M3GYM: A Large-Scale Multimodal Multi-view Multi-person Pose Dataset for Fitness Activity Understanding in Real-world Settings

Qingzheng Xu[1,3], Ru Cao[2], Xin Shen[1], Heming Du[1,3], Sen Wang[1], Xin Yu[1,3*]
[1]The University of Queensland [2]City University of Macau [3]Follow Me AI Pty LTD
{qingzheng.xu, xin.yu}@uq.edu.au

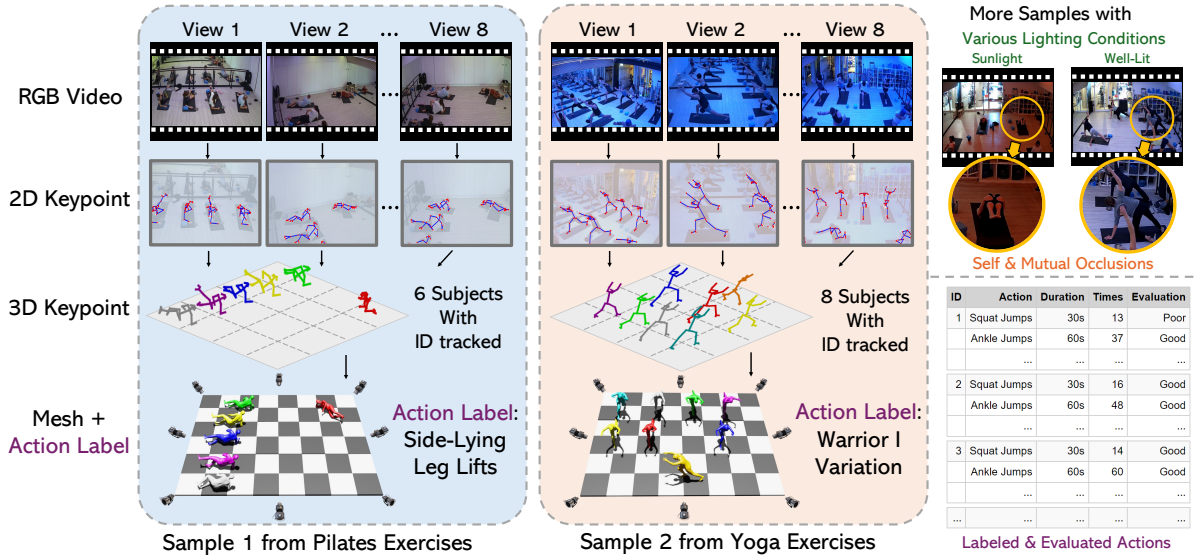

Figure 1. **Illustration of M3GYM.** For each sample, we provide multi-view videos with fine-grained annotations, including 2D keypoints, 3D keypoints, subject IDs, mesh, and action labels. M3GYM also includes expert assessments on action completeness for each subject, covers diverse lighting conditions and features scenes with complex self & mutual occlusions.

Abstract

Human pose estimation is a critical task in computer vision for applications in sports analysis, healthcare monitoring, and human-computer interaction. However, existing human pose datasets are collected either from custom-configured laboratories with complex devices or they only include data on single individuals, and both types typically capture daily activities. In this paper, we introduce the M3GYM dataset, a large-scale multimodal, multi-view, and multi-person pose dataset collected from a real gym to address the limitations of existing datasets. Specifically, we collect videos for 82 sessions from the gym, each session lasting between 40 to 60 minutes. These videos are gathered by 8 cameras, including over 50 subjects and 47 million frames. These sessions include 51 Normal fitness exercise sessions as well as 17 Pilates and 14 Yoga sessions. The exercises cover a wide range of poses and typical fitness activities, particularly in Yoga and Pilates, featuring poses with stretches, bends, and twists, e.g., *humble warrior, fire hydrants and knee hover side twists. Each session involves multiple subjects, leading to significant self-occlusion and mutual occlusion in single views. Moreover, the gym has two symmetric floor mirrors, a feature not seen in previous datasets, and seven lighting conditions. We provide frame-level multimodal annotations, including 2D&3D keypoints, subject IDs, and meshes. Additionally, M3GYM uniquely offers labels for over 500 actions along with corresponding assessments from sports experts. We benchmark a variety of state-of-the-art methods for several tasks,* i.e., *2D human pose estimation, single-view and multi-view 3D human pose estimation, and human mesh recovery. To simulate real-world applications, we also conduct cross-domain experiments across Normal, Yoga, and Pilates sessions. The results show that M3GYM significantly improves model generalization in complex real-world settings. The project is available here.*

*Corresponding author.

1. Introduction

Human pose estimation plays a pivotal role in computer vision for capturing and analyzing human movements through visual data. It supports a broad range of applications such as action analysis [3, 51, 52], augmented & virtual reality [58, 82], healthcare [55, 56], human body animation [31, 62, 72], and human-computer interaction [9, 23]. These applications highlight the critical importance of developing pose estimation algorithms and compiling datasets under various conditions [32, 80].

However, existing human pose datasets focus on single-person actions [6, 25, 34, 44, 47, 54, 60, 67, 73], limiting the applicability to real-world scenarios. Meanwhile, although many multi-person datasets are collected, they also have various limitations, such as lacking complex occlusions [28, 57], containing few subjects per scene [19, 24, 77], being confined to laboratory environments [28], or relying on unrealistic synthetic data [4, 48, 69]. Therefore, a large-scale, diverse, multi-person pose dataset with complex poses reflecting real-world environments is urgently needed.

In this paper, we introduce M3GYM, a large-scale, multimodal, multi-view, and multi-person pose dataset collected in authentic gym environments. We collect multimodal and multi-view videos that captures the dynamic and complex interactions among multiple people engaged in a variety of fitness activities. Our goal is to address the limitations of existing datasets and to provide a dataset that facilitates the development of human pose estimation algorithms in realistic settings, specifically within gym scenarios with multiple subjects. As demonstrated in Figure 1, we mount 8 cameras in a circular arrangement on the ceiling of the gym. Then, we collect videos across 82 sessions, each lasting between 40 to 60 minutes, capturing over 50 subjects and totaling 47 million frames.

