Dynamic Neural Surfaces for Elastic 4D Shape Representation and Analysis

Supplementary Material

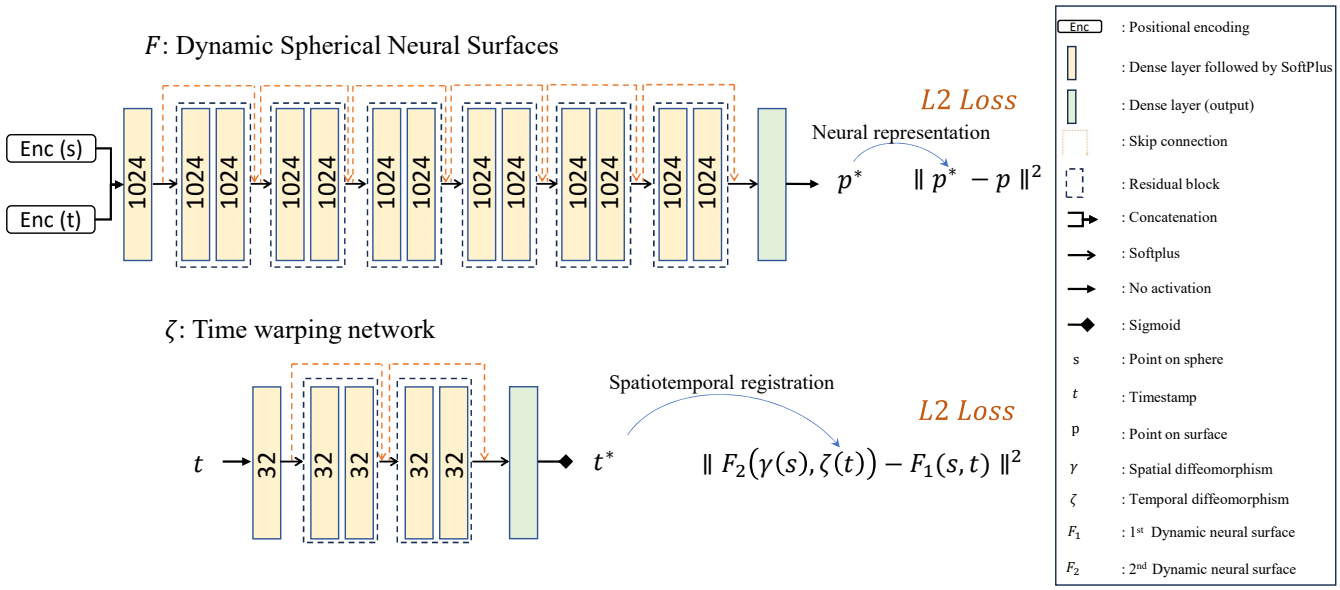


Figure 11. Detailed architecture of the neural networks used for the Dynamic Spherical Neural Surfaces (D-SNS) and for the time warping.

The Supplementary Material is organized as follows; Section A discusses the limitations and potential directions for future work. Section B describes the implementation details of our neural framework. Section C includes details on the dataset. Section D presents additional results of our neural framework.

A. Limitations and future work

Higher genus surfaces. D-SNS is based on spherical parameterization. Thus, it is limited to closed genus-0 surfaces. The representation, however, can be easily extended to open genus-0 surfaces, which can be parametrized on open domains such as a disk. Extending the representation to higher genus surfaces would require a complex parameterization, *e.g.,* using charts or even a volumetric domain. This will be investigated in future work.

