Supplementary - Adjoint Rigid Transform Network: Task-conditioned Alignment of 3D Shapes

Keyang Zhou[1,2] Bharat Lal Bhatnagar[1,2] Bernt Schiele[2] Gerard Pons-Moll[1,2]
[1]University of Tübingen, Germany
{keyang.zhou,gerard.pons-moll}@uni-tuebingen.de
[2]Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
{bbhatnag, schiele}@mpi-inf.mpg.de

In this supplementary material we provide further details about our method such as architecture and rotation representation used in ART. Next, we provide more results for shape reconstruction and alignment using ART on ShapeNet [3]. We also provide more qualitative results for human mesh registration and pose interpolation.

1. Architecture Details

In this section we describe the architecture of Adjoint Rigid Transform (ART) Network for point clouds and meshes respectively.

Point cloud The architecture for point cloud inputs resembles the T-net in PointNet [8]. Specifically, we first learn point-wise features by three 1D convolutional layers of size [64, 128, 1024]. A max-pooling layer then aggregates features over all points and produces a global feature vector. It is subsequently mapped to the rotation representation by three fully-connected layers with size [512, 256, 6]. We apply batch normalization [6] and ReLU to every layer except the input and output. The number of training samples in each ShapeNet category ranges from 3000 to 8000, while we train on 20000 samples for human registration.

Mesh For mesh inputs, we assume that they are registered to a common template and thus have the same connectivity. To keep the architecture simple, we only use mesh down-sampling layers and fully-connected layers. We first simplify the mesh to $\frac{1}{16}$ of its original resolution based on the quadric error metrics proposed in [4]. This down-sampling rate is shown to work well for meshes parameterized by SMPL [7], but it might need to be tuned on other meshes. Then we flatten the down-sampled mesh and feed it to fully-connected layers of size [128, 64, 6]. Human pose transfer is trained on 120000 samples.

2. Rotation Representation

We choose to use the continuous rotation representation proposed by Zhou *et al*. [10] since it was shown to be superior to other representations such as quaternions and Euler angles in rotation regression tasks. Let R_A be a learnable function. We have

$$[\boldsymbol{a}_1 \ \boldsymbol{a}_2] = R_A(\mathbf{X}) \tag{1}$$

where $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ are vectors in $\mathbb{R}^3$. Then we apply Gram-Schmidt process to obtain the orthogonal matrix $\mathbf{R}$.

$$\mathbf{R} = [\boldsymbol{r}_1 \ \boldsymbol{r}_2 \ \boldsymbol{r}_3] \tag{2}$$

$$= \left[\frac{\boldsymbol{a}_1}{\|\boldsymbol{a}_1\|} \ \frac{\boldsymbol{a}_2 - (\boldsymbol{a}_2^T \boldsymbol{r}_1)\boldsymbol{r}_1}{\|\boldsymbol{a}_2 - (\boldsymbol{a}_2^T \boldsymbol{r}_1)\boldsymbol{r}_1\|} \ \boldsymbol{r}_1 \times \boldsymbol{r}_2 \right] \tag{3}$$

Note that in the last column of $\mathbf{R}$ we take cross product of the first two columns to ensure $\det(\mathbf{R}) = 1$.

3. Scale Alignment

Besides orientation alignment, we can also extend ART to align the scale of shapes by predicting a scaling factor for each input shape, and enforcing scaling equivariance in the same manner as rotation equivariance. The new equivariance constraint then becomes

$$S_A(\boldsymbol{X})\, R_A(\boldsymbol{X})\, \boldsymbol{X} = s S_A(s\boldsymbol{R}\boldsymbol{X})\, R_A(s\boldsymbol{R}\boldsymbol{X})\, \boldsymbol{R}\boldsymbol{X}, \tag{4}$$

where $s \in \mathbb{R}^+, \boldsymbol{R} \in \mathrm{SO}(3)$ are random scaling factor and rotation, S_A and R_A perform scaling and rotation canonicalization respectively. We trained the extended model on the SMAL [11] dataset and example alignment results are shown in Fig. 1.