Unlike most existing datasets that focus on daily activities, M3GYM includes a variety of fitness exercises. We record 51 sessions featuring normal fitness exercises, as well as 17 Pilates and 14 Yoga sessions. Human poses in Pilates and Yoga exercises, such as humble warrior, fire hydrants, and knee hover side twists, include a wide range of poses, particularly stretches, bends, and twists. The complex movements in Pilates and Yoga often lead to self-occlusion, where a part of the body is hidden from the single camera view. Meanwhile, the presence of multiple participants in each session causes substantial mutual occlusion, realistically reflecting the challenges faced in real-world pose estimation applications.

Furthermore, our dataset is unique as it incorporates elements commonly seen in gyms and everyday life, such as mirrors, which are absent in other human pose datasets. The gym has two symmetric floor mirrors, presenting significant challenges for existing 2D and 3D pose estimation methods due to reflections. Additionally, the M3GYM dataset includes seven different lighting conditions, such as well-lit, sunlight with backlighting, and various ambient lighting setups. These features make our dataset more diverse and contribute to the development of pose estimation algorithms that are effective in real-world scenarios.

To ensure annotation accuracy, we design a semi-automated pipeline for error-checking, including time-based frame alignment, camera calibration with pixel alignment results, multi-view 2D keypoint voting with state-of-the-art detectors, triangulation for 3D keypoints, physics-based auto-calibration, inter-frame smoothing, and manual inspection of 3D keypoints. Using the adjusted 3D keypoints, we generate meshes and incorporate notes from sports experts in each session to achieve ground truth annotations. Building on this pipeline, M3GYM provides fine-grained, per-subject multimodal annotations for each time segment, covering 2D & 3D keypoints, subject IDs, mesh, action labels, and expert assessments of action quality.

M3GYM supports a wide range of pose estimation tasks, spanning 2D human pose estimation, single-view and multi-view 3D human pose estimation, and human mesh recovery. We provide benchmarks using state-of-the-art and widely-used methods, along with cross-domain experiments across three session types in M3GYM. The results show that M3GYM challenges models and enhances their performance in complex real-world settings. The main contributions of this research are summarized as follows:

- **We introduce M3GYM, a large-scale, multimodal, multi-view, multi-person pose dataset** with fine-grained, time-span-based annotations for each subject's actions, capturing authentic gym activities.
- **We establish benchmarks for multiple pose estimation tasks on M3GYM**, demonstrating the challenging nature of M3GYM and its effectiveness in enhancing model performance for real-world human activity analysis.

2. Related Works

2.1. Human Pose Dataset

Human pose estimation relies on comprehensive 2D and 3D pose datasets for effective model training and evaluation. **2D human pose datasets** play a foundational role. Some, like COCO [39], CrowdPose [33], and MPII [1], focus on isolated images, providing essential benchmarks for pose detection. Others, such as PoseTrack [2], introduce temporal context across video sequences, enabling the study of motion and improving pose tracking over time.

Beyond 2D pose datasets, **3D human pose datasets** further support depth-based analysis, such as HumanEva [54], Human3.6M [25], MPI-INF-3DHP [47], FreeMan [60], Fit3D [20], AIST++ [34], HUMBI [73], AMASS [44], MM-Fi [67], HuMMan [6]. Although essential for 3D pose estimation, these datasets focus on **single-person** data.

Table 1. **Comparison of multi-view real-world human pose datasets.** "#Subj Range" represents the subject range in each scene, and "Act Type" refers to the primary types of actions included. Only HD camera data is included for CMU Panoptic [28]. MPI-INF-3DHP [47] and FreeMan [60] include outdoor scenes, primarily in well-lit and backlighting conditions.

Dataset	#Act	#Frame	#Camera	#Subj Range	#Light	Act Type	Envir	Modalities					
								Video	2D Kpt	3D Kpt	Mesh	Act Label	Act Assmt
HumanEva [54]	6	80K	7	1	-	Daily	Laboratory	✓	✓	✓	✗	✓	✗
Human3.6M [25]	15	3.6M	4	1	-	Daily	Laboratory	✓	✓	✓	✗	✓	✗
MPI-INF-3DHP [47]	8	1.3M	14	1	2	Daily	Real-Scene	✓	✓	✓	✗	✓	✗
CMU Panoptic [28]	5	154M	31	1-8	-	Daily	Laboratory	✓	✓	✓	✗	✓	✗
CHI3D [19]	8	315K	4	2	-	Daily	Laboratory	✓	✓	✓	✓	✓	✗
Fit3D [20]	47	1.96M	4	1	-	Fitness	Laboratory	✓	✓	✓	✓	✓	✗
AIST++ [34]	10	10.1M	9	1	-	Dance	Laboratory	✓	✓	✓	✓	✓	✗
HuMMan [6]	500	60M	11	1	-	Daily	Laboratory	✓	✓	✓	✓	✓	✗
MM-Fi [67]	27	320K	3	1	-	Daily	Laboratory	✓	✓	✓	✗	✓	✗
FreeMan [60]	123	11.3M	8	1	2	Daily	Real-Scene	✓	✓	✓	✓	✓	✗
M3GYM (Ours)	502	47M	8	1-10	7	Fitness	Real-Scene	✓	✓	✓	✓	✓	✓

However, **multi-person** 3D datasets introduce additional challenges. CMU Panoptic [28] captures multi-person activities but is limited by its controlled lab setting and simple occlusions. 3DPW [57] includes outdoor scenes but lacks dense occlusions, limiting its utility for crowded scenarios. Datasets like CHI3D [19], EgoBody [77], and RICH [24] focus on human-object interactions but have few subjects per scene, reducing their effectiveness in high-density settings. Synthetic datasets like AGORA [48], SynBody [69], and BEDLAM [4] simulate interactions, but their artificiality may limit real-world applicability.

In Table 1, we compare M3GYM with existing multi-view human pose datasets captured in realistic environments. Unlike other datasets, M3GYM focuses on fitness activities with diverse, multi-person scenes involving up to 10 subjects, resulting in significant mutual occlusions. Additionally, 43.7% of the data centers on Yoga and Pilates, introducing unique self-occlusions. Compared to another fitness dataset, Fit3D [20], M3GYM includes multi-person scenarios, more subjects & action categories, and expert assessments. Particularly, Fit3D does not cover movements provided in M3GYM's Yoga and Pilates sessions. M3GYM further adds complexity through its unique overhead camera views, two floor mirrors, and seven lighting conditions.