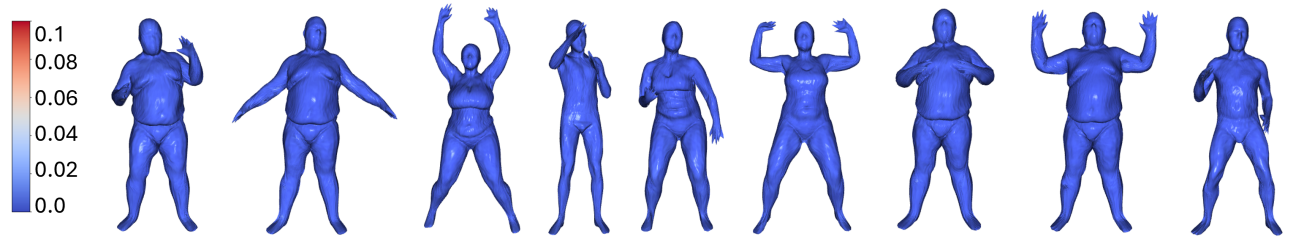
Computation efficiency. A key limitation of our neural framework is its high computation time, particularly when training individual D-SNS. Although the representation provides a continuous representation of surfaces, and thus all the differential properties can be computed analytically, the D-SNS needs to be fitted to every single 4D surface. Thus, it is computationally expensive when analyzing a large number of 4D surfaces. We plan in the future to explore shape-agnostic representations, *e.g.,* by following an approach similar to DeepSDF.

B. Implementation details

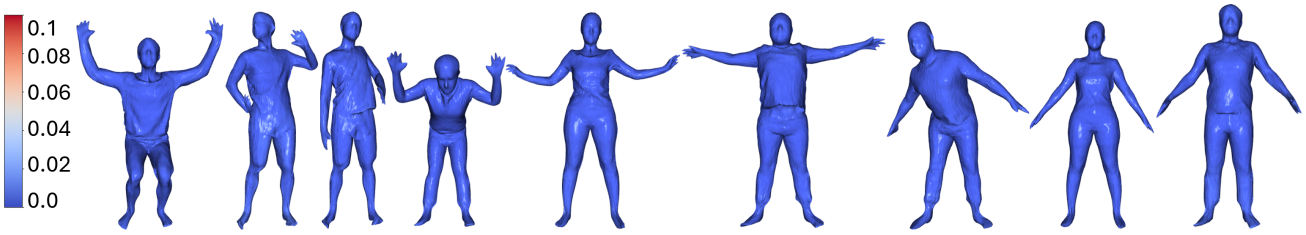
B.1. D-SNS network

We employ a Multi-Layer Perceptron (MLP) composed of six residual blocks. Each block consists of two layers of 1024 neurons each. We use positional encoding of both space and time. The output layer of each block uses SoftPlus as activation function to represent smooth and continuous 4D surfaces. Figure 11 summarizes the detailed architecture.

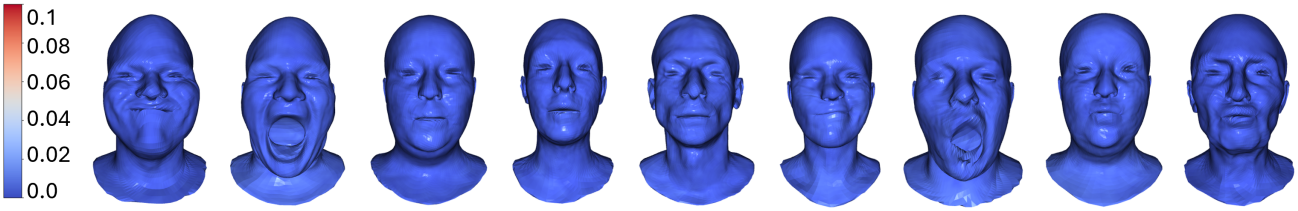
Training. We learn a continuous representation F of a discrete 4D surface using D-SNS. We first spherically parameterized the 4D surfaces, which consist of a set of triangular meshes, with the approach of [14]. We then map the mesh sequences to a temporal domain and allocate a time value in the range of $[0, 1]$. Next, we train D-SNS for this discrete 4D surface by defining a batch size of $80,000$ surface points, which are randomly selected. For each point p we have its associated point on sphere s and a time instance t. We then parse this batch to the D-SNS network, which outputs the predicted points on the surface p^*. We minimize the $\mathbb{L}^2$ loss between the D-SNS represented points p^* and



