A Comprehensive Review of Real-Time Multi-View Multi-Person Markerless Motion Capture

Pierre Nagorny[1], **Bart Kevelham**[1], **Sylvain Chagué**[1], **and Caecilia Charbonnier**[1]

[1]Artanim Foundation, Meyrin, Switzerland

August 27, 2025

Abstract

Markerless human body motion capture promises to remove markers from capture studios, thus simplifying its diverse application fields, from life science to virtual reality. This comprehensive review examines recent advances in real-time markerless motion capture systems from 2020 to 2024, focusing on real-time multi-view, multi-person tracking solutions. Recent advancements, particularly driven by neural network-based pose estimation, have enabled real-time tracking with minimal latency, achieving at least 25 frames per second. Through systematic analysis, we evaluate these methods based on three key metrics: accuracy in pose reconstruction, end-to-end latency, and computational efficiency. Special attention is given to how architectural decisions impact system scalability regarding the number of camera viewpoints and tracked individuals. While current methods show promise for applications like sports analysis and virtual reality, challenges remain in achieving optimal performance across all metrics. Through systematic analysis of leading real-time pipelines, we identify key technical advances and persistent challenges. This synthesis provides critical insights for researchers and practitioners working to develop more robust markerless motion capture systems, while outlining important directions for future research.

Keywords survey, motion capture, markerless, real-time, multi-view, multi-person, human pose, neural networks

1 Introduction

Recent advances in markerless motion capture technology have enabled real-time tracking of multiple people using only calibrated cameras. This capability is transforming applications ranging from virtual reality and sports analysis to healthcare and robotics, where capturing group interactions is essential.

While traditional marker-based systems require complex setups with retroreflective markers and controlled lighting conditions, markerless approaches promise simpler deployment using standard cameras. The integration of modern Deep Learning techniques, particularly Convolutional Neural Networks (CNNs) running on GPUs, has made markerless systems robust enough for real-world scenarios.

Our review focuses specifically on multi-view markerless methods that can: (1) track multiple people simultaneously, (2) use multiple cameras as input, (3) operate in real-time, at least 25fps (processing latency under 40ms), (4) handle occlusions caused by interactions between person, and (5) track a full set of body keypoints (at least 10 per person).

These requirements are like traditional motion-capture setups, and present specific technical challenges. Multi-person tracking introduces occlusions and identity association problems. Real-time operation constrains the available processing budget. Tracking numerous keypoints per person creates a computationally complex multi-view matching problem that is NP-hard [183].

The key contributions of this review are: (1) a comprehensive analysis of markerless motion capture evolution, from early multi-view approaches to current real-time methods, (2) a systematic evaluation of state-of-the-art real-time

Corresponding author: *pierre.nagorny@artanim.ch* DOI:10.1145/3757733

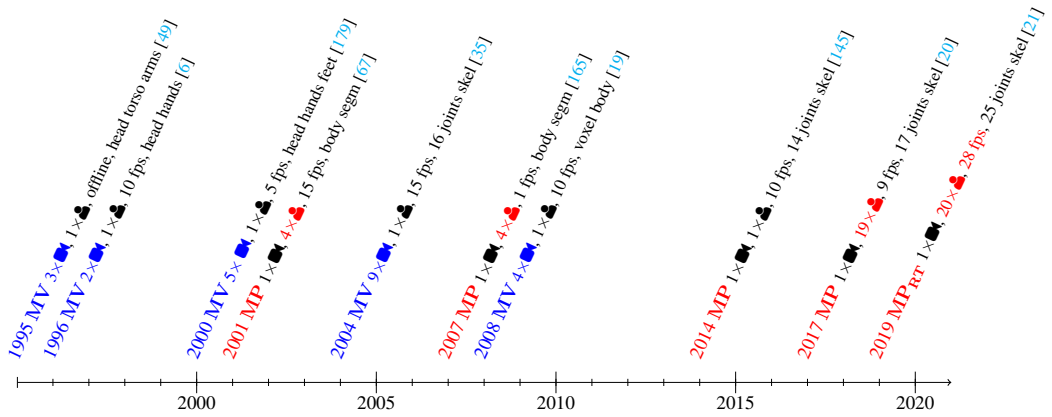


Figure 1: Timeline of the early real-time markerless methods. **MV** denotes multi-view methods, **MP** denotes multi-person methods, with the **RT** subscript for real-time multi-view $\mathbf{MV}_{\mathbf{RT}}$ or real-time multi-person $\mathbf{PV}_{\mathbf{RT}}$ methods, 📹 indicates the number of camera views, 👤 indicates the number of detected persons.

multi-view multi-person systems, including detailed accuracy and latency benchmarks, and (3) an in-depth discussion of current limitations and promising future research directions.

The review is organized as follows: Section §2 presents a historical perspective of the field, covering both early methods (§2.1) and recent deep learning approaches (§2.2), along with previous surveys (§2.3). Section §3 defines the markerless pose estimation problem, evaluation metrics, and benchmark datasets. Section §4 analyzes current leading approaches, their architectures, and limitations. Through this structure, we provide researchers and practitioners with a thorough understanding of this rapidly evolving field.

2 Comprehensive Survey of Literature

This section presents a structured analysis of markerless motion capture, a field that has generated thousands of research articles and numerous reviews. We organize our discussion chronologically and thematically into three key periods: §2.1 *Early Foundations 1985-2012*: we examine pioneering approaches that established core principles and algorithms, focusing on multi-view reconstruction techniques, early real-time and multi-person methods, §2.2 *Modern Real-time Methods 2013-Present*: we analyze the transformation brought by deep learning, highlighting CNN-based pose estimation architectures, real-time optimization and multi-person tracking innovations, §2.3*Survey Analysis and Synthesis*: we provide a systematic comparison of previous review works that present the evolution of key technical approaches and identify remaining challenges.

Throughout this section, we maintain focus on systems meeting our core requirements: real-time processing (>25 fps), multi-view capture, and multi-person tracking capabilities. This structured chronological and thematic organization aims to provide insight into both the historical evolution and the key technical innovations that enabled today's real-time markerless systems.

2.1 Early foundations (1985-2012)

Markerless motion capture emerged in 1985 with Lee and Chen [82], followed by multi-view approaches in 1995 [49] that achieved 10fps tracking rates through advances in cameras and algorithms. While early works claimed "real-time" performance, we distinguish between "fast" (10-24fps) and truly "real-time" (25fps) methods. The field progressed rapidly through improved computing power, GPU acceleration, optimized algorithms, and large datasets. Figure 1 presents a timeline of landmark early methods, selected based on their pioneering technical contributions and lasting impact on the field. Each entry represents the first publication to achieve a specific performance milestone in terms of speed, accuracy, or multi-person tracking capability.

2.1.1 Early real-time multi-view markerless

Multi-view systems enabled accurate 3D position estimation through geometric triangulation, avoiding single-view depth ambiguities. Gavrila and Davis [49] pioneered real-time 3D limb detection with 3 synchronized cameras by matching geometric primitives (head, torso, arms) to multi-view contours, establishing a paradigm for future multi-view

methods despite computational demands. In 1996, Azarbayejani and Pentland [6] achieved 20-30fps head and hand tracking using stereo cameras and 2D Gaussian blob modeling. Running on dual Silicon Graphics workstations, it achieved 1.5cm Mean Position Error §3.2.2 and included self-calibration. Their *Pfinder* extension [164] enabled 10fps monocular tracking of head, hands and feet. Early multi-camera systems focused on improving framerates with more cameras. In 2000, Yonemoto et al. [179] achieved 5fps with 5 cameras by detecting head, hands and feet as bounding boxes and fitting a 13-joint skeleton. The system used color blob tracking and epipolar triangulation, distributed across a 12-machine cluster for parallel processing. Though, the 200ms latency limited real-world applications. By 2004, Date et al. [35] reached 15fps with nine 640x480 cameras. Their method detected key body parts and fit a more complex 16-joint skeleton using inverse kinematics, but the background subtraction requirement restricted use to controlled environments.

In 2008, Caillette et al. [19] pioneered volumetric voxel representation for multi-view tracking. Their method reconstructed body visual hulls with gaussian blob tracking for temporal consistency, and avoiding background subtraction. Using four 320x240 cameras, it achieved 10fps on a single CPU - a milestone for volumetric reconstruction.

2.1.2 Early real-time multi-person markerless - pictorial structures

Multi-person tracking emerged in 2001 with Isard and MacCormick [67]'s particle filters for segmentation and cylindrical body models running at 15fps (160x120), identifying key challenges: occlusion handling §3.1.3, identity preservation, and re-identification after disappearance. In 2003, pictorial structures [113] detected limbs via parallel contrast lines. Ramanan et al. [114] improved this with Canny edge detection, multiclass pictorial classifiers and limb patch learning. The system clustered and matched limbs iteratively to identify distinct people, resulting in linear computational scaling with the number of people. In 2007, Wu and Nevatia [165] introduced bottom-up part grouping §3.1.1 using IoU (1fps), later extended to 3D [16] §2.2.2. The *Kinect* [129] achieved 200fps tracking (5ms latency) on Xbox 360 GPU using depth sensing.

2.2 Modern Deep Learning-based real-time methods (2013-present)

Deep Learning revolutionized pose estimation with *DeepPose* [145] in 2013. It introduced direct joint coordinate regression using convolutional neural networks and demonstrated successful transfer learning from *AlexNet* [78]. The network architecture used seven layers of convolutions and rectified linear units (ReLU), followed by a cascade of refinement stages that iteratively improved joint predictions. Despite not being real-time (100ms on 12 CPU cores), its superior accuracy and speed established Deep Learning as the new paradigm for pose estimation.

2.2.1 Real-time multi-person markerless methods

Multi-person pose estimation evolved into two approaches: top-down (detect persons then estimate poses) and bottom-up (detect keypoints then group into poses), each with different trade-offs between accuracy, speed, and scalability. **Top-down** developments include *Mask R-CNN* [59] extending *Faster R-CNN* [118] with keypoint prediction, and *RMPE* [44] with Symmetric Spatial Transformer Network. *AlphaPose* [45] added hierarchical feature pyramids and pose-guided proposals, reaching 25fps (2080Ti) but scaling linearly with persons. For **bottom-up**, *Associative Embedding* [103] pioneered single-stage keypoint detection with learned grouping embeddings, achieving 6fps (V100) [77] using a stacked hourglass network with skip connections to preserve both high-resolution spatial details and semantic information.

OpenPose [20] pioneered real-time multi-person pose estimation using a two-branch architecture with keypoint heatmaps and *Part Affinity Fields* (PAFs) for efficient skeleton assembly. Using *VGG-19* [134] backbone, it achieved 8.8fps on GTX-1080 Ti when trained on *COCO* [87] and *CMU Panoptic* [72]. The improved version [21] reached 28fps through optimizations, while its temporal extension *STAF* [112] maintained 27fps with temporal tracking. *PifPaf* [76] introduced *Part Intensity Fields* and PAFs with a *ResNet50* [58] backbone, achieving 14fps on V100 GPU after temporal tracking [77]. The same year, *FastPose* [181] unified detection, pose estimation and re-identification using *ResNet18* [58] with parallel heads, achieving 29.4fps (Titan X) through shared computation.

ROMP [140] introduced single-stage direct regression of *SMPL* [92] §3.1.2 parameters using *HRNet* [139, 30] and transformer-based regression. Its parallelized approach avoided iterative optimization, achieving 30fps on 1070Ti GPU with 60.5mm MPJPE on *3DPW* §3.2.1 and constant inference time regardless of person count.

Finally, single-stage methods based on *YOLO* [116] proposed significant speed improvements by treating keypoint detection as a direct regression problem. Unlike two-stage detectors that first propose regions then classify them, *YOLO* divides the image into a grid and directly predicts bounding boxes and keypoint coordinates in a single forward pass. *YOLOv7* [155] combined an efficient backbone with PANet-based multi-scale fusion and parallel keypoint prediction branches, achieving 56fps on an RTX3090 GPU.

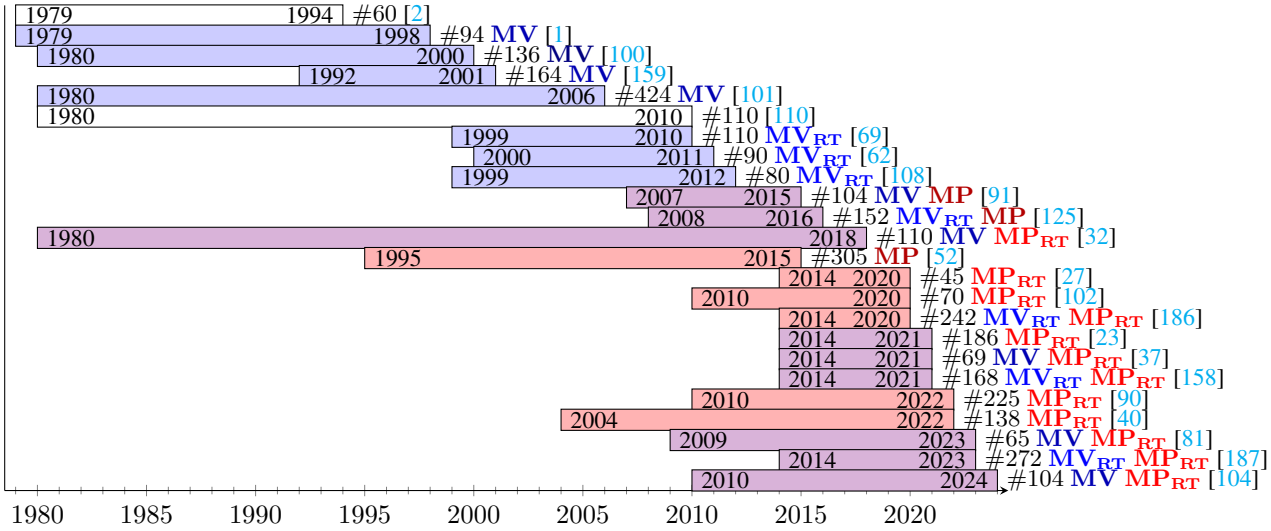


Figure 2: Timeline of previous markerless surveys.
#nb indicates the number of methods in the review, **MV** denotes survey reporting multi-view methods, **MP** denotes survey reporting multi-person methods, with the $_{RT}$ subscript for real-time multi-view **MV**$_{RT}$ or real-time multi-person **PV**$_{RT}$ methods.

2.2.2 Real-time monocular 2D to 3D pose estimation

VNect [95] pioneered real-time single-person 3D pose estimation with a two-branch CNN (*ResNet50* backbone) for 2D heatmaps and 3D regression, achieving 30fps on GTX 1080 through kinematic fitting. Mehta et al. [96] extended this to multi-person with unified heatmaps, PAFs and 3D regression. *XNect* [97] later achieved 30fps multi-person tracking through lightweight detection, efficient 3D regression, and temporal fitting, with person-count invariant performance.

The 3D pictorial structures approach [16, 10] modeled bodies as probabilistic graphs with Conditional Random Fields (CRF) constraints. Despite temporal extensions [11], its quadratic complexity limited real-time applications to 1fps for 3 persons.

Mvpose [39] combined *Cascaded Pyramid Network* [26] with epipolar geometry and re-ID, achieving 10% higher PCP3D on *Campus*/*Shelf* at 10fps (4 persons, 5 views) on 1080Ti.

Recent advances include Zhou et al. [189]'s two-stage framework using pose-guided transformers and adaptive feature selection, achieving 28.6mm MPJPE on *Human3.6M* with strong generalization (68.9mm on *MPI-INF-3DHP*). Cai et al. [17]'s diffusion method decomposed poses into bone length/direction, achieving 39.0mm MPJPE on *Human3.6M*, outperforming prior methods [156, 51] by 10.0% and 1.3% respectively, though computational costs limit real-time multi-person applications.

2.2.3 Pose Estimation from IMUs

While this review focuses on camera-based markerless capture, IMUs provide direct limb orientation measurement through acceleration and angular velocity sensors. von Marcard et al. [150]'s *SIP* achieved 4cm error using 6-15 IMUs with *SMPL* model optimization, though not real-time at 450ms/frame. Huang et al. [65]'s *Deep Inertial Poser* used 6 IMUs with *BIRNN* and *SMPL* to achieve 6cm MPJPE at 29fps, but required 25 past and 5 future frames, exceeding real-time latency criteria. Hybrid IMU-video approaches emerged to combine complementary strengths. von Marcard et al. [151]'s *Video Inertial Poser* fused data by optimizing *SMPL* pose, achieving 26.3mm MPJPE on *3DPW* with 13 IMUs + video, or 39.6mm with 6 IMUs + video. The authors introduced the *3DPW* dataset for multi-person IMU-video evaluation.

Zhang et al. [185] proposed real-time fusion through geometric optimization, achieving 24.6mm MPJPE on *Total Capture* at 150ms/frame. While IMU-camera fusion handles occlusions well, challenges remain with sensor drift, magnetic interference, calibration complexity, and the impracticality of requiring multiple IMUs per person.

2.3 Survey analysis and synthesis

Figure 2 provides a chronological overview of major markerless motion capture surveys, highlighting key developments in the field. While these reviews documented important methodological advances, most did not emphasize latency considerations - a gap we address by focusing on real-time applications.

2.3.1 1994 to 2012: pre-Deep Learning markerless era

Early reviews by Aggarwal et al. [2], Aggarwal and Cai [1] documented the transition to markerless approaches using color histograms and optical flow. Key innovations included the first 14-joint skeleton model [29] and real-time stereo tracking [6] §2.1.1. Cédras and Shah [22] and Aggarwal et al. [3] established core technical challenges. Wang et al. [159] analyzed 164 papers, defining key pipeline tasks: detection, pose estimation, and action recognition. Moeslund et al. [101] added *Initialization*, feature-based *Pose estimation*, *Tracking*, and *Recognition*, documenting the first neural network application [121].

