Supplementary Material for "Self-supervised Learning of Implicit Shape Representation with Dense Correspondence for Deformable Objects"

Baowen Zhang[1,2] Jiahe Li[1,2] Xiaoming Deng[1,2*] Yinda Zhang[3*]
Cuixia Ma[1,2] Hongan Wang[1,2]
[1]Institute of Software, Chinese Academy of Sciences [2]University of Chinese Academy of Sciences
[3]Google

In this supplementary material, we first introduce the details on network architecture (Section A) and loss functions (Section B). Then we show additional experiments and ablation study (Section C). Finally, we give the closed-form analytical solution to get the minimal alignment error of our piece-wise rigid constraint (Section D.1), analyze the least square solution of rotation (Section D.2.1), and reveal the relationship between ARAP loss and the local rigid constraint (Section D.2.2).

A. Details on Network Architecture

In this section, we describe more details of our network. It consists of three modules: an encoder-decoder network, part prediction networks, and an SDF prediction module.

Encoder-Decoder Network. The encoder-decoder network predicts correspondences on template (Figure 2 of the main paper). The encoder receives a point $\mathbf{p}$ from the target space S_i, along with its predicted SDF value from the SDF prediction module as input. It then produces a vector $\mathbf{l}(\mathbf{p}) \in \mathbb{R}^8$. Following that, the decoder takes $\mathbf{l}(\mathbf{p})$ as input and outputs the corresponding point $D_{i\to tmpl}(\mathbf{p})$ in template space. Our encoder and decoder consist of 5 and 4 fully connected layers, respectively.

Part Prediction Networks. The networks ψ_e and ψ_d predict part probabilities, *i.e.* $\mathbf{P}_h$ in Eq. 6 of the main paper, for each point. Each network divides the target shape into 20 parts, totally 40 parts together for calculating piece-wise rigid constraint, *i.e.*, $N_P = 40$ in Eq. 6 of the main paper. Our part prediction networks ψ_e and ψ_d have 4 and 3 fully connected layers, respectively. We use SoftMax to normalize the probabilities predicted by each network. Both networks are trained in a self-supervised manner by optimizing piece-wise rigid constraint.

For each point $\mathbf{p}$ in a target space S_i, ψ_e takes the output vector $\mathbf{l}(\mathbf{p}) \in \mathbb{R}^8$ of the encoder as input, and ψ_d takes the correspondence point $D_{i\to tmpl}(\mathbf{p}) \in S_{tmpl}$ in template space as input. Since the correspondences are consistent across shapes deformed from the same template, the part segmentation learned by ψ_d is also consistent across all shapes. The segmentation results are shown in Figure 4 in our main paper and Figure C in the supplementary material. However, the correspondences are not learned well at the beginning of training. Conceptually, the prediction of ψ_d highly depends on learned correspondences and template, so it cannot be effectively trained at the beginning of training, with highly-undertrained optimization of correspondences and template. In order to address this issue, we use ψ_e to predict part segmentation, which does not depend on correspondences on template. During training stage, we observe that ψ_e provides valid rigid constraint earlier than ψ_d and enables the network to converge faster.

SDF Prediction Module. The SDF prediction module Φ is in charge of modeling template SDF field as well as SDF fields of other shapes in training set. With dense correspondence predicted by the encoder-decoder network, we can query SDF values from template field to reconstruct target shapes (Eq. 2 in the main paper). Inspired by Atzmon *et al.* [1] that the initial scalar field contributes greatly to shape representation learning, we add the distance of an input point $\mathbf{p}$ to center $(0, 0, 0)$ to the output of the neural implicit SDF function Φ and achieve similar initialization to Atzmon *et al.* [1]. The output of SDF prediction module is formulated like [6] as $\Phi(\mathbf{p}|\alpha) = \phi(\mathbf{p}|\alpha) + \|\mathbf{p}\|_2$, where ϕ denotes a neural network for SDF prediction. The network ϕ consists of 5 fully connected layers.

Other Details. Similar to the previous works [4, 9], the parameters of encoder, decoder, and SDF prediction module are all predicted by Hyper-Nets, while part probability networks have their own parameters.All the Hyper-Nets in our method consist of 5 fully connected layers with $relu$ as activation function. The dimension of hidden features is 256 in Hyper-Nets, and is 128 in other modules. The dimension

[*] indicates corresponding author

of latent code α for each shape is 128.

We use the sine activation function proposed by Sitzmann *et al*. [8] for encoder, decoder, SDF module and part probability networks, because it has excellent property of representing complex signal and its derivative [8]. The sine activation function is in form of $f(\mathbf{x}) = sin(\omega \mathbf{x})$, and larger ω usually indicates output with higher frequency. In our experiments, ω is set to 15 in part probability networks, and set to 30 in encoder, decoder and SDF prediction networks.

B. More Details on Loss Functions

In Section 3.4 of our main paper, we follow the idea of L_{sdf} to supervise queried SDF values from template. In the following, we show the detailed formulations of the constraints. The L_{sdf} is used to supervise SDF values $\Phi(\mathbf{p}|\alpha_\mathbf{i})$, while the following constraints are used to supervise SDF values queried from template field $\Phi(D_{i\to tmpl}(\mathbf{p})|\alpha_{tmpl})$.

SDF Regression Constraints. The SDF regression constraints have the similar formulation to L_{sdf}. In order to constrain the queried SDF to have the same sign of the ground truth, we design the constraint for queried SDF value formulated as

$$L_{pbs} = \sum_{\mathbf{p}\in S_i} |\hat{s}(\mathbf{p})|,$$

$$\hat{s}(\mathbf{p}) = \begin{cases} \Phi(D_{i\to tmpl}(\mathbf{p})|\alpha_{tmpl}), \text{ if } \bar{s}\cdot\Phi(D_{i\to tmpl}(\mathbf{p})|\alpha_{tmpl}) \leq 0 \\ 0, \ otherwise \end{cases} \tag{1}$$

where $\bar{s}$ is the ground truth SDF value. The loss weight of L_{pbs} is w_s, which is the same as w_s in the main paper (the first term of Eq. 8).

In order to supervise the normal on a represented shape, we constrain the gradient of the queried SDF field to align with ground truth normal:

$$L_{pbn} = \sum_i \sum_{\mathbf{p}\in S_i^0} (1 - S_c(\nabla_\mathbf{p}\Phi(D_{i\to tmpl}(\mathbf{p})|\alpha_{tmpl}), \bar{\mathbf{n}})), \tag{2}$$

where $\bar{\mathbf{n}}$ is the ground truth normal, and S_c is cosine similarity. Note that the L_{pbn} is different from L_{pfn} (Eq. 9 in the main paper). The loss L_{pbn} supervises normals on the represented shape, while L_{pfn} supervises normals on template, which is proved to be crucial for shape representation learning by Deng *et al*. [4]. The weight of L_{pbn} and L_{pfn} is w_n (which is also the weight of $\sum_{\mathbf{p}\in S_i^0}(1 - S_c(\nabla\Phi(\mathbf{p}|\alpha_i), \bar{\mathbf{n}})$ of Eq. 8 in the main paper).

We also constrain the gradient of template field to satisfy Eikonal equation: $\sum_{\mathbf{p}\in S_{tmpl}} |\|\nabla\Phi(\mathbf{p}|\alpha_{tmpl})\|_2 - 1|$, and apply ρ (the fourth term of Eq. 8 of the main paper) on template field to encourage off-surface points on template to have larger SDF values. Their weights are same as w_{Eik} and w_ρ in the main paper.

