



Review

Human Pose Estimation in 2D and 3D: A Survey of Analytical Methods, Benchmarking Frameworks, and Engineering Applications

Rojan Shrestha [†] , Aroudra Syamantak Thakur [†] and Chenxi Wang ^{*}

Department of Computer Science and Engineering, The University of Texas at Arlington,
701 South Nedderman Drive, Arlington, TX 76019, USA; rxs7318@mavs.uta.edu (R.S.);
axt7255@mavs.uta.edu (A.S.T.)

^{*} Correspondence: chenxi.wang@uta.edu

[†] These authors contributed equally to this work.

Abstract

This survey presents a comprehensive review of Human Pose Estimation spanning 2D and 3D settings, unifying prior work through a taxonomy of body representations (2D keypoints, 3D skeletons, dense meshes), processing flows (top-down vs. bottom-up), problem formulations (regression vs. detection/heatmaps), and modern learning architectures (CNNs, Transformers, GCNs). We compare reported benchmark results of representative methods across widely used datasets (e.g., COCO, MPII, Human3.6M, 3DPW) and evaluation metrics (AP/OKS, PCK/AUC, MPJPE/PA-MPJPE, PVE), highlighting trade-offs between accuracy, robustness, and efficiency. Despite substantial progress driven by deep learning and temporal modeling, we identify persistent challenges, including costly and biased annotations, domain shift, occlusion, depth ambiguity, multi-person association, and real-time constraints on edge devices. We synthesize emerging directions that target these gaps, data-centric learning, stronger temporal and kinematic priors, and whole-body modeling, and outline deployment-oriented frontiers including generative motion priors, model compression, and on-device inference, framing their implications for engineering systems that demand reliable, low-latency human motion analysis.

Keywords: human pose estimation; 2D pose; 3D pose; computer vision; dataset; convolutional neural networks; transformers; graph convolutional networks; evaluation metrics



Academic Editors: Shuhui Li and
Marco Rossi

Received: 3 May 2026

Revised: 12 June 2026

Accepted: 29 June 2026

Published: 5 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Human Pose Estimation (HPE) is a computer vision task that estimates the spatial configuration of anatomical keypoints from sensory data, most commonly monocular images or videos. HPE constitutes a core methodology for quantifying human motion in engineering-relevant domains including biomedical monitoring, human-robot interaction, industrial ergonomics, and immersive systems [1,2]. Newer pose systems increasingly prioritize scalability and deployment: one-stage real-time multi-person pipelines reduce per-person inference cost in crowds, whole-body pose is becoming a default target, and category-agnostic keypoint detection is emerging to generalize beyond fixed skeleton definitions [3–5].

Recent progress in HPE is best understood as a shift in output representation and priors. For 2D HPE, likelihood-style outputs (heatmaps or coordinate distributions) replaced direct coordinate regression in many systems because they preserve spatial uncertainty and

stabilize optimization under clutter and partial visibility. Architecturally, high-resolution feature learning [6] and attention-based global context modeling improved localization of small or occluded keypoints, while post-processing and decoding refinements reduced discretization/quantization error in converting continuous joints to discrete outputs. For 3D HPE, monocular reconstruction remains ill-posed due to depth/scale ambiguity, so modern methods use constraints via temporal modeling, kinematic structure, multi-view geometry, and parametric body models rather than relying on per-frame regression alone.

This paper presents a comprehensive survey of HPE spanning 2D keypoint estimation, 3D joint reconstruction, and mesh recovery. Rather than treating HPE methods only as benchmark entries, this survey emphasizes how representation choices, inference pipelines, sensing modalities, model architectures, datasets, and evaluation metrics affect deployable engineering systems. It provides a structured taxonomy covering body representations, processing pipelines, learning formulations, training styles, and dimensional settings, and relates these method families to commonly used datasets, metrics, and benchmarking protocols.

The main contributions of this survey are:

- A unified taxonomy connecting 2D keypoint estimation, 3D skeleton reconstruction, and mesh recovery through shared design dimensions.
- An engineering-oriented comparison of HPE pipelines, including accuracy, robustness, scalability, latency, and sensing constraints.
- A consolidated review of datasets and metrics, emphasizing how benchmark design affects real-world reliability.
- A discussion of unresolved deployment challenges and emerging directions such as whole-body modeling, temporal priors, generative motion models, multimodal sensing, and edge inference.

Section 2 reviews foundational and contemporary works that have shaped the evolution of HPE and positions this survey relative to prior reviews. Section 3 presents the taxonomy used throughout the paper, while Sections 4 and 5 analyze representative 2D and 3D HPE methods, respectively. Later sections review datasets, evaluation metrics, efficiency considerations, deployment challenges, applications, and emerging research directions.

To clarify scope, this survey covers 2D keypoint estimation, 3D joint reconstruction, and mesh recovery, and it relates these method families to commonly used datasets, evaluation metrics, and benchmarking protocols. We focus on method design choices and failure modes that affect robustness and deployment (e.g., occlusion, domain shift, crowding, and compute constraints), rather than only reporting leaderboard results. Topics such as full motion-capture system design and calibration, clinical and regulatory validation in healthcare, and privacy/security policy frameworks are referenced only briefly when they impact benchmarking or deployment considerations.

To summarize the historical evolution of HPE, Figure 1 presents an overview roadmap of the field, highlighting major paradigm shifts across 2D single-person, 2D multi-person, 3D pose estimation, and mesh recovery.

The remainder of the paper is organized around this engineering perspective: we first position the survey relative to prior work, then define a taxonomy of HPE design choices, analyze 2D and 3D method families, review datasets and metrics, and finally discuss efficiency, deployment challenges, applications, and future directions.

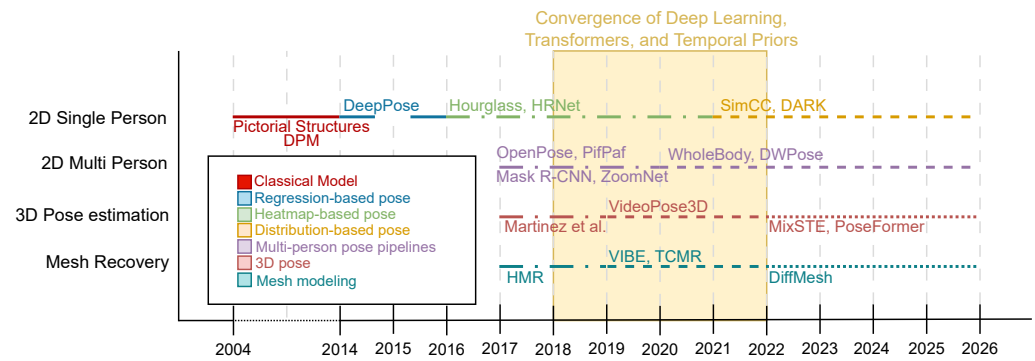


Figure 1. Grand overview roadmap of human pose estimation, illustrating major paradigm shifts across 2D single-person [7–12], 2D multi-person [13–17], 3D pose estimation [18–21], and mesh recovery [22–25].

2. Related Works and Contextual Information

HPE has been surveyed from multiple viewpoints, including broad 2D/3D overviews, 2D deep pose estimation–focused surveys, and more recent summaries that incorporate temporal modeling, whole-body settings, and mesh recovery. Early surveys such as [1,2,26] provide foundational coverage of HPE pipelines and design dimensions, with [1] emphasizing a structured breakdown of HPE dimensions and [2] offering an extensive technical review spanning both 2D and 3D settings. More recent surveys targeting deep learning–centric 2D HPE [27] summarize the rapid evolution of heatmap and regression-based formulations under modern backbones, while other works provide additional perspectives on deep learning–based HPE and its major categories [28,29]. Taken together, these surveys establish the standard taxonomy vocabulary and track the shift from classical structured models to modern deep architectures; however, they often treat 2D and 3D as largely separate narratives, discuss mesh recovery and whole-body modeling as extensions rather than first-class outputs, or summarize datasets and metrics without explicitly tying them to the method design choices and real deployment failures. Table 1 contrasts these representative surveys with the present work across the dimensions we emphasize: 2D, 3D, and mesh coverage, temporal modeling, dataset and metric treatment, and deployment/efficiency analysis.

This survey is positioned to extend prior work by using a single taxonomy consistently across 2D, 3D, and mesh-based HPE, explicitly connecting representation choice (2D joints vs. 3D skeletons vs. dense meshes), multi-person inference flow (top–down vs. bottom–up), learning formulation (direct regression vs. likelihood/distribution-based outputs), and supervision/temporal priors to the datasets, evaluation protocols, and robustness characteristics that dominate practical settings.

Earlier pictorial structure models [7] captured limb connectivity and enabled interpretable inference but were computationally expensive and fragile under occlusion, clothing variation, and background clutter. Discriminative part-based pipelines reduced dependence on explicit generative fitting by learning part detectors and assembling them into full-body configurations; deformable part models [30] are a representative example. While these methods established the importance of compositional structure, their performance depended heavily on local detector quality and degraded substantially in complex poses and crowded scenes.

A consistent theme across the contemporary literature is that benchmark improvements do not fully resolve deployment-critical bottlenecks. Multi-person HPE remains sensitive to crowd overlap and association errors; monocular 3D reconstruction remains ill-posed under severe occlusion and unusual viewpoints; and supervision quality (anno-

tation noise, label bias, limited domain diversity) materially affects training stability and cross-dataset generalization. These issues motivate recent directions such as data-centric robustness, whole-body keypoint modeling, and stronger temporal and kinematic priors. Recent real-time pipelines also shift the classic accuracy–throughput balance: RTMO demonstrates that one-stage designs can approach two-stage accuracy while sustaining high FPS in crowded scenes [3]. In parallel, whole-body and open-world keypoint detection are expanding the task definition; RTMW targets real-time whole-body pose and X-Pose introduces a unified large-scale keypoint dataset (UniKPT) to support promptable keypoint detection across many keypoint definitions [4,5]. In the remainder of this survey, we organize methods using a unified taxonomy (Section 3) to connect model families to their supervision regimes, datasets, metrics, and trade-offs in accuracy, robustness, and efficiency. This engineering-centric approach distinguishes the present survey from purely algorithm-focused reviews: rather than treating benchmark accuracy as the sole evaluation criterion, we explicitly connect method families to the measurement reliability and deployment feasibility required in real engineering systems, including edge-device constraints, sensor modality trade-offs, and cross-domain generalization.

Table 1. Comparison of existing HPE surveys with this work.

Survey	Year, Venue	2D	3D	Mesh	Temporal	Datasets & Metrics	Deployment/ Efficiency	Key Limitation
Dang et al. [27]	2019, Tsinghua Sci. Technol.	✓	×	×	⊙	✓	×	2D only; predates transformer-era methods.
Li et al. [1]	2020, Comput. Graph.	✓	✓	×	×	⊙	×	Structured breakdown of design dimensions, but no benchmarking or deployment analysis.
Munea et al. [2]	2020, IEEE Access	✓	⊙	×	×	✓	×	Primarily a 2D model taxonomy; no mesh or deployment perspective.
Zheng et al. [28]	2023, ACM Comput. Surv.	✓	✓	⊙	⊙	✓	⊙	Mesh treated as an extension; no per-method efficiency comparison.
Tharatipyakul et al. [26]	2024, Heliyon	✓	⊙	×	⊙	⊙	⊙	Domain-specific (movement feedback); not a general taxonomy.
Neupane et al. [29]	2025, Artif. Intell. Rev.	×	✓	⊙	✓	✓	×	3D only; 2D treated as an input stage.
This Survey	2026	✓	✓	✓	✓	✓	✓	—

Note: ✓ = covered in depth; ⊙ = partial coverage; × = not covered. Deployment denotes explicit treatment of latency, model complexity, resource requirements, and failure modes alongside accuracy.

Accordingly, the present survey uses prior algorithmic progress as a foundation but evaluates HPE methods through the lens of system design, deployment constraints, and failure modes.

3. Taxonomy

HPE systems can be categorized by the design decisions that most directly affect deployment: target setting, output dimensionality, input modality, inference pipeline, output representation, and architecture family. This section uses these dimensions to organize HPE methods around engineering trade-offs such as latency, scalability, robustness to occlusion, annotation cost, and suitability for resource-constrained hardware. An overview of the full taxonomy adopted in this survey is presented in Figure 2.

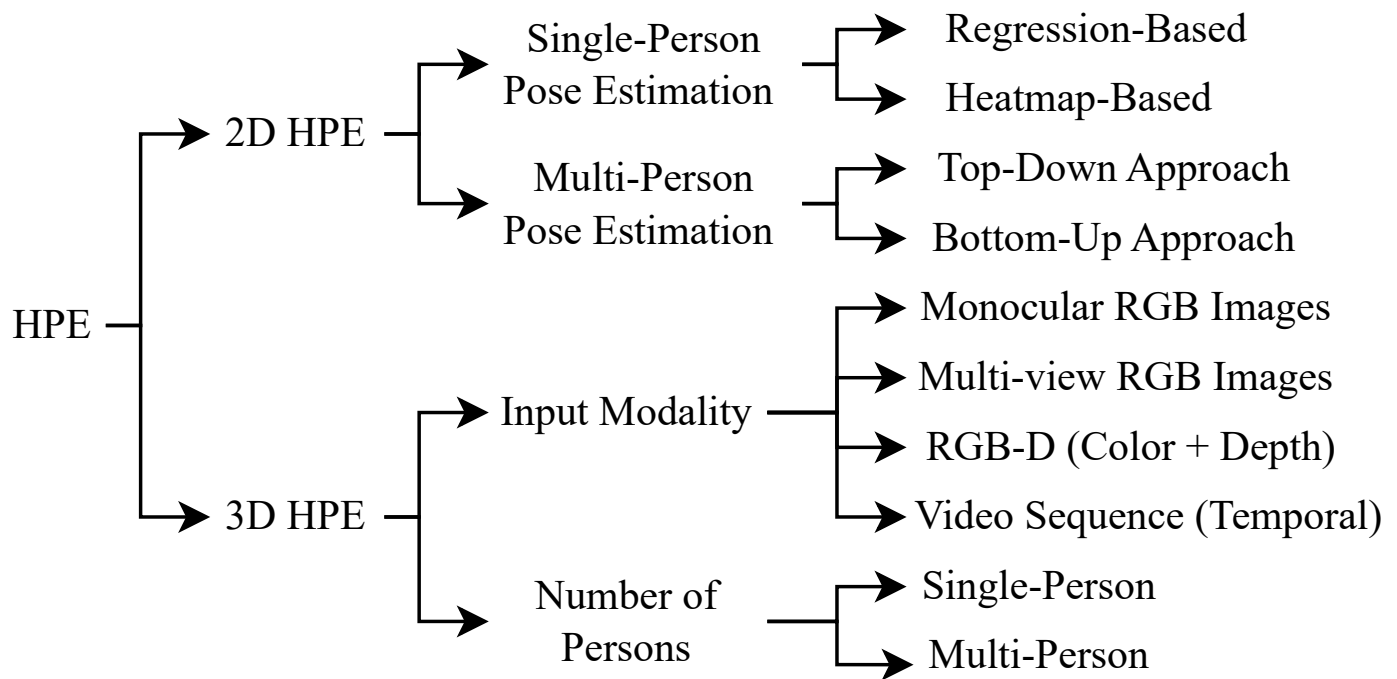


Figure 2. Taxonomy of this survey.

3.1. Target Setting: Single-Person vs. Multi-Person

The target category determines whether the method must solve instance separation. Single-person methods assume one subject and focus on localization accuracy, whereas multi-person methods additionally require robust instance association under overlap and occlusion. There are mainly two categories of architectures under the target category: single-person pose estimation and multi-person pose estimation. As the name suggests, single-person pose estimation assumes only a single person is present within the image provided and extracts the keypoints for the individual. Common examples are DeepPose [8] and Stacked Hourglass [9]. Similarly, the multi-person model estimates the poses for all the people present in the image, regardless of the number. Multi-person models employ processing flow approaches such as top-down and bottom-up methods to improve the processing complexity of the task. Common examples of multi-person pose estimation are OpenPose [13] and Mask R-CNN [14]. Single-person settings typically avoid association, but they still require scale robustness and occlusion handling; multi-person settings add the association problem, which dominates errors in crowds.