Poppe [110] introduced *top-down* and *bottom-up* paradigms (§3.1.1). Ji and Liu [69] focused on view-invariant representations, noting one fast multi-view method [19] §2.1.1.

In 2012, Holte et al. [62] compared body representations and evaluated 18 methods on *INRIA IXMAS* [162]. Notable real-time achievements included Date et al. [35] §2.1.1 at 15fps, Dahmane and Meunier [34] at 16fps for detection, and Caillette et al. [19] §2.1.1 at 10fps with multi-view tracking. These optimizations laid the groundwork for modern real-time systems.

2.3.2 2012 to 2018: Deep Learning advances

Deep learning emerged as the dominant paradigm during 2012-2018. Liu et al. [91] first systematically compared traditional and deep learning approaches on *LSP* [71] and *FLIC* [124] datasets. Early CNN-based methods [25, 143] demonstrated 5-10% accuracy gains over hand-crafted features, effectively establishing deep learning as the new state-of-the-art. Sarafianos et al. [125] analyzed 152 methods and highlighted 3 key datasets: *Human3.6M* [66] §3.2.1, *CMU Panoptic* [72] §3.2.1, and *SynPose300* §3.2.1. The review introduced standardized evaluation using Mean Per Joint Position Error (MPJPE) §3.2.2 to compare four markerless methods [174, 190]. The 3D pictorial structures representation [16] §2.2.2 emerged as foundational, while *Kinect* [129] §2.1.2 demonstrated commercial viability.

By 2018, Colyer et al. [32] evaluated biomechanics methods using *HumanEva* dataset [132]. Two key innovations emerged: *OpenPose* [20] §2.2.1 for real-time multi-person detection and *SMPL* [14] §3.1.2 parametric body model. Comparing 8 methods (2010-2016) using MPJPE on *HumanEva*, the authors found that markerless systems had not yet achieved the precision needed for sports and rehabilitation applications, noting challenges in marker-based validation due to body shape variations and marker interference.

2.3.3 2018 to 2023: Modern multi-view 3D pose estimation reviews

Our review focuses on real-time multi-view multi-person methods as alternatives to marker-based systems. Despite less research activity in multi-view compared to monocular approaches [52, 27, 23, 90, 40], the trend shows that achieving marker-based quality with low latency might be possible with further research.

Zheng et al. [186]'s analysis of 242 papers highlighted *EpipolarPose* [74] §4.4 with epipolar constraints, *VoxelPose* [147] §4.5.1's volumetric learning, real-time methods [117, 24] §4.2, and adversarial vulnerabilities [88, 127] §6.1.

Desmarais et al. [37] and documented MPJPE improvements from 100mm to under 20mm in a decade, highlighting *Epipolar Transformers* [60] §4.5.3 and temporal tracking [63, 154, 89]. The same year, Wang et al. [158] analyzed open-source implementations, highlighting how code sharing has accelerated progress since 2016 §6.1. Their evaluation identified *Mvpose* [39] as achieving an optimal accuracy-speed trade-off at 10fps.

Lam et al. [81] proposed a screened literature search of 65 papers on clinical applications. While Kinect [129] §2.1.2 dominated medical use, smartphone-based setups gained traction due to simplified setup, hygiene benefits from markerless tracking, and lower costs. Markerless methods are simpler to use, as they do not need the time-consuming marker placement procedure, and therefore avoid all the hygiene constraints associated with contact with the patient's body.

Recent advances are highlighted in Zheng et al. [187]'s review of 30 novel methods, covering multi-view fusion (*MvP* [160] §4.5.3), temporal modeling (*4D Association Graph* [183]), and volumetric processing (*Faster-VoxelPose* [176]). [104]'s evaluation of 9 methods on *CMU Panoptic* [72] §3.2.1 showed *Tessetrack* [115] §4.5.1 achieving best-in-class performance (MPJPE 7.3mm, AP 99.1% §3.2.2, §3.2.2). While thorough on accuracy metrics, this survey did not focus on computational efficiency and real-time performance - key aspects we address.

3 Problem statement

Real-time markerless pose estimation systems are characterized by their ability to track multiple subjects across multiple views without physical markers. This section defines the core problem and scope of the research within this domain.

3.1 Definition of a real-time markerless pose estimation method

From multi-view camera inputs, the task is to detect and track persons while predicting their full-body pose as joint coordinates or 3D meshes. The method should achieve an MPJPE under 10mm with latency below 40ms (25fps). While less precise than marker-based methods (250fps, MPJPE < 2mm §3.2.2 [98]), these specifications represent an achievable target that balances the inherent trade-off between accuracy and computational latency.

3.1.1 *Top-down* vs *bottom-up* approaches

Multi-person pose estimation follows two paradigms introduced by Poppe [110] and refined by Chen et al. [27]: *top-down* methods detect person boxes then estimate individual poses, achieving higher precision through cropped regions but struggling with occlusions; *bottom-up* methods detect and group keypoints directly, providing better robustness during close interactions by leveraging full scene context §2.2.1.

3.1.2 Human body representation and model

In 1997, Aggarwal and Cai [1] introduced the different human body representations that a markerless method can output: a simplified joint-stick skeleton, the 2D contour of the human body, or a simplified volumetric limb-cylinders body skeleton. Since then, the body representation has evolved into three main categories: keypoint-based, part-based volumetric, and parametric models.

Keypoint-based representations This approach represents the body as 3D keypoint coordinates connected to form an anatomical skeleton. It enables efficient storage, processing and 3D triangulation from multi-view 2D coordinates. For multi-person scenes, methods like *Associative Embedding* [103] §2.2.1 and *Part Affinity Fields* [21] §2.2.1 handle keypoint association. The keypoint definitions come from training datasets. For exemple, Hidalgo et al. [61] entend *COCO* §3.2.1 to include feet keypoints. Synthetic data also enables flexible keypoint definitions, detailed in Section §3.2.1.

3D parametric human models Statistical models represent body shape and pose variation. The widely-used *SMPL* model [92, 14] uses a deformable mesh controlled by shape (β, 10D), pose (θ, 72D for 24 joints), and Linear Blend Skinning, trained on 3000 *CAESAR* scans [119]. Extensions include hand modeling (*SMPL+H*/*MANO* [120]), face and hands (*SMPL-X* [107]), and *STAR* [105] which reduces parameters by 75% while maintaining quality through sparse decomposition. *VIBE* [75] achieved 25fps for 5 persons (RTX2080Ti), while *ROMP* [140] reached 30fps for 15 persons (GTX1070Ti). *GHUM* [170], trained on 60,000+ scans, uses variational autoencoders with 17x fewer parameters than *SMPL* while maintaining quality. However, these models struggle with extreme poses and multi-person computational costs. While offline methods use *SMPL* for multi-view multi-person tracking [85, 180], real-time systems like *spatiotemporal 4D association graph* [183] use it only for refinement, balancing detailed modeling with performance constraints.

3.1.3 Constraints of a real-time markerless setup

Real-time systems must process frames within 40ms (25fps), including per-view pose estimation and cross-view reconstruction. We discuss the main issues that impact the development of real-time methods.

Acquisition issues Several acquisition issues can limit the accuracy of the pose estimation: image noise, motion blur, and various occlusions (object, self, multi-person). These can cause missed detections, false positives, or incorrect limb associations, impacting the accuracy of 2D pose estimation and subsequent cross-view reconstruction. The markerless method will need to robustly overcome these hurdles to output accurate per-view 2D poses or 2D features, so that the cross-view solving might succeed.

Occlusion handling Occlusion is a common challenge in multi-person motion capture that modern methods address through several strategies:

- *Multi-view redundancy*: *VoxelPose* [147] §4.5.1 reconstructs occluded parts through volumetric feature fusion across multiple views.

- *Part association*: *OpenPose* [21] §2.2.1 uses *Part Affinity Fields* and *Associative Embedding* [103] §2.2.1 learns persistent pose embeddings.
- *Temporal modeling*: *4D Association Graph* [183] §4.3 tracks poses across frames.
- *Learning-based completion*: *TEMPO* [31] §4.5.1 predicts occluded joints using temporal priors learned from *CMU Panoptic* [72] §3.2.1.

Multi-view redundancy provides reliable handling but increases computation, whereas learning methods work with fewer views but face accuracy and domain adaptation challenges §3.1.3 for novel environments and motions.

Latency constraints Real-time markerless systems employ several strategies to minimize latency: *Lightweight architectures* like *MobileNet* [64] and *EfficientNet* [141] use depth-wise separable convolutions for faster inference. *HRNet* [139, 30] maintains accuracy via parallel multi-resolution streams. *Single-stage methods* like *ROMP* [140] perform direct 3D pose regression, while *feature sharing* approaches like *XNect* [97] reduce redundant computation across stages. *Early rejection* filtering [21] and multi-GPU parallelization [97] further reduce latency, though the latter increases system complexity. *Top-down* methods scale linearly with subject count, while *bottom-up* approaches maintain more consistent latency. Cross-view matching adds minimal overhead when using efficient algorithms. Figure 8 shows how different methods scale with scene complexity. While ideal systems would maintain constant processing time, practical implementations must balance optimization strategies based on application priorities.

Computational constraint Real-time methods require modern GPUs for inference [97, 140], with multi-view systems often needing multiple GPUs to parallelize per-view processing. While commercial marker-based setups use hundreds of cameras, markerless systems remain limited in capture volume. The largest reported markerless setup, *VoxelTrack* [182] §4.5.1, achieved 15fps with five cameras in a $10 \times 10 \times 4$ meters volume.

Dataset curation Large-scale dataset annotation is resource intensive - the *COCO* dataset [87] §3.2.1 required 19 minutes per image for 328,000 images. To address imperfect human annotations §6.1, the *CMU Panoptic* dataset [72] §3.2.1 introduced multi-view bootstrapping using 480 cameras and pose estimation from [161]. Their iterative bootstrapping approach [133] improved both dataset quality and model performance through repeated detection-annotation-retraining cycles, achieving state-of-the-art results after three iterations. This technique generalizes to any multi-view dataset with occluded keypoints or small objects.

Domain adaptation and cross-dataset generalization Learning-based markerless methods face challenges in generalizing across different datasets and domains, particularly with varied camera viewpoints, lighting, body shapes, clothing, and environmental contexts (indoor vs outdoor, controlled vs in-the-wild). Most methods currently perform best when trained or fine-tuned on the target scene and camera layout.

Several approaches have been proposed to improve cross-dataset generalization: (1) using robust off-the-shelf 2D pose estimators trained on large datasets like *COCO* §3.2.1, (2) multi-domain training with dataset mixing [53], (3) domain-invariant feature learning [171], (4) data augmentation with random perturbations [157], (5) synthetic-to-real domain adaptation [83].

Deng et al. [36] demonstrated domain adaptation that builds upon *VoxelPose* [147] using adversarial training from a pre-trained model on *CMU Panoptic* §3.2.1. It achieving 6.9% PCP3D §3.2.2 improvement on the *Campus* dataset [10] and 2.7% on *Shelf* compared to the pre-trained model. While these techniques help bridge domain gaps between controlled and real-world environments, robust generalization remains challenging.

Due to limited multi-view datasets, comprehensive cross-dataset evaluations are not standard in benchmarks, making it difficult to assess generalization across different capture scenarios. Future work should prioritize developing standardized cross-dataset evaluation protocols.

3.2 Comparison criteria: body representation and dataset

Most markerless methods represent the human body as a joints-limbs kinematic skeleton, with *OpenPose* §2.2.1 and *Faster-VoxelPose* [176] §4.5.1 being leading examples for monocular and multi-view estimation respectively. No real-time multi-view method currently regresses volumetric bodies. Key datasets include *MPI-INF-3DHP* [94], *COCO* [87], *Human3.6M* [66] and *CMU Panoptic* [72]. For fair comparison, we benchmark methods on *CMU Panoptic* §3.2.1. Standard metrics include Percentage of Correct Keypoints (PCK) and Average Precision (AP) §3.2.2 for 2D pose, and PCP §3.2.2 and MPJPE §3.2.2 for 3D pose. Building on Moeslund and Granum [100]'s criteria, we evaluate methods on their latency in milliseconds, accuracy on *CMU Panoptic* (PCP3D, MPJPE), and the computational scalability to the number of subjects and viewpoints. A detailed comparison of real-time methods is provided in §5.

Dataset	synth	frames	views	persons	motions	annotations
Campus and Shelf [10]	no	$6k$	$4-5$	$3-4$	≈ 5	25 kpts
Panoptic Studio [72]	no	$1.5M$	480	$1-5$	≈ 120	25 kpts
MuPoTS-3D [96]	no	$8k$	8	$1-5$	20	14 kpts
4D association [183]	no	$15k$	6	$3-4$	≈ 5	25 kpts from mocap
CHI3D [48]	no	$728k$	4	2	120	25 kpts, *GHUM*, *SMPL*
MultiHuman [188]	no	150	128	$1-3$	8	scans
ExPI [54, 55]	no	$30k$	68	2	16	36 kpts from mocap
Hi4D [178]	no	$11k$	8	2	100	25 kpts, *SMPL*, scans
HSPACE [9]	yes	$1M$	$1-5$	$1-16$	100	33 kpts, *GHUM* §3.1.2
SynBody [175]	yes	$1.2M$	5	$1-4$	$1,187$	190 kpts, *SMPL* §3.1.2

Table 1: Multi-view multi-person datasets. The *synth* column reports synthetically generated dataset. The *persons* column reports the number of persons in a scene.

3.2.1 Datasets for training and evaluation

Markerless pose estimation relies on several key datasets. For 2D pose estimation, *COCO* [87] is the standard benchmark with 107,000 annotated real-life images. *CrowdPose* [84] provides 20,000 images focused on occlusions and close interactions. *AI Challenger* [166] extends COCO's scope with 300,000 diverse annotated images.

Single view and multi-view datasets with IMUs Two datasets combine IMUs with video: *3DPW* [151] and *Total Capture* [146]. *3DPW* uses smartphone video with 9-17 IMUs per person to track 1-2 subjects outdoors across 51,000 frames, providing synchronized video, IMU data, and *SMPL* [92] fits. *Total Capture* uses 8 calibrated cameras and 13 IMUs per person in a controlled setup, but only contains single-person sequences. Neither dataset is ideal for multi-view multi-person scenarios - a dataset combining synchronized multi-view video and IMUs for multiple interacting subjects remains needed.

Synthetically generated and annotated datasets *SURREAL* (Synthetic hUmans foR REAL) [148] generated 6M frames with 145 body morphologies using *SMPL* [92] §3.1.2, with random backgrounds, lighting and camera positions. As a synthetic dataset, each frame includes body part segmentation, depth maps, optical flow and surface normals. This work proved the suitability of the synthetic generation of training datasets, which will be even more interesting for multi-view annotations. *Joint Track Auto* [43] provided 460K images with 10M poses generated from the video game *Grand Theft Auto V*.

Multi-view single-person datasets While many 2D pose datasets exist, multi-view datasets remain limited. For comprehensive reviews of single-person datasets, see [9, 175]. *Human3.6M* [66] contains 3.6M marker-based 3D poses from 4 viewpoints, capturing 11 subjects performing 17 actions in a motion capture studio, thus not representative of real-life spaces.

SynPose300 [125] was the first synthetic multi-view dataset, featuring 8 subjects with varied body types captured from 3 viewpoints at 2 distances, each performing 3 motions against a white background.

Multi-view multi-person datasets Multi-view multi-person datasets remain limited, posing evaluation challenges. Table 1 summarizes available datasets, which fall into three categories: marker-based, manually annotated, and synthetic.

The *4D association* [183] and *ExPI* [54, 55] datasets used motion capture markers for ground-truth, while *CMU Panoptic* §3.2.1 used multi-view boosting. *MuPoTS-3D* [96] used commercial markerless systems for 14-keypoint skeleton annotation. While this enables precise quantitative evaluation, the markers themselves can affect natural motion and appearance, potentially biasing results. Additionally, the studio environments may not reflect real-world conditions. The *Campus* and *Shelf* datasets [10] serve as challenging benchmarks with 4-5 cameras and close interactions. *CMU Panoptic* provides 1.5M poses across 521 viewpoints, making it invaluable for evaluating crowded scene handling. Leading approaches like *OpenPose* [21] §2.2.1 and *VoxelPose* [147] §4.5.1 use it extensively. Recent datasets Fieraru et al. [47] *CHI3D* and Yin et al. [177] *Hi4D* proposed close interactions between two persons, but with a limited number of 4 views.

Synthetic datasets like *HSPACE* [9] and *SynBody* [175] use statistical body models (*SMPL* [92] §3.1.2 and *GHUM* [170]) to generate diverse data with perfect annotations, enabling systematic evaluation of method robustness. However, they require domain adaptation for real-world use.

Looking forward, future datasets should combine real-world variety, dense camera coverage, complex interactions, high-quality ground truth, and standardized evaluation protocols, while matching *Panoptic*'s scale and the diversity of synthetic data.

3.2.2 Metrics

Following the overview of the reference datasets, we focus on metrics used to quantitatively compare the accuracy of multi-person multi-view markerless methods.

PCP3D: Percentage of Correct Part in 3D The PCP measures the limb correctness based on the distance between the predicted limb keypoints and the ground truth limb keypoints. Burenius et al. [16] extends PCP to 3D by measuring limb correctness based on distance between predicted and ground truth keypoints:

$$\frac{\|\hat{s}_n - s_n\| + \|\hat{e}_n - e_n\|}{2} \leq \alpha \|\hat{s}_n - \hat{e}_n\|$$

where $\hat{s}_n$, $\hat{e}_n$ are ground truth start/end points, s_n e_n are predictions, and the α threshold is typically 0.5 in recent benchmarks. The PCP3D complements MPJPE but doesn't penalize false positives or account for small limbs.

