

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

## Supplementary Material

### A. Necessity of the M3GYM

#### A.1. Selection of fitness movements

Fitness movements are chosen by biomechanics experts to enhance body balance and flexibility. These movements usually feature repetitions, supporting tasks like repetition counting and action localization. Despite being widely adopted in gyms, fitness movements remain challenging for current models due to their complexity. To bridge this gap, M3GYM includes more complex and uncommon movements, such as Yoga, to support fitness activity analysis. Along with “good” or “poor” ratings for each action, biomechanics specialists also provide explanations on poorly performed actions. These explanations can be employed to train exercise report generation models.

#### A.2. Occlusion frame ratios:

We calculate occlusion frame ratios by measuring the overlap of bounding boxes and projecting annotated 3D meshes (if available) to the image planes. M3GYM’s ratio is 82.1%, significantly higher than 51.7% (CMU Panoptic) and 62.3% (CHI3D), highlighting the challenges of M3GYM.

### B. Details of the Semi-automated Pipeline

#### B.1. Camera Calibration

A commercial video recording system synchronizes the eight cameras via hardware synchronization, and two annotators review all video segments to ensure frame-level alignment and viewpoint consistency.

We determine the camera parameters using a chessboard-based calibration method applied in MV-Pose [13]. For intrinsic calibration, we move a chessboard with a 9x6 inner corner pattern and a grid size of 0.1 meters per square across the field of view of all eight cameras. By capturing approximately 240,000 frames, we extract the 2D coordinates of the chessboard corners for each camera. These measurements are then used to calculate intrinsic parameters, including focal length, principal point, and lens distortion coefficients. For extrinsic calibration, we place the chessboard at the center of the gym, ensuring visibility across all cameras.

To verify the accuracy of the camera parameters, we select one video segment from each view in every session. These sequences are used for triangulation to evaluate the quality of 3D reconstruction. Specifically, we analyze 82\*8 video segments across the eight cameras and adjust the sam-

pling rate based on reconstruction results. Inspired by FreeMan [65], we minimize potential recording errors by first aligning frames across views, then capturing synchronized frames from each camera, and using LightGlue [43] to compute dense feature correspondences between images. These correspondences provide additional constraints for pixel-level adjustments of the camera parameters. Together, these processes enhance alignment accuracy across views, optimizing calibration parameters for precise multi-view pose estimation in subsequent tasks.

#### B.2. 2D Keypoint

From the synchronized videos, we extract frames and apply multiple 2D pose estimation models [8, 19, 23, 29, 48, 64, 71, 73, 75] to generate diverse 2D annotations for each frame. To standardize the output, all annotations are converted into the BODY25 [8] format, ensuring consistency across different methods. This conversion process accounts for differences in keypoint definitions and formats between the models. For models such as AlphaPose [19] that provide whole-body keypoints, the BODY25 keypoints are directly mapped to their corresponding locations. For models that output formats like COCO17 [42], specific adjustments are required. BODY25 includes six additional foot keypoints absent in COCO17, filled as null values [0, 0, 0]. Additionally, certain keypoints, such as the Neck and Mid Hip, are calculated as midpoints: the Neck from the shoulders points and the Mid Hip from the hips points.

We align subjects across 2D pose estimation model outputs using Intersection over Union (IoU) of bounding boxes and Euclidean distance between hip keypoints. A match is valid when the IoU exceeds  $S_{IoU} = 0.7$  and the hip distance is below  $S_{hip}$ , set to one-twentieth of the image width. IoU ensures spatial overlap, while hip distance adds anatomical consistency. When hip keypoints are missing, IoU alone determines the match. A subject is considered valid only if it is detected by more than half of the models.

