

SOMA: From Surface Observations to Muscle Anatomy

Eduardo Alvarado¹, Emily Kim¹, Gerrit Nolte², Friedemann Runte², Mario Botsch², Marc Habermann¹, and Christian Theobalt¹

¹ Max Planck Institute for Informatics, Saarland Informatics Campus

² TU Dortmund University

<https://vcai.mpi-inf.mpg.de/projects/SOMA/>

1 SKIM Dataset Details

In this section, we provide a detailed breakdown of the SKIM dataset, including suit manufacturing, subject demographics, the specific motion sequences captured, and the specifications of the anatomical templates.



Fig. 1: SKIM. Our recorded subjects (S1 to S5).

1.1 Suit Design and Marker Encoding

The motion capture suit is custom-manufactured using a stretchable, matte fabric to minimize specular reflections and ensure tight contact with the skin surface. To capture high-frequency deformations across the entire body, the suit requires a dense marker layout. It contains a total of 4×4 2310 ArUco markers [1]. However, the standard ArUco dictionary is limited to a maximum of $M = 1000$ unique IDs. To achieve the required density, we employ a multi-instance encoding scheme where specific IDs might be repeated once. We distinguish between duplicates using an instance index $i \in \{0, 1\}$. Formally, the full set of landmarks $\mathcal{C}$ on the suit is defined as:

$$\mathcal{C} = \{\mathbf{n}_{m,i,c} \mid m \in \{0, \dots, M-1\}, i \in \{0, 1\}, c \in \{0, 1, 2, 3\}\}, \quad (1)$$

Table 1: Subject demographics. Physical attributes of the 5 subjects in SKIM.

ID	Sex	Height (cm)	Weight (kg)	BMI
S1	M	178	82	25.9
S2	M	182	92	27.8
S3	M	181	90	27.5
S4	F	168	81	28.7
S5	F	162	60	23.1

Table 2: Mesh Specifications. Vertex and face counts for the canonical templates.

Component	Vertices	Faces
Skin Bind $\mathcal{S}_{\text{bind}}$	23,752	47,500
Muscle Bind $\mathcal{M}_{\text{bind}}$	23,752	47,500
Bone Bind $\mathcal{B}_{\text{bind}}$	23,752	47,500
Individual Muscles $\{\mathcal{M}_{\text{bind}}^m\}$ (Total)	320,256	118,758

where each landmark $\mathbf{n}$ is a specific marker corner defined by the dictionary ID m , the instance index i , and the corner index c . This encoding allows us to map the 2310 physical markers to a unique 3D location in the canonical template.

1.2 Subject Demographics and Templates

SKIM contains data from 5 subjects (3 Male, 2 Female) (see Fig. 1) with varying body shapes and BMIs to ensure anatomical diversity. Tab. 1 summarizes the physical attributes of each participant. As described in the main paper, we generate individualized volumetric templates for each subject. Tab. 2 details the resolution of these meshes, which are shared across all subjects.

1.3 Acquisition Protocol

Each subject underwent a capture session lasting approximately 45 minutes, while they were equipped with our custom-fit suit embedded with markers. The session was designed to capture how different muscles change by observing the skin surface during intense physical exertion. The protocol was divided into three main distinct phases:

1. **Part I: Extended Warm-Up & Calibration.** Subjects performed a standardized 10-minute routine including walking, jogging, and dynamic body-weight exercises (squats, lunges, push-ups) to mobilize joints and establish a physiological baseline. This phase concluded with arbitrary stretching.
2. **Part II: Weighted Resistance Training.** The core of the dataset targets key muscle groups (i.e., arms, shoulders, back, legs) using dumbbells and barbells. Subjects performed two rounds of five specific exercises (see Tab. 3). The first round utilized lighter weights, while the second employed increased loads, encouraging muscle bulging and facilitate capture.
3. **Part III: Post-exercise Comparison.** Immediately following the resistance training, subjects repeated the initial warm-up routine, allowing for a direct comparison of surface geometry before and after exercise, capturing effects such as muscle pumping (increased volume due to blood flow).