Reconstruction Loss. We give the detailed formulation of L_{recon} in Section 3.4 of the main paper as

$$L_{recon} = \sum_i \sum_{\mathbf{p}\in S_i} \|\mathbf{p} - D_{i\to i}(\mathbf{p})\|^2 + \sum_{\mathbf{p}\in S_{tmpl}} \|\mathbf{p} - D_{tmpl\to tmpl}(\mathbf{p})\|^2. \tag{3}$$

The weight of each loss term remains the same across all subjects, specially $w_s = 3\times10^2$, $w_n = 50$, $w_{Eik} = 5$, $w_\rho = 50$, $w_{recon} = 5\times10^3$, $w_{reg} = 1\times10^5$, $w_{lr} = 10$, $w_{nbr} = 5\times10^4$, $w_{pr} = 3\times10^3$.

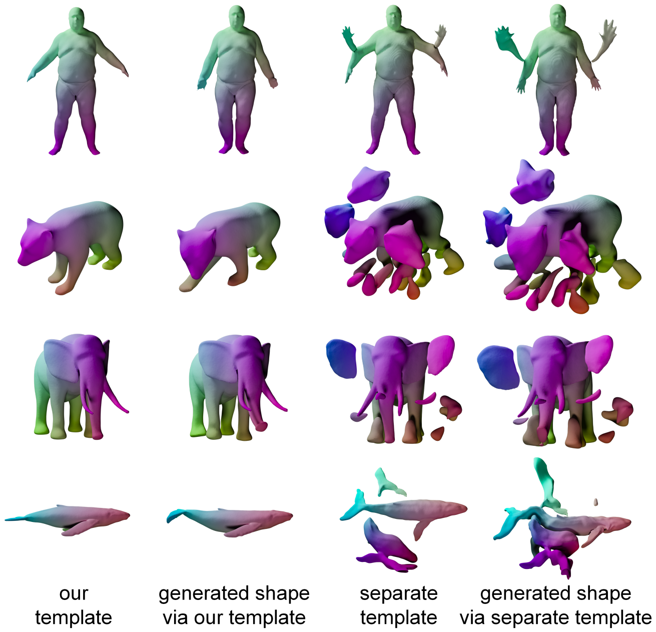
our template
generated shape via our template
separate template
generated shape via separate template

Figure A: More qualitative experiments to demonstrate the ability of our method to learn template. The learned template shapes with the separate template representation are all not reasonable, and our template enables significantly better results of generated shapes than those with the separate template.

C. Additional Experiments

C.1. Ablation Study

Effect of Our Template Representation. In this section, we use the same method as the main paper to further investigate the ability of our novel template shape representation architecture and show more results. Instead of representing the template shape as a latent code, we test an ablation case where a separate network only predicts template SDF like DIF. The architecture of the new template SDF module is the same as the origin SDF prediction module in our main

	w/o ψ_e	w/o SDF input	full model
CD ↓	1.174	0.783	**0.687**
corr ↓	0.0165	0.0265	**0.0141**

Table A: Ablation study on subject 50026 from D-FAUST[3] dataset. "w/o ψ_e" represents the model without part probabilities prediction network ψ_e. "w/o SDF input" represents the input of our encoder only contains coordinate $\mathbf{p}$. It shows that ψ_e can improve the performance on both Chamfer distance and geodesic distance. Using SDF as input of encoder can greatly improve the performance of our method.

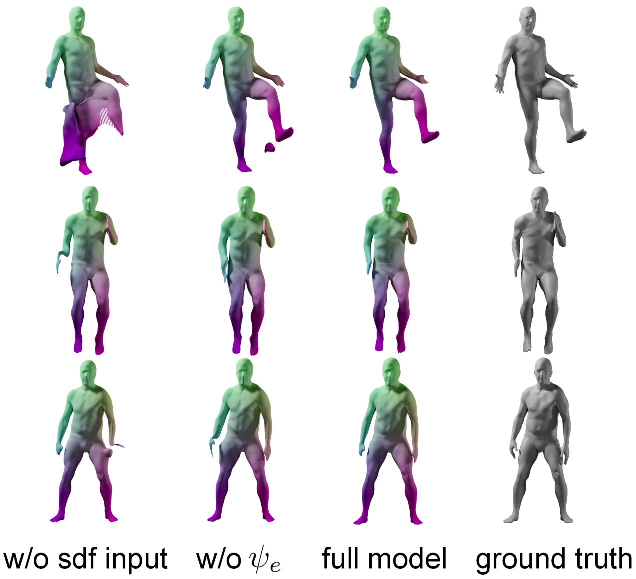

Figure B: Qualitative results of ablation study. Colors indicate the dense correspondences. SDF can provide encoder with geometric clues to predict correspondence. The part prediction network ψ_e is an essential component of the rigid constraint, which helps the network converge.

paper. Other parts of our network and the loss functions remain the same. More results are shown in Figure A. The learned template shapes with the separate template representation are all not reasonable, and our template enables significantly better results of generated shapes than those with the separate template.

Effect of ψ_e in Piece-wise Rigid Constraint. In the main paper, we conduct the experiments to demonstrate the significance of piece-wise rigid constraint. In this ablation, we investigate the effect of the part probabilities predicted by the networks ψ_e (See Section A). Although we cannot attain consistent part segmentation across shapes with ψ_e only, we demonstrate ψ_e plays a crucial role in our method. We compare the reconstruction and correspondence results of our method without ψ_e on subject 50026 in D-FAUST [3] in Table A and Figure B. Subject 50026 is selected for evaluation due to its large range of motion. Experiments show that ψ_e can deal with this challenge and improve the performance of shape reconstruction and correspondence prediction. In contrast, we observe that method without ψ_e cannot converge as well as our full model, especially in the end points of the body with the large range of motion.

Effect of SDF input to Encoder. In order to evaluate the effect of SDF input to encoder, we compare the Chamfer distance of shape reconstruction and geodesic distance of predicted correspondences by removing the SDF input. We also show the results using our method without input SDF on subject 50026 in D-FAUST [3] in Table A and Figure B. We observe that method without SDF as input to encoder fails in some poses and generates shapes with bad geometry.

C.2. More Evaluations on Model Capacity

In this section, we show more model capacity comparisons with DIF [4] and 3D-CODED [5] using reconstructed shapes in the training set. Figure E and Figure F shows that our method outperforms both of them.

C.3. Reconstruction from Full Observation

In this section, we will show more qualitative experiments of our method compared with DIF [4] and 3D-CODED [5]. We generate point cloud of each subject by simulating multiple depth cameras, and then fit our shape representation model by minimizing Eq. 10 in the main paper. Figure G and Figure H show the results of humans and animals. We can observe that our method outperforms DIF and 3D-CODED, and achieves plausible shape reconstruction and correspondence results. Our method can fit shapes with large deformation effectively. Conceptually, 3D-CODED and DIF lack sufficient rigid constraints, so they cannot model subjects with large deformation properly. Although DIF can learn template SDF field, the learned shape is out of the distribution of the training data. Therefore, there are many floating components on the reconstructed shapes.

C.4. Reconstruction from Partial Observation

We generate point cloud of each subject by simulating a single depth camera, and then fit our representation model by minimizing Eq. 10 in the main paper. Figure I and Figure J show the qualitative experiments of shape reconstruction from partial point cloud. Our model can reconstruct shapes from partial point cloud while 3D-CODED [5] fails. Therefore, we only compare with DIF [4]. Results show that our method outperforms DIF by a large margin for partial point clouds.

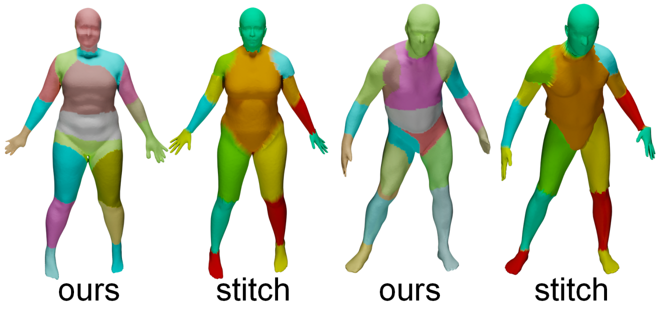

Figure C: Comparison with the stitched puppet [13].