3.2. Output Representation

Output representation defines what the HPE system predicts and directly influences model design, supervision requirements, evaluation metrics, and deployment suitability. The most common representations include 2D keypoints, 3D skeletons, dense meshes, whole-body keypoints, and parametric body models.

3.2.1. 2D Keypoints

2D keypoint representations estimate anatomical landmarks such as shoulders, elbows, wrists, hips, knees, and ankles in the image plane [8–10]. They are widely used because they require only image-level annotations and can be evaluated efficiently using metrics such as PCK, OKS, or AP. However, 2D keypoints do not encode depth, body shape, or metric scale, which limits their use in applications requiring physical motion understanding.

3.2.2. 3D Skeletons

3D skeleton representations estimate joint positions in three-dimensional space, usually as camera-relative or root-relative coordinates [18]. They provide richer motion information than 2D keypoints and are useful for biomechanics, robotics, animation, and motion analysis. Their main limitation is that 3D supervision is harder to obtain and monocular 3D reconstruction remains affected by depth ambiguity, scale uncertainty, and occlusion.

3.2.3. Dense Meshes

Dense mesh representations estimate the full body surface instead of only sparse joints [31]. This enables shape-aware applications such as AR/VR, virtual try-on, animation, human-object interaction, and biomechanical analysis. However, mesh recovery requires more complex models, stronger priors, and more difficult evaluation because surface accuracy, physical plausibility, and temporal consistency must all be considered.

3.2.4. Whole-Body Keypoints

Whole-body HPE extends standard body keypoints to include the face, hands, feet, and other fine-grained landmarks [15,16]. This representation is important for expressive motion understanding, sign language analysis, gesture recognition, human-computer interaction, and avatar control. The engineering challenge is that whole-body estimation greatly increases the number of keypoints, introduces scale imbalance between large body joints and small hand or facial landmarks, and requires higher-resolution features.

3.2.5. Parametric Body Models

Parametric body models, such as SMPL-style representations [32], encode human pose and shape using a low-dimensional set of parameters that generate a full 3D body mesh. These models provide anatomical structure and shape consistency, making them useful for mesh recovery and motion synthesis. Their limitations include dependence on model priors, possible mismatch with diverse body shapes or clothing, and difficulty handling severe occlusion or unusual poses.

From an engineering perspective, output representation should be selected based on the downstream task. 2D keypoints are efficient and suitable for image-plane analysis, 3D skeletons support physical motion reasoning, dense meshes provide surface-level detail, whole-body keypoints capture fine-grained interaction cues, and parametric body models offer compact structure for pose and shape recovery.

3.3. Input Modality

Input modality determines the type of sensor data available to the HPE system and strongly affects accuracy, deployment cost, robustness, and hardware requirements. While 2D HPE is commonly performed from monocular RGB images, 3D HPE often benefits from additional temporal, depth, geometric, or inertial information to reduce ambiguity and improve stability.

3.3.1. Monocular RGB

Monocular RGB images are the most widely used input modality because standard cameras are inexpensive, widely available, and easy to deploy. However, monocular inputs provide only image-plane evidence, making depth, scale, and occlusion difficult to resolve, especially for 3D reconstruction.

3.3.2. Video Sequences

Video-based HPE uses temporal information across consecutive frames to improve motion consistency, reduce jitter, and recover partially occluded joints [18,22]. The main engineering cost is increased memory, latency, and sequence-processing complexity, especially when long temporal windows are required.

3.3.3. RGB-D

RGB-D systems combine color images with depth measurements, reducing monocular depth ambiguity and improving 3D localization [33]. Their limitations include additional hardware cost, sensor noise, limited operating range, and reduced reliability outdoors or on reflective and transparent surfaces.

3.3.4. Multi-View RGB

Multi-view RGB systems use synchronized cameras from multiple viewpoints to reconstruct more accurate 3D poses through geometric consistency or triangulation [34]. They provide strong 3D accuracy but require camera calibration, synchronization, higher setup cost, and controlled capture environments.

3.3.5. IMU-Based and Vision-IMU Fusion

Inertial measurement units provide acceleration and orientation cues that can improve motion tracking when visual evidence is missing or unreliable [35]. IMU-based or vision-IMU fusion systems are useful for motion capture, biomechanics, and AR/VR, but they require wearable sensors, calibration, and careful handling of drift.

3.3.6. Event Cameras

Event cameras asynchronously capture changes in brightness rather than full image frames, making them useful for high-speed motion, low-latency sensing, and challenging lighting conditions [36]. However, event-based HPE remains less mature than RGB-based HPE because event data is sparse, requires specialized processing, and has fewer large-scale annotated benchmarks.

From an engineering perspective, input modality selection is a system-level decision. Monocular RGB offers the lowest deployment cost, video improves temporal reliability, RGB-D and multi-view systems reduce 3D ambiguity through additional sensing, IMUs provide motion cues beyond vision, and event cameras support high-speed or low-light scenarios. The appropriate modality depends on the target environment, hardware budget, accuracy requirement, and acceptable setup complexity.

3.4. Learning Formulation

Learning formulation defines how the pose estimation problem is represented during training and inference. It determines the supervision signal, output tensor structure, optimization behavior, post-processing requirements, and robustness to ambiguity. Common formulations include direct coordinate regression, heatmap prediction, distribution-based coordinate classification, detection-style one-stage prediction, 2D-to-3D lifting, and generative or diffusion-based reconstruction.

3.4.1. Direct Coordinate Regression

Direct regression methods predict keypoint coordinates directly from visual features [8]. This formulation is compact and avoids large heatmap outputs, making it attractive for lightweight or low-latency systems. However, direct regression can be difficult to optimize because the mapping from image appearance to joint coordinates is highly non-linear and ambiguous, especially under occlusion, rare poses, or multi-person interactions.

3.4.2. Heatmap Prediction

Heatmap-based methods predict a spatial likelihood map for each keypoint, where high-response regions indicate probable joint locations [9]. This formulation preserves spatial uncertainty and usually improves localization accuracy compared with direct regression. The main engineering cost is increased memory and computation from high-resolution output maps, along with the need for coordinate decoding and post-processing.

3.4.3. Distribution-Based Coordinate Classification

Distribution-based formulations represent keypoint localization as classification over coordinate bins or probability distributions rather than direct coordinate regression [12]. These methods can reduce quantization error and achieve sub-pixel localization while avoiding some of the cost of full 2D heatmaps. They are useful when high localization precision is needed but memory or decoding overhead must remain manageable.

3.4.4. Detection-Style and One-Stage Prediction

Detection-style formulations integrate person detection and keypoint prediction into a unified pipeline [3]. One-stage methods can improve throughput by avoiding separate detector and pose-estimator stages, making them attractive for real-time multi-person HPE. A fully end-to-end variant of this formulation has recently emerged: query-based methods such as PETR [37], ED-Pose [38], and GroupPose [39] follow the DETR paradigm, predicting all person instances and their keypoints in a single forward pass without person detectors or keypoint-grouping heuristics, which simplifies deployment and avoids per-person scaling costs. Their limitation is that localization, confidence estimation, and association must be solved jointly, which can become difficult in crowded or heavily occluded scenes.

3.4.5. 2D-to-3D Lifting

2D-to-3D lifting methods first estimate 2D keypoints and then infer 3D joint coordinates from the 2D pose sequence or skeleton [18]. This formulation is modular and can reuse strong 2D pose estimators, but it is vulnerable to error propagation from inaccurate 2D detections. It also remains affected by depth and scale ambiguity, especially when only monocular inputs are available.

3.4.6. Generative and Diffusion-Based Reconstruction

Generative formulations model pose or mesh recovery as sampling, refinement, or denoising from a learned motion or body prior [25]. Diffusion-based and generative approaches are promising for resolving ambiguity, producing plausible motion, and improving mesh recovery under missing or noisy observations. Their engineering challenges include higher inference cost, more complex training, and the need to ensure that generated poses remain physically plausible and useful for downstream tasks.

From an engineering perspective, the learning formulation determines the trade-off between accuracy, memory use, latency, ambiguity handling, and post-processing complexity. Heatmaps and distributions improve localization reliability, regression and one-stage prediction reduce output complexity, 2D-to-3D lifting improves modularity, and generative methods provide stronger priors for ambiguous 3D reconstruction.

3.5. Architecture Type

As discussed earlier in Section 2, architecture type has revolutionized the field of HPE and has provided a faster, more accurate, and efficient way of extracting features through the use of automated feature extraction algorithms and improved resistance to occlusion.

3.5.1. CNN-Based Methods

CNN-based methods can be grouped into high-resolution localization networks, lightweight real-time networks, and hybrid CNN-attention models. Arguably the most widely used architecture in HPE, CNN-based architectures have revolutionized 2D and 3D pose estimation among other fields. YOLO-pose [40] is a novel heatmap-free approach for keypoint detection. YOLO-pose has achieved state-of-the-art results on COCO validation and test-dev sets in the domain of bottom-up approaches in a single forward pass. RTMPose [41], which systematically studies architecture/training/deployment factors and reports strong COCO accuracy with mobile/CPU throughput, making it a common reference for deployment-oriented comparisons. Extensions such as A-HRNet [42] incorporate channel-attention modules to improve accuracy. Similarly targeting parameter efficiency, SOCA-PRNet [43] infuses spatially oriented attention and global context blocks into a ResNet34-based PoseResNet, reporting 90.9 PCKh@0.5 on the MPII validation set with a substantially smaller model than high-resolution alternatives. Yang et al. (2023) [15] introduce a novel approach to enhance whole-body pose estimation models. The proposed method, named DWPose, integrates a two-stage distillation process to improve the effectiveness and efficiency of pose estimators. It provides an efficient training strategy and leverages diverse datasets to improve model performance in complex, real-world scenarios. Papaioannidis et al. (2023) [44] present a CNN architecture designed for real-time, high-accuracy single-person 2D HPE, optimized for autonomous systems. The architecture features two parallel neural heads: one for human body structure modeling and another for 2D body keypoint regression, enabling efficient processing and integration of outcomes. Not only do CNN-based methods provide high accuracy and efficient feature extraction, but they are also straightforward to integrate into other networks [29]. Yu et al. (2023) [45] is an excellent example of a hybrid CNN-Transformer-based approach to HPE using spatial context and spatial awareness along with self-supervised techniques.

3.5.2. Transformer-Based Methods

Transformers improve HPE by modeling long-range spatial/temporal dependencies via self-attention, which is especially effective for video-based 3D lifting and motion-coherent prediction (e.g., MixSTE [19], multi-hypothesis refinement [46], motion representation learning [47]). Transformer variants increasingly combine global self-attention with explicit modeling of skeletal structures; MotionAGFormer [48] integrates attention with graph-based joint reasoning to better capture kinematic dependencies and improve monocular 3D pose estimation. At the backbone level, recent human-centric foundation models show that large-scale pretraining can substantially improve pose transfer and high-resolution inference, complementing task-specific transformer heads [49]. This becomes particularly relevant when dealing with flexible input representations and global context modeling. Transformers and attention mechanisms provide improved occlusion handling and better multi-person feature extraction and estimation. Wang et al. (2021) [50] constructed the first domain-specific human keypoint detection dataset for lower body rehabilitation, comprising approximately 1.8 million annotated poses across 31 therapeutic exercises, and proposed three lightweight HRNet-based models incorporating EESP and attention mechanisms, achieving nearly 4x reduction in computational complexity with comparable accuracy. Li et al. (2020) [51] added attention branches in the proposed model with very little extra computation operations, which extract additional features by compressing and recovering the channels of the input feature map. Zhang et al. (2022) [19] introduced MixSTE, a transformer-based sequence-to-sequence (Seq-2-Seq) model designed to lift 2D keypoint sequences. Li et al. (2022) [46] address the inherent depth ambiguity and self-occlusion in monocular video-based 3D HPE by generating and refining multiple plau-

sible pose hypotheses, then synthesizing them into a single accurate pose estimate. Zhao et al. (2022) [52] provide a novel transformer architecture designed via embedding graph convolution layers after a multi-head attention block, designed to better exploit spatial relationships in the human skeleton. Wei et al. (2022) [53] introduce MPS-Net (Motion Pose and Shape Network), a video-based architecture designed to estimate both 3D human pose and body shape from monocular videos, with a special focus on temporal coherence and motion smoothness. Zhu et al. (2023) [47] introduce a unified two-stage framework to learn universal human motion representations that can generalize across multiple downstream tasks, including 3D HPE, action recognition, and mesh recovery. In addition, Xiong et al. (2022) [54] utilized pre-trained Swin Transformer to extract features, and leverage a feature pyramid structure to extract and fuse feature maps from different stages. Wang et al. (2023) [55] developed a novel model TransNet using a convolutional neural network design with a parallel transformer encoder branch to capture the long-range spatial dependency simultaneously while fusing it with the local features extracted from the input images. Hybrid designs of this kind are now well established: HRFormer [56] embeds local-window self-attention within HRNet's multi-resolution topology, retaining high-resolution representations at transformer scale, while TokenPose [57] and TransPose [58] pair convolutional feature extractors with transformer encoders operating over keypoint tokens. Beyond hybrid CNN–Transformer designs, fully Transformer-based architectures have proven remarkably effective: ViTPose [59] shows that a plain, non-hierarchical Vision Transformer with a lightweight decoder achieves state-of-the-art COCO accuracy, and ViTPose++ [60] extends this across body types and scales.

3.5.3. Graph Convolutional Networks (GCNs)

GCNs are most useful when the intermediate representation is already skeletal, such as 2D-to-3D lifting, temporal pose refinement, action recognition, or occlusion recovery. They are less directly suited to raw image feature extraction unless combined with CNN or Transformer visual encoders. Unlike CNNs and general Transformers that treat images or sequences as grids or token sequences, GCNs model the human body as a graph, keypoints are treated as nodes, and bones are edges. GCNs encode the skeleton as a kinematic graph and propagate information along bones, providing an explicit structural prior under occlusion (e.g., FlexGCN [61], GLA-GCN [62]). Zhao et al. (2024) [63] improve robustness in single-view 3D HPE by modeling spatiotemporal skeleton sequences and handling occlusions effectively. Shahjahan and Hamze (2024) [61] introduce FlexGCN, which extends conventional one-hop GCNs by aggregating information from both immediate and secondary neighbor keypoints, capturing more global structure when resolving ambiguous or occluded poses. Yu et al. (2023) [62] present GLA-GCN, which models global spatiotemporal context while maintaining fine-grained local keypoint representations. It achieves up to a 17% reduction in error over prior state-of-the-art on datasets like Human3.6M and MPI-INF-3DHP.

3.6. Engineering Trade-Offs

The comparative performance of CNNs, Transformers, and GCNs (Table 2) highlights that no single architecture dominates across all settings:

- CNNs excel in image-based inputs and are highly data-efficient with pre-trained backbones.
- Transformers outperform CNNs in capturing global context and temporal dependencies, making them highly effective in 2D-to-3D lifting tasks.
- GCNs are most effective when body structure and keypoint connectivity are critical, such as in skeleton-based action recognition and occlusion-heavy scenes.

While AP, PCK, and MPJPE quantify pose accuracy, many modern HPE deployments (mobile AR/VR, HRI, robotics, smart cameras) are constrained by on-device latency, memory, and energy. In practice, model choice is often driven by a trade-off between accuracy and real-time throughput (FPS) under a fixed compute budget. Architecturally, CNN backbones are typically compute-efficient and scale linearly with image resolution, while Transformer-based designs can incur higher cost as attention grows with the number of tokens (and for video, with temporal window length). In multi-person settings, top-down pipelines additionally scale with the number of detected people (detector cost plus per-person pose inference), whereas bottom-up methods run a single forward pass and then perform grouping, often improving throughput in crowded scenes. Recent on-device systems therefore report efficiency metrics such as parameter count, GFLOPs, and latency/FPS alongside accuracy, and often provide model variants (for example lite/full/heavy) to target different real-time budgets.

