# MOVERSE: STARTING UP A **MO**TION UNI**VERSE**

STARTUP SUBMISSION

**Nikolaos Zioulis**
Moverse
Thessaloniki, Greece
nick@moverse.ai

**ABSTRACT**

Capturing human motion has been a long standing digitization challenge. Machine learning advances are now making it possible to capture human motion without suits and also generate motion on demand. Moverse is a start-up company whose goal is to accelerate human motion capture, generation and downstream workflow use. In this manuscript, we present the technical design choices of its motion capture technology, and the supporting platform and services, outlining the company's progress in this challenging sector.

***Keywords*** Motion Capture · Machine Learning · Computer Vision · Computer Graphics

## 1 Introduction

The digitization of human motion and performance allows researchers and creators to understand and reproduce human behavior. It is the cornerstone in many different fields, and its technological advances have been driving growth and changes in said fields. In life sciences, motion capture is used to personalize treatment for individuals with neuromuscular disorders, assess mobility, and monitor patient recovery progress. In the domain of sports science, performance digitization is used to apply biomechanical analysis and gain insight into the improvement and training of athletes, in terms of skill and physical traits, and to prevent injuries by identifying injury-inducing conditions such as fatigue. For academics in biomechanics research, motion capture systems are essential for studying human movement and the development of biomechanical models. As researchers seek to create specialized models for certain populations, the need to acquire motion capture data at scale is high.

In creative sectors, motion capture has been used to create lifelike content with characters moving naturally and realistically. Capturing the performance of actors has allowed film directors, animators, and game developers to produce high-quality content. The entertainment industry is also seeing significant advances through real-time performance capture. New workflows that promote creative flexibility, such as virtual production, boost the creative process, reducing costs and time. The integration of motion capture systems into popular game engines have also streamlined new digital content creation workflows, empowering storytellers.

Up to now motion capture technology has been primarily sensor driven. By acquiring data from multiple sensors, such as inertial or optical, and modeling human motion, it was possible to explain the observations by fitting the human kinematic model, and thus capture human motion. To achieve this, though, it requires the use of wearables, either the sensors themselves, or passive items like retro-reflective markers that could be robustly imaged. The emergence of modern data-driven techniques, allows us to lift this restriction, capturing humans unencumbered by suits and wearables. Such markerless capture technology will improve the efficiency of capture sessions, open up participatory uses of the technology, and enable unobstructed performance capture. Further, generative models take a step beyond capturing, as we can now receive motion data on demand, without specifically capturing human performances. This is expected to strike a new balance across the various fields, where capturing technology will only be used when there is a need for individualization, and generative technology will replace it for most other uses.

In this manuscript, we present the efforts of Moverse[1] in the motion digitization space, a startup company that develops tools and technology to capture and generate human motion. The focus lies on the technical design behind the technology and platform being developed, outlining the core design choices and their interplay. We also present the implementation status and current realization of this design, which is a step towards delivering the company's vision, namely, a set of tools and services spanning capture and generation that will accelerate the users' downstream workflows.

## 2 Building on a Unified Representation

Capturing human motion involves fitting an articulated skeleton model to any available real-world observations, under reasonable assumptions about the underlying model. This is typically formulated as a numerical optimization problem estimating the articulation parameters $\boldsymbol{\theta} \in \mathbb{R}^{\Theta \times \mathbb{SO}(3)}$, the scale parameters $\boldsymbol{\beta} \in \mathbb{R}^{B}$, and the global transform $\mathbf{T} = \left[\begin{smallmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{smallmatrix}\right] \in \mathbb{SE}(3)$ that minimize a combined error at each point in time $t$:

$$\underset{\boldsymbol{\theta}^t, \boldsymbol{\beta}^t, \mathbf{T}^t}{\operatorname{argmin}} \mathcal{E}^t_{data} + \mathcal{E}^t_{prior}, \tag{1}$$

with $\mathcal{E}_{data}$ and $\mathcal{E}_{prior}$ the error terms corresponding to the observations and assumptions, respectively.

