GHOST: Gaussian splatting for Human Osteoarticular Structure Tracking from Video

Guillaume Le Guludec[1], Corentin Hardy[1], Laurent Albera[2], Charles Pontonnier[1] and Pierre Hellier[1]
[1]Univ Rennes, Inria, IRISA [2]Univ Rennes, Inserm, LTSI
firstname.lastname@irisa.fr

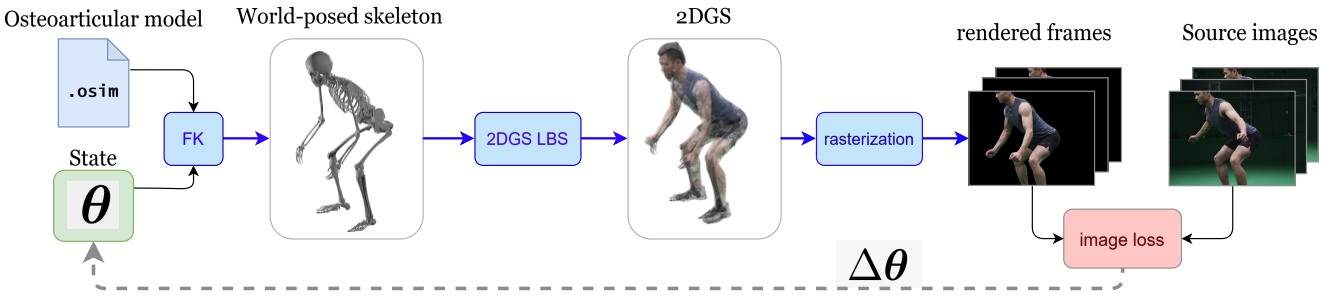

Figure 1. We propose a method to obtain the joint angles of an osteoarticular model by coupling it to a differentiable appearance with 2DGS, and optimizing from a photometric loss in an end-to-end manner. Our pipeline first computes the bone frames in world space through forward kinematics, then computes the location of Gaussian elements attached to the frames with linear blend skinning, and compares the rasterized 2DGS to a ground truth multi-view video to update the state with gradient descent.

Abstract

Accurate extraction of biomechanical quantities from multi-view video remains a challenging problem. Current markerless motion capture pipelines often rely on staged processing: extracting 2D keypoints before triangulating and fitting to a biomechanical model. This process depends on Human Pose Estimation (HPE), might accumulate errors, ignore rich photometric surface information, and might also be unable to retrieve distal rotations. In this paper, we propose a fully differentiable biomechanical-visual model that directly couples photometric appearance with underlying osteoarticular structure. We model the human surface using 2D Gaussian Splatting (2DGS), which is driven by a standard osteoarticular model through a parametric blend-shape formulation. By maintaining end-to-end differentiability, our method allows for the direct optimization of joint angles using photometric loss from multiple camera views. Preliminary results are encouraging for both biomechanical quantities retrieval and 3D reconstruction, paving the way for a new paradigm for markerless biomechanical analysis in the wild.

1. Introduction

The extraction of accurate biomechanical parameters (such as joint kinematics, segmental moments, joint torques, muscle forces and joint reaction forces) from multi-view video is a critical goal in sports science, clinical rehabilitation, and computer animation. Traditional approaches require invasive or restrictive marker-based optical motion capture systems.

Recently, markerless pipelines have gained traction. Open-source pipelines like Pose2Sim [18] utilize off-the-shelf 2D pose estimators to extract keypoints and perform inverse kinematics, relying on the common assumption that the keypoints are rigidly attached to bones. While robust, this staged approach is bottlenecked by the accuracy of the 2D keypoint detector and fundamentally ignores dense pixel data (photometry) that conveys subtle tissue and surface deformations. Cotton *et al.* [5] and Unger *et al.* [24] propose a differentiable pipeline and optimize the joint angles directly from the location of the 2D keypoints. Although it removes the need to rely on a two-stage procedure, this method is equally limited by the quality of the detected keypoints and still falls short of utilizing the rich pixel data.

Simultaneously, another line of work has been leveraging the recent advances in differentiable rendering [7, 9, 16] to solve the problem of modelling human visual appearance [19, 28]. While these methods usually provide compelling visual representations, they usually do not provide biomechanical information as they usually rely on body shape models like the Skinned Multi-Person Linear model (SMPL) body model [13] that lack the underlying osteoarticular structure that is required for classical biomechanical applications such as sports gesture analysis or clinical evaluation.

