

AthletePose3D: A Benchmark Dataset for 3D Human Pose Estimation and Kinematic Validation in Athletic Movements

Calvin Yeung¹ Tomohiro Suzuki¹ Ryota Tanaka¹ Zhuoer Yin¹ Keisuke Fujii^{1,2}

¹ Graduate School of Informatics, Nagoya University, Nagoya, Japan

² Center for Advanced Intelligence Project, RIKEN, Osaka, Japan

{yeung.chikwong, suzuki.tomohiro, tanaka.ryota, yin.zhuoer, fujii}@g.sp.m.is.nagoya-u.ac.jp

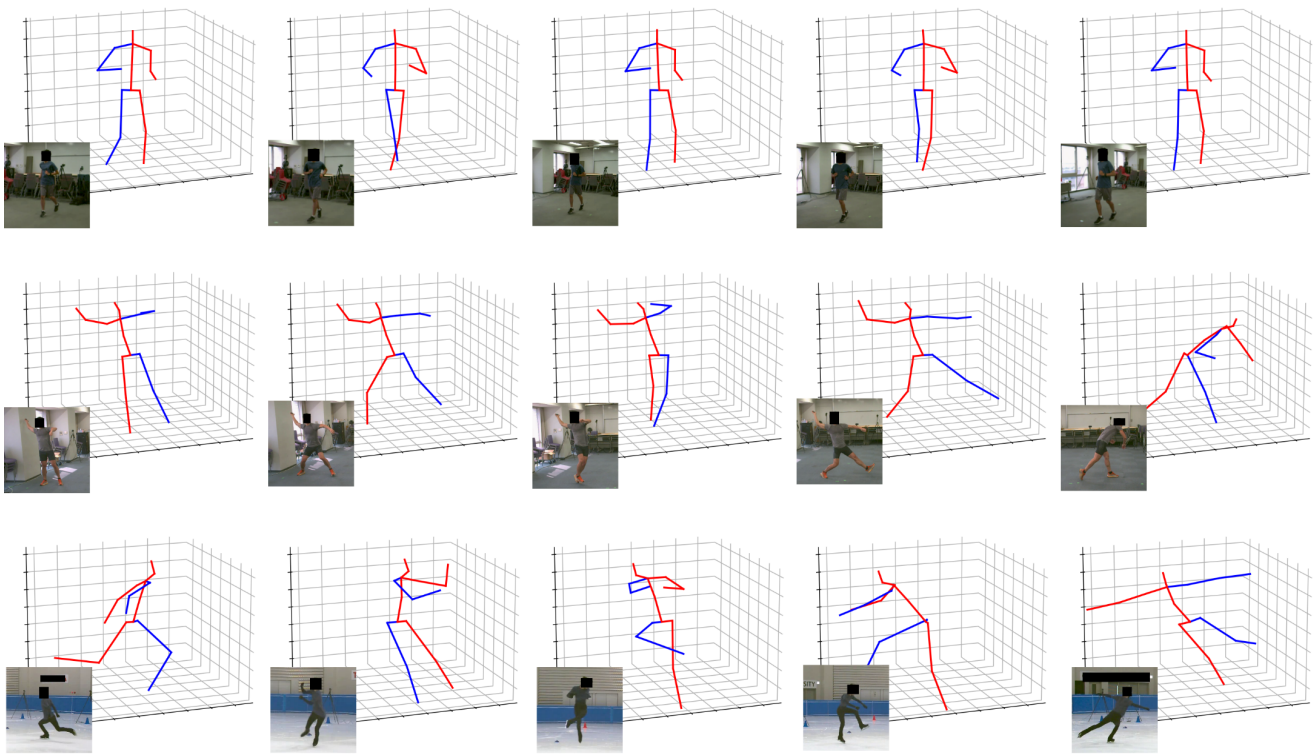


Figure 1. Examples of AthletePose3D data for running (top), track and field (middle), and figure skating (bottom). The poses are represented in centralized camera coordinates and rotated such that the motion progresses from left to right. The cropped image in the bottom left shows the posture as captured by the camera, with faces blacked out for privacy.

Abstract

Human pose estimation is a critical task in computer vision and sports biomechanics, with applications spanning sports science, rehabilitation, and biomechanical research. While significant progress has been made in monocular 3D pose estimation, current datasets often fail to capture the complex, high-acceleration movements typical of competitive sports. In this work, we introduce **AthletePose3D**, a

novel dataset designed to address this gap. AthletePose3D includes 12 types of sports motions across various disciplines, with approximately 1.3 million frames and 165 thousand individual postures, specifically capturing high-speed, high-acceleration athletic movements. We evaluate state-of-the-art (SOTA) monocular 2D and 3D pose estimation models on the dataset, revealing that models trained on conventional datasets perform poorly on athletic motions. However, fine-tuning these models on AthletePose3D no-

ably reduces the SOTA model mean per joint position error (MPJPE) from 214mm to 65mm—a reduction of over 69%. We also validate the kinematic accuracy of monocular pose estimations through waveform analysis, highlighting strong correlations in joint angle estimations but limitations in velocity estimation. Our work provides a comprehensive evaluation of monocular pose estimation models in the context of sports, contributing valuable insights for advancing monocular pose estimation techniques in high-performance sports environments. The dataset, code, and model checkpoints are available at: <https://github.com/calvinYeungck/AthletePose3D>

1. Introduction

Human pose estimation, a critical task in both computer vision [15, 37, 44, 61] and sports biomechanics [16, 18, 47, 55], has seen significant advancements in recent years. Accurate 3D pose estimation plays a crucial role in understanding and analyzing human movement across various domains, including sports science [16, 18], rehabilitation [19, 46], and biomechanical research [1]. Several 3D human pose datasets are available, such as Human3.6M [21], MPI-INF-3DHP [45], and SportsPose [20] (see Section 2.1 for a detailed overview).

Existing datasets [21, 45, 58, 68] provide valuable insights into human motion, consisting primarily of movements performed in controlled laboratory environments and daily activities. There are also sports-specific motion datasets [20, 49], which feature motions performed by amateurs. However, these datasets often fail to capture the intricate, high-acceleration movements that characterize competitive sports. Athletes engage in complex motions involving rapid changes in velocity, precise body positioning, and extraordinary biomechanical coordination—elements that are underrepresented in current datasets.