4. Shape Alignment

We show the distribution curves of pairwise alignment error for all categories under random azimuthal rotation

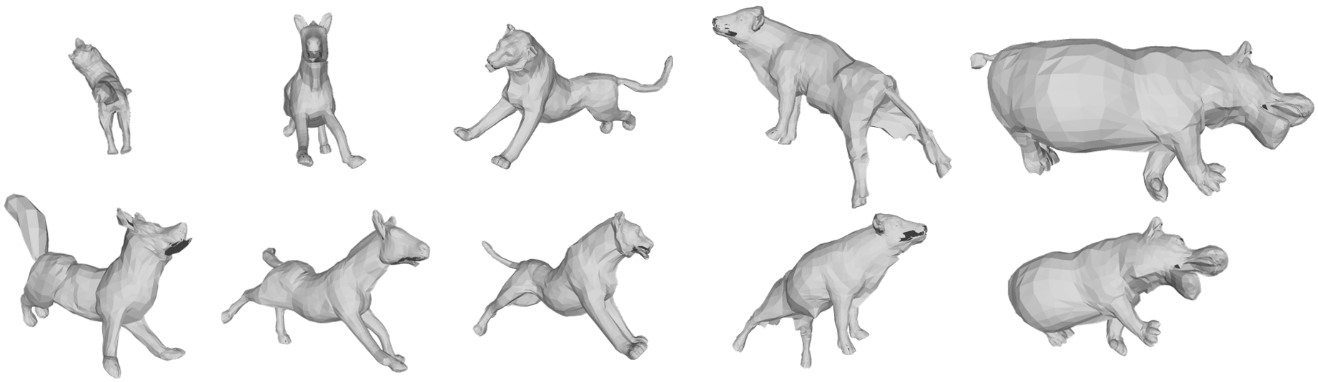

Figure 1. Top: Input SMAL shapes from five different categories. Bottom: ART-aligned shapes with consistent orientations and scales.

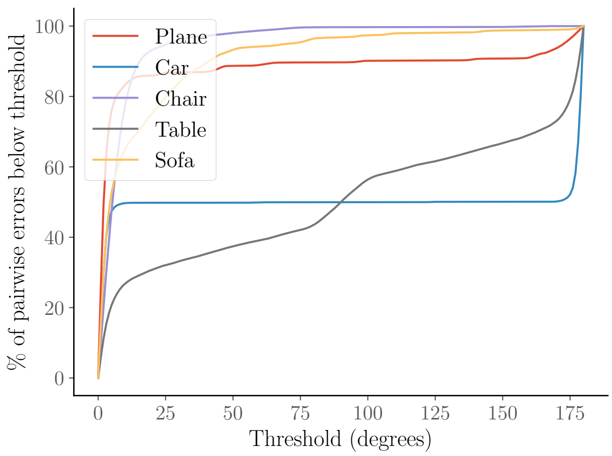

Figure 2. Percentage of shape pairs with an angular distance less than the given thresholds.

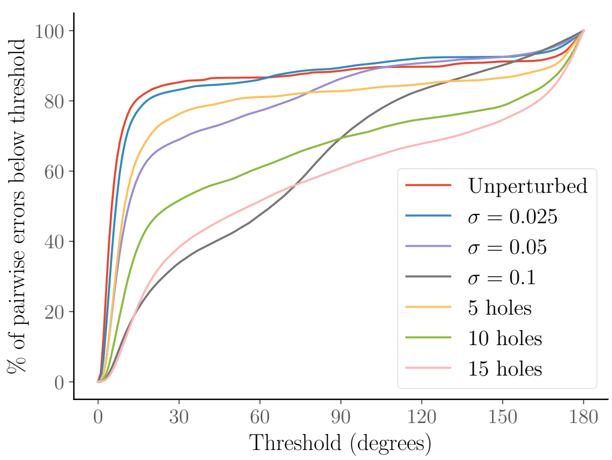

Figure 3. Effect of data perturbation on alignment.