MPJPE: Mean Per Joint Position Error The Mean Per Joint Position Error measures the Euclidean distance between predicted and ground truth 3D joint positions. It is the reference metric for 3D pose estimation, and it can be detailed for each keypoint type to find the granular accuracy of methods on different body parts. Ji and Liu [69] noticed that the MPJPE error is not dependent on the person size in the frame. As such, a low MPJPE does not always indicate an accurate pose estimation, whereas PCP3D is invariant to limb scaling.

$$MPJPE = \frac{1}{N} \sum_{i=1}^{N} \|J_i - J_i^*\|_2$$

where N is joint count, J_i is ground truth, and J_i^* is predicted position.

AP_K: Average Precision The 2D *Average Precision* metric initially defined in *DeepCut* [109] is extended to 3D by considering a pose accurate if its MPJPE is under K millimeters. Common thresholds are AP_{25}, AP_{50}, AP_{100}, and AP_{150}.

4 Architecture of a real-time multi-view multi-person markerless pose estimation method

This section analyzes three real-time markerless pipelines achieving sub-40ms latency: (1) *Top-down* approaches detect person bounding boxes then estimate poses, trading computation scaling with person count for accuracy, (2) *Bottom-up* methods detect and associate body parts in one pass with constant inference time, (3) end-to-end volumetric approaches directly regress 3D poses from multi-view features using learnable voxel representations. Each architecture optimizes different speed-accuracy-scalability trade-offs through neural network design.

4.1 A taxonomy of real-time multi-view multi-person markerless methods

We represent the three different architectures that achieved a real-time latency of less than 40ms in a unified taxonomy in Figure 3. All methods take as input multi-view frames from a calibrated camera system, then use different methods of feature extraction and 3D pose regression.

In Figure 4, we show a schema illustrating a subset of the different architectures discussed in this review, focusing on the most common real-time multi-view pipelines. Together, these figures demonstrate the key steps and characteristics of the three main architectures of current real-time methods. The figures do not include non-real-time architectures previously reviewed in the paper, such as 3D pictorial structures §2.2.2 and parametric body mesh models [92, 170] §3.1.2. Each pipeline processes multi-view inputs to output 3D poses while optimizing accuracy and latency with different computational complexity scaling properties. For a comprehensive comparison of important multi-person architectures reviewed in this paper in previous section §2, we include non-real-time and monocular methods in Table 2. The computational complexity of each key architecture regarding the number of people in the scene is shown, with voxel-based methods currently being the most efficient with constant time complexity. However, the latency performance between voxel-based and top-down methods is close. Table 4 will complete Table 2 with recent real-time multi-view multi-person markerless methods evaluated on the *Campus* [10] benchmark.

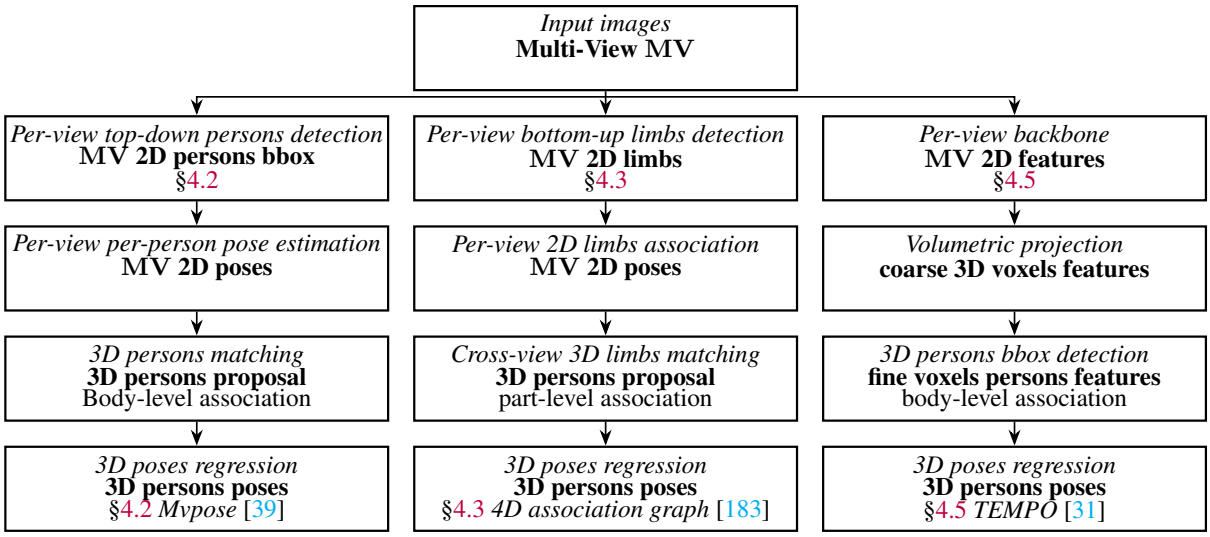


Figure 3: Taxonomy of real-time multi-view multi-person markerless methods.
MV denotes independent variables per-view.

4.2 Per-view *top-down* 2D poses detection then 3D pose regression

The per-view *top-down* architecture applies monocular pose estimation independently to each camera view before fusing to 3D. The pipeline has three stages: (1) per-view 2D pose detection, (2) cross-view pose matching, (3) 3D pose triangulation. We refer the reader to Figure 4a for a visual representation.

***Mvpose*: Efficient Multi-View Pose Estimation** *Mvpose* [39] introduced several algorithmic innovations: (1) two-stage matching with epipolar geometry and appearance features, (2) RANSAC-based triangulation for outlier robustness, (3) parallel pose detection. The mnethod achieves 20fps (4 persons, 5 views) on a GTX 1080Ti GPU, with latencies of 35ms/view for detection, 10ms matching, 10ms triangulation. On *CMU Panoptic* [72] §3.2.1, it achieves 83.30% 3D *PCP* §3.2.2 and 105.63mm MPJPE (5 views, 5 persons) using [33].

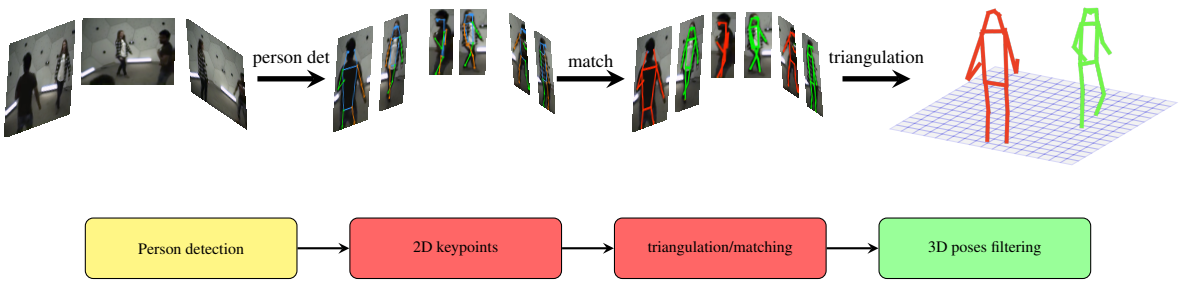
Asynchronous Multi-View Processing Chen et al. [24] proposed an asynchronous processing approach that processes views sequentially rather than in batch. The system uses a linear 3D motion model with timestamp-based penalties to forecast joint positions, updating 3D poses incrementally as new 2D detections arrive. While matching *Mvpose*'s accuracy [39] §2.2.2, it achieved 34 fps with 28 cameras and 16 persons. However, it had a latency of 300ms for the 3d pose inference, but the method technically ran at 34 fps for 28 cameras with 16 persons in the scene. Authors measured that the computational cost of this asynchronous approach scales linearly with camera count, making it suitable for large camera arrays.

4.3 Per-view bottom-up 2D limb detection then 3D pose regression

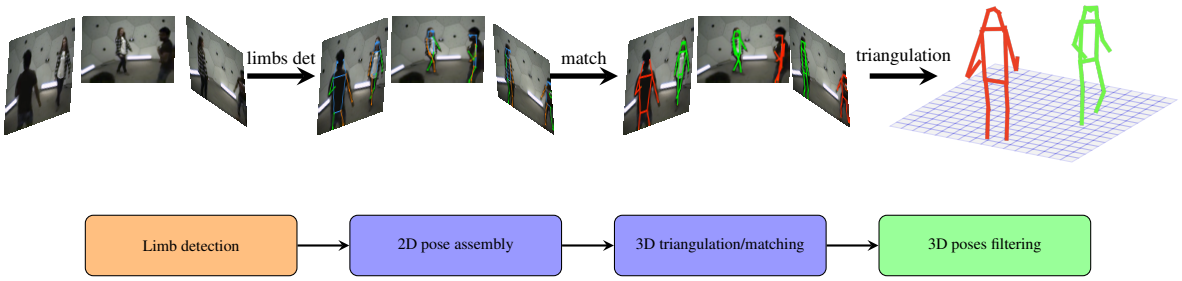
Bottom-up approaches detect limbs independently in each view, then associate them across views to reconstruct 3D poses. The main challenge is to correctly match corresponding limbs between views. We refer the reader to Figure 4b for a visual representation of this architecture.

Stoll et al. [138] pioneered real-time tracking using 3D Gaussian color blobs with 58 joints, extending *Pfinder* [6] §2.1.1. The 3D Gaussians projected efficiently to 2D circles for cross-view matching. Elhayek et al. [41] simplified this to 25 joints and added a 13-joint CNN detector [143], matching the 3D pictorial structures' accuracy [11] §2.2.2 at 1fps on *HumanEva* [132].

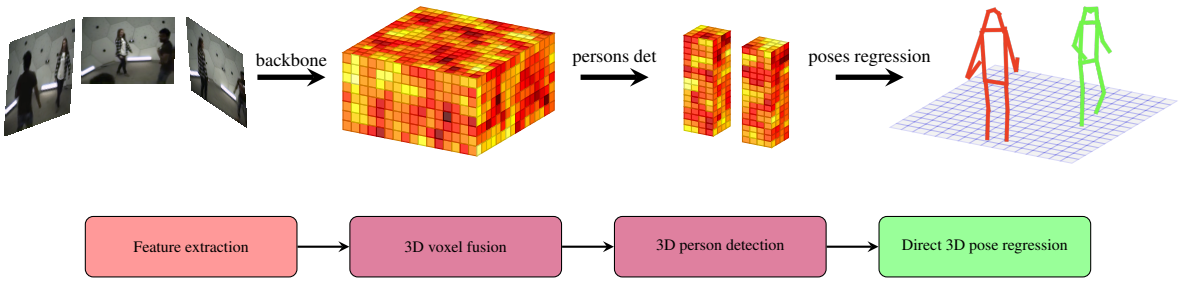
Real-time multi-view multi-person architectures Schwarcz and Pollard [126] first proposed multi-view aggregation of *OpenPose* detections, achieving 10% PCP3D improvement over *3D pictorial structures*, but it required offline processing due to full sequence graph optimization. In 2019, Kadkhodamohammadi and Padoy [73] developed a real-time method combining *OpenPose* detection with cross-view multi-person 3D pose regression. The two-stage design enabled independent optimization using specialized datasets. Their regression method required each person to be visible in at least two views and used: (1) cross-view matching with a 20-pixel threshold, (2) CNN-based 3D pose



(a) Per-view *top-down* persons detection and matching §4.2



(b) Per-view *bottom-up* limbs detection and matching §4.3



(c) Voxel features person detection and poses regression §4.5

Figure 4: Real-time multi-view multi-person markerless architectures with computational complexity scaling: (a) Top-down O(n): higher precision, linear person scaling; (b) Bottom-up O(1)-O(log n): robust to occlusions, sub-linear scaling; (c) Voxel-based O(1): optimal efficiency, constant scaling

reconstruction from matched 2D detections. Training on *Human3.6M* with injected noise achieved 4.7cm MPJPE error on *Human3.6M*. By parallelizing *OpenPose* across GPUs, the system achieved 28fps (37ms latency), making it one of the first truly real-time multi-view multi-person systems.

Real-time spatiotemporal 4D association graph In 2020, Zhang et al. [183] proposed a 4D association graph for joint cross-view and temporal matching optimization. The method runs at 25fps with 5 cameras and 5 persons, using *OpenPose* detections weighted in a 4D graph with epipolar distance and temporal distance losses. A Kruskal matching algorithm [79] assigns limbs to persons, with bone lengths used for identity tracking. Then, a parametric *SMPL* [92] §3.1.2 skeleton is fitted to filter the 3D joints. Processing times are 22.9ms for *OpenPose* (for a batch of 5 cameras on an Nvidia Titan X GPU), 11ms for graph solving, and 4ms for *SMPL* optimization. The method scales linearly with cameras but non-linearly with persons due to graph size. It achieves 2% higher PCP than *Mvpose* [39] §2.2.2 on *Shelf* dataset. The following year, Zhang et al. [184] extended this to include hand and face pose estimation, making it the first real-time body-face-hands multi-view multi-person method.

Method	Type	MPJPE(mm)	Lat(ms)	Pers.	Views	Key Features
3DPictorial §2.2.2	Pictorial	-	105	$O(n^2)$	Multi	+ explicit constraints - complexity
OpenPose §2.2.1	CPM	-	40	$O(1)$	Single	+ bottom-up + speed
4D Graph §4.3	CPM	51.3 (*Camp*)	40	$O(n)$	Multi	+ bottom-up limb + cross-view
Mvpose §2.2.2	CPM	57.3 (*Pano*)	50	$O(n)$	Multi	+ Efficient matching + RANSAC triangulation
VIBE §2.3.2	SMPL	56.9 (*3DPW*)	16.1	$O(n)$	Single	+ body model - top-down bottleneck
ROMP §2.2.1	SMPL	60.5 (*3DPW*)	33.3	$O(1)$	Single	+ constant time
VoxelPose §4.5.1	Voxel	19.5 (*Pano*)	110	$O(1)$	Multi	+ volumetric fusion + constant time
FastVoxPose §4.5.1	Voxel	18.26 (*Pano*)	33	$O(1)$	Multi	+ speed
TEMPO §4.5.1	Voxel	14.68 (*Pano*)	34	$O(1)$	Multi	+ volumetric + multi-view fusion
SelfPose3d	Voxel	24.5 (*Pano*)	80	$O(1)$	Multi	+ voxel-keypoint fusion + constant time
VoxelKeypointFusion	Voxel	47.8 (*Pano*)	238	$O(1)$	Multi	+ voxel-keypoint fusion + constant time

Table 2: Comparison of multi-person markerless architectures with computational complexity scaling properties

4.4 Epipolar feature pooling for 3D pose regression

Following monocular *top-down* approaches, some methods detect 2D poses, match features across views, and regress 3D poses. *EpipolarPose* [74] pioneered multi-view single-person pose estimation using epipolar matching of 2D joints for 3D regression. However, its volumetric convolutions with cubic complexity prevented real-time performance. This approach was later extended to multi-view multi-person methods.

Graph optimization for cross-view feature pooling Wu et al. [167] proposed a two-step method using graph models to first localize body centers and then regress 3D joint positions. Their Multi-view Matching Graph Module aggregates features across cameras using epipolar geometry and body topology, followed by a Center Refinement Graph Module that efficiently samples 2D features. Compared to *VoxelPose* [147] §4.5.1 which requires 128,000 queries per frame, this method needs only 1,830 queries while achieving 15.84mm MPJPE on *CMU Panoptic* [72]. The method runs at 10fps with 4 persons, but it decreases to 8fps with 5 persons due to linear scaling of the person refinement cost (6.8ms per person).

Real-time multi-view joints clustering In 2022, *QuickPose* [191] proposed a fast multi-view multi-person matching framework using only *OpenPose* [21] §2.2.1 2D joint positions. The method enumerates possible skeletons from 2D joints in each view using a tree-structured graph. Multi-view joint association is done by clustering joints into persons based on maximizing a *skeleton affinity score*, built from epipolar joints distance [183] §4.3 and part-association scores. Then, 3D poses are computed via triangulation after clustering. The single-threaded clustering algorithm achieves 30fps (30ms) for 8 views with *OpenPose* parallelized on four RTX2080 GPUs at 800x600 resolution. Computational cost scales linearly with cameras (1ms per camera) and persons (1ms per person). While ten times faster than Zhang et al. [183] on *Shelf* benchmark, MPJPE accuracy on standard datasets is lower. Authors argued about the accuracy-latency trade-off. Moreover, their method was not dependent on a learning of the scene to extract 3D poses. While this lowers accuracy, it also makes this method generalizable and easy to deploy in any unseen scene, whereas learning-based methods need to retrain.

FusionFormer [18] proposed a calibration-free multi-view transformer leveraging temporal information. Using *ViTPose* [172, 173] for 2D pose estimation, it achieved 15.1mm MPJPE on *Human3.6* [66] §3.2.1 with 4 views and 27 frames, outperforming *Cascaded Pyramid Network* [26] (25.4mm MPJPE). On *TotalCapture* [146], it reached 21.7mm MPJPE with *ResNet101* [58]. While more efficient than *MTF-Transformer+* [130], it lacks multi-person support and latency metrics.

4.5 Direct multi-view voxel representation to 3D pose regression

In comparison with the previous 2D detection methods, some methods directly lift backbone features to a 3D voxel representation, bypassing 2D detection. The voxel space discretization requires balancing precision and speed for the scene size. We refer to Figure 4c for the architecture.

