arXiv:2506.13040v3 [cs.CV] 1 Dec 2025

MAMMA: Markerless Accurate Multi-person Motion Acquisition

Hanz Cuevas-Velasquez[1], Anastasios Yiannakidis[1], Soyong Shin[2], Giorgio Becherini[1], Markus Höschle[1], Joachim Tesch[1], Taylor Obersat[1], Tsvetelina Alexiadis[1], Eni Halilaj[2], Michael J. Black[1]

[1]Max Planck Institute for Intelligent Systems, Germany [2]Carnegie Mellon University, USA



Figure 1. From dense surface landmarks to SMPL-X: MAMMA accurately reconstructs pose and shape from synchronized multi-view videos.

Abstract

We present MAMMA, a markerless motion-capture pipeline that accurately recovers SMPL-X parameters from multi-view video. Traditional motion-capture systems rely on physical markers. Although they offer high accuracy, their requirements of specialized hardware, manual marker placement, and extensive post-processing make them costly and time-consuming. Recent learning-based methods attempt to overcome these limitations, but most are designed for single-person capture, rely on sparse keypoints, or struggle with occlusions and physical interactions. In this work, we introduce a method that predicts dense 2D surface landmarks conditioned on segmentation masks, enabling person-specific correspondence estimation even under heavy occlusion. We employ a novel architecture that exploits learnable queries for each landmark. We demonstrate that our approach can handle complex person–person interaction and offers greater accuracy than existing methods. To train our network, we construct a large, synthetic multi-view dataset combining human motions from diverse sources, including extreme poses, hand motions, and close interactions. Our dataset yields high-variability synthetic sequences with rich body contact and occlusion, and includes SMPL-X ground-truth annotations with dense 2D landmarks. The result is a system capable of accurately capturing human motion without the need for markers. Our approach offers competitive reconstruction quality compared to commercial marker-based motion-capture solutions, without the extensive manual cleanup. Finally, we address the absence of common benchmarks for dense-landmark prediction and markerless motion capture by introducing two evaluation settings built from real multi-view sequences. We will release our dataset, method, code, and model weights for research purposes.

1. Introduction

Traditional marker-based motion-capture (mocap) methods [40, 44, 50, 62] are widely used to capture 3D human movement for varied applications. These systems rely on expensive, specialized hardware and optical markers that are time-consuming to apply. Most require trained personnel to configure and operate, and involve laborious procedures to clean up noisy results, especially with complex poses and close multi-person interactions due to markers falling off or occlusions. Close interactions are common in daily life, from hugging to interpersonal activities such as martial arts or dancing. Yet such challenging motions remain underrepresented in existing motion-capture datasets [35].

To address these issues, numerous attempts have

been made to replace marker-based systems with markerless ones. These capture a subject using multiple video cameras and apply various computer vision algorithms to track the 3D-body motion across frames. While commercial markerless systems exist [5, 41, 42, 58], they remain expensive and are generally considered to be less accurate than marker-based systems (the "gold standard").

Given the rapid advances in computer vision, we ask whether it is possible to "roll your own" markerless system that achieves accuracy comparable to commercial marker-based systems without the tedious manual clean up. Making such a system available for academic research would democratize motion capture and increase the scale at which complex motion data could be captured.

Here we develop such a system, called MAMMA (Markerless Accurate Multi-person Motion Acquisition), and make it freely available for academic research purposes. MAMMA takes multi-camera video, and outputs SMPL-X [47] bodies at each frame for every person. We focus on the two-person case with close interactions where close contact is problematic for marker-based methods.

To achieve this, we introduce several key innovations and design choices. We adopt a two-stage approach of estimating virtual markers in all camera views and then solving for the optimal human body that fits these markers. Previous work has taken a similar approach, either estimating sparse anatomical markers like 2D joint locations [61] or dense surface markers [16, 45] from every camera view. We follow the latter direction and train a dense 2D-landmark detector. However, landmark detectors are usually trained on images of people in isolation, making it difficult to guide the network to identify the correct person or correctly assign body parts when subjects have close-range interactions. Also, for real datasets it is hard to obtain fine-grained information like contact or visibility.

We address this in two ways. First, we extend the synthetic human video dataset BEDLAM [3] by assembling close-interaction motions (e.g., dancing) and rendering realistic multi-view sequences. Then, we define dense surface landmarks [16, 45] and construct a new dataset, MammaSyn.

Second, rather than train the model to directly regress keypoints, we formulate a novel transformer-based dense landmark estimator, MammaNet, which learns distinct embeddings for each keypoint, which we refer to as landmark queries. We then train a transformer decoder that takes these queries together with the image patch embeddings and outputs landmark locations, together with uncertainty, visibility and contact information. We find visibility and contact estimates to be critical for dealing with close person–person interaction and to prevent interpenetration. We demonstrate that our approach is more robust than traditional landmark estimation.

Our method can also be optionally conditioned on additional information to help resolve ambiguities in the multi-person case. In particular, we use SAM 2 [51] to compute and track masks corresponding to the people, and condition the network on these. This highlights a key advantage of vision-based systems over marker-based ones – richer pixel-based supervisory signals can be incorporated to reduce ambiguity.

From the estimated landmarks, uncertainty, visibility and contact, we fit SMPL-X by projecting landmark locations on the body mesh into each calibrated image, and optimizing the pose and shape parameters to minimize 2D reprojection error. Thanks to the accuracy of our landmarks, optimization does not rely on pose priors or pose initialization. Our resulting system allows us to process complex movements, with a level of accuracy that is visually indistinguishable from that of a marker-based system (see Fig. 1).

We perform extensive ablations to evaluate which design choices are important and evaluate performance on video datasets with both a single subject and two-person interaction. The latter includes Harmony4D [26], CHI3D [12], and Hi4D [73]. We find that our MAMMA outperforms all academic methods.

However, a key question remains: how does the method measure up against a commercial, marker-based system such as Vicon? To answer this, we perform a *unique quantitative evaluation* of our pipeline versus one that fits SMPL-X to Vicon markers using MoSh++ [35]. We use held-out 3D markers to evaluate how well each method captures the 3D body shape and pose. We find that the difference in held-out marker error between the methods is only 0.862mm and that resulting animations are visually indistinguishable. This suggests that our markerless approach can be used as "ground truth," replacing marker-based systems for SMPL-X capture.

2. Related Work

Traditional Marker-Based Motion Capture. Commercial marker-based motion-capture systems, such as Vicon [62], OptiTrack [44], Qualisys [50], Motion Analysis Corporation [40], and PhaseSpace [48], are widely used in biomechanics, animation, and virtual reality to capture precise 3D motion data. These systems employ either passive or active markers to track movement, providing high-fidelity data. However, placing markers, calibrating for both the environ-

ment and subject, and cleaning the data introduces substantial overhead in the capture process. Captured sequences often require extensive manual post-processing to fix noisy, missing, or swapped markers, a task that can take from minutes to hours depending on the complexity of the motion. In addition, extra processing is required to convert the raw capture data into the parameters of a 3D skeleton or parametric body model [32]. Our pipeline removes the need for these manual steps by exploiting rich information present in the pixels to directly regress dense 2D surface landmarks, recovering the shape, pose, and translation parameters of the SMPL-X model from multi-view video of one or two subjects.

Vision-Based Markerless Motion Capture. Several industry solutions for markerless motion capture have been proposed in recent years [5, 41, 42, 58]. However, with the exception of Moverse [42], no such methods are capable of producing the parameters of a parametric body model such as SMPL [33]. They also do not report accuracy on standard benchmarks, making it difficult to compare solutions or evaluate their accuracy. Unfortunately these implementations are not open source. Consequently, here we focus on academic systems.

Markerless motion capture has a long history. Many vision-based approaches focus on detecting sparse 2D human-body keypoints from monocular video sequences, either for individuals [49, 66] or crowds [4, 10]. More-recent monocular methods adopt parametric body models such as SMPL [33] and SMPL-X [47] as priors and regress the parameters of these models from a single view [14, 24, 45]; these struggle with depth ambiguity and occlusion when people closely interact. To capture accurate 3D human pose, multi-view systems triangulate 2D detections or learn cross-view feature representations [21, 29, 55, 64]. As public multi-view datasets [19, 23, 37] typically annotate only sparse keypoints, such methods stop at the skeleton level, and require an additional model-fitting stage, necessitating strong priors, to obtain full surface parameters. Instead, we predict a dense set of surface landmarks, rich enough to solve for both SMPL-X body and hand pose and shape, without heavy manual priors. By conditioning the landmark regressor on an instance-segmentation mask, our proposed approach is able to cleanly separate overlapping subjects, remaining robust in scenes with close human interactions.

Motion-Capture Datasets. Current progress in human pose and shape estimation has enabled the creation of a number of human-centric datasets. Several of them provide single-person sequences with mesh pseudo-ground truth annotations [17, 20, 25, 54, 59, 60, 63]. Others, provide labeled sequences of human–object interaction [2, 9, 15, 56]. There are datasets that provide human–human interactions [22, 36], however these interactions are limited to simple actions where often both people are apart from each other. Creating datasets with heavy interaction between people remains challenging due to frequent occlusions and such datasets are restricted to laboratory environments [1, 12, 23, 26, 38, 73].