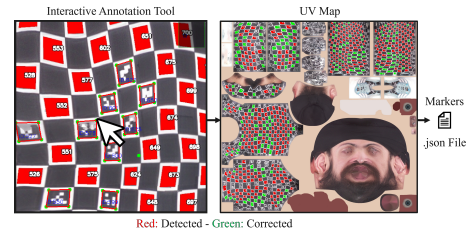
Table 3: Motion Sequences. Standardized exercises performed by all subjects.

Category	Motion	Muscle Focus
Warm-Up	Walking / Jogging	Locomotion
	Bodyweight Squats	Quadriceps / Glutes
	Lunges	Legs / Balance
	Push-ups	Pectorals / Triceps
Routine	Dumbbell Bicep Curls	Biceps Brachii
	Overhead Tricep Extension	Triceps Brachii
	Standing Military Press	Deltoids / Shoulders
	Barbell Bent-Over Rows	Upper Back / Lats
	Barbell Deadlifts	Posterior Chain

1.4 ArUco Annotation

As described in the main paper (Sec. 3.2), we establish the canonical marker template $\mathcal{P}_{\text{bind}}$ by detecting ArUco markers on the unwrapped 2D texture atlas $\mathcal{T}$ of the bind shape $\mathcal{S}_{\text{bind}}$. While standard ArUco detection libraries are robust, ensuring a complete set of markers is challenging due to two primary factors: (1) fail detections due to local distortions introduced by the UV unwrapping process or varying lighting conditions, and (2) texture seams, as unwrapping the 3D body into a 2D plane requires cuts (seams) and markers can be physically located on these boundaries. Standard algorithms, which expect contiguous quadrilaterals, inherently fail to detect these markers.

To overcome these limitations, we developed a custom interactive annotation tool to visualize the UV map $\mathcal{T}$ and overlay the successfully detected markers (see Fig. 2). It enables the user to manually annotate missing markers by selecting their four corners directly on the image. The tool is designed to handle seam artifacts: it allows the user to select corners that are spatially separated in the 2D UV map (e.g., two corners on one UV island and two on another) and groups them under a single, consistent Marker ID. The system automatically validates that each ID is associated with exactly four corner coordinates and merges the manual annotations with the automatic detections. This process yields the final corrected JSON structure required to build the dense 3D canonical point cloud $\mathcal{P}_{\text{bind}}$ via barycentric interpolation.

**Fig. 2: Interactive Annotation Tool.**

1.5 Data Structure and Size

The SKIM dataset is organized by subject (‘S1’ to ‘S5’). For each subject, we provide the canonical anatomical definition, the specific layer meshes, the raw sequence data, and a pre-processed version optimized for training:

- /S[1-5]/
 - /canonical_model/ – Data defining the subject’s bind configuration.
 - * canonical_markers_tpose.json – Canonical point-cloud $\mathcal{P}_{\text{bind}}$ in T-Pose.
 - * lbs_weights_markers.json – Weights $\mathbf{w}_k$ binding markers to skeleton.
 - * lbs_weights_skin.json – Weights binding the skin mesh $\mathcal{S}_{\text{bind}}$ to the skeleton, shared across $\mathcal{M}_{\text{bind}}$ and $\mathcal{B}_{\text{bind}}$.
 - /layers/ – High-resolution anatomical templates in bind pose.
 - * skin_bind.obj – Surface scan mesh $\mathcal{S}_{\text{bind}}$.
 - * muscle_bind.obj – Global volumetric muscle mesh $\mathcal{M}_{\text{bind}}$.
 - * bone_bind.obj – Global volumetric bone mesh $\mathcal{B}_{\text{bind}}$.
 - * /individual_muscles/ – Directory with individual muscle meshes $\{\mathcal{M}_{\text{bind}}^m\}$.
 - /raw/ – Captured motion sequences.
 - * /shot_XXX/
 - motion.bvh – GT skeletal motion θ_t .
 - residuals.json – Surface residuals $\delta_{k,t}$.
 - masks.json – Visibility masks $m_{k,t}$.
 - /processed/ – Data pre-processed for training.
 - * shot_XXX_frame_XXXX.npy – Consolidated frame-based state vectors (pose, residuals, masks).