To overcome these limitations, we introduce a fully differentiable biomechanical-visual model. Our core insight is to represent the photometric appearance of the subject using 2D Gaussian Splatting [7], which provides superior surface modeling compared to 3DGS [9], and couple it directly to a biomechanical model representing the osteoarticular structure of the subject. We bridge the biomechanical space and the visual space using a Linear Blend Skinning (LBS) formulation. This allows the 2D Gaussians to be deformed by underlying rigid joint kinematics while taking into account motion arising from soft-tissues and clothing deformation, rendering multi-view images in a fully differentiable manner.

Our main contributions are:

- We propose the first end-to-end differentiable pipeline coupling a standard biomechanical model with 2DGS.
- We introduce a LBS formulation that maps biomechanical parameters (joint angles) directly to 2DGS primitives
- We show first promising results of retrieving biomechanical parameters directly from photometric loss.

2. Related Work

Markerless Biomechanics. The transition from marker-based to markerless biomechanics has heavily relied on deep learning. Pose2Sim [18] provides a standard framework by linking OpenPose with OpenSim [6]. However, it operates in a top-down, non-differentiable manner, and relies on a two-stage pipeline of keypoint extraction and then inverse kinematics. Optimization methods, such as Cotton *et al.* approach [4, 5] and [24], incorporate kinematic and biomechanical constraints directly into the tracking optimization, and improve the temporal consistency by parameterizing the joint angles trajectories with a MLP. These methods still rely on HPE algorithms, and thus are not end-to-end differentiable from pixels to joint angles, making them unable to leverage the pixel-level photometric error signal.

Neural Rendering of Human Bodies. Novel view synthesis for humans has been revolutionized by NeRFs and Gaussian Splatting. SNARF [2] and the more recent Fast-SNARF [3] introduced a skeleton-driven neural radiance field, showcasing the potential of combining anatomy with neural rendering. However, SNARF and Fast-SNARF rely on implicit density fields which are computationally heavy and additionally lack direct parameterization compatible with standard biomechanical applications. Switching from NeRF to the much faster 3DGS, several works [17, 22] have simultaneously improved the visual quality of the human reconstruction while making it real-time. Very recently, [28] achieved a new state of the art with a hybrid approach utilize spatially distributed MLPs to predict the attributes of Gaussians. Their method relies on SMPL-X, an expressive body shape parameterization that is however unsuitable for biomechanical use.

2DGS has recently shown great promise in accurately reconstructing continuous surfaces, making it an ideal candidate for modeling the skin boundary layer over a skeletal rig.

SKEL-based methods. The recent SKEL model, introduced by Keller *et al.* [8] has provided a step towards bridging the gap between osteoarticular structure and visual appearance by re-rigging SMPL with a biomechanical model. By migrated the existing SMPL-annotated image and video datasets to SKEL-annotated datasets, they have paved the way for a variety of work that can predict the biomechanical SKEL parameters directly from images. HSMR [26] and SKEL-CF [12] regress the SKEL parameters directly from the image but do not include appearance modelling and rely solely on the SKEL-annotated data to train their model. In contrast, the recent GST [20], although it employs SMPL and not SKEL, jointly regresses the parameters along with the mesh-bound offsets of 3D Gaussians to model the appearance, using a photometric loss to drive training. Their use of SMPL however, precludes direct biomechanical applications.

3. Methodology

Our goal is to extract biomechanical quantities, specifically the joint state vector $\boldsymbol{\theta} \in \mathbb{R}^D$ of a biomechanical model directly from a set of C synchronized multi-view video frames $\mathcal{I} = \{I_1, I_2, \ldots, I_C\}$. We achieve this by optimizing a fully differentiable rendering pipeline, as shown in Fig. 1. We first define a mapping from the joint state vector to the rendered multi-view images using a calibration time frame or sequence of frames, and then leverage this differentiable mapping to retrieve the state vectors on the subsequent frames.