We refine 2D keypoint annotations using median voting and non-max suppression. Median voting considers keypoints with confidence above  $\tau_{vote} = 0.5$ , defining these as high-confidence keypoints. Keypoints with confidence below  $\tau_{vote} = 0.5$  are discarded. For an odd number of matches, the median point is selected. While for an even number, the higher-confidence middle point is chosen. Non-max suppression removes redundant bounding boxes with IoU exceeding 0.7, prioritizing those with more high-confidence keypoints and larger areas. Subjects with fewer

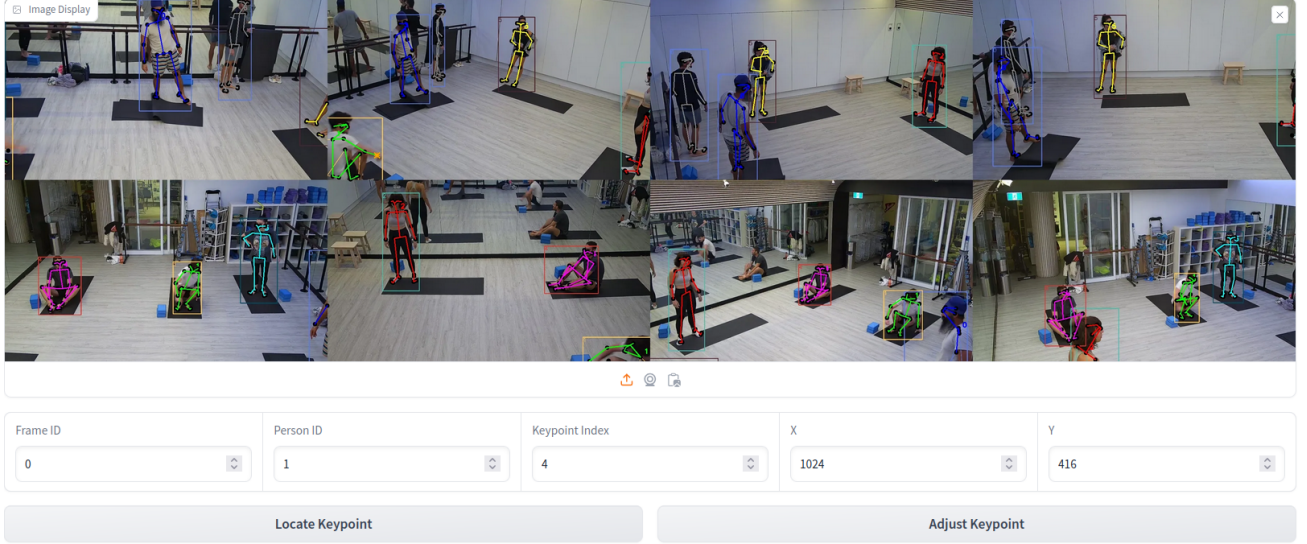


Figure 5. **Illustration of the Gradio-based 3D keypoint adjustment tool.** It displays 3D keypoints reprojected onto 2D views. Users can locate the keypoint by clicking on the image and adjust its position.