Each of the 5 subjects was recorded for approximately 45 minutes at 25 Hz, in a motion capture studio equipped with 120 synchronized cameras. This results in over 330,000 unique time steps of motion and nearly 40 million individual 2D images processed for marker detection. Tab. 4 provides a detailed breakdown of the recorded data volume per subject.

Table 4: Dataset Size Breakdown.

Subject	Duration (min)	2D Detected Frames	3D Detected Frames
S1	≈ 45	≈ 4,76M	≈ 50K
S2-S5 (each)	≈ 45	≈ 7,24M	≈ 54K
Total	225	33,7M	266K

2 Marker Refinement

The observed point-cloud across the motion sequence $t \in \{1, \dots, T\}$ is defined as:

$$\mathcal{P}_{\text{GT}}(t) = \{\tilde{\mathbf{p}}_{k,t}^* \mid k \in \mathcal{K}_t\}, \quad (2)$$

where $\mathcal{K}_t \subset \{1, \dots, \mathcal{C}\}$ denotes the subset of canonical markers successfully triangulated at time t . We apply different refinement strategies to address possible incompleteness due to self-occlusions, explained below:

2.1 2D and 3D Linear Interpolation

To address possible short-term dropouts in marker tracking due to their small size, we apply linear interpolation in both the 2D and 3D domains. After performing 2D detection, gaps of length $\delta t \leq \tau_{\max}$ in a camera stream j are identified and filled using linear interpolation between the surrounding valid detections. To preserve data plausibility, $\tau_{\max}$ is kept low (i.e., 2 frames). The same strategy is applied to the 3D trajectories $\mathbf{p}_{k,t}^*$ obtained after triangulation. For any missing frame at time $t \in (t_1, t_2)$, the position is estimated as:

$$\tilde{\mathbf{x}}_t = \frac{t_2 - t}{t_2 - t_1} \mathbf{x}_{t_1} + \frac{t - t_1}{t_2 - t_1} \mathbf{x}_{t_2}, \quad \text{for } t_1 < t < t_2 \quad (3)$$

where $\mathbf{x}$ denotes either 2D image coordinates or 3D world coordinates.

2.2 Temporal Smoothing

To suppress high-frequency flickering in marker trajectories, we apply a temporal median filter to the 3D corner positions. For each marker corner k , we consider its coordinate time series $\{\mathbf{p}_{k,t}^*\}_{t=1}^T$. We apply a median filter with an odd window size w , centered at each time step. The filtered trajectory $\tilde{\mathbf{p}}_{k,t}^*$ is computed component-wise as:

$$\tilde{\mathbf{p}}_{k,t}^* = \text{median} \left(\left\{ \mathbf{p}_{k,t'}^* \mid t' \in \left\{ t - \frac{w-1}{2}, t + \frac{w-1}{2} \right\} \right\} \right) \quad (4)$$

2.3 Local Motion Consensus Filtering

Although unlikely, spatial outliers may occur. Therefore, we apply a local motion consensus filter to evaluate the consistency of a marker's displacement relative to its spatial neighbors. For each marker m at time t , we compute its centroid $\mathbf{c}_m^{(t)} \in \mathbb{R}^3$ as the mean of its four corner positions. Using a k -d tree constructed from the centroids in the previous frame, we identify the K nearest neighbors of marker m and compute their respective displacement vectors $\delta \mathbf{c}_n = \mathbf{c}_n^{(t)} - \mathbf{c}_n^{(t-1)}$.

The marker's own displacement $\delta \mathbf{c}_m$ is then compared against the median displacement of its neighborhood, $\mathbf{v}_{\text{med}} = \text{median}(\{\delta \mathbf{c}_n\})$. A marker is classified as an outlier and excluded from the frame if the discrepancy d_m exceeds a predefined threshold τ_{out} :

$$d_m = \|\delta \mathbf{c}_m - \mathbf{v}_{\text{med}}\| > \tau_{\text{out}} \quad (5)$$

3 Experimental Setup

3.1 Parameters Details

Across all experiments, we adopt the following empirical parameters to train the SOMA framework and extract the resulting deformations.