The development of synthetic datasets has played a crucial role in advancing research on human pose and shape estimation, particularly in addressing the limitations of real-world data in terms of scale, diversity, and annotation accuracy. Synthetic data makes it possible to obtain large-scale, densely labeled samples under diverse conditions, free from the constraints of physical capture. However, most existing synthetic datasets do not contain person–person interaction [3, 16, 28, 46, 57]. Among those that do [11, 72], the range of activities and motion diversity is limited.

3. Dataset

3.1. MAMMA Synthetic Training Data

We train MammaNet exclusively using synthetic data. To extend the coverage of existing data, targeting (i) multi-view coverage, (ii) extreme pose, (iii) high-fidelity hands, and (iv) two-person interaction scenarios, we introduce MammaSyn, an extension of BEDLAM [3], containing 2.5M samples (crops).

We extend the original BEDLAM pipeline to a 32-camera multi-view configuration using a virtual camera setup. We sample motion sequences from available datasets and capture two additional marker-based (Vicon) datasets (Fig. 2): (i) Latin-Dance (2 subjects, 10 dancing sequences), and (ii) Interacting Couples (2 subjects, 48 two-person interaction sequences). We collect these to supply high-quality interaction sequences missing from BEDLAM. We represent all the 3D humans in SMPL-X [47] format. See SupMat for rendering details. We save per-subject segmentation masks, depth maps, and per-vertex visibility. We also generate per-vertex contact labels for floor and subject interactions by computing the Sign Distance Function (SDF) along with the surface normals. The resulting dataset consists of 955k images (Fig. 2).

For the dense-landmark ground truth, we sample 512 vertices from the SMPL-X body using Farthest Point Sampling (FPS), assigning more weight to vertices on the hands, feet, and head. This provides the network more information for smaller and more articulated body regions, as seen in Fig. 3. Our curated dataset can be categorized into three sub-datasets:



Figure 2. Cropped samples from MammaSyn dataset.

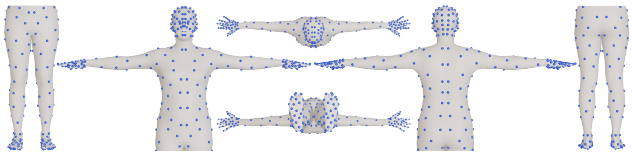


Figure 3. 512 Landmarks sampled from the SMPL-X body.

MammaSyn-S (Single Person): We sample SMPL-X motion sequences from the BEDLAM data pool, which offers a wide range of motions and body shapes. To broaden the pose distribution, we augment it with sequences from the MOYO training set.

MammaSyn-I (Interactions): In addition to our own motion-capture recordings of Latin-dance routines and other two-person interactions, we include sequences from the Hi4D [73], Harmony4D [26], and Inter-X [69] training sets. The resulting collection covers a diverse range of contact-rich interactions, from social exchanges to coupled dances.

MammaSyn-H (Hands): To improve the diversity of the sampled hand articulations we add SignAvatars [74]. We also introduce hand motions from Interhand2.6M [39], which contains captured sequences of hand-to-hand interaction. We merge these hand poses with the body parameters of BEDLAM by fitting SMPL-X to the MANO [52] hand sequences (Fig. 2).

3.2. MAMMA Real Evaluation Data

We captured three datasets to evaluate our method: MammaEval-(S)ingles (16 views), MammaEval-(D)ance (32 views), and MammaEval-Extra (32 views). All datasets were captured with a marker-based system (Vicon) synchronized with a multiview RGB system [18]. Marker data was manually cleaned and then processed with MoSh++ [35] to obtain SMPL-X annotations paired with multi-view images.

Each dataset serves a different evaluation purpose. MammaEval-Singles consists of 3 individuals in 22 single-subject sequences, including object interaction. MammaEval-Dance includes 18 West Coast Swing dance sequences performed by a professional dance couple, whose close partner interaction is intrinsic to the genre. MammaEval-Extra was acquired explicitly to benchmark our method against the Vicon-MoSh marker-based pipeline. In addition to using a standard Vicon marker set (FrontWaist10Fingers), we add 37 additional markers on each of the 3 subjects. These extra markers are not used in fitting SMPL-X with MoSh++, and provide independent ground-truth, enabling fair comparison between marker-based and markerless methods. Subjects performed 4 sequences each, including walking, dancing, and basic sports motions. SupMat has more details regarding the marker-based capture pipeline and configuration.

As part of our contribution, we additionally release a new dance dataset captured solely using MAMMA, covering Bachata, West Coast Swing, Breakdance and Ballroom. It consists of 9 subjects performing for approximately 1 hour in total, recorded at 30 FPS.

4. Method

4.1. Dense Landmark Estimation

Our novel dense 2D landmark estimator, MammaNet, extracts image features using ViT-Base and includes an additional CNN to process the mask (Fig. 4). It decodes $N = 512$ surface landmarks using a Transformer decoder. Unlike CameraHMR [45], which uses a single learnable embedding input to estimate the entire set of landmarks, $\mathbf{L}$, we learn N embeddings (landmark queries), with one corresponding to each landmark. By designing the network to decode landmarks from individual queries, cross-attention layers learn to match each landmark to the most relevant image patches. At the same time, the self-attention layers in the decoder are able to learn the pairwise correlations between landmarks. We design the network so that the image and mask features are encoded to the same space, where we combine them with element-wise summation to integrate mask conditioning [27]. We find the proposed network architecture enables the model to generalize to challenging poses (see Fig. 5).

For each landmark index $i \in \{0, ..., N-1\}$, the network predicts pixel coordinates $\mu_i = [x_i, y_i]$ and the associated uncertainty σ_i [16, 45, 67]. We observed that the predicted uncertainties are often high on rigid regions (e.g. torso) that are not visible by the camera. Due to this, we additionally task the network with estimating the visibility probability p_i for each landmark. We also introduce a person–person contact pc_i and a floor contact fl_i prediction layers. To train the dense landmark detector, we use a Gaussian negative log likelihood loss. For the visibility probability, we use binary cross entropy. For the contact probabilities, we use focal loss [30], because there are more landmarks that do not have contact. All the losses have per-landmark weights λ^*. We denote $\mu'_i = [x'_i, y'_i]$ and $p'_i \in \{0, 1\}$ as

ground truth landmark and visibility, respectively.

4.2. Model Fitting

Our method fits the SMPL-X neutral body, $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{t})$, to a multi-view sequence of one or two people using L-BFGS [31]. We initialize all the SMPL-X bodies at the 3D point that best minimizes the distance between the rays projected from the camera centers. Unlike other methods [16, 45], we do not initialize the pose and shape of the person with any regression method, as our surface landmarks carry enough information about the person and the scene. We assume the calibration parameters are given. We use 16 shape coefficients ($\boldsymbol{\beta}$).

Our optimization process is as follows. The first step involves finding the translation and rotation of the bodies by minimizing the reprojection error:

$$E_{\text{ldmks}} = \frac{1}{C} \sum_{t,c,l} \rho\left(\frac{\|\boldsymbol{\mu}_{t,c,l} - \Pi(\mathbf{V}_{t,l}, \mathbf{Q}_c)\|}{\sigma_{t,c,l}} \right) p_{t,c,l} \quad (1)$$

where E_{ldmks} is the energy term that penalizes the reprojection error between $\mathbf{L}$ and the projected SMPL-X vertices $\mathbf{V}_{t,l}$ using camera parameters $\mathbf{Q}_c$ and where $[\boldsymbol{\mu}_{tcl}, \sigma_{tcl}, p_{tcl}]$ is the predicted 2D location, uncertainty and visibility probability for the l^{th} landmark. ρ is the robust Geman-McClure function [13]. The second stage optimizes the pose, shape, and translation parameters with E_{ldmks} and an L2 regularizer on the body priors E_{shape}. In the third stage we use the same energies but update the uncertainty values. Sometimes, the network is not confident about some predictions, even when the location of the landmark is correct. Thus, we use the reprojection error e_i as evidence to weight down the uncertainties, $\sigma_i' = \sigma_i \cdot \min\left(\max\left(\frac{e_i}{\tau}, 0\right), 1\right)$, where $\tau = 10$px. This is possible because the second stage result is close to local optimum. We also penalize large joint accelerations E_{temp}.

As final stage we optimize the contact. Assuming that each per-view network output provides an approximately unbiased and calibrated estimate of the local contact probability, conditioned on that view, we average the contact probabilities to estimate the expected contact probability. We do this for foot and person–person contact predictions and use them to minimize $E_{\text{cont}} = E_{\text{p}} + E_{\text{c}}$. where $E_{\text{p}} = \frac{1}{N} \sum_{i=1}^{N} \left[\min\left(0,\ \text{SDF}_{\text{other}}(\mathbf{v}_i) + \delta\right) \right]^2$ is a repulsion term that penalizes all the vertices that are inside the body of the other person. δ allows for limited interpenetration to account for soft-tissue deformation. $E_{\text{c}} = \frac{1}{N} \sum_{i=1}^{N} p_i \left[\max\left(0, \text{SDF}_{\text{other}}(\mathbf{v}_i)\right) \right]^2$ is a contact term that attracts the points that are above the surface, this term is weighted by the predicted contact probabilities and their corresponding vertices. SDF is the Sign Distance Function from [43]. We re-implemented the SDF with a custom CUDA kernel for full GPU acceleration.

