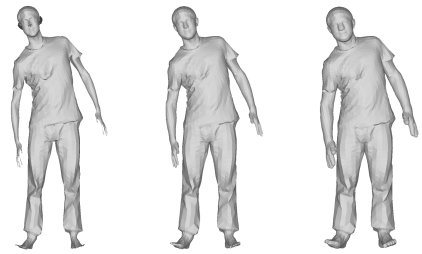
Supplementary Material: Spatio-Temporal Human Shape Completion With Implicit Function Networks

Boyao Zhou*† Jean-Sébastien Franco* Federica Bogo‡ Edmond Boyer *
{boyao.zhou, jean-sebastien.franco, edmond.boyer}@inria.fr febogo@microsoft.com



(a) o = +0.5 (b) o = 0 (c) o = -0.5

Figure 1. Point sampling strategy. From left to right, (a) shrunk surface, (b) origin surface, and (c) expanded surface. o stands for occupancy value.

1. Point Sampling

In the Section 3.4 of the paper, we discuss our point sampling technique with 3 surfaces: original ground truth surface, shrunk surface with negative normal displacement and expanded surface with positive normal displacement. One example of CAPE [2] dataset is shown in Fig. 1. Note that we also sample the points from inner part of shrunk surface and from outer part of expanded surface. In Tab. 3 of the paper, our point sampling strategy improves the reconstruction of our static method with respect to Gaussian sampling. In Fig. 3, we give some qualitative results. It is clear that we are able to retrieve more details with our sampling technique.

2. Network Details

The global network architecture is shown in Fig. 2 of the paper. Here, the details of the three blocks are shown in Fig. 2. In practice, 4 depth images with the dimension of $(1, reso, reso)$ are processed with U-GRU Encoder. Then, 4 feature maps and 1 temporal channel are concatenated along the channel dimension. The constant temporal channel is expanded with $c_t \times t$. So the input to the Feature Interpolation phase is with the dimension of $(4 \times 128 + 1, reso, reso)$. Once the feature map is interpolated at time stamp t, the feature of point p can be queried on the feature map with the horizontal and vertical coordinates. This feature and the depth information, $c_d \times d$, are concatenated, which are sent to the Occupancy Decoder.

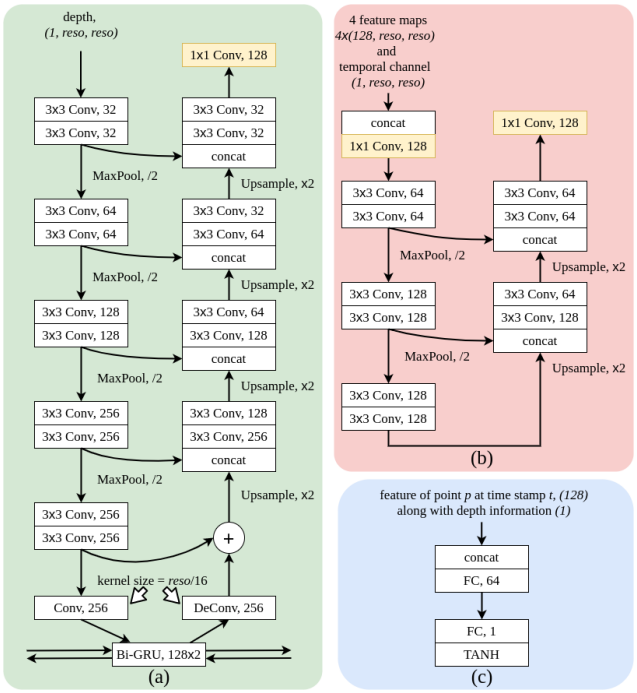


Figure 2. STIF-Nets architecture for the three blocks of paper Fig. 2(same block color): (a) U-GRU Encoder, (b) Temporal Feature Interpolation U-Net and (c) Occupancy Decoder. Conv, FC, reso and concat stand for 2D convolution, fully-connected layer, resolution and concatenation operation. The three yellow blocks are not activated with any activation function while the other Conv layers in (a) and (b) and the the first FC in (c) are activated with RELU. The output is activate with TANH to bound the occupancy prediction in the interval [-1, 1]. Here, we ignore the dimension of batch size.

*Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP (Institute of Engineering Univ. Grenoble Alpes)
†Microsoft Research-Inria Joint Center
‡Microsoft Zürich, Switzerland

3. More Qualitative Results

(a) (b) (c) (d) (e)

Figure 3. Qualitative result for completion task. From left to right, (a) front-view partial scan, the reconstruction (b) our static method with Gaussian sampling, (c) our static method with our sampling, (d) the naive dynamic baseline and (e) our STIF-Nets.

The new qualitative comparison is shown in Fig. 3. We include three remaining subjects, in CAPE [2] and DFAUST [1] dataset, left out of the paper in the interest of space. We also prepare a naive dynamic baseline which drops the GRU and the dimensionality in the simpler Temporal Feature Interpolation is (32, 32, 32, 64, 64, 64, 32, 32, 32, 128). On one hand, the naive dynamic baseline is not able to handle the temporal information without U-GRU Encoder. On the other hand, our STIF-Nets improve significantly the qualitative result over all baselines: higher frequency facial & surface details and input similarity, less oversmoothing, less spurious or discretization-related artifacts including the back surface. Remark that the missing part of head in the last example of Fig. 3 degrades the reconstruction of the two static baselines, but the STIF-Nets could fix it by considering temporal information.

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*, pages 6233–6242, 2017. 2

[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 Conference on Computer Vision and Pattern Recognition*, June 2020. 1, 2