in Fig. 2. Since the table category suffers from ambiguity of rotation symmetry, the quantitative measure on tables is not conclusive and we only include it here for completeness. We can observe that the alignment accuracy for planes, chairs and sofas are good, with more than 80% of shape pairs differing by less than 30°. However, ART was still confused by the front and back of cars and learned two modes of canonical orientations, as can be seen from the blue curve. Qualitative results for random 3D alignment are demonstrated in Fig. 4.

Shapes captured from real world sensor data are often accompanied with noise and holes. Hence we also evaluate the robustness of ART alignment to various shape perturbations. We apply random Gaussian noise with different standard deviation σ as an approximation of real noise. To simulate holes, we randomly sample a given number of points from the point cloud as hole centers, and remove all points within a radius of 0.2 from the centers. The alignment error distribution plot is shown in Fig. 3. ART is robust to a reasonable amount of noise and holes. Even in the worst case the alignment accuracy is comparable to PCA (see Fig. 7 in main paper).

5. Point Cloud Auto-encoding

ART improves the performance of existing methods [1] on point cloud reconstruction by aligning the data to a common global orientation. We show more qualitative examples for point cloud reconstruction using ART in Fig. 5.

6. Human Body Registration

We show more qualitative examples for human mesh registration using ART in Fig. 6. This experiment clearly highlights the general applicability of ART for non-rigid objects. It can also be seen that human meshes with varying poses can still be aligned to a common global orientation with ART.

7. Human Pose Interpolation

The human pose interpolation method used by Zhou *et al*. [9] can have squeezing artifacts when the source and the target differs by a nontrivial global rotation. ART mitigates this problem by explicitly factoring out and interpolating global rotations. Additional qualitative results are shown in Fig. 7.

8. Limitations and Future Work

ART is a simple yet effective module that helps a wide variety of 3D networks to achieve satisfactory performance when training on data without pre-alignment. However, it can still be improved in several aspects. First, although the formulation of ART subsumes general rotations in $SO(3)$, currently it performs much better on shapes perturbed by random azimuthal rotations only. We hypothesize that this limitation is related to architecture capacity. Hence we plan to explore using SotA point cloud architectures as the backbone of ART. Besides, ART learns semantic features for shape alignment completely from scratch. In future work, we plan to utilize more geometric properties such as the plane of reflection so that ART can have more clues to make accurate predictions.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pages 40–49. PMLR, 2018. 2

[2] Federica Bogo, Javier Romero, Matthew Loper, and Michael J. Black. FAUST: Dataset and evaluation for 3D mesh registration. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Piscataway, NJ, USA, June 2014. IEEE. 6

[3] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 1

[4] Michael Garland and Paul S Heckbert. Surface simplification using quadric error metrics. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 209–216, 1997. 1

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

[6] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 1

[7] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 1

[8] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1

[9] Keyang Zhou, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Unsupervised shape and pose disentanglement for 3d meshes. In *European Conference on Computer Vision (ECCV)*, August 2020. 2, 7

[10] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 1

[11] Silvia Zuffi, Angjoo Kanazawa, David Jacobs, and Michael J. Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1