The final energy function is $E(\boldsymbol{\Phi}; \mathbf{Q}, \mathbf{L}) = E_{\text{ldmks}} + E_{\text{shape}} + E_{\text{temp}} + E_{\text{cont}}$, where $\boldsymbol{\Phi} = \{\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{t}\}$ are the SMPL-X parameters, $\mathbf{Q}$ represents the camera parameters, and abusing notation $\mathbf{L}$ represents the predicted landmarks, uncertainty, visibility and contact probabilities.

4.3. Multiview Correspondence

Accurate subject identification is essential within each camera view and across multiple views, for convergence in the model fitting. We first obtain the masks with SAM2 [51] which can be initialized once in several ways depending on the setting: either with automatic subject detection using a Grounding LLM or a bounding-box detector, or manually with a few clicks on the the subjects in a single frame.

We use SAM2 segmentation propagation as a way to track a person across the sequence. For each tracked person, we predict their 2D dense landmarks per frame $f \in \{0, \ldots, F-1\}$. Therefore, for each view $v \in \{0, \ldots, V-1\}$, each person $p \in \{0, \ldots, P-1\}$ has $\mathbf{x} \in \mathbb{R}^{F \times N \times 2}$ 2D landmarks. If the person is not visible in some frames, we assign a dummy value.

For each tracked person, we predict 2D dense landmarks for every frame $f \in \{0, \ldots, F-1\}$, view $v \in \{0, \ldots, V-1\}$, and person $p \in \{0, \ldots, P-1\}$, resulting in $\mathbf{x}_{v,p} \in \mathbb{R}^{F \times N \times 2}$. If the person is not visible in a given frame, the corresponding value is ignored.

All detected people across all views are compared pairwise to establish cross-view correspondences. Given two sets of landmarks $\mathbf{x}_a, \mathbf{x}_b$ from two views, their geometric affinity is defined as $A_g(\mathbf{x}_a, \mathbf{x}_b) = \exp(-D_g/\lambda) \in [0, 1]$ where λ is a temperature parameter and D_g (Eq. (2)) is the symmetric epipolar distance [6]. To avoid possible wrong predictions from non-visible landmarks, we only compute the affinity between the landmarks that are visible by both views.

$$D_g = \frac{1}{2FN} \sum_{i=1}^{FN} \left(d(\mathbf{x}_b^i,\ \mathbf{F}_{ba} \mathbf{x}_a^i) + d(\mathbf{x}_a^i,\ \mathbf{F}_{ab} \mathbf{x}_b^i) \right) \quad (2)$$

$\mathbf{F}_{ba}$ is the fundamental matrix between two views, $d(\mathbf{x}, \mathbf{l})$ denotes the point–line distance in pixel coordinates. If fewer than a minimum number of landmarks are visible or if D_g exceeds a threshold, the affinity is set to zero. The pairwise affinities between all detections in views a and b form the matrix $\mathbf{A}_{ab} \in [0, 1]^{M_a \times M_b}$, where M_a and M_b denote the number of detected persons in each view. Optimal

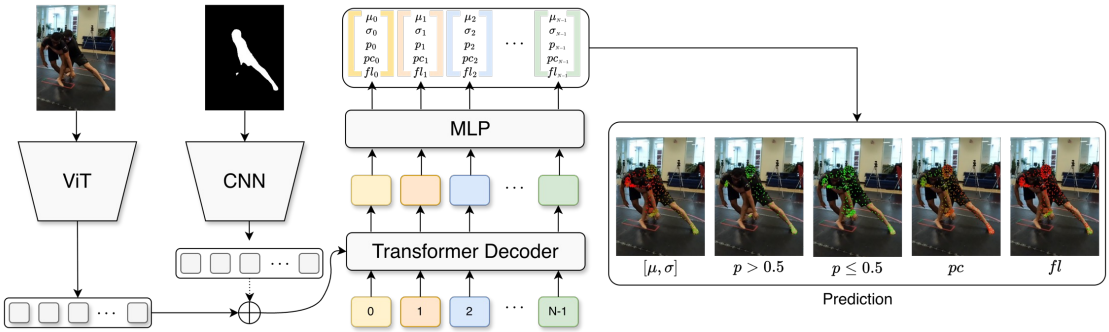


Figure 4. MammaNet. The input to the network is the image and mask. It predicts per landmark visibility probability p (green is visible, red not visible), landmark locations μ, uncertainties σ (red means highly uncertain), person–person pc and floor contact fl probabilities (red means no contact and green contact).

one-to-one matches are obtained by solving a Hungarian assignment on the cost matrix $1 - \mathbf{A}_{ab}$. All high-affinity matches are then linked across all view pairs to form a cycle-consistent correspondence graph, whose connected components represent consistent multi-view groups corresponding to the same person across views.

5. Experiments

5.1. 2D Landmark Network Comparison

We compare MammaNet with two existing dense-landmark models: CameraHMR [45], which uses the same backbone as ours and a transformer decoder with a single learnable token and output for all the landmarks. The second model is a re-implementation of Hewitt et al. [16]; their code is not publicly available. For fair comparison, we remove the pose and shape heads and only use the landmark head. We refer to the resulting network as Look-Ma*.

The input resolution is set to 512×384, consistent with the transformer-based models [70]. As the other models do not predict contact, we trained a version of our model that only predicts uncertainty and visibility. For implementation details see SupMat.

All models are trained on cropped images from BEDLAM and evaluated on real datasets. For single-person evaluation, we use RICH [17] and MOYO [59]. For two-person interaction, we evaluate on Harmony4D [26] and CHI3D [12]. We also report results on our datasets: MammaEval-S and MammaEval-D.

Tab. 1 and Tab. 2 show the average 2D error between the ground truth 2D landmarks and the predicted landmarks across images for single and two-person datasets. Tab. 4 reports standard 3D errors after fitting the model to the predicted markers; i.e. Mean Per-Joint Position Error (MPJPE) and Per-Vertex Error (PVE). For both 2D marker prediction and 3D fitting, MammaNet is more accurate that prior methods.

Table 1. Dense landmark evaluation on single person datasets. Mean 2D Euclidean distance error (in pixels) between GT and predicted landmarks. Bold is the most accurate and underline is the most accurate without mask.

Model	RICH	MOYO	MammaEval-S
Look-Ma*	13.26	22.43	10.25
CameraHMR	8.84	12.53	6.32
MammaNet (ours)	**8.55**	11.40	**6.09**
MammaNet +masks SAM2 (ours)	8.83	**11.04**	6.16

Table 2. Two-person datasets dense landmark evaluation. We use the images where IOU > 0.5 between two people.

Model	Harmony4D	CHI3D	MammaEval-D
Look-Ma*	31.45	8.77	15.01
CameraHMR	32.84	6.30	10.21
MammaNet (ours)	31.96	6.22	9.87
MammaNet +masks SAM2 (ours)	**18.33**	**4.36**	**7.70**

Our network can also generalize to complex and unseen poses, unlike CameraHMR and Look-Ma*, and predict the limbs of the person more accurately. Fig. 5 and Tab. 1 shows that MammaNet correctly estimates the landmarks from the yoga dataset (MOYO) even though it was never trained on such poses.

We additionally train a variant of our network conditioned on segmentation masks. During evaluation, we use masks obtained using SAM2. While masks can provide a marginal benefit to the performance of the network in the single-person case (Tab. 1), their main impact is on two-person interactions (Tab. 2), where masks help to disambiguate the target individual. For Tab. 2, we restrict the evaluation to images where the Intersection-over-Union (IoU) between two people is greater than 0.5. See SupMat for examples.

5.2. Contact Prediction from 2D Landmarks

The previous evaluation shows that our network is capable of differentiate between two people given a noisy mask. However, we cannot obtain contact information from the BEDLAM dataset as it has single people

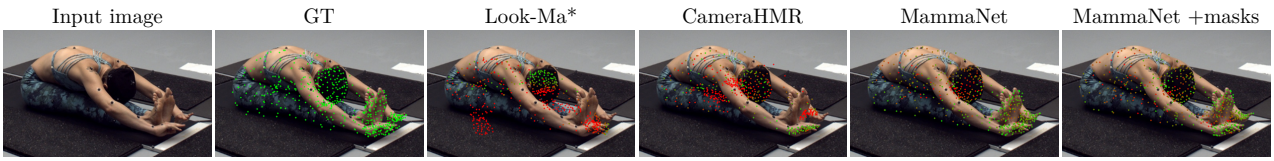
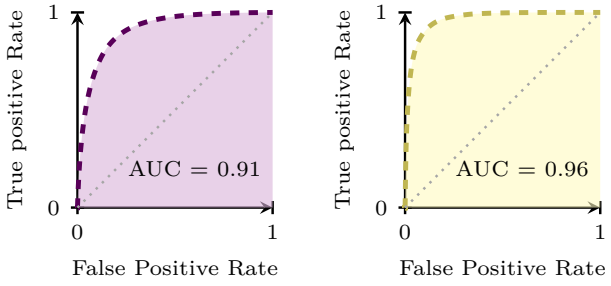


Figure 5. Comparison on extreme poses. Ground-truth landmarks are shown in green. For each prediction, landmarks are color-coded: red indicates higher pixel error, green indicates lower pixel error. We compare networks trained on BEDLAM.