C.5. Comparison with LoopReg

In this section, we compare our method with LoopReg [2] on training set. LoopReg creates a self-supervised loop to register a corpus of scans to a common 3D human model (*i.e.*, SMPL [7]), which can model correspondences between human pairs. As shown in Table B, our method outperforms LoopReg on both IoU and *corr*.

	LoopReg	Our method
IoU ↑	0.726	**0.881**
corr ↓	0.1087	**0.0304**

Table B: Capacity evaluation on D-FAUST with LoopReg [2]

C.6. Qualitative Experiment on Part Segmentation

In this section, we compare our method with the stitched puppet [13]. The stitched puppet [13] is a shape representation method that manually segments the represented shape into multiple parts and combines the parts into human body shapes with different poses. As shown in Figure C, our self-supervised method achieves comparable results.

C.7. Failure Cases

We show several failure cases in Figure D where floating components make Chamfer distance increase. Although these shapes have fine human surface geometry, they have large Chamfer distance because of the floating components far from the body.

D. Details on Rigid Constraint

In this section, we will give details on the closed-form solution of piece-wise rigid constraint in Section 3.3 of the main paper. Then, we will give theoretic analysis on our local rigid constraint, in which we elaborate on the relationship between the proposed constraint with implicit representation in Section 3.2 of the main paper and the traditional As-Rigid-As-Possible loss [11] that was originally proposed for discrete mesh deformation.

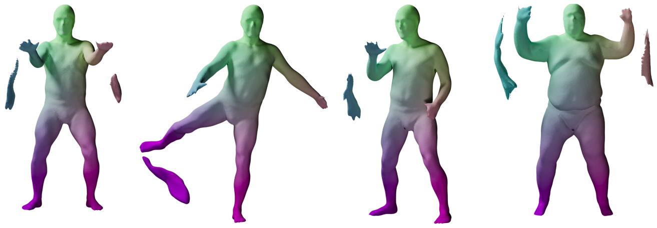

Figure D: Several failure cases where floating components make Chamfer distance increase.

D.1. Closed-Form Solution of Piece-wise Rigid Constraint

We follow Sorkine-Hornun *et al.* [11] to give a closed-form solution of the minimal rigid transformation error of our piece-wise rigid constraint L_{pr} (Eq. 6 of our main paper). Detailed proof can be found in [11]. In this section, we use the same notions as [11] for easy understanding.

Denote $\mathcal{P} = \{\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_n\}$ and $\mathcal{Q} = \{\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_n\}$ to be corresponding points in $\mathbb{R}^d$. Therefore, the optimal rigid transformation $(\mathbf{R}, \mathbf{t})$ between $\mathcal{P}$ and $\mathcal{Q}$ can be estimated by minimizing the following alignment error as

$$\begin{aligned} L &= \min_{\mathbf{R},\mathbf{t}} F(\mathbf{R}, \mathbf{t}) \\ F &= \sum_{i=1}^{n} w_i \|\mathbf{R}\mathbf{p}_i + \mathbf{t} - \mathbf{q}_i\|^2, \end{aligned} \tag{4}$$

where $\mathbf{R} \in SO(d)$ is rotation matrix and $\mathbf{t} \in \mathbb{R}^d$ is translation.

First, Sorkine-Hornun *et al.* [11] proved that the optimal translation $\mathbf{t}$ can be expressed as

$$\mathbf{t} = \bar{\mathbf{q}} - \mathbf{R}\bar{\mathbf{p}}, \tag{5}$$

where $\bar{\mathbf{q}}$ and $\bar{\mathbf{p}}$ are the centroid of $\mathcal{Q}$ and $\mathcal{P}$

$$\bar{\mathbf{p}} = \frac{\sum_{i=1}^{n} w_i \mathbf{p}_i}{\sum_{i=1}^{n} w_i}, \quad \bar{\mathbf{q}} = \frac{\sum_{i=1}^{n} w_i \mathbf{q}_i}{\sum_{i=1}^{n} w_i}. \tag{6}$$

Incorporate the optimal $\mathbf{t}$ into Eq. 4, and then we get the loss function F as

$$F = \sum_{i=1}^{n} w_i \|\mathbf{R}(\mathbf{p}_i - \bar{\mathbf{p}}) - (\mathbf{q}_i - \bar{\mathbf{q}})\|^2. \tag{7}$$

Giving the definitions as follows

$$\mathbf{x}_i := \mathbf{p}_i - \bar{\mathbf{p}}, \quad \mathbf{y}_i := \mathbf{q}_i - \bar{\mathbf{q}}, \tag{8}$$

we can set the translation $\mathbf{t}$ to be zero, and then focus on the estimation of $\mathbf{R}$ by optimizing the following equivalent loss function

$$L = \min_{\mathbf{R}} \sum_{i=1}^{n} w_i \|\mathbf{R}\mathbf{x}_i - \mathbf{y}_i\|^2. \tag{9}$$

Denote $\mathbf{W} = diag(w_1, ..., w_n)$, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]$, $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n]$. Then the loss function L can be effectively calculated with the following closed-form solution

$$L = \sum_{i=1}^{n} w_i(\|\mathbf{x}_i\|^2 + \|\mathbf{y}_i\|^2) - 2S_\sigma(\mathbf{X}\mathbf{W}\mathbf{Y}^T)$$
$$S_\sigma = \begin{cases} \sigma_1 + \sigma_2 + ... + \sigma_{d-1} + \sigma_d, \ if \ det(\mathbf{U}\mathbf{V}^T) = 1 \\ \sigma_1 + \sigma_2 + ... + \sigma_{d-1} - \sigma_d, \ if \ det(\mathbf{U}\mathbf{V}^T) = -1, \end{cases} \tag{10}$$

where $\mathbf{U}$, $\mathbf{V}$ are the left and the right singular matrices of $\mathbf{X}\mathbf{W}\mathbf{Y}^T$, and $\{\sigma_i\}$ are singular value of $\mathbf{X}\mathbf{W}\mathbf{Y}^T$ in descending order. Moreover, the gradient of S_σ is a rotation matrix, which does not contain large value and enables stable learning process.

In order to calculate the 3D alignment error L, we only need several efficient operations, such as solving SVD of 3×3 square matrix $\mathbf{X}\mathbf{W}\mathbf{Y}^T$ and conducting point-wise additions and multiplications.

In our problem, $\mathcal{P}$ and $\mathcal{Q}$ are consisting of the points in the target space $\{\mathbf{p}\}$ and the correspondence $\{D_{i\rightarrow tmpl}(\mathbf{p})\}$ in the template space, respectively. Therefore, the piecewise rigid loss L_{pr} (Eq. 6 of our main paper) can be expressed in closed-form as

$$\min_{\mathbf{R}_h, \mathbf{t}_h} \sum_{\mathbf{p}\in S_i^0 \cup S_i^-} \mathbf{P}_h(\mathbf{p})\|(\mathbf{R}_h\mathbf{p} + \mathbf{t}_h) - D_{i\rightarrow tmpl}(\mathbf{p})\|_2^2$$
$$= \sum_{\mathbf{p}\in S_i^0 \cup S_i^-} \mathbf{P}_h(\mathbf{p})(\|\mathbf{x}\|_2 + \|\mathbf{x}_{i\rightarrow tmpl}\|_2) - 2S_\sigma(\mathbf{X}\mathbf{W}_h\mathbf{X}^T_{i\rightarrow tmpl}) \tag{11}$$