Data-driven models have opened up new ways of acquiring observations. Instead of requiring constrained environments and the attachment of sensors/markers on the subjects' bodies, we can now infer said observations directly from color images. This has the potential to streamline workflows, enable new types of participatory experiences, democratize the technology itself, and allow for wider capture scenarios. Nevertheless, its reliance on precollected data and the datasets' traits have produced various models, each delivering different types of observations that also suffer from intra-type variability.

There exist AI models trained on the aforementioned dataset variants that infer 2D keypoints Fang et al. [2022], Cao et al. [2019], with varying placement across models, or models that estimate 3D joint locations Pavlakos et al. [2017], Sárándi et al. [2020] or rotations Pavlakos et al. [2018], Choutas et al. [2020], Goel et al. [2023]. More recent models have tried to address this variability by training on multiple datasets and designing ways to overcome the challenges associated with conflicting datasets Sárándi et al. [2023], Jeong et al. [2025]. Our solution to this relies on the use of a unified motion representation based on an articulated parametric body model $\mathcal{B}$. Such models Wang et al. [2020], Pavlakos et al. [2019] define a skinned mesh $(\mathbf{v}, \mathbf{f}) = \mathcal{B}(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{T})$, comprising vertices $\mathbf{v} \in \mathbb{R}^{V \times 3}$ and faces $\mathbf{f} \in \mathbb{N}^{F \times 3}$ that is simultaneously shaped, posed and positioned by parameters $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ and $\mathbf{T}$, respectively. The mesh-based representation allows us to unify all different types of constraints and simultaneously exploit direct and indirect annotations when training the observation-inferring models. Given its explicit and dense nature, it can also accommodate dense observations, such as silhouettes Ke et al. [2022], Ravi et al. [2024], Khirodkar et al. [2024], vertices Kolotouros et al. [2019], or UV maps Güler et al. [2018], Khirodkar et al. [2024].

There are many ways to solve Eq.(1), each with different capabilities and computational requirements. Still, the underlying model can be either kinematic or dynamic, and thus, two classes of approach exist, inverse kinematics Aristidou et al. [2018] or inverse dynamics Shimada et al. [2020]. While the former is more straightforward, the latter can better integrate contact forces within its underlying formulation and additionally estimates forces like torque. However, proper modeling of said forces requires knowledge about the subjects' mass and its distribution, a challenging measure to quantify. Using a mesh-based representation, we can acquire a proxy for mass via simple volume-based density estimation Tripathi et al. [2023], which improves the dynamics model and its accuracy.

While there exist many parametric body models like Google's GHUM Xu et al. [2020], and DeepDaz Yan et al. [2021], the simple and linear nature of SMPL Loper et al. [2023] renders it as the most efficient and effective one, and our choice as a unified representation. Still, it comes with a number of challenges that we need to overcome. Although primarily designed for computer graphics use, the model lacks straightforward engine integration without disregarding its pose corrective blendshapes. Also, its joints' positions are not completely symmetric, and their corresponding rotations are all defined in the same global coordinate frames. This makes it highly problematic for direct use in digital content creation tools. It also makes it hard to define proper rotation constraints and/or limits as the global reference frame prevents association with human biomechanics definitions.

Consequently, to actually benefit from the constraint unification necessary for $\mathcal{E}_{data}$ of the selected parametric body model $\mathcal{B}$, it is necessary to accompany it with a retargeting technology to adapt its skeleton/rig configuration on demand Tak and Ko [2005]. Reference frame adaptation can be achieved either explicitly or in closed form. On the other hand, complete rig topology adaptation requires numerical solutions similar to Eq. (1), and can be accomplished by exploiting the aforementioned flexible constraint definition on $\mathcal{B}$. This way, it is possible to integrate anthropometric limits as

