

Article

Techniques of 2D Human Pose Estimation Based on Computer Vision: A Survey

Deyu Lin ^{1,2}, Yujie Zhang ¹, Yang Yu ¹, Shuaibo Gao ¹, Lu Zhou ¹ and Yufei Zhao ^{2,*} 

¹ School of Software, Nanchang University, Nanchang 330047, China; deylin@ncu.edu.cn (D.L.); yjzhang@ncu.edu.cn (Y.Z.); yyu@ncu.edu.cn (Y.Y.); sgao@ncu.edu.cn (S.G.); luzhou@ncu.edu.cn (L.Z.)

² School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore

* Correspondence: yufei.zhao@ntu.edu.sg

Abstract

Two-dimensional (2D) human pose estimation is one of the key research directions in Computer Vision (CV), which has wide application prospects in behavior recognition, such as gesture tracking, intelligent monitoring, and identity recognition. Therefore, it has recently attracted extensive attention from academia and industry. However, although a large amount of literature has been published, existing reviews often lack a unified theoretical perspective and fail to capture the latest paradigm shifts brought by foundation models. To this end, this paper reviews the applications of deep learning in the domain of 2D body pose estimation from 2010 to 2025 through a cascading approach. First, the mainstream body pose datasets and related evaluation metrics are introduced in a comprehensive and convincing way through mathematical formulas. Subsequently, an in-depth analysis of the performance of the algorithms in single-person and multi-person scenarios, and a comprehensive comparative analysis of the strengths and weaknesses of each algorithmic model, are conducted. A comprehensive comparative analysis encompassing both traditional architectures and the latest deep learning breakthroughs are provided, specifically highlighting Vision Foundation Models (VFM), generative Diffusion processes, and State Space Models (SSMs). Finally, the current state of research in the field of 2D human pose estimation is summarized, and three main challenges, emerging solutions, and expected development trends are pointed out. This survey is an exhaustive compilation of existing research in 2D human pose estimation, providing a blueprint for researchers in the field and laying the foundation for future research work.

Keywords: computer vision; 2D human pose estimation; deep learning; convolutional neural network; joint point detection



Academic Editor: George A. Tsihrintzis

Received: 5 April 2026

Revised: 18 May 2026

Accepted: 20 May 2026

Published: 25 June 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 refers to the technology of using intelligent devices, such as computers, cameras, sensors, etc., to analyze images in order to understand and obtain the positional information and directions of different parts of the human body, thus achieving human postures detection [1]. As one of the most crucial research areas in Computer Vision (CV), it aims to determine the locations of keypoints of the human body from images or videos, which usually include major joints such as the head, shoulders, elbows, wrists, hips, knees, and ankles. By means of predicting the location of these major joints, human pose estimation can be achieved successfully [2]. The accuracy in human pose estimation

results is a prerequisite for further action behavior detection and analysis. Recently, human pose estimation has gained extensive applications in action recognition [3], advanced human–computer interaction [4], intelligent surveillance [5,6], image-based behavior annotation [7], and gait-based identity recognition [8]. However, the structural complexity of the human body and the diversity of behavioral actions, occlusions, environmental factors, and viewpoint variations pose considerable challenges to its further development and application.

Before the application of Deep Learning in human pose estimation, most of the studies put emphasis on the graphical structures (such as the improvement of human–computer interaction, Augmented Reality (AR), Virtual Reality (VR) applications, etc.) [9]. Early feature extraction methods in human pose estimation, often referred to as ‘hand-crafted’ methods, focused on low-level feature areas, such as edges, colors, and light streams. Another common approach was to use foreground masks as features when the background could be removed. However, these hand-crafted methods also have limitations. For example, when dealing with background removal, there is a risk of inadequate image segmentation, which usually leads to the loss of intricate feature-based information. To address this, researchers proposed techniques such as Scale-Invariant Feature Transform (SIFT) [9] and Histogram of Oriented Gradients (HOG) [10] to obtain features that are more expressive and effectively compressed in the feature space. Despite these advancements, the manual methods remain challenged in pinpointing the exact location of the body part. In contrast, the Deep Learning-based (DL-based) method extracts image data as well as features through CV, which has obvious advantages [11] over the traditional manual method [12,13].

Research on the theory of Deep Learning has mushroomed and has been successfully applied in various fields, such as speech and image recognition [14,15]. The combination of human pose estimation with Deep Learning is of profound significance. Generally speaking, traditional image recognition methods struggle to handle such complex tasks. The advent of Convolutional Neural Networks (CNNs) makes it possible to analyze human pose actions not only in still images, but also in video-based human pose action analysis in some models, with excellent performance [16]. With the adoption of Deep Learning, researchers can precisely extract human pose information from static images and progressively extend this methodology to the temporal video poses analysis. In addition, algorithms that fuse spatial–temporal features have solved the problem of feature extraction in time series data effectively, enabling Deep Learning models to perform well in human pose estimation [17]. Compared with the traditional image recognition method, the human pose estimation based on Deep Learning can quickly fit human pose information in the sample labels, thus producing a model with the capability of pose analysis.