(a) Hi4D person contact (b) MOYO floor contact

Figure 6. ROC curve evaluation of our contact predictions.

motions. Therefore, we retrain our network with our curated MammaSyn and added two branches. One to predict the probability of contact between two people and the second one to predict the probability of floor contact. Note that unlike [53, 65] this is not only foot contact, but any contact that the body has on the floor. To evaluate contact, we compute the Receiver Operating Characteristic curve (ROC) and report the area under the curve (AUC) where AUC=0.5 is a random guess and higher AUC is better. For floor contact, we use the MOYO test set, since its yoga poses involve complex floor contact. For person–person contact we use Hi4D contact ground truth. For the latter, we retrain our network without the Hi4D sequences. For both methods we subsample the SMPL-X vertices to the number our network predicts. As Fig. 6 reports, the AUC for both predictions is more than 90% which indicates that our model correctly identifies most contact points with few false positives.

5.3. Correspondence Evaluation

We run our correspondence method on all three interaction datasets. The input is the dense landmark and visibility predictions. We obtain a 100% accuracy in matching the identity across views; see SupMat for ablation. This demonstrates that our dense landmarks provide rich and accurate geometric information, removing the need for an extra identity feature computed by external networks similar to classical approaches [6]. Note that the resulting masks from SAM2 can miss some parts of the body, especially when two people are too close, which might lead to sub-optimal landmark predictions. However, our method remains robust.

Table 3. Mean Penetration (M.P.) depth (mm) and vertices on Harmony4D, CHI3D, and MammaEval-D.

Model	M.P.Depth	M.P.Verts.
Look-Ma*	13.73	990.61
CameraHMR	13.41	678.10
Ground truth	9.84	479.94
MAMMA (ours)	10.50	456.05
MAMMA-C (ours)	**8.46**	**378.02**

5.4. Markerless Motion Capture Evaluation

We fit SMPL-X parameters to the test sequences with our method, Look-Ma* and CameraHMR [45], as well as a multi-view implementation [8] of SMPLify-X [47]. For MAMMA, we test two versions, one with the contact optimization stage (MAMMA-C) and another without it. For a fair comparison, we use the ground truth bounding boxes for all the methods, and SAM2 masks for the mask-conditioned network.

As a proxy for ground truth for our MammaEval datasets, we use Vicon markers fit using MoSh++ [35], and for all methods, we predict pose, shape and global translation. Tab. 4 reports the MPJPE (mm) for each of the 55 SMPL-X joints along with the PVE. We report the error for the body and hands separately in SupMat. MAMMA, even without optimizing contact achieves better result than previous methods. However, human reconstruction metrics do not tell the whole story as the ground truth presents some interpenetration which can affect the score. For this, we computed the average depth penetration and number of penetrating vertices across the interaction datasets. In Tab. 3 we can see that MAMMA results in bodies that have less penetration, indicating that overall MAMMA-C achieves the best results. See Fig. 7 and SupMat for more results. We also noticed that optimizing for floor contact does not improve the results, suggesting that our network landmarks are already robust enough without it.

We also compare MAMMA on the Hi4D dataset. For a fair evaluation, we retrained our network and removed the sequences that used Hi4D. The dataset is in SMPL format, so we transform our SMPL-X mesh to SMPL, and regress 19 joints following Lu et al. [34].

Table 4. Benchmark 3D fitting errors (mm).

Model	RICH		Harmony4D		CHI3D		MammaEval-S		MammaEval-D		MOYO	
	MPJPE	PVE	MPJPE	PVE	MPJPE	PVE	MPJPE	PVE	MPJPE	PVE	MPJPE	PVE
SMPLfiy	96.18	71.42	-	-	67.68	51.79	47.15	35.42	53.92	43.08	62.15	44.68
LookMa*	39.52	30.29	59.37	45.6	46.47	39.36	25.97	23.94	27.98	24.89	60.15	53.82
CameraHMR	25.61	21.36	58.59	42.0	40.8	34.61	15.25	18.43	20.41	21.06	33.75	33.74
MAMMA	22.20	19.76	**45.26**	**34.02**	38.01	32.84	12.96	17.18	**17.71**	19.80	22.95	25.48
MAMMA-C	**22.20**	**19.76**	45.35	34.05	**37.96**	**32.82**	**12.96**	**17.18**	17.73	**19.78**	**22.95**	**25.48**

Table 5. MPJPE on the Hi4D dataset for 19 SMPL joints.

Method	MPJPE (mm)
MvP [64]	92.77
Graph [68]	89.62
Faster VoxelPose [71]	68.40
MVPose* [6]	53.05
MVPose [7]	42.63
4DAssociation [76]	41.29
AvatarPose [34]	32.10
MAMMA-C (Ours)	**12.44**

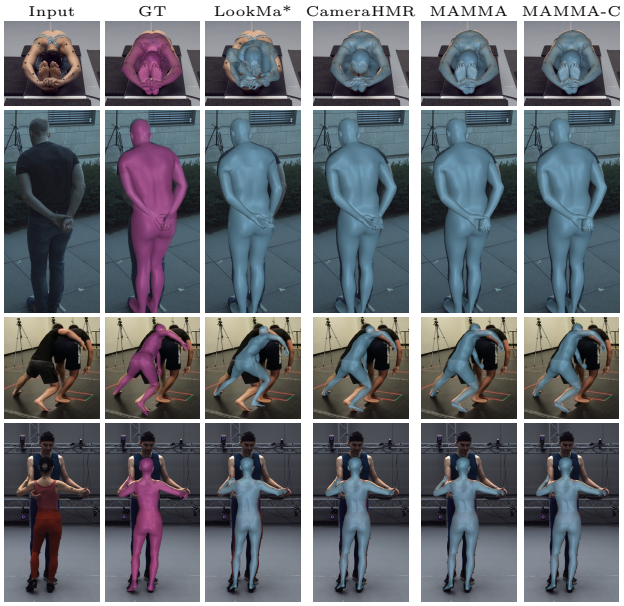


Figure 7. Mesh fitting comparison. Top to bottom, MOYO, RICH, Harmony4D, MammaEval-D. Notice how MAMMA has better fit to the image than the GT body shape (first, third and fourth row). Also the hand contact on the fourth.

Tab. 5 shows that our method outperforms previous methods by a large margin, even though our network was not trained on any sequence of Hi4D. In SupMat we compare against Harmony4D and show that our method produces better fits. We do not compare with pure feed-forward approaches because they need to be retrained for each multi-view configuration. We also do not compare with single-frame camera methods because they can not recover the global translation and rotation of the mesh.

5.5. VICON Markers Comparison

But how do the results match up against those of Vicon? Vicon believably claims sub-millimeter accuracy in locating markers. However, the estimated body is subject to noise from soft-tissue artifacts and marker misplacement, and it is the accuracy of this inferred model that is important. To evaluate this, we capture MammaEval-Extra with a marker-based system and fit SMPL-X to a standard Vicon marker set (FrontWaist10Fingers) using MoSh++ [35]. For each of the 3 subjects, we recorded 4 sequences using both the standard marker set and with 37 additional markers that are not used by MoSh++. Using the SMPL-X bodies estimated by MoSh++ and MAMMA, we predict the held-out marker locations and compute the average 3D Euclidean distance error. The Vicon + MoSh++ method has an error of 21.619mm, while our method has an error of 22.481mm.

Our takeaway is that there is only a minimal, 0.862mm difference between our method and the process used to construct the AMASS dataset, which has been widely used for many tasks. This suggests that our markerless approach is sufficiently accurate to replace the traditional pipeline. Our method reduces the full capture-to-SMPL-X pipeline from $\sim$ 72 hours to 26 hours on a consumer GPU—making it nearly three times faster than the manual marker-based workflow. See SupMat for details.

6. Conclusions

We have demonstrated a novel architecture for multi-view markerless capture of human motion in SMPL-X format. We address the challenging case of close multi-person interaction through algorithm design choices and novel datasets. MAMMA is able to predict dense 2D landmarks robustly even with extreme poses and occlusion using novel per-landmark learnable queries. We provide a unique quantitative comparison between MAMMA and the gold-standard marker-based solution and find that our approach achieves competitive reconstruction quality with significantly lower cost. Code and data will be available for research.

Supplementary Material

7. Landmark Networks Training Details

Our network, CameraHMR [45] and LookMa*[16] are trained for 300K iterations using 4 NVIDIA A100 GPUs, with a per-GPU batch size of 24 and gradient accumulation set to 2 steps. The training time took around 3 days. We use the Adam optimizer with 500 warm-up iterations. For MammaNet and CameraHMR, we initialize the transformer backbone with ViTPose-B weights pretrained at a resolution of 256×192. Following Zhang et al. [75], we interpolate the positional embeddings to increase the effective input resolution to 512×384. Hewitt et al. [16] use HRNet-W48 as the backbone. We follow their configuration, except we initialize the network using pose HRNet weights trained on the COCO dataset. The input resolution is set to 512×384, consistent with the transformer-based models. As the other models do not predict contact, we trained a version of our model that only predicts uncertainty and visibility.

8. MammaSyn Dataset

Each rendered sequence contains 2–6 synthetic subjects, placed at random positions within the capture volume. They are randomly assigned one of 100 skin textures which are overlaid with one of 1.7K garment textures.

In total, we render 2.8k sequences (5.5 hours). Each sequence is rendered from 8 views, where the first view is randomly selected and the rest via the Farthest Point Sampling (FPS) algorithm on our 32-camera setup, maximizing spatial coverage and projected-pose diversity. For lighting and background variety, we use a pool of 95 HDR images. We render at 6 fps, 2056×1504px, roughly twice the resolution of BEDLAM, to ensure much finer detail in the hand regions. in Tab. 6 we show the dataset composition by number of images.