[1] https://www.moverse.ai

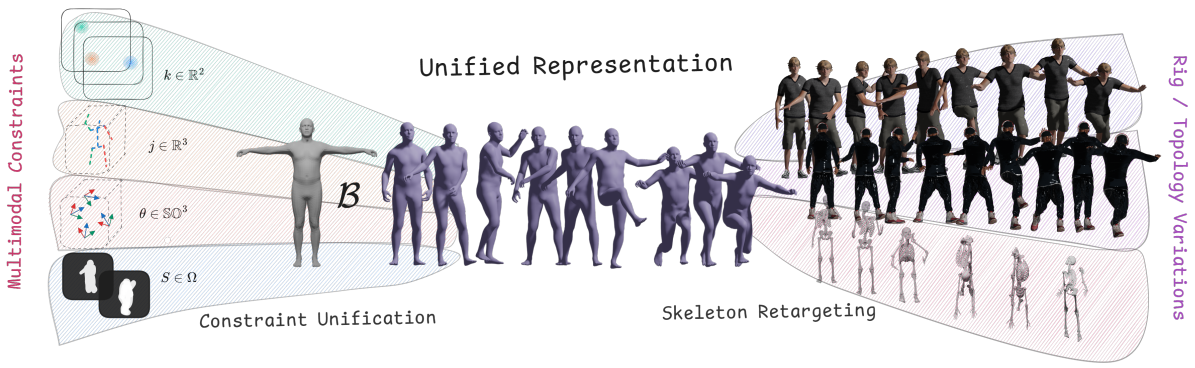


Figure 1: *Unified Representation*: The parametric body model $\mathcal{B}$ unifies multimodal constraints (keypoints $k$, joints $j$, rotations $\theta$, and silhouettes $S$) but needs to be supported by retargeting technology to overcome its topological definition weaknesses and allow for straightforward downstream use for both animation and biomechanics.

constraints Keller et al. [2023] that participate in $\mathcal{E}_{prior}$, retarget to biomechanical models Rajagopal et al. [2016] or to animation-friendly rig topologies. Closing, a unified body $\mathcal{B}$ representation supplemented by a retargeting technology, aimed at circumventing its downstream weaknesses, serves as a solid backbone for Moverse's horizontal human motion capture and generation offerings, and is illustrated in Figure 1.

## 3 Designing for Scalable, Smart Capture

Recent advances in machine learning have made monocular motion capture feasible, robust, and accurate. Still, relying on a single viewpoint, despite its simplicity and flexibility, renders single-camera motion capture a challenging task. Lifting 2D observations to 3D will always be ill-posed, with multiple solutions satisfying Eq.(1), thus suffering from ambiguity. At the same time, the inherent occlusions require strong priors (*e.g.* temporal, scale or data-driven), that will nevertheless, hallucinate and produce plausible but not accurate or faithful solutions. Although this field advances rapidly and greatly benefits from the availability of more data, synthetic or real, accurate reconstructions of human performance are necessary in many downstream motion capture applications, in fields like sports, rehabilitation, and cinematic animation.

Adding more viewpoints provides multiple benefits, as additional views help resolve occlusion ambiguity. In addition, multiview geometric constraints increase accuracy and provide robustness to outliers. However, deploying more viewpoints increases deployment complexity and, more importantly, system complexity. More cameras require higher data bandwidth, precise synchronization, and incur high computational costs. The capturing context also increases the demand for more viewpoints as larger volumes require enough viewing coverage, and multi-actor performances need more views to address higher occlusion rates. While deployment complexity increases linearly with the number of cameras, gains in robustness and accuracy are far from linear. This adds a system design requirement, namely that the system should be capable of gracefully scaling, to adapt to various scenarios of use. Even though most systems will strive to reduce the number of sensors they require both from economical and business perspectives, there will always be use cases that require more views to be deployed.