In recent years, reviews on human pose estimation based on Computer Vision have attracted extensive attention from academia and industry all over the world [18–20]. For example, Murphy-Chutorian et al. reviewed the research on head pose estimation, discussed the difficulties in head pose estimation and highlighted the advantages and disadvantages of various algorithms [21]. A review on gesture recognition was conducted by Saroja et al., who comprehensively analyzed various studies related to gesture recognition through cameras and sensors with regard to their principles, implementation, algorithms, performance, advantages, and challenges [22]. Mitra et al. discussed the application of various techniques in gesture recognition, including hidden Markov models, particle filtering and coalescence, finite state machines, optical flow, skin color, and connectivity models in detail [23]. Liu et al. presented a review of body part parsing methods for human pose estimation [24]. All of the aforementioned tutorials focused on the specific area of human postures. Zhang et al. emphasized approaches that achieve human pose detection using depth and RGB

image data [25]. Gong et al. paid attention to the aspect of monocular images and found that most of the research was conducted through handcrafted features and models [26]. Sun et al. conducted a more comprehensive survey on human pose estimation from 2D monocular images but did not systematically analyze or compare the advantages, disadvantages, and applicability of different computational methods [27]. Zheng et al. provided an overview on deep learning-based 2D and 3D posture estimation over the last few years, including the datasets, the evaluation metrics, and the applications [28]. They pointed out the challenges of human pose estimation and the future research directions concerning these challenges. While these surveys analyzed 2D and 3D human pose estimation technologies in detail, the rapid ascent of 3D posture assessment has led some to question the continued relevance of 2D research. We argue, however, that 2D estimation remains the foundational and irreplaceable cornerstone of the field. Specifically, 2D keypoints act as the critical “perceptual front-end” for most 3D lifting pipelines, where 2D precision directly dictates the theoretical ceiling of 3D reconstruction accuracy. Furthermore, for real-time edge deployment on resource-constrained devices—such as mobile edge computing and IoT-based monitoring—lightweight 2D frameworks remain the only viable choice compared to computationally prohibitive 3D mesh recovery. Considering the global prevalence of monocular RGB infrastructures lacking depth sensors, robust 2D estimation is practically essential for extracting semantic value from legacy video streams.

Despite its foundational importance, the existing literature—including the most recent comprehensive reviews—often exhibits critical shortcomings. Many surveys lack the corresponding mathematical expressions for the presented evaluation metrics, resulting in an analysis that is short in intuition and persuasiveness. Additionally, in the review of heatmap-based methods for 2D single-person posture estimation, the development history and the specific incremental improvements of each method are often not clearly categorized, which obscures the logical evolution of the field. Furthermore, while Zheng et al. identified three major challenges currently faced by HPE—detecting individuals under severe occlusion, improving computational efficiency, and handling insufficient data—the corresponding research tendencies and solutions were not illustrated in detail.

Table 1 summarizes the aforementioned reviews or surveys in terms of the characteristics, advantages, and disadvantages.

To bridge the aforementioned gaps, this paper moves beyond a simple chronological summary to provide a critical synthesis of deep learning-based 2D human pose estimation techniques from 2010 to 2025. Our core objective is to uncover the underlying design philosophies of these algorithms, specifically exploring how they navigate the inherent trade-offs among localization accuracy, computational efficiency, and robustness in complex scenes. Through this comparative analysis, we aim to establish a clear architectural roadmap and offer practical insights for future algorithm selection and deployment. To be specific, the main contributions of this paper are listed as follows:

- (1) The mainstream body pose datasets and related evaluation metrics are introduced in a comprehensive way through mathematical formulas, accompanied by a critical assessment of their applicability in complex scenarios, which can help researchers and practitioners to understand the benchmark of model training and evaluation;
- (2) An in-depth analysis on the performance of these algorithms in single-person and multi-person scenarios is conducted, providing a comprehensive comparative framework that evaluates not only the strengths, but also the structural limitations and computational complexity of each algorithmic model;
- (3) The current state of research in the field of 2D human pose estimation is summarized, and three main challenges and proposed solutions and expected development trends

are pointed out, with the aim to direct future research efforts and help stakeholders understand where the technology is headed.

The rest of this paper is organized as follows. Section 2 compares the relevant dataset and the evaluation metrics in human pose estimation, and Section 3 divides the human pose estimation algorithms into two categories, namely, single-person pose estimation and multi-person pose estimation. In Section 4, the challenges faced by 2D human pose estimation are discussed in detail. Finally, the future research direction of human pose estimation is also envisaged in terms of dataset optimization, data generation, and joint multi-mission exercises.

Table 1. Characteristics, advantages and disadvantages in the study of human pose estimation in the last decade.