Table 2. Key relevant characteristics of CNN, transformer, and GCN architectures.

Characteristic	CNN	Transformer	GCN
Input Type	Image (Grid)	Patches/Tokens	Graph (Joints as nodes)
Strengths	Spatial locality, efficient	Captures global dependencies	Models anatomical structure
Temporal Modeling	Limited (ConvLSTM/3D CNN)	Excellent via self-attention	Good when extended to spatiotemporal GCN
Occlusion Handling	Moderate	Strong (contextual reasoning)	Strong (joint-level reasoning)
Computational Complexity	Scales approximately linearly with resolution	Attention scales quadratically with tokens or sequence length	Low; sparse graph operations
Memory Consumption	Moderate	High for long sequences or high-resolution inputs	Low
Real-Time Inference	Strong; mature mobile variants	Possible but requires optimization such as token pruning or distillation	Fast, but depends on upstream 2D detector
Multi-Person Scalability	Pipeline-dependent (per-person cost in top-down methods)	Query-based designs handle multi-person estimation end-to-end	Per-skeleton processing; scales with person count
Robustness to Noise	Moderate; local features degrade under occlusion	Strong due to global contextual reasoning	Strong structural prior but inherits 2D detector errors
Examples	HRNet [10], DeepPose [8], SimpleBaseline [20]	MixSTE [19], PoseFormer [21], TokenPose [57]	ST-GCN [64], GLA-GCN [62], FlexGCN [61]

3.7. Engineering Implications of the Taxonomy

The taxonomy in Figure 2 is not only organizational; it also reflects deployment constraints that shape practical HPE system design. Top-down pipelines often provide strong localization but scale poorly with the number of people because each detected person requires separate pose inference. Bottom-up and one-stage methods improve throughput in crowded scenes but shift the difficulty to keypoint grouping and association. Heatmap-based methods improve localization robustness but introduce memory and decoding overhead, while coordinate or distribution-based formulations can reduce post-processing cost. Similarly, CNNs remain attractive for real-time and edge deployment, Transformers improve global and temporal reasoning at higher computational cost, and GCNs provide explicit skeletal priors when joint connectivity is central to the task. These trade-offs suggest that HPE model selection should be guided by the target operating environment, not benchmark accuracy alone. Together, Tables 2 and 3 show that HPE design choices should be evaluated not only by benchmark accuracy, but also by latency, scalability, sensing cost, robustness, and downstream deployment requirements. With this taxonomy in place, we next review modern 2D HPE frameworks, which form the

foundation of many contemporary pose pipelines and serve as the primary input for many 3D reconstruction methods.

Table 3. Trade-offs across HPE design choices.

Design Choice	Advantage	Limitations	Best-Fit Deployment
<i>Axis 1: Multi-person inference flow (choose one)</i>			
Top-down pipeline	Strong per-person localization; naturally supports high-resolution single-person refinement.	Latency scales with the number of detected people; depends heavily on detector quality.	Low-to-moderate crowd density, whole-body pose, applications where localization precision is prioritized.
Bottom-up pipeline	Processes all people in a single forward pass; more scalable in crowded scenes.	Requires reliable keypoint grouping and association; errors increase under severe overlap.	Crowded scenes, surveillance, sports footage, and real-time multi-person monitoring.
<i>Axis 2: Output representation (choose one)</i>			
Heatmap output	Preserves spatial uncertainty and improves keypoint localization stability.	Requires larger output tensors, more memory, and coordinate decoding/post-processing.	High-accuracy 2D keypoint estimation, medical/biomechanics analysis, and systems where precision matters.
Coordinate regression	Compact output representation with lower memory and simpler post-processing.	More sensitive to ambiguity, occlusion, and complex poses; can be harder to optimize.	Lightweight or low-latency applications where approximate pose is acceptable.
<i>Axis 3: Backbone architecture (choose one)</i>			
CNN backbone	Efficient spatial feature extraction; mature pretrained models; strong real-time variants.	Limited global context unless combined with multi-scale fusion, attention, or temporal modules.	Mobile, embedded, edge, and real-time camera systems.
Transformer backbone	Strong global context modeling and temporal dependency learning.	Higher compute and memory cost, especially for high-resolution inputs or long video sequences.	Server-side inference, video-based 3D lifting, occlusion-heavy scenes, and long-range reasoning applications.
GCN skeleton model	Explicitly models anatomical connectivity and joint relationships.	Usually requires reliable detected keypoints inputs; less suitable as a standalone visual encoder.	2D-to-3D lifting, temporal pose refinement, action analysis, and occlusion recovery.
<i>Axis 4: Input modality (choose one)</i>			
Monocular RGB	Low-cost, widely available, and easy to deploy with standard cameras.	Depth and scale ambiguity; sensitive to viewpoint, lighting, and occlusion.	Fitness apps, mobile systems, web applications, and general consumer camera deployments.
RGB-D	Provides explicit depth and reduces monocular ambiguity.	Additional hardware cost; limited outdoor reliability; noisy depth on reflective or transparent surfaces.	Indoor robotics, gesture recognition, rehabilitation monitoring, and controlled smart environments.
Multi-view RGB	Improves 3D accuracy through geometric triangulation and complementary viewpoints.	Requires multiple synchronized cameras, calibration, and controlled setup.	Motion capture studios, sports biomechanics, laboratory studies, and 3D reconstruction.
<i>Axis 5: Temporal and output extensions (optional)</i>			
Temporal video model	Improves motion consistency, reduces jitter, and helps recover occluded joints.	Increases memory, latency, and sequence-processing complexity.	Sports analytics, rehabilitation, animation, surveillance, and applications using pose trajectories.
Mesh recovery	Provides full-body surface, shape, and richer human representation beyond sparse joints.	Higher model complexity; harder evaluation; can produce physically implausible meshes under occlusion.	AR/VR, animation, virtual try-on, biomechanics, and human-object interaction.

Note: The five axes are orthogonal rather than mutually exclusive: a deployed system combines one choice per axis, with temporal and mesh extensions added as required.

4. 2D HPE Approaches

2D HPE estimates predefined anatomical keypoints in the image plane from images or video frames. In single-person settings, the primary challenge is precise keypoint localization; in multi-person settings, the system must additionally separate individuals and associate keypoints with the correct person instance [65]. Most recent improvements come from better uncertainty representation, stronger multi-scale feature backbones, and decoding strategies that reduce quantization error under scale variation and occlusion. Generally, 2D HPE falls into single-person and multi-person categories based on people count.

4.1. Single-Person Pose Estimation

Single-person HPE methods primarily differ in how keypoints are represented and supervised. A representative single-person pipeline is illustrated in Figure 3.

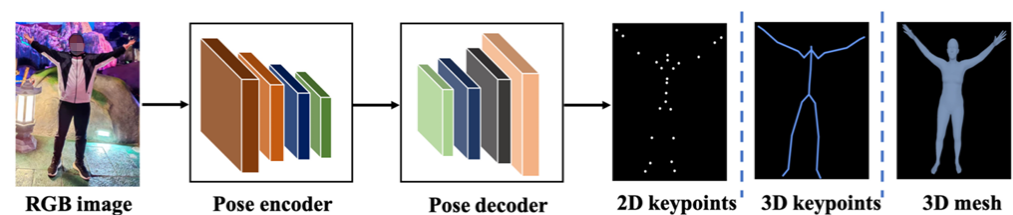


Figure 3. Typical framework for single-person pose estimation [66].

4.1.1. Regression Methods

Regression methods directly predict the (X, Y) coordinates of body keypoints from an input image, essentially learning a direct mapping from the image to the keypoint positions. DeepPose [8] used a deep neural network to directly output keypoint coordinates, often refining these predictions through a cascade of regressors. While groundbreaking, direct coordinate regression proved to be a challenging task because it is highly non-linear and can be unstable, especially for complex poses, which often led to less accurate results compared to heatmap-based methods and spurred further research into alternative representations. TFPose [67], instead of relying on heatmaps, treats pose estimation as a sequence prediction problem, directly predicting a sequence of keypoint coordinates. It uses a standard CNN to extract features, which are then processed by a transformer encoder and decoder. Its query-to-query attention explicitly encodes inter-keypoint structure, reducing regression-based feature misalignment and implicitly learning anatomical dependencies.

From a deployment perspective, regression-based methods can reduce output memory and post-processing cost, but they are often more sensitive to ambiguity and complex poses than likelihood-based representations.

4.1.2. Heatmap-Based Methods

Instead of directly predicting coordinates, heatmap-based methods estimate 2D “heatmaps” for each body keypoint. These heatmaps are probability maps where brighter pixels indicate a higher likelihood of a keypoint being located at that position. This approach is widely used because it preserves spatial information and makes the training process smoother and more stable compared to direct regression.

A foundational contribution in this area is Stacked Hourglass Networks [9]. This architecture is designed like an hourglass, with successive steps of pooling (downsampling to capture global context) and upsampling (recovering resolution for precise localization). It repeatedly processes features across all scales, using “skip connections” to preserve fine-grained spatial details. By stacking multiple hourglass modules and applying “intermediate

supervision” (losses at different stages), the network iteratively refines its predictions, significantly improving accuracy, especially for difficult-to-detect keypoints like knees and ankles.

Building on this, HRNet [10] introduced a new paradigm by maintaining high-resolution representations throughout the entire network, rather than recovering them from low-resolution features. HRNet connects multi-resolution sub-networks in parallel and repeatedly exchanges information across these parallel streams. This “multi-scale fusion” ensures that high-resolution representations are continuously enriched with information from all scales, leading to more accurate and spatially precise keypoint heatmaps. TokenPose [57] uses Transformer layers to learn both visual cues and anatomical constraints. By explicitly embedding keypoints as tokens, the model can effectively capture global context and gradually focus on local regions, leading to competitive performance with fewer parameters and computational cost compared to traditional CNNs.

To address the “quantization error” inherent in converting continuous coordinates to discrete heatmaps, Distribution-Aware Coordinate Representation (DARK) [12] was proposed which focused on improving how coordinates are represented and decoded from heatmaps. It uses a “distribution-aware decoding” approach that uses the Gaussian distribution of heatmaps and Taylor expansion to more accurately pinpoint sub-pixel keypoint locations. It also introduces “unbiased coordinate encoding” to create more precise ground-truth heatmaps for training, making it a model-agnostic plug-in that boosts performance across various existing models.

Similarly, SimCC [11], reformulated 2D HPE as two independent classification tasks, one for horizontal (X) coordinates and one for vertical (Y) coordinates. SimCC divides each pixel into multiple “bins” and achieves sub-pixel localization precision and significantly reduces quantization error. By omitting expensive upsampling and post-processing, this approach streamlines the architecture and yields an efficient pipeline, particularly for low-resolution inputs. This makes SimCC attractive for deployment because it reduces reliance on expensive 2D heatmap upsampling while preserving sub-pixel localization. Building on the coordinate-classification formulation, AECA-PRNetCC [68] augments a PoseResNet backbone with adaptive efficient channel attention to improve coordinate classification accuracy at low computational cost.

From an engineering perspective, heatmap methods improve localization robustness but increase memory use and decoding overhead, especially at high output resolutions. The drawback is that heatmap resolution affects memory and computation. Higher-resolution heatmaps improve localization but increase output tensor size and decoding cost.

4.2. Multi-Person Pose Estimation

In multi-person HPE, frameworks are commonly categorized into top-down and bottom-up pipelines, as defined in Section 3. Typical multi-person processing flows are illustrated in Figure 4.

4.2.1. Top-Down Approach

As described in Section 3, top-down methods rely on person detection followed by per-instance pose estimation. These pipelines typically achieve strong localization accuracy when detections are reliable, but their computational cost scales with the number of detected people and they are sensitive to detector errors, which directly propagate to pose estimates. A representative example is SimpleBaseline [69], which applies a ResNet backbone with deconvolutional upsampling to generate high-resolution keypoint heatmaps. Extensions such as ZoomNet [16] expand top-down pipelines to whole-body pose estima-

tion by selectively zooming into hands and face regions, improving fine-grained keypoint localization while preserving a single-network design.

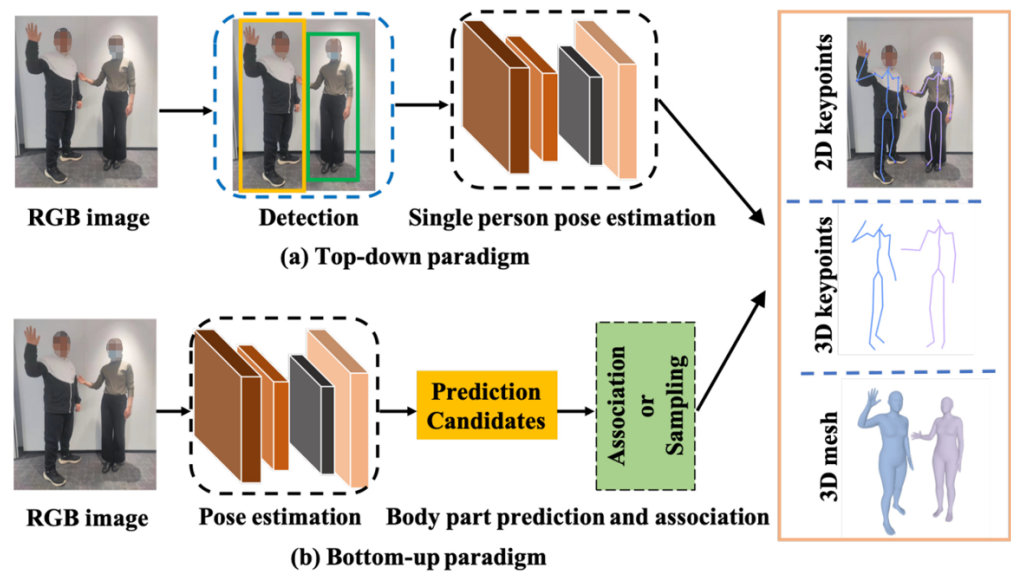


Figure 4. Typical frameworks for multi-person pose estimation [66].

In deployed multi-person systems, top-down pipelines are attractive when person counts are low or detection quality is high, but their latency increases as the number of people in the scene grows.

4.2.2. Bottom-Up Approach

In contrast, bottom-up methods follow the taxonomy outlined in Section 3 by first detecting all keypoints jointly and then grouping them into person-specific skeletons. This design avoids per-person inference and therefore offers better scalability in crowded scenes, although grouping errors can limit accuracy under heavy occlusion. PifPaf [17] exemplifies this class by using composite fields, including Part Intensity Fields (PIF) for keypoint localization and Part Association Fields (PAF) for grouping. Its association-aware design and Laplace-based regression losses improve robustness in low-resolution and occlusion-heavy scenarios.

Bottom-up methods are often more scalable in crowded scenes, but their reliability depends on robust keypoint grouping under occlusion and body overlap.

4.3. Summary for 2D HPE

Modern 2D HPE has converged on two core design choices: *structured output representations* that model spatial uncertainty, and *multi-scale, high-resolution feature backbones* that preserve localization accuracy under scale variation and occlusion. These choices reflect a field-level consensus that representation and decoding quality are as critical as backbone capacity for accurate keypoint estimation.

Stacked Hourglass formalized repeated bottom-up/top-down refinement with intermediate supervision to improve hard keypoint localization [9], while HRNet showed that *maintaining* high-resolution representations throughout the network yields stronger keypoint localization than recovering resolution late in the pipeline [10].

Recent improvements focus on *reducing quantization and decoding error* rather than only increasing backbone capacity. DARK explicitly models the heatmap distribution and applies distribution-aware decoding to recover sub-pixel accuracy [12], while SimCC re-frames localization as two 1D classification problems, achieving precise coordinate recovery

without heavy upsampling [11]. Transformer-based 2D estimators introduce global context modeling and explicit keypoint tokenization (e.g., TokenPose) or direct transformer regression (e.g., TFPose), improving long-range reasoning and keypoint dependency modeling compared to purely convolutional encoders [57,67].

