Unsupervised Shape and Pose Disentanglement for 3D Meshes - Supplementary Material

Keyang Zhou, Bharat Lal Bhatnagar, and Gerard Pons-Moll

Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
{kzhou,bbhatnag,gpons}@mpi-inf.mpg.de

1 Architecture Details

In the following we summarize the model architecture we used to train on AMASS [4], SMAL [9], COMA [5] and MANO [6]. Our model consists of a shape branch, a pose branch and a decoder. Each model component is a 4-layer neural network with alternated spiral convolution layers and sampling layers. The number of filters for convolutional layers and the sampling rates for sampling layers are listed in Table 1.

	Shape Dim	Pose Dim	Shape Branch	Pose Branch	Decoder	Sampling
AMASS	16	112	4, 8, 16, 32	12, 24, 48, 96	128, 64, 32, 16	4, 4, 4, 4
SMAL	36	60	6, 12, 24, 48	10, 20, 40, 80	128, 64, 32, 16	4, 4, 4, 4
COMA	4	4	4, 8, 16, 32	8, 16, 32, 64	128, 64, 32, 16	4, 4, 4, 4
MANO	4	16	4, 8, 16, 16	8, 16, 32, 64	64, 64, 32, 16	4, 2, 2, 2

Table 1: Model architecture for different datasets.

2 COMA Exprapolation Experiment

We show the detailed comparison of our method to Jiang et al. [1] and FLAME [2] for all 12 COMA extrapolation experiments in Table 2. Notably, our method requires no manual processing but outperforms Jiang et al. for all the cross-validation experiments.

3 Shape Space Exploration

Our method used cross-consistency loss to implicitly enforce that shape codes of the same subject are consistent under swapping. We believe that imposing the loss in mesh space gives the network more flexibility to choose the optimal structure of latent space. It remains an interesting question how far are these shape codes away from each other. It turns out that on AMASS test set, the average intra-subject shape code distance is 0.092, while the average inter-subject shape code distance is 0.218. To qualitatively demonstrate it, we sampled 20

	Ours	Jiang et al.'s	FLAME
bareteeth	**1.472**	1.695	2.002
cheeks in	**1.425**	1.706	2.011
eyebrow	**1.151**	1.475	1.862
high smile	**1.359**	1.714	1.960
lips back	**1.379**	1.752	2.047
lips up	**1.256**	1.747	1.983
mouth down	**1.155**	1.655	2.029
mouth extreme	**1.473**	1.551	2.028
mouth middle	**1.128**	1.757	2.043
mouth open	**1.098**	1.393	1.894
mouth side	**1.341**	1.748	2.090
mouth up	**1.174**	1.528	2.067
average	**1.284**	1.643	2.001

Table 2: Mean errors of 12-fold cross-validation for COMA expression extrapolation. All numbers are in millimeters. The results of Jiang et al. and FLAME are taken from [1].

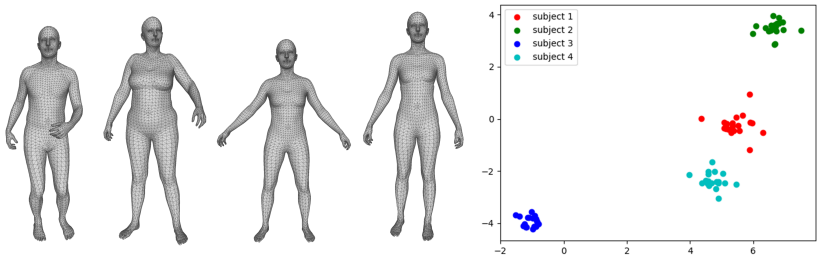


Fig. 1: Left: four subjects from AMASS test set. Right: t-SNE plot of shape codes of different deformations of these subjects.

poses for each of the four subjects in AMASS test set. We encoded all meshes into shape space and visualized with t-SNE plot, as shown in Fig. 1. The shape codes of every subject clearly form a cluster. This justifies using shape codes for retrieval.