In Fig. 8 and Fig. 9, we show more cropped samples of our dataset, and in Fig. 15 and Fig. 16, we show the contact labels.

9. Dataset Evaluation

We evaluate our proposed dataset, MammaSyn, on landmark prediction accuracy, particularly in cases involving occlusions caused by interactions. We use the best-performing network, MammaNet +masks SAM2, pre-trained on BEDLAM (B), and train it for another 300K iterations on our dataset. We compared against the network trained only on BEDLAM but with an additional 300K iterations (BEDLAM*). We also train our network conbining our dataset and BEDLAM



Figure 8. Crops from MammaSyn-S (Single person).



Figure 9. Crops from MammaSyn-I (Interactions).

Subset	# minutes	Datasets	% of images
MammaSyn-S	86	BEDLAM	9.9%
		MOYO	15.9%
MammaSyn-I	211	Harmony4D	17.7%
		Hi4D	8.0%
		Inter-X	14.6%
		InteractionsCouple (ours)	19.4%
		LatinDance (ours)	3.4%
MammaSyn-H	36	InterHand2.6M	10.0%
		SignAvatars	1.0%

Table 6. Composition of the MammaSyn dataset. Percentages are computed over all MammaSyn images.

Table 7. **Evaluation of datasets.** Mean 2D Euclidean distance error (in pixels) between ground truth and predicted landmarks; includes visible and invisible landmarks.

Dataset	RICH	Harmony4D	CHI3D	MammaEval-S	MammaEval-D	MOYO
BEDLAM	8.83	18.33	4.36	6.16	7.70	11.04
BEDLAM*	9.35	19.11	4.40	6.49	7.60	11.92
MammaSyn	8.68	19.09	4.74	6.53	7.71	7.05
MammaSyn +B	**8.09**	**17.34**	**4.09**	**5.62**	**6.66**	**6.95**

(ALL+B).

Tab. 7 shows that training longer with BEDLAM does not improve performance (in fact, it degrades slightly). However, when we use our datasets along with BEDLAM, the performance of the network improves, especially on challenging poses such as those in MOYO and the interaction sequences.

10. MammaEval Dataset

We captured our evaluation datasets (MammaEval-Singles, MammaEval-Dance, and MammaEval-Extra) using a Vicon marker-based motion capture system, synchronized with a multi-view RGB camera system (IO Industries (IOI)). The motion capture setup consists of 54 VICON Vantage V16 cameras. The IOI setup includes 16 Victorem and 17 Volucam cameras. Both systems are mounted on a truss structure with the dimensions of 9m by 7m with a height of 5m. Within that space, the IOI cameras are positioned to cover a 3m×3m×3m space with the subject's full body visible in all views (see Fig. 10). Custom 12K Lux lighting by NORKA provides enough light to avoid shadows and motion blur at high capture frame rates. The lights are not constantly powered but are triggered to only turn on together with the IOI system when the shutter is open during each capture frame. This protects the subjects' vision from being permanently exposed to high-intensity light and prevents overheating while providing sufficient lighting for the captured images. All systems are synchronized using transistor-transistor Logic (TTL), precision time protocol (PTP), and linear timecode (LTC) to ensure accurate calibration and a frame-to-frame correspondence of captured data. A key factor in creating high-accuracy data across both camera systems is our custom calibration procedure. The VICON calibration wand is utilized to capture a calibration sequence that is processed to determine camera parameters in a joint space with a common origin. All sequences are captured at 30 fps for the IOI, with the framerate for VICON varying from 60 fps for MammaEval-Dance and MammaEval-Extra to 120 fps for MammaEval-Single.

For our evaluation comparing Vicon with MAMMA, we need to know ground truth marker locations on the surface of the SMPL-X body. To that end, we place markers on a subject and then capture their 3D body shape using a 3dMD body scanner. We fit the SMPL-X mesh to this scan to get the true body shape in SMPL-X topology. We then capture the subject moving in the motion capture system and run MoSh++ with the ground 3D shape template. This lets us establish the marker locations on the 3D surface of the body. We use these known locations when evaluating held-out marker error in the comparison section in the main document. Note that we do not use the ground truth body shape when in MAMMA or when running MoSh++ for the comparison with MAMMA. In both cases, the methods estimate the body shape directly from the images or markers, respectively.

11. Comparison with Harmony4D

The closest work to ours is Harmony4D. However, we cannot directly compare with that method since the test set was generated using their own method. Instead, we evaluate the silhouette reprojections against SAM2 [51] mask predictions. For this comparison, we reproject the Harmony4D meshes into each view and use the resulting silhouettes as masks for our network. However, as illustrated in the main paper and Fig. 11, the provided ground-truth meshes are often misaligned or do not perfectly conform to the subject's body, leading to noisy and imperfect input for our network. We achieve greater mean IoU (mIoU) regardless, 74.28% vs 69.80%. Fig. 11 illustrates that our method overlaps better with the silhouette of the people. We also highlight that Harmony4D uses an off-the-shelf method to obtain the body shapes, meanwhile our network does not have any special initialization or body shape and pose prior. Note that although SAM2 is generally good in extracting masks of the desired person, it sometimes misses parts of the body during video segmentation. This limitation is the main reason why the average IoU is around 70% rather than higher.

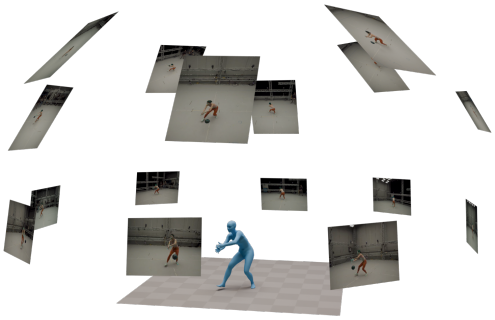
12. Multiview Matching Evaluation

In the main paper we show that our Multiview algorithm is able to match people across views with a success rate of 100% in the evaluation datasets. The datasets Harmony4D and MammaEval-D have around 20 cameras, whereas CHI3D has only 4.

Here, we evaluate the robustness of our algorithm by reducing the number of cameras from 16 down to 2 in powers of two (16, 8, 4, 2). We used MammaEval-D dataset. After the evaluation we observe that even with 2 cameras our method is capable to match correctly the people in the scene 100% of the times. This shows that our dense landmarks plus temporal information (from



(a) MammaEval-Dance



(b) MammaEval-Singles



(c) Sample views from the IOI cameras.

Figure 10. Our MammaEval capture setup.

Table 8. Full Benchmark 3D fitting errors (mm). We evaluate the error for the full body, only for the body, and only for the hands.

Model	RICH		Harmony4D		CHI3D		MammaEval-S		MammaEval-D		MOYO	
	MPJPE	PVE	MPJPE	PVE	MPJPE	PVE	MPJPE	PVE	MPJPE	PVE	MPJPE	PVE
SMPLfiy	96.18	71.42	-	-	67.68	51.79	47.15	35.42	53.92	43.08	62.15	44.68
LookMa*	39.52	30.29	59.37	45.6	46.47	39.36	25.97	23.94	27.98	24.89	60.15	53.82
CameraHMR	25.61	21.36	58.59	42.0	40.8	34.61	15.25	18.43	20.41	21.06	33.75	33.74
MAMMA	22.20	19.76	**45.26**	**34.02**	38.01	32.84	12.96	17.18	**17.71**	19.80	22.95	25.48
MAMMA-C	**22.20**	**19.76**	45.35	34.05	**37.96**	**32.82**	**12.96**	**17.18**	17.73	**19.78**	**22.95**	**25.48**
Only body												
SMPLfiy	71.34	63.99	-	-	40.05	44.21	31.4	30.81	37.69	38.92	47.81	39.66
LookMa*	37.78	28.89	49.09	42.69	33.48	35.18	25.34	22.7	23.81	23.41	62.02	53.57
CameraHMR	28.18	21.28	46.74	38.67	30.58	30.83	18.27	18.5	19.17	20.43	33.0	33.37
MAMMA	27.32	20.44	**40.84**	**32.58**	28.36	29.25	17.61	17.88	18.86	19.95	24.72	25.89
MAMMA-C	**27.32**	**20.44**	40.87	32.6	**28.35**	**29.25**	**17.61**	**17.88**	**18.82**	**19.93**	**24.72**	**25.89**
Only hands												
SMPLfiy	112.73	93.83	-	-	86.1	74.64	57.65	49.29	64.73	55.62	71.71	59.80
LookMa*	40.67	34.5	66.23	54.37	55.14	51.95	26.38	27.68	30.76	29.34	58.9	54.56
CameraHMR	23.91	21.61	66.5	52.05	47.62	45.98	13.24	18.21	21.24	22.97	34.24	34.87
MAMMA	18.79	17.72	**48.2**	**38.33**	44.44	43.65	9.85	15.06	**16.94**	**19.33**	21.77	24.26
MAMMA-C	**18.79**	**17.72**	48.34	38.41	**44.36**	**43.6**	**9.85**	**15.06**	17.01	19.34	**21.77**	**24.26**

SAM2) provide rich and accurate geometric information that is useful for matching the identity of a person



Figure 11. Mesh reprojection comparision between Harmony4D (red) and ours (blue).

across views. Note that for the single-frame case, our method can fail, as wrong mask predictions or heavy occlusions can confuse the network and lead to incorrect landmark predictions. This is particularly true when the limbs are heavily occluded. In those cases, the network has to "guess" the location of those parts, which can make them not view-consistent.

