Supplementary for Locally Aware Piecewise Transformation Fields for 3D Human Mesh Registration

Shaofei Wang[1], Andreas Geiger[2,3], Siyu Tang[1]
[1]ETH Zürich [2]Max Planck Institute for Intelligent Systems, Tübingen [3]University of Tübingen

Abstract

In this supplementary material, we first elaborate on the difference between the original NASA [5] and our modified NASA. Next, we present detailed model architectures of PTF-Piecewise and PTF-FC which are introduced in the main paper. Then we present additional quantitative and qualitative results on the CAPE dataset, as well additional qualitative results on the BUFF dataset. Lastly, we discuss limitations of our approach.

1. Difference Between the Original NASA [5] and Our Modified NASA

In the original NASA [5] paper, the bone transformations $\{\mathbf{B}'_b\}$ are defined as:

$$\mathbf{B}'_b = G_b(\boldsymbol{\beta}, \boldsymbol{\theta}) \tag{1}$$

$$G_b(\boldsymbol{\beta}, \boldsymbol{\theta}) = \prod_{b' \in \mathbf{A}(b)} \begin{bmatrix} \mathbf{R}_{b'}(\boldsymbol{\theta}) & \mathbf{t}_{b'} - \mathbf{t}_{\bar{b}'} \\ \mathbf{0} & 1 \end{bmatrix} \tag{2}$$

$$\{\mathbf{t}_b\} = \mathcal{J}(\boldsymbol{\beta}) \tag{3}$$

The $\{\mathbf{B}'_b\}$ are relative to *the origin*, whereas in our formulation we have:

$$\mathbf{B}_b = G_b(\boldsymbol{\beta}, \boldsymbol{\theta}) G_b(\boldsymbol{\beta}, \mathbf{0})^{-1} \tag{4}$$

$$G_b(\boldsymbol{\beta}, \boldsymbol{\theta}) = \prod_{b' \in \mathbf{A}(b)} \begin{bmatrix} \mathbf{R}_{b'}(\boldsymbol{\theta}) & \mathbf{t}_{b'} - \mathbf{t}_{\bar{b}'} \\ \mathbf{0} & 1 \end{bmatrix} \tag{5}$$

$$\{\mathbf{t}_b\} = \mathcal{J}(\boldsymbol{\beta}) \tag{6}$$

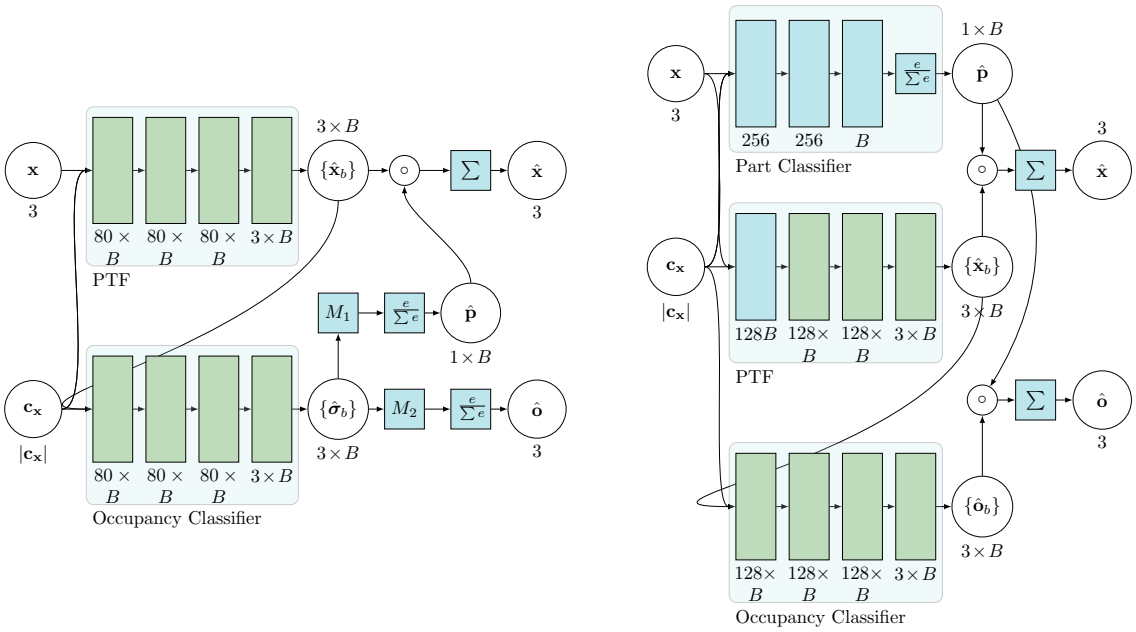
Eq. (4)-(6) are the same as Eq. (2)-(4) in the main paper. The $\{\mathbf{B}_b\}$ are relative to *the rest-pose joints*.