2.2. Human Pose Estimation

Rapid progress in human pose estimation [41] is inseparable from the advances of deep network models [13–16, 18, 22, 63, 64]. For **2D pose estimation**, OpenPose [8] enables real-time keypoint detection with a bottom-up approach. MediaPipe [43] offers a cross-platform framework for mobile devices. DEKR [21] proposes a bottom-up method that decouples detection and regression, refining localization precision. AlphaPose [17] combines a region proposal network with pose-guided refinement for multi-person accuracy. ViTPose [66] uses Vision Transformers for global context, achieving top results. YOLO-Pose [45] and YOLOv7-Pose [59] adapt YOLO for pose estimation, balancing speed and accuracy. ED-Pose [68] enhances detection in complex scenes with edge-aware networks. RTM-Pose [27] optimizes for mobile. DWPose [70] boosts efficiency with depth-wise separable convolution. Sapiens [30] provides a versatile model family for human-centric tasks, including 2D pose estimation.

Building on these 2D approaches, **3D pose estimation** methods use 2D keypoints for 3D reconstruction [26, 53, 75, 76]. **Multi-view** approaches, like MV-Pose [11, 12], leverage synchronized 2D detections for efficient multi-person 3D estimation, while PlaneSweepPose [36] applies depth regression through cross-view consistency. Faster VoxelPose [71] employs a fast voxel-based technique, re-projecting feature volumes for speed and accuracy. While **single-view** methods reconstruct 3D keypoints from single-view images, often relying on temporal or structural cues. SimpleBaseline [46] uses a basic network to lift 2D to 3D keypoints. Video-based methods, like VideoPose3D [49], PoseFormer [78, 79], and MHFormer [35] integrate temporal information, while MotionBERT [81] further captures complex motion dynamics, enhancing 3D pose accuracy.

2.3. Human Mesh Recovery

Human mesh recovery focuses on reconstructing 3D body meshes from visual data, capturing both shape and pose [50, 61]. HMR [29] pioneers this setting with an end-to-end framework. METRO [38] uses a transformer to enhance vertex regression while PyMAF-X [74] refines alignment with a pyramidal feedback loop. OSX [37] integrates body, face, and hand estimation into a unified model. SMPLer-X [7] achieves robust generalization across diverse datasets. SMPLer [65] applies decoupled attention in a Transformer for efficient 3D shape and pose estimation.

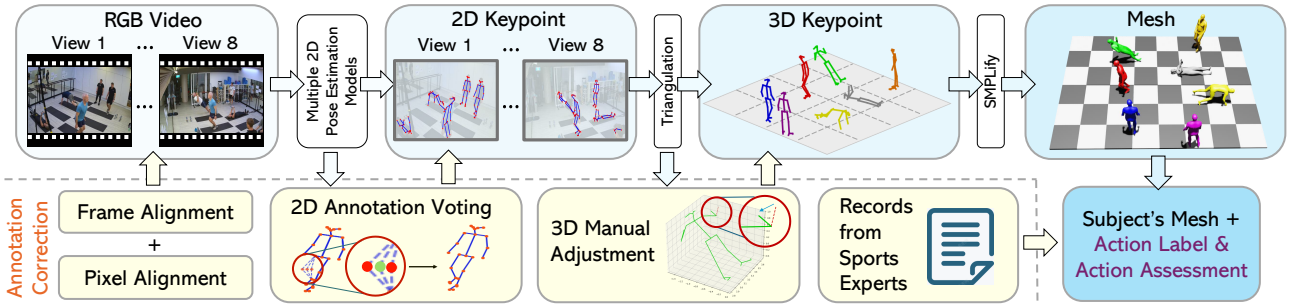

Figure 2. **Illustration of M3GYM Semi-automated Pipeline.** For each sample, we align videos for each view through frame and pixel alignment. Multiple 2D keypoint detection methods are applied across all 8 viewpoints for each frame, and median voting is used to obtain a basic 2D keypoint set. Triangulation then produces 3D keypoints, which are refined with physics-based auto-calibration and inter-frame smoothing before manual checking and adjustment, resulting in the final ground truth 3D keypoints. Using SMPLify, we generate meshes, incorporating sports expert notes from each session to obtain the final mesh, action label, and action assessment for each subject.

3. M3GYM

This section introduces the recording setup and workflow of M3GYM, including the overall semi-automated annotation pipeline and detailed statistics of this dataset.

3.1. Recording Setup

To curate M3GYM, we establish a partnership with a real gym. The gym offers new training sessions designed to capture the movements of participants, and thus, all participants are actual gym clients. Before joining, participants need to review the experiment details and sign a written consent form. With consent from everyone involved, including clients and sports experts, we place eight fisheye overhead cameras around the gym. Cameras are positioned at each corner and the midpoint of each side of the rectangular gym, recording 1920x1080 videos at 25 fps.

We obtain the parameters of each camera through a chessboard-based calibration applied in MV-Pose [11]. By recording the moving chessboard across all cameras, we capture a total of 240,000 frames, which we use to derive the intrinsic parameters for the eight cameras. Next, we fix the chessboard at the center of the gym to calculate the extrinsic parameters. To ensure the accuracy of camera parameters, we select 82 * 8 video segments from each session for initial triangulation and adjust the sampling rate based on reconstruction results.

3.2. Action Set

The action set in M3GYM comes from real training sessions conducted at the partner gym. These sessions include three types: Normal, Pilates, and Yoga sessions. In each 45-minute **Normal session**, sports experts design training routines for each participant, with actions like squat jumps, standing calf raises, mountain climbers, and burpees. Participants train by completing a set number of repetitions for each action. Due to variations in fitness levels and training routines, participants often perform different actions simultaneously during normal sessions. Sports experts record the specific training flow for each participant and assess their action completeness.

Each Pilates and Yoga session lasts about an hour and includes more complex movements with significant self-occlusion. **Pilates sessions** involve movements like knee hover side twists, pendulums, bicep curls, and straight leg side toe reaches, while **Yoga sessions** include poses like humble warrior, knee hovers, fire hydrants, and side-lying tree. Unlike the customized routines in normal training sessions, these two types of sessions are led by sports experts who guide participants to perform each action for specific time intervals, resulting in synchronized actions among participants. Sports experts assess each participant's performance at various stages.

3.3. Semi-automated Pipeline

We design a semi-automated annotation pipeline for M3GYM as shown in Figure 2. This pipeline extracts 2D keypoints, 3D keypoints, and mesh results from RGB video. Each modality includes a corresponding error correction step, covering frame and pixel alignment, 2D annotation voting, 3D manual adjustment, and action labeling refinement based on records from sports experts.