Network Architecture The core deformation networks representing the skin and muscle layers ($\mathcal{S}_{\text{bind}}$ and $\mathcal{M}_{\text{bind}}$) are implemented as U-Net-style encoder-decoder architectures with skip connections. Both networks consist of a two-layer encoder with hidden dimensions of 512 and 1024, a 1024-dimensional bottleneck, and a two-layer decoder mirroring the encoder. To preserve the original pose signals and prevent vanishing gradients, the features from the encoder stages are concatenated to their corresponding decoder layers. We utilize GELU as the non-linear activation function across all hidden blocks, coupled with 1D Batch Normalization and Dropout ($p = 0.1$). The input dimension, corresponding to the flattened relative rotation matrices of the 23 non-root joints, is 207, while the output layer linearly projects the decoder features to the $3N$ vertex displacements.

Training Hyperparameters The models are trained end-to-end using the Adam optimizer. We train for a total of 20 epochs with a batch size of 128 frames per iteration. The learning rate is held constant at 1×10^{-4} . Gradient clipping with a maximum norm of 1.0 is applied to prevent gradient explosion during optimization.

Loss Formulation and Regularization Weights To guarantee anatomically plausible deformations, continuous surface topology, and to prevent interpenetration between the individual muscles, our total loss function is governed by several weighted terms. The empirical weights for the full model are set as follows:

- **Data/Reconstruction Loss** (λ_{data}): Set to 50, this weight drives the primary vertex-to-vertex alignment between our network predictions and the ground-truth simulations in the canonical space.
- **Spatial Smoothness** (λ_{smooth} and $\lambda_{\text{biharmonic}}$): We penalize high-frequency spatial gradients and unnatural bending using first-order Laplacian and second-order biharmonic constraints. For the muscle layer, both weights are set to 100. To enforce the skin layer to act as a cohesive, thick tissue, its smoothness weights are set higher at 200.
- **Stretch / Edge Length Preservation** (λ_{spring}): Set heavily to 1000 for both the skin and muscle layers. This ensures the discretization of the meshes is maintained by penalizing deviations from the rest-pose edge lengths.
- **Tangential Sliding Penalty** (λ_{tangent}): To prevent vertices from freely sliding along the surface tangent, we apply a weight of 1.0 for the muscle layer and 0.1 for the skin layer.
- **Volume Preservation** (λ_{vol}): Set to 1000 for both layers, aggressively mitigating unrealistic scaling and maintaining the geometric volume of the individual overlapping muscle prisms during severe articulations.

3.2 Metrics

We evaluate surface accuracy and internal anatomical plausibility using the following metrics:

- **Global Surface Tracking:** Mean (MPME), Median (MedPME), and 90th percentile (P90) Euclidean errors between predicted ($\mathbf{p}_k^{\text{pred}}(\boldsymbol{\theta})$) and ground-truth ($\mathbf{p}_{k,t}^*$) markers. P90 specifically captures extreme deformations.

- **Dynamic Region Error (DRE):** To isolate complex soft-tissue dynamics from rigid regions, we compute tracking error exclusively on highly dynamic markers where baseline LBS error exceeds $\tau \in \{15, 20, 25\}$ mm.
- **Intersection Ratio (%):** The percentage of inner vertices penetrating the outer layer, indicating anatomical collision avoidance.
- **Volume Preservation (%):** The mean absolute percentage change from rest volume ($\frac{|V-V_0|}{V_0} \times 100$), measuring the isochoric plausibility of the generated biological tissues.

3.3 Baselines

We compare our method against:

- **LBS:** Applied to the subject’s static 3D scan $\mathcal{S}_{\text{bind}}$ and internal canonical muscle prediction $\mathcal{M}_{\text{bind}}$ without pose-dependent corrections.
- **HIT [2] (via SMPL [4]):** Given fitted shape parameters and pose θ , HIT infers the SMPL skin and internal meshes. Comparing against this pipeline allows us to simultaneously evaluate the limitations of standard parametric surface blending (SMPL) and uncoupled volumetric prediction (HIT) against our explicitly coupled, high-resolution physical formulation.