	Characteristics	Advantages	Disadvantages
Zhang, H et al. [25]	Identity-invariant head pose estimation	A comprehensive review of head pose estimation methods was presented and compared	Only head pose estimation has been studied, and the study is limited in scope
Gong, W et al. [26]	Human posture estimation based on cameras and sensors	These two methods of human posture estimation were described in detail and compared	No comparisons of the accuracy of the estimation methods were presented
Sun, J et al. [27]	Gesture recognition	Discussed in detail the application of various techniques in gesture recognition	The different gesture recognition methods were not compared, and their performance aspects were not specifically analyzed and compared
Zheng, C et al. [28]	Focus on the body part parsing methods	Discussed in detail the application of human pose estimation and the limitations of existing methods	These methods are very limited for irregular postures and do not specifically investigate solutions
Sapp, B et al. [29]	Methods using depth and RGB image data	Focused on the specific area of human pose estimation, depth and RGB image data were utilized to achieve human pose detection	The focus on depth and RGB image data did not fully encompass other emerging technologies or sensors being developed for human posture estimation
Lin, T et al. [30]	Human posture estimation from monocular images	Detailed description and comparison of the two methods from the images were conducted	No specific comparisons with other methods and the scope of the investigation was not comprehensive enough
Johnson et al. [31]	Human pose estimation from 2D monocular images	Comprehensive survey on human pose estimation from 2D monocular images was conducted	Did not systematically analyze and compare the advantages, disadvantages, and applicability of different computational methods
Andriluka et al. [32]	Deep learning-based 2D and 3D posture estimation	2D and 3D human posture estimation were described in detail	Research trends and solutions corresponding to the three main challenges were not indicated in detail

2. Relevant Datasets and Evaluation Metrics

In 2D human pose estimation, the performance of data-driven models is significantly affected by data quality. Therefore, data is a kind of essential resource for training and testing. In addition, standardized metrics are required to accurately assess algorithmic performance. Consequently, the current mainstream datasets and commonly used

evaluation metrics are introduced before formally reviewing the algorithms of human posture estimation.

Since the quality of data plays a decisive role in the final prediction results, the selection of the dataset is crucial. However, the collection, compilation, and labeling of datasets are very time-consuming. Therefore, most studies are based on the existing datasets. The data sources of these datasets vary from Hollywood movie footage, e.g., FLIC, to images downloaded from search engines, such as Google and Bing, to MSCOCO [29,30]. Consequently, these differing data sources cater to distinct application scenarios.

In 2D human pose estimation, the joint is deemed to be detectable when the distance between the predicted position of joints and the ground truth (the true position of the joint obtained by manual labeling) in 2D human pose estimation is within a predefined threshold. The accuracy varies when different thresholds are utilized. Depending on the selected benchmark, commonly used evaluation metrics include the Percentage of Correct Parts (PCP), Percentage of Correct Keypoints (PCK), Object Keypoint Similarity (OKS), and Average Precision (AP).

2.1. Related Datasets

Owing to limited resources during the early stages of research and disparities in the quantitative description of human poses, most early datasets primarily concentrated on annotating specific body parts of individual subjects. In recent years, many datasets for human pose estimation have been published. Rather than merely listing these datasets chronologically, we can observe their structural evolution through distinct developmental stages, reflecting the field's progression from constrained single-person tasks to unconstrained foundation-level stress tests.

In the formative years of deep learning, progress was predominantly catalyzed by single-person datasets in constrained environments. The Leeds Sports Pose (LSP) [31] dataset, for instance, provided a foundational collection of athletic annotations, while FLIC [32] offered annotations derived from Hollywood movie frames. Although these datasets established early baselines, their limited sample sizes, constrained poses (mostly unobstructed or unfolded), and relatively low image quality no longer meet the requirements of modern high-resolution algorithms. Thus, they have been fundamentally abandoned in contemporary evaluations [33,34].

As algorithms advanced, the academic focus shifted toward multi-person estimation in complex, daily environments. The MPII Human Pose Dataset [35] significantly broadened the research scope by extracting distinct human activities from YouTube videos. Subsequently, the introduction of the MSCOCO [36] dataset marked a critical watershed. With hundreds of thousands of instances annotated with 17 keypoints, it became the absolute gold standard for evaluating deterministic regression models. Similarly, the HKD (AI-Challenger) [37] dataset contributed massive high-resolution images to further challenge keypoint detection in unconstrained scenarios.

To capture kinematic continuity, researchers further expanded into the temporal domain with video-based datasets. Early video-level single-person datasets like Penn Action and J-HMDB paved the way for more complex temporal benchmarks [38,39]. PoseTrack [40] emerged as the first extendable dataset for multi-person pose tracking in videos, while Human-in-Events (HiEve) [41] provided large-scale tracking data for dense crowds and complex real-world events, heavily driving multi-person action recognition.