4.5.1 Voxel features representation

These methods fuse multi-view features into a unified voxel space for global scene understanding. Due to memory constraints, voxel resolution must be reduced for large scenes. All methods follow a two-stage top-down approach with person detection then pose estimation, typically using pre-trained 2D detectors like *COCO* for initial feature extraction.

Bottom-up 3D part-affinity field *Light3DPose* [42] extends 2D *Part-Affinity Fields* to 3D by projecting 2D heatmaps into voxel space. Using a lightweight *MobileNet-V1* backbone [64, 106] and *V2V* volumetric projection layer, it generates 3D joint heatmaps and *Part-Affinity Fields*. With a simplified 12-limb skeleton, it achieves 5fps (146ms) for 5 views and 4 persons on GTX 1080. The *V2V* layer has a fixed 125ms cost, while backbone cost grows linearly but slowly (3ms per view). As a bottom-up method, performance should be person-count invariant, though this was not evaluated. Unfortunately, with no source code released, we cannot benchmark the method in our comparison.

***VoxelPose* voxels features representation** [147] is a two-step end-to-end method using voxel scene representation. Features from each view are aggregated into 3D voxels with epipolar projection of 2D joint heatmaps, similar to *Learnable Triangulation* [68]. Due to memory constraints, it uses dual resolutions: a low-resolution (8m $\times$ 8m $\times$ 2m at 80 $\times$ 80 $\times$ 20 voxels) for person detection via a Cuboid Proposal Network, and high-resolution (2m $\times$ 2m $\times$ 2m at 64 $\times$ 64 $\times$ 64 voxels) for per-person joint regression via a Pose Regression Network. The method nearly matched the MPJPE accuracy of [68] with 19mm reported instead of 17.7mm, all while being faster. However, this method is not real-time with 320ms per frame (3fps, measured by [160] as the original paper did not report speed) on *CMU Panoptic* with an Nvidia RTX 2080Ti GPU, with a linear scaling per person due to the per-person pose regression.

Recently, Song et al. [135] proposed a two-step approach: first, a local then a global optimization network are used to optimize the 3D joint positions, benchmarked on *Campus* and *Shelf*. Srivastav et al. [137]'s *SelfPose3d* modified *VoxelPose* to enable self-supervised training without 3D ground truth by using *HRNet* 2D poses and cross-view geometric constraints. The method achieves 96.4% AP50 and 24.5mm MPJPE on *CMU Panoptic*, comparable to fully-supervised approaches despite not using any 2D or 3D ground truth labels. Key innovations include: (1) adaptive supervision attention with hard attention for L1 joint loss and soft attention for L2 heatmap loss to handle noisy 2D poses, (2) cross-affine-view consistency with random rotations and scaling for geometric constraints, (3) self-supervised 3D root localization using synthetic data and root-only regression that enables 10fps inference on an Nvidia RTX4090. The method demonstrates strong cross-dataset generalization, achieving 95.1% PCP on *Shelf* [10] without fine-tuning, outperforming both optimization-based methods like [39] §2.2.2 (96.9% PCP) and fully-supervised approaches like *VoxelPose* (96.9% PCP) on this cross-dataset evaluation. Bermuth et al. [12] proposed a learning-free algorithmic approach that similarly used *RTMPose* [70] with voxel-keypoint fusion. The method achieves 47.8mm MPJPE at 8.4fps (RTX3090 GPU), with strong cross-dataset generalization demonstrated on *Human3.6M* [66] §3.2.1 (64.3mm MPJPE), *Shelf* [10] (51.3mm MPJPE), and *Campus* (84.4mm MPJPE) without any retraining. Key innovations include: (1) person-id images for joint association, (2) bottom-up joint detection with voxel-based triangulation, (3) outlier filtering in 3D space. The authors also evaluated depth sensor integration through voxel masking, which reduced invalid detections, but it decreased precision and framerate due to imperfect depth-color synchronization.

***VoxelTrack* bottom-up 3D joints from voxels features** Zhang et al. [182] extends *VoxelPose* with temporal tracking using heatmaps and person Re-ID features fused into 3D voxels. An occlusion detector triggers Re-ID tracking when needed. The method uses sparse 3D convolutions for efficiency and follows a bottom-up approach: detecting 3D joints before grouping them with an *Ambiguity Resolution Network* in a 32^3 voxel space. Processing cost increases by only 2.72ms per person, plus 2.5ms for Re-ID. Operating in a 10 $\times$ 10 $\times$ 4m space (160 $\times$ 160 $\times$ 64 voxels). A fast version is proposed that leverages a *MobileNet-V2* backbone [123] and it achieved 15fps with 5 views on an Nvidia 2080Ti GPU.

***TesseTrack* spatiotemporal voxels** Reddy et al. [115] introduced 4D CNNs to process voxel representations of 5 previous frames. Following *VoxelPose*'s pipeline, it first detects persons in 3D voxel features, then crops and aggregates 4D spatiotemporal voxels (*tesseract*) per person for 3D pose regression. The method achieved remarkable 7.3mm MPJPE on *CMU Panoptic* (13.1mm without temporal tracking), but ran below 1fps on two 32GB Tesla V100 GPUs due to 4D CNN costs. View scaling matches *VoxelPose* (+30ms/view), but person scaling is higher (+40ms/person)

from 4D processing. We refer the reader to the original paper for further metric comparison, given that the method's source code is not open sourced, and we therefore cannot integrate it into our comparison.

Real-time Faster-VoxelPose The first real-time voxel representation method was *Faster-VoxelPose* [176], which ran at 30fps on the *CMU Panoptic* dataset [72] §3.2.1, with 5 persons in the scene, with a MPJPE of 18.26mm. It represents a significant milestone in real-time methods and will likely see further improvements as models continue to advance. The method is an improvement of *VoxelPose* [147], detailed in §4.5.1. It represented the voxel features as three separate and orthogonal 2D planes. The same two step approach is kept: a coarse person detector and a fine joint estimation, but the 2D plane representation makes the method ten times faster than *VoxelPose*. The expensive 3D CNNs are replaced by 2D CNNs for a massive speedup. For the joint estimation step all features outside the detected human body's bounding box are ignored for efficiency. We notice that the computational cost scales linearly with the number of persons in the scene but stays below 15ms for up to 5 persons.

Shuai et al. [131] proposed a similar method to *FasterVoxelPose* to solve close human interactions, but is not real-time. The approach was evaluated on the [47] and [177] datasets which focus on close interactions between 2 persons. The method introduced a novel two-stage pose estimation network to handle the challenges of close interactions. First, a cleaned 3D heatmap volume of all keypoints are filtered by a CNN, then a second stage estimates per-person keypoint probability volumes while suppressing responses from other individuals. The method proposed a simple keypoint temporal filter with a threshold of a 5cm motion between frames. It achieved significant improvements in accuracy with a MPJPE of 20.28mm and *PCK@50* of 98.29% on the difficult *Hi4D* dataset. Recently, Zhuang and Zhou [192] improved upon *FasterVoxelPose* by introducing two key innovations: a depth-wise projection decay (*DPD*) and an encoder-decoder network (*EDN*). The *DPD* addresses depth ambiguity by applying a Gaussian decay centered on the estimated root depth when constructing voxel features, rather than using uniform epipolar projection. The *EDN* processes re-projected voxel features through parallel decoders with deconvolution skip-connections for richer multi-scale information fusion. The method achieved a MPJPE of 17.42mm on the *CMU Panoptic* dataset [72] §3.2.1 while maintaining real-time speed at 30.5fps on an Nvidia RTX 2080 Ti GPU. The improvements were particularly significant with fewer cameras, reducing MPJPE by 19% and 46% with two and one views respectively compared to *FasterVoxelPose* [176]. The method also demonstrated strong performance on other datasets, achieving 96.4% and 97.7% PCP3D on *Campus* and *Shelf* [10] respectively. Unfortunately the code of this method is not open sourced, so we cannot integrate it into our comparison.

Real-time spatiotemporal voxel representation Recently, *TEMPO* [31] introduced a fast spatiotemporal voxel representation of features, similar in spirit to *TesseTrack*'s [115] spatiotemporal 4D CNNs, but real-time (30 times faster), running at 29.3fps on an Nvidia A100 GPU, and with a 10% lower MPJPE of 14.68mm on the *CMU Panoptic* dataset. It has the same speed as *FasterVoxelPose* [176] but with a lower MPJPE. The method is a three-step method based on *VoxelPose*'s voxel representation for person detection, *FasterVoxelPose*'s 2D CNNs, with *TesseTrack*'s temporal tracking with a simple recurrent network. First, a backbone extracts features on each view, which are unprojected to a 3D voxel representation, from which only a *top-down* bird's-eye view is kept with a maxpooling. From these top-view features, a 2D CNN produced a person body center heatmap, which is sampled to get the top K locations of body center points. At this point coordinate, the vector of features in the 3D voxel representation is passed to a 1D CNN to regress the body joint's height location, so that the body center's 3D position is inferred. Similarly for the network training, each person's body bounding box is regressed from the body center's 3D position with a multi-headed 2D CNN on the 3D voxel representation. Each bounding box is defined by its width, length and center. Each 2D view's features are cropped based on the bounding boxes, then unprojected to three separate orthogonal planes, like in *FasterVoxelPose*. Finally, three separate 2D CNNs infer the joint positions heatmaps in each plane, and top-k parsing gives the full 3D body joint positions, like in *FasterVoxelPose*.

TEMPO introduced temporal tracking in the form of the *top-down* bounding box of each person, tracked with the off-the-shelf *SORT* [13] Kalman-filtered tracker. From the bounding boxes in the previous frames, each person's three 2D separate orthogonal unprojected feature planes are taken as the tracker input. Spatial Gated Recurrent Units [8] take the tracker features, center positions and the cropped three separate orthogonal features planes from the previous frame, and output the current and the next frame 3D body joint positions. Authors noted that *TEMPO* has a similar tracking accuracy as *VoxelTrack* §4.5.1, but without any reidentification features (Re-ID) and at a fraction of the computational cost as it only used the *top-down* view's person bounding box for tracking, whereas *VoxelTrack* tracked every joint for every view. The method is trained on *CMU Panoptic* and the generalization to other camera layouts is evaluated on *Human3.6M* [66] §3.2.1. The authors found that the accuracy is worse when the training is not done on the evaluation scene, but this can be mitigated by training on both the *CMU Panoptic* and *Human3.6M* datasets, with a reasonable 30mm MPJPE §3.2.2. *TEMPO* is the first method to also perform next-frame prediction in a multi-view multi-person setup, whereas *VoxelTrack* only tracked on past frames. The pose prediction capability of the model makes its output

more accurate with temporally smoothed poses. Thus, the *AP50* §3.2.2 accuracy of 89% and the MPJPE of 14.68mm are better than *FasterVoxelPose*.

4.5.2 Plane sweep depth representation

PlaneSweepPose [86] introduced depth cues features from pair of stereoscopic cameras as an alternative to voxels for avoiding cross-view matching. 2D poses are projected onto 64 virtual planes, with a coarse-to-fine approach that first regresses full body depth, then individual joint depths within $[-1, 1]$ meters. Finally, 3D poses are computed by back-projecting 2D poses with depths. This avoids voxel memory costs through viewpoint-dependent depth planes. On *Shelf* (5 views, 4 persons), it achieved 16.75mm MPJPE, slightly worse than *VoxelPose*'s 17.68mm, but it ran 20x faster. On *CMU Panoptic*, it ran at 4.3fps with 5 views [31] and a mean of number of 3.4 persons in the scene on an Nvidia A100 GPU with a *ResNet50* [58] backbone.

4.5.3 Multi-view multi-person transformers representation methods

Transformer architectures with Attention modules [149] provide efficient multi-view feature pooling as an alternative to to the voxel representation. The attention mechanism fuses features across views by querying selected locations, enabling learned cross-view pooling at minimal computational cost, similar to epipolar parsing, but with more understanding gathered from the learning of the multi-view setup and the human body knowledge.

***MvP* direct regression approach with Transformers** In comparison with *VoxelPose* which needed to project each view's heatmap into the voxel representation, *MvP* (Multi-View Pose transformer) [160] directly regressed multiple person's 3D poses by using a new Projective Attention model that queries features at the supposed 3D joint positions. An implicit body skeleton model is learned, so that the Transformer module can query cross-view features at the right position. For each joint, a positional embedding is learned (named joint query) and a Projective Attention module fused multi-view features and assigned each joint to its corresponding person. The Projective Attention is a constrained Attention module that projects the estimated 3D joint location as the anchor point in each view, then queries a limited set of per view 2D local feature points (2 or 4 points), near the projected 3D point, with a deformable convolution. This scheme is efficient, but would not generalize to different camera layouts, thus an augmentation strategy is added: a weighted multi-view global pooling of each feature, done for each query. For each query specific to a view, retrieved features will have some features of from the other views weighted in. The method is not real-time, running at 6fps (170ms per frame) on an Nvidia RTX 2080Ti, with a constant latency regardless of the number of persons, thanks to the transformer queries, which is a big improvement over *VoxelPose*. Finally, the *MvP* method has a 9% better accuracy than *VoxelPose* on *CMU Panoptic* [72] §3.2.1 and a 12% lower MPJPE §3.2.2, with 15.8mm.

Epipolar Transformers features fusion Transformers have been a promising neural network-based representation since their 2017 definition [149], as they can manipulate complex and weakly correlated inputs, replacing convolutions or recurrent networks. The *Epipolar Transformers* [60] method proposed a single-person multi-view 3D pose estimation. It leveraged epipolar lines to sample a fixed number of feature vectors from a backbone network in other views, and weighted them based on their similarity, like [111] but more efficient. All features are fused across all the views, and a final 3D feature is built for the original 2D detection point. From this 3D features representation, any human body representation can be regressed, here with a 3D pictorial structures model [16] §2.2.2. Cameras must be calibrated to apply the epipolar geometry. The authors evaluated the method on *Human3.6M* with a MPJPE §3.2.2 of 19mm, that nearly matches *LearnableTriangulation*'s [68] 17.7mm at a lower computational cost. However, this method cannot work with multiple persons, as a person detector and person matching would need to be integrated before the Epipolar Transformers module.

Recently, *PETR* [128] introduced a novel transformer-based architecture for end-to-end monocular multi-person pose estimation. The method views pose estimation as a hierarchical set prediction problem, using multiple pose queries to directly reason about full-body poses. A pose decoder first predicts instance-aware poses, followed by a joint decoder that refines the poses by exploring kinematic relations between body joints. The attention mechanism allows the model to adaptively attend to features most relevant to target keypoints, helping overcome feature misalignment issues. The *PETR-R50* model achieves 67.6 AP on *COCO* test-dev §3.2.1 with a *ResNet50* [58] backbone, outperforming previous multi-stage methods like *PifPaf* [76] with a 3 times faster speed at 89ms per frame on an NVIDIA V100 GPU. While currently designed for monocular pose estimation, the transformer-based architecture of *PETR* could be extended to multi-view scenarios by incorporating cross-view attention mechanisms similar to *Epipolar Transformers*. The hierarchical pose queries could be augmented to reason about 3D poses across multiple views, potentially offering an efficient end-to-end approach for multi-view pose estimation.

***VtP*: volumetric transformer** Transformers are a powerful learning-based method for multi-dimension manipulation. *VtP* [28] introduced Transformer-based voxel pooling instead of classical 3D CNNs. The first part of the method is similar to *VoxelPose*, a Cuboid Proposal Network regresses each person's global position. From the cropped feature voxels of each person, a transformer regressed each joint's 3D poses. The authors discussed the quadratic cost to the data size of the transformer self-attention. Their solution is to limit the use of the transformer to a low dimension embedding of the voxel features, but the embedding might lose some information with a small size. Unfortunately, the paper did not disclose any latency timing, but the accuracy metrics are well reported, and we choose to integrate this method into our comparison as the only volumetric transformer. As a *top-down* method, the computational scaling should be linear to the number of persons in the scene. The backbone with the volumetric projection should also be similar to *VoxelPose*, thus the scaling to the number of viewpoints should be less than ideal. For these reasons real-time performance is currently not achievable.

5 Comparative Analysis

Following the overview of the different architectures, we propose an exhaustive comparison based on the same reference datasets. We will evaluate the accuracy, latency, and the computational scalability of the different methods.

5.1 Accuracy and latency trade-off

The main markerless trade-off is to achieve the best accuracy with the lowest latency. Here, we establish clear benchmarking criteria to enable fair comparisons between methods. We benchmark methods on the reference multi-view multi-person *CMU Panoptic* dataset [72], detailed in §3.2.1. To ensure consistent evaluation, we follow the standardized benchmark procedure from [147], using a fixed set of 5 cameras (index 3, 6, 12, 13, 23), test sequences: *160906_pizza1*, *160422_haggling1*, *160906_ian5*, *160906_band4* (excluding *160906_band3*), input resolution: 960×512 pixels, backbone network: *ResNet50* [58], input resolution: 960×512 pixels, backbone network: *ResNet50* (unless otherwise noted), Nvidia RTX 4090 GPU.

The benchmark was reproduced using the *XRMocap* open-source toolbox [33]. We evaluate methods using three standardized accuracy metrics: *Average Precision* (AP §3.2.2), *Mean Per-Joint Position Error* (MPJPE §3.2.2), *Percentage of Correct Parts* (PCP3D §3.2.2). Results are presented in Table 3 for *CMU Panoptic* and Table 4 for *Campus* and *Shelf* benchmarks. Some methods deviate from our standard evaluation setup: *VoxelTrack* §4.5.1 uses a *DLA-34* backbone instead of *ResNet50* [58]. We include it as the only *bottom-up* voxel method, noting that *DLA-34* has comparable computational cost to *ResNet50* on the RTX4090. *TesseTrack* §4.5.1 is excluded as its source code is not released and it uses the more computationally intensive *HRNet* backbone. *QuickPose* §4.4 uses *OpenPose* [21] §2.2.1 for 2D joint detection. While its code is not open, detailed supplementary materials allow inclusion. We apply a latency penalty as results were reported on an Nvidia RTX2080Ti rather than our reference RTX4090.