where $\mathbf{x}$ and $\mathbf{x}_{i\rightarrow tmpl}$ are the points in the target space and their correspondence in template after removing their respective centroid as

$$\begin{aligned} \mathbf{x} &= \mathbf{p} - \bar{\mathbf{p}} \\ \mathbf{x}_{i\rightarrow tmpl} &= D_{i\rightarrow tmpl}(\mathbf{p}) - \bar{D}_{i\rightarrow tmpl}(\mathbf{p}), \end{aligned} \tag{12}$$

and $\bar{\mathbf{p}}$ and $\bar{D}_{i\rightarrow tmpl}(\mathbf{p})$ are their respective centroid as Eq. 6. The j-th column of $\mathbf{X} \in \mathbb{R}^{3\times n}$ is a $\mathbf{x}$ derived from the j-th point $\mathbf{p}$, each column of $\mathbf{X}_{i\rightarrow tmpl} \in \mathbb{R}^{3\times n}$ is the correspondence $\mathbf{x}_{i\rightarrow tmpl}$ of the j-th point $\mathbf{p}$, $\mathbf{P}_h(\mathbf{p})$ is the probability that point $\mathbf{p}$ belongs to h-th part, and $\mathbf{W}_h$ is a $n \times n$ diagonal matrix, its j-th diagonal element is the predicted part probability $\mathbf{P}_h$ of the j-th point $\mathbf{p}$.

D.2. Analysis on Local Rigid Constraint

D.2.1 Analysis on Least Square Solution of Rotation

In this section, we give further analysis on the formulation of closest rotation matrix of $J(D_{i\rightarrow tmpl})$. With singular value decomposition (SVD), we get $J(D_{i\rightarrow tmpl}) = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. Given the properties of the determinant, we know that $det(J(D_{i\rightarrow tmpl})) = det(\mathbf{U})det(\mathbf{\Sigma})det(\mathbf{V}^T)$. According to the definition of the singular value, the singular values of $J(D_{i\rightarrow tmpl})$ (*i.e.* diagonal items of $\mathbf{\Sigma}$) are always positive. Therefore, $det(J(D_{i\rightarrow tmpl}))$ has the same sign as $det(\mathbf{U})det(\mathbf{V}^T)$, *i.e.* $det(\mathbf{U}\mathbf{V}^T)$. When $det(J(D_{i\rightarrow tmpl})) < 0$, its closest orthogonal matrix $\mathbf{U}\mathbf{V}^T$ has negative determinant. However, a rotation matrix must have positive determinant. To this end, previous method [12] figured out the closet rotation that has positive determinant. $\mathbf{R} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ ($\mathbf{R}$ have positive determinant) of $J(D_{i\rightarrow tmpl})$ with a diagonal matrix $\mathbf{S} = diag(1, 1, det(\mathbf{U}\mathbf{V}^T))$.

D.2.2 Equivalence between Local Rigid Constraint and ARAP

In this section, we will prove that our implicit local rigid constraint is equivalent to traditional As-Rigid-As-Possible (ARAP) loss in infinite small scope. ARAP loss [10] is generally defined on discrete representations such as mesh, while we find that with the closed form of alignment error Eq. 10 ARAP loss can be extended to continuous implicit representation for infinite small scope.

According to Sorkine *et al.* [10], ARAP loss on mesh is defined as

$$E = \min_{\mathbf{R}} \sum_{j\in\mathcal{N}(i)} w_{ij}\|(\mathbf{p}'_i - \mathbf{p}'_j) - \mathbf{R}_i(\mathbf{p}_i - \mathbf{p}_j)\|^2. \tag{13}$$

In our method, we represent the shape as implicit field instead of mesh in original ARAP [10], so there is not explicit adjacency relation for our shape representation. It is barely addressed and highly challenging to constrain ARAP in the continuous implicit shape representation.

For a sampling point $\mathbf{p}$, we assume the adjacent points of $\mathbf{p}$ are uniformly distributed on the surface of a sphere centered at $\mathbf{p}$, which can be formulated as $\mathbf{p} + \omega s$, where ω is an arbitrary unit vector. For simplicity, we consider w_{ij} as 1.

Considering the adjacent points within the infinitely-small volume, we denote adjacent points as

$$\mathbf{p} + \omega ds, \tag{14}$$

where ds is infinitely small length.

Denote the mapping from $\mathbf{p}$ to $\mathbf{p}'$ as $D(\mathbf{p})$ and $\frac{\partial \mathbf{p}'}{\partial \mathbf{p}^T}$, *i.e.* $J(D_{i\rightarrow tmpl})(\mathbf{p})$, as J. In our case, $\mathbf{p}$ is in the target shape space and $\mathbf{p}'$ is in the template shape space. Then we have the following equation by Taylor expansion

$$D(\mathbf{p} + \omega ds) = D(\mathbf{p}) + J\omega ds + o(ds). \tag{15}$$

According to Eq. 10, we can also get a closed-form solution for ARAP loss in Eq. 13. Because ω is evenly distributed on the sphere, $\sum_\omega \omega = \mathbf{0}$, the centroid of $\mathbf{p} + \omega ds$

is $\mathbf{p}$ and the centroid of $D(\mathbf{p}) + J\omega ds$ is $D(\mathbf{p})$. After incorporating $\mathbf{p}+\omega ds$ and $D(\mathbf{p}) + J\omega ds$ and their centroids into Eq. 8, we get $\mathbf{x}_i = \omega ds$ and $\mathbf{y}_i = J\omega ds + o(ds)$, and incorporate $\mathbf{x}_i$ and $\mathbf{y}_i$ into Eq. 10, then we can get the following equation by ignoring the infinitesimal of higher order

$$\begin{aligned} E =& \sum_{\omega} \|\omega ds\|^2 + \|J\omega ds\|^2 - 2S_\sigma(\sum_{\omega} \omega\omega^T J^T ds^2) \\ =& ds^2 \sum_{\omega}(\|\omega\|^2 + \|J\omega\|^2) - 2ds^2 S_\sigma(\sum_{\omega} \omega\omega^T J^T). \end{aligned} \tag{16}$$

Since ω is uniformly distributed, we use integration instead of summation.

$$\begin{aligned} E = ds^2(\int_{\mathcal{S}^2} \|\omega\|^2 d\omega +& \int_{\mathcal{S}^2} \|J\omega\|^2 d\omega \\ & -2S_\sigma(\int_{\mathcal{S}^2} \omega\omega^T J^T d\omega)), \end{aligned} \tag{17}$$

where S^2 represents the surface of a unit sphere embedded in the 3-dimensional space, and each term will be analyzed in the following part.

Since $\|\omega\|^2 = 1$, the first term can be easily calculated as the area of the sphere, *i.e.* 4π.