RGB video. We process raw videos through frame and pixel alignment to obtain synchronized RGB videos. The eight cameras operate under a unified control system. By recording each device's start time on this central console, we determine the starting frame for each view (frame alignment). To further minimize potential recording errors, inspired by FreeMan [60], we capture corresponding frames from each view after frame alignment and use LightGlue [40] to compute dense correspondences across views. These correspondences are used to further refine the camera parameters (pixel alignment).

2D keypoint. From the synchronized videos, we ex-

Figure 3. **Illustration of samples from M3GYM. (a)** Diverse lighting conditions. **(b)** Complex self-occlusions. **(c)** Mutual occlusions in dense sessions. **(d)** Realistic gym setting with symmetric floor mirrors. **(e)** Unique overhead view, where a single camera may miss some participants or fail to capture a subject in full.

tract frames and apply multiple 2D pose estimation models [8, 17, 21, 27, 45, 59, 66, 68, 70] to obtain several 2D annotations for each frame. All annotations are converted into the BODY25 [8] format, where we omit the six keypoints for foot during conversion when converting outputs in COCO17 [39] format. Using IoU bounding box calculations and hip distance between subjects, we match subjects across model outputs. For each subject, we apply median voting and non-max suppression to filter the points, producing initial 2D keypoint annotations. Among the nine tested models, DWPose [70], RTMPose [27], and ViTPose [66] show relatively lower rates of missed and false detections compared to the ground truth 2D keypoint annotations, making them suitable for the subsequent pipeline stages to obtain the raw 2D keypoint annotations with confidence levels across 8 views, denoted as $Kpt_{2D} \in \mathbb{R}^{8\times25\times3}$. To ensure the accuracy of triangulation results, we further filter these raw 2D keypoint annotations. All points with a confidence level below the threshold τ (we set $\tau = 0.5$) are marked as low confidence and removed. Subjects with fewer than m high-confidence points (we set $m = 5$) or missing essential keypoints (we set these as the left and right shoulders) are also excluded, resulting in filtered 2D keypoint annotations.

3D keypoint. We apply triangulation on the filtered 2D keypoint annotations to obtain 3D keypoint with confidence level for each subject, denoted as $Kpt_{3D} \in \mathbb{R}^{25\times4}$. After non-max suppression filtering, we use bone length and smoothing constraints introduced in HuMMan [6] to optimize the 3D keypoint annotations. To ensure annotation accuracy, we build a 3D calibration tool in Blender [10]. This tool visualizes each subject's 3D poses in the scene and reprojects annotations across the eight 2D views. Details of the tool are provided in the supplementary materials. Through extensive manual inspection and adjustment, we produce ground truth 3D keypoint annotations, with the corresponding reprojected 2D coordinates in each view serving as ground truth 2D keypoint annotations.

Mesh output. Inspired by Freeman [60], we use SMPLify [5] to fit the SMPL [42] model to our ground truth 3D keypoints, generating mesh annotations and placing the mesh in the 3D scene based on subject coordinates. Based on notes taken by sports experts in each session, we segment each subject's meshes by time duration into distinct actions, assigning labels accordingly. By calculating changes in the relative distances between segmented skeletons, we further refine the action time spans. This allows us to manually check and correct any possible errors in action counts noted by the sports experts.

The final output of the M3GYM semi-automated pipeline includes the mesh annotations for each subject within specified time spans, along with the corresponding action labels, and expert assessments of each action. Action assessments are classified as *good* or *poor*, where for each *poor* action, M3GYM also provides text-format feedback identifying areas for improvement.

Table 2. **Key statistics of M3GYM by training session type.** These three session categories share overlapping action types and lighting conditions. "Mean #Subj" represents the average number of subjects per session. We determine the subject count for each session by taking the number of individuals present for the majority of the session duration.

	#Session	#Frame	#Act	#Lighting	Mean #Subj
Normal	51	26,500,189	50	4	5.29
Pilates	17	10,822,689	207	4	5.59
Yoga	14	9,710,460	254	7	6.21
Total	82	47,033,338	502	7	5.51

3.4. Dataset Statistics

M3GYM contains 82 multi-person training sessions, each lasting 40 to 60 minutes, captured by 8 overhead cameras for a total of 47,033,338 frames. As shown in Table 2, M3GYM includes 3 session types, Normal, Pilates, and Yoga. Normal sessions constitute the majority, encompassing 51 sessions and producing over 26 million frames, with an average of 5.29 subjects. Pilates sessions, with 17 sessions, involve more actions, resulting in 207 labeled activities across different lighting conditions. Yoga sessions, al-

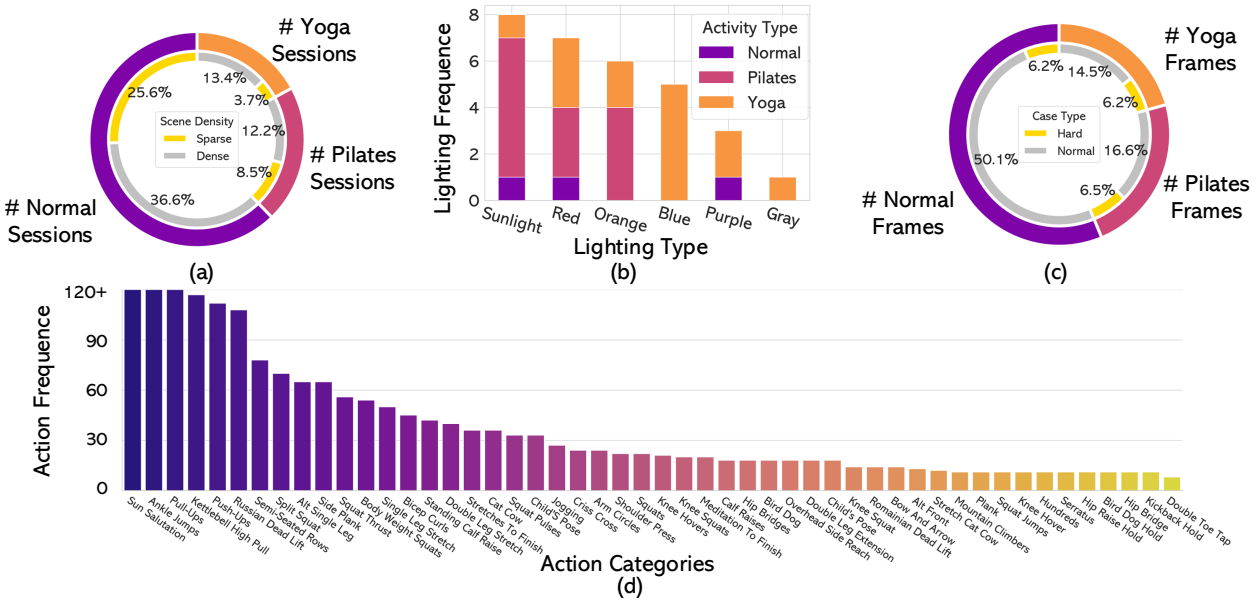

Figure 4. **Detailed statistics of M3GYM. (a)** The proportion of dense and sparse sessions across session types. **(b)** The distribution of lighting conditions across session types. The well-lit condition is common and not shown here. **(c)** The proportion of hard case frames across session types. **(d)** The distribution of the 50 most frequent action types in M3GYM.