4 Additional Pose Transfer Results

In Fig. 2 we show additional pose transfer results at a higher resolution. While our method produces natural pose-transferred reconstructions in general, artefacts are sometimes incurred at certain difficult parts such as fingers of human body. It is probably the consequence of using a simple L1 loss, which does not take into account local curvature and vertex density. A more sophisticated loss function for geometry reconstruction is however beyond the scope of this paper.

5 Limitations

Pose transfer or auto-encoding with our model typically fails for two situations: either the orientation or the pose is uncommon in the training data. Fig. 3 shows a few such examples. In particular, Fig. 3a, 3c, 3d are bent-over poses, while Fig. 3b is upside down compared to the template. Augmenting training data with sufficient meshes of such kinds could alleviate this problem.

Another limitation of our model is that it cannot interpolate between a pair of meshes involving a nontrivial global rotation, as illustrated in Fig. 4. This implies that the latent pose space learned by our model is not linear. This is not a problem in practice, because we can always manually align the meshes before interpolation. However, a linear pose space is still useful and we leave this to future work.

6 Discussion

Compared with parametric 3D models such as SMPL [3], our model is trained on a much larger dataset. This is due to the fact that SMPL is based on linear blend skinning, which requires limited model capacity, while our model learns deformations and articulations completely from scratch. Our data-driven approach is proven to be effective in unsupervised learning of shape-dependent and pose-dependent deformations. However it is still inferior to the state-of-the-art parametric models in terms of surface details and reconstruction quality. One potential remedy is to embed domain prior into deep neural networks, as was done in [8,7]. Moreover, despite having more training samples, AMASS dataset lacks the huge variation in subject shape which is present in SMPL's training scans. For reproducibility concern, we only train on publicly available datasets. But we expect our model to learn a richer shape space once we complement the training data with more extreme body shapes.

References

1. Jiang, Z.H., Wu, Q., Chen, K., Zhang, J.: Disentangled representation learning for 3d face shape. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11957–11966 (2019)
2. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. ACM Transactions on Graphics **36**(6), 194:1–194:17 (Nov 2017), two first authors contributed equally
3. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM transactions on graphics (TOG) **34**(6), 1–16 (2015)
4. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: IEEE International Conference on Computer Vision (ICCV). IEEE (oct 2019)
5. Ranjan, A., Bolkart, T., Sanyal, S., Black, M.J.: Generating 3d faces using convolutional mesh autoencoders. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 704–720 (2018)

6. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics (ToG) **36**(6), 245 (2017)
7. Tan, Q., Gao, L., Lai, Y.K., Xia, S.: Variational autoencoders for deforming 3d mesh models. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5841–5850 (2018)
8. Tretschk, E., Tewari, A., Zollhöfer, M., Golyanik, V., Theobalt, C.: Demea: Deep mesh autoencoders for non-rigidly deforming objects. arXiv preprint arXiv:1905.10290 (2019)
9. Zuffi, S., Kanazawa, A., Jacobs, D.W., Black, M.J.: 3d menagerie: Modeling the 3d shape and pose of animals. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6365–6373 (2017)