For multi-person HPE, the literature consistently shows a trade-off between accuracy and scalability. Top-down pipelines (person detection → single-person pose) remain the accuracy leader in many settings and naturally extend to whole-body estimation via region zooming [16,69]. However, their computational cost scales with the number of detected people, making crowded scenes expensive. Bottom-up methods (detect all keypoints → group into instances) improve scalability and can be competitive in dense crowds when association is robust; OpenPose introduced part affinity fields (PAFs) for keypoint grouping [13], and PifPaf proposed composite fields (PIF/PAF) that are more resilient to low resolution and occlusions via association-field design and regression losses [17]. Heatmap-free approaches that integrate detection-style losses indicate a parallel trend toward simplifying pipelines for speed [40].

Unresolved problems. Across the cited works, three practical bottlenecks remain. First, *occlusion and crowd overlap* still degrade both localization and association, particularly for small extremities and rare poses; specialized occlusion training and occlusion-aware frameworks are still active research directions [70,71]. Second, *annotation noise and label bias* are increasingly recognized as limiting factors for benchmark-driven progress; learning with inaccurate labels and analyzing faulty annotations show measurable effects on pose training stability and evaluation [72,73]. Third, *efficiency constraints* matter for deployment: high-accuracy backbones can be costly, motivating lightweight and real-time CNN designs for autonomous/embedded settings [44] and more efficient training strategies such as distillation for whole-body models [15].

Near-term trends. The surveyed literature suggests that 2D HPE is moving toward *whole-body and fine-grained keypoints* as a default target (face/hands/body/feet) [15,16], *data-centric robustness* (label cleaning, uncertainty modeling, occlusion-aware training) [71,73], and *edge-friendly architectures* that explicitly report complexity or enable real-time inference in safety-critical applications [40,44].

While 2D HPE provides accurate localization of body keypoints in the image plane, many applications require full 3D understanding of human motion. We therefore next review 3D human pose estimation frameworks, which extend 2D keypoints into three-dimensional space using geometric, temporal, and generative modeling.

5. 3D HPE Approaches

3D HPE estimates the three-dimensional configuration of human joints, and in mesh-based settings the full body surface, from images, videos, or additional sensor modalities [74,75]. 3D HPE often requires a more complex architecture as compared to 2D HPE as it requires depth information which is often more difficult to obtain without adequate tools and sensors. Accurate depth sensors are often expensive and not as widely available as compared to 2D images, making depth ambiguity a significant obstacle for 3D HPE [76]. Datasets that include depth information are also often limited; common examples are Human3.6M, MPI-INF-3DHP, both of which are recorded in controlled indoor labs. 3D HPE can be classified on the basis of several key factors such as input modality, number of persons, output representation, model type, and learning supervision style.

5.1. Input Modalities for 3D HPE

5.1.1. Monocular RGB Images

Monocular RGB images and videos are the most common input sources for 3D HPE due to their widespread availability. However, they present significant challenges like depth ambiguity and self-occlusions, as a single 2D view inherently loses depth information. It seeks to predict 3D spatial locations of the body keypoints from an RGB image. Obtaining 3D HPEs from monocular 2D images and videos often requires 2D-to-3D lifting. Extracting 3D human poses from monocular RGB images is often used in low-cost systems, fitness monitoring, and mobile apps because of its simpler architecture and lower costs. Refs. [20,31] both use monocular RGB images for HPE in these studies.

Monocular RGB is the most deployable input modality because it requires only standard cameras, but it transfers the burden of depth recovery to learned priors and temporal constraints.

5.1.2. Multi-View RGB Images

Multiple synchronized cameras RGB images taken from different camera angles serve as the primary input source in this input modality. As the same person or scenery is observed from multiple angles and viewpoints, the 3D keypoint positions can be reconstructed more accurately by triangulating the corresponding 2D keypoint locations from each camera view. Multi-camera setups mitigate depth ambiguity via complementary views, yielding more precise and robust 3D localization than monocular methods. However, it is less practical in real-world settings, as it demands controlled setups, multiple cameras, and tight synchronization. Therefore, multi-view RGB images are often used in professional motion capture studios for films and games [77,78], in biomechanical analysis of movement for sports or medical research [79,80], or for animation [81,82] and virtual reality systems.

5.1.3. RGB-D (Color + Depth)

Collecting ground truth data is often expensive and requires a combination of sensors and specialized motion capture (MoCap) systems which are expensive and impractical for large-scale or in-the-wild data. RGB-D cameras are a type of markerless MoCap systems that can be used for lower-cost or research-focused applications. These cameras provide explicit depth but can degrade outdoors because sunlight interferes with active infrared sensing. RGB-D cameras typically offer low spatial resolution and noisy measurements, complicating data processing; they also struggle with transparent or reflective surfaces.

These cameras output an RGB image and a per-pixel depth map, acquired by depth sensors (e.g., Microsoft Kinect, Intel RealSense, Azure Kinect). The depth data provides direct 3D structure, allowing models to bypass monocular ambiguity while greatly improving accuracy in keypoint position estimation, especially along the z-axis. Common applications of RGB-D cameras are gesture recognition [83,84], robotics [85], and virtual reality [86]. Refs. [33,87] provide examples of models that use RGB-D cameras for HPE.

RGB-D sensing reduces depth ambiguity but introduces hardware cost, range limits, outdoor degradation, and sensor-specific noise.

5.1.4. Video Sequence (Temporal)

Rather than processing frames independently, temporal methods estimate 3D human pose from video sequences, exploiting motion continuity to improve stability and robustness. By modeling temporal dependencies, these approaches reduce frame-to-frame jitter

and recover occluded or missing keypoints using context from neighboring frames. Temporal modeling is particularly effective for dynamic activities and unconstrained settings, and is widely used in sports analytics, surveillance, and motion capture. Representative examples include VideoPose3D, which applies temporal convolutions for efficient 2D-to-3D lifting [18], and VIBE, which combines temporal encoders with adversarial priors to recover temporally consistent SMPL pose and shape from monocular video [22].

Closely related to video-based estimation is multi-person pose tracking, which extends per-frame estimation by maintaining identity consistency across frames, associating each detected pose with the same individual over time. Top-down trackers such as PoseFlow [88] link per-frame estimates into temporally coherent pose sequences, LightTrack [89] provides a generic online framework unifying pose estimation with identity association, and Simple-Baseline [20] couples its estimator with an optical-flow-based tracking pipeline. Tracking quality is evaluated on the PoseTrack benchmark [90] using Multi-Object Tracking Accuracy (MOTA) alongside per-frame AP, capturing identity switches and missed associations that frame-level metrics ignore. Reliable tracking is a prerequisite for the temporal models discussed above, which assume identity-consistent keypoint sequences as input.

5.2. Number of Persons

5.2.1. Single-Person

Single person 3D HPE primarily focuses on predicting the pose of a single individual in the camera frame or the scene. These models assume that the area of interest is a tightly focused box around the subject or person in the input frame and hence is simpler because it removes inter-person association and is less affected by crowd-induced occlusion, although self-occlusion and object occlusion remain important failure modes. This allows more accurate keypoint localization, making single-person tracking common in fitness apps, virtual try-on, video editing/blurring, and animation, where only one subject is present.

Zhang et al. (2025) [23] focuses on a novel generative HMR framework that addresses depth and scale ambiguity by jointly encoding bounding-box data, camera parameters, and geometric cues into a ray-based representation. Li et al. (2025) [91] efficiently reconstructs 3D body, face, and hands using a binary network architecture, reducing model size and inference cost while maintaining high accuracy. Hoang et al. (2024) [92] propose robust temporal lifters trained on corrupted datasets (Human3.6M-C, HumanEva-I-C). Their augmentation (TAGN) and confidence-aware convolution (CA-Conv) significantly increase resilience of 2D-to-3D lifting models under real-world disturbances and occlusions.

5.2.2. Multi-Person

Multi-person 3D HPE aims to estimate poses for all individuals in a scene, even under crowded or occluded conditions. This is considerably more challenging due to overlapping bodies and occlusion, which make it harder to correctly associate detected keypoints to specific individuals. Similar to 2D HPE, multi-person 3D HPE relies on a top-down approach or a bottom-up approach. Refer to Sections 2 and 4 for more information on top-down and bottom-up approaches. Multi-person HPE is the most common setting, since most videos and images naturally contain multiple people (surveillance, entertainment), while single-person scenes are rare. For that reason, some of the most widely used datasets, such as COCO, Posetrack, CrowdPose, and MuPoTS-3D, are built around multi-person annotations. These reflect the natural complexity found in sports, surveillance, and social environments.

In the field of Biomechanics, Koleini et al. (2025) [93] have innovated a novel framework BioPose for predicting biomechanically accurate 3D human pose directly from monocular videos. BioPose utilizes a combination of three key components: a Multi-Query Human Mesh Recovery model (MQ-HMR), a Neural Inverse Kinematics (NeurIK) model, and a 2D-informed pose refinement technique. Reported results show that BioPose’s MQ-HMR component outperforms prior state-of-the-art mesh-recovery methods across benchmarks: on Human3.6M it achieves 42.5 mm MPJPE and 28.5 mm PA-MPJPE (vs. 44.8/33.6 for HMR2.0); on the in-the-wild 3DPW dataset, 69.0 mm MPJPE and 39.5 mm PA-MPJPE (vs. 70.0/44.5); and on EMDB, 92.5 mm MPJPE and 52.1 mm PA-MPJPE (vs. 97.8/61.5) [93].

Zheng et al. (2025) [25] introduced DeepMesh, a novel diffusion framework for human mesh recovery from a video. It connects diffusion models with human motion, resulting in the efficient generation of highly precise and seamlessly smooth output mesh sequences. It is shown to perform impressively with good results on the Human3.6M and 3DPW datasets. However, DiffMesh still struggles to combat occlusions and produces unrealistic mesh outputs in scenarios with significant occlusions.

5.3. Summary for 3D HPE

From the papers cited above, modern 3D HPE can be grouped into three major solution families: *2D-to-3D lifting* (skeleton regression from detected 2D keypoints), *direct 3D regression/triangulation* from monocular or multi-view RGB, and (iii) *mesh recovery* using parametric body models such as SMPL. The simple fully-connected lifting baseline remains influential because it isolates the core difficulty of 3D reconstruction from 2D evidence [20], while monocular single-image mesh recovery (HMR) demonstrated that jointly estimating pose and shape is feasible but sensitive to ambiguity and priors [31]. In parallel, real-time monocular 3D systems (e.g., VNect [94]) highlight the practical value of lightweight inference but also expose the accuracy limits of purely monocular cues under occlusion and unusual viewpoints [94].

A consistent conclusion across recent work is that *temporal context is one of the strongest “free” priors* for 3D HPE: temporal convolutions and semi-supervised learning improve stability and reduce jitter compared to per-frame lifting [18]. Video-based mesh methods such as VIBE use temporal encoders and adversarial/learned priors to regularize shape and motion, improving in-the-wild robustness compared to single-frame mesh recovery [22]. Transformer sequence models further strengthen long-range temporal reasoning for 3D skeleton estimation, and multi-hypothesis or motion-representation learning reflects a shift from frame-wise regression toward *motion modeling* as the central representation [19,46,47]. Graph-oriented transformers and GCN variants explicitly exploit kinematic structure, strengthening spatial constraints under occlusion by propagating information along the skeletal graph [52,61–63].

Multi-view and self-supervised geometry provide an alternative route to reduce monocular ambiguity. Multi-view pipelines triangulate consistent 2D observations (effective but hardware-intensive), while epipolar geometry enables self-supervised 3D learning without full 3D labels [78]. Recent “robustness-first” works emphasize that model ranking can change substantially under corrupted 2D detections and real-world noise, motivating benchmarks and training strategies explicitly designed for disturbance tolerance [92].

Unresolved problems. Even with strong temporal models, monocular 3D remains fundamentally ill-posed: depth ambiguity, scale uncertainty, and self-occlusion persist, and surveys dedicated to monocular 3D emphasize these as long-standing covariates [75,76,95]. Mesh recovery adds additional failure modes, including physically implausible surfaces under heavy occlusion or truncation, which motivates stronger priors and generative regularization. Diffusion-based video mesh recovery represents a concrete step toward

generative motion priors, but reported limitations still include occlusion-induced artifacts and unrealistic reconstructions in extreme cases [25]. Domain-specific requirements also challenge generic benchmarks: biomechanics-focused pipelines emphasize physically valid articulation and kinematic plausibility, which standard MPJPE-centric evaluation may not fully capture [93].

Near-term trends. Three directions emerge from the cited literature. First, *generative and motion-prior models* (transformers and diffusion) are increasingly used to regularize ambiguous 3D inference, suggesting a shift from “pose regression” to “pose-and-motion modeling” [25,47]. Second, *multimodal fusion* is positioned as a practical solution to monocular ambiguity: video + IMU fusion improves robustness and absolute motion recovery when vision fails [35,96], and event-based sensing is emerging for extreme lighting and high-speed motion [36,97,98]. Third, *deployment constraints* are becoming first-class: binarized or efficient whole-body mesh models and metric-scale recovery (MetricHMR) reflect demand for models that are simultaneously accurate, scalable, and feasible on limited hardware [23,91].

6. Datasets and Evaluation Metrics

6.1. Datasets

The evolution of HPE is deeply connected to the availability of large-scale datasets and standardized evaluation metrics. These benchmarks not only provide the necessary data for training robust models but also enable fair and consistent comparison between different approaches. This section provides an overview of the most influential datasets and the commonly used evaluation metrics for both 2D and 3D HPE.

Dataset choice is an engineering decision because capture conditions, annotation type, sensor modality, subject diversity, and scene complexity determine the failure modes a model is likely to learn or ignore.

6.1.1. 2D Datasets

Early HPE frameworks relied heavily on hand-crafted body-part features and generally struggled to generalize across pose variation, clothing, occlusion, and background clutter. Deep learning shifted the field toward learned feature representations, enabling stronger robustness when sufficient annotated data is available. A fundamental principle of deep learning models is their inherent reliance on vast amounts of data; their performance is directly proportional to the quantity and quality of the training data provided. While several datasets for 2D HPE were introduced before 2014, many have been replaced by larger and more diverse datasets better suited for training deep learning models. The following are some of the most prominent large-scale 2D human pose datasets.

Max Planck Institute for Informatics (MPII) Human Pose Dataset (2014) [99]

This dataset comprises approximately 25,000 images, featuring around 40,000 individuals. It is split into roughly 28,821 images for training and 11,701 for testing. MPII covers a broad spectrum of everyday activities, including recreational, occupational, and household tasks, captured from diverse viewpoints. Each annotation includes 15 human keypoints and occlusion labels. The design was informed by a taxonomy of over 800 human activities derived from YouTube, ensuring comprehensive coverage of both common and rare poses.

Microsoft Common Objects in Context (COCO) Dataset [100]

COCO is a widely recognized large-scale benchmark for HPE. It features over 200,000 annotated images, containing 250,000 person instances. Each instance includes

labeling of 17 keypoints, along with bounding boxes and body segmentation masks. The dataset is considered particularly challenging because it captures people in diverse real-world conditions, featuring multiple interacting individuals, wide variations in pose, scale, and viewpoint, frequent occlusions, and complex, cluttered backgrounds. These factors make COCO an essential yet difficult benchmark for evaluating the robustness and generalization of modern multi-person HPE models. Its real-world diversity makes it useful for multi-person robustness, but it remains limited to 2D annotations and cannot directly evaluate depth consistency.

PoseTrack Dataset (2017/2018) [90]

PoseTrack stands as a significant large-scale video-level dataset, specifically designed for multi-person HPE and articulated tracking. It presents demanding scenarios, including occluded individuals within dense environments across various video clips. For example, PoseTrack2018 contains 1138 videos and 153,615 pose annotations. Each person in the dataset is annotated with 15 keypoints and corresponding visibility data. This dataset is especially vital for assessing the temporal consistency of pose estimation models.