2. Model Architectures

We provide illustrations for our model architectures in Fig. 1. Note that throughout our implementations, we use 1D convolutions with kernel size 1 to substitute the fully-connected layers used in typical occupancy networks. This enables us to use the off-the-shelf implementation of grouped convolutions [7] in PyTorch [8] to represent piecewise occupancy classifiers $\{\bar{\mathcal{O}}^b_\omega\}$ and PTF $\{\bar{\mathcal{T}}^b_\psi\}$.

Fig. 1a shows our fully-piecewise model, PTF-Piecewise. The key observation we make here is that, the pre-activation occupancy logits $\{\boldsymbol{\sigma}_b\} \in \mathbb{R}^{3\times B}$ can be used to compute the softmax probability over parts. To construct inputs to the occupancy classifier module, we concatenate the local point cloud feature $\mathbf{c}_\mathbf{x}$ to each $\hat{\mathbf{x}}_b, \forall b \in \{1, \cdots, B\}$, thus a piecewise occupancy function $\bar{\mathcal{O}}^b_\omega$ of bone b takes the concatenation of $\mathbf{c}_\mathbf{x}$ and $\hat{\mathbf{x}}_b$ as the input and outputs the occupancy logits $\hat{\boldsymbol{\sigma}}_b$. We set $B = 24$ for PTF-Piecewise.

Fig. 1b shows our PTF-FC model. Here we use a similar structure as in IPNet [2] which predicts part probabilities with a separate part classifier. The difference between IPNet and our PTF-FC is that we apply PTF to query point $\mathbf{x}$ before feeding it to the occupancy classifier. The first layer (in blue) of the PTF module takes the concatenation of $\mathbf{c}_\mathbf{x}$ and $\mathbf{x}$ as the input and outputs a $128B$ dimensional feature for later stages. We set $B = 14$ for PTF-FC.



(a) PTF-Piecewise

(b) PTF-FC

Figure 1: Illustration for model architectures. Blue rectangles indicate 1D convolution layers, green rectangles indicate 1D group-convolutions with B groups, $\frac{e}{\sum e}$ squares indicate the softmax function, $\circ$ indicates the Hadamard product with broadcasting, $\sum$ squares indicate summation over part dimension, M_1 and M_2 squares indicate $\max$ over the occupancy dimension and part dimension, respectively. Both PTF-Piecewise and PTF-FC take a query point $\mathbf{x}$ and the corresponding local point feature $\mathbf{c_x}$ as input, and output a multi-class occupancy probability $\hat{\mathbf{o}}$ and the rest-pose correspondence $\hat{\mathbf{x}}$.

Method	Outer Err.	Max Outer Err.	Inner Err.	Max Inner Err.
IPNet	28.2 mm	562.7 mm	28.3 mm	544.2 mm
IPNet-128	26.4 mm	564.4 mm	26.9 mm	546.5 mm
Stitched Puppet [11]	36.1 mm	454.9 mm	NA	NA
3D-CODED [6]	23.7 mm	614.8 mm	NA	NA
PTF-FC	23.1 mm	**76.6 mm**	23.1 mm	88.9 mm
PTF-FC-128	**21.2 mm**	83.6 mm	**21.4 mm**	**82.3 mm**

Table 1: Additional registration evaluation on the CAPE dataset. IPNet-128 and PTF-FC-128 indicate IPNet model and our model trained with encoder resolution of 128×3, where size of the input point cloud remains 5K.

3. Additional Quantitative Results on the CAPE Dataset

In this section we compare our approach to other related baselines, namely 3D-CODED [6] and the Stitched Puppet [11]. Note that to our best knowledge, IPNet is the only published method that addresses automatic **model-fitting** of **both** under-cloth body and clothed body, given unoriented, sparse point-clouds of **clothed** humans. Both IPNet and ours can produce controllable parameters and clothed avatars from input point clouds.

3D-CODED [6] only focuses on correspondence prediction. Their method can not produce controllable parameters for under-cloth or clothed body models. Furthermore, 3D-CODED uses exhaustive initial rotation search and is allowed to deform the template freely, while we can only optimize for $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, $\mathbf{t}$ and $\mathbf{D}$, where $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ are limited to SMPL parameter space, and $\mathbf{D}$ is in the A-pose and regularized to be small.