From a data transmission point of view, adding more views challenges contemporary hardware, considering their color stream payload. The contradictions are many: **i)** video encoding is efficient, but produces motion prediction artifacts with fast motions, **ii)** covering larger areas or capturing details such as finger motion require higher resolutions, **iii)** multiple higher resolution streams cannot be decoded at high rates in most GPUs, **iv)** image coding preserves quality and does not suffer from motion prediction artifacts but is highly inefficient, preventing scaling without excessive hardware and, **v)** lowering bitrates with any form of encoding reduces quality for downstream machine learning tasks Ehrlich et al. [2021], Dodge and Karam [2016]. One option for such a system would be to operate purely with high-resolution on-camera recordings. This ensures data quality, but offers a highly rigid workflow that lacks flexibility and delivers inferior user experience. To support real-time scenarios, a supporting system would be necessary to handle live streams, increasing complexity, especially when considering that cameras that record on-device and live stream are rare. Another option would be a system that supports live video streaming to support both real-time processing and simultaneous recording. Scaling this to more views and higher resolutions will greatly increase the computational demand for the receiving machine to decode multiple video streams, and then process these at high rates. Increasing resolution would be necessary to ensure data quality for fast motion and detailed areas such as the human face, but this would increase

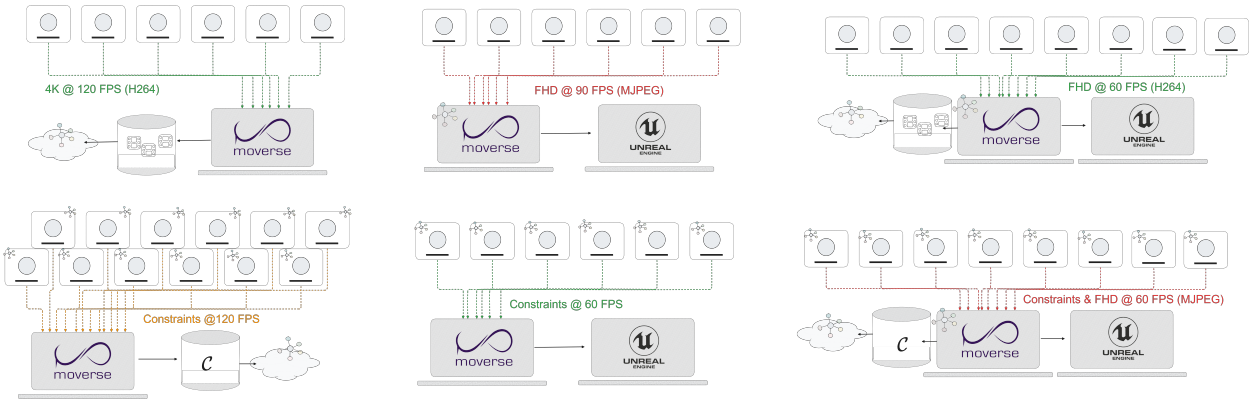


Figure 2: *System Scalability*: By distributing the processing on the edge AI cameras, and exploiting their programmable nature, a wide range of deployment scenarios can be supported. **i)** *Top left*: Pure high quality and high frame rate video recorded for offline processing, with low bandwidth requirements; **ii)** *Top middle*: High bandwidth but also high rate image coded streams processed on the central compute and streamed live; **iii)** *Top right*: Low bandwidth mid resolution video streamed and decoded for live processing, and simultaneously stored for offline reprocessing; **iv)** *Bottom left*: Large number of edge AI constraints streamed live at high rate but medium bandwidth, stored for offline processing; **v)** *Bottom middle*: Lightweight edge AI constraints streamed and processed live; and **vi)** *Bottom right*: High bandwidth dual streams of image coded color and edge AI constraints to be further processed live, streamed for pre-viz, and stored for offline reprocessing. The system can adapt its bandwidth via the sensor's programmability and distribute processing power when needed to add more viewpoints, or allow simultaneous storage and streaming, managing computational and hardware trade-offs.

complexity downstream to the centralized processor. Overall, such a system would not be highly scalable as it would need to limit the resolution, frame rate, and number of cameras to meet its computational constraints.