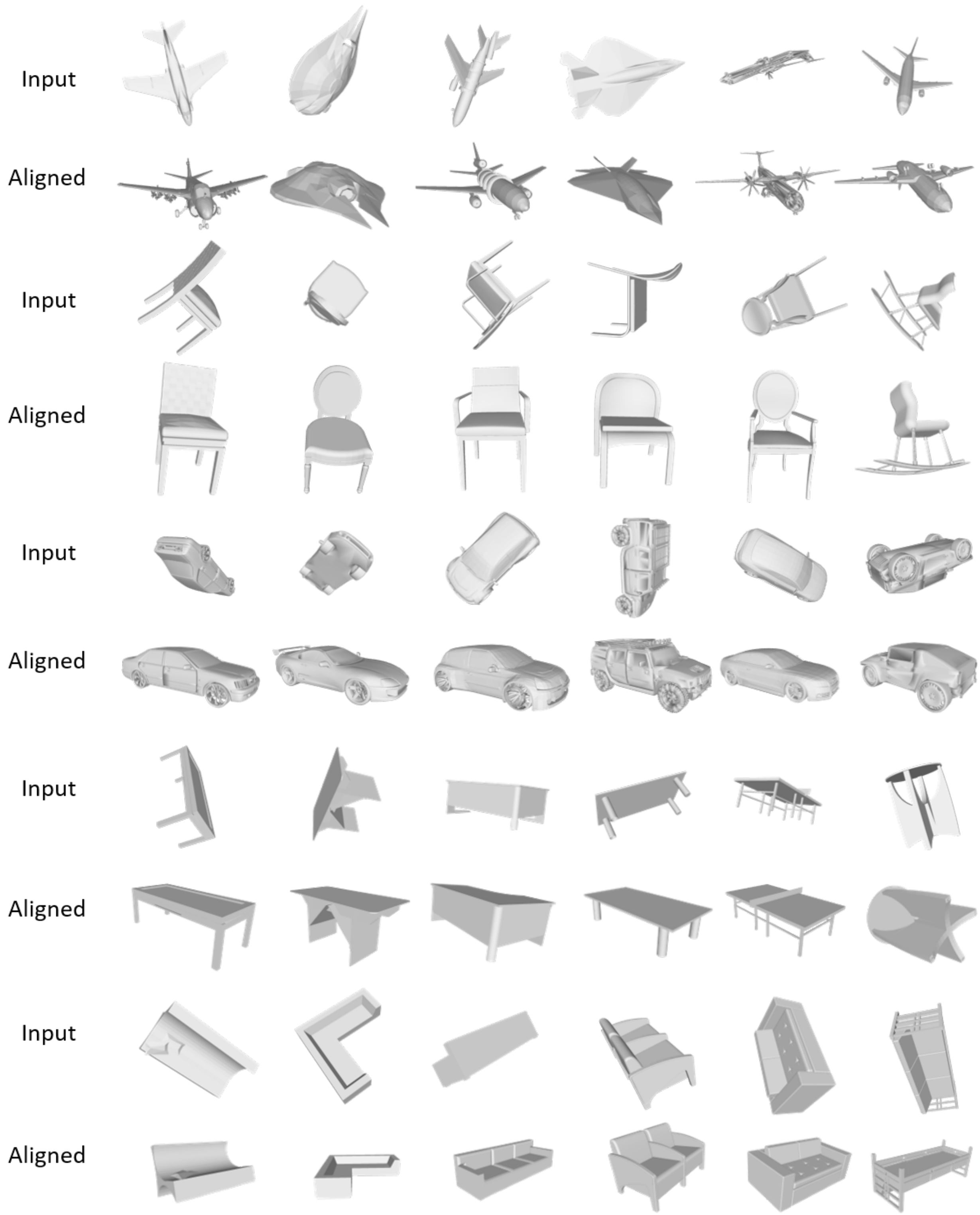

Figure 4. Alignment of shapes perturbed by random 3D rotations with ART. The last column shows failure cases.