The Stitched Puppet [11] tackles model-fitting as a pure optimization problem via belief propagation. Since [11] is based on SCAPE [1] topology, we fit it to FAUST [4] scans with the neutral pose, and then obtain correspondence from SCAPE vertices to SMPL vertices via barycentric interpolation.

We report the numbers in Tab. 1. For clothing surface registration, 3D-CODED produces slightly inferior results than ours

(a) IPNet surface (b) IPNet SMPL (c) IPNet SMPL+D (d) Ours surface (e) Ours SMPL (f) Ours SMPL+D

Figure 2: One common failure case of IPNet [2] happens when the global orientation deviates significantly from zero. This often results in catastrophic failures in the subsequent (b) SMPL fit and (c) SMPL+D fit.

(a) IPNet surface (b) IPNet SMPL (c) IPNet SMPL+D (d) Ours surface (e) Ours SMPL (f) Ours SMPL+D

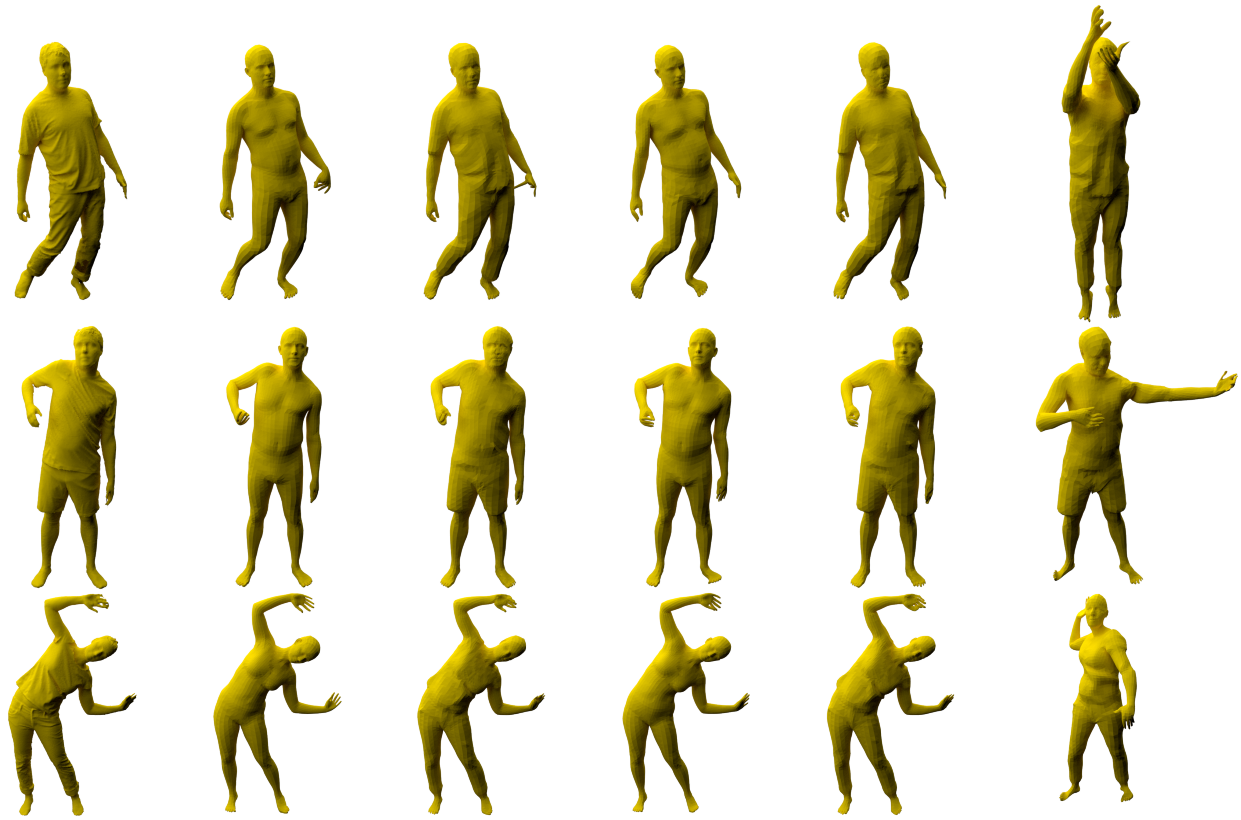
Figure 3: Another type of failure cases of IPNet [2] is caused by poor generalizability. As can be seen in (a), although IPNet has been trained using random augmentation to global orientation, it still struggles to generalize to the case in which the person's two arms are roughly aligned with the z-axis. IPNet misclassifies both arms of the person as the left-arm (same color), this results in catastrophic failure in the subsequent (b) SMPL fit and (c) SMPL+D fit. In (d) our model generalizes better and correctly distinguishes the left-arm and the right-arm under this rare pose, resulting in much more accurate registrations (e+f).