though smaller in number, feature the highest average subject count (6.21) and the most diverse lighting variations, creating richer scenarios for self-occlusions. Overall, our dataset offers 502 distinct actions across varied lighting conditions and subject counts.

Samples from M3GYM. Figure 3 presents some samples from M3GYM, emphasizing its unique characteristics. **(a)** Unlike existing human pose datasets, M3GYM features a wide range of lighting conditions. In contrast to Freeman [60], which categorizes lighting simply as backlit or well-lit, our dataset provides clearly defined lighting settings, including well-lit, sunlight with backlighting, and colored lights used during training to create specific atmospheres. **(b)** M3GYM, especially in the Yoga and Pilates sessions, includes complex actions with significant self-occlusion. Even state-of-the-art methods exhibit notable errors when identifying such *hard cases*. **(c)** All videos in our dataset contain scenes with multiple participants. This leads to extensive mutual occlusions, challenging models to leverage multi-view data for accurate human pose estimation. **(d)** M3GYM replicates a realistic gym environment with two symmetric floor mirrors, creating additional challenges for pose estimation. **(e)** Unlike most datasets filmed at eye-level, ours provides a unique overhead view, where a single camera may miss subjects or fail to capture a subject in full, introducing further complexity.

Detailed statistic of M3GYM. In Figure 4, we present more detailed statistics of M3GYM. **(a)** Previous datasets, such as EgoBody [77] and RICH [24], focus on scenes with up to 2 people, while CHI3D [19] includes activities with a maximum of 3 subjects. Based on this context, we define sessions with more than 4 subjects as *dense* sessions. Figure 4 (a) shows the proportion of three session types and the proportion of *dense* sessions within each. The figure shows that *dense* sessions overwhelmingly dominate across all three session types, highlighting M3GYM's emphasis on crowded scenarios. **(b)** In addition to standard indoor lighting, our dataset includes six additional lighting conditions, including sunlight and various ambient lights. Figure 4 (b) shows the distribution of these extra lighting conditions across different session types. **(c)** Our dataset includes numerous complex actions with significant occlusions, causing substantial errors in many frames even for state-of-the-art methods. Based on manual annotations, frames requiring detailed review and adjustment are labeled as *hard cases*. Figure 4 (c) displays the proportion of frame count for each session type and the percentage of *hard case* frames. The high ratio of *hard case* frames highlights M3GYM's challenge for pose estimation. **(d)** Our dataset includes 502 distinct action types. Figure 4 (d) presents the top 50 most frequent action labels in M3GYM, with frequency calculated based on the presence of each action label per session for each subject. This distribution highlights the diversity of actions in M3GYM.

4. M3GYM Benchmark

This section presents and analyses benchmark results for various human pose estimation tasks on M3GYM. Additional experiments and details are provided in the supplementary material.

Table 3. **2D pose estimation baseline on M3GYM.** We present the results of widely used and state-of-the-art methods, showing inference results and performance after fine-tuned. **Bold** highlights the top value within each data type.

Method	Inference						Fine-tuned on M3GYM					
	AP	AP^{50}	AP^{75}	AR	AR^{50}	AR^{75}	AP	AP^{50}	AP^{75}	AR	AR^{50}	AR^{75}
DEKR [21]	19.9	26.4	21.0	76.1	90.6	79.9	53.6	76.8	64.5	76.4	91.0	80.1
YOLO-Pose [45]	62.6	75.5	66.5	76.4	90.7	80.2	63.3	80.4	72.3	77.8	90.8	80.4
ED-Pose [68]	14.0	18.8	14.9	74.4	90.0	78.5	53.2	76.1	60.9	74.8	90.2	78.6
DWPose [70]	63.8	79.7	68.0	76.7	89.5	80.0	72.6	90.3	77.8	77.2	90.8	81.7
RTMPose [27]	64.0	80.5	69.0	76.9	90.3	81.0	72.8	**90.7**	**78.9**	77.5	91.1	82.4
ViTPose [66]	**66.3**	**80.9**	**70.7**	**79.0**	**90.9**	**82.6**	**73.1**	90.5	78.2	**79.1**	**91.2**	**82.8**

4.1. Human Pose Estimation Tasks

2D pose estimation. This task involves detecting 2D keypoints (Kpt_{2D}) from single-view images. Models process individual frames to identify joint positions in the image plane, capturing the spatial configuration of the subject.

Multi-view 3D pose estimation. This task utilizes images from multiple synchronized cameras to estimate 3D poses. By integrating information from different viewpoints, models aim to reconstruct the 3D positions of joints (Kpt_{3D}), addressing occlusions and perspective ambiguities.

Single-view 3D pose estimation. This task requires models to predict 3D joint positions (Kpt_{3D}) from video sequences captured by a single camera. The challenge lies in inferring depth information from monocular inputs, and reconstructing the spatial arrangement of joints in 3D space.

Human mesh recovery. This task involves reconstructing a 3D mesh of the human body from images. Models aim to capture surface geometry with detail in both body shape and pose, offering a complete representation of the subject.

4.2. Evaluation Metric

AP^k and **AR^k** are common metrics in 2D pose estimation. For example, AP^{50} and AR^{50} denote Average Precision (AP) and Average Recall (AR) at a 50% Object Keypoint Similarity (OKS) threshold, per COCO evaluation [39]. AP and AR without suffixes represent averages across OKS thresholds from 0.5 to 0.95 in 0.05 increments.

MPJPE (Mean Per Joint Position Error) is commonly used in 3D pose estimation, measuring the mean Euclidean distance between predicted and ground truth joint positions.