Then the second term can be simplified as

$$\begin{aligned} & \int_{\mathcal{S}^2} \|J\omega\|^2 d\omega = \int_{\mathcal{S}^2} \omega^T J^T J \omega d\omega \\ =& \int_{\mathcal{S}^2} tr(\omega^T J^T J \omega) d\omega = \int_{\mathcal{S}^2} tr(\omega\omega^T J^T J) d\omega \\ =& \, tr\left(\int_{\mathcal{S}^2} J^T J \omega\omega^T d\omega\right) = tr\left(J^T J \int_{\mathcal{S}^2} \omega\omega^T d\omega\right). \end{aligned} \tag{18}$$

To solve the above function, we need to know the result of $\int_{\mathcal{S}^2} \omega\omega^T d\omega$. We use spherical coordinates to calculate the integration. $\omega = (sin\theta cos\phi, sin\theta sin\phi, cos\theta)^T = (sin\theta cos\phi, sin\theta sin\phi, 0)^T + (0, 0, cos\theta)^T$

$$\begin{aligned} & \int_{\mathcal{S}^2} \omega\omega^T d\omega = \\ & \int_0^\pi \int_0^{2\pi} sin\theta \left(\begin{pmatrix} 0 \\ 0 \\ cos\theta \end{pmatrix} + \begin{pmatrix} sin\theta cos\phi \\ sin\theta sin\phi \\ 0 \end{pmatrix} \right) \\ & \qquad \left(\begin{pmatrix} 0 \\ 0 \\ cos\theta \end{pmatrix} + \begin{pmatrix} sin\theta cos\phi \\ sin\theta sin\phi \\ 0 \end{pmatrix} \right)^T d\phi d\theta \end{aligned} \tag{19}$$

Consider $\int_0^{2\pi} sin\phi \; d\phi = 0$ and $\int_0^{2\pi} cos\phi \; d\phi = 0$:

$$\begin{aligned} =& \int_0^\pi \int_0^{2\pi} sin\theta \Bigg(\begin{pmatrix} 0 \\ 0 \\ cos\theta \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ cos\theta \end{pmatrix}^T \\ & \qquad + \begin{pmatrix} sin\theta cos\phi \\ sin\theta sin\phi \\ 0 \end{pmatrix} \begin{pmatrix} sin\theta cos\phi \\ sin\theta sin\phi \\ 0 \end{pmatrix}^T \Bigg) d\phi d\theta \\ =& \int_0^\pi \int_0^{2\pi} sin\theta \begin{pmatrix} sin^2\theta cos^2\phi & \frac{sin^2\theta sin2\phi}{2} & 0 \\ \frac{sin^2\theta sin2\phi}{2} & sin^2\theta sin^2\phi & 0 \\ 0 & 0 & cos^2\theta \end{pmatrix} d\phi d\theta \end{aligned} \tag{20}$$

Consider the periodicity of $\int_0^{2\pi} sin2\phi \; d\phi = 0$:

$$\begin{aligned} =& \int_0^\pi \int_0^{2\pi} sin\theta \begin{pmatrix} sin^2\theta cos^2\phi & 0 & 0 \\ 0 & sin^2\theta sin^2\phi & 0 \\ 0 & 0 & cos^2\theta \end{pmatrix} d\phi d\theta \\ =& \begin{pmatrix} \frac{4\pi}{3} & 0 & 0 \\ 0 & \frac{4\pi}{3} & 0 \\ 0 & 0 & \frac{4\pi}{3} \end{pmatrix}. \end{aligned} \tag{21}$$

Denote singular value decomposition (SVD) of J to be $J = \mathbf{U\Sigma V}^T$. Then Eq. 18 can be calculated as

$$\begin{aligned} & tr(J^T J \int_{\mathcal{S}^2} \omega\omega^T d\omega) = tr(J^T J \mathbf{I} \frac{4\pi}{3}) \\ =& \frac{4\pi}{3} tr(J^T J) = \frac{4\pi}{3} tr(\mathbf{V\Sigma U}^T \mathbf{U\Sigma V}^T) \\ =& \frac{4\pi}{3} tr(\mathbf{V\Sigma\Sigma V}^T) = \frac{4\pi}{3} tr(\mathbf{V}^T \mathbf{V\Sigma\Sigma}) \\ =& \frac{4\pi}{3} tr(\mathbf{\Sigma}^2) = \frac{4\pi}{3}(\sigma_1^2 + \sigma_2^2 + \sigma_3^2), \end{aligned} \tag{22}$$

where σ_1, σ_2, σ_3 are singular values in descending order.

With the above result of $\int_{\mathcal{S}^2} \omega\omega^T d\omega$, the third term of Eq. 17 can be calculated as

$$\begin{aligned} & S_\sigma\left(\int_{\mathcal{S}^2} \omega\omega^T J^T d\omega\right) \\ =& S_\sigma\left(\int_{\mathcal{S}^2} (\omega\omega^T J^T)^T d\omega\right) \\ =& S_\sigma\left(J \int_{\mathcal{S}^2} \omega\omega^T d\omega\right) = S_\sigma\left(\frac{4\pi}{3} J\right) \\ =& \frac{4\pi}{3}(\sigma_1 + \sigma_2 + det(\mathbf{UV^T})\sigma_3). \end{aligned} \tag{23}$$

We can simplify Eq. 17 with the above results of its three items as

$$\begin{aligned} E =& \frac{4\pi}{3} ds^2 (3 + \sigma_1^2 + \sigma_2^2 + \sigma_3^2 - 2(\sigma_1 + \sigma_2 + det(\mathbf{UV^T})\sigma_3)) \\ =& \frac{4\pi}{3} ds^2 ((\sigma_1 - 1)^2 + (\sigma_2 - 1)^2) + (\sigma_3 - det(\mathbf{UV^T}))^2). \end{aligned} \tag{24}$$

Our L_{arap} in Section 3.2 of our main paper has the following formulation

$$\begin{aligned} L_{arap} &= smoothL1(\sigma_1, 1) + smoothL1(\sigma_2, 1) \\ &+ smoothL1(\sigma_3, det(\mathbf{U}\mathbf{V}^T)) \end{aligned} \tag{25}$$

If σ_1, σ_2 and σ_3 are close to 1 and $\mathbf{U}\mathbf{V}^T = 1$, the smoothL1 loss becomes L2 loss. By ignoring the scale term, L_{arap} has the same form as Eq. 24.

Therefore, our implicit local rigid constraint is equivalent to traditional As-Rigid-As-Possible (ARAP) loss in infinite small scope.

References

[1] Matan Atzmon and Yaron Lipman. Sal: Sign agnostic learning of shapes from raw data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2565–2574, 2020.

[2] 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. *Advances in Neural Information Processing Systems*, 33:12909–12922, 2020.

[3] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Dynamic FAUST: Registering human bodies in motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[4] Yu Deng, Jiaolong Yang, and Xin Tong. Deformed implicit field: Modeling 3d shapes with learned dense correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10286–10296, June 2021.

[5] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan Russell, and Mathieu Aubry. 3d-coded : 3d correspondences by deep deformation. In *The European Conference on Computer Vision (ECCV)*, 2018.

[6] P. Liu, K. Zhang, Tateo D., Jauhri S., Peters J., and Chalvatzaki G. and. Regularized deep signed distance fields for reactive motion generation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022.

[7] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):1–16, 2015.

[8] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020.

[9] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019.

[10] Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In *Symposium on Geometry Processing*, volume 4, pages 109–116, 2007.

[11] Olga Sorkine-Hornung and Michael Rabinovich. Least-squares rigid motion using svd. *Computing*, 1(1):1–5, 2017.

[12] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(04):376–380, 1991.

[13] Silvia Zuffi and Michael J Black. The stitched puppet: A graphical model of 3d human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3537–3546, 2015.