Despite the historical significance of the aforementioned benchmarks, modern Vision Foundation Models (VFMs) routinely exceed 80 mAP on datasets like MSCOCO. This performance saturation indicates that classical benchmarks no longer fully expose the true bottlenecks of real-world deployment. Consequently, the focus has shifted to-

ward evaluating robustness under extreme semantic ambiguity and out-of-distribution (OOD) scenarios. This necessity has driven the emergence of next-generation stress test datasets from the IEEE community. For instance, UBody [42] (IEEE CVPR, 2023) provides a million-scale dataset specifically designed to evaluate whole-body estimation under severe occlusion across diverse real-world scenes. Complementary to this, HumanArt [43] (IEEE CVPR, 2023) introduces a novel dimension for OOD generalization by testing algorithms across 20 non-photorealistic domains, ranging from cartoons to historical artworks. These benchmarks represent the definitive touchstones for the transition toward zero-shot and foundation-driven paradigms.

The sample annotations and keypoint distributions of these foundational single-person benchmarks are visualized in Figure 1. Figure 2 displays the typical keypoint configurations for the MPII and MSCOCO datasets under unconstrained daily environments. The detailed body part annotations and high-resolution configurations of the AI-Challenger dataset are depicted in Figure 3.

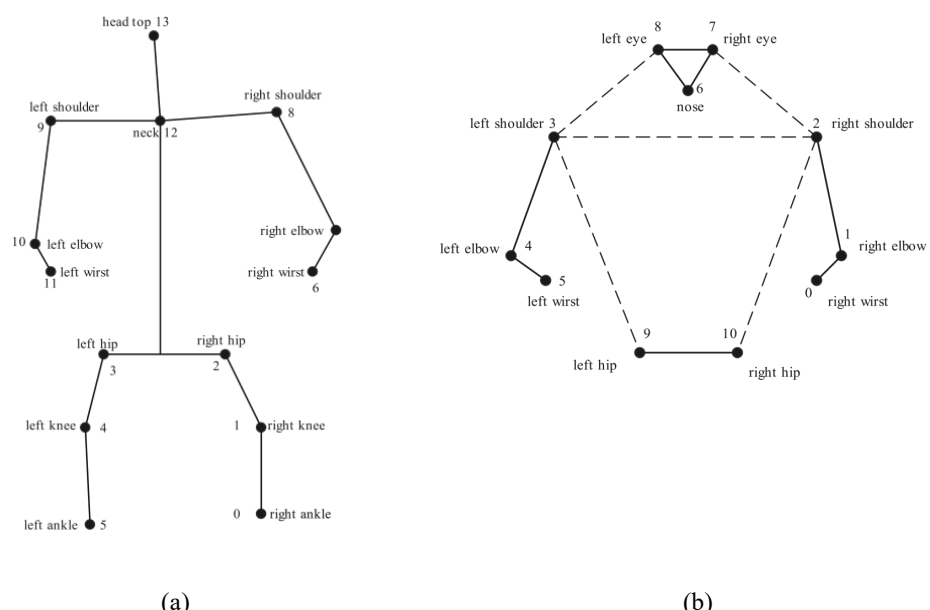


Figure 1. (a) Annotations of LSP dataset [6] and (b) annotations of FLIC [32].

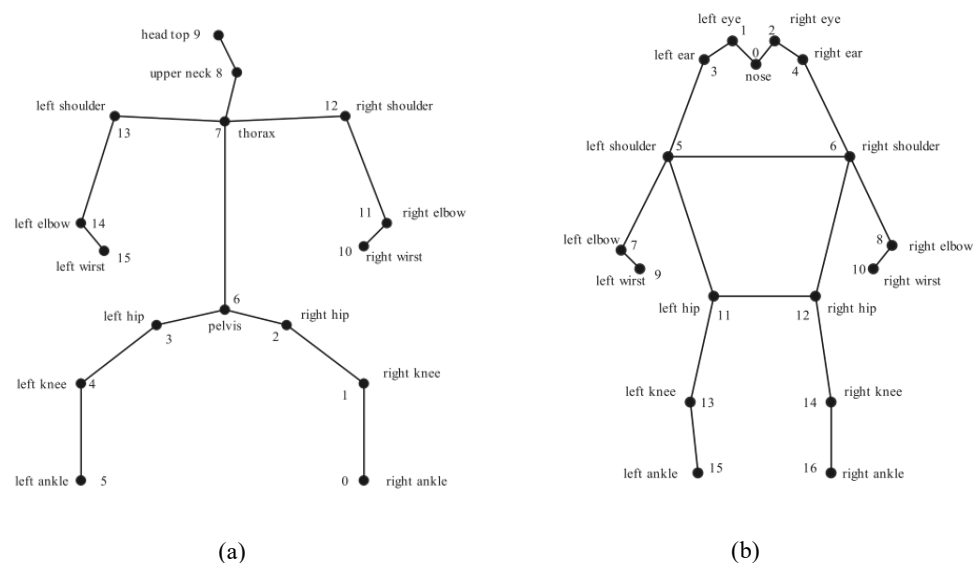


Figure 2. (a) Annotations of MPII dataset [36] and (b) annotations of MSCOCO [30].