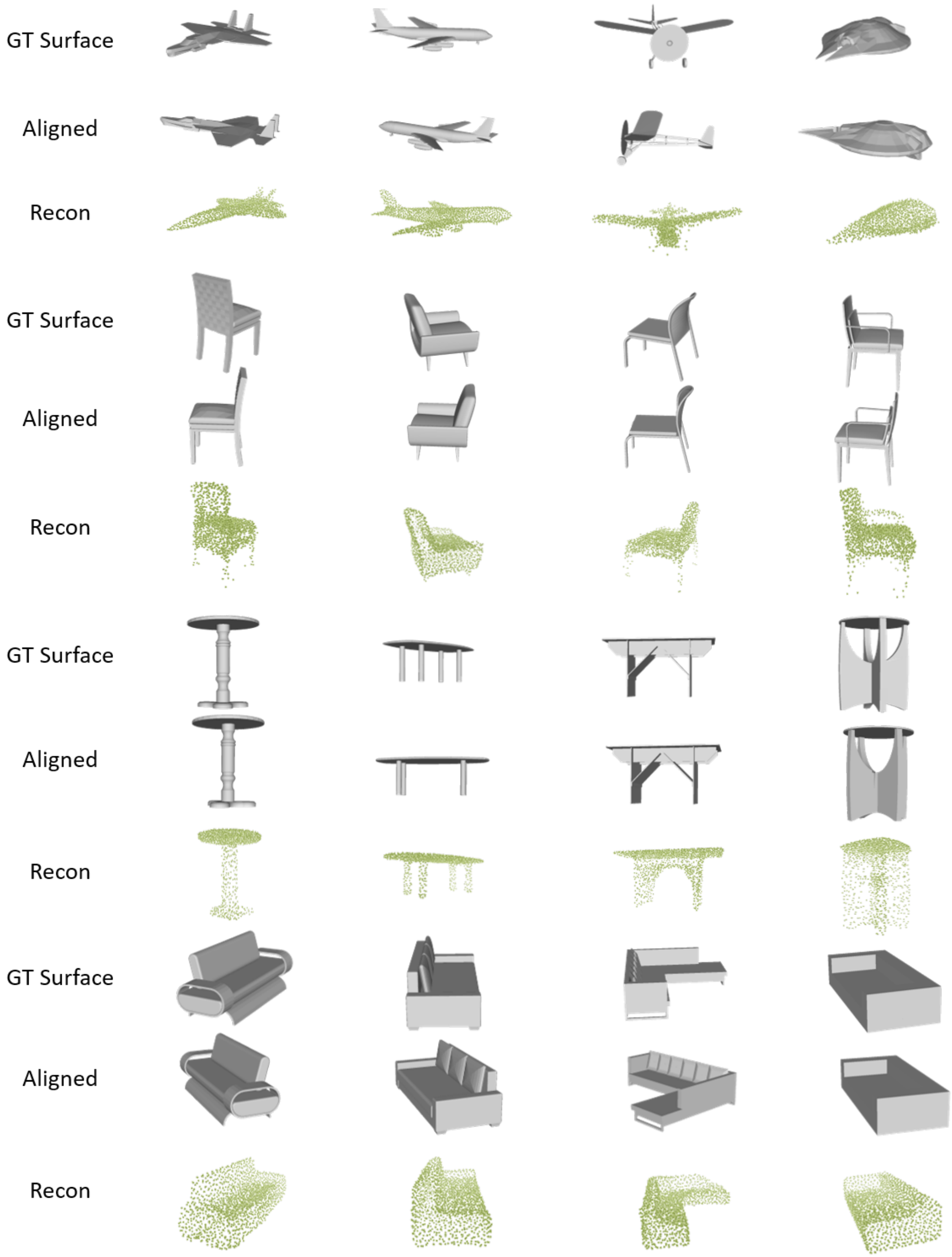

Figure 5. Reconstruction of ShapeNet surfaces with ART. We show the groundtruth surfaces, surfaces aligned by ART, and point cloud reconstructions in order.