Additionally, capturing and analyzing sports movements presents significant challenges. Traditional motion capture systems, whether marker-based or markerless, are often expensive and time-consuming. While monocular pose estimation using deep learning models could offer a more effective solution, these models are typically trained and evaluated on datasets with limited movement complexity. Which could hinder their performance in the context of competitive sports. Furthermore, the kinematic estimates derived from state-of-the-art (SOTA) monocular 3D pose estimation models have not been sufficiently validated.

In response to these challenges, we introduce **AthletePose3D**, a sports motion dataset designed to bridge the gap in pose estimation data. This dataset includes 12 types of sports motions performed by athletes, comprising approximately 1.3 million frames and 165 thousand unique postures across disciplines such as running, track and field,

and figure skating. Compared to conventional benchmarking datasets, AthletePose3D features notably higher speeds and accelerations, making it uniquely suited for athletic movement analysis.

Additionally, we provide a comprehensive evaluation of SOTA monocular 2D and 3D pose estimation models on AthletePose3D, demonstrating significant performance variations before and after fine-tuning on the dataset. We further validate the kinematic estimation capabilities of SOTA models through detailed waveform analysis. By establishing these benchmarks, this work aims to lay a solid foundation for advancing monocular pose estimation techniques in high-performance sports contexts.

Our key contributions are as follows:

- **Athlete Motion Dataset:** We introduce AthletePose3D, the first large-scale dataset capturing high-speed and high-acceleration 3D human poses across 12 unique sports actions performed by athletes.
- **Comprehensive Model Evaluation:** We systematically benchmark state-of-the-art monocular 2D and 3D pose estimation models on our dataset, offering critical insights into their performance with athletic movements. Our experiments reveal that models trained exclusively on existing datasets perform poorly on athletic movements. However, after finetuning on AthletePose3D, error rates could be reduced by over 69%, underscoring the need for sports-specific training data.
- **Kinematic Validation:** We validate the estimated pose kinematics by comparing the kinematic waveforms against ground truth motion capture data, assessing the accuracy and reliability of monocular pose estimation techniques in sports contexts. Our results show strong correlations in joint angle estimations, though some limitations remain in velocity estimation.

The remainder of this paper is organized as follows. Section 2 offers an overview of related works. The methods are detailed in Section 3, followed by the details of AthletePose3D in Section 4. Section 5 presents the experimental results. Finally, Section 6 concludes the paper.

2. Related work

2.1. 3D human pose datasets

Large-scale 3D human pose datasets have been fundamental for training and evaluating monocular 3D pose estimation models. Table 1 provides a detailed summary of existing 3D human pose datasets. Significant efforts have been made to develop large-scale datasets, with Human3.6M [21] and MPI-INF-3DHP [45] being among the most well-known benchmarks for monocular 3D pose estimation. Broadly, 3D human pose datasets are collected using two mainstream methods: marker-based motion capture, which employs optical tracking systems and infrared cameras (e.g.,

Dataset	Environment	Subjects	Keypoints	Poses	Cameras	Markerless	Sync	FPS	Frames
Human3.6M [21]	lab	11	26	900K	4	×	hw	50	3.6M
MPI-INF-3DHP [45]	lab & outdoor	8	28	93K	14	✓	hw	25/50	1.3M
3DPW [70]	lab & outdoor	7	24	49K	1	×	sw	30	51K
HumanEva-I [58]	lab	6	15	78K	7	×	sw	60	280K
HumanEva-II [58]	lab	6	15	3K	4	×	hw	60	10K
TotalCapture [68]	lab	5	25	179K	8	×	hw	60	1.9M
CMU Panoptic [25]	lab	8	18	1.5M	31	✓	hw	30	46.5M
AIST++ [31]	lab	30	17	1.1M	9	✓	sw	60	10.1M
ASPset-510 [49]	outdoor	17	17	110K	3	✓	sw	50	330K
SportsPose [20]	lab & outdoor	24	17	177K	7	✓	hw	90	1.5M
AthletePose3D (ours)	lab & ice rink	8	55/86	165K	4/8/12	✓	hw	60/120	1.3M

Table 1. Comparison of the AthletePose3D dataset with other 3D pose datasets for monocular 3D pose estimation. The configuration of AthletePose3D (Keypoints, Cameras, FPS) was adapted to suit the specific environment and sports captured.

Human3.6M [21], 3DWP [70], HumanEva [58], TotalCapture [68]), and markerless motion capture, which relies on industrial-grade cameras and computer vision algorithms (e.g., MPI-INF-3DHP [45], CMU Panoptic [25], SportsPose [20], FS-Jump3D [63], WorldPose [24]).

Additionally, alternative approaches exist, including inferring 3D poses from 2D annotations (e.g., 3DSP [80]), estimating 3D poses via triangulation from multiple 2D pose detections (e.g., ASPset-510 [49]), and reconstructing 3D poses from real-world 3D human pose data (e.g., AIST++ [31]). Whilst various methods exist for capturing 3D human posture, marker-based and markerless approaches remain the gold standard and are widely used in sports science [42, 62] and biomechanics [47, 55] studies.

For 3D human posture in sports, SportsPose [20] and ASPset-510 [49] are currently the largest datasets designed to facilitate monocular 3D human posture estimation. Other datasets, such as 3DSP [80] and WorldPose [24], provide posture data from professional soccer matches for posture analysis and pose tracking, respectively, while FS-Jump3D [63] focuses on enhancing figure skating 3D pose estimation.

However, existing datasets are often not specifically designed for monocular 3D pose estimation or lack high-speed, high-acceleration movements that are essential in competitive sports. To address these limitations, we introduce AthletePose3D, a comprehensive dataset encompassing 12 types of sports actions performed by athletes ranging from amateur athletes (athletes without remuneration) to professional athletes. See Section 4 for more details.

2.2. Monocular 2D & 3D human pose estimations

In computer vision, both monocular 2D & 3D human pose estimations are a well-established research topic. 2D human pose estimation approaches are generally categorized into bottom-up and top-down approaches. Bottom-up methods [3, 6, 17] first detect keypoints in an image and then

group them to form 2D poses, whereas top-down methods [23, 60, 79] begin by detecting individuals using object detection models before estimating poses within cropped bounding boxes. Due to their effectiveness, top-down approaches are often preferred.

In 2D pose estimation models, early methods were primarily CNN-based, proposing architectures such as ResNet [74], HRNet [60], and SwinPose [39]. Later approaches introduced transformers [69] to refine feature decoding from CNN backbones, including cascade transformers [28], TransPose [77], TokenPose [35], and UniFormer [29]. Further advancements replaced CNNs entirely with transformer-based architectures, as seen in HRFormer [81] and ViTPose [75], which employ vision transformers [12] for pose estimation. More recently, the resurgence of CNN models [38, 40, 53, 76], has led to a renewed focus on CNN-based methods like RTMPose [23] and MogaNet [32].