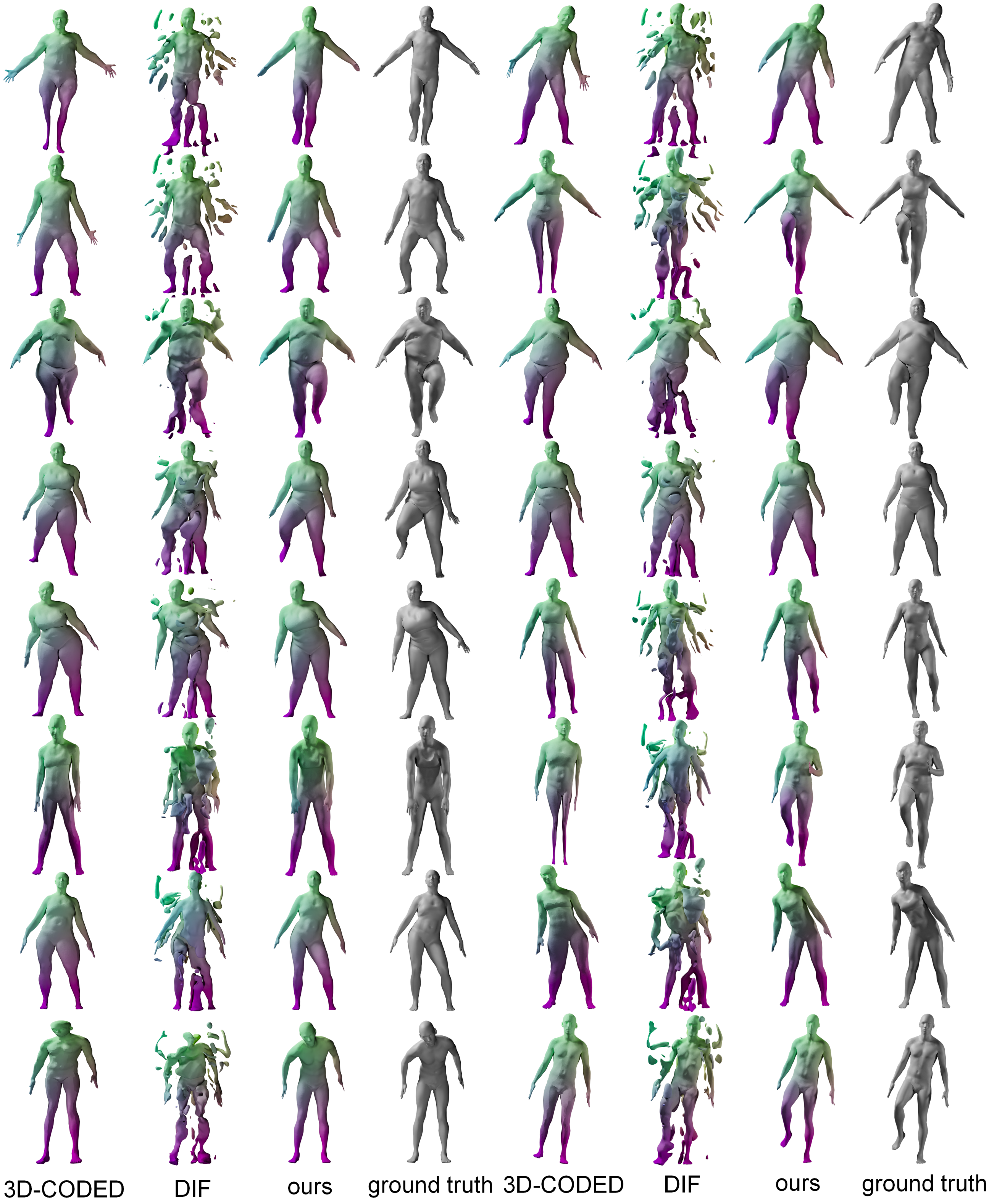

Figure E: Reconstruction from training set of humans. We compare our method with DIF [4] and 3D-CODED [5]. Our method reconstructs shapes with multiple poses and large deformations. The characteristics of each subject is represented well.

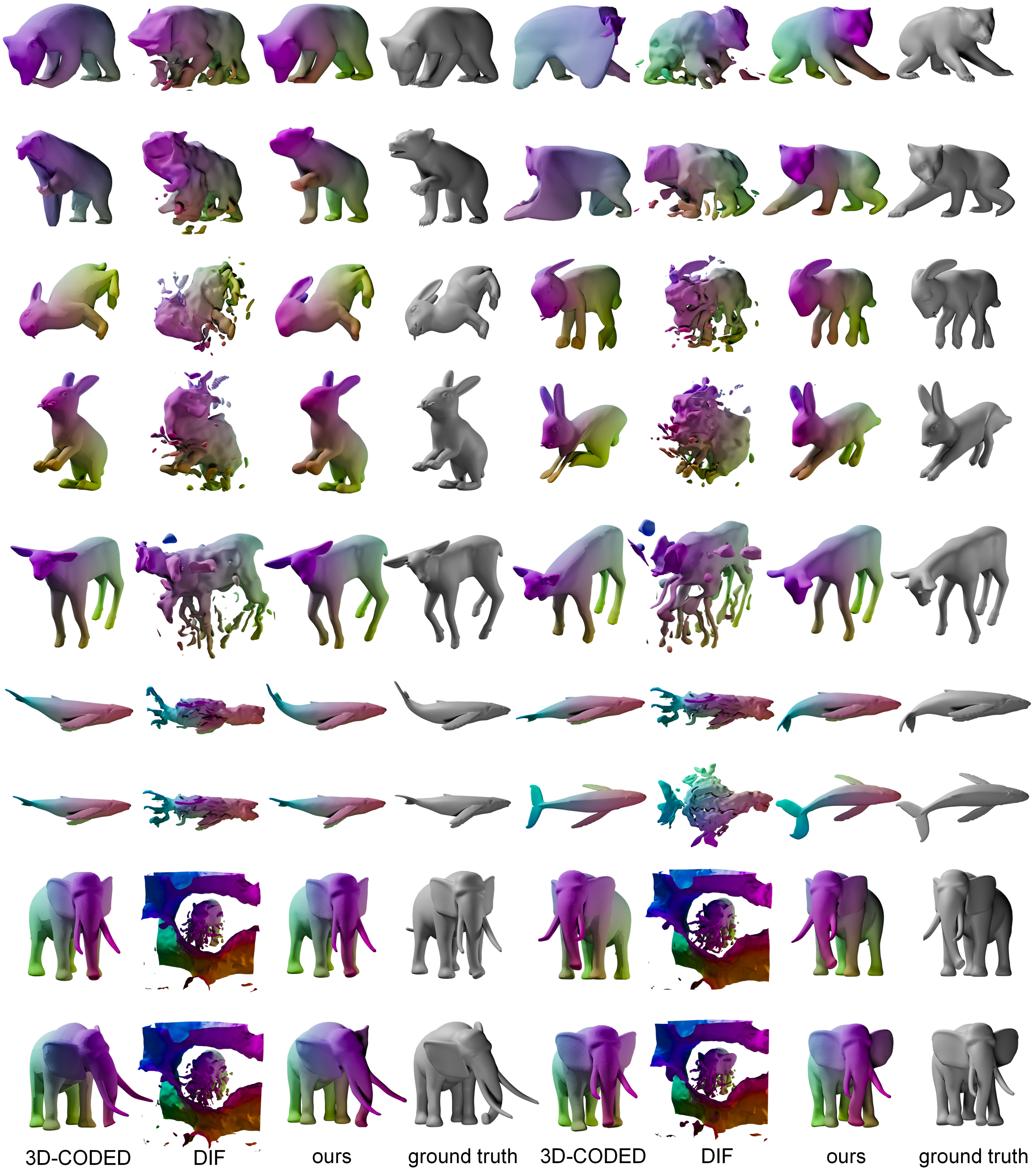

Figure F: Reconstruction from training set of animals. We compare our method with DIF [4] and 3D-CODED [5]. Our method reconstructs shapes with multiple poses and large deformations. The characteristics of each subject is represented well.