PA-MPJPE (Procrustes-Aligned MPJPE) is a variant of MPJPE that first aligns the predicted pose to the ground truth, removing global misalignment.

4.3. Benchmark Results

Among 82 sessions (51/17/14 for Normal/Pilates/Yoga), 9/4/3 are allocated for testing, 4/2/2 for validation, and the rest for training, with frames subsampled to maintain data balance and efficiency, ensuring no session-level overlap across different sets while also minimizing identity overlap.

2D pose estimation. We evaluate several 2D pose estimation models [21, 27, 45, 66, 68, 70] on M3GYM. These models cover a range of approaches, with top-down methods like RTMPose and ViTPose generally excelling in accuracy metrics, while bottom-up models such as ED-Pose achieve strong recall. As shown in Table 3, ViTPose performs highest across most metrics, with RTMPose also achieving top scores in AP^{50} and AP^{75}. The lower inference performance of DEKR and ED-Pose may result from their bottom-up design, which faces challenges in highly occluded and varied poses in M3GYM. Notably, models fine-tuned on M3GYM show substantial improvement over direct inference, highlighting the dataset's value in enhancing pose estimation accuracy.

Multi-view 3D pose estimation. Multi-view 3D pose estimation accuracy relies strongly on the precision of 2D detection results. As shown in Table 3, ViTPose and RTMPose achieve the best performance. To evaluate M3GYM's impact on 3D keypoint accuracy, we use MV-Pose [11] to reconstruct 3D keypoints from the multi-view 2D keypoint outputs of ViTPose and RTMPose, then measure their deviation from the ground-truth 3D keypoints. As shown in Table 4, fine-tuning on M3GYM improves performance for both methods, underscoring M3GYM's ability to support precise multi-view 3D reconstructions.

Single-view 3D pose estimation. We evaluate single-view 3D pose estimation models on M3GYM, including SimpleBaseline [46], VideoPose3D [49], and MotionBERT [81]. Given M3GYM's unique overhead view, only subjects with their full ground-truth bounding boxes within the current view are used in testing, ensuring accurate evaluation. For multi-person scenes, we test each subject individually to prevent errors from relative coordinate differences. As shown in Table 5, MotionBERT achieves the lowest MPJPE after fine-tuning (123.7), while VideoPose3D attains the lowest PA-MPJPE (76.6). Fine-tuning on M3GYM substantially enhances each model's performance, demonstrating the dataset's utility in challenging gym settings.

Human mesh recovery. We evaluate human mesh recovery models on the M3GYM dataset, including PyMAF [74], OSX-SMPL [37], and SMPLer-L [65]. The evaluation setup follows the same approach as single-view 3D pose estimation, selecting only frames where the subject's full ground-truth bounding box is within the overhead view and

Table 4. **Multi-view 3D pose estimation baseline on M3GYM.**

Method	Status	MPJPE (mm)	PA-MPJPE (mm)
RTMPose [27]	Inference	132.4	76.9
	Fine-tuned	113.5	62.7
ViTPose [66]	Inference	123.6	77.8
	Fine-tuned	**112.3**	**61.7**

Table 5. **Single-view 3D pose estimation baseline on M3GYM.**

Method	Status	MPJPE (mm)	PA-MPJPE (mm)
SimpleBaseline [46]	Inference	201.4	127.6
	Fine-tuned	134.5	86.1
VideoPose3D [49]	Inference	165.4	115.1
	Fine-tuned	126.4	**76.6**
Motionbert [81]	Inference	158.9	111.2
	Fine-tuned	**123.7**	77.9

testing each subject individually in multi-person scenes to avoid relative coordinate discrepancies. As shown in Table 6, SMPLer-L achieves the best results, with the lowest MPJPE and PA-MPJPE after fine-tuning. OSX-SMPL and PyMAF also improve significantly with M3GYM fine-tuning, underscoring the dataset's effectiveness in enhancing mesh recovery accuracy in complex environments.

Cross-domain experiments. We conduct cross-domain evaluations across session types in M3GYM to assess model generalization. As shown in Table 7, models trained on Yoga sessions perform best on Pilates, achieving the highest AP and AR scores (90.8 and 91.9). Models trained on Normal and Pilates sessions show limited generalization to Yoga, likely due to Yoga's unique poses and higher self-occlusion. However, the Yoga-trained model's performance decreases significantly on Normal sessions (AP of 56.1), indicating varied adaptability across session types. These results highlight the distinct characteristics of each session type and underscore M3GYM's value in testing model adaptability.

5. Discussions and Insights

M3GYM provides key insights for advancing human pose estimation in real-world settings. **We establish a benchmark for testing models in realistic conditions.** Its diverse scenarios, including varied lighting, complex occlusions, floor mirrors, and unique fitness poses, allow researchers to evaluate model robustness against challenges typical of gym environments yet often absent in existing datasets. **M3GYM demonstrates the importance of multi-person and multi-view data for realistic applications.** Multi-person scenes reflect authentic gym settings, while multi-view perspectives are essential for handling occlusions in complex actions and crowded scenarios. **Additionally, M3GYM emphasizes the need for activity-specific data to enhance model adaptability.** Cross-domain results reveal performance drops when models trained on one session type are tested on others, underscoring the value of specialized datasets for different activities. Together, these insights establish M3GYM as a critical resource for developing adaptive algorithms suited to diverse, gym-based applications.

Table 6. **Human mesh recovery baseline on M3GYM.**

Method	Status	MPJPE (mm)	PA-MPJPE (mm)
PyMAF [74]	Inference	189.3	118.2
	Fine-tuned	155.1	94.5
OSX-SMPL [37]	Inference	153.5	104.7
	Fine-tuned	123.8	69.4
SMPLer-L [65]	Inference	149.6	94.3
	Fine-tuned	**116.4**	**67.2**

Table 7. **Cross-domain evaluation of 2D pose estimation across three session types in M3GYM.**

Train	Test	AP	AP^{50}	AP^{75}	AR	AR^{50}	AR^{75}
Normal	Pilates	32.0	67.3	24.6	39.4	72.8	34.9
	Yoga	16.6	30.8	14.7	17.5	31.3	15.9
Pilates	Normal	16.0	40.4	10.7	21.7	49.4	17.0
	Yoga	11.9	22.4	11.3	13.2	24.5	12.9
Yoga	Normal	56.1	83.2	59.6	59.9	84.4	63.7
	Pilates	90.8	98.0	93.8	91.9	98.0	94.6

Despite its strengths, M3GYM still has certain limitations. For example, Pilates and Yoga sessions make up only 31 out of 82 sessions, and lighting types beyond well-lit conditions are relatively rare and unevenly distributed across session types. However, Pilates and Yoga sessions make up a substantial 43.7% of the total frames, providing extensive data due to the long duration of each sequence. Although the gray lighting condition appears only once, it still includes 720,000 frames, offering sufficient data for model training despite its limited occurrence.