3.1. Biomechanical-visual coupling

Kinematics. We employ a biomechanical model defining a kinematic tree to serve as the osteoarticular structure of our model. The state vector $\boldsymbol{\theta}$ includes the joint angles of the model as well as all other degrees of freedom like the global position. Given the state $\boldsymbol{\theta}$, the forward kinematics

(FK) function computes for each bone j the global transformations $\mathcal{F}(\theta) = \mathbf{T} = (\mathbf{T}_j)_{1 \leq j \leq J} \in SE(3)^J$.

2DGS reference appearance. We use 2DGS to model the photometric appearance, *i.e.* skin and clothing. In contrast to the original 3DGS [9], 2DGS are flat, and therefore very amenable to representing surfaces. 2DGS represents a scene as a collection of many Gaussian elements, each characterized by a 3D mean $\boldsymbol{\mu}_n$, a 2D scaling vector $\boldsymbol{s}_n$, a rotation defined by a quaternion q_n, as well as an opacity α_n and spherical harmonics c_n for view-dependent color.

We define the coupling of the biomechanics to the appearance as finding a mapping from the global skeletal transforms $\mathbf{T}$ to the attributes of the N Gaussian elements $\boldsymbol{g}$:

$$\boldsymbol{g}(\mathbf{T}) = (\boldsymbol{\mu}_n(\mathbf{T}), \boldsymbol{s}_n(\mathbf{T}), q_n(\mathbf{T}), \alpha_n(\mathbf{T}), c_n(\mathbf{T}))_{1 \leq n \leq N} \tag{1}$$

In order to build this coupling, a reference appearance is first found by fitting the 2DGS on the views $\mathcal{I}^0 = \{I_1^0, I_2^0, \ldots, I_C^0\}$ at $t = 0$ of the multi-view videos, using a photometric objective regularized by a consistency and a depth distortion loss as prescribed by the authors [7]. We use $\sum_{c=1}^{C} \|\hat{I}_c - I_c^0\|_1$ as the photometric objective, where the rasterized images $\hat{I}_c = \mathcal{R}(\boldsymbol{g}, \mathbf{K}_c, \mathbf{V}_c)$ are differentiable w.r.t. their arguments, permitting gradient-based optimization of the 2DGS. Here $\mathbf{K}_c$ and $\mathbf{V}_c$ denote the intrinsic and extrinsic matrices of camera c.

In order to avoid modeling the background, we run a segmentation algorithm, Sapiens [11], to get a mask for each ground truth view. We further leverage the obtained body-part segmentation to endow each Gaussian with a body part identity b_n, analogous to a view-independent color. We learn these additional parameters by adding a pixel-wise cross-entropy loss on the segmented images, similar to a photometric loss. We then use these attributes as a heuristic to filter out Gaussians that do not contribute to the appearance of our upper limbs model. We use the Monte-Carlo Markov Chain method of Kheradmand *et al.* [10] as the Gaussian densification strategy, an instrumental part of the optimization. We discuss the design choices in A.1.

The obtained reference appearance $\boldsymbol{g}^0 = (\boldsymbol{\mu}_n^0, \boldsymbol{s}_n^0, q_n^0, \alpha_n^0, c_n^0)_{1 \leq n \leq N}$ is then rigged to an initial estimation of the reference global transformations $\mathbf{T}^0 = (\mathbf{T}_j^0)_{1 \leq j \leq J}$. We detail how we obtain $\mathbf{T}^0$ in section 4.1.

Rigging the Linear Blend Skinning model. The rigging is performed with a Linear Blend Skinning (LBS) formulation. This choice is motivated by the need to model non-rigid motion that arises because of soft-tissues. For the sake of simplicity, we make only the means and quaternions depend on the skeletal transforms $\mathbf{T}$, and keep the scalings, opacities, and spherical harmonics of the 2DGS constant regardless of $\mathbf{T}$.

LBS, originally used with mesh, is naturally extended to Gaussian blending, by interpolating both means and quaternions. We define the re-posed means as:

$$\boldsymbol{\mu}_n(\mathbf{T}) = \sum_j w_{j,n} \mathbf{T}_j ({\mathbf{T}^0}_j)^{-1} \boldsymbol{\mu}_n^0 \tag{2}$$

while the re-posed 2DGS quaternions are computed using weighted Markley averaging [14]:

$$q_n(\mathbf{T}) = \text{markley}((q_n^1, w_{1,n}), \ldots, (q_n^J, w_{J,n})) \tag{3}$$

where $q_n^j = \text{mat2vec}(\mathbf{T}_j ({\mathbf{T}^0}_j)^{-1}) q_n^0$ corresponds to the world orientation of Gaussian element n when rigidly moved from the reference segment frame $\mathbf{T}_j^0$ to the new segment frame $\mathbf{T}_j$.