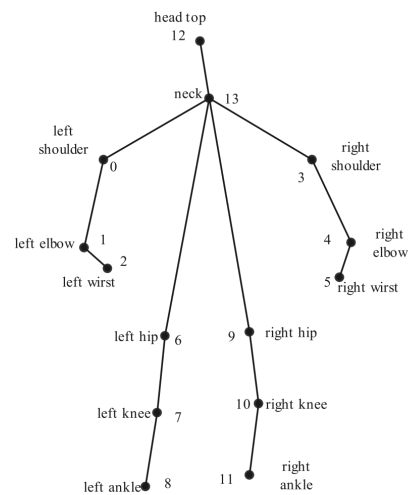


Figure 3. Annotations of AI-challenger dataset [33].

Table 2 exhibits several globally recognized datasets and details the suitable classification, the count of annotated nodes, the total samples, and associated references.

Table 2. Dataset of human pose estimation.

Dataset	Year	Single/Multiple	Number of Joints	Number of Samples/103	Description
LSP [31]	2010	Single	14	2	Full body pose image downloaded from Flickr
FLIC [32]	2013	Single	10	20	Video frames captured by a Hollywood movie
MPII [35]	2014	Single, Multiple	16	25	YouTube downloaded video, manually select the screen in the video
MSCOCO [36]	2014	Multiple	17	300	Pictures downloaded by Google, Bing, and Flickr
HKD [37]	2017	Multiple	14	300	Daily images from internet
Penn Action [39]	2013	Single	13	2	YouTube downloaded videos
PoseTrack [40]	2018	Multiple	15	0.5	Extends MPII and can be used for pose tracking
HiEve [41]	2020	Multiple	14	50	9 real scenes such as the airport, restaurant, and school

2.2. Evaluation Metrics

2.2.1. Percentage of Correct Parts

Percentage of Correct Parts (PCP) is a metric for early pose estimation and mainly adopted to assess the localization accuracy of a limb [43]. In general, a limb is correctly localized if both endpoints of a limb fall within the threshold of the corresponding actual endpoints. To be specific, it can be mathematically expressed as follows.

$$PCP = \frac{\sum_i \delta(d_i < kL_{norm})\delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \quad (1)$$

In Equation (1), d_i represents the Euclidean distance between the predicted location of keypoint i and its ground truth. L_{norm} denotes the standard limb dimension used to delimit the threshold. The indicator function $\delta(\cdot)$ returns 1 if the spatial distance d_i is below a specific proportion k (often 0.5) of L_{norm} , and $v_i > 0$ signifies that the node i is annotated as a visible keypoint.

2.2.2. Percentage of Correct Keypoints (PCK)

PCK indicates the proportion of the normalized distance d_i between the estimated nodes and their corresponding ground truth (i.e., true position), which is less than a set threshold value [44]. On dataset MPII, the threshold is set to the head length L_{head} and is referred to as PCKh, which is defined by (2) as follows.

$$PCKh = \frac{\sum_i \delta(d_i < kL_{head}) \cdot \delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \quad (2)$$

The threshold is normalized by the head length L_{head} (referred to as PCKh), where k is a predefined coefficient. The indicator function $\delta(\cdot)$ validates the prediction as correct if $d_i < k * L_{head}$ is satisfied for visible joints ($v_i > 0$). PCP may focus more on overall limb parts, whereas PCK focuses more on the accuracy of individual keypoints, which is used to evaluate the accuracy of human articulation point localization. When the candidate articulation point ranges within the threshold pixel of the actual one, it signifies the correctness of the candidate articulation point.

2.2.3. Object Keypoint Similarity (OKS)

OKS is a metric designed for multi-individual pose estimation. The similarity between detected joints and real human joint positions is assessed by means of adding a quality prediction module and calculating weighted Euclidean distances for joint positions [45]. It evaluates the similarity between the detected joint point and the corresponding true labeled data according to the weighted Euclidean distance of the joint point positions, which is described as follows.

$$O_{OKS.P} = \frac{\sum_i \exp\left\{-d_{pi}^2 / 2s_p^2 \cdot \delta_i^2\right\} R(v_{pi} = 1)}{\sum_i R(v_i = 1)} \quad (3)$$

At the time of testing, the Euclidean distance (denoted as d_{pi}) represents the distance between the location of the detected joint point and the location of the annotated one in the dataset. The frame number of the detected human is represented by p , while i is the number assigned to each joint point; s is the square root of the area of the detection human frame (human scale factor); and δ_i is the skeletal joint point normalization factor (standard deviation of the offset of the manually annotated joint point location). The state of person number pi is denoted by v_{pi} concerning the i th joint point, encompassing visibility, invisibility, or presence outside the bounding box, etc. $R(v_{pi} = 1)$ is the Kronecker function. In (3), only the visible nodes (when condition $v_{pi} = 1$ is met) are considered in the process of the posture assessment.