CrowdPose Dataset (2019) [101]

This dataset was specifically designed to address the challenges posed by crowded scenarios. It contains 20,000 images featuring approximately 80,000 individual persons, with samples selected based on a “Crowd Index” derived from other public benchmarks. CrowdPose effectively tackles the major challenges of inter-person occlusion and overlap.

Leeds Sports Pose (LSP) Dataset (2010) [102]

This dataset contains 2000 images of full body poses, annotated with 14 keypoints. It is split into 1000 images for training and 1000 for testing. The images were sourced from Flickr and tagged with various sports activities, including athletics, badminton, baseball, gymnastics, parkour, soccer, tennis, and volleyball. A subsequent extension, LSP-Extended (2011) [72], expanded the training set to over 10,000 images. Despite their foundational role, these datasets are relatively small and primarily focus on specific sports activities, which limits their generalizability to broader scenarios.

Frames Labeled in Cinema (FLIC) Dataset (2013) [103]

This dataset comprises approximately 5000 images, with 4000 designated for training and 1000 for testing. The images are extracted from Hollywood movies and are annotated with 10 upper body keypoints. Although FLIC features diverse poses and clothing, its relatively small size and specific domain (movie scenes) restrict its applicability.

6.1.2. 3D Datasets

3D HPE datasets are fundamentally more challenging compared to 2D HPE datasets. This is primarily because 3D data requires precise depth information, often necessitating specialized motion capture systems. The following are some of the most prominent large-scale 3D human pose datasets.

Human3.6M (2013) [104]

Human3.6M is one of the largest and most widely used indoor 3D HPE datasets. It comprises 3.6 million 3D human poses, along with synchronized high-resolution videos captured at 50 frames per second from 4 distinct camera views. The dataset showcases 11 professional actors performing 17 varied activities, such as discussions, smoking, and photography. A marker-based motion capture system was used to guarantee the high

accuracy of the 3D keypoint positions. Its controlled indoor capture makes it valuable for precise 3D benchmarking but limits direct transfer to unconstrained outdoor scenes.

MPI-INF-3DHP (2016) [105]

This dataset offers 1.3 million frames captured from 8 actors performing 8 distinct action sets. It leverages a commercial marker-less motion capture system, which allows actors to wear everyday clothes, thereby enhancing data diversity. Additionally, it provides chroma-key masks to facilitate foreground and background augmentation.

MoVi (2020) [106]

MoVi is a comprehensive single-person video dataset featuring synchronized 3D body pose and mesh annotations. It includes data from 90 subjects executing 20 distinct actions, along with detailed subject characteristics such as age, height, and BMI. MoVi uses IMUs and multi-perspective motion capture cameras to guarantee precise 3D pose acquisition.

3D Poses in-the-Wild (3DPW) (2018) [107]

A pioneering dataset for single-view, multi-person, “in-the-wild” 3D HPE. It includes 60 video sequences, totaling approximately 51,000 frames, depicting a rich variety of activities captured in diverse real-world scenes like forests, streets, and shopping malls. This dataset leverages IMUs and handheld phone cameras to achieve accurate 3D poses despite the inherent complexity of the scenes.

SURREAL (Synthetic humans for REAL tasks) (2017) [108]

SURREAL is a large-scale synthetic dataset comprising 6 million RGB frames. These frames are generated by rendering textured SMPL (Skinned Multi-Person Linear) models onto various background images, driven by 3D motion capture data. SURREAL provides extensive annotations, including 3D pose, depth, body parts, optical flow, and surface normal information.

Additional benchmarks in Table 4 extend coverage to specific regimes: AIC [109] provides large-scale crowded 2D annotations; Penn Action [110] and HiEve [111] target video-based estimation; JTA [112] offers synthetic urban scenes; and Total Capture [35] fuses IMU and Vicon ground truth for 3D evaluation.

6.2. Evaluation Metrics

Evaluating HPE models is the process of quantifying how closely the predicted positions of human keypoints or body structures match the ground truth. Depending on whether the task is 2D or 3D, different metrics are used, though many share conceptual foundations. Table 5 summarizes which metrics apply to the 2D and 3D settings and what each measures. Although standard metrics enable reproducible benchmarking, they do not fully capture engineering reliability. A model with strong AP or low MPJPE may still fail under motion blur, occlusion, unusual viewpoints, sensor noise, or real-time latency constraints. For deployment, pose accuracy should therefore be reported alongside latency, memory footprint, throughput, energy use, failure-case robustness, and task-level utility.

6.2.1. 2D Evaluation Metrics

To accurately evaluate the performance of HPE models needs defined metrics. These metrics have progressed from straightforward geometric approximations to more sophisticated, detection-focused benchmarks that consider factors like scale, visibility, and the individual significance of keypoints. Due to this many evaluation metrics have been used in 2D HPE. Here we summarise some of the most common ones. Below are some of the most common examples.

Table 4. Comprehensive Summary of HPE Datasets.

Name	Year	Type	#	Train	Val	Test	Total	Key Characteristics
Image-Based 2D Datasets								
LSP [102]	2010	SP	14	1000	-	1000	2000	Relatively small, sports focus, scaled images
LSP (Extended) [72]	2011	SP	14	10,000	-	-	10,000	Larger training set, similar characteristics to LSP
FLIC [103]	2013	SP	10	4000	-	1000	5000	Relatively small, movie-specific, diverse poses
MPII [99]	2014	SP	15	28,821	3000	11,701	~43k	Wide variety of activities, viewpoints, occlusion labels, activity labels.
COCO [100]	2016 2017	MP	17	118,287	5000	40,670	~200k	Diverse poses, varying body scales, occlusion patterns, bounding boxes, body segmentation
AIC(HKD) [109]	2017	MP	14	210,000	30,000	60,000	300,000	One of the largest, high-resolution images, 700,000 human instances.
CrowdedPose [101]	2019	MP	14	~10,000	~2000	~8000	~20,000	Specifically designed for crowded scenarios, addresses inter-person occlusion
Penn Action [110]	2013	SP	13	1258	-	1068	2326	Unconstrained videos, includes visibility and camera viewpoints
Video Based 2D Datasets								
PoseTrack [90]	2017 2018	MP	15	2017: 514 Vids (16,219 poses) 2018: 1138 Vids (153,615 poses)				Challenging crowded environments, occlusion, temporal consistency evaluation
HiEve [111]	2020	MP	14	31 Vids (1,099,357 poses)				Largest video-based HPE dataset, integrates HPE, tracking, action recognition
Video-Based 3D Datasets								
HumanEva-I [113]	2010	SP	15	~6800	~6800	~24,000	40K Fr	Indoor, multi-view, markerless MoCap, 4 subj., 6 actions
Human3.6M [104]	2013	SP	17	3.6M poses @50 FPS Train/Val 7 Subj., Test 4 Subj.				Indoor, marker-based MoCap, 11 actors, 4 camera views, 17 scenarios
MPI-INF 3DHP [105]	2016	SP	15	-	-	-	1.3M Fr	Markerless MoCap, 8 actors, 14 cameras, chroma key masks for augmentation
Total Capture [35]	2017	SP	26	-	-	-	1.9M Fr	No markers, loose clothes, IMU + Vicon data, 5 subjects, 5 actions
3DPW [107]	2018	MP	18	24 Seq	24 Seq	12 Seq	~51K Fr 60 Seq	First in-the-wild, IMUs + handheld phone camera, rich activities, real-world scenes
MoVi [106]	2020	SP	44	~9.75k	250	-	1057 Seq	Synchronized 3D pose/mesh, detailed subject characteristics (BMI, age), IMU + MoCap
JTA [112]	2018	MP	14	256		256	500K Fr 512 HD Vids	Synthetic urban traffic dataset; dense 2D/3D annotations (keypoints, occlusion, segmentation, depth); designed for crowded surveillance scenarios

Abbreviations: # = Number of keypoints; SP = single-person; MP = multi-person; Vids = videos; Seqs = sequences; Fr = frames.

Table 5. Comparison of evaluation metrics in 2D vs. 3D HPE, with keypoints considered.

Metric	2D HPE	3D HPE	Description
MPJPE	✗	✓	Average Euclidean distance between predicted and ground truth all 3D keypoints in millimeters.
PA-MPJPE	✗	✓	MPJPE after rigid alignment (Procrustes Analysis) to remove global pose/scale/rotation errors, computed over all 3D keypoints.
PCK	✓	✓	Proportion of all annotated keypoints falling within a specified distance threshold. In 2D HPE, distances are normalized using head or torso length; in 3D HPE, they are measured in absolute millimeters.
OKS	✓	✗	COCO metric computed over all annotated keypoints, normalizing distances by object size and keypoint visibility.
Mesh-based	✗	✓	Intersection over Union between predicted and ground truth meshes, evaluated over all mesh vertices rather than sparse keypoints.
PCP	✓	✗	Percentage of Correct Parts: proportion of limbs (pairs of keypoints) whose predicted endpoints are within a threshold distance from ground truth.

Note: ✓ indicates applicability of the metric; ✗ indicates not applicable.

Percentage of Correct Parts (PCP)

PCP [114] is an early metric which measures the detection rate of limbs. A limb is considered correctly detected if the distance between its two predicted keypoints and the ground truth keypoint is less than a fraction of the limb length (between 0.1 to 0.5). PCP@0.5 is a common threshold.

$$\text{PCP}@{\alpha} = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} \mathbf{1}(\max(\|\hat{J}_{l,1} - J_{l,1}\|_2, \|\hat{J}_{l,2} - J_{l,2}\|_2) \leq \alpha L_l) \quad (1)$$

where $\mathcal{L}$ is the set of limbs, $J_{l,1}$ and $J_{l,2}$ are the ground-truth endpoints of limb l , $\hat{J}_{l,1}$ and $\hat{J}_{l,2}$ are the corresponding predicted endpoints, L_l is the ground-truth limb length, and $\mathbf{1}(\cdot)$ denotes the indicator function. This measure is used for single-person HPE evaluation. However, PCP has been criticized for its tendency to be overly strict on shorter limbs, which are hard to detect.

Percentage of Correct Keypoints (PCK)

PCK measures the precision of keypoint localization. A detected keypoint is considered accurate if its predicted position is within a specified distance from the true keypoint, relative to a reference size like torso diameter or head segment length [115]. This threshold is normalized by a measure of the person's scale to ensure fairness across different image resolutions and person sizes. The most common variant, PCKh, normalizes the distance by the head segment length, which provides a more stable reference across different body poses compared to earlier normalizations based on torso or bounding box dimensions.

$$\text{PCK}@{\alpha} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}(\|\hat{J}_i - J_i\|_2 \leq \alpha d_{\text{ref}}) \quad (2)$$

where d_{ref} denotes the reference scale used for normalization, corresponding to the torso diameter in classic PCK and the head-segment length in PCKh. PCK is intuitive and easy to interpret, but as a threshold-based metric it does not distinguish fine-grained localization errors, motivating its use alongside distance-based metrics such as MPJPE.

Object Keypoint Similarity (OKS)

OKS [100] is used to measure the similarity between a predicted and ground-truth keypoint, playing a role similar to how Intersection over Union (IoU) functions in object detection. The OKS for a single person is calculated as:

$$\text{OKS} = \frac{\sum_i \exp\left(-\frac{d_i^2}{2s^2k_i^2}\right) \delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \quad (3)$$

where d_i is the Euclidean distance between the predicted and ground-truth keypoint, s is the scale of the object (e.g., the square root of the person's segmentation area), v_i is a visibility flag, and k_i is a per-keypoint constant defined in the COCO evaluation protocol that controls the falloff of the similarity for keypoint type i , set from the empirical annotation variance of that keypoint (larger for hips than for eyes).

Average Precision (AP)

AP [100] is the primary metric used in the COCO challenge. It is calculated by averaging precision values across Object Keypoint Similarity (OKS) thresholds, typically ranging from 0.50 to 0.95 in increments of 0.05. This approach provides a comprehensive

evaluation of a model's ability to both detect keypoints (i.e., achieve a minimum OKS) and localize them precisely (i.e., achieve a high OKS).

$$AP = \frac{1}{|T|} \sum_{t \in T} \text{Prec}(\text{OKS} > t), \quad T = \{0.50, 0.55, \dots, 0.95\} \quad (4)$$

where $\text{Prec}(\text{OKS} > t)$ is the precision computed at threshold t , such that a predicted pose is counted as a true positive when its Object Keypoint Similarity (OKS) score (Equation (3)) exceeds t . Common variants, such as AP50 (AP at an OKS threshold of 0.50), AP75, APM (for medium-sized objects), and APL (for large objects), offer a more detailed analysis of model performance.

6.2.2. 3D Evaluation Metrics

3D evaluation metrics are generally similar to 2D evaluation metrics but they fundamentally differ in the spatial domain they operate in and the attributes they evaluate. While 2D metrics assess keypoint accuracy in the image plane, for example pixels, 3D metrics evaluate keypoint positions in real-world or camera space (usually in millimeters). 3D metrics also need to take into consideration the depth ambiguity and camera viewpoint, both of which are unnecessary for 2D evaluation metrics.

MPJPE (Mean Per Joint Position Error)

MPJPE is one of the most widely adopted metrics for evaluating 3D HPE. It focuses on the Euclidean distance, in millimeters, between each predicted keypoint and the corresponding ground truth keypoint.

$$\text{MPJPE} = \frac{1}{N} \sum_{i=1}^N \|\hat{J}_i - J_i\|_2 \quad (5)$$

For every keypoint, the model predicts the 3D keypoint coordinates for all keypoints, calculates the distance between predicted and actual keypoint positions for each keypoint, and then averages these distances across all keypoints. One of the key advantages of MPJPE is that it directly evaluates the spatial accuracy of the keypoint predictions made by the model. The lower the distance or the cost, the higher the accuracy of the model. It also provides a simple and interpretable form of evaluation. MPJPE is commonly used in datasets like Human3.6M, MPI-INF-3DHP, and 3DPW. However, on the other hand, MPJPE is sensitive to rotation, scale, and translation. This means that even though the pose estimation might be accurate, the distance produced from this method would be higher if the pose is in a different scale or rotated. It penalizes poses even if their shape is correct but placement is wrong.

Procrustes-Aligned MPJPE (Protocol 2)

PA-MPJPE, also known as P-MPJPE, is a variant of the standard MPJPE metric that compensates for global misalignments between the predicted and ground-truth 3D poses. It attempts to combat the translation and rotation problems that inherently exist in MPJPE. PA-MPJPE addresses this by aligning the predicted pose to the ground truth first, ensuring that the metric evaluates the pose structure itself, not its global placement in space.

$$\text{PA-MPJPE} = \frac{1}{N} \sum_{i=1}^N \|(s^* R^* \hat{J}_i + t^*) - J_i\|_2, \quad (s^*, R^*, t^*) = \arg \min_{s, R, t} \sum_{i=1}^N \|(s R \hat{J}_i + t) - J_i\|_2^2 \quad (6)$$

where the similarity transformation parameters: scale s , rotation matrix $\mathbf{R}$, and translation vector $\mathbf{t}$ are estimated using Procrustes analysis before the error is computed. It does this by applying alignment to the pose and then computing the Euclidean distance between each aligned keypoint and the ground truth as formalized in Equation (6). Not only

does the PA-MPJPE measure the structural similarity between poses, but it also ignores absolute position and resolves the misalignment problem within MPJPE. It is ideal for evaluating pose reconstruction quality, especially in monocular 3D HPE, where depth and scale ambiguity are common. While it is suitable to eliminate translation misalignment of poses, it completely masks depth or translation errors, so it is often reported alongside MPJPE. It is also often not suitable if absolute position matters, for example in real-world tracking and localization.

Mesh-Based Metrics

For mesh-based human pose and shape estimation, evaluation extends beyond sparse keypoints to the full body surface. Metrics such as Intersection over Union (IoU) assess the overlap between predicted and ground-truth meshes, capturing surface-level reconstruction quality that joint-based errors cannot reflect. While mesh metrics are valuable for applications requiring accurate silhouettes or surface continuity, they are sensitive to small misalignments and depend on the availability of ground-truth meshes. Consequently, mesh-based measures are typically reported alongside vertex- or joint-level errors (e.g., PVE or MPJPE) rather than used in isolation.