13. Perceptual Study

To further evaluate the perceptual quality of our reconstructions, we conducted an Amazon Mechanical Turk study in which 33 participants rated the realism of rendered motion from GT and MAMMA-C results using a five-point Likert scale. A one-sided Wilcoxon signed-rank test revealed that our reconstructions were perceived as significantly more realistic than the ground truth on CHI3D ($p = 2 \times 10^{-6}$, $\Delta = 1.36$) and Harmony4D ($p = 2.4 \times 10^{-4}$, $\Delta = 0.48$). For MOYO, the reconstruction also slightly outperformed the GT ($p \approx 0.0010$, $\Delta = 0.15$), although the effect size is smaller given that both versions are already highly realistic (means $\approx$ 4.2–4.4). In contrast, no significant advantage was observed for RICH, MammaEval-D, or MammaEval-S, where GT and MAMMA-C results received similarly high ratings (all $p > 0.3$, $|\Delta| \leq 0.11$).

14. Visualizing the Contact Probabilities

During optimization, we use contact probabilities averaged over all views. Figure 12 shows a visual example of these averaged values.

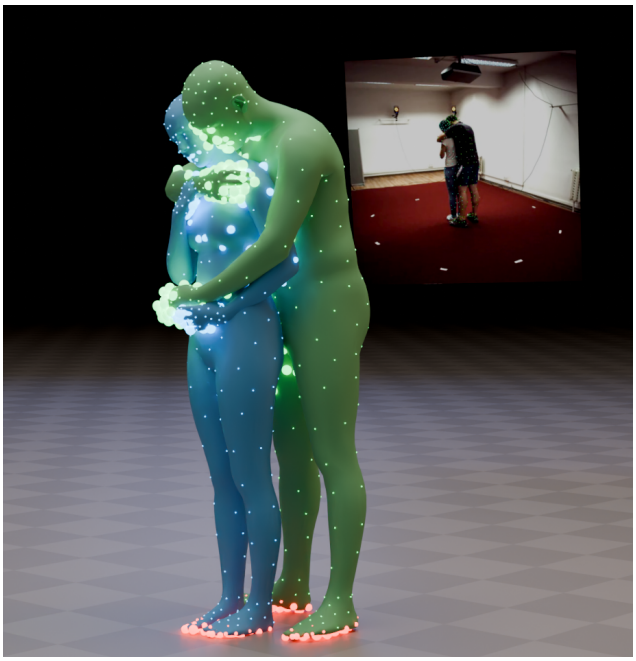


Figure 12. Contact probabilities for the 512 landmarks, averaged over per-view predictions. Brighter and larger points indicate higher contact probability.

15. Capture Protocol

Participants in the MammaEval-Single and MammaEval-Extra were selected to have a variety of body shapes and ethnicities, and gender balance. The MammaEval-Dance subjects were chosen based on their proficiency as dancers. All subjects were informed about the details of the capture protocol and the use of their data in advance and gave written informed consent.

The capture procedure and data processing steps have been approved by an ethics committee. All data collection, storage, and processing activities are compliant with privacy regulations. Participants wore

Table 9. IoU comparison between our method and Harmony4D on the Harmony4D test set.

Sequence	Harmony4D	Ours w/ mask
002_hugging	75.72	**77.96**
025_grappling2	66.26	**71.94**
028_grappling2	72.26	**78.34**
030_grappling2	70.26	**77.18**
031_grappling2	74.35	**80.44**
032_grappling2	54.79	**60.29**
033_grappling2	68.47	**74.95**
034_grappling2	64.16	**71.48**
035_grappling2	69.42	**76.16**
036_grappling2	67.82	**74.00**
037_grappling2	65.71	**72.25**
038_grappling2	71.93	**78.71**
039_grappling2	66.44	**72.78**
040_grappling2	73.41	**78.62**
041_grappling2	71.59	**77.72**
042_grappling2	71.62	**77.47**
043_grappling2	71.94	**78.61**
009_sword2	68.25	**69.19**
010_sword2	65.16	**66.03**
001_sword3	**70.70**	70.36
002_sword3	71.38	**72.52**
003_sword3	71.57	**72.52**
004_sword3	71.39	**72.02**
005_sword3	71.84	**72.30**
006_sword3	70.71	**71.41**
007_ballroom2	67.72	**72.11**
008_ballroom2	69.51	**73.29**
009_ballroom2	69.52	**72.84**
010_ballroom2	69.41	**73.11**
016_mma4	65.36	**76.18**
001_mma5	70.69	**74.86**
002_mma5	72.19	**77.27**
003_mma5	71.19	**75.85**
004_mma5	71.86	**76.57**
005_mma5	71.79	**76.77**
009_mma5	71.40	**75.69**
011_mma5	71.40	**76.07**
013_mma5	72.11	**76.09**
016_mma5	71.02	**74.95**
mIoU	69.80	**74.28**

tight-fitted clothing in distinctive colors to enable good quality body scanner data and the processing of the multi-view captures. In the body scanner, each subject was recorded performing several sequences and still poses.

For all captures, the *FrontWaist10Fingers* marker set from Vicon was used with a total of 73 markers distributed on the body and hands of the participant. In line with the Vicon data capture process, subjects were calibrated by performing a range-of-motion sequence.

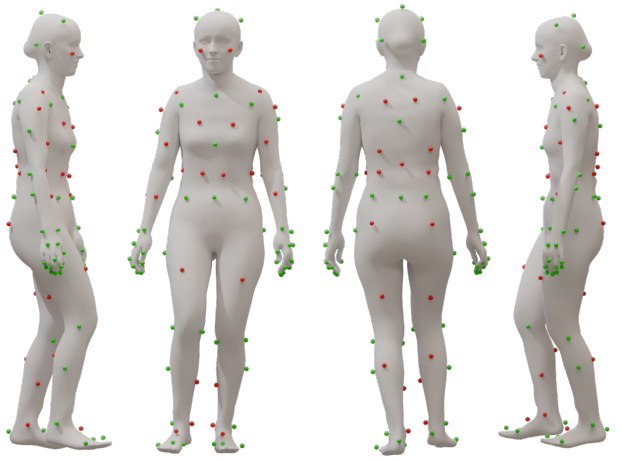


Figure 13. To evaluate MAMMA predictions on the Vicon ground truth, we use the additional held-out 37-marker layout (shown in red) alongside the default 73 FrontWaist10Fingers marker set from Vicon (green).

For the MammaEval-Extra, the additional 37 markers were attached after subject calibration to enable both the standard Vicon post-processing and the tracking of added markers that were labelled manually. The positions of the 37 extra markers were selected to be distinct from the standard marker positions, to avoid occlusion, and to include soft-tissue areas that are excluded from standard motion capture templates (Fig. 13). Reference images of the subjects after marker placement were captured to allow for an accurate replacement of fallen markers and support further processing steps.

For the Vicon Marker-based comparison experiment, we excluded one subject due to calibration errors of the Vicon-IOI system and discarded the first 5 frames to account for trigger-light delay.

15.1. Computing marker positions

We use MoSh++ to create the GT SMPL-X fits of our MammaEval datasets and to compute the marker positions of the methods used in our Vicon Markers comparison experiment. MoSh++ requires an initialization step that assigns each motion-capture marker to a vertex on the SMPL-X body model. In its first optimization stage, the marker positions are allowed to vary slightly to better align with the input data. While standard marker layouts often come with predefined marker-to-vertex mappings, real-world usage introduces significant variability. Due to human error and the diverse range of body shapes and clothing conditions, the actual marker placements for the same layout can differ substantially across captures, sometimes by several tens of centimeters. MoSh++ is

highly sensitive to initialization quality, and an accurate marker-to-vertex mapping is critical for achieving optimal results. To ensure reliable initialization, we capture 3D scans of subjects wearing markers, fit the SMPL-X model to the scans, and manually select the most appropriate vertex for each marker.

During the first optimization stage, MoSh++ refines marker positions using a local coordinate system defined as follows: (a) the closest SMPL-X vertex to the marker serves as the origin; (b) the coordinate axes are constructed from two edges of the triangle incident to this vertex, along with their cross product to form an orthonormal basis. The final marker position is stored as a combination of the vertex index and three displacement values, each corresponding to an offset along one of the local coordinate axes. The origin and local coordinate axes obtained during this optimization stage enable us to regress marker positions for any SMPL-X body shape.

15.2. MammaEval-Extra Protocol

Table 10. Vicon Markers comparison experiment: Mean per-marker distance (mm) of MoSh and MAMMA on the Vicon held-out 37 markers of our MammaEval-Extra dataset.

Subject	Action	MoSh	MAMMA
00202	solo_dancing	24.368	25.464
00202	calib_routine	20.081	21.259
00202	walking	23.384	25.841
00202	warmup	21.888	24.108
00219	solo_dancing	23.967	22.889
00219	calib_routine	23.338	22.670
00219	walking	23.507	24.503
00219	warmup	23.898	21.864
00236	solo_dancing	19.864	21.876
00236	calib_routine	17.832	19.050
00236	walking	18.791	21.002
00236	warmup	18.514	19.248
		21.619	22.481

The process described in Section 15.1, allows us to compute the 37 additional held-out Vicon marker positions from SMPL-X vertices (Fig. 13). This enables us to quantitatively compare markerless to marker-based methods. As a baseline, we use MoSh++, which takes as input only the 73 markers (*FrontWaist10Fingers* marker set). In Tab. 10, we report the mean per-marker distance error of MoSh++ (baseline) and MAMMA, per-sequence, with both methods evaluated without using a GT body shape (i.e. without using the 3D scan).