4 Prism Volume

We use E_{vol} to regularize soft volume preservation. Let P be a prism consisting of a lower triangle $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ and an upper triangle $(\mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6)$. Using the FEM shape functions as defined in, e.g., Landau et al. [3], the interior of this prism is given as the set of all points that can be written as

$$\mathbf{x} = \xi (\alpha \mathbf{x}_1 + \beta \mathbf{x}_2 + \gamma \mathbf{x}_3) + (1 - \xi) (\alpha \mathbf{x}_4 + \beta \mathbf{x}_5 + \gamma \mathbf{x}_6) \quad (6)$$

with the constraints that $0 \leq \alpha, \beta, \gamma, \xi \leq 1$ and $\alpha + \beta + \gamma = 1$. The signed volume of a prism can be obtained as the integral over the volume defined by the prism’s shape function:

$$\int_{\xi=0}^1 \int_{\beta=0}^1 \int_{\alpha=0}^{1-\beta} \det(\mathbf{J}(\alpha, \beta, \xi)) \, d\alpha \, d\beta \, d\xi, \quad (7)$$

where $\mathbf{J}$ is the Jacobian of the interpolation function

$$\mathbf{J}(\alpha, \beta, \xi) = \begin{bmatrix} \frac{\partial \mathbf{x}}{\partial \alpha} & \frac{\partial \mathbf{x}}{\partial \beta} & \frac{\partial \mathbf{x}}{\partial \xi} \end{bmatrix} \quad (8)$$

and the partial derivatives are given by

$$\frac{\partial \mathbf{x}}{\partial \alpha} = \xi (\mathbf{x}_1 - \mathbf{x}_3) + (1 - \xi) (\mathbf{x}_4 - \mathbf{x}_6), \quad (9)$$

$$\frac{\partial \mathbf{x}}{\partial \beta} = \xi (\mathbf{x}_2 - \mathbf{x}_3) + (1 - \xi) (\mathbf{x}_5 - \mathbf{x}_6), \quad (10)$$

$$\frac{\partial \mathbf{x}}{\partial \xi} = \alpha (\mathbf{x}_1 - \mathbf{x}_4) + \beta (\mathbf{x}_2 - \mathbf{x}_5) + (1 - \alpha - \beta) (\mathbf{x}_3 - \mathbf{x}_6). \quad (11)$$

The determinant of this matrix is quadratic in ξ , which occurs in two columns, and linear in α and β , which occur in only one column.

The integral can therefore be evaluated exactly using Gauss Legendre quadrature with two nodes for the integral over ξ and one node for α and β , whose domain forms a triangle. The integral thus evaluates to

$$V = \frac{1}{4} (\det(\mathbf{J}(\omega_1)) + \det(\mathbf{J}(\omega_2))), \quad (12)$$

with nodes:

$$\omega_1 = \left(\frac{1}{3}, \frac{1}{3}, \frac{1-\sqrt{\frac{1}{3}}}{2} \right)^\top \quad \omega_2 = \left(\frac{1}{3}, \frac{1}{3}, \frac{1+\sqrt{\frac{1}{3}}}{2} \right)^\top \quad (13)$$

5 Marker Visibility

Fig. 3 reports mean marker detection coverage across all frames and 5 subjects. We show a per-motion breakdown: Taking as an example arm-focused exercises, we retain 85–95% coverage. We complement this with a histogram of arm-related markers detected during curls. The temporary reduction in *biceps* and *brachialis* visibility at peak flexion is a geometric consequence of the pose, yet lateral bulging remains visible, and the *brachioradialis* (a synergistic flexor) stays well-tracked.