Building on these datasets and evaluation protocols, we next present a comparative analysis of benchmark results reported in the original publications for representative 2D and 3D HPE frameworks.

7. Comparison and Benchmarking

All quantitative results in Tables 6–8 are collected from the original papers under their own evaluation protocols; we did not re-run any model. Each entry carries its source citation and protocol, and differences across rows therefore reflect both method and protocol.

Benchmarking protocol: We report standard metrics for each benchmark to avoid mixing incompatible evaluation protocols: (i) COCO: AP (OKS), AP@0.5, AP@0.75, AP(M), AP(L), AR; (ii) MPII: PCKh@0.5 (and per-joint scores where available); (iii) Human3.6M: MPJPE (Protocol 1) and PA-MPJPE (Protocol 2); (iv) 3DPW: MPJPE, PA-MPJPE, and PVE.

7.1. 2D HPE Benchmarking

Tables 6 and 7 compare representative 2D HPE methods on the COCO (multi-person) and MPII (single-person) benchmarks respectively, reporting both accuracy metrics and qualitative efficiency indicators. CNN-based models such as HRNet and its variants achieve strong accuracy due to their high-resolution feature representations. For example, HRNet-W48 (384×288) reports 76.3 AP on COCO and 92.3 PCKh@0.5 on MPII, placing it among the strongest widely adopted baselines. Regression refinements that reduce coordinate quantization error, including DARK, further improve performance, with DARK (HRNet-W48) reaching 76.2 AP on COCO and 90.6 PCKh@0.5 on MPII. Similarly, RSN ($4 \times$ RSN-50) and UDP (HRNet-W48) provide noticeable accuracy gains, with UDP achieving 76.5 AP on COCO.

Table 6. Comparison of 2D HPE frameworks on the COCO test-dev set (multi-person HPE).

Method	Backbone	Input	Params	GFLOPs	AP	AP.5	AP.75	AP(M)	AP(L)	AR	FPS/Latency
OpenPose [13]	CMU-Net	368×368	–	136.1	61.8	84.9	67.5	57.1	68.2	66.5	Medium–High
SimpleBaseline [69]	ResNet-152	384×288	68.6M	35.6	73.7	91.9	81.1	70.3	80.0	79.0	Medium
HRNet-W48 [10]	HRNet-W48	384×288	63.6M	32.9	76.3	90.8	82.9	72.3	83.4	81.2	Medium
RSN [116]	$4 \times$ RSN-50	384×288	111.8M	65.9	78.6	94.3	86.6	75.5	83.3	83.8	Low–Medium
UDP [117]	HRNet-W48	384×288	63.8M	33	76.5	92.7	84.0	73.0	82.4	81.6	Medium
DARK [12]	HRNet-W48	384×288	63.6M	32.9	76.2	92.5	83.6	72.5	82.4	81.1	Medium

Table 7. Comparison of 2D HPE frameworks on the MPII test set (single-person HPE).

Method	Backbone	Input	Params	PCKh	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle
Stacked Hourglass [9]	Hourglass	256 × 256	25.1	90.9	98.2	96.3	91.2	87.1	90.1	87.4	83.6
Adversarial PoseNet [118]	ResNet-50/10	256 × 256	–	93.1	98.5	94.0	89.8	87.5	93.9	94.1	93.0
HRNet-W32 [10]	HRNet-W32	256 × 256	28.5	92.3	98.6	96.9	92.8	89.0	91.5	89.0	85.7
FPD [119]	Hourglass	256 × 256	3.0	91.1	98.3	96.4	91.5	87.4	90.9	87.1	83.7
DARK [12]	HRNet-W48	384 × 288	63.6	90.6	97.2	95.9	91.2	86.7	89.7	86.7	84.0

Note: PCKh denotes PCKh@0.5.

Table 8. Comparison of 3D HPE frameworks.

<i>Human3.6M Test Set: Single-person/Indoor (Protocol 1 MPJPE, Protocol 2 PA-MPJPE)</i>							
Method	Backbone	Input	MPJPE ↓	PA-MPJPE ↓	Params	GFLOPs	FPS
VideoPose3D [18]	Temporal CNN	2D → 3D	46.8	36.5	16.95M	33.87	– ^a
PoseFormer [21]	Transformer	2D → 3D	44.3	34.6	9.6M	1358	269
MixSTE [19]	Spatio-temp. Transf.	2D → 3D	40.9	32.6	33.6M	1141	– ^b
VIBE [22]	ResNet50 + GRU	RGB → mesh	65.6	41.4	– ^c	–	– ^c
<i>3DPW Test Set: In-the-wild</i>							
Method	Backbone	Input	MPJPE ↓	PA-MPJPE ↓	PVE	Params	GFLOPs
HMR [31]	ResNet50 + SMPL	RGB	130.0	76.7	–	–	–
VIBE [22]	ResNet50 + GRU	Video	82.9	51.9	99.1	–	–

Efficiency note (hardware- and pipeline-dependent): Exact Params, MFLOPs, and runtime are not consistently reported under standardized conditions across the surveyed 3D HPE methods; “–” denotes values not reported by the original authors. Where architectural complexity is well established in the literature, we give a qualitative indication: –^a runtime unreported; lightweight temporal convolutions yield high relative throughput. –^b runtime unreported; full spatio-temporal attention over 243-frame sequences yields medium relative throughput. –^c parameter count unreported; SMPL mesh-recovery pipelines sit at the upper end of model complexity. MFLOPs are per output frame (MixSTE: 277.25 GFLOPs total over 243 frames ÷ 243). MPJPE follows Human3.6M Protocol 1; PA-MPJPE follows Protocol 2.

Consistent with prior benchmarks, top-down pipelines (e.g., SimpleBaseline, HRNet, DARK, UDP) outperform bottom-up approaches (e.g., OpenPose) in accuracy: SimpleBaseline (ResNet-152) attains 73.7 AP on COCO, compared to 61.8 AP for OpenPose. However, the qualitative FPS/latency indicators in Table 6, highlight an inherent trade-off between accuracy and efficiency. Bottom-up frameworks such as OpenPose are associated with higher relative throughput (Medium–High), as they perform a single forward pass followed by grouping, while top-down methods typically exhibit Medium throughput due to per-person inference and reliance on an upstream detector. Heavier architectures, such as RSN, fall into the Low–Medium throughput regime, reflecting their substantially higher parameter count and computational cost.

On the MPII benchmark, heatmap-based top-down methods dominate in accuracy, with Adversarial PoseNet (ResNet-50) achieving the strongest reported score of 93.1 PCKh@0.5, followed closely by HRNet and DARK variants. Overall, the combined accuracy and qualitative efficiency results in Table 9 illustrate that while high-resolution top-down frameworks remain the most accurate for 2D HPE, bottom-up and lightweight designs provide more favorable deployment characteristics in real-time or large-scale multi-person scenarios where throughput and scalability are critical.

Table 9. Qualitative efficiency comparison of 2D HPE framework families.

Model/Family	Setting	Params	GFLOPs	Latency	Throughput
Top-down CNN (HRNet-style [10])	COCO multi-person (per-person inference)	63.6–68.6	32.9–35.6	Medium	Medium
Bottom-up (OpenPose-style [13])	COCO multi-person (single-pass + grouping)	High	High	Medium	Medium–High
Lightweight (MediaPipe-style [120])	Mobile/edge real-time	Low	Low	Low	High

7.2. 3D HPE Benchmarking

On Human3.6M, the trend in Table 8 is clear: moving from frame-independent regression to temporal and transformer-based models steadily reduces error, at the cost of increased computational complexity. VideoPose3D [18] incorporates short-range temporal context through temporal convolutions, substantially reducing 3D pose estimation error to 46.8 MPJPE and 36.5 PA-MPJPE. These results demonstrate that even modest temporal modeling can significantly outperform per-frame 2D to 3D lifting approaches, while retaining a low parameter count, low GFLOPs, and high relative throughput.

Models that capture longer temporal dependencies further improve accuracy but introduce higher computational cost. PoseFormer [21], which replaces temporal convolutions with transformer-based sequence modeling, achieves 44.3 MPJPE and 34.6 PA-MPJPE with an 81-frame receptive field, while MixSTE improves further to 40.9 MPJPE and 32.6 PA-MPJPE by jointly modeling spatial and temporal dependencies over 243-frame sequences. These gains are accompanied by qualitatively higher parameter counts and GFLOPs, and correspondingly lower relative throughput compared to VideoPose3D. Notably, PA-MPJPE decreases more sharply than raw MPJPE across these models, indicating that temporal and transformer architectures are especially effective at refining relative joint configurations up to a similarity transform, even when absolute scale and translation remain challenging.

End-to-end RGB video pipelines that recover pose and shape exhibit a different accuracy–efficiency trade-off. VIBE, which combines temporal modeling with adversarial motion priors and SMPL supervision, reports 65.6 MPJPE and 41.4 PA-MPJPE on Human3.6M. While less accurate than keypoint-lifting transformer models in absolute joint error, VIBE operates in a higher-complexity regime due to mesh recovery and exhibits medium to medium-high latency, reflecting the additional cost of parametric body modeling.

On 3DPW, which presents a harder in-the-wild setting, single-frame methods struggle with depth ambiguity and motion inconsistency. HMR reports 130.0 MPJPE and 76.7 PA-MPJPE, reflecting the limitations of per-image inference under in-the-wild conditions. Injecting temporal information yields substantial gains on both metrics: VIBE reduces MPJPE to 82.9 and PA-MPJPE to 51.9 while adding mesh-level accuracy (99.1 PVE), with the temporal encoder and adversarial motion prior improving absolute joint accuracy and structural alignment alike. Although not all methods report standardized runtime or complexity metrics, the efficiency note accompanying Table 8 indicates that temporal CNN and transformer-based pipelines incur higher computational cost than frame-based CNNs, with diffusion and mesh-centric approaches lying at the upper end of the latency spectrum.

Overall, the benchmarking results reveal a clear progression in 3D HPE performance. Lightweight regression-based baselines provide strong foundations with high efficiency, but incorporating temporal context markedly improves accuracy. Transformer-based spatio-temporal models deliver the lowest MPJPE and PA-MPJPE on Human3.6M, at the expense of increased parameter count and reduced throughput. In contrast, SMPL-based video pipelines trade some joint-level accuracy for improved temporal stability and mesh quality, particularly on in-the-wild data. Across both datasets, temporal modeling is the domi-

nant factor for improving accuracy: in keypoint-lifting pipelines it sharpens relative joint configurations, while for SMPL-based video methods on in-the-wild data it substantially improves absolute accuracy, alignment, and temporal stability. Efficiency considerations increasingly shape the choice of architecture for real-time and edge-deployment scenarios.

7.3. Efficiency and Real-Time Performance on Edge Devices

Accuracy-centric metrics (AP/OKS, PCK, MPJPE) do not fully capture deployment-critical constraints such as latency, throughput, memory footprint, and energy consumption. To reflect practical feasibility, we additionally summarize commonly reported efficiency indicators, including parameter count, GFLOPs, and measured latency/FPS where available, and otherwise discuss qualitative accuracy-latency trade-offs.

As outlined in the benchmarking sections above, efficiency depends strongly on architectural and pipeline choices. In 2D multi-person HPE, top-down methods trade higher accuracy for per-person scaling costs, while bottom-up pipelines amortize computation across all instances. In 3D video-based HPE, temporal modeling introduces additional runtime overhead, with temporal CNNs generally offering lower latency than Transformer or diffusion-based sequence models. Owing to inconsistent reporting across works, Tables 6, 7, and 9 include qualitative efficiency indicators; for Table 8 these are consolidated in the accompanying note.

8. Challenges

Despite making remarkable progress in recent years, HPE is far from achieving perfection. HPE continues to face several challenges that limit its robustness and deployment in the global market. The challenges can be roughly grouped into 4 major categories.

8.1. Data-Related

One of the most pressing challenges in HPE lies in the availability and quality of data. Annotating large-scale pose datasets is expensive and time-consuming, especially for 3D data that often requires motion capture systems or multi-view setups. Even widely used datasets such as COCO, MPII, and Human3.6M suffer from limitations in diversity, with biases toward specific environments, clothing, or demographics that hinder generalization to real-world settings. Moreover, inconsistencies in manual annotations add noise that negatively affects training. Finally, models trained on curated benchmarks often face a significant domain gap when applied to unconstrained scenarios, making robust generalization an ongoing challenge.

8.2. Algorithmic Limitations

HPE faces algorithmic limitations that constrain accuracy and robustness. A central issue in 3D HPE is depth ambiguity, where different 3D poses can project to identical 2D keypoints, making monocular reconstruction inherently ill-posed. Frame-by-frame predictions often suffer from temporal inconsistency, producing jittery or unstable motion when applied to video. Handling multi-person interactions adds further complexity, as overlapping bodies and keypoint confusion challenge both top-down and bottom-up methods. Moreover, moving from body-level pose estimation to whole-body modeling, capturing hands, feet, and facial landmarks, significantly increases difficulty due to the greater precision requirements. These algorithmic hurdles express the need for stronger kinematic constraints, temporal reasoning, and holistic modeling to advance pose estimation performance.

8.3. Computational Limitations

Computational demands pose another challenge for HPE, especially when balancing accuracy with efficiency. High-performing models such as HRNet or transformer-based architectures often require large memory footprints and significant GPU resources, which limits their deployment in real-time or on edge devices. Top-down approaches, while highly accurate, scale poorly in multi-person settings since each detected individual must be processed separately. Conversely, lightweight or bottom-up models achieve faster inference but often sacrifice precision. Achieving real-time performance without compromising accuracy remains a key limitation, and designing architectures that are both computationally efficient and robust across diverse environments is an ongoing area of research.

9. Future Directions

HPE remains one of the younger fields in the domain of computer vision, with considerable room for improvement and growth. It is increasingly evaluated by whether it can transfer and remain stable outside the benchmarks. Recent surveys emphasize that domain shift, occlusion, and annotation quality still dominate failure in real settings [1,2,75]. A near-term direction is leveraging foundation-model style pretraining as a backbone [121] but adapting it to structured pose outputs with explicit anatomical priors rather than treating pose as generic regression. This motivates keypoint-token heads for 2D [57] and parametric constraints for mesh recovery [32], with evaluation focused on cross-dataset transfer and rare-pose/occlusion strata rather than only average AP/MPJPE [70,71].

Scalable supervision remains the bottleneck for 3D mesh. Self-supervised HPE should combine multi-view geometry and temporal consistency so that learning signals persist under missing joints and noisy detections: epipolar consistency enables label-free 3D learning [78], while temporal convolutional pipelines show that temporal priors reduce jitter and error compared to frame-wise lifting [18]. Motion-representation learning further suggests that learning priors over dynamics, not only per-frame pose, is a strong regularizer for monocular ambiguity [19,47]. For 2D, self-supervised spatial-aware training is promising but must be validated under occlusion/crowd stress tests [45,70].

Edge deployment constraints should be treated as a first-class objective with end-to-end reporting, not isolated per-image metrics. Heatmap-free single-pass designs and fast CNN estimators target real-time constraints [40,44], while efficient 3D mesh recovery increasingly explores hardware-friendly designs such as binarization and metric-scale recovery [23,91], and real-time monocular 3D remains a practical reference point [94]. Continual learning and test-time adaptation are necessary once models face evolving cameras and environments [122]. Evidence from label-noise studies shows that imperfect supervision can measurably distort training and evaluation [72,73].

Finally, HPE is increasingly a component of embodied AI, where pose quality must be judged by task utility rather than joint error alone. AR/VR and HRI pipelines prioritize stable, low-latency motion [81,82,85], while biomechanics demands physically valid articulation beyond MPJPE-centric optimization [79,80,93]. Multimodal fusion is a practical route to reliability under occlusion and adverse sensing and should be benchmarked under controlled sensor-failure cases [35,36,96,97].