6. Conclusion

In conclusion, M3GYM provides a comprehensive resource for multimodal, multi-view, and multi-person pose estimation, capturing authentic gym activities with varied scenarios. Its fine-grained, time-span-based annotations, supported by the M3GYM Semi-automated Pipeline, deliver reliable ground truth for each subject and action, supporting a wide range of tasks from 2D and 3D pose estimation to human mesh recovery. Benchmark results show that M3GYM challenge current models and improves model performance in complex real-world settings, making it an invaluable tool for advancing human activity analysis under conditions of substantial occlusion and diverse training contexts.

Acknowledgments: This research is funded by the Australian Research Council (ARC) through DECRA Grant DE230100477 and Discovery Project Grant DP220100800.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 2

[2] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5167–5176, 2018. 2

[3] Aritz Badiola-Bengoa and Amaia Mendez-Zorrilla. A systematic review of the application of camera-based human pose estimation in the field of sport and physical exercise. *Sensors*, 21(18), 2021. 2

[4] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8726–8737, 2023. 2, 3

[5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 561–578. Springer, 2016. 5

[6] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, Fangzhou Hong, Mingyuan Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. HuMMan: Multi-modal 4d human dataset for versatile sensing and modeling. In *17th European Conference on Computer Vision, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 557–577. Springer, 2022. 2, 3, 5

[7] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[8] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 3, 5

[9] Yalin Cheng, Pengfei Yi, Rui Liu, Jing Dong, Dongsheng Zhou, and Qiang Zhang. Human-robot interaction method combining human pose estimation and motion intention recognition. In *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 958–963, 2021. 2

[10] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 5

[11] Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation from multiple views. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7792–7801, 2019. 3, 4, 7

[12] 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 transactions on pattern analysis and machine intelligence*, 44(10):6981–6992, 2021. 3

[13] Heming Du, Xin Yu, and Liang Zheng. Learning object relation graph and tentative policy for visual navigation. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VII*, pages 19–34. Springer, 2020. 3

[14] Heming Du, Xin Yu, and Liang Zheng. Vtnet: Visual transformer network for object goal navigation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[15] Heming Du, Zi Huang, Scott Chapman, and Xin Yu. Toward a unified framework for RGB and RGB-D visual navigation. In *AI 2023: Advances in Artificial Intelligence - 36th Australasian Joint Conference on Artificial Intelligence, AI 2023, Brisbane, QLD, Australia, November 28 - December 1, 2023, Proceedings, Part II*, pages 363–375. Springer, 2023.

[16] Heming Du, Lincheng Li, Zi Huang, and Xin Yu. Object-goal visual navigation via effective exploration of relations among historical navigation states. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 2563–2573. IEEE, 2023. 3

[17] 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. 3, 5

[18] Xiaoyu Feng, Heming Du, Hehe Fan, Yueqi Duan, and Yongpan Liu. Seformer: Structure embedding transformer for 3d object detection. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 632–640. AAAI Press, 2023. 3

[19] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3, 6

[20] Mihai Fieraru, Mihai Zanfir, Silviu Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. Aifit: Automatic 3d human-interpretable feedback models for fitness training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9919–9928, 2021. 2, 3

[21] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *Proceedings of the IEEE/CVF*

conference on computer vision and pattern recognition, pages 14676–14686, 2021. 3, 5, 7

[22] Tianchen Guo, Heming Du, Huan Huo, Bo Liu, and Xin Yu. Who is being impersonated? deepfake audio detection and impersonated identification via extraction of id-specific features. In *International Conference on Algorithms and Architectures for Parallel Processing*, pages 301–320. Springer, 2024. 3

[23] Abdelfetah Hentout, Mustapha Aouache, Abderraouf Maoudj, and Isma Akli. Human–robot interaction in industrial collaborative robotics: a literature review of the decade 2008–2017. *Advanced Robotics*, 33(15-16):764–799, 2019. 2

[24] 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 *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 13274–13285, 2022. 2, 3, 6

[25] 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 Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. 2, 3

[26] Boyuan Jiang, Lei Hu, and Shihong Xia. Probabilistic triangulation for uncalibrated multi-view 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14850–14860, 2023. 3

[27] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*, 2023. 3, 5, 7, 8

[28] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 2, 3

[29] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 3

[30] 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, 2025. 3

[31] Laxman Kumarapu and Prerana Mukherjee. Animepose: Multi-person 3d pose estimation and animation. *Pattern Recognition Letters*, 147:16–24, 2021. 2

[32] Askat Kuzdeuov, Darya Taratynova, Alim Tleuliyev, and Huseyin Atakan Varol. Openthermalpose: An open-source annotated thermal human pose dataset and initial yolov8-pose baselines. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2024. 2

[33] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10863–10872, 2019. 2

[34] Ruilong Li, Sha Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13381–13392, 2021. 2, 3

[35] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13147–13156, 2022. 3

[36] Jiahao Lin and Gim Hee Lee. Multi-view multi-person 3d pose estimation with plane sweep stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11886–11895, 2021. 3

[37] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21159–21168, 2023. 3, 7, 8

[38] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1954–1963, 2021. 3

[39] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 2, 5, 7

[40] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. In *ICCV*, 2023. 4

[41] Huan Liu, Qiang Chen, Zichang Tan, Jiang-Jiang Liu, Jian Wang, Xiangbo Su, Xiaolong Li, Kun Yao, Junyu Han, Errui Ding, et al. Group pose: A simple baseline for end-to-end multi-person pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15029–15038, 2023. 3

[42] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 5

[43] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, et al. Mediapipe: A framework for perceiving and processing reality. In *Third workshop on computer vision for AR/VR at IEEE computer vision and pattern recognition (CVPR)*, 2019. 3

[44] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, 2019. 2

[45] Debapriya Maji, Soyeb Nagori, Manu Mathew, and Deepak Poddar. Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. In *Proceed-*