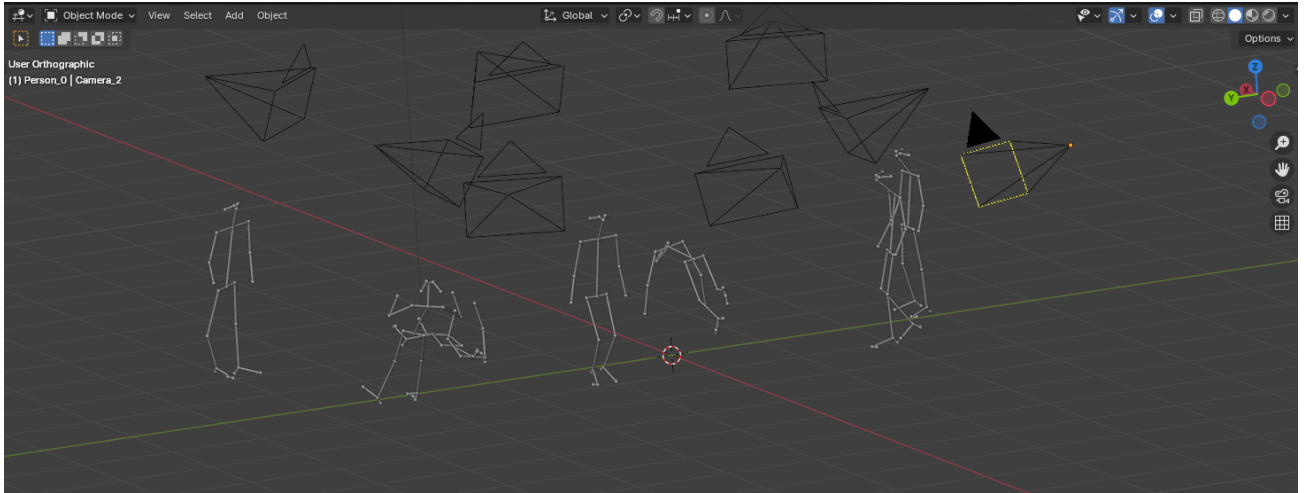


Figure 6. **Illustration of the Blender-based 3D keypoint adjustment tool.** It provides a 3D visualization of spatial relationships between subjects and keypoints. Users can perform detailed adjustments to achieve accurate annotations.

than  $m = 5$  high-confidence keypoints or missing the left or right shoulder are excluded. The dimension of *voted 2D keypoints* for each subject is  $8 \times 25 \times 3$ , voted by predictions of multiple 2D detectors across 8 views in BODY25 format. Then, we associate the *voted 2D keypoints* into 3D. After 3D annotations, we reproject 3D keypoints to each image plane to obtain *annotated 2D keypoints* of dimension  $25 \times 2$  for each view.

### B.3. 3D Keypoint

To generate 3D keypoint annotations, we perform triangulation using filtered 2D keypoint annotations from multiple views. This reconstructs 3D points by calculating inter-

sections of projection rays from different cameras. To refine the 3D keypoints, we apply bone length and smoothing constraints from HuMMan [6], ensuring anatomical proportions and temporal consistency. Despite these optimizations, errors can still occur in complex scenarios involving occlusions or challenging poses.

To address these issues, we develop 3D keypoint adjustment tools. First, we implement a Gradio-based tool, as shown in Figure 5, which reprojects 3D keypoints onto 2D views. Users can visually identify errors, click on the image to locate the nearest keypoint (highlighted with an orange star in the selected view), and view the corresponding person ID and keypoint index. By clicking “Adjust Keypoint”,