Figure 5 shows the trade-off between accuracy (AP_{50}) and latency on the *CMU Panoptic* benchmark. Recent methods like *TEMPO* [31] §4.5.1 and *Faster-VoxelPose* [176] §4.5.1 achieve the best balance, with more than 98% accuracy while maintaining latencies around 35ms on an RTX4090. This represents a significant improvement over earlier methods like *VoxelPose* [147] §4.5.1 which had higher latencies (>300ms) for similar accuracy levels. The graph demonstrates the clear trend toward methods that optimize both metrics simultaneously, rather than trading one for

Method	AP_{25} ↑	AP_{50} ↑	AP_{100} ↑	AP_{150} ↑	MPJPE ↓	Latency (ms)↓
Learnable Triangulation [68]	-	-	-	-	12.85	500
MVpose [39] §2.2.2	-	-	-	-	81.17	50
VoxelPose [147] §4.5.1	83.59	98.33	99.76	99.91	17.68	313
VoxelTrack† [182] §4.5.1	79.34	96.83	99.58	-	18.49	57
Wu et al. [167] §4.4	-	98.10	-	-	15.84	125
VtP [28] §4.5.3	83.79	97.14	98.15	98.40	17.62	-
PlaneSweepPose [86] §4.5.2	92.12	98.96	99.81	99.84	16.75	233
MvP [160] §4.5.3	**92.28**	96.6	97.45	97.69	15.76	278
QuickPose‡ [191] §4.4	-	27.5	-	-	27.5	30
Faster-VoxelPose [176] §4.5.1	86.66	98.08	99.26	99.53	18.41	**32**
TEMPO [31] §4.5.1	89.01	**99.08**	**99.76**	**99.93**	**14.68**	34

Table 3: Accuracy AP, MPJPE (mm) and latency comparison on CMU Panoptic, 5 views, *ResNet50* backbone, input resolution of 960×512. † indicates a *DLA-34* backbone and ‡ *OpenPose*, both on RTX2080Ti instead of *ResNet50* on RTX4090.

Method	Campus 3 cam, 3 pers		Campus 5 cam, 4 pers		
	PCP3D ↑	MPJPE ↓	PCP3D ↑	MPJPE ↓	Latency (ms)↓
MVpose [39] §2.2.2	96.3	84.5	96.9	55.3	50
VoxelPose [147] §4.5.1	96.7	78.2	97.0	57.3	313
VoxelTrack† [182] §4.5.1	-	96.7	-	97.1	57
Wu et al. [167] §4.4	-	-	97.7	-	125
VtP [28] §4.5.3	96.3	80.1	97.3	56.3	-
PlaneSweepPose [86] §4.5.2	97.0	-	97.9	-	233
MvP [160] §4.5.3	96.6	**64.1**	97.4	52.2	278
4D association [183] §4.3	81.5	287.8	96.4	**51.3**	40
QuickPose‡ [191] §4.4	-	-	98.1	-	30
Faster-VoxelPose [176] §4.5.1	96.9	-	97.6	-	**32**
TEMPO [31] §4.5.1	**97.3**	-	**98.0**	-	34

Table 4: PCP3D accuracy, MPJPE (mm) and latency on the Campus and Shelf datasets. *ResNet50* backbone, input resolution of 960×512. † indicates a *DLA-34* backbone and ‡ *OpenPose*, both on RTX2080Ti instead of *ResNet50* on RTX4090.

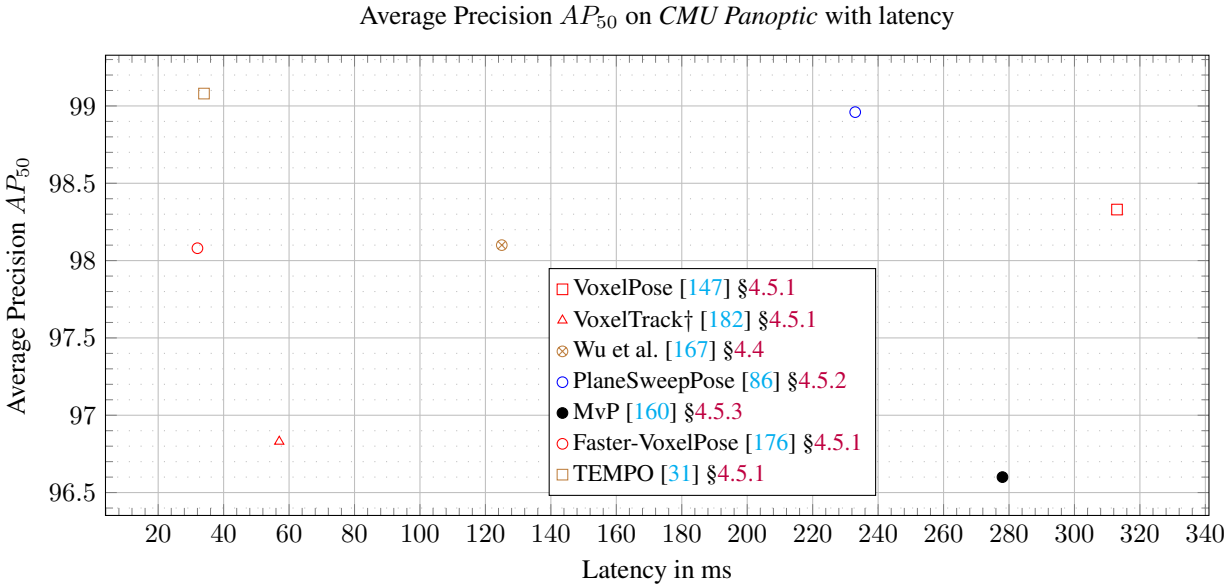


Figure 5: Average Precision AP_{50} on CMU Panoptic. *ResNet50* backbone input resolution of 960×512. † indicates a DLA-34 backbone instead of *ResNet50*.

the other. On the *CMU Panoptic* benchmark, in Figure 6, both *TEMPO* and *Faster-VoxelPose* exceed 98% AP_{50} with latencies around 35ms on an RTX4090. Similarly impressive results are seen on the *Campus* and *Shelf* datasets, Figure 7, where these methods achieve 97-98% PCP3D §3.2.2 scores. In terms of precision, they demonstrate excellent performance with MPJPE values around 10mm, significantly outperforming earlier approaches while maintaining real-time processing speeds.

5.1.1 Scaling to the number of persons

The computational cost scaling with the number of persons varies significantly between methods based on their architecture. Figure 8 shows how different methods scale with increasing number of people. *FasterVoxelPose* maintains consistent latency as views increase, with only small increases per additional person. In contrast, *MvP* has a fixed 170ms latency for 5 views but increases substantially with more views. Overall, *FasterVoxelPose* and *4D Association Graph* achieve the best latency scaling.

We can classify methods into two main categories: *Bottom-up* methods like *Mvpose* [39] §2.2.2 and *4D association* [183] §4.3 generally scale better with the number of persons since they detect all joints at once before associating

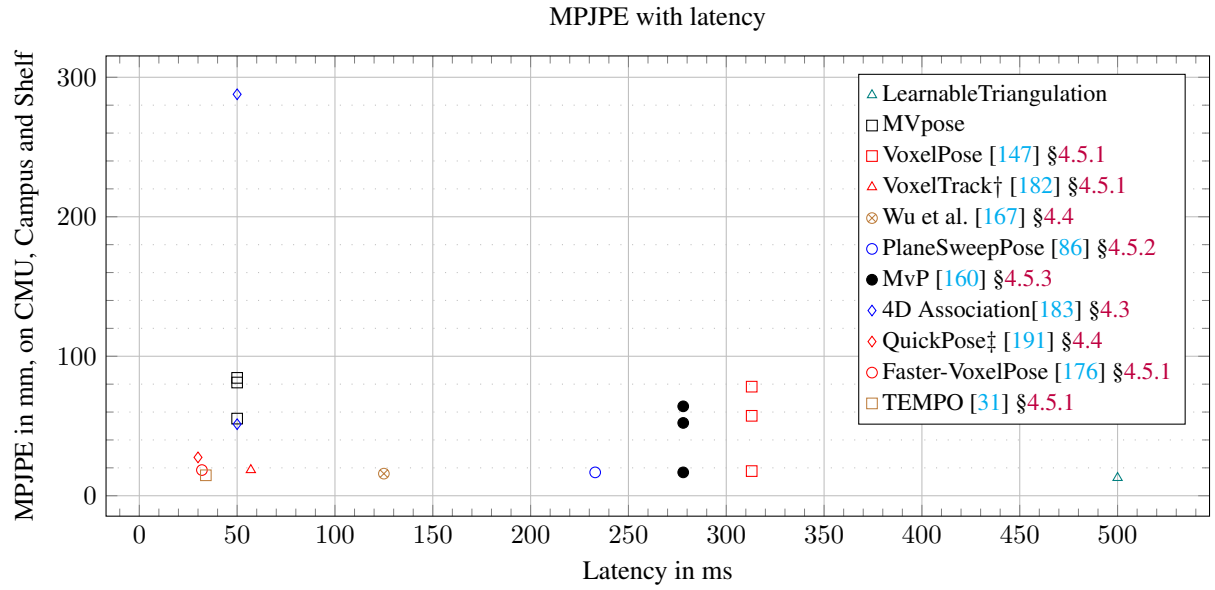


Figure 6: MPJPE on Campus, Shelf and CMU Panoptic. *ResNet50* backbone input resolution of 960×512. † indicates a *DLA-34* backbone and ‡ *OpenPose*, both on RTX2080Ti instead of *ResNet50* on RTX4090.

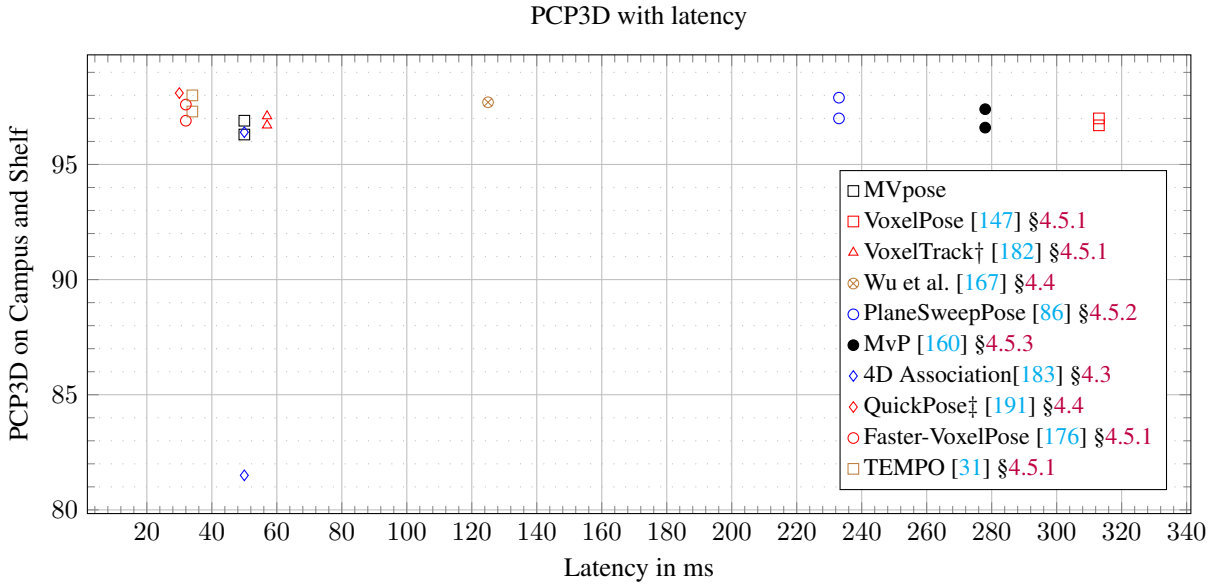


Figure 7: PCP3D on Campus and Shelf. *ResNet50* backbone input resolution of 960×512. † indicates a *DLA-34* backbone and ‡ *OpenPose*, both on RTX2080Ti instead of *ResNet50* on RTX4090.

them into skeletons. However, they can suffer from accuracy degradation with more people due to increased joint association complexity. *Top-down* methods like *VoxelPose* [147] §4.5.1 and *Faster-VoxelPose* [176] §4.5.1 have a more linear scaling since they process each detected person separately. For example, *Faster-VoxelPose* adds approximately 2.7ms per additional person. While this makes the computational cost more predictable, it can become prohibitive with many people. Recent methods have made progress in reducing this per-person overhead. *TEMPO* [31] §4.5.1 maintains real-time performance (>30fps) with up to 5 people by using efficient feature representations. However, even these optimized approaches will eventually hit performance limits as the number of tracked individuals grows.

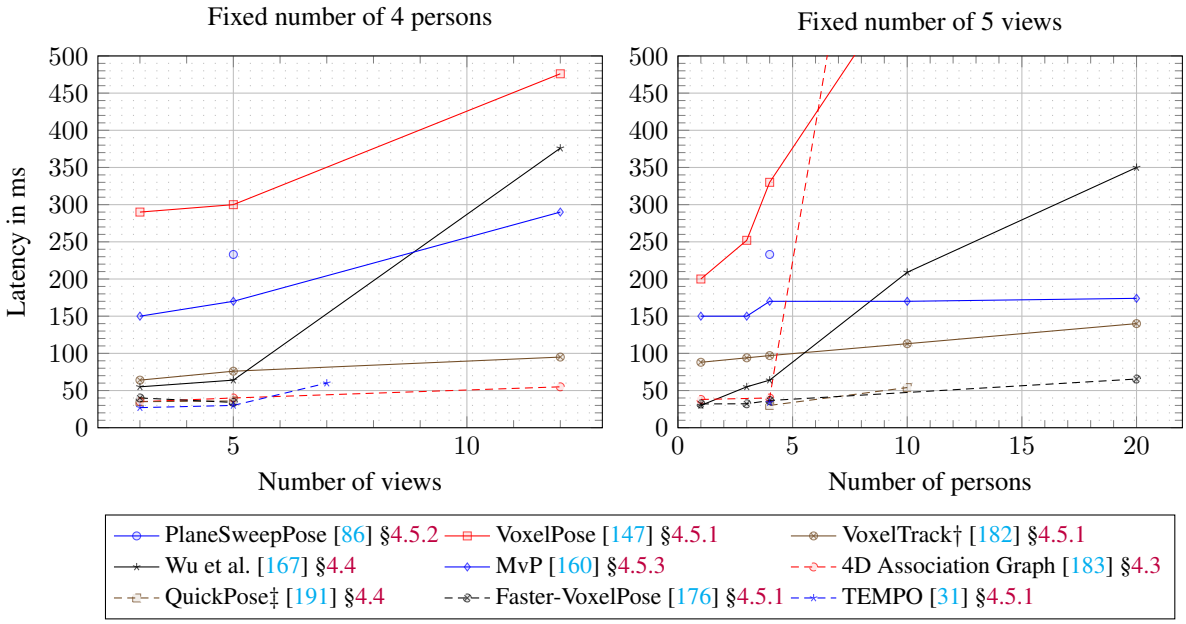


Figure 8: Methods latency for a constant number of 5 views or 4 persons. *ResNet50* backbone input resolution of 960×512. † indicates a *DLA-34* backbone and ‡ *OpenPose* instead of *ResNet50*.

5.1.2 Scaling to the number of viewpoints

The difference in scalability of methods is large, as it depends on their architecture. Figure 8 shows that methods scale differently with more viewpoints. *FasterVoxelPose* maintains low latency as views increase, while *MvP* scales poorly, taking 170ms for 5 views, and significantly more beyond that.

We notice that the 2D feature extraction cost is mostly linear to the number of views. On recent GPUs, like the Nvidia RTX 4090, with the default *ResNet50* [58] model, it takes 5ms to 10ms per view, but this could theoretically be improved with batch processing and GPU compilation methods (*TensorRT*). This cost can become negligeable with respect to the increase of cost as the number of detected persons grows. We noticed that *Faster-VoxelPose* [176] §4.5.1 *TEMPO* method [31] §4.5.1 are the most interesting method, with a slow increase of the latency with the number of views and persons in the scene.

6 Conclusion

This review analyzes real-time multi-view multi-person markerless motion capture, examining key methods from early works to current state-of-the-art, highlighting key architectural advances that have enabled significant improvements in accuracy and efficiency §2. Three dominant architectural approaches have emerged §4: volumetric methods like *VoxelPose* [147] §4.5.1, plane sweep methods leveraging multi-view geometry, and graph-based skeletal reasoning.

Recent advances have achieved both high accuracy and real-time performance. *VoxelTrack* [182] §4.5.1 introduced temporal consistency through feature tracking, while *Faster-VoxelPose* [176] §4.5.1 optimized volumetric processing for efficiency. *TEMPO* [31] §4.5.1 advanced temporal modeling and feature representations, achieving 34ms latency with 99.08% *AP50* accuracy on *CMU Panoptic* §3.2.1. Modern methods consistently achieve >97% PCP3D, though computational scaling remains challenging. Bottom-up methods like *MVpose* §2.2.2 show better theoretical scaling but lower accuracy compared to top-down approaches like *Faster-VoxelPose* and *TEMPO*.