To define the blend weights $w_{j,n}$ between a given 2DGS g_n and a bone segment j, we use the softmax of the negative squared distance from the center $\boldsymbol{\mu}_n$ to the proximal-distal line segment of the reference world-transformed bones. We then freeze in the subsequent steps. Although this heuristic does not guarantee optimal blend weights, it provides plausible values nonetheless. See A.2 for more details about this design choice.

3.2. Photometry-driven tracking

Our overall appearance model then combines the FK with the bones-to-Gaussians LBS to yield a fully-differentiable pose-dependent Gaussian model $\boldsymbol{g}(\mathcal{F}(\theta))$. By chaining it with the 2DGS rasterizer, we get a fully differentiable mapping from $\boldsymbol{\theta}$ to $\hat{I}$, allowing gradient to back-propagate directly from pixels to joint angles $\boldsymbol{\theta}$.

During inference (tracking), we keep the reference Gaussian parameters fixed and optimize with gradient descent a combination of a photometric loss and a biomechanical joint-limits regularization term:

$$\mathcal{L} = \frac{1}{T} \sum_{t=1}^{T} \left(\sum_{c=1}^{C} \|\hat{I}_c^t - I_c^t, \|_1 + \lambda_{\text{b}} \sum_{j=1}^{D} \left(\max(0, \theta_j^t - u_j) + \max(0, l_j - \theta_j^t) \right) \right) \tag{4}$$

where u_j and l_j are respectively the upper and lower bounds for joint angle j as per the model, and λ_{b} is a hyper-parameter controlling the strength of the constraint on joint angle limits. We learn an independent joint-angle vector $\boldsymbol{\theta}_t$ for each time frame t. In practice, rather than processing all time frames in parallel, we leverage the continuity of motion and process them sequentially, initializing each value joint angle $\boldsymbol{\theta}_t^{\text{init}}$ with the value $\boldsymbol{\theta}_{t-1}^{\text{final}}$ from the previously optimized time frame.

4. Experiments

4.1. Experimental Setup

We implemented our pipeline in PyTorch and use `gsplat` [27] for 2DGS rasterization. All experiments were run on an NVIDIA RTX A5000 graphics card. We use Adam for training.

Osteoarticular model. We use a simplified version of the MoBL-ARMS upper limb model [15, 21], although any biomechanical model could be used. In this model, the finger joints have been removed and the motion of the clavicle and scapula have been frozen so that the shoulder's motion is determined by the 3 degrees of freedom of the gleohumeral joint. Thus each arm has a total of 7 degrees of freedom. We include 6 additional degrees of freedom to characterize the orientation and position of the trunk in space, yielding an upper-limb model with 20 degrees of freedom.

This model, originally described in OpenSim [6], has been ported to the MuJoCo [23] file format and is accessible with the MyoSuite library [1]. Since MuJoCo only provides GPU acceleration with JAX, we use `Pytorch Kinematics` [29] instead to seamlessly incorporate the FK into our pipeline.

Dataset and baselines. We evaluate our method on scenes from the NHR dataset [25]. All scenes have 200 time frames and 56 views. Our 2DGS has around 50k Gaussians.

We compare our method to two baselines: Pose2Sim [18] and our own implementation of the method of Cotton *et al.* [4]. In order to get an initial estimate of the global skeletal transform $\mathbf{T}^0$, we use the initial pose estimated by Cotton's method. We set λ_b to 0.01.

4.2. Preliminary results

Methods	Cot. vs P2S	Cot. vs Ours	P2S vs Ours
Avg. corr.	0.621	0.600	0.731

Table 1. Average correlation across scenes and joint angles of the shoulders and elbows. Cot. referes to our re-implementation of [4], and P2S is Pose2Sim.