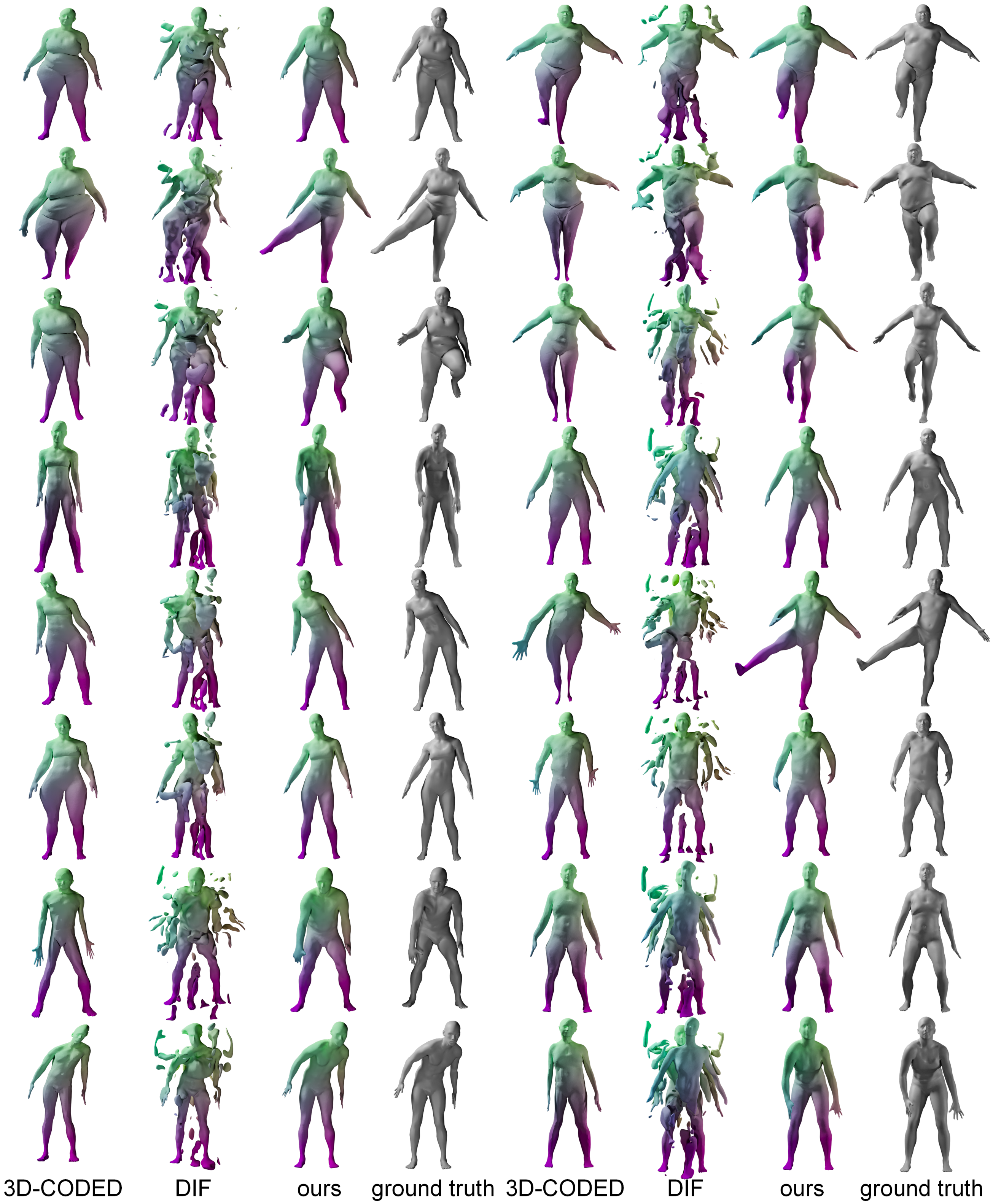

Figure G: Reconstruction from full observation of humans. We compare our method with DIF [4] and 3D-CODED [5]. Our method achieves plausible shape reconstructions and can predict correspondence across shapes.

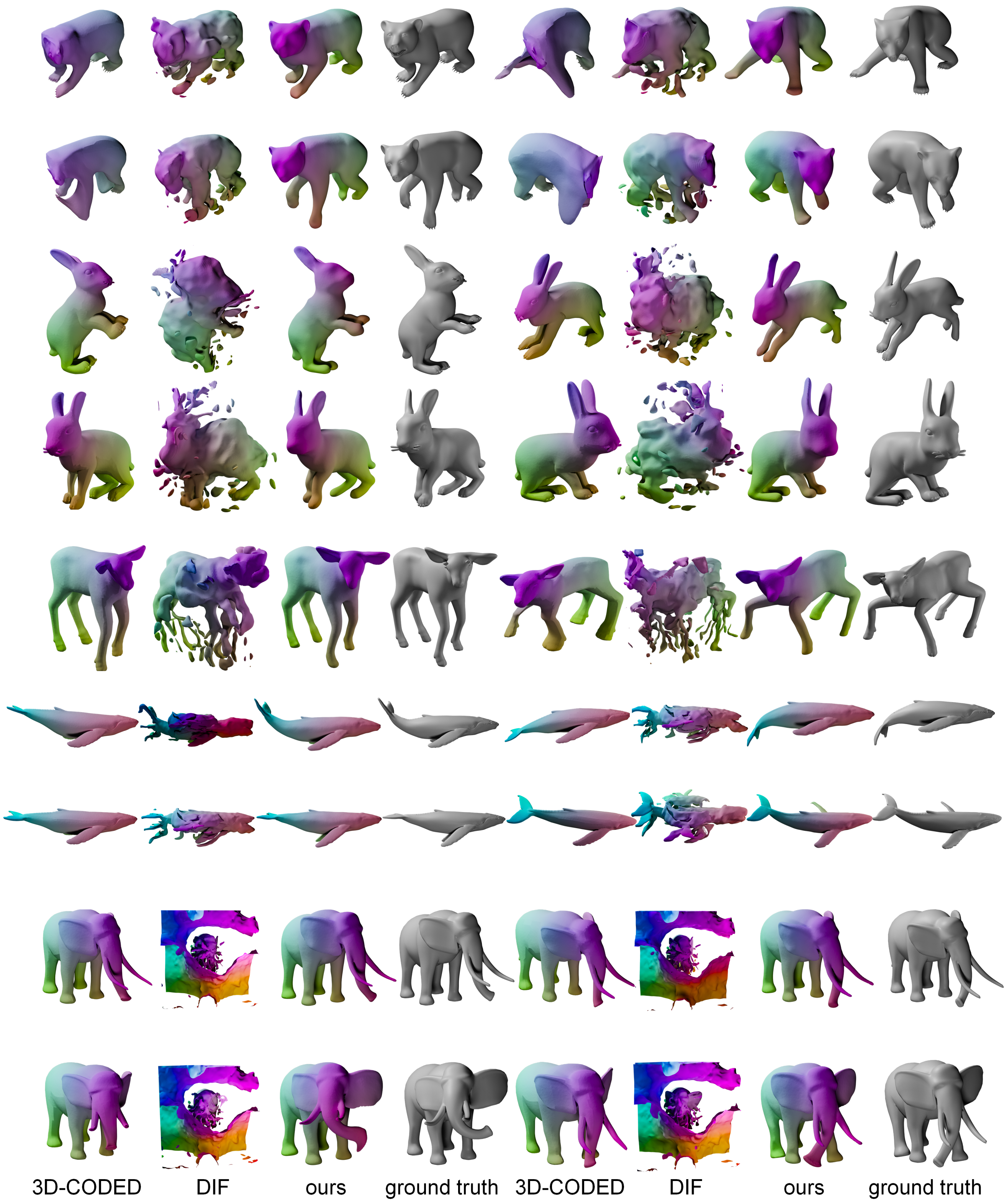

Figure H: Reconstruction from full observation of animals. We compare our method with DIF [4] and 3D-CODED [5]. Our method achieves plausible shape reconstructions and can predict reliable correspondence across shapes.