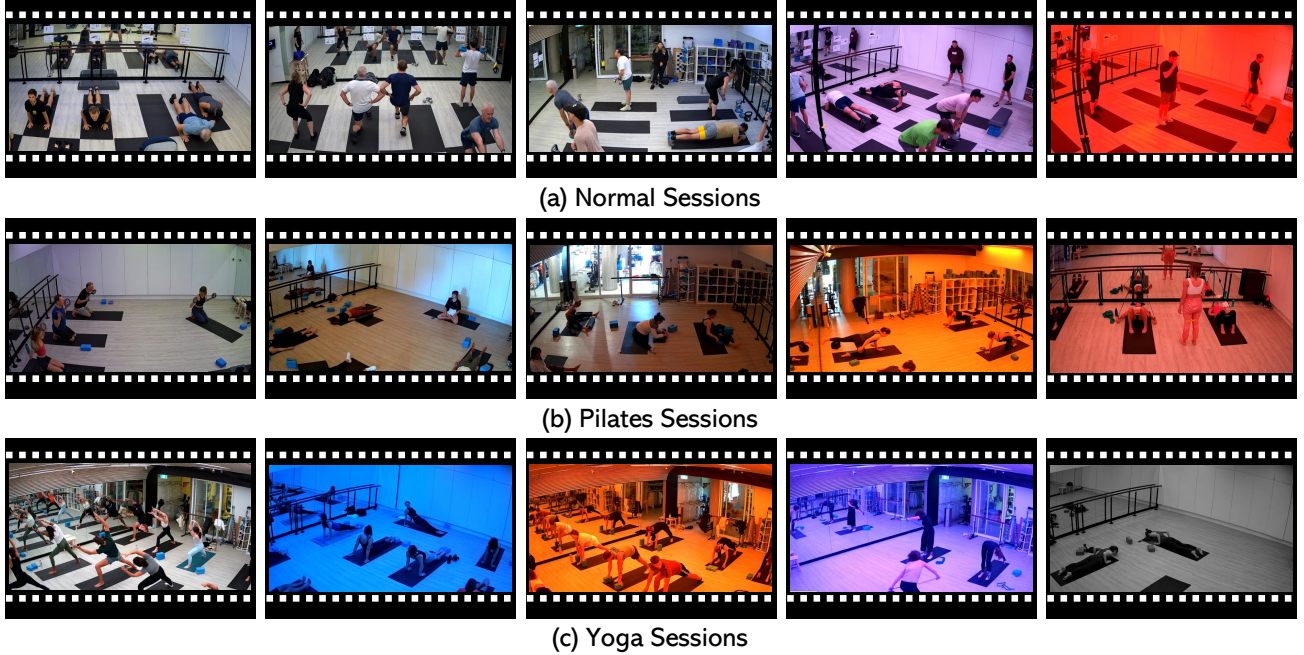


Figure 7. Illustration of samples from M3GYM across three session types.

the tool updates the located keypoint to the clicked position. For complex cases, such as off-center subjects where 2D re-projections lack sufficient detail, we design a Blender-based tool, shown in Figure 6. This tool visualizes the spatial relationships in 3D, enabling users to make precise adjustments and achieve high-quality final annotations.

## B.4. Mesh

Inspired by Freeman [65], we apply SMPLify [5] to fit the SMPL [45] model to ground truth 3D keypoints, generating body meshes for each subject and positioning them within the 3D scene. Sports experts provide detailed notes for each session, including the subject’s person ID, the sequence and labels of actions performed, the number of repetitions, time spent on each action type, an overall assessment (*Good* or *Poor*), and suggestions for improving *poor* actions. Based on this information, we divide each subject’s mesh data into individual actions based on time intervals, assigning the corresponding labels and assessments.

To improve the accuracy of these time-based segmentations, we calculate the relative distances between skeletons in consecutive meshes. This analysis refines the start and end times of actions, corrects any errors in the notes provided by the sports experts, and ensures that transitions between movements are accurately defined and the annotations align with the observed actions.

## B.5. More Annotation pipeline details:

Our pipeline integrates 2D annotation voting and 3D manual adjustment, followed by manual verification, while FreeMan does not involve manual 3D rectification and validation. Each session has been cross-validated by two annotators: 10% of frames are manually reviewed, and sessions with an error rate over 0.5% have been re-annotated. In the 2D median voting stage, our pipeline removes 1.91 false detections and recalls 1.07 inconsistent detections per frame on average. In the 3D adjustment stage, it modifies 45.8% of subjects in 18.9% of frames, reducing MPJPE by 65.7 mm. Due to severe self-occlusion, complex actions, such as Glute-Stretch, requires manual rectification of 71.4% of subjects in 85.1% of involved frames.

## C. Details of M3GYM

As shown in Figure 7, M3GYM includes diverse samples. This section provides a detailed analysis of the Normal, Pilates, and Yoga session types in M3GYM, along with an analysis of lighting conditions. For convenience, this section uses abbreviations to refer to the samples in Figure 7. For example, a-1 refers to the first sample of the Normal sessions, located in the first row, first column.

### C.1. Session Types

**Normal Session.** Normal sessions include basic fitness exercises such as squat thrusts, planks, and standing calf raises. In Pilates and Yoga sessions, sports experts guide



participants through each movement. In contrast, normal sessions involve coaches designing personalized fitness routines for each participant. Participants complete one set of an exercise and then move on to the next. Due to differences in physical ability and experience, participants typically perform the same exercise only during the initial stage (a-1). At other times, subjects in the same scene follow their own routines and perform different exercises, as shown in a-2, a-3, and a-4. Lighting conditions include well-lit settings (a-1, a-2 and a-3), sunlight, red (a-5), and purple (a-4).