Joint angles analysis. Since we do not have ground truth, we compute the correlation between trajectories for each pair of methods. Figure 2 illustrates the agreement between our method and the baselines. Note that our method produces trajectories that are somewhat less smooth than the baselines, which is expected as we do not use any filtering or temporal regularization. The baselines do not provide meaningful estimations of the forearm pronosupination and wrist motion because of the single keypoint on the wrist, so we exclude them from our comparison in Table 1. Our method detects motion in the wrist, but visual inspection reveals frequent failure cases. For instance on *sport_3*, the wrist motion fails to be recovered when the two hands are very close from one another. We hypothesize that this might be due to fixing the appearance once and for all at the beginning of the sequence, ignoring finger and shading changes. Figure 3 provides a visual assessment of the method. We further discuss the robustness of the approach in A.3.

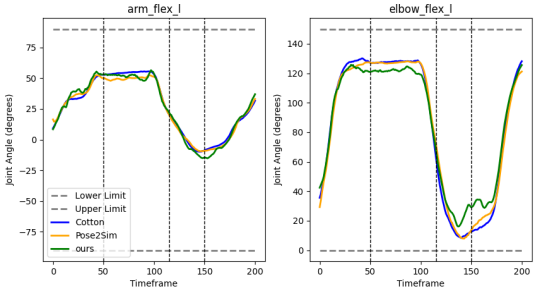

Figure 2. Comparison of the left arm and elbow flexion angular trajectories on *sport_3* where the subject is performing a squat.

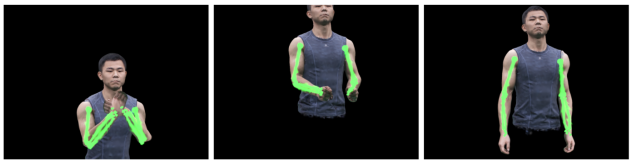

Figure 3. Visualisation of the rasterized images for *sport_3* at timeframes 50, 115 and 150. Estimated bones are shown as a green overlay.

5. Conclusion

In this paper, we presented a novel paradigm for extracting biomechanical quantities from multi-view video. By creating a fully differentiable pipeline that explicitly couples a standard osteoarticular model with 2D Gaussian Splatting via a LBS formulation, we enable direct optimization of biomechanical states — joint angles — from photometric data. While the current results are preliminary, the proposed method shows potential to recover motions that are challenging for keypoint-based approaches, such as arm pronosupination and wrist flexion. Furthermore, leveraging rich photometric cues together with an accurate appearance model may enable more robust reconstruction from only a few views compared with approaches that rely solely on HPE. Future work could be carried out to extend the current results: use of a full-body osteoarticular model and of corrective blendshapes, or of bones-to-skin models like SKEL. Besides, the method currently processes each time step sequentially and with fixed appearance from the first time frame; allowing parallel computation and continuous learning of the blend shapes and texture during a calibration phase would greatly improve the appearance model. The addition of physics-based priors could also further improve the continuity of the reconstructed movement.

References

[1] Vittorio Caggiano, Huawei Wang, Guillaume Durandau, Massimo Sartori, and Vikash Kumar. Myosuite–a contact-rich simulation suite for musculoskeletal motor control. *arXiv preprint arXiv:2205.13600*, 2022.

[2] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11594–11604, 2021.

[3] Xu Chen, Tianjian Jiang, Jie Song, Max Rietmann, Andreas Geiger, Michael J Black, and Otmar Hilliges. Fast-snarf: A fast deformer for articulated neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11796–11809, 2023.

[4] R James Cotton. Differentiable biomechanics unlocks opportunities for markerless motion capture. In *2025 International Conference On Rehabilitation Robotics (ICORR)*, pages 44–51. IEEE, 2025.

[5] R James Cotton, Allison DeLillo, Anthony Cimorelli, Kunal Shah, JD Peiffer, Shawana Anarwala, Kayan Abdou, and Tasos Karakostas. Optimizing trajectories and inverse kinematics for biomechanical analysis of markerless motion capture data. In *2023 International Conference on Rehabilitation Robotics (ICORR)*, pages 1–6. IEEE, 2023.

[6] Scott L. Delp, Frank C. Anderson, Allison S. Arnold, Peter Loan, Ayman Habib, Chand T. John, Eran Guendelman, and Darryl G. Thelen. OpenSim: Open-Source Software to Create and Analyze Dynamic Simulations of Movement. *IEEE Transactions on Biomedical Engineering*, 54(11):1940–1950, 2007.