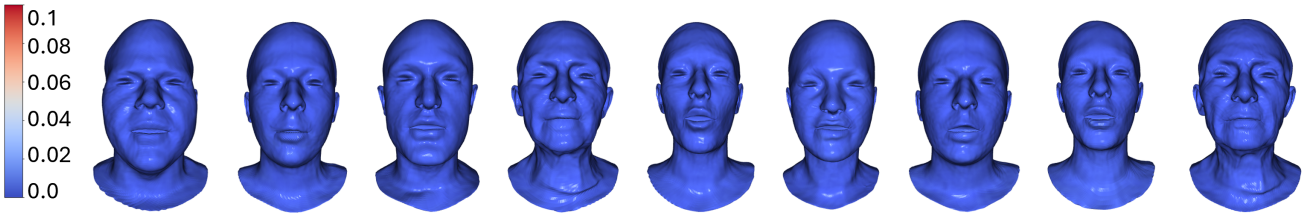
(a) DFAUST [5].



(b) CAPE [20].



(c) COMA [29].



(d) VOCA [9].

Figure 12. We measure the pointwise error between the proposed D-SNS and the ground truth discrete 4D surfaces. Here, we show some time frames with the error plotted as a heatmap. Observe that the proposed neural representation can accurately represent 4D surfaces, with a pointwise error that is less than 0.01%.

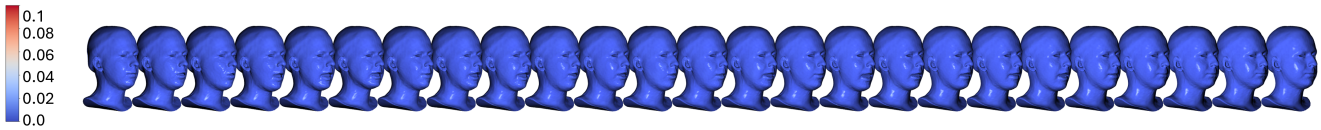
the discrete points p. We have noticed that using the random sampling of surface points from the meshes helps our D-SNS network to converge faster. It also results in smooth and continuous 4D surfaces.

B.2. Spatial diffeomorphism

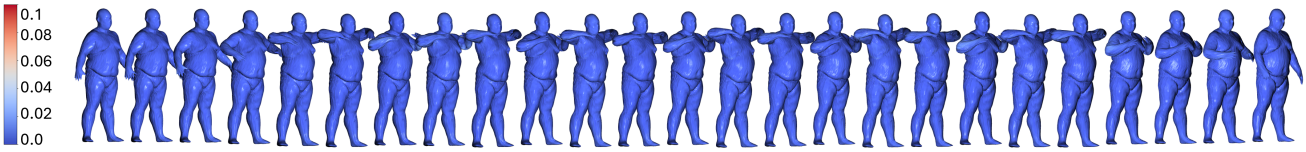
As discussed in Section 3.2.1, we select 3D instances f_1, f_2 from their 4D surfaces F_1, F_2. We then find the optimal rotation $O \in SO(3)$ and diffeomorsphism $\gamma \in \Gamma$ such that when $O(f_2 \circ \gamma)$ is spatially register f_2 onto f_1.

We use a gradient descent-based optimization method that finds the optimal diffeomorphism γ^* and rotation O^* in the SRNF space. As discussed in our approach, we apply the spatial registration framework directly to the neural functions. We keep optimizing for diffeomorphism γ using a weighted sum of spherical harmonic basis and rotation O using Singular Value Decomposition (SVD). We apply these on the unit sphere and parse the reparameterized unit sphere to the D-SNS network, which results in a spatially registered neural function f_2 that is as close as possible to f_1.

(a) COMA [29].

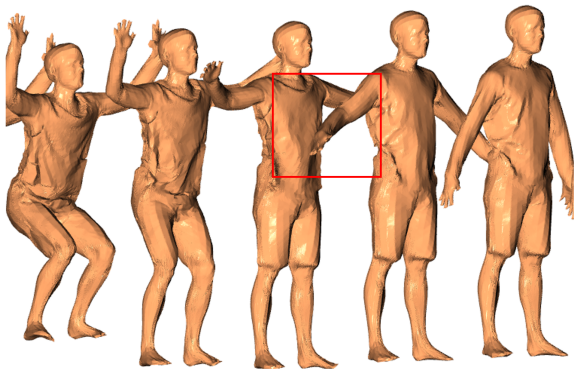


(b) VOCA [9].