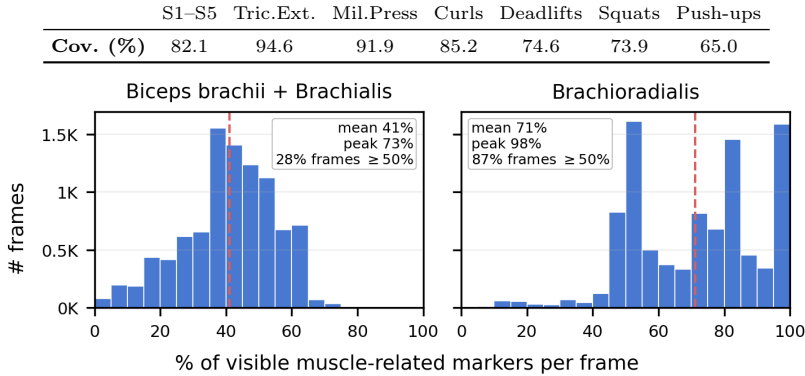


Fig. 3: Marker coverage and elbow flexors visibility for curls.

6 Additional Qualitative Results

In Fig. 4, we show different motions from our dataset, as well as the predicted inward (blue)/outward (red) deformations for both skin and muscle layers, $\mathcal{S}_{\text{bind}}$ and $\mathcal{M}_{\text{bind}}$ respectively. We can observe how the corresponding muscles activate based on the exercise, s.t. quadriceps in the upper-leg during squats, the triceps during push-ups, or the shoulders for military presses.