Marker-based systems addressed this by on-camera processing to extract marker positions, and image streams that are highly compressible after thresholding, whereas markerless systems need to transmit the entire image per viewpoint. Our scalability design is based on edge AI and software-defined on-camera processing[2]. Offloading computation on cameras improves scalability by lifting some computational costs. It can also overcome bandwidth limitations as extracting constraints on the camera can skip color image transmission, rendering the system scalable in terms of bandwidth. AI inference on the camera is also compression artifact-free and with separate hardware for video compression can be done simultaneously to sending – and recording – video feeds. A programmable camera pipeline can also be used to optimize the system to allow for joint support for rigid workflows that also record multiview video, to scalable real-time scenarios that only transmit constraints, to hybrid workflows that can both pre-visualize and record. The resulting user experience is significant as they benefit from a system that is both highly modular and scalable, adapting to their diverse capturing contexts. Overall, an edge AI based system design enables horizontal scalability without exponentially growing the central compute load, with indicative deployment scenarios illustrated in Figure 2.

## 4 Exploiting Data-driven Services

The flexibility and usability of markerless motion capture (§1), combined with a scalable system (§3), allows for significantly improved data collection throughput, and thus large data volumes. Considering the choice of a standardized representation §2, a new opportunity opens up, to offer data-driven services for collected data. The offering of such services for varying motion capture data is challenged by the differences in skeleton topology, bone / marker placement, degrees-of-freedom and, file formats. Such services may entail editing-oriented ones, like restyling, exaggerating, and blending, or may focus on capture-related support, like inpainting, filtering and, clean up.

We design such a service to reprocess the real-time captured data and simultaneously inpaint, filter and automatically clean up the motion captured data. Reprocessing is considered a standard feature of most motion capture systems to date. Motion capture sessions, even in tightly controlled environments, are prone to errors. Markers can fall off, be mislabeled, or be occluded from the cameras while inertial sensors can drift, misalign, or pick up unwanted magnetic interference noise. AI-based markerless systems are based on a different measurement basis, as their observations are not physically derived, but rather inferred from a trained model. Measurement granularity is also lower than that

[2] Luxonis Edge AI Cameras

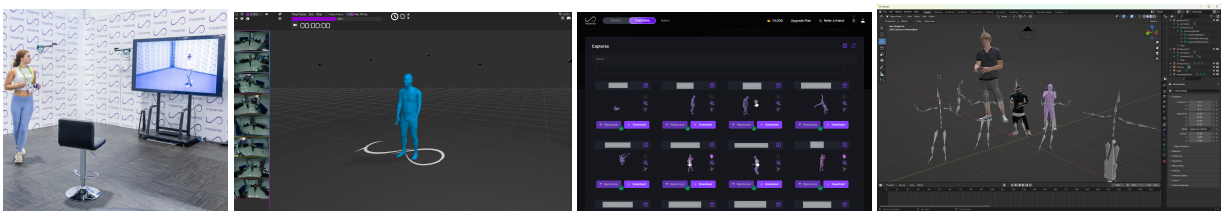

Figure 3: From left to right: live unreal streaming, the Moverse Capture Studio, the Moverse Portal, and exported assets imported in a digital content creation tool, Blender.

of marker-based high-resolution triangulations or high-rate inertial measurements. It is also an open research task to accurately calibrate the models' confidence and uncertainty, a fact that makes it mandatory to address higher outlier ratios. In addition, being optical-based means that they suffer from a subset of the marker-based system's problems, like occlusions and, improper or insufficient viewpoint coverage.