**Pilates Session.** Pilates focuses on controlled actions designed to enhance flexibility, strength, and balance. In M3GYM, Pilates sessions often involve using props such as small weight plates, resistance bands, and foam mats to assist participants in performing exercises. Compared to normal sessions, Pilates includes more unique self-occluding actions and features more sessions with special lighting conditions, such as red (b-5) and orange (b-4). Notably, Pilates sessions have the most scenes under sunlight conditions, as shown in b-2 and b-3.

**Yoga Session.** Yoga emphasizes flexibility, strength, and mindfulness through a variety of structured actions and controlled breathing techniques. In M3GYM, Yoga sessions include the most diverse and unique actions, often involving self-occluding poses. Yoga also has the highest average number of participants per scene, creating significant mutual occlusions. These sessions feature all seven lighting conditions, including distinctive cases such as blue (c-2) and gray (c-5). These attributes make Yoga sessions the most challenging and unique part of M3GYM.

## C.2. Lighting Conditions

**Sunlight in M3GYM.** In M3GYM, sunlight refers specifically to scenes illuminated only by natural sunlight, unlike well-lit conditions. These sessions are typically darker due to the indoor setting. Since sunlight enters only through the gym’s main entrance, some of the eight camera views feature significant backlighting. For example, b-2 shows a view without backlighting, while b-3 includes backlighting, making it harder to distinguish details of the subjects.

**Lighting Intensity.** Differences in light sources create varying lighting intensities across different camera views under the sunlight condition. Similarly, the five special lighting conditions beyond well-lit and sunlight produce different visual appearances across views. For example, in the red light condition, a-5 and b-5 appear distinct. Likewise, in the purple light condition, a-4 and c-4 show different intensities. These variations require detection models with stronger recognition capabilities.

Table 8. **Additional baselines for various tasks on M3GYM.** MPJPE/PA-MPJPE (mm) are used for evaluation.

Task	Method	Inference	Fine-tuned
Multi-view 3D	VoxelPose [61]	123.0 / 72.3	101.4 / 53.4
	SelfPose3d [58]	127.8 / 75.4	98.6 / 51.2
Single-view 3D on groups	3D Multi-Person Pose [10]	163.7 / 115.9	125.4 / 78.1
	RTMW3D [30]	159.0 / 114.3	123.9 / 76.7
Human mesh recovery	METRO [41]	153.6 / 104.3	124.3 / 70.5
	SMPLer-X [7]	151.2 / 103.8	122.7 / 67.9

## D. Experiments Settings

### D.1. Training Hardware

All experiments are performed on machines configured with four NVIDIA A100 GPUs, each offering 80GB of memory. This setup provides the computational power and memory capacity required to handle large-scale data processing, complex model training, and evaluation tasks efficiently.

### D.2. Evaluation Metrics

Below are the details of the evaluation metrics used in our experiments.

**AP<sup>k</sup>** and **AR<sup>k</sup>** evaluate 2D pose estimation based on Object Keypoint Similarity (OKS). OKS serves as an IoU-like metric for keypoints and is defined as:

$$\text{OKS} = \frac{\sum_i \exp\left(-\frac{d_i^2}{2s^2k_i^2}\right) \delta(v_i > 0)}{\sum_i \delta(v_i > 0)}, \quad (1)$$

where  $d_i$  is the Euclidean distance between detected and ground truth keypoints,  $s$  represents the object scale,  $k_i$  is a per-keypoint constant, and  $v_i$  indicates the visibility of the ground truth keypoint. This OKS metric measures similarity based on the spatial alignment of keypoints rather than traditional IoU.