10. Conclusions

In this survey, we reviewed HPE through the design choices that shape both benchmark performance and engineering deployment. We covered body representations, including 2D keypoints, 3D skeletons, and dense meshes; inference pipelines, including top-down, bottom-up, and one-stage methods; learning formulations, including coordinate regression, heatmap-based estimation, detection-style prediction, and distribution-based

representations; and modern neural architectures, including CNNs, Transformers, GCNs, and hybrid models. We also examined representative datasets and evaluation metrics for 2D and 3D HPE, showing how benchmark design influences what models learn, how they are compared, and where they may fail in practical settings.

Despite substantial progress from deep learning, temporal modeling, multi-view geometry, and hybrid sensing, HPE still faces persistent deployment challenges. These include expensive and biased annotations, domain shift, occlusion, depth and scale ambiguity, multi-person association errors, temporal instability, and the trade-off between accuracy and efficiency in real-time or resource-constrained systems. For engineering applications, pose accuracy is often an intermediate metric rather than the final objective. In rehabilitation, robotics, ergonomics, AR/VR, surveillance, and sports analytics, the central question is whether pose estimates support reliable downstream decisions under realistic sensing, latency, memory, and compute constraints.

We also highlighted emerging directions that aim to address these limitations, including data-centric learning, stronger temporal and kinematic modeling, whole-body estimation, uncertainty-aware prediction, generative motion priors, event-based sensing, and edge-oriented model design through compression, quantization, and distillation. These directions suggest that future HPE systems will increasingly be evaluated not only by key-point accuracy, but also by robustness, calibration, temporal consistency, sensing cost, and task-level utility.

As the field matures, future research should move beyond isolated benchmark accuracy toward system-level validation in realistic deployment environments. The central engineering lesson is that no HPE method is universally optimal: the appropriate design depends on the target task, sensing setup, person density, accuracy requirement, compute budget, and acceptable failure modes. By integrating methodological innovation with responsible attention to fairness, privacy, and reliability, HPE can progress toward trustworthy and scalable adoption across healthcare, AR/VR, sports analytics, human-computer interaction, robotics, and other engineering domains.

Author Contributions: Conceptualization, R.S. and A.S.T.; methodology, R.S., A.S.T. and C.W.; validation, R.S., A.S.T. and C.W.; formal analysis, R.S., A.S.T. and C.W.; investigation, R.S., A.S.T. and C.W.; resources, R.S., A.S.T. and C.W.; data curation, R.S. and A.S.T.; writing—original draft preparation, R.S. and A.S.T.; writing—review and editing, R.S., A.S.T. and C.W.; visualization, R.S.; supervision, C.W.; project administration, C.W.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Li, Y.; Cao, H.; Shi, H. A Review on Computer Vision-Based Human Pose Estimation: 2D and 3D Deep Learning Methods. *Comput. Graph.* **2020**, *91*, 131–150. [\[CrossRef\]](#)
2. Munez, T.L.; Jembre, Y.Z.; Weldegebriel, H.T.; Chen, L.; Huang, C.; Yang, C. The Progress of Human Pose Estimation: A Survey and Taxonomy of Models Applied in 2D Human Pose Estimation. *IEEE Access* **2020**, *8*, 133330–133348. [\[CrossRef\]](#)
3. Lu, P.; Jiang, T.; Li, Y.; Li, X.; Chen, K.; Yang, W. RTMO: Towards High-Performance One-Stage Real-Time Multi-Person Pose Estimation. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 1491–1500. [\[CrossRef\]](#)
4. Jiang, T.; Xie, X.; Li, Y. RTMW: Real-Time Multi-Person 2D and 3D Whole-body Pose Estimation. *arXiv* **2024**, arXiv:2407.08634.
5. Yang, J.; Zeng, A.; Zhang, R.; Zhang, L. X-Pose: Detecting Any Keypoints. *arXiv* **2024**, arXiv:2310.08530.
6. Wu, H.; Liang, C.; Liu, M.; Wen, Z. Optimized HRNet for Image Semantic Segmentation. *Expert Syst. Appl.* **2021**, *174*, 114532. [\[CrossRef\]](#)

7. Felzenszwalb, P.F.; Huttenlocher, D.P. Pictorial Structures for Object Recognition. *Int. J. Comput. Vis.* **2005**, *61*, 55–79. [[CrossRef](#)]
8. Toshev, A.; Szegedy, C. DeepPose: Human Pose Estimation via Deep Neural Networks. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2014; pp. 1653–1660. [[CrossRef](#)]
9. Newell, A.; Yang, K.; Deng, J. Stacked Hourglass Networks for Human Pose Estimation. In *Proceedings of the Computer Vision—ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer: Cham, Switzerland, 2016; pp. 483–499. [[CrossRef](#)]
10. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. *arXiv* **2019**, arXiv:1902.09212.
11. Li, Y.; Yang, S.; Liu, P.; Zhang, S.; Wang, Y.; Wang, Z.; Yang, W.; Xia, S.T. SimCC: A Simple Coordinate Classification Perspective for Human Pose Estimation. *arXiv* **2022**, arXiv:2107.03332.
12. Zhang, F.; Zhu, X.; Dai, H.; Ye, M.; Zhu, C. Distribution-Aware Coordinate Representation for Human Pose Estimation. *arXiv* **2019**, arXiv:1910.06278.
13. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *arXiv* **2017**, arXiv:1611.08050.
14. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 22–29 October 2017; pp. 2961–2969. [[CrossRef](#)]
15. Yang, Z.; Zeng, A.; Yuan, C.; Li, Y. Effective Whole-body Pose Estimation with Two-stages Distillation. *arXiv* **2023**, arXiv:2307.15880.
16. Jin, S.; Xu, L.; Xu, J.; Wang, C.; Liu, W.; Qian, C.; Ouyang, W.; Luo, P. Whole-Body Human Pose Estimation in the Wild. *arXiv* **2020**, arXiv:2007.11858.
17. Kreiss, S.; Bertoni, L.; Alahi, A. Pifpaf: Composite fields for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Munich, Germany, 8–14 September 2019; pp. 11977–11986.
18. Pavllo, D.; Feichtenhofer, C.; Grangier, D.; Auli, M. 3D Human Pose Estimation in Video with Temporal Convolutions and Semi-Supervised Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 15–20 June 2019; pp. 7753–7762. [[CrossRef](#)]
19. Zhang, J.; Tu, Z.; Yang, J.; Chen, Y.; Yuan, J. MixSTE: Seq2seq Mixed Spatio-Temporal Encoder for 3D Human Pose Estimation in Video. *arXiv* **2022**, arXiv:2203.00859.
20. Martinez, J.; Hossain, R.; Romero, J.; Little, J.J. A simple yet effective baseline for 3d human pose estimation. *arXiv* **2017**, arXiv:1705.03098.
21. Zheng, C.; Zhu, S.; Mendieta, M.; Yang, T.; Chen, C.; Ding, Z. 3D Human Pose Estimation with Spatial and Temporal Transformers. *arXiv* **2021**, arXiv:2103.10455.
22. Kocabas, M.; Athanasiou, N.; Black, M.J. VIBE: Video Inference for Human Body Pose and Shape Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 13–19 June 2020; pp. 5252–5262. [[CrossRef](#)]
23. Song, C.; Zhang, H.; Yuan, H.; Lin, H.; Tao, J.; Zhang, H.; Yu, T. MetricHMSR: Metric Human Mesh and Scene Recovery from Monocular Images. *arXiv* **2026**, arXiv:2506.09919.
24. Choi, H.; Moon, G.; Chang, J.Y.; Lee, K.M. Beyond Static Features for Temporally Consistent 3D Human Pose and Shape from a Video. *arXiv* **2021**, arXiv:2011.08627.
25. Zheng, C.; Liu, X.; Peng, Q.; Wu, T.; Wang, P.; Chen, C. DiffMesh: A Motion-aware Diffusion Framework for Human Mesh Recovery from Videos. *arXiv* **2025**, arXiv:2303.13397.
26. Tharatipyakul, A.; Srikaewsiew, T.; Pongnumkul, S. Deep learning-based human body pose estimation in providing feedback for physical movement: A review. *Heliyon* **2024**, *10*, e36589. [[CrossRef](#)] [[PubMed](#)]
27. Dang, Q.; Yin, J.; Wang, B.; Zheng, W. Deep Learning-Based 2D Human Pose Estimation: A Survey. *Tsinghua Sci. Technol.* **2019**, *24*, 663–676. [[CrossRef](#)]
28. Zheng, C.; Wu, W.; Chen, C.; Yang, T.; Zhu, S.; Shen, J.; Kehtarnavaz, N.; Shah, M. Deep Learning-based Human Pose Estimation: A Survey. *ACM Comput. Surv.* **2023**, *56*, 1–37. [[CrossRef](#)]
29. Neupane, R.B.; Li, K.; Boka, T.F. A survey on deep 3D human pose estimation. *Artif. Intell. Rev.* **2025**, *58*, 24. [[CrossRef](#)]
30. Felzenszwalb, P.; McAllester, D.; Ramanan, D. A discriminatively trained, multiscale, deformable part model. In *Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 24–26 June 2008; pp. 1–8. [[CrossRef](#)]
31. Kanazawa, A.; Black, M.J.; Jacobs, D.W.; Malik, J. End-to-end Recovery of Human Shape and Pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7122–7131. [[CrossRef](#)]
32. Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; Black, M.J. SMPL: A skinned multi-person linear model. *ACM Trans. Graph.* **2015**, *34*, 2818013. [[CrossRef](#)]

33. Shotton, J.; Fitzgibbon, A.; Cook, M.; Sharp, T.; Finocchio, M.; Moore, R.; Kipman, A.; Blake, A. Real-Time Human Pose Recognition in Parts from a Single Depth Image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 1297–1304. [\[CrossRef\]](#)
34. Isakov, K.; Burkov, E.; Lempitsky, V.; Malkov, Y. Learnable Triangulation of Human Pose. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7717–7726. [\[CrossRef\]](#)
35. Trumble, M.; Gilbert, A.; Malleson, C.; Hilton, A.; Collomosse, J. Total Capture: 3D Human Pose Estimation Fusing Video and Inertial Sensors. In Proceedings of the 28th British Machine Vision Conference, London, UK, 4–7 September 2017; pp. 1–13. Available online: <http://cvssp.org/data/totalcapture/> (accessed on 12 June 2026).
36. Cho, H.; Kim, T.; Jeong, Y.; Yoon, K.J. A Benchmark Dataset for Event-Guided Human Pose Estimation and Tracking in Extreme Conditions. In *Proceedings of the Advances in Neural Information Processing Systems*; Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2024; Volume 37, pp. 134826–134840. [\[CrossRef\]](#)
37. Shi, D.; Wei, X.; Li, L.; Ren, Y.; Tan, W. End-to-End Multi-Person Pose Estimation with Transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2022; pp. 11059–11068. [\[CrossRef\]](#)
38. Yang, J.; Zeng, A.; Liu, S.; Li, F.; Zhang, R.; Zhang, L. Explicit Box Detection Unifies End-to-End Multi-Person Pose Estimation. *arXiv* **2023**, arXiv:2302.01593.
39. Liu, H.; Chen, Q.; Tan, Z.; Liu, J.J.; Wang, J.; Su, X.; Li, X.; Yao, K.; Han, J.; Ding, E.; et al. Group Pose: A Simple Baseline for End-to-End Multi-person Pose Estimation. *arXiv* **2023**, arXiv:2308.07313.
40. Maji, D.; Nagori, S.; Mathew, M.; Poddar, D. YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss. *arXiv* **2022**, arXiv:2204.06806.
41. Jiang, T.; Lu, P.; Zhang, L.; Ma, N.; Han, R.; Lyu, C.; Li, Y.; Chen, K. RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose. *arXiv* **2023**, arXiv:2303.07399.
42. Li, Y.; Wang, C.; Cao, Y.; Liu, B.; Luo, Y.; Zhang, H. A-HRNet: Attention Based High Resolution Network for Human pose estimation. In *2020 Second International Conference on Transdisciplinary AI (TransAI)*; IEEE: Piscataway, NJ, USA, 2020; pp. 75–79. [\[CrossRef\]](#)
43. Zakir, A.; Salman, S.A.; Takahashi, H. SOCA-PRNet: Spatially Oriented Attention-Infused Structured-Feature-Enabled PoseRes-Net for 2D Human Pose Estimation. *Sensors* **2024**, *24*, 110. [\[CrossRef\]](#) [\[PubMed\]](#)
44. Papaioannidis, C.; Mademlis, I.; Pitas, I. Fast CNN-based single-person 2D human pose estimation for autonomous systems. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 1262–1275. [\[CrossRef\]](#)
45. Yang, X.; Niu, T. SSpose: Self-Supervised Spatial-Aware Model for Human Pose Estimation. *IEEE Trans. Artif. Intell.* **2024**, *5*, 5403–5417. [\[CrossRef\]](#)
46. Li, W.; Liu, H.; Tang, H.; Wang, P.; Gool, L.V. MHFormer: Multi-Hypothesis Transformer for 3D Human Pose Estimation. *arXiv* **2022**, arXiv:2111.12707.
47. Zhu, W.; Ma, X.; Liu, Z.; Liu, L.; Wu, W.; Wang, Y. MotionBERT: A Unified Perspective on Learning Human Motion Representations. *arXiv* **2023**, arXiv:2210.06551.
48. Mehraban, S.; Adeli, V.; Taati, B. MotionAGFormer: Enhancing 3D Human Pose Estimation with a Transformer-GCNFormer Network. *arXiv* **2023**, arXiv:2310.16288.
49. Khirodkar, R.; Bagautdinov, T.; Martinez, J.; Zhaoen, S.; James, A.; Selednik, P.; Anderson, S.; Saito, S. Sapiens: Foundation for Human Vision Models. *arXiv* **2024**, arXiv:2408.12569.
50. Wang, C.; Li, Y.; Xiong, Z.; Luo, Y.; Cao, Y. Lower body rehabilitation dataset and model optimization. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*; IEEE: Piscataway, NJ, USA, 2021; pp. 1–6. [\[CrossRef\]](#)
51. Li, Y.; Wang, C.; Cao, Y.; Liu, B.; Tan, J.; Luo, Y. Human pose estimation based in-home lower body rehabilitation system. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8. [\[CrossRef\]](#)
52. Zhao, W.; Tian, Y.; Ye, Q.; Jiao, J.; Wang, W. GraFormer: Graph Convolution Transformer for 3D Pose Estimation. *arXiv* **2021**, arXiv:2109.08364.
53. Wei, W.L.; Lin, J.C.; Liu, T.L.; Liao, H.Y.M. Capturing Humans in Motion: Temporal-Attentive 3D Human Pose and Shape Estimation from Monocular Video. *arXiv* **2022**, arXiv:2203.08534.
54. Xiong, Z.; Wang, C.; Li, Y.; Luo, Y.; Cao, Y. Swin-Pose: Swin Transformer Based Human Pose Estimation. In Proceedings of the 2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR), Virtual, 2–4 August 2022; pp. 228–233. [\[CrossRef\]](#)
55. Wang, C.; Xiong, Z.; Li, Y.; Cao, Y.; Luo, Y. TransNet: Parallel encoder architecture for human pose estimation. *J. Smart Health* **2023**, *28*, 100395. [\[CrossRef\]](#)