Monocular 3D pose estimation methods can be broadly classified into direct 3D pose estimation [8, 22, 54] and 2D-to-3D lifting [44, 60, 79], where the latter involves reconstructing 3D poses from 2D keypoint coordinates. Early approaches to monocular 3D pose estimation were based on Hidden Markov Models [27, 67] and graphical models [26, 59]. With the rise of deep learning, Fully Connected Networks [11, 43] were introduced, followed by temporal-based [7, 50] and graph-based [2, 9, 71] convolutional networks, which improved the modeling of sequential dependencies and inter-keypoint relationships.

More recently, transformer-based methods have become the dominant approach, achieving state-of-the-art performance. PoseFormer [85] pioneered this direction, leading to further advancements with models such as [13, 33, 34, 52, 56, 57, 64, 64, 82, 84]. The latest models, KTPFormer [51] integrates kinematics and trajectory priors to improve motion representation, MotionAGFormer [44] introduces a hybrid approach, combining self-attention with graph convolution to better capture both spatial and temporal depen-

dencies, and TCPFormer [37] further enhances temporal modeling by leveraging pose proxies to refine sequential feature learning. Additionally, self-supervised learning has been extensively explored in monocular 3D pose estimation [14, 30, 48, 72, 78, 86].

Although many monocular pose estimation models have been proposed, they are often benchmarked on datasets featuring everyday movements. High-speed sports movements, however, present unique challenges, and the performance of these models on such movements remains unclear. Therefore, in this study, we select the best-performing monocular 2D & 3D models from corresponding benchmark datasets and fine-tune them on the AthletePose3D dataset to evaluate their performance.

2.3. Validate kinematic from pose estimation

Kinematic analysis has been widely used in various fields, including sports analysis [16, 18], rehabilitation [19, 46], and biometric recognition [1]. As a result, numerous studies [5, 19, 46, 73] have aimed to validate the use of 2D and 3D motion estimation for kinematic analysis.

In sports research, [16] compares athletic and sports motion kinematics using a marker-based motion capture system and 3D pose estimation derived from triangulated OpenPose [4] 2D pose estimation across multiple cameras. Similarly, [18] investigates single-leg squat motions by comparing the waveforms of movements captured via OpenPose [4] 2D pose estimation and 3D motion capture.

Although models like OpenPose [4], Mediapipe [41], and HRNet [60] are frequently used in sports science and kinematic validation studies, they do not represent the state-of-the-art in 2D pose estimation. Moreover, monocular 3D pose estimation models have received limited attention in validation studies. Therefore, this study evaluates both state-of-the-art monocular 2D and 3D pose estimation models for sports motion kinematic waveform using the newly introduced AthletePose3D dataset.

3. Methodology

3.1. Motion capture setup

For data capture, we employed a multi-camera system consisting of 4, 8, and 12 high-speed cameras (Miquis Video, Qualisys Inc.) for running, track and field, and figure skating, respectively. The number of cameras was selected based on the complexity of the movements in each sport, ensuring sufficient coverage to accurately track the athletes' motion from multiple angles. The camera layout was strategically designed to minimize occlusions and accommodate the natural movement patterns of athletes without imposing any restrictions on their performance.

All cameras were hardware-synchronized to ensure precise temporal alignment of captured frames, which is cru-

cial for accurate 3D motion reconstruction. The calibration process was conducted using Qualisys Track Manager (Qualisys), employing a wand-type calibration kit to establish accurate spatial relationships between the cameras. This setup achieved a calibration error of less than 1 mm, ensuring high precision in 3D motion extraction.

The frame rates were chosen based on the movement characteristics of each sport. Running, which involves rapid limb movements and high-speed motion, was recorded at 120 FPS to capture fine details. Track and field events, which typically involve a mix of explosive and sustained movements, were recorded at 60 FPS to balance temporal resolution with data storage efficiency. Similarly, figure skating, which features both dynamic jumps and continuous fluid motion, was also recorded at 60 FPS, as it provides sufficient detail while maintaining manageable data volumes. The captured videos had a primary resolution of 1920×1080 , ensuring high-quality motion capture and enabling precise monocular 3D pose estimation.

3.2. Benchmarking monocular 2D & 3D pose estimation models

To benchmark both monocular 2D and 3D pose estimation models, we considered the best-performing models on representative datasets, taking into account their architectures. These models were fine-tuned and evaluated using the AthletePose3D dataset, which was split into training, validation, and test sets with a 60/20/20 ratio. The training set was used for model training, and the validation set results were reported.

For **2D pose estimation**, we selected HRNet [60] and SwinPose [39] as representatives of conventional CNN-based methods. ViTPose [75] and UniFormer [29] were chosen to represent transformer-based approaches, while MogaNet [32] was included as a recent CNN-based resurgence. All selected models were state-of-the-art (SOTA) on either the COCO [36] or OCHuman [83] datasets.

For finetuning COCO dataset pretrained 2D pose estimation model with AthletePose3D, we projected the 3D keypoints into 2D keypoints in the COCO format, omitting facial keypoints (kpt 1 to 4) and setting their loss weights to zero during training. The fine-tuning process followed [60], with training conducted for 20 epochs. The learning rate was initially set to 1×10^{-4} , reduced to 1×10^{-5} after 10 epochs, and further decreased to 1×10^{-6} after 15 epochs. All models used an input resolution of 384×288 , except for ViTPose, where the 384×288 variant was unavailable, and thus, the 256×192 resolution was used. Only the model variants with the highest reported Average Precision (AP) were selected.

The performance is assessed using the Percent of Detected Joints (PDJ) evaluation metric [66]. A keypoint is considered detected if the normalized distance between the

Sport	Env.	Act. Types	Subs	Kpts	Cams	Poses	FPS	Frames	Comp. Level	Representative
Running	Lab	1	3	55	4	40K	120	161K	Inter-university	University
Track & Field	Lab	5	1	86	8	52K	60	416K	International	National
Figure Skating	Ice rink	6	4	86	12	73K	60	700K	national	University

Table 2. Details of the 3D pose for various sports in AthletePose3D, including subjects’ competition experience and team representation.

predicted and ground truth keypoints is below a threshold. Following [65, 80], the normalization factor is defined as the distance between the center of the shoulders and the center of the hips. A threshold of 0.2 is selected as a cutoff, as higher thresholds tend to yield overly high PDJ values. Additionally, the area under the PDJ curve (AUC) is computed to evaluate performance across varying thresholds.