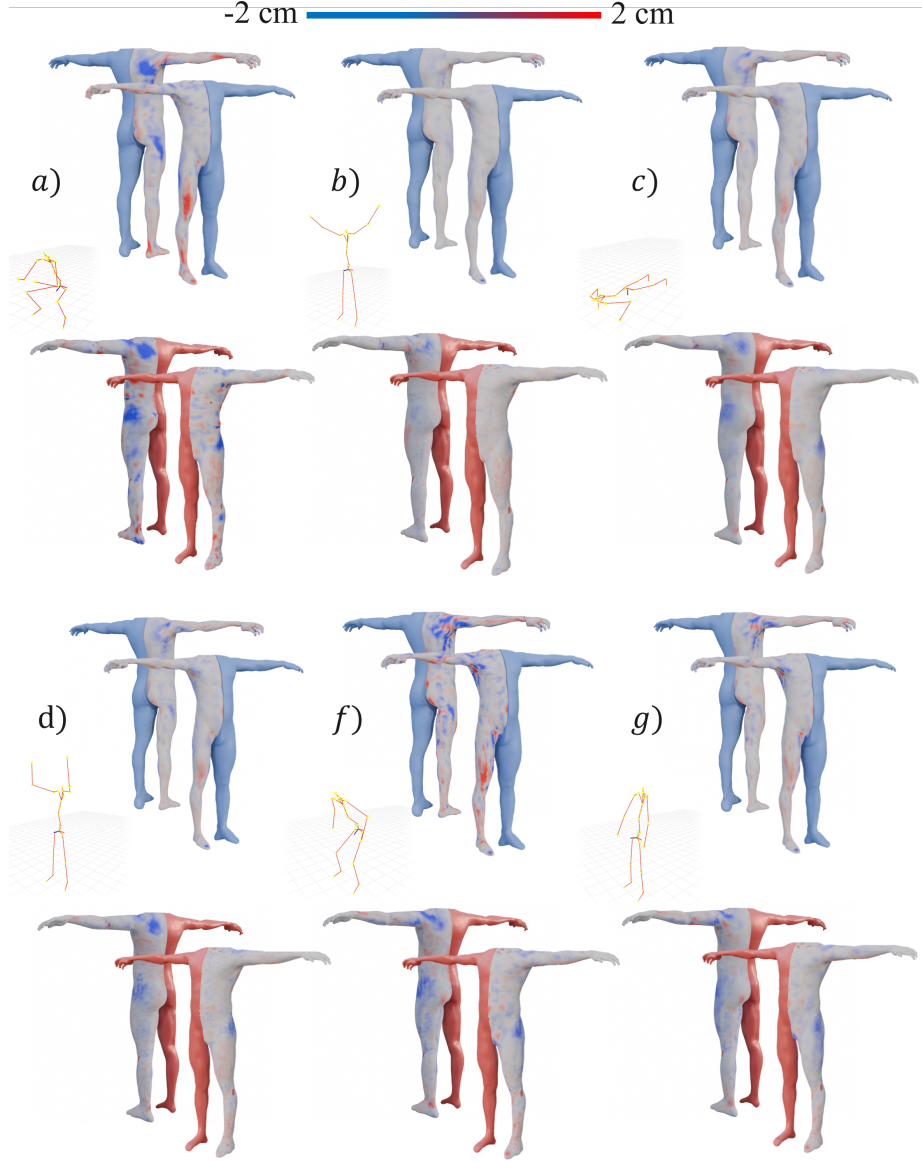


Fig. 4: Predicted muscle and skin meshes across various activities. (a) Squats, (b) Arm rotation, (c) Push-ups, (d) Military press, (e) Barbell bent-over rows, and (f) Deadlifts. For each activity, we visualize the displacement of both $\mathcal{S}_{\text{bind}}$ (top) and $\mathcal{M}_{\text{bind}}$ (bottom) from front and back viewpoints.

As an example of our method on a specific muscle, we show the T-pose canonical bicep deformation in Fig. 5, using the frames of minimum and maximum forearm–upper-arm angle during curls. Following the original convention for positive (blue) and negative (red) deformation, the belly bulges outward while concavities form near the elbow and shoulder.

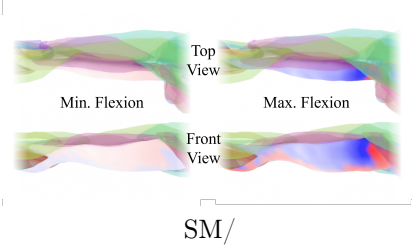


Fig. 5: Biceps. Deformation of biceps brachii during maximum and minimum flexion from top and front views.

7 Individual Muscle Deformation

To animate detailed individual anatomical structures $\{\hat{\mathcal{M}}_{pred}^{(m)}\}_{m=1}^M$, we propagate the learned continuous deformation from the unified boundary layer $\mathcal{M}_{bind}$ to the underlying high-resolution meshes.

In a pre-processing step, we compute a binding that anchors every vertex of the high-resolution individual muscles to the canonical boundary muscle mesh $\mathcal{M}_{bind}$. For each vertex $\mathbf{u} \in \mathcal{M}_{bind}^{(m)}$ in the rest pose, we identify the closest triangle on $\mathcal{M}_{bind}$. Let $\mathcal{I}_u = \{k_1, k_2, k_3\}$ denote the indices of the three vertices forming this triangle, and $\mathbf{b}_u = \{b_1, b_2, b_3\}$ be the corresponding barycentric weights such that the projection of $\mathbf{u}$ onto the surface is $\mathbf{u}_{proj} = \sum_{j=1}^3 b_j \mathbf{v}_{k_j}$. During inference, we reconstruct the posed individual muscles by transferring the barycentric-interpolated canonical displacement to each internal vertex:

$$\tilde{\mathbf{u}}(\boldsymbol{\theta}) = \mathbf{u} + \sum_{j=1}^3 b_j \mathbf{D}_{musc}(\mathbf{v}_{k_j}, \boldsymbol{\theta}). \quad (14)$$

This formulation ensures that the internal high-resolution muscles deform in physical synchrony with the volumetric constraints enforced on the overall boundary structure of $\mathcal{M}_{bind}$.

References

1. Garrido-Jurado, S., Muñoz-Salinas, R., Madrid-Cuevas, F.J., Marín-Jiménez, M.J.: Automatic generation and detection of highly reliable fiducial markers under occlusion. In: Pattern Recognition. Lecture Notes in Computer Science, vol. 8827, pp. 113–122. Springer (2014). https://doi.org/10.1007/978-3-319-16865-4_10
2. Keller, M., Aora, V., Abdelmouttaleb, D.: HIT: Estimating Internal Human Implicit Tissues from the Body Surface (2024)
3. Landau, L.D., Lifshitz, E.M., Atkin, R., Fox, N.: The theory of elasticity. In: Physics of Continuous Media, pp. 167–178. CRC Press (2020)
4. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graph. **34**(6), 248:1–248:16 (Oct 2015). <https://doi.org/10.1145/2816795.2818013>