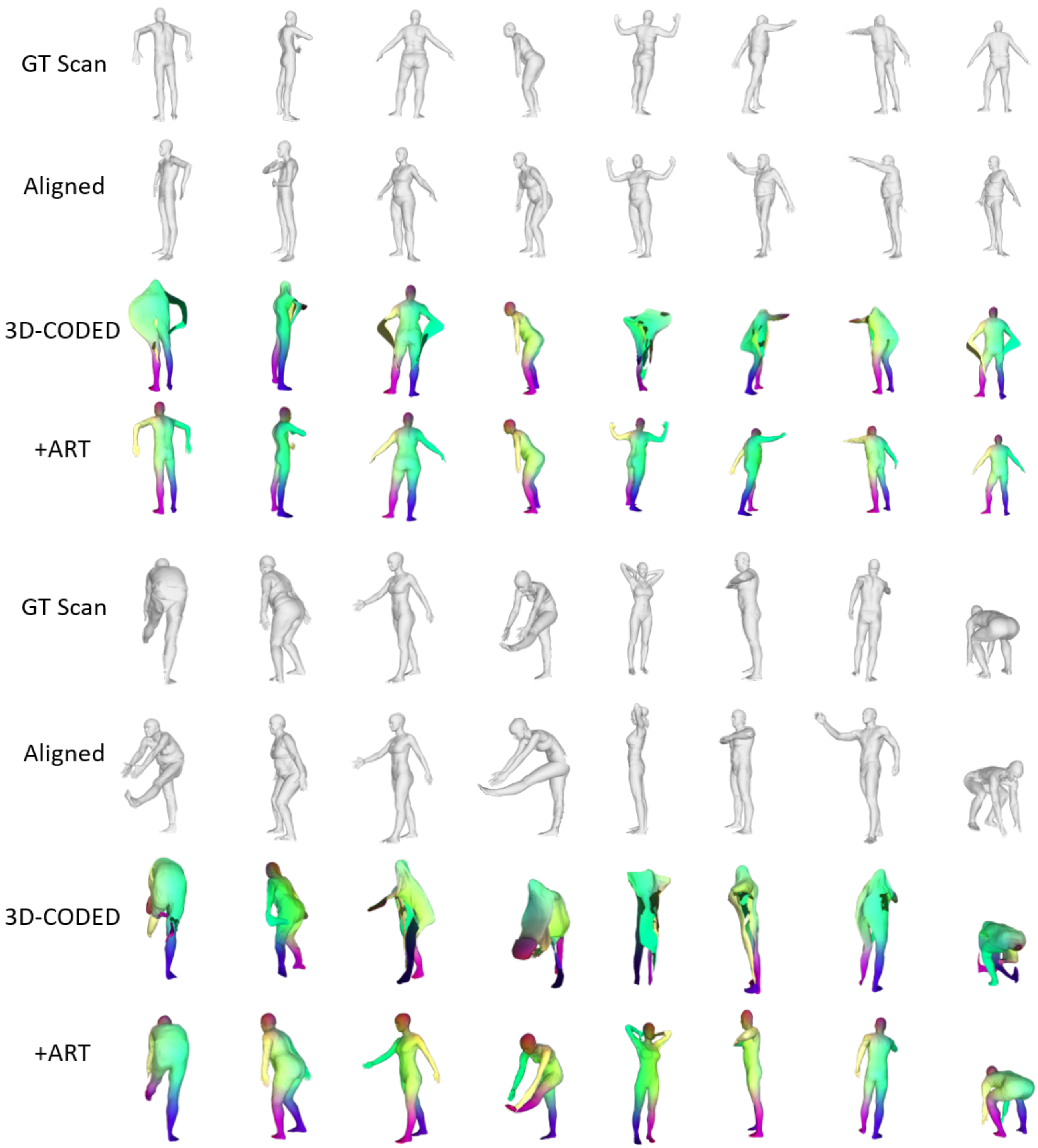

Figure 6. Registration of (rotated) raw FAUST [2] scans using 3D-CODED [5] and ART. Note that all results were obtained using single initialization.