2.2.4. Average Precision (AP)

AP is often adopted as a metric for dataset COCO, which is targeted to calculate the accuracy percentage of the testing set and can be used for both single-person and multi-person pose estimations [34]. When used for single-person stance estimation, the formula is listed as follows.

$$AP = \frac{\sum_P \delta(OKS_p > s)}{\sum_P \delta(p)} \quad (4)$$

where s is the OKS threshold, and $\sum_P \delta(p)$ represents the sum of all keypoints in a single-person instance.

For multi-person pose estimation, the formula for AP is the same as those for both top-down and bottom-up methods. They differ in the truth that the matching and grouping methods of keypoints are distinct. Consequently, the precision and recall obtained from

these two methods may be different, thus affecting the final AP value. Based on (4), when different values are set for the artificial threshold in the AP metrics, multiple AP metrics are obtained. Finally, multiple AP metrics are averaged to finally obtain the mean Average Precision (mAP).

Table 3 summarizes the characteristics of the assessment metrics mentioned above.

Table 3. Evaluation metrics and characteristics.

Evaluation Metrics	Characteristics
PCP [43]	A key assessment metric for early pose estimation; primarily used to assess the localization accuracy of a limb
PCK [44]	More widely used; the performance of the model is quantified by evaluating the distance between detected and real joints
OKS [45]	An evaluation of multi-person pose estimation is introduced; the performance of the model is assessed by means of calculating the weighted Euclidean distance between the detected joints and the labeled data
AP [34]	Assessment metrics specific to the COCO dataset; for single and multi-person pose estimation, a harmonized and standardized assessment methodology is provided

2.2.5. Limitations of Distance-Based Metrics

Although metrics like PCP, PCK, and OKS provide a solid mathematical basis for evaluating spatial accuracy, they share a common limitation: they evaluate each keypoint independently. By computing the weighted average of individual point deviations, these metrics essentially ignore the underlying anatomical structure of the human body.

As a result, an estimated pose could be biomechanically impossible—such as a limb bending at an unnatural angle or a joint severely detached from the torso—but still achieve a high OKS score as long as the absolute pixel errors remain small. This discrepancy becomes particularly evident when evaluating modern generative models, which are designed to learn structural priors rather than just local pixel patterns. Relying entirely on point-wise Euclidean distances is therefore insufficient for assessing real-world reliability. To address this, future evaluation protocols need to move beyond simple spatial measurements and incorporate kinematic constraints, ensuring that high accuracy scores actually reflect anatomically correct predictions.

3. Detailed Review Concerning 2D Human Pose Estimation Algorithms

2D human pose estimation lays the foundation for human behavior prediction. As for 2D human pose estimation, it mainly describes the human skeletal information through the 2D coordinates of the keypoints of the human body; therefore, an efficient method to obtain the coordinates of these keypoints is crucial. Deep Learning provides a powerful tool for 2D human pose estimation, which primarily utilizes the superior feature extraction capabilities of convolutional neural networks for keypoint detection and skeleton information recovery to achieve accurate human pose predictions.

Regarding the number of detected individuals, the relevant algorithms can be categorized into two groups, namely, single-person pose estimation and multi-person pose estimation. Multi-person pose estimation can be further divided into two-stage methods and one-stage methods, where two-stage methods include top-down, bottom-up, and top-down and bottom-up combination. The classifications are presented as shown in Figure 4. The strategies of multi-person pose estimation are closely related to the fundamentals of single-person pose estimation. To build a solid foundation for these strategies, it is essential to initially focus on single-person methods.

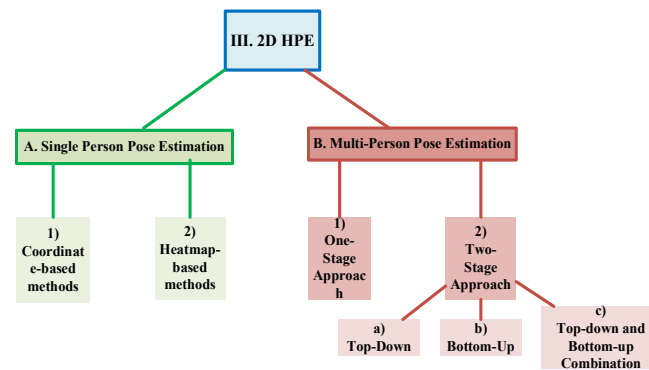


Figure 4. 2D keypoint detection method (Note: The lines represent the estimated skeletal connections between keypoints, and the different colors are used solely to visually distinguish different individuals and body parts, rather than denoting specific quantitative values).

3.1. 2D Single-Person Pose Estimation

In 2D single-person pose estimation, the goal is to estimate the pose of each individual from the input image. If an image contains multiple people, it is automatically segmented into several single-person sub-images through human detection algorithms and image segmentation. Human pose estimation is a branch of the area of Computer Vision. According to the ground truth in the dataset, single-person pose estimation can be divided into two categories, namely, coordinate-based methods [46–51] and heatmap-based methods [52–59].