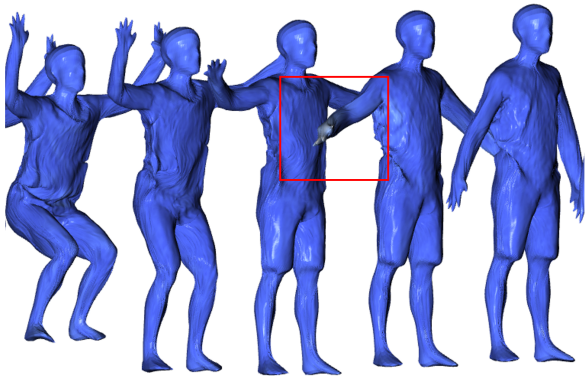


(c) DFAUST [5].

(d) CAPE [20].



(e) Ground truth.



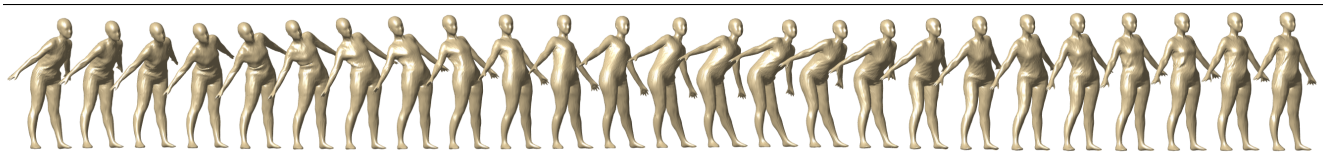
(f) Dynamic neural surface.

Figure 13. Example of the interpolation quality of D-SNS. We train D-SNS on a subset of 30 temporal samples and visualize the time interval as a heatmap. Note that the D-SNS is able to faithfully represent even the detailed clothed human (from the CAPE dataset) and interpolate the missing sequences. The last row shows a zoom on the region highlighted in red.

B.3. Time warping network

The time-warping network is an MLP that finds the optimal temporal alignment between the SRVFs q_1, q_2 of two curves α_1, α_2 obtained from spatially registered D-SNS F_1, F_2; see Figure 11 for the detailed architecture.

Training. We obtain the 4D surfaces F_1 and F_2 at a spherical resolution of 32×32 with 50 time samples $t\}_{i=1}^{50}$. First, using PCA, we map the surfaces F_1 and F_2 to a low dimensional space to obtain two curves $\alpha_1 = \mathcal{Z}(F_1), \alpha_2 = \mathcal{Z}(F_2)$. We then compute their SRVF $q_1 = Q(\alpha_1), q_2 = Q(\alpha_2)$. Since the $\mathbb{L}^2$ metric in the SRVF space mea-



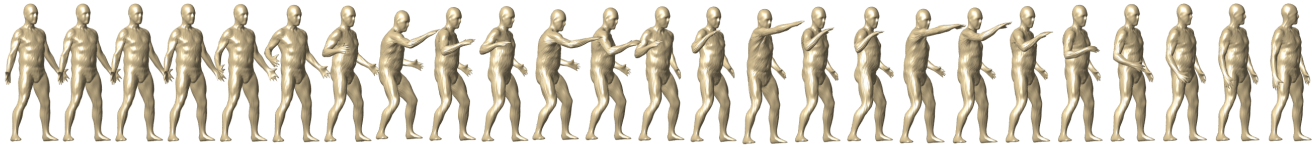
(a) Source.

(b) Target before spatiotemporal registration.

(c) Target after spatiotemporal registration.

(a) Source.

(b) Target before spatiotemporal registration.



(c) Target after spatiotemporal registration.