For Monocular **3D pose estimation**, we select Motion-AGFormer [44] and TCPFormer [37], as most monocular 3D pose estimation models are transformer-based. The selected models are SOTA on the Human3.6M [21] and MPI-INF-3DHP [45] datasets.

For fine-tuning, the 3D keypoints follow the Human3.6M format, and preprocessing is performed according to [10, 37, 44, 87]. Specifically, the 3D poses in world coordinates are converted to camera coordinates using the extrinsic parameters. The (X, Y) coordinates are then projected onto the image plane using the intrinsic parameters. The Z coordinate in image space is obtained by scaling the Z coordinate in camera space. The scaling factor is determined by minimizing the distance between the scaled (X, Y) camera coordinates and their corresponding image coordinates.

We compare three training strategies: training on Human3.6M, training on AthletePose3D, and training on Human3.6M followed by fine-tuning on AthletePose3D. Given the dataset size, both training and fine-tuning are conducted for 60 epochs. Furthermore, considering the typical video length of sports motion sequences (see Section 4.2), we use the model variant that processes 81 frames.

Performance is assessed using the Mean Per Joint Position Error (MPJPE), which measures the per-joint distance between the estimated 3D poses and the ground truth, after aligning both poses by setting the hip joint center as the origin. Additionally, the Procrustes-aligned MPJPE (P-MPJPE) applies a rigid transformation to the predicted pose before computing the error.

3.3. Estimated pose kinematic validation

In general, monocular pose estimation provides a more efficient alternative to conventional motion capture methods. Therefore, the best-performing monocular 2D and 3D pose estimation models on the AthletePose3D validation set (see Sections 5.1 and 5.2) were selected for evaluation. Using these models, 2D and 3D poses were estimated from clips in the AthletePose3D test set. Next, limb joint velocity and

joint angles were computed, as these joints are central to kinematic analysis and tend to have the highest errors. To smooth the kinematic data, a fourth-order Butterworth low-pass digital filter with zero-phase shift at 8 Hz was applied, following [42, 62].

Since coordination differences affect absolute kinematic values, the analysis focused on the waveform of kinematic data (e.g., the temporal series of wrist velocity in a javelin throw) rather than exact values, as the waveform can reveal valuable insights. To validate the motion waveform, paired t-test statistical parametric mapping (SPM) with a significance level of $\alpha = 0.05$ was performed, and Pearson correlation was reported for a more intuitive interpretation.

4. AthletePose3D dataset

4.1. Dataset statistics

The AthletePose3D dataset aims to capture high-speed, high-acceleration, and professional sports motions, with a detailed summary provided in Table 2. The collected motions are categorized into three sports—running, track and field, and figure skating—comprising 12 unique motion types. Specifically, track and field includes shot put, glide shot put, javelin throw, discus, and spin discus, while figure skating consists of Axel, Salchow, Toe Loop, Loop, Flip, and Lutz. Running and track and field are among the most widely studied sports in sports science, whereas figure skating involves highly dynamic and complex movements, making it particularly challenging for 3D motion estimation and a frequent benchmark in recent models.

A total of eight athletes participated in data collection: three for running, one for track and field, and four for figure skating. Their expertise ranged from university-level representatives to national-level representatives, with experience spanning inter-university and international competitions. The University’s ethics committee approved this data collection, and written informed consent was obtained from all participants. Building on previous measurements of running [61] and figure skating [63], we present previously unpublished data, including camera angles and monocular 3D pose estimation required information (e.g., camera parameters and valid frames), which were collected in our earlier research. In total, approximately 1.3 million frames of data were collected, capturing around 165K individual postures, including 40K from running, 52K from track and field, and 73K from figure skating.

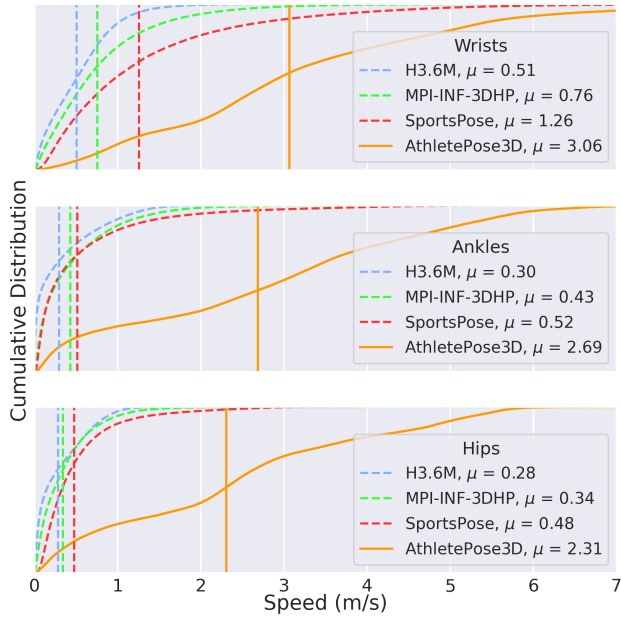


Figure 2. Comparison of the cumulative distribution function (CDF) of wrist, ankle, and hip speeds across 3D pose datasets. A lower CDF curve indicates a higher speed relative to the other curves, with the mean speed (μ) marked by a vertical line.

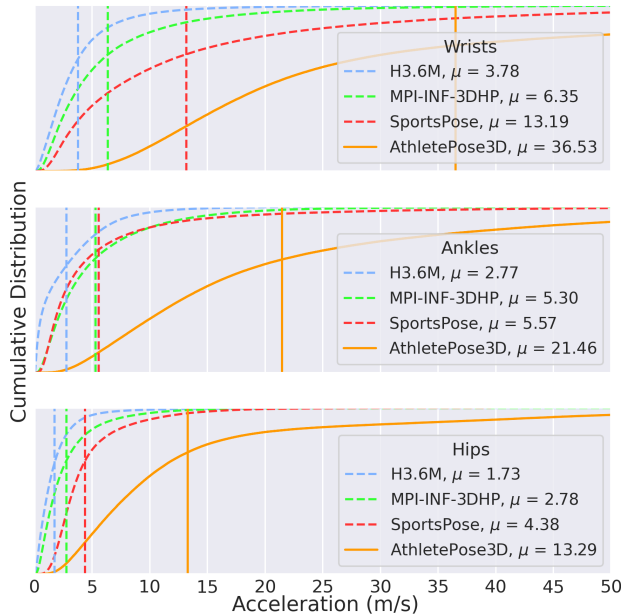


Figure 3. Comparison of the cumulative distribution function (CDF) of wrist, ankle, and hip acceleration across 3D pose datasets. A lower CDF curve indicates a higher acceleration relative to the other curves, with the mean speed (μ) marked by a vertical line.