Looking forward, we identify several promising research directions §6.2: (1) more efficient feature representations and backbones, (2) improved temporal modeling, and (3) hybrid architectures combining bottom-up scaling with top-down accuracy. While accuracy remains critical, latency, scalability and robustness challenges persist. Privacy and bias risks require careful consideration as markerless capture expands across disciplines from biomechanics to entertainment. In the following section, we propose an overview of the future direction of the markerless field.

6.1 Applications and future impact

While real-time multi-view multi-person systems are still limited in deployment, their increasing real-world applications drive research advances across domains.

Emerging applications Reviews by Sarafianos et al. [125] and Wang et al. [158] §2.3.3 identify key applications in behavior analysis, security, healthcare, sports, and entertainment. The markerless approach enables natural environment deployment, though cost remains a barrier - single/dual camera setups offer a more practical alternative to full studio systems.

Human-computer interaction Real-time markerless tracking enables natural full-body VR/AR control [7, 136] with sub-40ms latency for responsive embodiment. In robotics, it enhances human-robot interaction through accurate gesture interpretation and motion planning.

Real-time multi-person markerless tracking enables natural group interactions in VR/gaming. Low latency is essential for immersion [56], with multi-person tracking enabling social presence and collaborative experiences beyond single-user systems.

Body analysis for sports and healthcare Recent markerless systems enable novel applications in sports science and healthcare, combining accessibility with growing precision for biomechanical analysis. In sports, real-time tracking provides biomechanical feedback for performance optimization and injury prevention [5, 153]. Recently, [46] proposed joint-torques optimization using *Mujoco* [142] physics to refine 2-person capture, combining multi-view *YOLOv8* §2.2.1 and *ViT-Pose* [172] for 3D joints, with *SMPL* [92] §3.1.2 body fitting to compute physics-based joint-torques. While not real-time, this method shows the interest of realistic physics-based filtering for a high velocity use-case.

For healthcare applications, while early systems lacked clinical precision [32] §2.3.2, recent methods approach marker-based accuracy. D'Haene et al. [38] validated a 3-camera Stereolab's ZED2 setup with *YOLOv8x-pose-p6* §2.2.1 against *OptiTrack*, achieving RMSE $< 5°$ for hip/knee angles in gait analysis. While some constant biases remained due to sampling rate differences and kinematic computation methods, their results demonstrated the potential of markerless systems as accessible alternatives for clinical gait analysis.

Societal and ethical considerations Real-time markerless motion capture raises privacy concerns due to its ability to operate without consent. Motion data can reveal identifying biometric characteristics [168, 144]. Integration with autonomous systems like UAVs [122] further emphasizes the need for ethical frameworks governing deployment and transparency. Open-source markerless systems based of state-of-the-art research show the potential of the technology and support the ethical development of the field.

Adversarial attack on neural-networks Pose estimation networks show vulnerability to adversarial attacks. Liu et al. [88] demonstrated fooling action recognition by modifying *VNect* [95] pose outputs. [127] found heatmap-based architectures more robust than direct regression, though perturbed *HigherHRNet* [30] could generate false joints and hallucinated limbs. *Bottom-up* methods proved vulnerable due to dual attack surfaces, but simple defenses like image flipping showed effectiveness. These insights are crucial for developing more robust markerless systems, particularly for safety-critical applications.

Bias and Fairness Markerless systems face challenges with algorithmic bias and fairness due to training data imbalances §3.1.3. Current fairness research has been conducted on Deep Learning-based vision recognition systems [15, 163, 57, 4, 169]. These algorithms are a building block of human pose estimation, and progress in the field should limit issues in the markerless tracking. Analysis of datasets like *COCO* [87] §3.2.1 shows significant demographic skews [80], impacting detection accuracy for underrepresented groups like females, darker-skinned and older individuals. These biases directly impact system performance, leading to degraded accuracy and detection failures for underrepresented groups. This has far-reaching implications for applications like healthcare diagnostics and human-computer interaction. The field is addressing this through transparency frameworks like *model cards* [99] and dataset datasheets [50]. For example, *MoveNet* [152] documents performance variations across demographics in its model card.

6.2 Future research directions

Our comprehensive review reveals several key areas that will shape the future of markerless motion capture.

Machine Learning innovations Deep learning has fundamentally transformed markerless pose estimation, with three key innovations poised to drive future progress: (1) Transformer architectures §4.5.3 enable powerful multi-view pose

estimation (*MvP* [160], *Epipolar Transformers* [60]), though GPU optimization remains needed, (2) large synthetic datasets [148, 43] §3.2.1 with domain adaptation §3.1.3 enable training without manual annotation, (3) standardized evaluation protocols §3.2.2 are needed to systematically compare architectures across views and subjects.

Feature representation for pose estimation Our analysis reveals key architectural paradigms for multi-view pose estimation, each with distinct trade-offs: Volumetric approaches fuse multi-view information in voxel-space for superior accuracy but higher compute cost. Recent hybrid CNN-transformer architectures with temporal consistency (e.g., TEMPO [31]) achieve real-time performance. Top-down methods excel with few viewpoints by leveraging high resolution, but lose scene context through person-centric cropping. Bottom-up methods maintain global understanding with fixed computational scaling. Novel volumetric representations like Gaussian splatting [93] achieve 22.1mm MPJPE on *CMU Panoptic*, revisiting earlier work [138] with modern deep learning. The key challenge remains balancing accuracy, computational efficiency, and robustness across varying scene complexities.

Acknowledgments

The authors would like to thank the Artanim Foundation for the funding, the computing resources and the working conditions that made the writing of this large technical review possible.

References

[1] Jake K. Aggarwal and Quin Cai. Human motion analysis: A review. *Proceedings IEEE Nonrigid and Articulated Motion Workshop*, pages 90–102, 1997. doi:10.1109/NAMW.1997.609859. (Cited on pages 12, 13, and 14.)

[2] Jake K. Aggarwal, Quin Cai, Wen Liao, and Bikash Sabata. Articulated and elastic non-rigid motion: A review. *Proceedings of 1994 IEEE Workshop on Motion of Non-rigid and Articulated Objects*, pages 2–14, 1994. doi:10.1109/MNRAO.1994.346261. (Cited on pages 12 and 13.)

[3] Jake K. Aggarwal, Quin Cai, Wen Liao, and Bikash Sabata. Nonrigid Motion Analysis: Articulated and Elastic Motion. *Computer Vision and Image Understanding*, 70(2):142–156, May 1998. ISSN 1077-3142. doi:10.1006/cviu.1997.0620. (Cited on page 13.)

[4] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 249–260, New York, NY, USA, March 2021. Association for Computing Machinery. ISBN 978-1-4503-8309-7. doi:10.1145/3442188.3445888. (Cited on page 28.)

[5] Cortney Armitano-Lago, Dominic Willoughby, and Adam W. Kiefer. A SWOT Analysis of Portable and Low-Cost Markerless Motion Capture Systems to Assess Lower-Limb Musculoskeletal Kinematics in Sport. *Frontiers in Sports and Active Living*, 3, 2022. ISSN 2624-9367. (Cited on page 28.)

[6] Ali Azarbayejani and Alex Pentland. Real-time self-calibrating stereo person tracking using 3-D shape estimation from blob features. In *Proceedings of 13th International Conference on Pattern Recognition*, volume 3, pages 627–632 vol.3, August 1996. doi:10.1109/ICPR.1996.547022. (Cited on pages 10, 11, 13, and 18.)

[7] Huidong Bai, Lei Gao, Jihad El-Sana, and Mark Billinghurst. Markerless 3D gesture-based interaction for handheld augmented reality interfaces. In *SIGGRAPH Asia 2013 Symposium on Mobile Graphics and Interactive Applications*, SA '13, page 1, New York, NY, USA, November 2013. Association for Computing Machinery. ISBN 978-1-4503-2633-9. doi:10.1145/2543651.2543678. (Cited on page 28.)

[8] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving Deeper into Convolutional Networks for Learning Video Representations. (arXiv:1511.06432), March 2016. doi:10.48550/arXiv.1511.06432. (Cited on page 22.)

[9] Eduard Gabriel Bazavan, Andrei Zanfir, Mihai Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. HSPACE: Synthetic Parametric Humans Animated in Complex Environments. (arXiv:2112.12867), January 2022. doi:10.48550/arXiv.2112.12867. (Cited on page 16.)

[10] Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 3D Pictorial Structures for Multiple Human Pose Estimation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1669–1676, Columbus, OH, USA, June 2014. IEEE. ISBN 978-1-4799-5118-5. doi:10.1109/CVPR.2014.216. (Cited on pages 12, 15, 16, 17, 21, and 22.)

[11] Vasileios Belagiannis, Xinchao Wang, Bernt Schiele, Pascal Fua, Slobodan Ilic, and Nassir Navab. Multiple Human Pose Estimation with Temporally Consistent 3D Pictorial Structures. In Lourdes Agapito, Michael M. Bronstein, and Carsten Rother, editors, *Computer Vision - ECCV 2014 Workshops*, Lecture Notes in Computer Science, pages 742–754, Cham, 2015. Springer International Publishing. ISBN 978-3-319-16178-5. doi:10.1007/978-3-319-16178-5$_5$2. (Cited on pages 12 and 18.)

[12] Daniel Bermuth, Alexander Poeppel, and Wolfgang Reif. Voxelkeypointfusion: Generalizable multi-view multi-person pose estimation. 2024. doi:10.48550/arXiv.2410.18723. (Cited on page 21.)

[13] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple Online and Realtime Tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3464–3468, September 2016. doi:10.1109/ICIP.2016.7533003. (Cited on page 22.)

[14] 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. (arXiv:1607.08128), July 2016. doi:10.48550/arXiv.1607.08128. (Cited on pages 13 and 14.)

[15] Joy Buolamwini and Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, January 2018. (Cited on page 28.)

[16] Magnus Burenius, Josephine Sullivan, and Stefan Carlsson. 3D Pictorial Structures for Multiple View Articulated Pose Estimation. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3618–3625, June 2013. doi:10.1109/CVPR.2013.464. (Cited on pages 11, 12, 13, 17, and 23.)

[17] Qingyuan Cai, Xuecai Hu, Saihui Hou, Li Yao, and Yongzhen Huang. Disentangled diffusion-based 3d human pose estimation with hierarchical spatial and temporal denoiser. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(2):882–890, March 2024. ISSN 2159-5399. doi:10.1609/aaai.v38i2.27847. URL http://dx.doi.org/10.1609/aaai.v38i2.27847. (Cited on page 12.)

[18] Yanlu Cai, Weizhong Zhang, Yuan Wu, and Cheng Jin. Fusionformer: a concise unified feature fusion transformer for 3d pose estimation. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi:10.1609/aaai.v38i2.27849. URL https://doi.org/10.1609/aaai.v38i2.27849. (Cited on page 20.)

[19] Fabrice Caillette, Aphrodite Galata, and Toby Howard. Real-time 3-D human body tracking using learnt models of behaviour. *Computer Vision and Image Understanding*, 109(2):112–125, February 2008. ISSN 1077-3142. doi:10.1016/j.cviu.2007.05.005. (Cited on pages 10, 11, and 13.)

[20] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1302–1310, July 2017. doi:10.1109/CVPR.2017.143. (Cited on pages 10, 11, and 13.)

[21] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *arXiv:1812.08008 [cs]*, May 2019. (Cited on pages 10, 11, 14, 15, 16, 20, and 24.)

[22] Claudette Cédras and Mubarak Shah. Motion-based recognition a survey. *Image and Vision Computing*, 13(2): 129–155, March 1995. ISSN 0262-8856. doi:10.1016/0262-8856(95)93154-K. (Cited on page 13.)

[23] Haoming Chen, Runyang Feng, Sifan Wu, Hao Xu, Fengcheng Zhou, and Zhenguang Liu. 2D Human Pose Estimation: A Survey. *arXiv:2204.07370 [cs]*, April 2022. (Cited on pages 12 and 13.)

[24] Long Chen, Haizhou Ai, Rui Chen, Zijie Zhuang, and Shuang Liu. Cross-View Tracking for Multi-Human 3D Pose Estimation at over 100 FPS. (arXiv:2003.03972), July 2021. doi:10.48550/arXiv.2003.03972. (Cited on pages 13 and 18.)

[25] Xianjie Chen and Alan Yuille. Articulated Pose Estimation by a Graphical Model with Image Dependent Pairwise Relations. (arXiv:1407.3399), November 2014. doi:10.48550/arXiv.1407.3399. (Cited on page 13.)

[26] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded Pyramid Network for Multi-Person Pose Estimation. (arXiv:1711.07319), April 2018. doi:10.48550/arXiv.1711.07319. (Cited on pages 12 and 20.)

[27] Yucheng Chen, Yingli Tian, and Mingyi He. Monocular Human Pose Estimation: A Survey of Deep Learning-based Methods. *Computer Vision and Image Understanding*, 192:102897, March 2020. ISSN 10773142. doi:10.1016/j.cviu.2019.102897. (Cited on pages 12, 13, and 14.)

[28] Yuxing Chen, Renshu Gu, Ouhan Huang, and Gangyong Jia. VTP: Volumetric Transformer for Multi-view Multi-person 3D Pose Estimation. (arXiv:2205.12602), May 2022. doi:10.48550/arXiv.2205.12602. (Cited on pages 24 and 25.)

[29] Z. Chen and H.-J. Lee. Knowledge-guided visual perception of 3-D human gait from a single image sequence. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(2):336–342, March-April/1992. ISSN 00189472. doi:10.1109/21.148408. (Cited on page 13.)

[30] 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. *arXiv:1908.10357 [cs, eess]*, March 2020. (Cited on pages 11, 15, and 28.)

[31] Rohan Choudhury, Kris Kitani, and Laszlo A. Jeni. TEMPO: Efficient Multi-View Pose Estimation, Tracking, and Forecasting. (arXiv:2309.07910), September 2023. doi:10.48550/arXiv.2309.07910. (Cited on pages 15, 18, 22, 23, 24, 25, 26, 27, and 29.)

[32] Steffi L. Colyer, Murray Evans, Darren P. Cosker, and Aki I. T. Salo. A Review of the Evolution of Vision-Based Motion Analysis and the Integration of Advanced Computer Vision Methods Towards Developing a Markerless System. *Sports Medicine - Open*, 4(1):24, June 2018. ISSN 2198-9761. doi:10.1186/s40798-018-0139-y. (Cited on pages 12, 13, and 28.)

[33] XRMoCap Contributors. Openxrlab multi-view motion capture toolbox and benchmark. https://github.com/openxrlab/xrmocap, 2022. (Cited on pages 18 and 24.)

[34] Mohamed Dahmane and Jean Meunier. Real-time video surveillance with self-organizing maps. In *The 2nd Canadian Conference on Computer and Robot Vision (CRV'05)*, pages 136–143, May 2005. doi:10.1109/CRV.2005.65. (Cited on page 13.)

[35] Naoto Date, Hiromasa Yoshimoto, Daisaku Arita, and Rin-Ichiro Taniguchi. Real-time human motion sensing based on vision-based inverse kinematics for interactive applications. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 318–321 Vol.3, August 2004. doi:10.1109/ICPR.2004.1334531. (Cited on pages 10, 11, and 13.)

[36] Junli Deng, Haoyuan Yao, and Ping Shi. Enhanced 3d pose estimation in multi-person, multi-view scenarios through unsupervised domain adaptation with dropout discriminator. *Sensors*, 23(20), 2023. ISSN 1424-8220. doi:10.3390/s23208406. URL https://www.mdpi.com/1424-8220/23/20/8406. (Cited on page 15.)

[37] Yann Desmarais, Denis Mottet, Pierre Slangen, and Philippe Montesinos. A review of 3D human pose estimation algorithms for markerless motion capture. (arXiv:2010.06449), July 2021. doi:10.48550/arXiv.2010.06449. (Cited on pages 12 and 13.)

[38] Mathis D'Haene, Frédéric Chorin, Serge S. Colson, Olivier Guérin, Raphaël Zory, and Elodie Piche. Validation of a 3d markerless motion capture tool using multiple pose and depth estimations for quantitative gait analysis. *Sensors*, 24(22), 2024. ISSN 1424-8220. doi:10.3390/s24227105. URL https://www.mdpi.com/1424-8220/24/22/7105. (Cited on page 28.)

[39] Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and Robust Multi-Person 3D Pose Estimation from Multiple Views. (arXiv:1901.04111), January 2019. doi:10.48550/arXiv.1901.04111. (Cited on pages 12, 13, 18, 19, 21, 24, and 25.)

[40] Shradha Dubey and Manish Dixit. A comprehensive survey on human pose estimation approaches. *Multimedia Systems*, 29(1):167–195, February 2023. ISSN 1432-1882. doi:10.1007/s00530-022-00980-0. (Cited on pages 12 and 13.)

[41] A. Elhayek, E. De Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt. Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3810–3818, Boston, MA, USA, June 2015. IEEE. ISBN 978-1-4673-6964-0. doi:10.1109/CVPR.2015.7299005. (Cited on page 18.)

[42] Alessio Elmi, Davide Mazzini, and Pietro Tortella. Light3DPose: Real-time Multi-Person 3D PoseEstimation from Multiple Views. (arXiv:2004.02688), April 2020. doi:10.48550/arXiv.2004.02688. (Cited on page 21.)

[43] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Andrea Palazzi, Roberto Vezzani, and Rita Cucchiara. Learning to Detect and Track Visible and Occluded Body Joints in a Virtual World. (arXiv:1803.08319), September 2018. doi:10.48550/arXiv.1803.08319. (Cited on pages 16 and 29.)

[44] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2353–2362, 2017. doi:10.1109/ICCV.2017.256. (Cited on page 11.)