although it is less constrained. The results from the Stitched Puppet are inferior to all others probably due to 1) the sensitivity of belief propagation to noise and 2) the fact that clothed surfaces vary more than SCAPE can capture. Most importantly, our approach **completely avoids** catastrophic failures ($>$ 100 mm error) that happen for the baselines, demonstrating strong generalization.

4. Additional Qualitative Results on the CAPE Dataset

We show additional qualitative results on the CAPE dataset in Fig. 2 and Fig. 3. We observe that IPNet [2] often fails when the global orientation deviates too much from zero. As can be seen in Fig. 2, even if the local poses are relatively simple, IPNet still fails catastrophically; this is most likely due to the fact that, in order to make the optimization stable, the optimization objective often has regularization terms that penalize poses which deviate too much from the mean-pose. This choice is statistically meaningful, but without pose initialization, it will also make it impossible for the optimizer to converge to rare poses that are realistic but deviate too much from the mean-pose.

Another type of failure cases of IPNet is caused by its network's poor generalizability. Fig. 3 shows that IPNet fails to correctly distinguish between the left-arm and the right-arm under a rare pose. This is because IPNet learns the occupancy functions of the two arms in posed space, and thus these occupancy functions need to memorize all possible locations of arms in posed space. On the other hand, our model learns to canonicalize points before the occupancy classification, thus the occupancy functions of arms only need to memorize a small region in rest-pose space. This results in better generalizability and our model correctly distinguishes the left-arm and the left-arm in this case.



(a) Raw scan (b) IPNet SMPL (c) IPNet SMPL+D (d) Ours SMPL (e) Ours SMPL+D (f) Ours reposed

Figure 4: More qualitative results on the BUFF dataset. Note that our SMPLD/SMPL+D fits perform consistently better than IPNet [2], especially around faces and hands

GT Ours surface Ours SMPL GT Ours surface Ours SMPL

Figure 5: Failure cases: we note that implicit surface reconstruction often fails in self-contact or near-self-contact scenarios. This leads to failures in registration.

5. Additional Qualitative Results on the BUFF Dataset

We present more qualitative results on the BUFF dataset in Fig. 4. Note that the BUFF dataset [10] *does not* contain ground-truth pose parameters, thus it is not possible to evaluate quantitatively on registration errors as we did on the CAPE dataset. These qualitative results are meant to demonstrate generalization performance of our trained model to real scans, even though the model is only trained on synthetically sampled point clouds from registered dressed people.

6. Limitations

Our approach often fails in self-contact or near-self-contact scenarios. We show typical failure cases in Fig 5. Another limitation of our approach is that it requires fully-supervised training on accurate surface registration. This kind of data is very hard to acquire in practice, thus limiting the scalability of our approach. A straightforward improvement would be integrating the self-supervised loop of [3] into our pipeline, or utilizing the weakly supervised approach of [9] to generate training data using registered under-cloth SMPL body.

References

[1] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 24, 2005. 2

[2] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *Proc. of ECCV*, 2020. 1, 3, 4

[3] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. In *Proc. of NeurIPS*, 2020. 5

[4] Federica Bogo, Javier Romero, Matthew Loper, and Michael J. Black. FAUST: Dataset and evaluation for 3D mesh registration. In *Proc. of CVPR*, 2014. 2

[5] Boyang Deng, JP Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Nasa: Neural articulated shape approximation. In *Proc. of ECCV*, 2020. 1

[6] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan Russell, and Mathieu Aubry. 3d-coded : 3d correspondences by deep deformation. In *Proc. of ECCV*, 2018. 2

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. of NeurIPS*, 2012. 1

[8] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Proc. of NeurIPS*, 2019. 1

[9] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *Proc. of CVPR*, 2021. 5

[10] Chao Zhang, Sergi Pujades, Michael J. Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *Proc. of CVPR*, 2017. 4

[11] Silvia Zuffi and Michael J. Black. The stitched puppet: A graphical model of 3D human shape and pose. In *Proc. of CVPR*, 2015. 2