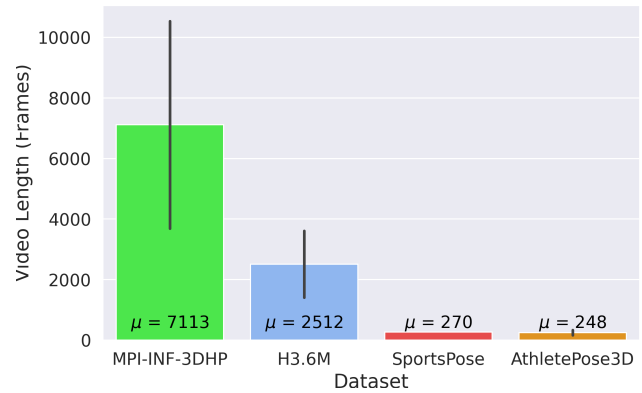


Figure 4. Video lengths of 3D pose datasets. The black bar represents the standard deviation, calculated using the Seaborn Python package.

4.2. Compare with existing dataset

When comparing AthletePose3D with existing 3D human pose datasets such as Human3.6M [21], MPI-INF-3DHP [45], and SportsPose [20], joint speed, joint acceleration, and video length are key considerations. Figures 2 and 3 show the cumulative distribution of joint speed and acceleration for the wrists, ankles, and hips, respectively. It can be observed that both the distribution and mean values in AthletePose3D are significantly higher than those in previous datasets, including SportsPose. This highlights the difference in motion speed and acceleration between everyday movements and sports motions performed by athletes in competitive-like settings, emphasizing the need for a dataset like AthletePose3D.

Another key distinction between 3D human pose datasets, often overlooked, is the video length, as shown in Figure 4. Datasets such as Human3.6M and MPI-INF-3DHP, which focus on everyday movements, typically feature videos that extend beyond a thousand frames. In contrast, datasets capturing sports motions, like SportsPose and AthletePose3D, generally consist of shorter videos, averaging around 250 frames. This reflects the dynamic nature of sports movements, which occur within seconds and are designed to capture specific actions of interest. Given this characteristic, we selected a shorter input size for the 3D pose estimation models utilized in this study, 81 frames.

5. Experiments results

5.1. 2D pose estimation models performance

The performance of various 2D pose estimation models on the AthletePose3D validation set was summarized in Table 3, ranked by their average PDJ across body parts. MogaNet [32] achieved the highest performance, with the best PDJ

Model	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	PDJ _[.1]	PDJ-AUC _[.0:.2]	AP _[.5:.95]
HRNet (CVPR'2019) [60]	88.3	77.9	86.7	90.1	93.4	96.0	88.7	74.8	94.0
SwinPose (ICCV'2021) [39]	93.9	84.2	86.3	95.0	89.4	95.5	90.7	76.2	95.2
VitPose (NeurIPS'2022) [75]	97.4	88.7	91.0	98.1	96.6	98.3	95.0	81.4	95.8
UniFormer (ICLR'2022) [29]	97.2	89.4	91.2	98.1	97.2	98.0	95.2	80.9	94.9
MogaNet (ICLR'2024) [32]	97.7	90.6	92.0	98.2	97.4	98.4	95.7	81.7	95.6

Table 3. 2D pose estimation performance on AthletePose3D validation set, with the top-performing result highlighted in bold. Ranked by PDJ in descending order. Each column for body parts indicates the average PDJ for the specified body parts, while the PDJ column denotes the mean PDJ across all 12 keypoints (keypoints 5-16 in COCO format). The subscript denotes the threshold used for the metrics.

Model	Dataset	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	MPJPE	P-MPJPE
MotionAGFormer (WACV'2024) [44]	H3.6M	223.47	197.46	235.68	62.09	302.74	535.56	237.43	91.04
	AP3D	76.95	73.82	58.49	17.00	79.06	122.64	67.23	10.00
	H3.6M+AP3D	75.90	72.47	55.60	15.47	76.67	119.35	65.55	8.37
TCPFormer (AAAI'2025) [37]	H3.6M	232.54	222.20	282.86	67.05	234.81	293.32	213.64	93.82
	AP3D	76.01	72.11	55.95	16.09	77.07	119.52	65.80	8.15
	H3.6M+AP3D	75.79	72.23	54.45	15.66	76.04	117.28	65.07	7.68

Table 4. Monocular 3D pose estimation performance on AthletePose3D validation set, with the top-performing result highlighted in bold. Each column for body parts indicates the average JPE for the specified body parts, while the MPJPE and P-MPJPE are also known as Protocols 1 and 2, respectively.

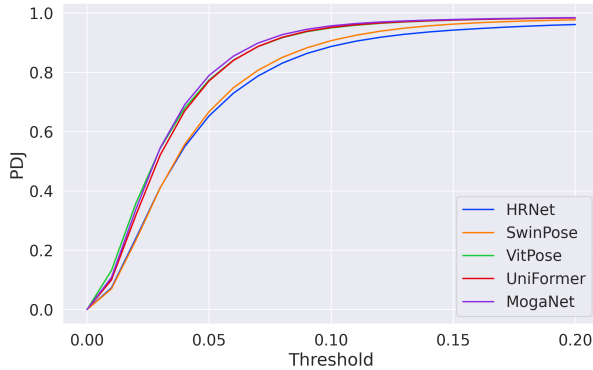


Figure 5. PDJ curve for 2D pose estimation model.

values for each joint and an overall PDJ of 95.7. UniFormer [29] followed closely, trailing slightly behind MogaNet in PDJ. ViTPose [75] and SwinPose [39] delivered competitive results, with ViTPose leading in AP. HRNet [60] lagged behind, particularly in Elbow predictions.

The PDJ curves in Figure 5 provide a more detailed comparison of model performance across different thresholds. As expected, increasing the threshold resulted in improved PDJ scores for all models, as larger thresholds allowed for greater error tolerance. MogaNet and UniFormer maintained strong performance throughout, consistently yielding the highest PDJ scores across all thresholds. ViTPose and SwinPose performed well but showed slightly lower results, with HRNet trailing behind in terms of joint position accu-

racy, especially at smaller thresholds.