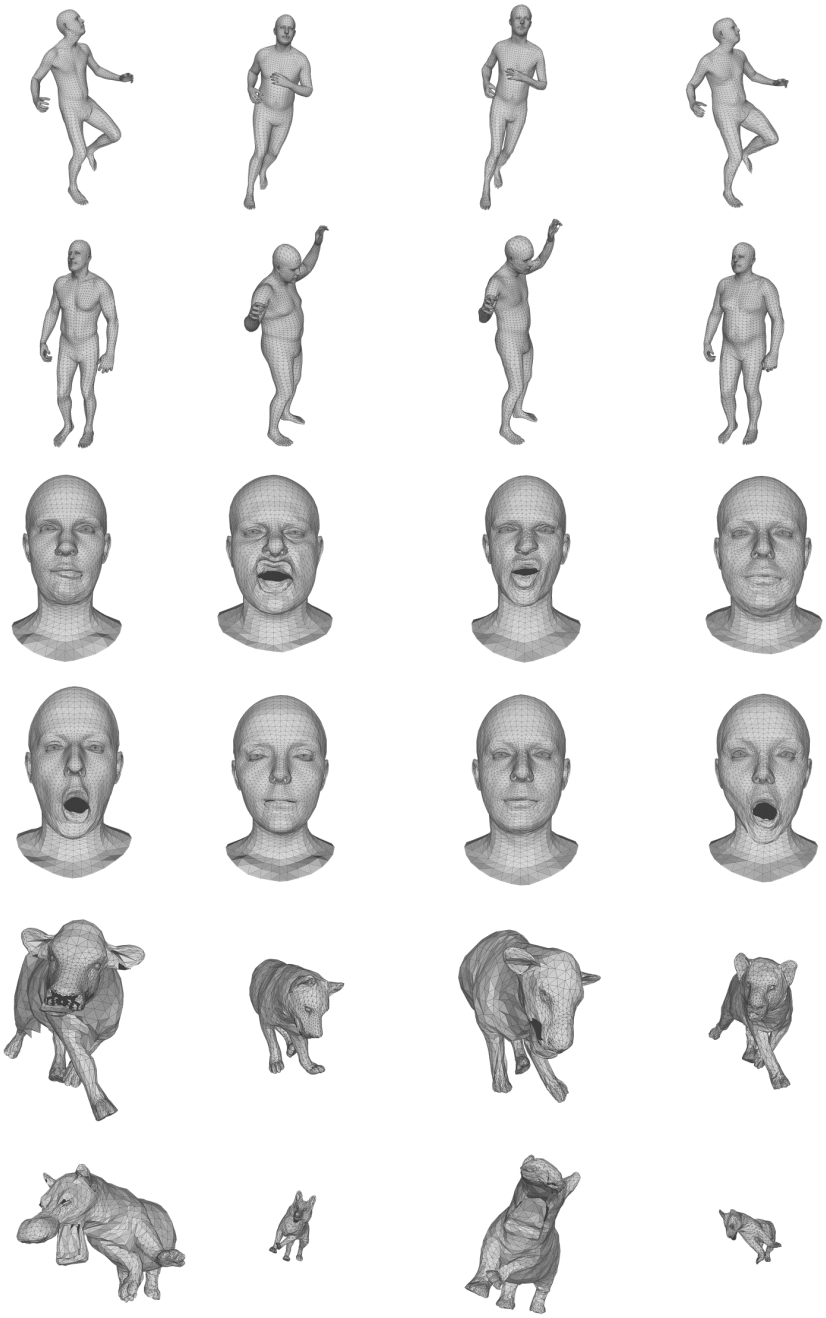
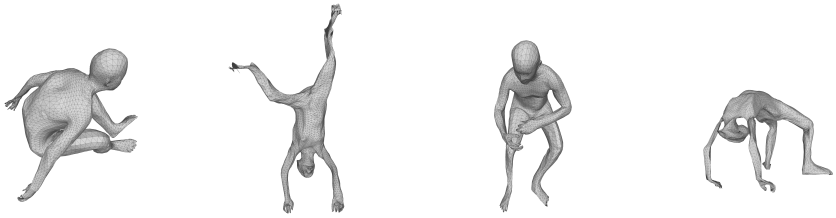


Fig. 2: Pose transfer results on AMASS, COMA and SMAL. From left to right: subject 1, subject 2, pose of subject 2 transferred to subject 1, pose of subject 1 transferred to subject 2.



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

Fig. 3: Failure cases of pose transfer. It mostly occurred on difficult orientations or poses.

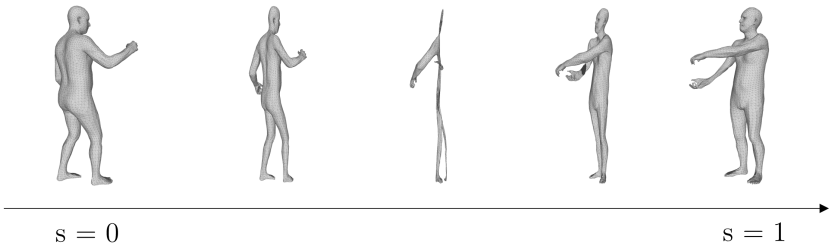


Fig. 4: Failure case of pose interpolation. Intermediate meshes are squeezed due to interpolating global rotations.