Similarly, we additionally tested our protocol on the MammaEval-Dance sequences, but this time we held-out 9 markers (from the 73 FrontWaist10Fingers marker set) distributed across the body and ran the same protocol. As a baseline, we use MoSh++, which takes as input the remaining 64 markers. The average error of our MAMMA is 27.590mm and 26.150mm for MoSh++ (baseline) with a difference of 1.44mm.

16. Usability and Cost

In addition to accuracy, the cost of setup and processing is important. We compare the time cost of using our automatic method vs. the marker-based pipeline which requires manually cleaning and labelling data. To estimate the cost of marker labelling, we recorded the performance of 3 marker post-processing technicians who had prior experience cleaning mocap frames, with experience ranging from 118,972 to 149,454 frames, with an average of 134,500 frames. Each technician cleaned 3 mocap sequences of 2 people dancing (West Coast Swing), while timing and documenting their workflow. Using the Vicon Shogun Post tool, the data was inspected frame-by-frame and all missing markers (gaps) and marker swaps were manually corrected. They cleaned 9 sequences with a total of 29,244 frames and a capture duration of 24 minutes and 31 seconds. Cleaning required 46 hours and 51 minutes. The most time-intensive step of the pipeline was fixing the swapped markers located on the fingers, with an average time of 4.95 hours per sequence. While marker swaps on body locations are typically less problematic than those on the fingers, the time to fix them took an average of 15 minutes. Gap filling was performed in Shogun Post and, while manual, the time is minimal compared with the labeling.

After cleaning we ran MoSh++ on the result, which took approximately 25 hours of computation. Ignoring the time to put on and take off markers, the time from capture to SMPL-X fits was approximately 72 hours. In contrast, running MAMMA takes approximately 26 hours on a single GTX-4090. SAM2 runs at 2 fps, MammaNet at 12 fps and the optimization on average takes 65 secs. for a sequence of 100 frames.

17. MAMMA Dance Release

In addition to MammaSyn, MammaEval-Singles, MammaEval-Dance and MammaEval-Extra, we release a new dataset captured using our MAMMA pipeline. The dataset includes Bachata, West Coast Swing, Breakdance and Ballroom, performed by 9 subjects for approximately 1 hour in total. We capture at 30 fps, from 32 synchronized cameras, using the multi-view setup shown in Fig. 14.

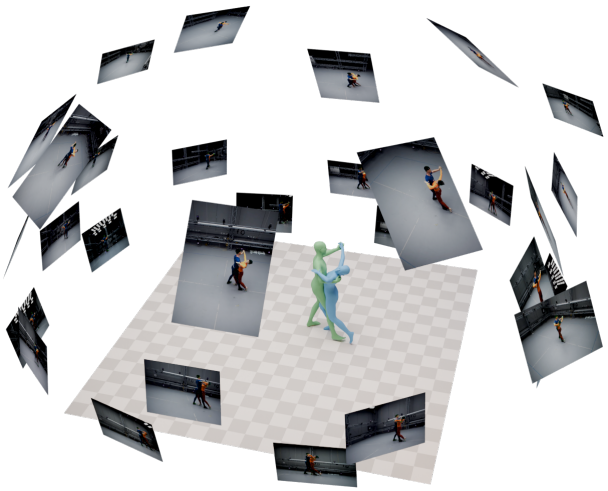


Figure 14. Multi-view setup of the MAMMA Dance sequences captured solely with MAMMA.

18. Limitations and Future Work

Our system recovers motion sequences of subjects interacting with each other from multiview cameras. Despite its competitive performance, MAMMA still has limitations. The predicted landmark uncertainties and visibility probabilities allow the optimizer to reduce the influence of low-confidence or occluded landmarks. As a result, the optimizer may ignore landmarks that are heavily occluded and uncertain across views, leading to flickering artifacts or over-smoothing, especially when smoothing regularization is heavily weighted.

The contact probability prediction of our network is conservative, the highest probability is around 60%. This is due to the single view ambiguity. If two people are interacting, and one occludes the other person, there is a chance that they are really close but not in contact. One fix to make the network more confident is weighting the loss for parts that are in contact, however, this comes at the cost of failing in the previous example.

Foot contact prediction is generally good. However, our network often predicts the foot landmarks based on the height of the shoes. If we then penalize floor contact errors too strongly during optimization, the optimizer may incorrectly pull the person downward to satisfy the contact constraint. In other situations, when the body is already touching the floor, the dense landmarks alone provide sufficient cues, so the additional contact term has little effect. Therefore, we recommend using the floor contact signal primarily for single-view cases or as a post-processing step to correct foot motion near the floor.

The accuracy of hand-motion recovery can still be improved. It is worth noting that most marker-based captures ignore the hands completely since they are too costly to capture and clean. We believe that image-based methods can be further improved with better training data.

We plan to extend the network to predict landmarks across multiple views jointly to improve inter-view consistency. Similarly, temporal modeling would enhance landmark stability over time. Another potential direction is the integration of diffusion-based human motion priors to refine predictions.

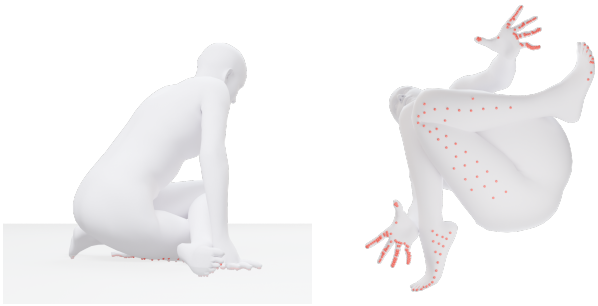


Figure 15. Computed floor contact from geometric cues, used in MammaSyn. We show the pose from a side-view and an under-view. SMPL-X vertices that are in contact are shown in red.

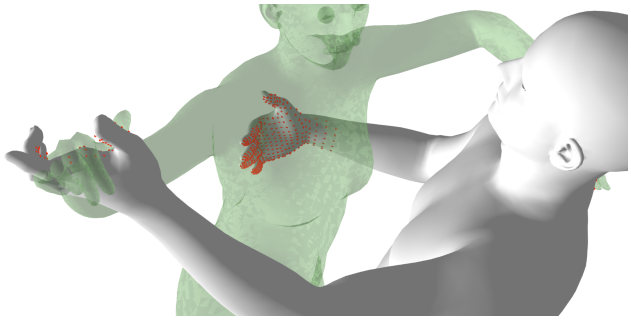


Figure 16. Computed contact between two subjects from geometric cues, used in MammaSyn. SMPL-X vertices of gray subject that are in contact with the green transparent subject are shown in red.

References

[1] Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 3D pictorial structures for multiple human pose estimation. In *CVPR*, pages 1669–1676, Los Alamitos, CA, USA, 2014. IEEE Computer Society. 3

[2] Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: Dataset and method for tracking human object interactions. In *CVPR*, pages 15935–15946, Los Alamitos, CA, USA, 2022. IEEE Computer Society. 3

[3] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *CVPR*, pages 8726–8737, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 2, 3

[4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE TPAMI*, 43(1):172–186, 2021. 3

[5] Captury. Captury. https://captury.com/, 2025. 2, 3

[6] Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3D pose estimation from multiple views. In *CVPR*, pages 7792–7801, Los Alamitos, CA, USA, 2019. IEEE Computer Society. 5, 7, 8

[7] Junting Dong, Qi Fang, Wen Jiang, Yurou Yang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3D pose estimation and tracking from multiple views. *IEEE TPAMI*, 44(10):6981–6992, 2021. 8

[8] EasyMocap. EasyMoCap – make human motion capture asier. Github, 2021. 7

[9] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *CVPR*, pages 12943–12954, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 3

[10] 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 TPAMI*, 45(6):7157–7173, 2023. 3

[11] Qi Fang, Yinghui Fan, Yanjun Li, Junting Dong, Dingwei Wu, Weidong Zhang, and Kang Chen. Capturing closely interacted two-person motions with reaction priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 655–665, 2024. 3

[12] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *CVPR*, pages 7214–7223, Los Alamitos, CA, USA, 2020. IEEE Computer Society. 2, 3, 6

[13] Stuart Geman and Donald E. McClure. Statistical methods for tomographic image reconstruction. In *Proceedings of the 46th Session of the International Statistical Institute, Bulletin of the ISI*, 1987. 5

[14] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, pages 14783–14794, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 3

[15] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *ICCV*, pages 2282–2292, Los Alamitos, CA, USA, 2019. IEEE Computer Society. 3

[16] Charlie Hewitt, Fatemeh Saleh, Sadegh Aliakbarian, Lohit Petikam, Shideh Rezaeifar, Louis Florentin, Zafiirah Hosenie, Thomas J. Cashman, Julien Valentin, Darren Cosker, and Tadas Baltrusaitis. Look Ma, no markers: Holistic performance capture without the hassle. *ACM TOG*, 43(6), 2024. 2, 3, 4, 5, 6, 9

[17] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *CVPR*, pages 13274–13285, Los Alamitos, CA, USA, 2022. IEEE Computer Society. 3, 6