[7] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 conference papers*, pages 1–11, 2024.

[8] Marilyn Keller, Keenon Werling, Soyong Shin, Scott Delp, Sergi Pujades, C Karen Liu, and Michael J Black. From skin to skeleton: Towards biomechanically accurate 3d digital humans. *ACM Transactions on Graphics (TOG)*, 42(6):1–12, 2023.

[9] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[10] Shakiba Kheradmand, Daniel Rebain, Gopal Sharma, Weiwei Sun, Yang-Che Tseng, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. 3d gaussian splatting as markov chain monte carlo. *Advances in Neural Information Processing Systems*, 37:80965–80986, 2024.

[11] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, pages 206–228. Springer, 2024.

[12] Da Li, Jiping Jin, Xuanlong Yu, Wei Liu, Xiaodong Cun, Kai Chen, Rui Fan, Jiangang Kong, and Xi Shen. Skel-cf: Coarse-to-fine biomechanical skeleton and surface mesh recovery. *arXiv preprint arXiv:2511.20157*, 2025.

[13] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023.

[14] F Landis Markley, Yang Cheng, John L Crassidis, and Yaakov Oshman. Averaging quaternions. *Journal of Guidance, Control, and Dynamics*, 30(4):1193–1197, 2007.

[15] Daniel C. McFarland, Emily M. McCain, Michael N. Poppo, and Katherine R. Saul. Spatial Dependency of Glenohumeral Joint Stability During Dynamic Unimanual and Bimanual Pushing and Pulling. *Journal of Biomechanical Engineering*, 141(5):051006, 2019.

[16] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106, 2021.

[17] Arthur Moreau, Jifei Song, Helisa Dhamo, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Human gaussian splatting: Real-time rendering of animatable avatars. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 788–798, 2024.

[18] David Pagnon, Mathieu Domalain, and Lionel Reveret. Pose2sim: An end-to-end workflow for 3d markerless sports kinematics. *Sensors*, 2022.

[19] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9050–9059, Nashville, TN, USA, 2021. IEEE.

[20] Lorenza Prospero, Abdullah Hamdi, Joao F Henriques, and Christian Rupprecht. Gst: Precise 3d human body from a single image with gaussian splatting transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6007–6017, 2025.

[21] Katherine R. Saul, Xiao Hu, Craig M. Goehler, Meghan E. Vidt, Melissa Daly, Anca Velisar, and

Wendy M. Murray. Benchmarking of dynamic simulation predictions in two software platforms using an upper limb musculoskeletal model. *Computer methods in biomechanics and biomedical engineering*, 18 (13):1445–1458, 2015.

[22] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. SplattingAvatar: Realistic Real-Time Human Avatars With Mesh-Embedded Gaussian Splatting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1606–1616, Seattle, WA, USA, 2024. IEEE.

[23] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033, 2012.

[24] Tim Unger, Arash Sal Moslehian, J.D. Peiffer, Johann Ullrich, Roger Gassert, Olivier Lambercy, R. James Cotton, and Chris Awai Easthope. Differentiable Biomechanics for Markerless Motion Capture in Upper Limb Stroke Rehabilitation: A Comparison With Optical Motion Capture. *IEEE Transactions on Medical Robotics and Bionics*, pages 1–1, 2025.

[25] Minye Wu, Yuehao Wang, Qiang Hu, and Jingyi Yu. Multi-view neural human rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1682–1691, 2020.

[26] Yan Xia, Xiaowei Zhou, Etienne Vouga, Qixing Huang, and Georgios Pavlakos. Reconstructing humans with a biomechanically accurate skeleton. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5355–5365, 2025.

[27] Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, and Angjoo Kanazawa. gsplat: An open-source library for gaussian splatting. *Journal of Machine Learning Research*, 26(34):1–17, 2025.

[28] Youyi Zhan, Tianjia Shao, Yin Yang, and Kun Zhou. Real-time high-fidelity gaussian human avatars with position-based interpolation of spatially distributed mlps. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26297–26307, 2025.

[29] Sheng Zhong, Thomas Power, Ashwin Gupta, and Peter Mitrano. PyTorch Kinematics, 2024.