Figure 14. Two examples of the spatiotemporal registration of 4D humans performing various actions. For each example, we show **(a)** the source 4D surface, **(b)** the target 4D surface before registration, and **(c)** the target 4D surface after registration using the proposed framework. Observe that the target neural surface after registration **(c)** is fully aligned with the source neural surface.

sures nonrigid deformations of curves in the original space, we train the time warping network using the $\mathbb{L}^2$ loss, *i.e.,* $\zeta^* = \| q_1 - q_2 \circ \zeta \|$. To ensure that ζ is a diffeomorphism, it needs to be a monotonically increasing function on the temporal domain $[0, 1]$. To enforce this, we apply a regularization term that enforces the first derivative of the network with respect to time t to be non-negative. We also apply the Sigmoid activation function to the output of the time-warping network to keep its output within the bounds of $[0, 1]$.

We initialize the training with the parameters of a pre-trained time-warping network that is overfitted, in an offline pre-processing step, to the identity diffeomorphism. We then refine the training for $2,000$ epochs and continuously change the timestamps after every 200 epoch. This way, the time-warping network is able to learn a continuous temporal representation that aligns F_2 to F_1.

C. Datasets

We have evaluated the proposed framework on:

(a) Source.

(b) Target before spatiotemporal registration.

(c) Target after spatiotemporal registration.

(a) Source.

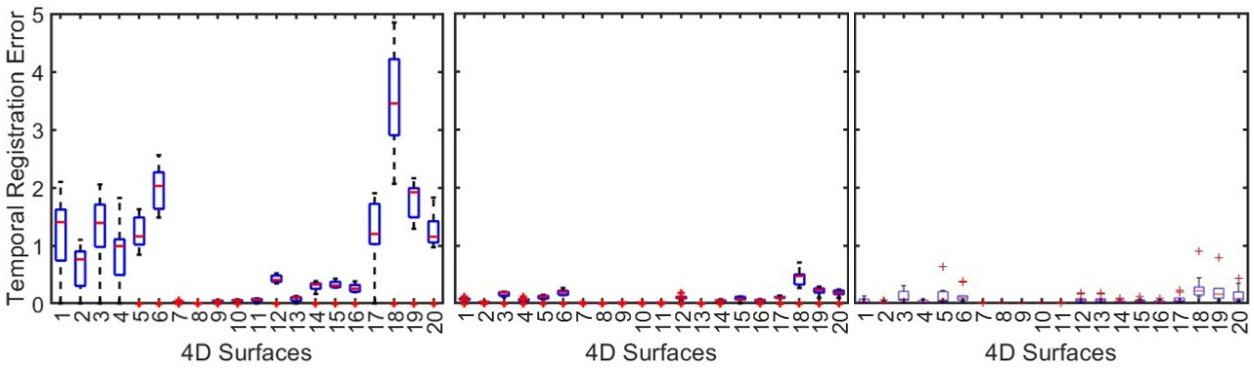
(b) Target before spatiotemporal registration.

(c) Target after spatiotemporal registration.

Figure 15. Two examples of the spatiotemporal registration of 4D faces from the VOCA dataset. For each example, we show **(a)** the source 4D surface, **(b)** the target 4D surface before registration, and **(c)** the target 4D surface after registration using the proposed framework. Observe that the target neural surface after registration **(c)** is fully aligned with the source neural surface.

			CAPE		
	squat	chicken wings	twist tilt left	punching	bend back and forth
4D Atlas	0.4607	1.1354	1.8219	1.404	1.6472
Ours	0.1491	0.3199	0.2193	0.6064	0.4925
			DFAUST		
	punching	punching	jumping jacks	jumping jacks	punching
4D Atlas	1.7044	1.661	4.6092	1.8665	2.0516
Ours	1.09	0.51	0.9023	0.637	0.721
			COMA		
	eyebrow	mouth extreme	high smile	lips back	mouth up
4D Atlas	0.0617	0.0493	0.0377	0.0626	0.0874
Ours	0.0137	0.0173	0.0136	0.0296	0.0640
			VOCA		
	sentence 3	sentence 4	sentence 1	sentence 2	sentence 3
4D Atlas	0.6934	0.383	0.4494	0.4606	0.5266
Ours	0.103	0.0691	0.0759	0.109	0.1400

Table 4. Comparison of the proposed spatiotemporal registration with 4D Atlas [17]. The evaluation showcases the individual pair performance on all four datasets. Table 3 provides the mean, standard deviation, and median for each dataset.