The results highlighted that CNN-based models like MogaNet and transformer-based models like UniFormer performed similarly well, suggesting that both architectural approaches could achieve high accuracy in 2D pose estimation tasks. This finding indicated that the choice between convolutional and transformer-based architectures could depend on factors such as computational efficiency and generalization to unseen poses, rather than absolute performance.

5.2. Monocular 3D pose estimation models performance

The results in Table 4 demonstrated the impact of different training datasets on monocular 3D pose estimation performance when evaluated on the AthletePose3D (AP3D) validation set. Models trained solely on Human3.6M (H3.6M) performed significantly worse than those trained on AP3D or a combination of both datasets. MotionAGFormer, when trained on H3.6M, exhibited high joint position errors, particularly for the wrist (235.68 mm) and ankle (535.56 mm), resulting in an MPJPE of 237.43 mm. Similarly, TCPFormer trained on H3.6M struggled with high errors, especially in the wrist (282.86 mm), although it achieved a slightly lower MPJPE of 213.64 mm.

Training on AP3D substantially improved performance for both models, reducing MPJPE by over 70% compared to training solely on H3.6M. TCPFormer achieved an MPJPE of 65.80 mm and a P-MPJPE of 8.15 mm, outperforming MotionAGFormer (67.23 mm MPJPE, 10.00 mm P-MPJPE). The most notable improvements were observed in

	2D Pose Estimation vs GT Motion Capture		3D Pose Estimation vs GT Motion Capture	
	r	p-value	r	p-value
Velocity				
Upper limbs	0.47	0.032	0.29	0.021
Lower limbs	0.45	0.027	0.12	0.004
Joint Angle				
Upper limbs	0.75	0.021	0.89	0.013
Lower limbs	0.66	<0.001	0.81	<0.001

Table 5. Limbs kinematic validation result from monocular pose estimation models. To perform the paired t-test SPM, the kinematic temporal sequences were interpolated linearly into the same length.

lower-body joints, particularly the knee and ankle, where errors dropped from over 300 mm (H3.6M-trained models) to approximately 76 mm and 119 mm, respectively.

Utilizing both H3.6M and AP3D for training led to further refinements, with TCPFormer achieving the lowest overall errors (MPJPE: 65.07 mm, P-MPJPE: 7.68 mm). The best per-joint accuracy was also observed in this setting, with notable improvements in the wrist (54.45 mm), knee (76.04 mm), and ankle (117.28 mm). These results highlighted that training on sports-specific datasets like AP3D was crucial for achieving accurate monocular pose estimation in athletic movements, and combining diverse datasets further enhanced robustness.

5.3. Estimated pose kinematic validation

Table 5 presents the results of the paired t-test SPM analysis and Pearson correlation (r), comparing the estimated kinematic waveforms from monocular 2D and 3D pose estimation model, MogaNet [32] and TCPformer [37], to ground-truth (GT) motion capture data. If the p-value is lower than the significance level (i.e., $p < 0.05$), it suggests a significant difference between the pose estimation waveforms and the GT motion capture waveform. Pearson correlation provides an additional measure of similarity, where higher values indicate a stronger relationship between estimated and GT waveforms.

For **velocity estimation**, the 2D pose model showed moderate correlation with GT data for both the upper ($r = 0.47$, $p = 0.032$) and lower limbs ($r = 0.45$, $p = 0.027$), despite significant differences detected by the paired t-test SPM. Similarly, the 3D pose model exhibited a lower correlation for upper limb velocity ($r = 0.29$, $p = 0.021$) and a weak correlation for lower limb velocity ($r = 0.12$, $p = 0.004$), with significant waveform differences. These results suggested that while the estimated velocity waveforms captured some temporal trends, they still significantly deviated from GT motion capture.

For **joint angles**, the correlations were substantially higher, particularly for the 3D pose model. The 2D model achieved strong correlations for upper limb ($r = 0.75$, $p = 0.021$) and lower limb ($r = 0.66$, $p < 0.001$) joint angles. The 3D model further improved upon this, showing very strong correlations for upper limb ($r = 0.89$, $p = 0.013$) and lower limb ($r = 0.81$, $p < 0.001$) joint angles. However, despite these high correlations, all joint angle waveforms still exhibited statistically significant differences from GT data according to the paired t-test SPM, indicating residual discrepancies in waveform shape.

Since all paired t-test results yielded $p < 0.05$, the null hypothesis of no difference was rejected for all comparisons, meaning that both monocular 2D and 3D pose estimation models produced waveforms that significantly differed from GT motion capture. However, high correlation values, particularly for joint angles, suggested that these models still captured meaningful kinematic trends. These results highlighted the potential of monocular pose estimation as an efficient alternative to conventional motion capture, though further refinements were necessary, especially for velocity estimation.

6. Conclusion

A key challenge in human pose estimation for competitive sports is the lack of datasets capturing high-speed, high-acceleration movements, as well as the insufficient validation of kinematic estimates from SOTA models. Existing datasets often fail to represent the complexities of athletic actions, leading to unreliable monocular pose estimations.

To address this, we introduced AthletePose3D, a large-scale dataset featuring 12 sports actions performed by athletes, with over 1.3 million frames and 165 thousand individual postures. Our evaluation of SOTA monocular 2D and 3D pose estimation model, MogaNet [32] and TCPformer [37], revealed poor performance on athletic movements, but fine-tuning on AthletePose3D improved accuracy, reducing error rates by over 70%. Kinematic validation showed strong joint angle correlations, though velocity estimation remains a challenge.

AthletePose3D provides a crucial resource for advancing monocular pose estimation in sports, offering a foundation for more accurate motion analysis in sports science, biomechanics, and rehabilitation. This can lead to improved performance optimization and injury prevention. Future work will focus on refining accuracy and expanding the dataset.

7. Acknowledgement

Supported by JSTSPRING Grant Number JPMJSP2125.