A. Additional details and discussion

A.1. 2DGS design choices

2DGS vs 3DGS. We adopt 2DGS instead of 3DGS as it is better suited for modeling surface-like structures such as skin and clothing. In practice, 2DGS provides a more adequate representation where the Gaussians are encouraged to concentrate on a surface rather than fill the whole inside of the envelope. Additionally, unlike 3DGS, 2DGS also outputs normal maps, which may readily be employed in future work to leverage normal annotations from human vision foundation models like Sapiens to improve the signal quality.

MCMC densification. Compared to the original densification strategy from [9], the MCMC densification strategy of [10] facilitates the pipeline by bypassing the need for a structure-from-motion step, which is otherwise necessary to obtain a good initialization of the Gaussians, without incurring any loss in reconstruction quality.

A.2. Blend weights

The Gaussian-to-bone blend weights are defined as $w_{j,n} = \frac{e^{-d_{j,n}/\tau}}{\sum_j e^{-d_{j,n}/\tau}}$ where $d_{j,n}$ is the distance between the mean of the nth Gaussian and the jth bone, which is approximated by the central segment of its mesh bounding cylinder. τ is a temperature hyper-parameter; a small τ will produce a winner-take-all effect resulting in a rigid poly-articulated model, while large values lead to smoother blending. This provides a simple way to smoothly attach Gaussians to their nearest bones, while ensuring smoothness in regions like the elbow, where bone membership is not clear. Note however that this remains merely a convenient heuristic, and further work may refine the blend weights, *e.g.* through optimization.

A.3. Robustness analysis

Fixed appearance. For the sake of simplicity, our model currently fixes the spherical harmonics determining the appearance of the Gaussians. This results in artifacts like baked-in shadows, potentially driving the purely photometric optimization to converge to bad local minima (as it will match for instance darker areas to darker areas and neglect finer details). In our experiments we use only the first time frame to create the appearance model, hindering the accurate prediction of change in color under a change in orientation. Learning an orientation-dependent color through a richer calibration phase might provide a way to mitigate this problem. Likewise, taking pose-dependent biomechanical deformations like muscle bulging into account might make the optimization more robust.

Impact of the number of views. In our experiments we tested our approach using all 56 available views. To assess the sensitivity of the method to reducing the number of views, we performed minimal ablation experiments. Fig. 4 compares the obtained trajectories for the arm and elbow

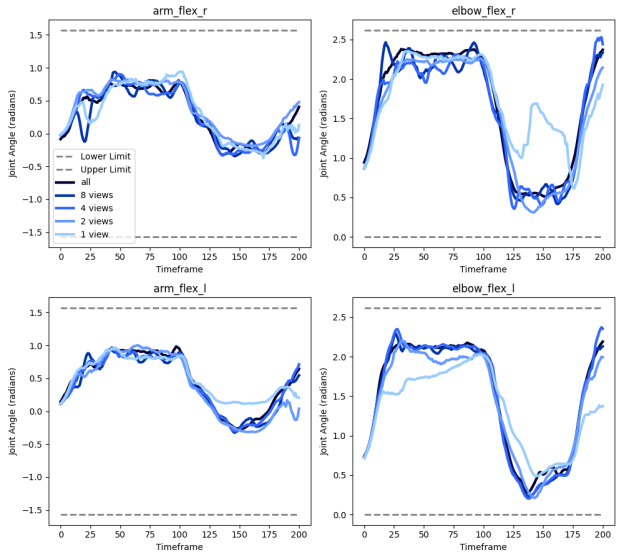

Figure 4. Comparison of the trajectories for the left and right shoulder and elbow flexion angles on *sport_3* for varying number of views during tracking: 1, 2, 4, 8 and 56.

flexions when using a different number of views for tracking (1, 2, 4, 8 and all). Note that we still use all 56 views to estimate the initial transform $\mathbf{T}^0$. We observe that the method reliably tracks shoulder and elbow motion with 4 views. Below that, tracking becomes somewhat unstable, although the overall tendency seems to be captured even with a single view. We did not however make an effort at tuning the hyper-parameters, and hypothesize that a more careful selection of the learning rate together with a stronger time regularization (*e.g.* using an anchoring prior from the previous time steps) may improve robustness.