Our reprocessing service is a bundle solver that processes the entire captured sequence using a data-driven representation. The latter is trained to provide a latent representation that can be plausibly interpolated so that the underlying solver can operate on manifold segments instead of temporal frames Albanis et al. [2023a]. In addition, the solver itself calibrates the uncertainty of the measurements simultaneously with the solve by jointly optimizing their likelihood Albanis et al. [2023b]. Finally, by defining the interpolation weights to uniformly sample the motion, the solver fills in gaps, produces smooth motions, and is very robust to higher outlier ratios, correcting any artifacts that may have come up during the live capturing session. The resulting services allow users to upload their recordings to the cloud, and automatically receive a cleaner, more accurate version of the original recording without user intervention, maximizing output quality. Such a step is necessary for certain downstream use of motion capture data, either for cinematic use, or where high precision is required, for example in biomechanics research.

In addition to reprocessing the captured data themselves, the availability of multiple data points creates opportunities for data aggregation services. Such services can be offered for statistical (*e.g.* clustering) or metadata/analytics visualization purposes. However, as mentioned above, the unified representation allows us to train and refine models on these data points. Exploiting this, we design a service that trains a small, specialized generative model using a small set of motions. The result is a Motif, a lightweight inference model that generates variants of its training motions. To be able to accomplish this in reasonable time and make it useful, we implement faster training schemes Roditakis et al. [2024] and a set of additional services on top of the Motif generative model. These allow for enforcing looping and controlling the generation process using pre-defined motion segments.

## 5 The Moverse Capture Studio and Portal

The previous sections laid out the core design concepts of the envisaged motion universe. Building on a unified representation, we homogenize constraint definition and integrate this into a capturing system designed to scale to multiple use cases. The captured data benefit from a set of services that exploit the unification and standardization of the captured data. The realization of the presented design is the Moverse Capture Studio, the software that runs our motion capture system, and the Moverse Portal, our cloud-based application that hosts the users' data and the data-driven services that users can run on them.

The **Moverse Capture Studio** supports multi-view video streaming from multiple Luxonis edge AI cameras. The cameras are calibrated by moving a marker board inside the capture volume using bundle adjustment, and the calibrated system becomes gravity aligned by finishing the calibration process after placing the board on the floor. The next step entails calibrating the actors by estimating their body proportions. Once the actors are calibrated, they are tracked by the system that captures their motion (*e.g.* 3D body articulation). The system operates at high rates ($\geq$60 frames per second) to capture high-fidelity motion. The camera resolution, frame-rate, sensor parameters and programmable pipelines can be adapted according to the needs of each capturing session, as explained in §3. For real-time, pre-viz and/or interactive contexts, the data is streamed to third-party software (*e.g.* Unreal Engine, TouchDesigner). For content creation purposes, the motion capture data are recorded and uploaded to the cloud.

The **Moverse Portal** hosts a number of services for the uploaded recordings. In addition to exporting to various industry formats – *e.g.* Filmbox (.fbx), GL Transmission (.gltf/.glb) and Track Row Column (.trc) – the portal offers re-targeting and reprocessing services. The former adapts the skeleton topology and bones' orientation to popular rigs used by creators as explained in §2, while the latter automatically cleans any artifacts, smooths and infills the recordings as presented in §4. The portal also hosts a unique library, the Moverse Motifs, where certain motion archetypes are

represented by generative models. Users can generate infinite variations of these archetypes, enforcing looping, or control the generation process by selecting specific motion patterns to be enforced at specific time points. Given the unification of the motion data, the aforementioned exporting and re-targeting services are straightforwardly and without loss of quality available for the Motifs as well.

An overview of all these integrated capabilites are presented in Figure 3.

## 6 Outlook

Between the Moverse Capture Studio and Portal, we have implemented a first version and the fundamentals of the envisaged motion universe. Users can capture performances or generate motions via a unique library, and benefit from a set of core unified services. Future evolution can expand both the directions of capture and generation. For the former, the interplay between multiview video and motion capture opens up the exploration of Gaussian avatars, specifically to create photorealitic reconstructions of the actor performances. For the latter, higher-level editing services that restyle or blend motions can exploit the recent advances in the literature. Finally, retargeting to biomechanical formats will allow for downstream use of the captured data in life science applications.