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2637–2646, 2022. 3, 5, 7

[46] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2640–2649, 2017. 3, 7, 8

[47] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal V. Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. *2017 International Conference on 3D Vision (3DV)*, pages 506–516, 2016. 2, 3

[48] 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 *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3

[49] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 7, 8

[50] Feng Qiu, Wei Zhang, Chen Liu, Rudong An, Lincheng Li, Yu Ding, Changjie Fan, Zhipeng Hu, and Xin Yu. Freeavatar: Robust 3d facial animation transfer by learning an expression foundation model. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 3

[51] Xin Shen, Shaozu Yuan, Hongwei Sheng, Heming Du, and Xin Yu. Auslan-daily: Australian sign language translation for daily communication and news. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 2

[52] Xin Shen, Heming Du, Hongwei Sheng, Shuyun Wang, Hui Chen, Huiqiang Chen, Zhuojie Wu, Xiaobiao Du, Jiaying Ying, Ruihan Lu, Qingzheng Xu, and Xin Yu. Mm-wlauslan: Multi-view multi-modal word-level australian sign language recognition dataset, 2024. 2

[53] Hui Shuai, Lele Wu, and Qingshan Liu. Adaptive multi-view and temporal fusing transformer for 3d human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4122–4135, 2022. 3

[54] Leonid Sigal, Alexandru Balan, and Michael Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87:4–27, 2010. 2, 3

[55] Jan Stenum, Kendra M. Cherry-Allen, Connor O Pyles, Rachel Reetzke, Michael F. Vignos, and Ryan T. Roemmich. Applications of pose estimation in human health and performance across the lifespan. *Sensors (Basel, Switzerland)*, 21, 2021. 2

[56] Lei Su, Jinhua She, and Chi Xu. Estimating human pose with both physical and physiological constraints. In *2021 4th IEEE International Conference on Industrial Cyber-Physical Systems (ICPS)*, pages 693–699, 2021. 2

[57] 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 *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X*, page 614–631, Berlin, Heidelberg, 2018. Springer-Verlag. 2, 3

[58] Thomas Waltemate, Dominik Gall, Daniel Roth, Mario Botsch, and Marc Erich Latoschik. The impact of avatar personalization and immersion on virtual body ownership, presence, and emotional response. *IEEE transactions on visualization and computer graphics*, 24(4):1643–1652, 2018. 2

[59] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 5

[60] Jiong Wang, Fengyu Yang, Bingliang Li, Wenbo Gou, Danqi Yan, Ailing Zeng, Yijun Gao, Junle Wang, Yanqing Jing, and Ruimao Zhang. Freeman: Towards benchmarking 3d human pose estimation under real-world conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21978–21988, 2024. 2, 3, 4, 5, 6

[61] Suzhen Wang, Weijie Chen, Wei Zhang, Minda Zhao, Lincheng Li, Rongsheng Zhang, Zhipeng Hu, and Xin Yu. Easycraft: A robust and efficient framework for automatic avatar crafting. *arXiv preprint arXiv:2503.01158*, 2025. 3

[62] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, 2022. 2

[63] Qingzheng Xu, Huiqiang Chen, Heming Du, Hu Zhang, Szymon Łukasik, Tianqing Zhu, and Xin Yu. M3a: A multimodal misinformation dataset for media authenticity analysis. *Computer Vision and Image Understanding*, 249: 104205, 2024. 3

[64] Qingzheng Xu, Heming Du, Huiqiang Chen, Bo Liu, and Xin Yu. Mmooc: A multimodal misinformation dataset for out-of-context news analysis. In *Australasian Conference on Information Security and Privacy*, pages 444–459. Springer, 2024. 3

[65] Xiangyu Xu, Lijuan Liu, and Shuicheng Yan. Smpler: Taming transformers for monocular 3d human shape and pose estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(5): 3275–3289, 2024. 3, 7, 8

[66] 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. 3, 5, 7, 8

[67] Jianfei Yang, He Huang, Yunjiao Zhou, Xinyan Chen, Yuecong Xu, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie. Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 2, 3

[68] Jie Yang, Ailing Zeng, Shilong Liu, Feng Li, Ruimao Zhang, and Lei Zhang. Explicit box detection unifies end-to-end

multi-person pose estimation. In *International Conference on Learning Representations*, 2023. 3, 5, 7

[69] Zhitao Yang, Zhongang Cai, Haiyi Mei, Shuai Liu, Zhaoxi Chen, Weiye Xiao, Yukun Wei, Zhongfei Qing, Chen Wei, Bo Dai, Wayne Wu, Chen Qian, Dahua Lin, Ziwei Liu, and Lei Yang. Synbody: Synthetic dataset with layered human models for 3d human perception and modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20282–20292, 2023. 2, 3

[70] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023. 3, 5, 7

[71] Hang Ye, Wentao Zhu, Chunyu Wang, Rujie Wu, and Yizhou Wang. Faster voxelpose: Real-time 3d human pose estimation by orthographic projection. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[72] Jae Shin Yoon, Lingjie Liu, Vladislav Golyanik, Kripasindhu Sarkar, Hyun Soo Park, and Christian Theobalt. Pose-guided human animation from a single image in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15039–15048, 2021. 2

[73] Jae Shin Yoon, Zhixuan Yu, Jaesik Park, and Hyun Soo Park. Humbi: A large multiview dataset of human body expressions and benchmark challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):623–640, 2021. 2

[74] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3, 7, 8

[75] Lijun Zhang, Kangkang Zhou, Feng Lu, Xiang-Dong Zhou, and Yu Shi. Deep semantic graph transformer for multi-view 3d human pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7205–7214, 2024. 3

[76] Lijun Zhang, Kangkang Zhou, Feng Lu, Zhenghao Li, Xiaohu Shao, Xiang-Dong Zhou, and Yu Shi. Esmformer: Error-aware self-supervised transformer for multi-view 3d human pose estimation. *Pattern Recognition*, 158:110955, 2025. 3

[77] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *European Conference on Computer Vision*, pages 180–200. Springer, 2022. 2, 3, 6

[78] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8877–8886, 2023. 3

[79] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11656–11665, 2021. 3

[80] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *ACM Comput. Surv.*, 56(1), 2023. 2

[81] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3, 7, 8

[82] Matko Šarić, Mladen Russo, Luka Kraljević, and Davor Meter. Extended reality telemedicine collaboration system using patient avatar based on 3d body pose estimation. *Sensors*, 24 (1), 2024. 2