References

- [1] Claudia Álvarez-Aparicio, Ángel Manuel Guerrero-Higueras, Miguel Ángel González-Santamarta, Adrián Campazas-Vega, Vicente Matellán, and Camino Fernández-Llamas. Biometric recognition through gait analysis. *Scientific Reports*, 12(1):14530, 2022. 2, 4
- [2] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2272–2281, 2019. 3
- [3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 3
- [4] 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. 4
- [5] Luca Ceriola, Ilaria Mileti, Marco Donati, and Fabrizio Patanè. Comparison of video-based algorithms for 2d human kinematics estimation: A preliminary study. In *Journal of Physics: Conference Series*, page 012002. IOP Publishing, 2023. 4
- [6] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5386–5395, 2020. 3
- [7] Yu Cheng, Bo Yang, Bo Wang, and Robby T Tan. 3d human pose estimation using spatio-temporal networks with explicit occlusion training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10631–10638, 2020. 3
- [8] Sungho Chun, Sungbum Park, and Ju Yong Chang. Learnable human mesh triangulation for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2850–2859, 2023. 3
- [9] Hai Ci, Chunyu Wang, Xiaoxuan Ma, and Yizhou Wang. Optimizing network structure for 3d human pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2262–2271, 2019. 3
- [10] Hai Ci, Xiaoxuan Ma, Chunyu Wang, and Yizhou Wang. Locally connected network for monocular 3d human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1429–1442, 2020. 5
- [11] Hai Ci, Mingdong Wu, Wentao Zhu, Xiaoxuan Ma, Hao Dong, Fangwei Zhong, and Yizhou Wang. Gfpose: Learning 3d human pose prior with gradient fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4800–4810, 2023. 3
- [12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 3
- [13] Moritz Einfalt, Katja Ludwig, and Rainer Lienhart. Uplift and upsample: Efficient 3d human pose estimation with uplifting transformers. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2903–2913, 2023. 3
- [14] Mark Endo, Kathleen L Poston, Edith V Sullivan, Li Fei-Fei, Kilian M Pohl, and Ehsan Adeli. Gaitforemer: Self-supervised pre-training of transformers via human motion forecasting for few-shot gait impairment severity estimation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 130–139. Springer, 2022. 4
- [15] Jiale Fang, Calvin Yeung, and Keisuke Fujii. Foul prediction with estimated poses from soccer broadcast video. *arXiv preprint arXiv:2402.09650*, 2024. 2
- [16] Takashi Fukushima, Patrick Blauger, Tiago Guedes Rusomanno, and Martin Lames. The potential of human pose estimation for motion capture in sports: a validation study. *Sports Engineering*, 27(1):19, 2024. 2, 4
- [17] 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
- [18] Lucas D Haberkamp, Micah C Garcia, and David M Bazett-Jones. Validity of an artificial intelligence, human pose estimation model for measuring single-leg squat kinematics. *Journal of Biomechanics*, 144:111333, 2022. 2, 4
- [19] Rebecca I Hamilton, Zornitza Glavcheva-Laleva, Md Imdadul Haque Milon, Yeshwin Anil, Jenny Williams, Peter Bishop, and Catherine Holt. Comparison of computational pose estimation models for joint angles with 3d motion capture. *Journal of Bodywork and Movement Therapies*, 40: 315–319, 2024. 2, 4
- [20] Christian Keilstrup Ingwersen, Christian Mikkelsen, Janus Nørtoft Jensen, Morten Rieger Hannemose, and Anders Bjorholm Dahl. Sportpose: A dynamic 3d sports pose dataset. In *Proceedings of the IEEE/CVF International Workshop on Computer Vision in Sports*, 2023. 2, 3, 6
- [21] 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, 2013. 2, 3, 5, 6
- [22] Karim Isakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7718–7727, 2019. 3
- [23] 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
- [24] Tianjian Jiang, Johsan Billingham, Sebastian Müksch, Juan Zarate, Nicolas Evans, Martin R Oswald, Marc Pollefeys, Otmar Hilliges, Manuel Kaufmann, and Jie Song. World-

- pose: A world cup dataset for global 3d human pose estimation. In *European Conference on Computer Vision*, pages 343–362. Springer, 2024. 3
- [25] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015. 3
- [26] Björn Krüger, Anna Vögele, Tobias Willig, Angela Yao, Reinhard Klein, and Andreas Weber. Efficient unsupervised temporal segmentation of motion data. *IEEE Transactions on Multimedia*, 19(4):797–812, 2016. 3
- [27] Andreas M Lehmann, Peter V Gehler, and Sebastian Nowozin. Efficient nonlinear markov models for human motion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1314–1321, 2014. 3
- [28] Ke Li, Shijie Wang, Xiang Zhang, Yifan Xu, Weijian Xu, and Zhuowen Tu. Pose recognition with cascade transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1944–1953, 2021. 3
- [29] Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatiotemporal representation learning. In *International Conference on Learning Representations*, 2022. 3, 4, 7
- [30] Linguo Li, Minsi Wang, Bingbing Ni, Hang Wang, Jiancheng Yang, and Wenjun Zhang. 3d human action representation learning via cross-view consistency pursuit. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4741–4750, 2021. 4
- [31] 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. 3
- [32] Siyuan Li, Zedong Wang, Zicheng Liu, Cheng Tan, Haitao Lin, Di Wu, Zhiyuan Chen, Jiangbin Zheng, and Stan Z. Li. Moganet: Multi-order gated aggregation network. In *International Conference on Learning Representations*, 2024. 3, 4, 6, 7, 8
- [33] Wenhao Li, Hong Liu, Runwei Ding, Mengyuan Liu, Pichao Wang, and Wenming Yang. Exploiting temporal contexts with strided transformer for 3d human pose estimation. *IEEE Transactions on Multimedia*, 25:1282–1293, 2022. 3
- [34] 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*, pages 13147–13156, 2022. 3
- [35] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 11313–11322, 2021. 3
- [36] 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: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 4
- [37] Jiajie Liu, Mengyuan Liu, Hong Liu, and Wenhao Li. Tcpformer: Learning temporal correlation with implicit pose proxy for 3d human pose estimation. *arXiv preprint arXiv:2501.01770*, 2025. 2, 4, 5, 7, 8
- [38] Shiwei Liu, Tianlong Chen, Xiaohan Chen, Xuxi Chen, Qiao Xiao, Boqian Wu, Tommi Kärkkäinen, Mykola Pechenizkiy, Decebal Mocanu, and Zhangyang Wang. More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity. *arXiv preprint arXiv:2207.03620*, 2022. 3
- [39] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3, 4, 7
- [40] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 3
- [41] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Ubaweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 4
- [42] Mizuki Makino and Kenji Tauchi. Kinematic factors related to forward and vertical release velocity in male javelin throwers. *International Journal of Sport and Health Science*, 20: 249–259, 2022. 3, 5
- [43] 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
- [44] Soroush Mehraban, Vida Adeli, and Babak Taati. Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 6920–6930, 2024. 2, 3, 5, 7
- [45] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017. 2, 3, 5, 6
- [46] Dimitrios Menychtas, Nikolaos Petrou, Ioannis Katsizoglou, Erasmia Giannakou, Athanasios Grekidis, Antonios Gasteratos, Vassilios Gourgoulis, Eleni Douda, Ilias Smilios, Maria Michalopoulou, et al. Gait analysis comparison between manual marking, 2d pose estimation algorithms, and 3d marker-based system. *Frontiers in Rehabilitation Sciences*, 4:1238134, 2023. 2, 4
- [47] Elzbieta Mirek, Monika Rudzińska, and Andrzej Szczudlik. The assessment of gait disorders in patients with parkinson’s disease using the three-dimensional motion analysis system vicon. *Neurologia i neurochirurgia polska*, 41(2):128–133, 2007. 2, 3

