motion dataset and we use the pre-trained model presented in the paper without retraining on the new data. Even if the temporal consistency is no longer held and the geometry is challenging in THuman2.0, our proposed method can still give physically plausible results.

7 Limitations

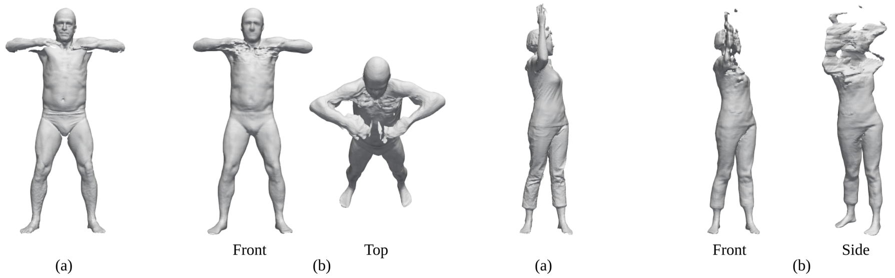

Fig. 5. Failure cases. (a) Partial scan and (b) completion with our method.

It can be observed that some noise may be introduced by our approach on certain datasets which present self-occlusions. This is a common limitation of monocular approaches with 2D convolutional representations [8]. The performance degrades rather gracefully inasmuch as no observations of the occluded surface are available and the problem is ill-posed (Fig. 5(left)). More global humanness constraints could be added in the future to improve over this. Stronger degradations such as the one observed in Fig. 5(right) may occur when strong self-occlusions are compounded with poses that are far from the ones compiled in the training set. Note that the results shown in Table 2 of the paper include the reconstructions of this figure (Fig. 5), and still demonstrate a significant improvement over all state-of-the-art approaches compared to.

References

1. Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259 (2014)
2. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
3. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
4. Lewiner, T., Lopes, H., Vieira, A.W., Tavares, G.: Efficient implementation of marching cubes' cases with topological guarantees. Journal of graphics tools **8**(2), 1–15 (2003)

5. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Transaction on Graphics (TOG) **34**(6), 248:1–248:16 (Oct 2015)
6. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. ACM siggraph computer graphics **21**(4), 163–169 (1987)
7. Ma, Q., Yang, J., Ranjan, A., Pujades, S., Pons-Moll, G., Tang, S., Black, M.J.: Learning to dress 3d people in generative clothing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (Jun 2020)
8. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2304–2314 (2019)
9. Yu, T., Zheng, Z., Guo, K., Liu, P., Dai, Q., Liu, Y.: Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (June 2021)
10. Zhou, B., Franco, J.S., Bogo, F., Boyer, E.: Spatio-temporal human shape completion with implicit function networks. In: Proceedings of the International Conference on 3D Vision (2021)