**MPJPE (Mean Per Joint Position Error)** measures the mean Euclidean distance between predicted and ground truth joint positions in 3D pose estimation, defined as:

$$\text{MPJPE} = \frac{1}{N} \sum_{i=1}^N \|\mathbf{J}_{\text{pred},i} - \mathbf{J}_{\text{gt},i}\|, \quad (2)$$

where  $N$  is the number of joints,  $\mathbf{J}_{\text{pred},i}$  is the predicted 3D position of the  $i$ -th joint, and  $\mathbf{J}_{\text{gt},i}$  is the ground truth 3D position of the  $i$ -th joint.

**PA-MPJPE (Procrustes-Aligned MPJPE)** calculates MPJPE after aligning the predicted pose to the ground truth using Procrustes alignment, removing global translation and rotation errors. The error is then calculated as:

$$\text{PA-MPJPE} = \frac{1}{N} \sum_{i=1}^N \|\text{Align}(\mathbf{J}_{\text{pred})}_i - \mathbf{J}_{\text{gt},i}\|, \quad (3)$$

where  $\text{Align}(\mathbf{J}_{\text{pred}})$  represents the predicted joint positions after Procrustes alignment to the ground truth joints.

### D.3. Models and Frameworks

We use publicly available frameworks and models as baselines in our experiments and express our gratitude to the authors of these works for their contributions. The sources for the models are listed below:

- **2D Pose Estimation:** OpenPose [8] (link), DEKR [23] (link), AlphaPose [19] (link), ViTPose [71] (link), YOLO-Pose [48] (link, link), YOLOv7-Pose [64] (link), ED-Pose [73] (link), DWPose [75] (link), RTMPose [29] (link), and MMPose [12] (link) for model training.
- **3D Pose Estimation:** MV-Pose [13] (link), Simple-Baseline [49] (link), VideoPose3D [52] (link), MotionBERT [86] (link).
- **Human Mesh Recovery:** PyMAF [79] (link), OSX-SMPL [40] (link), and SMPLer-L [70] (link).

### D.4. More benchmark results

In Table 8, we provide more state-of-the-art 3D pose estimation and mesh recovery methods to emphasize the challenges of M3GYM across diverse tasks.

### D.5. More benchmark analysis and insights

Under consistent lighting and visible angles, a model trained on Normal sessions and tested on Glute-Stretch (from Yoga) suffers a 9.3% AP drop. This indicates current methods still struggle to estimate complex and heavily self-occluded poses. Particularly, occluded keypoints cause a 5.4% AP drop. Moreover, we observe illumination conditions also affect pose estimation performance on M3GYM (2.6% AP drop) even though simple actions are presented. Since M3GYM contains various occluded and uncommon poses, MPJPE on M3GYM is higher than that of other datasets, implying the challenges of our M3GYM.

## E. Consent Form of M3GYM Recording

As mentioned in Section 3.1, before participating, all individuals review the experiment details and sign a consent form, as shown in Figure 8. The consent process is essential due to the inclusion of body and facial information in our dataset. While we record body movements and facial data, no personally identifiable details, such as names, ages, or occupations, will be released. Facial features are anonymized to ensure privacy and prevent identification. The dataset is exclusively intended for academic research and is not permitted for commercial use.

## F. Future work

Outdoor scene collection often requires more sophisticated camera systems and dedicated venues. Pose estimation will

Consent Form for M3GYM Recording	
<b>Purpose of the Project</b>	
This project aims to collect data for the M3GYM dataset, supporting research in multi-view, multi-person pose estimation and fitness activity analysis. The dataset contributes to advancements in pose estimation, activity recognition, and motion analysis in gym environments.	
<b>Participation Details</b>	
You will be recorded while performing fitness exercises. These recordings may include your body movements and facial information.	
<b>Privacy and Data Use</b>	
The recorded data will be used solely for academic research and will not be employed for any commercial purposes. All data will be anonymized to protect your identity and ensure privacy. The recordings may be used in academic conferences, research publications, or educational materials.	
<b>Consent</b>	
1. I have read and understood the purpose and details of this project. 2. I agree to participate in the recording sessions. 3. I understand that my participation is voluntary and that I can withdraw at any time without consequences. 4. I agree to the use of my body movements and facial information for academic research.	
<b>Participant Information</b>	
• Name: _____ • Signature: _____ • Date: _____	
For further information or questions about this project, please contact:	
Contact Person: [Coordinator's Name]	
Email: [Coordinator's Email]	
Phone: [Coordinator's Phone]	