- [48] Mahdiar Nekoui and Li Cheng. Enhancing human motion assessment by self-supervised representation learning. In *BMVC*, page 322, 2021. 4
- [49] Aiden Nibali, Joshua Millward, Zhen He, and Stuart Morgan. ASPset: An outdoor sports pose video dataset with 3D keypoint annotations. *Image and Vision Computing*, page 104196, 2021. 2, 3
- [50] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7753–7762, 2019. 3
- [51] Jihua Peng, Yanghong Zhou, and PY Mok. Ktpformer: Kinematics and trajectory prior knowledge-enhanced transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1123–1132, 2024. 3
- [52] Xiaoye Qian, Youbao Tang, Ning Zhang, Mei Han, Jing Xiao, Ming-Chun Huang, and Ruei-Sung Lin. Hstformer: Hierarchical spatial-temporal transformers for 3d human pose estimation. *arXiv preprint arXiv:2301.07322*, 2023. 3
- [53] Yongming Rao, Wenliang Zhao, Yansong Tang, Jie Zhou, Ser Nam Lim, and Jiwen Lu. Hornet: Efficient high-order spatial interactions with recursive gated convolutions. *Advances in Neural Information Processing Systems*, 35: 10353–10366, 2022. 3
- [54] N Dinesh Reddy, Laurent Guigues, Leonid Pishchulin, Jayan Eledath, and Srinivasa G Narasimhan. Tesseract: End-to-end learnable multi-person articulated 3d pose tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15190–15200, 2021. 3
- [55] Øyvind Sandbakk, Gertjan Ettema, and Hans-Christer Holmberg. The influence of incline and speed on work rate, gross efficiency and kinematics of roller ski skating. *European journal of applied physiology*, 112:2829–2838, 2012. 2, 3
- [56] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. In *European Conference on Computer Vision*, pages 461–478. Springer, 2022. 3
- [57] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao. Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14761–14771, 2023. 3
- [58] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1):4–27, 2010. 2, 3
- [59] Yang Song, Luis Goncalves, and Pietro Perona. Unsupervised learning of human motion models. *Advances in Neural Information Processing Systems*, 14, 2001. 3
- [60] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 3, 4, 7
- [61] Tomohiro Suzuki, Ryota Tanaka, Kazuya Takeda, and Keisuke Fujii. Pseudo-label based unsupervised fine-tuning of a monocular 3d pose estimation model for sports motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3315–3324, 2024. 2, 5
- [62] Hiroko Takigawa and Kenji Tauchi. Factors in javelin throw that result in differences in throwing records between throwers with similar approach velocities. *International Journal of Sport and Health Science*, 21:153–159, 2023. 3, 5
- [63] Ryota Tanaka, Tomohiro Suzuki, and Keisuke Fujii. 3d pose-based temporal action segmentation for figure skating: A fine-grained and jump procedure-aware annotation approach. In *Proceedings of the 7th ACM International Workshop on Multimedia Content Analysis in Sports*, pages 17–26, 2024. 3, 5
- [64] Zhenhua Tang, Zhaofan Qiu, Yanbin Hao, Richang Hong, and Ting Yao. 3d human pose estimation with spatio-temporal criss-cross attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4790–4799, 2023. 3
- [65] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. *Advances in neural information processing systems*, 27, 2014. 5
- [66] Alexander Toshev and Christian Szegedy. DeepPose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014. 4
- [67] Dorra Trabelsi, Samer Mohammed, Faicel Chamroukhi, Latifa Oukhellou, and Yacine Amirat. An unsupervised approach for automatic activity recognition based on hidden markov model regression. *IEEE Transactions on automation science and engineering*, 10(3):829–835, 2013. 3
- [68] Matt Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *2017 British Machine Vision Conference (BMVC)*, 2017. 2, 3
- [69] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3
- [70] 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 *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018. 3
- [71] Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. Motion guided 3d pose estimation from videos. In *European conference on computer vision*, pages 764–780. Springer, 2020. 3
- [72] Peng Wang, Jun Wen, Chenyang Si, Yuntao Qian, and Liang Wang. Contrast-reconstruction representation learning for self-supervised skeleton-based action recognition. *IEEE Transactions on Image Processing*, 31:6224–6238, 2022. 4

- [73] Edward P Washabaugh, Thanikai Adhithiyan Shanmugam, Rajiv Ranganathan, and Chandramouli Krishnan. Comparing the accuracy of open-source pose estimation methods for measuring gait kinematics. *Gait & posture*, 97:188–195, 2022. 4
- [74] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018. 3
- [75] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. 3, 4, 7
- [76] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks. *Advances in Neural Information Processing Systems*, 35:4203–4217, 2022. 3
- [77] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Keypoint localization via transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11802–11812, 2021. 3
- [78] Zhuoqian Yang, Wentao Zhu, Wayne Wu, Chen Qian, Qiang Zhou, Bolei Zhou, and Chen Change Loy. Transmomo: Invariance-driven unsupervised video motion retargeting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5306–5315, 2020. 4
- [79] 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
- [80] Calvin Yeung, Kenjiro Ide, and Keisuke Fujii. Autosoccerpose: Automated 3d posture analysis of soccer shot movements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3214–3224, 2024. 3, 5
- [81] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution transformer for dense prediction. *arXiv preprint arXiv:2110.09408*, 2021. 3
- [82] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Jun-song Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13232–13242, 2022. 3
- [83] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2seg: Detection free human instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 889–898, 2019. 4
- [84] 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*, pages 8877–8886, 2023. 3
- [85] 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
- [86] Wentao Zhu, Zhuoqian Yang, Ziang Di, Wayne Wu, Yizhou Wang, and Chen Change Loy. Mocanet: Motion retargeting in-the-wild via canonicalization networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3617–3625, 2022. 4
- [87] 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. 5