56. Yuan, Y.; Fu, R.; Huang, L.; Lin, W.; Zhang, C.; Chen, X.; Wang, J. HRFormer: High-Resolution Vision Transformer for Dense Prediction. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 7281–7293.
57. Li, Y.; Zhang, S.; Wang, Z.; Yang, S.; Yang, W.; Xia, S.T.; Zhou, E. TokenPose: Learning Keypoint Tokens for Human Pose Estimation. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 11293–11302. [\[CrossRef\]](#)
58. Yang, S.; Quan, Z.; Nie, M.; Yang, W. TransPose: Keypoint Localization via Transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2021; pp. 11782–11792. [\[CrossRef\]](#)
59. Xu, Y.; Zhang, J.; Zhang, Q.; Tao, D. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. *arXiv* **2022**, arXiv:2204.12484.
60. Xu, Y.; Zhang, J.; Zhang, Q.; Tao, D. ViTPose++: Vision Transformer for Generic Body Pose Estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 1212–1230. [\[CrossRef\]](#) [\[PubMed\]](#)
61. Shahjahan, A.T.M.; Hamza, A.B. Flexible graph convolutional network for 3D human pose estimation. *arXiv* **2024**, arXiv:2407.19077.
62. Yu, B.X.B.; Zhang, Z.; Liu, Y.; Zhong, S.H.; Liu, Y.; Chen, C.W. GLA-GCN: Global-local Adaptive Graph Convolutional Network for 3D Human Pose Estimation from Monocular Video. *arXiv* **2023**, arXiv:2307.05853.
63. Zhao, J.; Li, J.; Chen, W.; Wang, W.; Yuan, P.; Zhang, X.; Peng, D. 3D-UGCN: A Unified Graph Convolutional Network for Robust 3D Human Pose Estimation from Monocular RGB Images. *arXiv* **2024**, arXiv:2407.16137.
64. Yan, S.; Xiong, Y.; Lin, D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. *arXiv* **2018**, arXiv:1801.07455.
65. Chen, H.; Feng, R.; Wu, S.; Xu, H.; Zhou, F.; Liu, Z. 2D Human Pose Estimation: A Survey. *arXiv* **2022**, arXiv:2204.07370.
66. Liu, W.; Bao, Q.; Sun, Y.; Mei, T. Recent advances of monocular 2D and 3D human pose estimation: A deep learning perspective. *ACM Comput. Surv.* **2022**, *55*, 1–41. [\[CrossRef\]](#)
67. Mao, W.; Ge, Y.; Shen, C.; Tian, Z.; Wang, X.; Wang, Z. TFPose: Direct Human Pose Estimation with Transformers. *arXiv* **2021**, arXiv:abs/2103.15320.
68. Zakir, A.; Salman, S.A.; Benitez-Garcia, G.; Takahashi, H. AECA-PRNetCC: Adaptive Efficient Channel Attention-based PoseResNet for Coordinate Classification in 2D Human Pose. In *2023 38th International Conference on Image and Vision Computing New Zealand (IVCNZ)*; IEEE: Piscataway, NJ, USA, 2023; pp. 1–6. [\[CrossRef\]](#)
69. Xiao, B.; Wu, H.; Wei, Y. Simple baselines for human pose estimation and tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 466–481.
70. Cheng, Y.; Yang, B.; Wang, B.; Tan, R.T. 3D Human Pose Estimation using Spatio-Temporal Networks with Explicit Occlusion Training. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020. [\[CrossRef\]](#)
71. Xu, L.; Zhao, L.; Sun, X.; Wang, D.; Li, G.; Yan, K. A Comprehensive Framework for Occluded Human Pose Estimation (DAG Framework). *arXiv* **2024**, arXiv:2401.00155.
72. Johnson, S.; Everingham, M. Learning effective human pose estimation from inaccurate annotation. In Proceedings of the CVPR 2011, Providence, RI, USA, 20–25 June 2011; pp. 1465–1472. [\[CrossRef\]](#)
73. Schwarz, A.; Hernadi, L.; Biessmann, F.; Hildebrand, K. The Influence of Faulty Labels in Data Sets on Human Pose Estimation. In *Proceedings of the Computer Science Research Notes*; University of West Bohemia: Pilsen, Czech Republic, 2025. [\[CrossRef\]](#)
74. Alzughairbi, A.; Chaczko, Z. Human detection model using feature extraction method in video frames. In Proceedings of the 2016 International Conference on Image and Vision Computing New Zealand (IVCNZ), Auckland, New Zealand, 23–24 November 2016; pp. 1–6. [\[CrossRef\]](#)
75. Guo, Y.; Gao, T.; Dong, A.; Jiang, X.; Zhu, Z.; Wang, F. A Survey of the State of the Art in Monocular 3D Human Pose Estimation: Methods, Benchmarks, and Challenges. *Sensors* **2025**, *25*, 2409. [\[CrossRef\]](#) [\[PubMed\]](#)
76. Zhang, S.; Wang, C.; Dong, W.; Fan, B. A Survey on Depth Ambiguity of 3D Human Pose Estimation. *Appl. Sci.* **2022**, *12*, 10591. [\[CrossRef\]](#)
77. Ponton, J.L.; Yun, H.; Aristidou, A.; Andujar, C.; Pelechano, N. SparsePoser: Real-time Full-body Motion Reconstruction from Sparse Data. *ACM Trans. Graph.* **2023**, *43*, 1–14. [\[CrossRef\]](#)
78. Kocabas, M.; Karagoz, S.; Akbas, E. Self-Supervised Learning of 3D Human Pose using Multi-view Geometry. *arXiv* **2019**, arXiv:1903.02330.
79. Gomez, L.F.; Garrido-Lopez, G.; Fierrez, J.; Morales, A.; Tolosana, R.; Rueda, J.; Navarro, E. Comparison of Visual Trackers for Biomechanical Analysis of Running. In Proceedings of the 2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG), Clearwater, FL, USA, 26–30 May 2025; pp. 1–8. [\[CrossRef\]](#)
80. Cotton, R.J.; DeLillo, A.; Cimorelli, A.; Shah, K.; Peiffer, J.D.; Anarwala, S.; Abdou, K.; Karakostas, T. Markerless Motion Capture and Biomechanical Analysis Pipeline. *arXiv* **2023**, arXiv:2303.10654.
81. Anvari, T.; Park, K.; Kim, G. Upper Body Pose Estimation Using Deep Learning for a Virtual Reality Avatar. *Appl. Sci.* **2023**, *13*, 2460. [\[CrossRef\]](#)

82. Yang, J.; Chen, T.; Qin, F.; Lam, M.S.; Landay, J.A. HybridTrak: Adding Full-Body Tracking to VR Using an Off-the-Shelf Webcam. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 30 April–5 May 2022. [\[CrossRef\]](#)
83. He, G.F.; Kang, S.K.; Song, W.C.; Jung, S.T. Real-time gesture recognition using 3D depth camera. In Proceedings of the 2011 IEEE 2nd International Conference on Software Engineering and Service Science, Beijing, China, 15–17 July 2011; pp. 187–190. [\[CrossRef\]](#)
84. Tran, D.T.; Truong, M.N.; Cho, Y.S. Real-Time Hand Gesture Spotting and Recognition Using RGB-D Camera and 3D Convolutional Neural Network. *Appl. Sci.* **2020**, *10*, 722. [\[CrossRef\]](#)
85. Kim, D.; Kim, D.; Choi, J.; Woo, W. Augmented Pointing Gesture Estimation for Human-Robot Interaction. *Sensors* **2022**, *22*, 3293. [\[CrossRef\]](#) [\[PubMed\]](#)
86. Gao, B.; Li, L.; Zhang, R. Human–Machine Interaction and Implementation on the Upper Limb Movement of Humanoid Robot Using Kinect and MediaPipe. *Discov. Appl. Sci.* **2024**, *6*, 152. [\[CrossRef\]](#)
87. González-Jiménez, H.; Castaño-Candamil, O.; de-la Higuera, P.C.; García-Varea, I. Real-Time Human Pose Recognition Using RGB-D Cameras: A Review. *Sensors* **2019**, *19*, 728. [\[CrossRef\]](#) [\[PubMed\]](#)
88. Xiu, Y.; Li, J.; Wang, H.; Fang, Y.; Lu, C. Pose Flow: Efficient Online Pose Tracking. *arXiv* **2018**, arXiv:1802.00977.
89. Ning, G.; Huang, H. LightTrack: A Generic Framework for Online Top-Down Human Pose Tracking. *arXiv* **2019**, arXiv:1905.02822.
90. Andriluka, M.; Iqbal, U.; Insafutdinov, E.; Pishchulin, L.; Milan, A.; Gall, J.; Schiele, B. PoseTrack: A Benchmark for Human Pose Estimation and Tracking. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5167–5176. [\[CrossRef\]](#)
91. Li, Z.; Zhang, Y.; Lin, J.; Qin, H.; Gu, J.; Yuan, X.; Kong, L.; Yang, X. BinaryHPE: 3D Human Pose and Shape Estimation via Binarization. *arXiv* **2025**, arXiv:2311.14323
92. Hoang, T.; Zehni, M.; Phan, H.; Vo, D.; Do, M. Improving the Robustness of 3D Human Pose Estimation: A Benchmark Dataset and Learning from Noisy Input. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2024), Seattle, WA, USA, 17–18 June 2024; pp. 113–123. [\[CrossRef\]](#)
93. Koleini, F.; Saleem, M.U.; Wang, P.; Xue, H.; Helmy, A.; Fenwick, A. BioPose: Biomechanically-accurate 3D Pose Estimation from Monocular Videos. *arXiv* **2025**, arXiv:2501.07800.
94. Mehta, D.; Sridhar, S.; Sotnychenko, O.; Rhodin, H.; Shafiei, M.; Seidel, H.P.; Xu, W.; Casas, D.; Theobalt, C. VNect: Real-Time 3D Human Pose Estimation With a Single RGB Camera. *ACM Trans. Graph. (TOG)* **2017**, *36*, 44. [\[CrossRef\]](#)
95. Sarafianos, N.; Boteanu, B.; Hu, X.; Kakadiaris, I.A. 3D Human Pose Estimation: A Review of the Literature and Analysis of Covariates. *Comput. Vis. Image Underst.* **2016**, *152*, 1–20. [\[CrossRef\]](#)
96. Xefferis, V.; Syropoulou, A.C.; Pistola, T.; Kasnesis, P.; Poullos, I.; Tsanousa, A.; Symeonidis, S.; Diplaris, S.; Goulianas, K.; Chatzimisios, P.; et al. Multimodal fusion of inertial sensors and single RGB camera data for 3D human pose estimation based on a hybrid LSTM-Random forest fusion network. *Internet Things* **2024**, *29*, 101465. [\[CrossRef\]](#)
97. Rebecq, H.; Gallego, G.; Delbruck, T.; Scaramuzza, D. Events-to-Frame Reconstruction and Event-Based Visual-Inertial Odometry Using Spiking Neural Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 1125–1140. [\[CrossRef\]](#)
98. Chakravarthi, B.; Verma, A.A.; Daniilidis, K.; Fermüller, C. Recent Event Camera Innovations: A Survey. In *Computer Vision Advances*; Springer: Berlin/Heidelberg, Germany, 2024; Chapter 21. [\[CrossRef\]](#)
99. Andriluka, M.; Pishchulin, L.; Gehler, P.; Schiele, B. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3686–3693. [\[CrossRef\]](#)
100. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014; p. 740.
101. Li, J.; Wang, C.; Zhu, H.; Mao, Y.; Fang, H.S.; Lu, C. CrowdPose: Efficient Crowded Scenes Pose Estimation and a New Benchmark. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 10855–10864. [\[CrossRef\]](#)
102. Johnson, S.; Everingham, M. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. In *Proceedings of the British Machine Vision Conference*; BMVA Press: Durham, UK, 2010; pp. 12.1–12.11. [\[CrossRef\]](#)
103. Sapp, B.; Taskar, B. MODEC: Multimodal Decomposable Models for Human Pose Estimation. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3674–3681. [\[CrossRef\]](#)
104. Ionescu, C.; Papava, D.; Olaru, V.; Sminchisescu, C. Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 1325–1339. [\[CrossRef\]](#)
105. Mehta, D.; Rhodin, H.; Casas, D.; Fua, P.; Sotnychenko, O.; Xu, W.; Theobalt, C. Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision. *arXiv* **2017**, arXiv:1611.09813.
106. Ghorbani, S.; Mahdavian, K.; Thaler, A.; Kording, K.; Cook, D.J.; Blohm, G.; Troje, N.F. MoVi: A large multi-purpose human motion and video dataset. *PLoS ONE* **2021**, *16*, e0253157. [\[CrossRef\]](#) [\[PubMed\]](#)

107. Marcard, T.; Henschel, R.; Black, M.; Rosenhahn, B.; Pons-Moll, G. Recovering Accurate 3D Human Pose in the Wild Using IMUs and a Moving Camera. In *Proceedings of the 15th European Conference, Munich, Germany, 8–14 September 2018, Proceedings, Part X*; Springer: Cham, Switzerland, 2018; pp. 614–631. [\[CrossRef\]](#)
108. Varol, G.; Romero, J.; Martin, X.; Mahmood, N.; Black, M.J.; Laptev, I.; Schmid, C. Learning from Synthetic Humans. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2017; pp. 4627–4635. [\[CrossRef\]](#)
109. Wu, J.; Zheng, H.; Zhao, B.; Li, Y.; Yan, B.; Liang, R.; Wang, W.; Zhou, S.; Lin, G.; Fu, Y.; et al. Large-Scale Datasets for Going Deeper in Image Understanding. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*; IEEE: Piscataway, NJ, USA, 2019; pp. 1480–1485. [\[CrossRef\]](#)
110. Zhang, W.; Zhu, M.; Derpanis, K.G. From Actemes to Action: A Strongly-Supervised Representation for Detailed Action Understanding. In *2013 IEEE International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2013; pp. 2248–2255. [\[CrossRef\]](#)
111. Lin, W.; Liu, H.; Liu, S.; Li, Y.; Xiong, H.; Qi, G.; Sebe, N. HiEve: A Large-Scale Benchmark for Human-Centric Video Analysis in Complex Events. *Int. J. Comput. Vis.* **2023**, *131*, 2994–3018. [\[CrossRef\]](#)
112. Fabbri, M.; Lanzi, F.; Calderara, S.; Palazzi, A.; Vezzani, R.; Cucchiara, R. Learning to Detect and Track Visible and Occluded Body Joints in a Virtual World. *arXiv* **2018**, arXiv:1803.08319. [\[CrossRef\]](#)
113. Sigal, L.; Balan, A.O.; Black, M.J. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *Int. J. Comput. Vis.* **2010**, *87*, 4–27. [\[CrossRef\]](#)
114. Eichner, M.; Marin-Jimenez, M.; Zisserman, A.; Ferrari, V. 2D Articulated Human Pose Estimation and Retrieval in (Almost) Unconstrained Still Images. *Int. J. Comput. Vis.* **2012**, *99*, 190–214. [\[CrossRef\]](#)
115. Ning, G.; Zhang, Z.; He, Z. Knowledge-Guided Deep Fractal Neural Networks for Human Pose Estimation. *IEEE Trans. Multimed.* **2018**, *20*, 1246–1259. [\[CrossRef\]](#)
116. Cai, Y.; Wang, Z.; Luo, Z.; Yin, B.; Du, A.; Wang, H.; Zhang, X.; Zhou, X.; Zhou, E.; Sun, J. Learning Delicate Local Representations for Multi-Person Pose Estimation. *arXiv* **2020**, arXiv:2003.04030.
117. Huang, J.; Zhu, Z.; Guo, F.; Huang, G.; Du, D. The Devil is in the Details: Delving into Unbiased Data Processing for Human Pose Estimation. *arXiv* **2020**, arXiv:1911.07524.
118. Chen, Y.; Shen, C.; Wei, X.S.; Liu, L.; Yang, J. Adversarial PoseNet: A Structure-aware Convolutional Network for Human Pose Estimation. *arXiv* **2017**, arXiv:1705.00389.
119. Zhang, F.; Zhu, X.; Ye, M. Fast Human Pose Estimation. *arXiv* **2019**, arXiv:1811.05419.
120. Bazarevsky, V.; Grishchenko, I.; Raveendran, K.; Zhu, T.; Zhang, F.; Grundmann, M. BlazePose: On-device Real-time Body Pose tracking. *arXiv* **2020**, arXiv:2006.10204.
121. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. DINOv2: Learning Robust Visual Features without Supervision. *arXiv* **2024**, arXiv:2304.07193.
122. Khan, M.S.U.; Stricker, D. PoseAdapt: Sustainable Human Pose Estimation via Continual Learning Benchmarks and Toolkit. *arXiv* **2025**, arXiv:2409.20469.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.