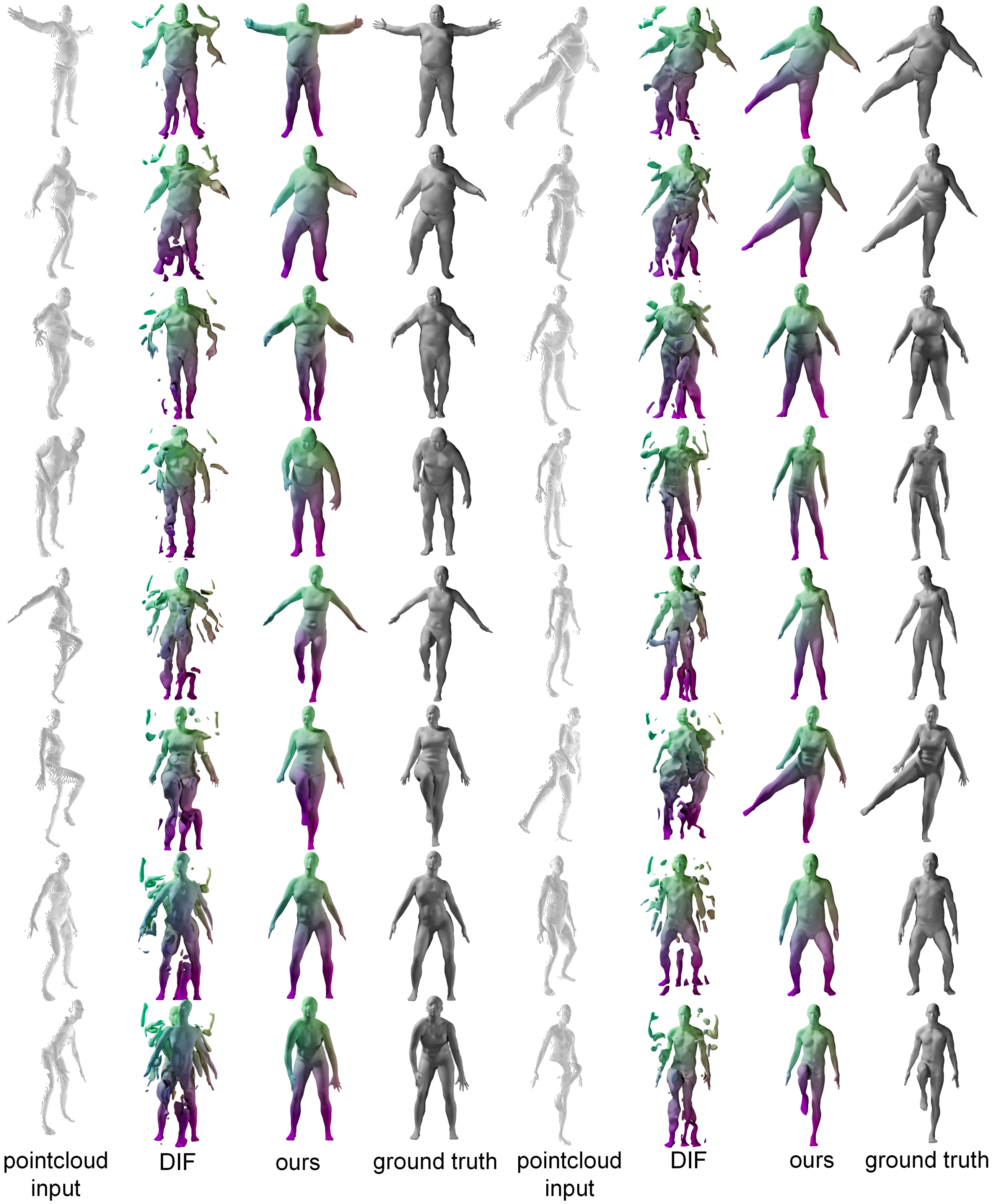

Figure I: Reconstruction from partial observation of humans. We compare our method with DIF [4]. Our method reconstructs shapes with multiple poses and large deformations. The characteristics of each subject is represented well.

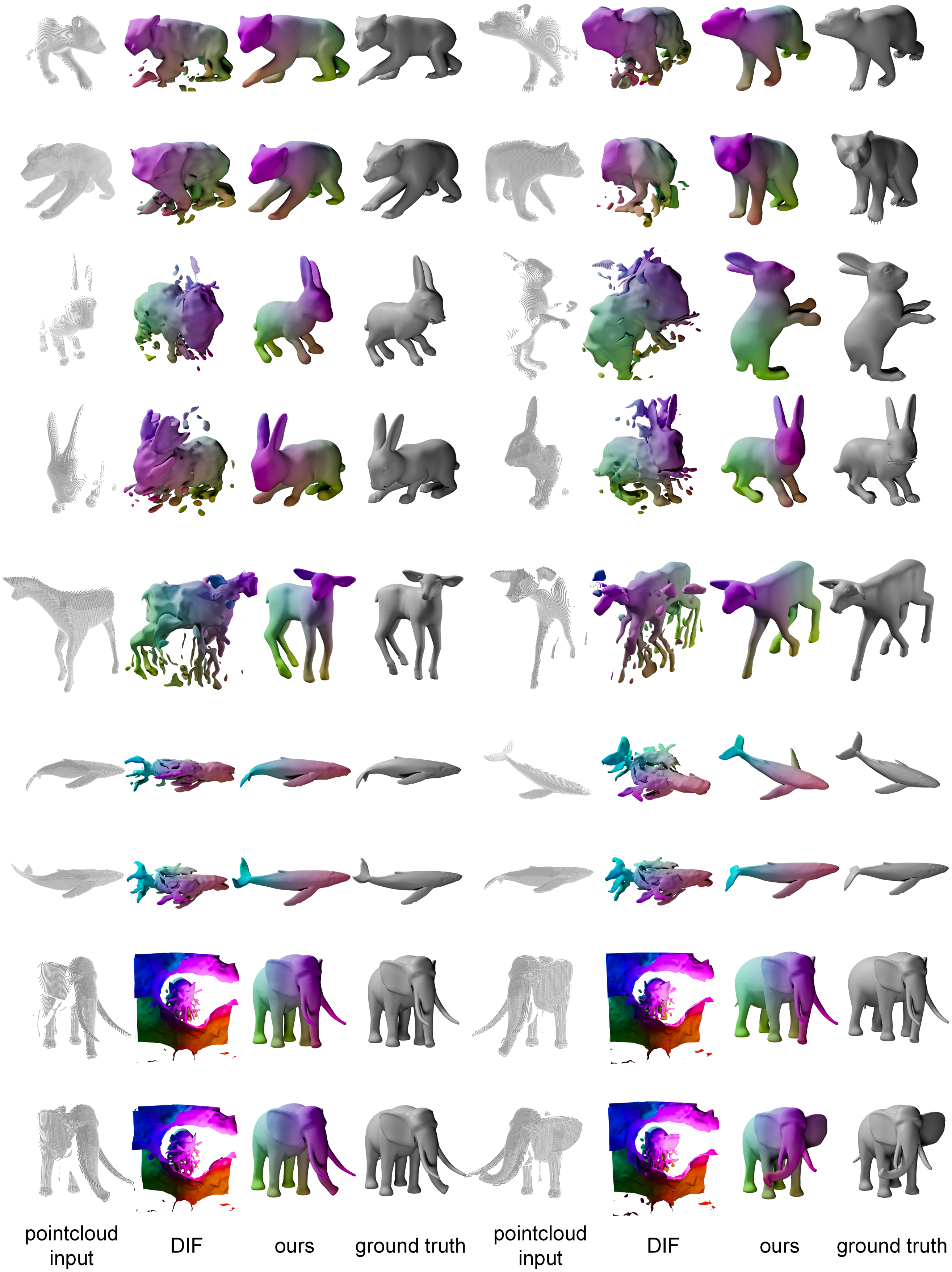

Figure J: Reconstruction from partial observation of animals. We compare our method with DIF [4]. Our method reconstructs shapes with multiple poses and large deformations. The characteristics of each subject is represented well.