(a) Before registration. (b) After registration 4D Atlas. (c) After registration Ours.

Figure 16. Boxplot visualization of the spatiotemporal registration experiment performed in Figure 8 in the main manuscript. We show the alignment error **(a)** before spatiotemporal registration, **(b)** after spatiotemporal registration using 4D Atlas [17], and **(c)** after spatiotemporal registration using our framework.

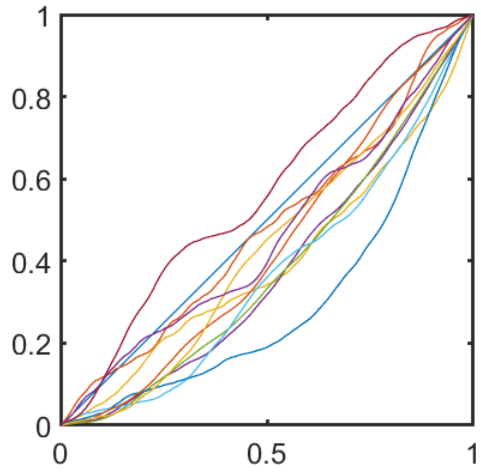


Figure 17. The 10 temporal diffeomorphisms applied on the same 4D surfaces for the evaluation of our framework with 4D Atlas as shown in Figure 8. Note that the range for each temporal diffeomorphism is from $[0, 1] \rightarrow [0, 1]$.

- MPI DFAUST [5], which contains high-resolution 4D body scans of 10 human subjects in motion, captured at 60 fps;
- VOCA [9], which contains high-resolution 4D facial scans of 12 subjects speaking various sentences;
- MPI COMA [29], which contains 4D facial scans of 12 subjects performing various facial expressions; and
- MPI 4D CAPE [20], which contains high-resolution 4D full-body scans of 10 male and 5 females in clothing.

The datasets come with registered triangular meshes. We spherically parameterize these datasets using the implementation [14] of Praun and Hoppe's approach [28]. We then generate random diffeomorphisms to simulate non-registered surfaces.

D. Results

In this section, we show additional results of our neural framework that could not fit within the page limit of the main manuscript. It also reproduces the figures of the main manuscript in high resolution.

D.1. Dynamic Spherical Neural Surfaces

Figure 12 provides more quantitative results on the four datasets. Similar to Figure 6 in main manuscript, here we measure the representation capability of the proposed neural representation using the pointwise error between the neural surfaces and the original ground-truth surfaces. As one can see in Figure 12, the error is smaller than 0.01%. Note that all the surfaces have been normalized for scale to fit within a unit sphere centered at the origin.

Figure 13, on the other hand, demonstrates the interpolation ability of our representation; see Table 2 in the main manuscript for a quantitative evaluation. In this experiment, the neural representation was trained only on 30 temporal samples of the entire sequences. Yet, the method is able to interpolate the missing frames and generate a plausibly smooth 4D surface. For example, the clothed 4D human from the CAPE dataset with high clothing wrinkles is accurately represented and faithfully interpolated using the proposed D-SNS representation.

D.2. Spatiotemporal registration

Figures 14 and 15 show examples of the spatiotemporal registration of pairs of 4D surfaces. In Figure 14, we show two examples of 4D humans before and after their temporal registration. Figure 15 shows two examples of 4D faces before and after their temporal registration. These examples demonstrate that our neural framework is able to spatiotemporally register complex body articulations and facial

(a) Before spatiotemporal registration.

(b) After spatiotemporal registration.

Figure 18. Example of a geodesic **(a)** before and **(b)** after registration between two 4D faces from the COMA dataset. In each example, the first row corresponds to the source 4D surface, the last row corresponds to the target 4D surface, and the three intermediate rows correspond to intermediate 4D surfaces sampled at equidistance along the geodesic path between the source and target. Observe that before registration, the geodesics paths are not well-defined in the highlighted sequences. The highlighted row corresponds to the mean 4D surface.

expressions.