Figure 8. Consent Form of M3GYM Recording.

be more likely affected by illuminations and object resolution, and tracking across cameras would impose new challenges. Since indoor scenes already present many challenges that existing methods cannot fully solve, we aim to address these difficulties first and extend our work to outdoor scenes in the future.

## 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, 4
- [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, 5
- [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, 2, 6
- [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] Yu Cheng, Bo Wang, and Robby Tan. Dual networks based 3d multi-person pose estimation from monocular video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 5
- [11] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 5
- [12] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. <https://github.com/open-mmlab/mmpose>, 2020. 6
- [13] 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, 2, 6
- [14] 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
- [15] 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
- [16] 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.
- [17] 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.
- [18] 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
- [19] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alpha-pose: 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, 2, 6
- [20] 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
- [21] 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

- [22] 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
- [23] 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, 2, 6
- [24] 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
- [25] 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
- [26] 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
- [27] 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
- [28] 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
- [29] 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, 2, 6
- [30] Tao Jiang, Xinchun Xie, and Yining Li. Rtmw: Real-time multi-person 2d and 3d whole-body pose estimation. *arXiv preprint arXiv:2407.08634*, 2024. 5
- [31] 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
- [32] 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
- [33] 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
- [34] Laxman Kumarapu and Prerana Mukherjee. Animepose: Multi-person 3d pose estimation and animation. *Pattern Recognition Letters*, 147:16–24, 2021. 2
- [35] Askat Kuzdeuov, Darya Taratynova, Alim Tleuliyeu, 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
- [36] 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
- [37] 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
- [38] 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
- [39] 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
- [40] 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, 6
- [41] 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, 5
- [42] 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
- [43] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. In *ICCV*, 2023. 4, 2
- [44] Huan Liu, Qiang Chen, Zichang Tan, Jiang-Jiang Liu, Jian Wang, Xiangbo Su, Xiaolong Li, Kun Yao, Junyu Han, Er-rui 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
- [45] 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, 4
- [46] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuoling 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
- [47] 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
- [48] Debapriya Maji, Soyeb Nagori, Manu Mathew, and Deepak Poddar. Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2637–2646, 2022. 3, 5, 7, 2, 6
- [49] 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, 6
- [50] 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
- [51] 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
- [52] 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, 6
- [53] 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
- [54] 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
- [55] 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
- [56] 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
- [57] 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
- [58] Vinkle Srivastav, Keqi Chen, and Nicolas Padoy. Selfpose3d: Self-supervised multi-person multi-view 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2502–2512, 2024. 5
- [59] 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
- [60] 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
- [61] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In *European Conference on Computer Vision (ECCV)*, 2020. 5
- [62] 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
- [63] 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
- [64] 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, 2, 6
- [65] 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
- [66] 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
- [67] 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
- [68] 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



- [69] 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](#)
- [70] Xiangyu Xu, Lijuan Liu, and Shuicheng Yan. Smpier: Tampering transformers for monocular 3d human shape and pose estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(5): 3275–3289, 2024. [3](#), [7](#), [8](#), [6](#)
- [71] 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](#), [2](#), [6](#)
- [72] 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](#)
- [73] 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](#), [2](#), [6](#)
- [74] 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](#)
- [75] 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](#), [2](#), [6](#)
- [76] 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](#)
- [77] 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](#)
- [78] 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](#)
- [79] 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](#), [6](#)
- [80] 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](#)
- [81] 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](#)
- [82] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Ego-body: 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](#)
- [83] 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](#)
- [84] 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](#)
- [85] 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](#)
- [86] 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](#), [6](#)
- [87] 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](#)