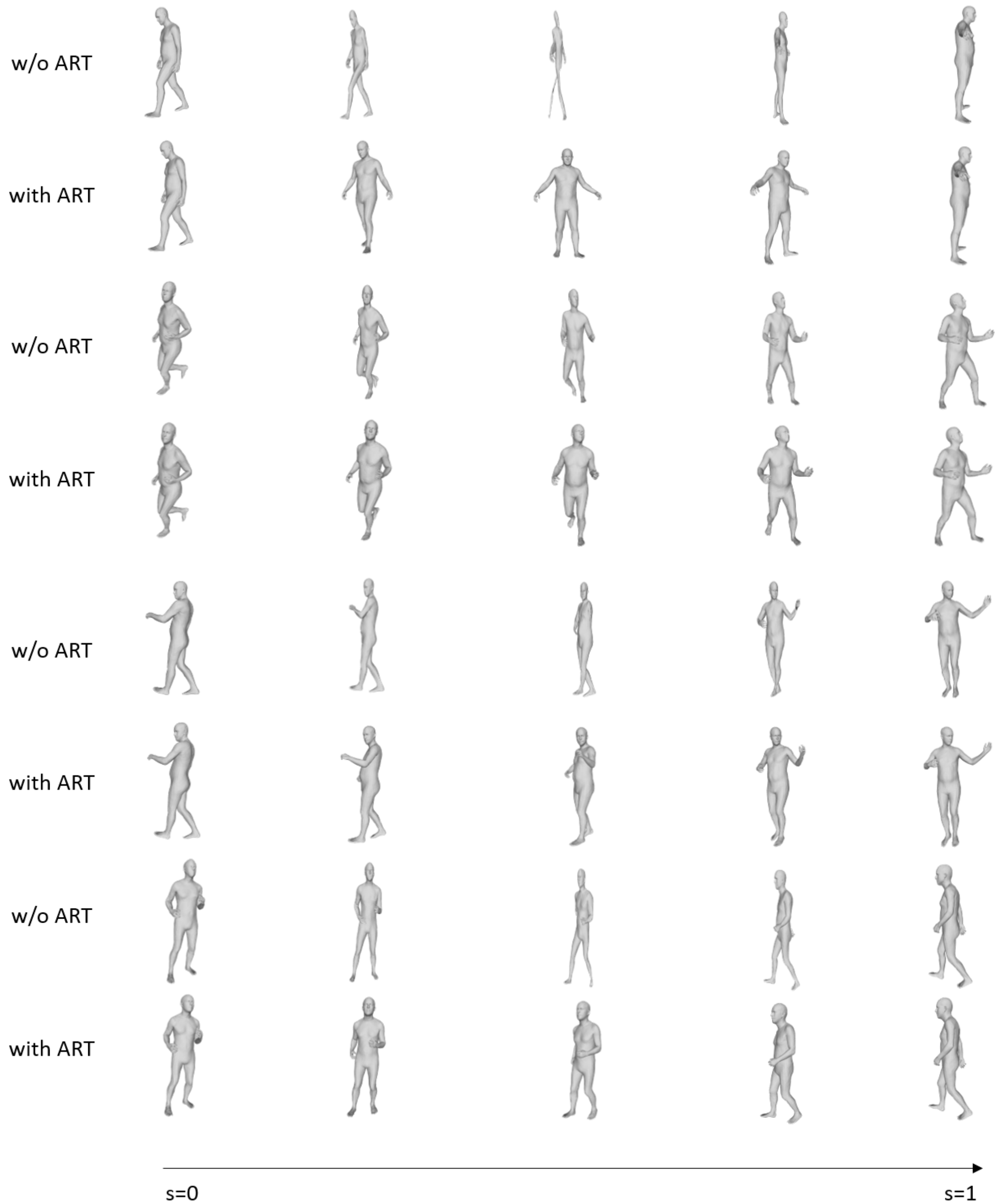

Figure 7. Human pose interpolation using Zhou *et al*. [9]. The source pose is at $s = 0$ and target pose at $s = 1$. We show three intermediate poses at uniform time steps.