Figure 16 shows the quantitative evaluation of the temporal registration on the same 4D surfaces as the ones shown in Figure 8 in the main manuscript. In this exper-

Figure 19. Example of a geodesic before registration between two 4D surfaces performing a jumping action. In this example, the first row corresponds to the source 4D surface, the last row corresponds to the target 4D surface, and the three intermediate rows correspond to intermediate 4D surfaces sampled at equidistance along the geodesic path between the source and target. Observe that the geodesics paths are not well-defined in the highlighted sequences. The highlighted row corresponds to the mean 4D surface.

iment, we use the evaluation framework proposed in 4D Atlas [17]. Note that, we have changed the range of the Y axis from $0-1$ to $0-5$ to faithfully represent the error before and after registration.

Figure 17, on the other hand, shows the 10 temporal diffeomorphisms applied to perturb the same 4D surfaces for quantitative evaluation performed in Figure 8 of the main manuscript.

Table 4 expands the results of Table 3 in the main manuscript by providing the error on each individual 4D surface. In this experiment, we measure the geodesic distance between the registered 4D surfaces using our method and 4D Atlas method. Note that the smaller the geodesic distance is, the better is the alignment.

D.3. 4D geodesics

Figure 18 shows an example of a geodesic of 4D faces from COMA dataset. Figure 19 and Figure 20, on the other hand, show a high-resolution version of the example of Figure 9 in the main manuscript. Figure 19 shows the 4D surfaces before registration; notice how misaligned is the mean 4D surface (highlighted in red) with the input surfaces. Figure 20 shows the same geodesic after spatiotemporal registration of the source and target 4D surfaces.

D.4. Co-registration and mean 4D surfaces

Figure 21 and Figure 22 shows an example of the 4D mean surface of a set of 4D neural surfaces, from the CAPE dataset. In this document, we show the same example before registration (Figure 21) and after registration (Figure 22). The 4D surfaces in these two figures perform a squat action, and the following two rows perform a bending action. Note that squat action is repetitive, and the number of cycles differs from one 4D surface to another. In particular, the 4D surface in the first row performs two squats while the remaining 4D surfaces perform a single one. Despite this complexity, the proposed approach is able to co-register the 4D surfaces and compute a plausible 4D mean that is as close as possible to all the other 4D surfaces.

Similarly, Figure 23 shows the co-registration and 4D mean surface of six 4D faces, from the VOCA dataset, speaking sentences. In this example, the first four rows

After spatiotemporal registration.

Figure 20. Example of a geodesic after registration between two 4D surfaces performing a jumping action. In this example, the first row corresponds to the source 4D surface, the last row corresponds to the target 4D surface, and the three intermediate rows correspond to intermediate 4D surfaces sampled at equidistance along the geodesic path between the source and target. Observe that the geodesics paths are well-aligned in the highlighted sequences. The highlighted row corresponds to the mean 4D surface.

speak a different sentence than the last two rows. The facial expressions on each 4D surface vary depending on their speaking style. Note that our neural framework is able to accurately co-register them and compute the 4D mean surface.

Figure 21. Example of a mean 4D surface (highlighted in red) computed on six 4D surfaces from the CAPE dataset **before their spatiotemporal registration**. The input 4D surfaces perform different actions: the 4D surfaces in the first four rows perform a squat action, while the last two perform a back and forward bending action. Observe how misaligned the 4D surfaces are before their co-registration and mean computation.