[45] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7157–7173, June 2023. ISSN 1939-3539. doi:10.1109/tpami.2022.3222784. URL http://dx.doi.org/10.1109/TPAMI.2022.3222784. (Cited on page 11.)

[46] Hossein Feiz, David Labbé, and Sheldon Andrews. Markerless Multi-view Multi-person Tracking for Combat Sports . In Victor Zordan, editor, *Eurographics/ ACM SIGGRAPH Symposium on Computer Animation - Posters*. The Eurographics Association, 2024. ISBN 978-3-03868-263-9. doi:10.2312/sca.20241162. (Cited on page 28.)

[47] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. (Cited on pages 16 and 22.)

[48] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-Dimensional Reconstruction of Human Interactions. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7212–7221, June 2020. doi:10.1109/CVPR42600.2020.00724. (Cited on page 16.)

[49] D. Gavrila and L. Davis. Towards 3-D model-based tracking and recognition of human movement: A multi-view approach. 1995. (Cited on page 10.)

[50] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for Datasets. (arXiv:1803.09010), December 2021. doi:10.48550/arXiv.1803.09010. (Cited on page 28.)

[51] Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13041–13051. IEEE, June 2023. doi:10.1109/cvpr52729.2023.01253. URL http://dx.doi.org/10.1109/CVPR52729.2023.01253. (Cited on page 12.)

[52] Wenjuan Gong, Xuena Zhang, Jordi Gonzàlez, Andrews Sobral, Thierry Bouwmans, Changhe Tu, and El-hadi Zahzah. Human Pose Estimation from Monocular Images: A Comprehensive Survey. *Sensors*, 16(12):1966, December 2016. ISSN 1424-8220. doi:10.3390/s16121966. (Cited on pages 12 and 13.)

[53] Hengkai Guo, Tang Tang, Guozhong Luo, Riwei Chen, Yongchen Lu, and Linfu Wen. Multi-Domain Pose Network for Multi-Person Pose Estimation and Tracking. 11130:209–216, 2019. doi:10.1007/978-3-030-11012-$3_1$7. (Cited on page 15.)

[54] Wen Guo, Xiaoyu Bie, Francesc Moreno-Noguer, and Xavier Alameda-Pineda. ExPI Dataset, October 2021. (Cited on page 16.)

[55] Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-Person Extreme Motion Prediction. (arXiv:2105.08825), June 2022. doi:10.48550/arXiv.2105.08825. (Cited on page 16.)

[56] Eunchong Ha, Gongkyu Byeon, and Sunjin Yu. Full-Body Motion Capture-Based Virtual Reality Multi-Remote Collaboration System. *Applied Sciences*, 12(12):5862, January 2022. ISSN 2076-3417. doi:10.3390/app12125862. (Cited on page 28.)

[57] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. Towards a Critical Race Methodology in Algorithmic Fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 501–512, January 2020. doi:10.1145/3351095.3372826. (Cited on page 28.)

[58] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. URL https://arxiv.org/abs/1512.03385. (Cited on pages 11, 20, 23, 24, and 27.)

[59] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, October 2017. doi:10.1109/ICCV.2017.322. (Cited on page 11.)

[60] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I. Yu. Epipolar Transformers. (arXiv:2005.04551), May 2020. doi:10.48550/arXiv.2005.04551. (Cited on pages 13, 23, and 29.)

[61] Gines Hidalgo, Yaadhav Raaj, Haroon Idrees, Donglai Xiang, Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Single-Network Whole-Body Pose Estimation. *arXiv:1909.13423 [cs]*, September 2019. (Cited on page 14.)

[62] Michael B. Holte, Cuong Tran, Mohan M. Trivedi, and Thomas B. Moeslund. Human Pose Estimation and Activity Recognition From Multi-View Videos: Comparative Explorations of Recent Developments. *IEEE Journal of Selected Topics in Signal Processing*, 6(5):538–552, September 2012. ISSN 1941-0484. doi:10.1109/JSTSP.2012.2196975. (Cited on pages 12 and 13.)

[63] Mir Rayat Imtiaz Hossain and James J. Little. Exploiting temporal information for 3D pose estimation. volume 11214, pages 69–86. 2018. doi:10.1007/978-3-030-01249-6$_5$. (Cited on page 13.)

[64] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. 2017. (Cited on pages 15 and 21.)

[65] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J. Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. 2018. URL https://arxiv.org/abs/1810.04703. (Cited on page 12.)

[66] 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, July 2014. ISSN 1939-3539. doi:10.1109/TPAMI.2013.248. (Cited on pages 13, 15, 16, 20, 21, and 22.)

[67] Michael Isard and John MacCormick. BraMBLe: A Bayesian multiple-blob tracker. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 34–41 vol.2, July 2001. doi:10.1109/ICCV.2001.937594. (Cited on pages 10 and 11.)

[68] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable Triangulation of Human Pose. (arXiv:1905.05754), May 2019. doi:10.48550/arXiv.1905.05754. (Cited on pages 21, 23, and 24.)

[69] Xiaofei Ji and Honghai Liu. Advances in View-Invariant Human Motion Analysis: A Review. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(1):13–24, January 2010. ISSN 1558-2442. doi:10.1109/TSMCC.2009.2027608. (Cited on pages 12, 13, and 17.)

[70] 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. 2023. doi:10.48550/arXiv.2303.07399. (Cited on page 21.)

[71] Sam Johnson and Mark Everingham. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. In *Procedings of the British Machine Vision Conference 2010*, pages 12.1–12.11, Aberystwyth, 2010. British Machine Vision Association. ISBN 978-1-901725-40-7. doi:10.5244/C.24.12. (Cited on page 13.)

[72] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic Studio: A Massively Multiview System for Social Interaction Capture. (arXiv:1612.03153), December 2016. doi:10.48550/arXiv.1612.03153. (Cited on pages 11, 13, 15, 16, 18, 20, 22, 23, and 24.)

[73] Abdolrahim Kadkhodamohammadi and Nicolas Padoy. A generalizable approach for multi-view 3D human pose regression. (arXiv:1804.10462), October 2019. (Cited on page 18.)

[74] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Self-Supervised Learning of 3D Human Pose using Multi-view Geometry. (arXiv:1903.02330), April 2019. doi:10.48550/arXiv.1903.02330. (Cited on pages 13 and 20.)

[75] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video Inference for Human Body Pose and Shape Estimation. (arXiv:1912.05656), April 2020. doi:10.48550/arXiv.1912.05656. (Cited on page 14.)

[76] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. PifPaf: Composite Fields for Human Pose Estimation. (arXiv:1903.06593), April 2019. doi:10.48550/arXiv.1903.06593. (Cited on pages 11 and 23.)

[77] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. OpenPifPaf: Composite Fields for Semantic Keypoint Detection and Spatio-Temporal Association. *arXiv:2103.02440 [cs]*, September 2021. (Cited on page 11.)

[78] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. (Cited on page 11.)

[79] Joseph B. Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7(1):48–50, 1956. ISSN 0002-9939, 1088-6826. doi:10.1090/S0002-9939-1956-0078686-7. (Cited on page 19.)

[80] Julienne LaChance, William Thong, and Shruti Nagpal Alice Xiang. A case study in fairness evaluation: Current limitations and challenges for human pose estimation. In *AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI on (R${\hat{}}{2}$HCAI)*, 2023. (Cited on page 28.)

[81] Winnie W. T. Lam, Yuk Ming Tang, and Kenneth N. K. Fong. A systematic review of the applications of markerless motion capture (MMC) technology for clinical measurement in rehabilitation. *Journal of NeuroEngineering and Rehabilitation*, 20(1):57, May 2023. ISSN 1743-0003. doi:10.1186/s12984-023-01186-9. (Cited on pages 12 and 13.)

[82] Hsi-Jian Lee and Zen Chen. Determination of 3D human body postures from a single view. *Computer Vision, Graphics, and Image Processing*, 30(2):148–168, May 1985. ISSN 0734189X. doi:10.1016/0734-189X(85)90094-5. (Cited on page 10.)

[83] Chen Li and Gim Hee Lee. From Synthetic to Real: Unsupervised Domain Adaptation for Animal Pose Estimation. (arXiv:2103.14843), March 2021. doi:10.48550/arXiv.2103.14843. (Cited on page 15.)

[84] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. CrowdPose: Efficient Crowded Scenes Pose Estimation and A New Benchmark. (arXiv:1812.00324), January 2019. doi:10.48550/arXiv.1812.00324. (Cited on page 16.)

[85] Junbang Liang and Ming C. Lin. Shape-Aware Human Pose and Shape Reconstruction Using Multi-View Images. (arXiv:1908.09464), August 2019. doi:10.48550/arXiv.1908.09464. (Cited on page 14.)

[86] Jiahao Lin and Gim Hee Lee. Multi-View Multi-Person 3D Pose Estimation with Plane Sweep Stereo. (arXiv:2104.02273), April 2021. doi:10.48550/arXiv.2104.02273. (Cited on pages 23, 24, 25, 26, and 27.)

[87] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context. (arXiv:1405.0312), February 2015. doi:10.48550/arXiv.1405.0312. (Cited on pages 11, 15, 16, and 28.)

[88] Jian Liu, Naveed Akhtar, and Ajmal Mian. Adversarial Attack on Skeleton-based Human Action Recognition. (arXiv:1909.06500), September 2019. doi:10.48550/arXiv.1909.06500. (Cited on pages 13 and 28.)

[89] Ruixu Liu, Ju Shen, He Wang, Chen Chen, Sen-ching Cheung, and Vijayan Asari. Attention Mechanism Exploits Temporal Contexts: Real-Time 3D Human Pose Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5064–5073, August 2020. (Cited on page 13.)

[90] Wu Liu, Qian Bao, Yu Sun, and Tao Mei. Recent Advances of Monocular 2D and 3D Human Pose Estimation: A Deep Learning Perspective. *ACM Computing Surveys*, 55(4):80:1–80:41, November 2022. ISSN 0360-0300. doi:10.1145/3524497. (Cited on pages 12 and 13.)

[91] Zhao Liu, Jianke Zhu, Jiajun Bu, and Chun Chen. A survey of human pose estimation: The body parts parsing based methods. *Journal of Visual Communication and Image Representation*, 32:10–19, October 2015. ISSN 10473203. doi:10.1016/j.jvcir.2015.06.013. (Cited on pages 12 and 13.)

[92] 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, October 2015. doi:10.1145/2816795.2818013. (Cited on pages 11, 14, 16, 17, 19, and 28.)

[93] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3D Gaussians: Tracking by Persistent Dynamic View Synthesis. (arXiv:2308.09713), August 2023. doi:10.48550/arXiv.2308.09713. (Cited on page 29.)

[94] 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 *International Conference on 3D Vision (3DV 2017)*, 2017. (Cited on page 15.)

[95] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera. In *ACM Transactions on Graphics (SIGGRAPH 2017)*, 2017. (Cited on pages 12 and 28.)

[96] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-Shot Multi-Person 3D Pose Estimation From Monocular RGB. (arXiv:1712.03453), August 2018. doi:10.48550/arXiv.1712.03453. (Cited on pages 12 and 16.)

[97] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. XNect: Real-time Multi-Person 3D Motion Capture with a Single RGB Camera. *ACM Transactions on Graphics*, 39(4), August 2020. ISSN 0730-0301, 1557-7368. doi:10.1145/3386569.3392410. (Cited on pages 12 and 15.)

[98] Pierre Merriaux, Yohan Dupuis, Rémi Boutteau, Pascal Vasseur, and Xavier Savatier. A study of vicon system positioning performance. *Sensors*, 17(7), 2017. ISSN 1424-8220. doi:10.3390/s17071591. URL https://www.mdpi.com/1424-8220/17/7/1591. (Cited on page 14.)

[99] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 220–229, New York, NY, USA, January 2019. Association for Computing Machinery. ISBN 978-1-4503-6125-5. doi:10.1145/3287560.3287596. (Cited on page 28.)

[100] Thomas B. Moeslund and Erik Granum. A Survey of Computer Vision-Based Human Motion Capture. *Computer Vision and Image Understanding*, 81(3):231–268, March 2001. ISSN 1077-3142. doi:10.1006/cviu.2000.0897. (Cited on pages 12 and 15.)

[101] Thomas B. Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2):90–126, November 2006. ISSN 1077-3142. doi:10.1016/j.cviu.2006.08.002. (Cited on pages 12 and 13.)

[102] Tewodros Legesse Munea, Yalew Zelalem Jembre, Halefom Tekle Weldegebriel, Longbiao Chen, Chenxi Huang, and Chenhui Yang. The Progress of Human Pose Estimation: A Survey and Taxonomy of Models Applied in 2D Human Pose Estimation. *IEEE Access*, 8:133330–133348, July 2020. ISSN 2169-3536. doi:10.1109/ACCESS.2020.3010248. (Cited on page 12.)

[103] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative Embedding: End-to-End Learning for Joint Detection and Grouping. (arXiv:1611.05424), June 2017. doi:10.48550/arXiv.1611.05424. (Cited on pages 11, 14, and 15.)

[104] Ana Filipa Rodrigues Nogueira, Hélder P. Oliveira, and Luís F. Teixeira. Markerless multi-view 3d human pose estimation: a survey. 2024. doi:10.48550/arXiv.2407.03817. (Cited on pages 12 and 13.)

[105] Ahmed A. A. Osman, Timo Bolkart, and Michael J. Black. *STAR: Sparse Trained Articulated Human Body Regressor*, pages 598–613. Springer International Publishing, 2020. ISBN 9783030585396. doi:10.1007/978-3-030-58539-6$_3$6. URL http://dx.doi.org/10.1007/978-3-030-58539-6_36. (Cited on page 14.)

[106] Daniil Osokin. Real-time 2D Multi-Person Pose Estimation on CPU: Lightweight OpenPose. *arXiv:1811.12004 [cs]*, November 2018. (Cited on page 21.)

[107] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10967–10977. IEEE, June 2019. doi:10.1109/cvpr.2019.01123. URL http://dx.doi.org/10.1109/CVPR.2019.01123. (Cited on page 14.)

[108] Xavier Perez-Sala, Sergio Escalera, Cecilio Angulo, and Jordi Gonzàlez. A Survey on Model Based Approaches for 2D and 3D Visual Human Pose Recovery. *Sensors*, 14(3):4189–4210, March 2014. ISSN 1424-8220. doi:10.3390/s140304189. (Cited on page 12.)

[109] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation. (arXiv:1511.06645), April 2016. doi:10.48550/arXiv.1511.06645. (Cited on page 17.)

[110] Ronald Poppe. Vision-based human motion analysis: An overview. *Computer Vision and Image Understanding*, 108(1):4–18, October 2007. ISSN 1077-3142. doi:10.1016/j.cviu.2006.10.016. (Cited on pages 12, 13, and 14.)

[111] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. Cross View Fusion for 3D Human Pose Estimation. (arXiv:1909.01203), September 2019. doi:10.48550/arXiv.1909.01203. (Cited on page 23.)

[112] Yaadhav Raaj, Haroon Idrees, Gines Hidalgo, and Yaser Sheikh. Efficient Online Multi-Person 2D Pose Tracking with Recurrent Spatio-Temporal Affinity Fields. *arXiv:1811.11975 [cs]*, June 2019. (Cited on page 11.)

[113] Deva Ramanan and David A. Forsyth. Finding and tracking people from the bottom up. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–II, June 2003. doi:10.1109/CVPR.2003.1211504. (Cited on page 11.)

[114] Deva Ramanan, David A. Forsyth, and Andrew Zisserman. Strike a pose: Tracking people by finding stylized poses. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 271–278 vol. 1, June 2005. doi:10.1109/CVPR.2005.335. (Cited on page 11.)

[115] N Dinesh Reddy, Laurent Guigues, Leonid Pishchulin, Jayan Eledath, and Srinivasa G. Narasimhan. TesseTrack: End-to-End Learnable Multi-Person Articulated 3D Pose Tracking. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15185–15195, Nashville, TN, USA, June 2021. IEEE. ISBN 978-1-66544-509-2. doi:10.1109/CVPR46437.2021.01494. (Cited on pages 13, 21, and 22.)

[116] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. (arXiv:1506.02640), May 2016. doi:10.48550/arXiv.1506.02640. (Cited on page 11.)

[117] Edoardo Remelli, Shangchen Han, Sina Honari, Pascal Fua, and Robert Wang. Lightweight Multi-View 3D Pose Estimation through Camera-Disentangled Representation. *arXiv:2004.02186 [cs]*, June 2020. (Cited on page 13.)

[118] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 91–99, Cambridge, MA, USA, 2015. MIT Press. (Cited on page 11.)

[119] Kathleen Robinette, Sherri Blackwell, Hein Daanen, Mark Boehmer, Scott Fleming, Tina Brill, David Hoeferlin, and Dennis Burnsides. Civilian American and European Surface Anthropometry Resource (CAESAR) Final Report AFRL-HE- WP-TR-2002-0169. Technical report, US Air Force Research Laboratory, 2002. (Cited on page 14.)

[120] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Trans. Graph.*, 36(6), November 2017. ISSN 0730-0301. doi:10.1145/3130800.3130883. URL https://doi.org/10.1145/3130800.3130883. (Cited on page 14.)

[121] Ròmer Rosales and Stan Sclaroff. Learning and synthesizing human body motion and posture. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 506–511, March 2000. doi:10.1109/AFGR.2000.840681. (Cited on page 13.)

[122] Nitin Saini, Elia Bonetto, Eric Price, Aamir Ahmad, and Michael J. Black. AirPose: Multi-View Fusion Network for Aerial 3D Human Pose and Shape Estimation. *IEEE Robotics and Automation Letters*, 7(2):4805–4812, April 2022. ISSN 2377-3766, 2377-3774. doi:10.1109/LRA.2022.3145494. (Cited on page 28.)