3.1.1. Coordinate-Based Methods

The coordinate-based single-person pose estimation model aims to obtain the coordinate of each skeletal joint point by means of training a neural network. This approach was widely applied in early research on deep learning-based single-human skeletal joint point detection, known as Coordinate Net (as illustrated in Figure 5). Based on their regression strategies, these models can be further classified into multi-stage direct regression and multi-stage stepwise regression. This classification is illustrated in Figure 6.

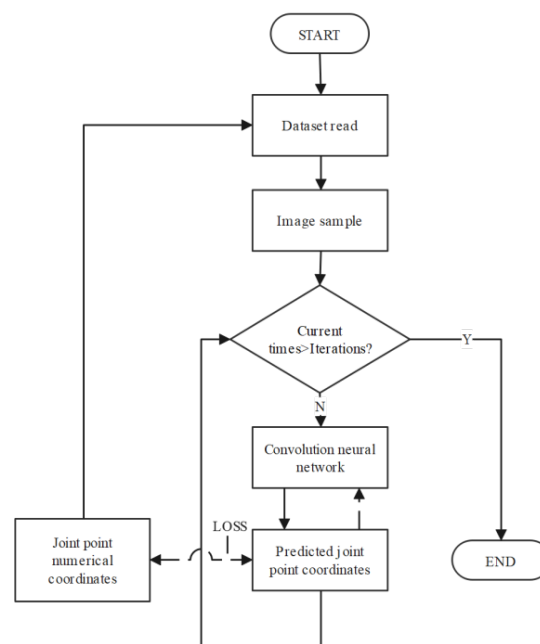


Figure 5. Flowchart of single-person attitude estimation based on coordinate regression (Note: The solid arrows indicate the direction of data flow and sequential processing steps within the framework).

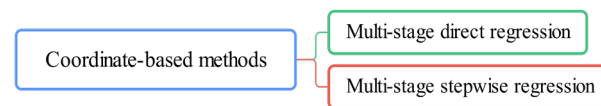


Figure 6. Classification diagram of single-person attitude estimation based on coordinate regression (Note: The different colors are used solely to visually distinguish different individuals and body parts, and do not represent specific quantitative values).

- Multi-stage Direct Regression

DeepPose is one of the early typical Coordinate Nets, which was proposed to apply deep learning research methods to single-person pose estimation [60]. During the initial detection phase of this algorithm, a prediction of the approximate position of the human joint point is made through a Deep Neural Network (DNN). In the refinement phase, multiple DNN-based regressors further optimize these predictions using local sub-images around the initial joints. The algorithm centers around the initial-stage 2D coordinates of human body keypoints, then small image patches are cropped from the domain around the center point and fed into the regression operation of this stage. This process provides the neural network model with more detailed information about the joint point imagery, thus facilitating the correction of the 2D coordinate values. The specific algorithm flow is presented as shown in Figure 7.

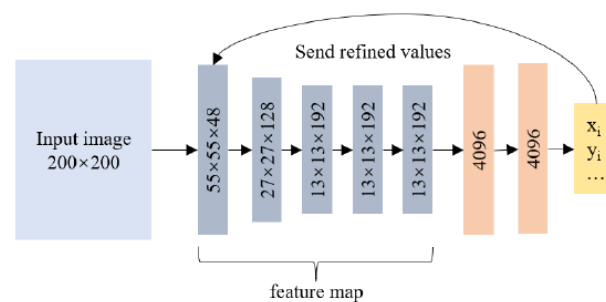


Figure 7. Deep pose structure diagram (Note: The solid arrows indicate the direction of data flow and sequential processing steps within the framework).

- Multi-stage stepwise regression

The Iterative Error Feedback (IEF) model is another variant of coordinate-based regression networks [61]. In contrast to DeepPose, which uses multi-stage direct regression to obtain two-dimensional coordinates of joints, IEF employs a novel approach of multi-stage stepwise regression. IEF is divided into four phases, and the objective of each phase is regressed through three iterations. The results with one-third error are fed back to the neural network. Compared to multi-stage direct regression, this stepwise approach yields more accurate predictions for local joint coordinates. The detailed workflow of IEF is presented in Figure 8.

In the regression-based work described above, the majority of methods primarily focus on the minimization of the positional errors of individual joint points, disregarding the internal composition of the pose. As a result, the potential interdependence among joints remains underutilized. In general, skeletal information is more primitive, more stable and easier to learn than joint points. Sun et al. proposed a structure-aware regression method named Compositional Pose Regression (CPR), which reparametrizes the skeletal pose to replace the joint points and defines a combined loss function to encode the interdependence among different bones through the joint connection structure, which is not only applicable to human 2D pose estimation but also can be extended to human 3D pose estimation [62]. However, whether using direct joint regression like IEF or structural

regression like CPR, the entire family of coordinate-based networks ultimately faced an insurmountable theoretical ceiling. Although early coordinate-based methods established the baseline for deep learning in pose estimation, directly regressing 2D numerical coordinates from high-dimensional image pixels forces the network to learn highly non-linear mappings. This process typically relies on fully connected layers, which inherently discard the spatial topologies and translation invariance of convolutional feature maps. The subsequent domain-wide shift toward heatmap-based representations was not merely an empirical choice, but a structural necessity. By casting pose estimation as a dense spatial probability distribution, heatmaps allow fully convolutional networks to preserve local context and spatial hierarchies from end to end. However, this transition introduced a new theoretical bottleneck: the quantization error caused by the downsampling strides of CNNs, which physically bounds the sub-pixel accuracy of keypoint localization. It is precisely this trade-off—sacrificing absolute coordinate precision for spatial robustness—that catalyzed the development of multi-scale feature fusion and high-resolution architectures in the following years [63].