## References

Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):7157–7173, 2022. 2

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019. 2

Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7025–7034, 2017. 2

István Sárándi, Timm Linder, Kai Oliver Arras, and Bastian Leibe. Metrabs: metric-scale truncation-robust heatmaps for absolute 3d human pose estimation. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(1): 16–30, 2020. 2

Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 459–468, 2018. 2

Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Monocular expressive body regression through body-driven attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 20–40. Springer, 2020. 2

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14783–14794, 2023. 2

István Sárándi, Alexander Hermans, and Bastian Leibe. Learning 3d human pose estimation from dozens of datasets using a geometry-aware autoencoder to bridge between skeleton formats. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2956–2966, 2023. 2

Uyoung Jeong, Jonathan Freer, Seungryul Baek, Hyung Jin Chang, and Kwang In Kim. Posebh: Prototypical multi-dataset training beyond human pose estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12278–12288, 2025. 2

Haoyang Wang, Riza Alp Güler, Iasonas Kokkinos, George Papandreou, and Stefanos Zafeiriou. Blsm: A bone-level skinned model of the human mesh. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 1–17. Springer, 2020. 2

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 2

Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson WH Lau. Modnet: Real-time trimap-free portrait matting via objective decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1140–1147, 2022. 2

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2

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. 2

Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4501–4510, 2019. 2

Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018. 2

Andreas Aristidou, Joan Lasenby, Yiorgos Chrysanthou, and Ariel Shamir. Inverse kinematics techniques in computer graphics: A survey. In *Computer graphics forum*, volume 37, pages 35–58. Wiley Online Library, 2018. 2

Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *ACM Transactions on Graphics (ToG)*, 39(6):1–16, 2020. 2

Shashank Tripathi, Lea Müller, Chun-Hao P Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. 3d human pose estimation via intuitive physics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4713–4725, 2023. 2

Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6184–6193, 2020. 2

Haonan Yan, Jiaqi Chen, Xujie Zhang, Shengkai Zhang, Nianhong Jiao, Xiaodan Liang, and Tianxiang Zheng. Ultrapose: Synthesizing dense pose with 1 billion points by human-body decoupling 3d model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10891–10900, 2021. 2

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. 2

Seyoon Tak and Hyeong-Seok Ko. A physically-based motion retargeting filter. *ACM Transactions on Graphics (ToG)*, 24(1):98–117, 2005. 2

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. 3

Apoorva Rajagopal, Christopher L Dembia, Matthew S DeMers, Denny D Delp, Jennifer L Hicks, and Scott L Delp. Full-body musculoskeletal model for muscle-driven simulation of human gait. *IEEE transactions on biomedical engineering*, 63(10):2068–2079, 2016. 3

Max Ehrlich, Larry Davis, Ser-Nam Lim, and Abhinav Shrivastava. Analyzing and mitigating jpeg compression defects in deep learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2357–2367, 2021. 3

Samuel Dodge and Lina Karam. Understanding how image quality affects deep neural networks. In *2016 eighth international conference on quality of multimedia experience (QoMEX)*, pages 1–6. IEEE, 2016. 3

Georgios Albanis, Nikolaos Zioulis, and Kostas Kolomvatsos. Bundlemocap: Efficient, robust and smooth motion capture from sparse multiview videos. In *Proceedings of the 20th ACM SIGGRAPH European Conference on Visual Media Production*, pages 1–9, 2023a. 5

Georgios Albanis, Nikolaos Zioulis, Spyridon Thermos, Anargyros Chatzitofis, and Kostas Kolomvatsos. Noise-in, bias-out: Balanced and real-time mocap solving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4237–4247, 2023b. 5

Konstantinos Roditakis, Spyridon Thermos, and Nikolaos Zioulis. Towards practical single-shot motion synthesis. *arXiv preprint arXiv:2406.01136*, 2024. 5