[123] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. 2018. (Cited on page 21.)

[124] Ben Sapp and Ben Taskar. MODEC: Multimodal Decomposable Models for Human Pose Estimation. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3681, Portland, OR, USA, June 2013. IEEE. ISBN 978-0-7695-4989-7. doi:10.1109/CVPR.2013.471. (Cited on page 13.)

[125] Nikolaos Sarafianos, Bogdan Boteanu, Bogdan Ionescu, and Ioannis A. Kakadiaris. 3D Human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding*, 152:1–20, November 2016. ISSN 1077-3142. doi:10.1016/j.cviu.2016.09.002. (Cited on pages 12, 13, 16, and 28.)

[126] Steven Schwarcz and Thomas Pollard. 3D Human Pose Estimation from Deep Multi-View 2D Pose. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2326–2331, August 2018. doi:10.1109/ICPR.2018.8545631. (Cited on page 18.)

[127] Sahil Shah, Naman Jain, Abhishek Sharma, and Arjun Jain. On the Robustness of Human Pose Estimation. (arXiv:1908.06401), June 2021. doi:10.48550/arXiv.1908.06401. (Cited on pages 13 and 28.)

[128] Dahu Shi, Xing Wei, Liangqi Li, Ye Ren, and Wenming Tan. End-to-end multi-person pose estimation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11069–11078, June 2022. (Cited on page 23.)

[129] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124, January 2013. ISSN 0001-0782. doi:10.1145/2398356.2398381. (Cited on pages 11 and 13.)

[130] 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, April 2023. ISSN 1939-3539. doi:10.1109/tpami.2022.3188716. URL http://dx.doi.org/10.1109/TPAMI.2022.3188716. (Cited on page 20.)

[131] Qing Shuai, Zhiyuan Yu, Zhize Zhou, Lixin Fan, Haijun Yang, Can Yang, and Xiaowei Zhou. Reconstructing close human interactions from multiple views. *ACM Transactions on Graphics*, 42(6):1–14, December 2023. ISSN 1557-7368. doi:10.1145/3618336. URL http://dx.doi.org/10.1145/3618336. (Cited on page 22.)

[132] 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, March 2010. ISSN 1573-1405. doi:10.1007/s11263-009-0273-6. (Cited on pages 13 and 18.)

[133] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand Keypoint Detection in Single Images using Multiview Bootstrapping. (arXiv:1704.07809), April 2017. doi:10.48550/arXiv.1704.07809. (Cited on page 15.)

[134] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. (Cited on page 11.)

[135] Jucheng Song, Chi-Man Pun, Haolun Li, Rushi Lan, Jiu-Cheng Xie, and Hao Gao. Local optimization networks for multi-view multi-person human posture estimation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3995–3999, 2024. doi:10.1109/ICASSP48485.2024.10445922. (Cited on page 21.)

[136] Maurício Sousa, Daniel Mendes, Rafael Kuffner Dos Anjos, Daniel Medeiros, Alfredo Ferreira, Alberto Raposo, João Madeiras Pereira, and Joaquim Jorge. Creepy Tracker Toolkit for Context-aware Interfaces. In *Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces*, ISS '17, pages 191–200, New York, NY, USA, October 2017. Association for Computing Machinery. ISBN 978-1-4503-4691-7. doi:10.1145/3132272.3134113. (Cited on page 28.)

[137] Vinkle Srivastav, Keqi Chen, and Nicolas Padoy. Selfpose3d: Self-supervised multi-person multi-view 3d pose estimation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2502–2512. IEEE, June 2024. doi:10.1109/CVPR52733.2024.00242. (Cited on page 21.)

[138] Carsten Stoll, Nils Hasler, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt. Fast articulated motion tracking using a sums of Gaussians body model. In *2011 International Conference on Computer Vision*, pages 951–958, Barcelona, Spain, November 2011. IEEE. ISBN 978-1-4577-1102-2 978-1-4577-1101-5 978-1-4577-1100-8. doi:10.1109/ICCV.2011.6126338. (Cited on pages 18 and 29.)

[139] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep High-Resolution Representation Learning for Human Pose Estimation. (arXiv:1902.09212), February 2019. doi:10.48550/arXiv.1902.09212. (Cited on pages 11 and 15.)

[140] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J. Black, and Tao Mei. Monocular, One-stage, Regression of Multiple 3D People. (arXiv:2008.12272), September 2021. doi:10.48550/arXiv.2008.12272. (Cited on pages 11, 14, and 15.)

[141] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/tan19a.html. (Cited on page 15.)

[142] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033, 2012. doi:10.1109/IROS.2012.6386109. (Cited on page 28.)

[143] Jonathan Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. (arXiv:1406.2984), September 2014. doi:10.48550/arXiv.1406.2984. (Cited on pages 13 and 18.)

[144] Luke K. Topham, Wasiq Khan, Dhiya Al-Jumeily, and Abir Hussain. Human Body Pose Estimation for Gait Identification: A Comprehensive Survey of Datasets and Models. *ACM Computing Surveys*, 55(6):120:1–120:42, December 2022. ISSN 0360-0300. doi:10.1145/3533384. (Cited on page 28.)

[145] Alexander Toshev and Christian Szegedy. DeepPose: Human Pose Estimation via Deep Neural Networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1653–1660, June 2014. doi:10.1109/CVPR.2014.214. (Cited on pages 10 and 11.)

[146] 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. (Cited on pages 16 and 20.)

[147] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. VoxelPose: Towards Multi-Camera 3D Human Pose Estimation in Wild Environment. (arXiv:2004.06239), August 2020. doi:10.48550/arXiv.2004.06239. (Cited on pages 13, 14, 15, 16, 20, 21, 22, 24, 25, 26, and 27.)

[148] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from Synthetic Humans. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4627–4635, July 2017. doi:10.1109/CVPR.2017.492. (Cited on pages 16 and 29.)

[149] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *arXiv:1706.03762 [cs]*, June 2017. (Cited on page 23.)

[150] T. von Marcard, B. Rosenhahn, M. J. Black, and G. Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. *Computer Graphics Forum*, 36(2):349–360, 2017. doi:https://doi.org/10.1111/cgf.13131. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13131. (Cited on page 12.)

[151] 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)*, September 2018. (Cited on pages 12 and 16.)

[152] Ronny Votel and Na Li. Next-generation pose detection with movenet and TensorFlow. js. *TensorFlow Blog*, 4, 2021. (Cited on page 28.)

[153] Logan Wade, Laurie Needham, Polly McGuigan, and James Bilzon. Applications and limitations of current markerless motion capture methods for clinical gait biomechanics. *PeerJ*, 10:e12995, February 2022. ISSN 2167-8359. doi:10.7717/peerj.12995. (Cited on page 28.)

[154] Bastian Wandt and Bodo Rosenhahn. RepNet: Weakly Supervised Training of an Adversarial Reprojection Network for 3D Human Pose Estimation. (arXiv:1902.09868), March 2019. doi:10.48550/arXiv.1902.09868. (Cited on page 13.)

[155] 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. (arXiv:2207.02696), July 2022. doi:10.48550/arXiv.2207.02696. (Cited on page 11.)

[156] Guangming Wang, Honghao Zeng, Ziliang Wang, Zhe Liu, and Hesheng Wang. Motion projection consistency-based 3-d human pose estimation with virtual bones from monocular videos. *IEEE Transactions on Cognitive and Developmental Systems*, 15(2):784–793, June 2023. ISSN 2379-8939. doi:10.1109/tcds.2022.3185146. URL http://dx.doi.org/10.1109/TCDS.2022.3185146. (Cited on page 12.)

[157] Jiahang Wang, Sheng Jin, Wentao Liu, Weizhong Liu, Chen Qian, and Ping Luo. When Human Pose Estimation Meets Robustness: Adversarial Algorithms and Benchmarks. (arXiv:2105.06152), May 2021. doi:10.48550/arXiv.2105.06152. (Cited on page 15.)

[158] Jinbao Wang, Shujie Tan, Xiantong Zhen, Shuo Xu, Feng Zheng, Zhenyu He, and Ling Shao. Deep 3D human pose estimation: A review. *Computer Vision and Image Understanding*, 210:103225, September 2021. ISSN 1077-3142. doi:10.1016/j.cviu.2021.103225. (Cited on pages 12, 13, and 28.)

[159] Liang Wang, Weiming Hu, and Tieniu Tan. Recent developments in human motion analysis. *Pattern Recognition*, 36(3):585–601, March 2003. ISSN 0031-3203. doi:10.1016/S0031-3203(02)00100-0. (Cited on pages 12 and 13.)

[160] Tao Wang, Jianfeng Zhang, Yujun Cai, Shuicheng Yan, and Jiashi Feng. Direct Multi-view Multi-person 3D Pose Estimation. *arXiv:2111.04076 [cs]*, November 2021. (Cited on pages 13, 21, 23, 24, 25, 26, 27, and 29.)

[161] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional Pose Machines. (arXiv:1602.00134), April 2016. doi:10.48550/arXiv.1602.00134. (Cited on page 15.)

[162] Daniel Weinland, Remi Ronfard, and Edmond Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 104(2):249–257, November 2006. ISSN 1077-3142. doi:10.1016/j.cviu.2006.07.013. (Cited on page 13.)

[163] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. Predictive Inequity in Object Detection. (arXiv:1902.11097), February 2019. doi:10.48550/arXiv.1902.11097. (Cited on page 28.)

[164] Christopher R. Wren, Ali Azarbayejani, Trevor Darrell, and Alex Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, July 1997. ISSN 1939-3539. doi:10.1109/34.598236. (Cited on page 11.)

[165] Bo Wu and Ram Nevatia. Detection and Tracking of Multiple, Partially Occluded Humans by Bayesian Combination of Edgelet based Part Detectors. *International Journal of Computer Vision*, 75(2):247–266, November 2007. ISSN 1573-1405. doi:10.1007/s11263-006-0027-7. (Cited on pages 10 and 11.)

[166] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, Yizhou Wang, and Yonggang Wang. AI Challenger : A Large-scale Dataset for Going Deeper in Image Understanding. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1480–1485, July 2019. doi:10.1109/ICME.2019.00256. (Cited on page 16.)

[167] Size Wu, Sheng Jin, Wentao Liu, Lei Bai, Chen Qian, Dong Liu, and Wanli Ouyang. Graph-Based 3D Multi-Person Pose Estimation Using Multi-View Images. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11128–11137, Montreal, QC, Canada, October 2021. IEEE. ISBN 978-1-66542-812-5. doi:10.1109/ICCV48922.2021.01096. (Cited on pages 20, 24, 25, 26, and 27.)

[168] Zifeng Wu, Yongzhen Huang, Liang Wang, Xiaogang Wang, and Tieniu Tan. A Comprehensive Study on Cross-View Gait Based Human Identification with Deep CNNs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(2):209–226, February 2017. ISSN 1939-3539. doi:10.1109/TPAMI.2016.2545669. (Cited on page 28.)

[169] Alice Xiang. Being 'Seen' vs. 'Mis-Seen': Tensions between Privacy and Fairness in Computer Vision. (4068921), February 2022. doi:10.2139/ssrn.4068921. (Cited on page 28.)

[170] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3D Human Shape and Articulated Pose Models. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6183–6192, Seattle, WA, USA, June 2020. IEEE. ISBN 978-1-72817-168-5. doi:10.1109/CVPR42600.2020.00622. (Cited on pages 14, 16, and 17.)

[171] Xixia Xu, Qi Zou, and Xue Lin. Alleviating Human-level Shift : A Robust Domain Adaptation Method for Multi-person Pose Estimation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2326–2335, October 2020. doi:10.1145/3394171.3414040. (Cited on page 15.)

[172] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 38571–38584. Curran Associates, Inc., 2022. doi:10.48550/arXiv.2204.12484. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/fbb10d319d44f8c3b4720873e4177c65-Paper-Conference.pdf. (Cited on pages 20 and 28.)

[173] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose++: Vision transformer for generic body pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1212–1230, February 2024. ISSN 1939-3539. doi:10.1109/tpami.2023.3330016. URL http://dx.doi.org/10.1109/TPAMI.2023.3330016. (Cited on page 20.)

[174] Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR 2011*, pages 1385–1392, June 2011. doi:10.1109/CVPR.2011.5995741. (Cited on page 13.)

[175] 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. (arXiv:2303.17368), September 2023. doi:10.48550/arXiv.2303.17368. (Cited on page 16.)

[176] Hang Ye, Wentao Zhu, Chunyu Wang, Rujie Wu, and Yizhou Wang. Faster VoxelPose: Real-time 3D Human Pose Estimation by Orthographic Projection. (arXiv:2207.10955), July 2022. doi:10.48550/arXiv.2207.10955. (Cited on pages 13, 15, 22, 24, 25, 26, and 27.)

[177] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17016–17027. IEEE, June 2023. doi:10.1109/cvpr52729.2023.01632. URL http://dx.doi.org/10.1109/CVPR52729.2023.01632. (Cited on pages 16 and 22.)

[178] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4D: 4D Instance Segmentation of Close Human Interaction. (arXiv:2303.15380), March 2023. doi:10.48550/arXiv.2303.15380. (Cited on page 16.)

[179] Satoshi Yonemoto, Daisaku Arita, and Rin-Ichiro Taniguchi. Real-time visually guided human figure control using IK-based motion synthesis. In *Proceedings Fifth IEEE Workshop on Applications of Computer Vision*, pages 194–200, December 2000. doi:10.1109/WACV.2000.895422. (Cited on pages 10 and 11.)

[180] Zhixuan Yu, Linguang Zhang, Yuanlu Xu, Chengcheng Tang, Luan Tran, Cem Keskin, and Hyun Soo Park. Multiview Human Body Reconstruction from Uncalibrated Cameras. *Advances in Neural Information Processing Systems*, 35:7879–7891, December 2022. (Cited on page 14.)

[181] Jiabin Zhang, Zheng Zhu, Wei Zou, Peng Li, Yanwei Li, Hu Su, and Guan Huang. FastPose: Towards Real-time Pose Estimation and Tracking via Scale-normalized Multi-task Networks. (arXiv:1908.05593), August 2019. doi:10.48550/arXiv.1908.05593. (Cited on page 11.)

[182] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenyu Liu, and Wenjun Zeng. VoxelTrack: Multi-Person 3D Human Pose Estimation and Tracking in the Wild. (arXiv:2108.02452), August 2021. doi:10.48550/arXiv.2108.02452. (Cited on pages 15, 21, 24, 25, 26, and 27.)

[183] Yuxiang Zhang, Liang An, Tao Yu, Xiu Li, Kun Li, and Yebin Liu. 4D Association Graph for Realtime Multi-person Motion Capture Using Multiple Video Cameras. *arXiv:2002.12625 [cs]*, February 2020. (Cited on pages 9, 13, 14, 15, 16, 18, 19, 20, 25, 26, and 27.)

[184] Yuxiang Zhang, Zhe Li, Liang An, Mengcheng Li, Tao Yu, and Yebin Liu. Lightweight Multi-person Total Motion Capture Using Sparse Multi-view Cameras. *arXiv:2108.10378 [cs]*, August 2021. (Cited on page 19.)

[185] Zhe Zhang, Chunyu Wang, Wenhu Qin, and Wenjun Zeng. Fusing wearable imus with multi-view images for human pose estimation: A geometric approach. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2197–2206. IEEE, June 2020. doi:10.1109/cvpr42600.2020.00227. URL http://dx.doi.org/10.1109/cvpr42600.2020.00227. (Cited on page 12.)

[186] 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. (arXiv:2012.13392), December 2020. doi:10.48550/arXiv.2012.13392. (Cited on pages 12 and 13.)

[187] 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. (arXiv:2012.13392), July 2023. doi:10.48550/arXiv.2012.13392. (Cited on pages 12 and 13.)

[188] Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. DeepMultiCap: Performance Capture of Multiple Characters Using Sparse Multiview Cameras. (arXiv:2105.00261), August 2021. doi:10.48550/arXiv.2105.00261. (Cited on page 16.)

[189] Feng Zhou, Jianqin Yin, and Peiyang Li. Lifting by image - leveraging image cues for accurate 3d human pose estimation. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi:10.1609/aaai.v38i7.28596. URL https://doi.org/10.1609/aaai.v38i7.28596. (Cited on page 12.)

[190] Xiaowei Zhou, Spyridon Leonardos, Xiaoyan Hu, and Kostas Daniilidis. 3D Shape Estimation from 2D Landmarks: A Convex Relaxation Approach. (arXiv:1411.2942), June 2015. doi:10.48550/arXiv.1411.2942. (Cited on page 13.)

[191] Zhize Zhou, Qing Shuai, Yize Wang, Qi Fang, Xiaopeng Ji, Fashuai Li, Hujun Bao, and Xiaowei Zhou. QuickPose: Real-time Multi-view Multi-person Pose Estimation in Crowded Scenes. In *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH '22, pages 1–9, New York, NY, USA, July 2022. Association for Computing Machinery. ISBN 978-1-4503-9337-9. doi:10.1145/3528233.3530746. (Cited on pages 20, 24, 25, 26, and 27.)

[192] Zonghuang Zhuang and Yue Zhou. FasterVoxelPose+: Fast and accurate voxel-based 3D human pose estimation by depth-wise projection decay. In Berrin Yanikoğlu and Wray Buntine, editors, *Proceedings of the 15th Asian Conference on Machine Learning*, volume 222 of *Proceedings of Machine Learning Research*, pages 1763–1778. PMLR, 11–14 Nov 2024. (Cited on page 22.)