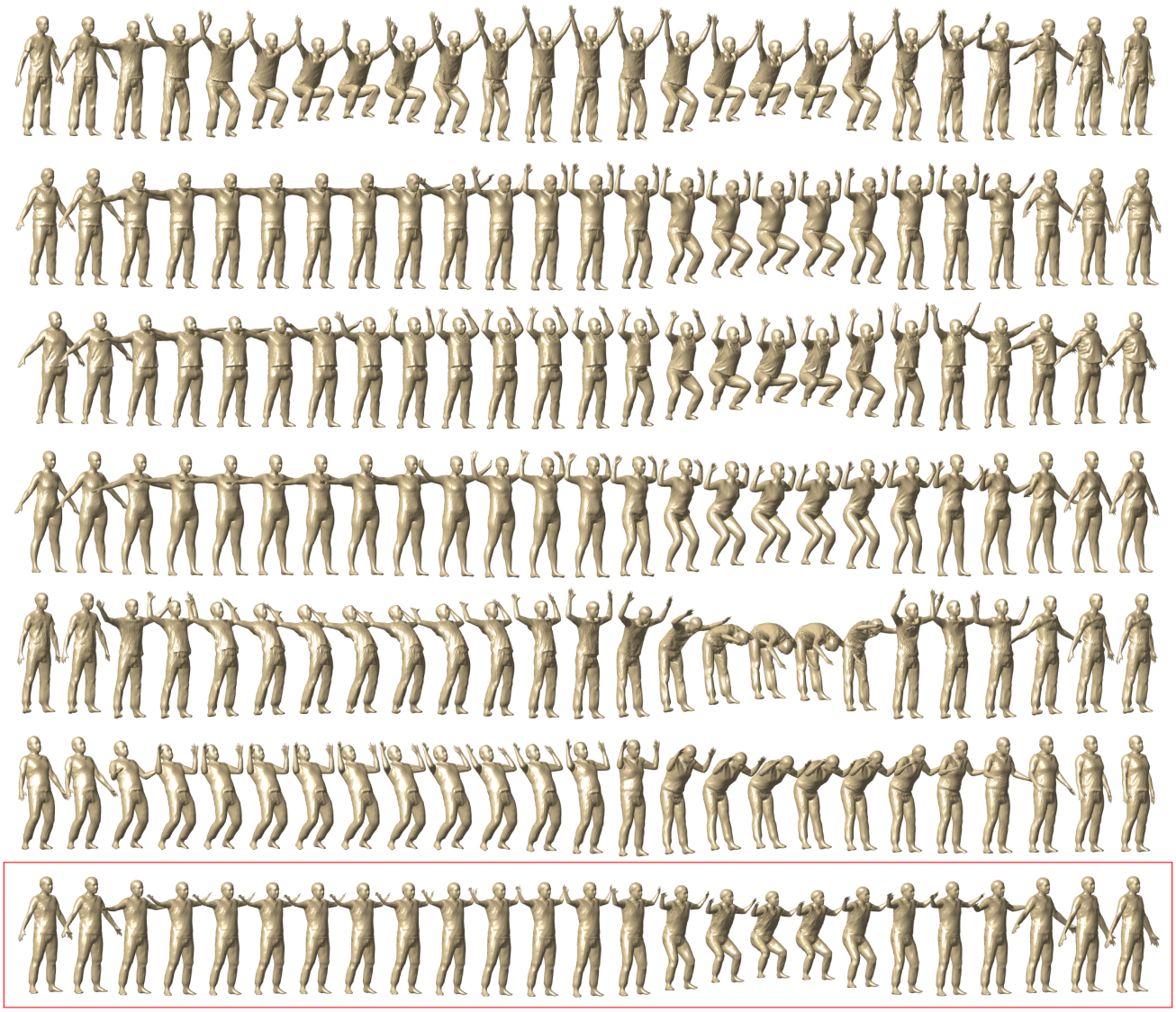


Figure 22. Example of a mean 4D surface (highlighted in red) computed on six 4D surfaces from the CAPE dataset **after their spatiotemporal registration**. The input 4D surfaces perform different actions: the 4D surfaces in the first four rows perform a squat action while the last two perform a back and forward bending action. Observe how aligned the 4D surfaces become after their co-registration and mean computation compared to Figure 21.

(a) Before spatiotemporal registration.

(b) After spatiotemporal registration.

Figure 23. Example of a mean 4D surface (highlighted in red) computed on six 4D surfaces from the VOCA dataset **(a)** before and **(b)** after their spatiotemporal registration. The input 4D surfaces speak different sentences: the 4D surfaces in the first four rows speak the same sentence while the last two rows speak a different sentence. Observe how aligned the 4D surfaces become after their co-registration and mean computation.