[18] IOI Industries. Ioi. https://www.ioindustries.com/, 2025. 4

[19] Catalin Ionescu, Fuxin Li, and Cristian Sminchisescu. Latent structured models for human pose estimation. In *ICCV*, pages 2220–2227, Los Alamitos, CA, USA, 2011. IEEE Computer Society. 3

[20] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE TPAMI*, 36(7):1325–1339, 2014. 3

[21] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *ICCV*, pages 7718–7727, Los Alamitos, CA, USA, 2019. IEEE Computer Society. 3

[22] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, 2015. 3

[23] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE TPAMI*, 41(01):190–204, 2019. 3

[24] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131, Los Alamitos, CA, USA, 2018. IEEE Computer Society. 3

[25] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate,

and Otmar Hilliges. EMDB: The electromagnetic database of global 3D human pose and shape in the wild. In *ICCV*, pages 14632–14643, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 3

[26] Rawal Khirodkar, Jyun-Ting Song, Jinkun Cao, Zhengyi Luo, and Kris Kitani. Harmony4D: A video dataset for in-the-wild close human interactions. *NeurIPS*, 37:107270–107285, 2024. 2, 3, 4, 6

[27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *ICCV*, pages 4015–4026, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 4

[28] Muhammed Kocabas, Ye Yuan, Pavlo Molchanov, Yunrong Guo, Michael J. Black, Otmar Hilliges, Jan Kautz, and Umar Iqbal. Pace: Human and camera motion estimation from in-the-wild videos. pages 397–408, Los Alamitos, CA, USA, 2024. IEEE Computer Society. 3

[29] Ziwei Liao, Jialiang Zhu, Chunyu Wang, Han Hu, and Steven L. Waslander. Multiple view geometry transformers for 3D human pose estimation. In *CVPR*, pages 708–717, Los Alamitos, CA, USA, 2024. IEEE Computer Society. 3

[30] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 4

[31] Dong C. Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1):503–528, 1989. 5

[32] Matthew Loper, Naureen Mahmood, and Michael J. Black. MoSh: Motion and shape capture from sparse markers. *ACM TOG*, 33(6), 2014. 3

[33] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM TOG*, 34(6): 248:1–248:16, 2015. 3

[34] Feichi Lu, Zijian Dong, Jie Song, and Otmar Hilliges. AvatarPose: Avatar-guided 3d pose estimation of close human interaction from sparse multi-view videos. In *ECCV*, pages 215–233, Cham, 2025. Springer Nature Switzerland. 7, 8

[35] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, pages 5442–5451, Los Alamitos, CA, USA, 2019. IEEE Computer Society. 1, 2, 4, 7, 8

[36] Claire McLean, Makenzie Meendering, Tristan Swartz, Orri Gabbay, Alexandra Olsen, Rachel Jacobs, Nicholas Rosen, Philippe de Bree, Tony Garcia, Gadsden Merrill, et al. Embody 3d: A large-scale multimodal motion and behavior dataset. *arXiv preprint arXiv:2510.16258*, 2025. 3

[37] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. pages 506–516, Los Alamitos, CA, USA, 2017. IEEE Computer Society. 3

[38] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3D pose estimation from monocular RGB. pages 120–130, Los Alamitos, CA, USA, 2018. IEEE Computer Society. 3

[39] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image. In *ECCV*, pages 548–564, Cham, 2020. Springer International Publishing. 4

[40] Motion Analysis. Motion analysis. https://www.motionanalysis.com/, 2025. 1, 2

[41] Move AI Ltd. Move AI ltd. https://www.move.ai/, 2025. 2, 3

[42] Moverse. Moverse. https://moverse.ai/, 2025. 2, 3

[43] Lea Müller, Ahmed A. A. Osman, Siyu Tang, Chun-Hao P. Huang, and Michael J. Black. On self-contact and human pose. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recogßnition (CVPR)*, 2021. 5

[44] Natural Point. OptiTrack. https://www.optitrack.com/, 2025. 1, 2

[45] Priyanka Patel and Michael J. Black. CameraHMR: Aligning people with perspective. Los Alamitos, CA, USA, 2025. IEEE Computer Society. 2, 3, 4, 5, 6, 7, 9

[46] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *CVPR*, pages 13468–13478, Los Alamitos, CA, USA, 2021. IEEE Computer Society. 3

[47] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, pages 10975–10985, Los Alamitos, CA, USA, 2019. IEEE Computer Society. 2, 3, 7

[48] PhaseSpace. PhaseSpace. https://www.phasespace.com/, 2025. 2

[49] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. DeepCut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*, pages 4929–4937, Los Alamitos, CA, USA, 2016. IEEE Computer Society. 3

[50] Qualisys. Qualisys. https://www.qualisys.com/, 2025. 1, 2

[51] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM

2: Segment anything in images and videos. In *ICLR*, Red Hook, NY, USA, 2025. Curran Associates, Inc. 2, 5, 10

[52] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM TOG*, 36(6):245:1–245:17, 2017. 4

[53] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 7

[54] Leonid Sigal, Alexandru Balan, and Michael J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87(1):4–27, 2010. 3

[55] Vinkle Srivastav, Keqi Chen, and Nicolas Padoy. SelfPose3d: Self-supervised multi-person multi-view 3D pose estimation. In *CVPR*, pages 2502–2512, Los Alamitos, CA, USA, 2024. IEEE Computer Society. 3

[56] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *ECCV*, pages 581–600, Cham, 2020. Springer International Publishing. 3

[57] Joachim Tesch, Giorgio Becherini, Prerana Achar, Anastasios Yiannakidis, Muhammed Kocabas, Priyanka Patel, and Michael J. Black. BEDLAM2.0: Synthetic humans and cameras in motion. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. 3

[58] Theia Markerless. Theia markerless. https://www.theiamarkerless.com/, 2025. 2, 3

[59] Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Taheri Omid, Michael J. Black, and Dimitrios Tzionas. 3D human pose estimation via intuitive physics. In *CVPR*, pages 4713–4725, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 3, 6

[60] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total capture: 3D human pose estimation fusing video and inertial sensors, 2017. 3

[61] Scott D. Uhlrich, Antoine Falisse, Łukasz Kidziński, Julie Muccini, Michael Ko, Akshay S. Chaudhari, Jennifer L. Hicks, and Scott L. Delp. OpenCap: Human movement dynamics from smartphone videos. *PLOS Computational Biology*, 19(10):1–26, 2023. 2

[62] Vicon Motion Systems. Vicon. https://www.vicon.com/, 2025. 1, 2

[63] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *ECCV*, pages 614–631, Cham, 2018. Springer International Publishing. 3

[64] Tao Wang, Jianfeng Zhang, Yujun Cai, Shuicheng Yan, and Jiashi Feng. Direct multi-view multi-person 3D pose estimation. In *NeurIPS*, pages 13153–13164, Red Hook, NY, USA, 2021. Curran Associates, Inc. 3, 8

[65] Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J Black, and Muhammed Kocabas. Prompthmr: Promptable human mesh recovery. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 7

[66] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, pages 4724–4732, USA, 2016. IEEE Computer Society. 3

[67] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljević, Daniel Wilde, Stephan Garbin, Toby Sharp, Ivan Stojiljković, et al. 3D face reconstruction with dense landmarks. In *ECCV*, pages 160–177, Cham, 2022. Springer, Springer Nature Switzerland. 4

[68] Size Wu, Sheng Jin, Wentao Liu, Lei Bai, Chen Qian, Dong Liu, and Wanli Ouyang. Graph-based 3D multi-person pose estimation using multi-view images. In *ICCV*, pages 11148–11157, Los Alamitos, CA, USA, 2021. IEEE Computer Society. 8

[69] Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, et al. Inter-X: Towards versatile human-human interaction analysis. In *CVPR*, pages 22260–22271, Los Alamitos, CA, USA, 2024. IEEE Computer Society. 4

[70] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*, 2022. 6

[71] Hang Ye, Wentao Zhu, Chunyu Wang, Rujie Wu, and Yizhou Wang. Faster VoxelPose: Real-time 3D human pose estimation by orthographic projection. In *ECCV*, pages 142–159, Cham, 2022. Springer, Springer Nature Switzerland. 8

[72] Wanqi Yin, Zhongang Cai, Ruisi Wang, Fanzhou Wang, Chen Wei, Haiyi Mei, Weiye Xiao, Zhitao Yang, Qingping Sun, Atsushi Yamashita, et al. WHAC: World-grounded humans and cameras. In *ECCV*, pages 20–37, Cham, 2024. Springer, Springer Nature Switzerland. 3

[73] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Zarate, Jie Song, and Otmar Hilliges. Hi4D: 4D instance segmentation of close human interaction. In *CVPR*, pages 17016–17027, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 2, 3, 4

[74] Zhengdi Yu, Shaoli Huang, Yongkang Cheng, and Tolga Birdal. SignAvatars: A large-scale 3D sign language holistic motion dataset and benchmark. In *ECCV*, pages 1–19, Cham, 2025. Springer Nature Switzerland. 4

[75] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *ICLR*, Red Hook, NY, USA, 2023. Curran Associates, Inc. 9

[76] Yuxiang Zhang, Liang An, Tao Yu, Xiu Li, Kun Li, and Yebin Liu. 4D association graph for realtime multi-person motion capture using multiple video cameras. In *CVPR*, pages 1324–1333, Los Alamitos, CA, USA, 2020. IEEE Computer Society. 8