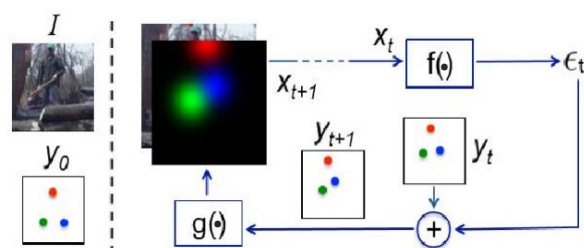


Figure 8. IEF structure diagram (Note: In this figure, the dots represent the detected skeletal keypoints, and the arrows indicate [the direction of data flow/relative displacement vectors]. The different colors are used solely to visually distinguish different instances and body parts. All variables correspond to their definitions in the main text).

3.1.2. Heatmap-Based Methods

The heatmap-based detection method was developed from the body part detection method. For each location, the method based on heatmaps generates a score that represents the level of confidence regarding its association with a critical point. By examining the heatmap formed by the above steps, the probability distribution of keypoints and the location information of keypoints are obtained. The spatial location information is better preserved in the heatmap-based detection method, which is more in line with the design characteristics of the Convolutional Neural Network (CNN). Therefore, it achieves better prediction accuracy (as shown by the flowchart in Figure 9). The classification in this subsection is presented as shown in Figure 10.

As illustrated in Figure 9, instead of directly regressing the numerical coordinates, the model predicts a spatial probability distribution. Specifically, for each ground truth joint located at coordinates (x, y) , a 2D Gaussian heatmap G is generated. The response value at pixel (u, v) in the heatmap is calculated as $G(u, v) = \exp\left(-\frac{(u-x)^2 + (v-y)^2}{2\sigma^2}\right)$, where σ represents the standard deviation that controls the spatial spread of the heatmap response. This methodology, which transforms discrete coordinate points into continuous spatial probability distributions, was pioneered by Tompson et al. [64].

- Adding Prior Information of Human Body Structures

In contrast to Coordinate Net’s direct regression algorithm that relies on two-dimensional coordinates, Heatmap Net incorporates both of the probability distribution-based ground truth constructions and incorporates structural information among different body parts. After generating the heatmap of each node, Heatmap Net can explicitly create

a network connection relationship, in the form of graph or tree structures, by combining the probabilistic graph model with the connectivity among human nodes and inputting it into the neural network. As a result, the neural network acquires a priori information about the human structure prior to its trainings. This is equivalent to artificially controlling the propagation direction of the feature information of each joint point in the network. This approach is able to enhance the sensitivity to such information and consequently improve the detection ability of the joint point through integrated learning. Heatmap Net is further divided into two methods, namely, explicitly adding a priori structures and implicitly learning information structures. In this paper, the two methods will be introduced individually, with a detailed explanation provided for each.

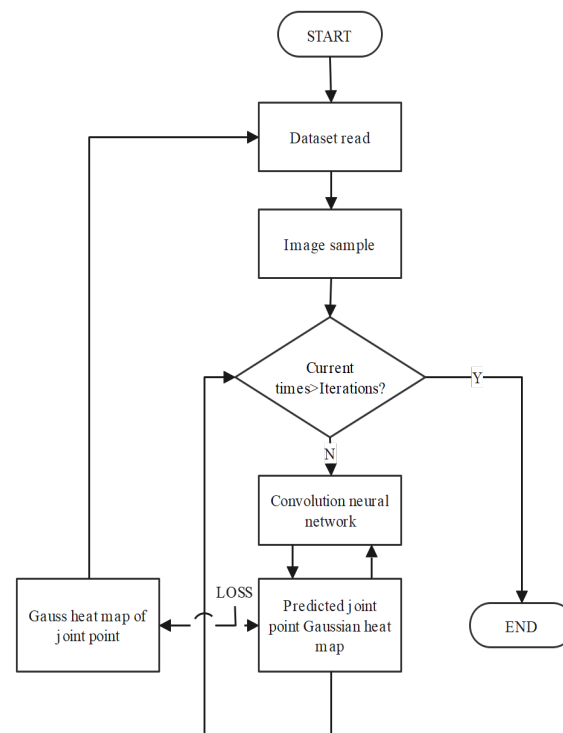


Figure 9. Flow chart of single-person attitude estimation based on heatmap monitoring (Note: The solid arrows indicate the direction of data flow and the sequential processing steps within the heatmap-based pose estimation pipeline).

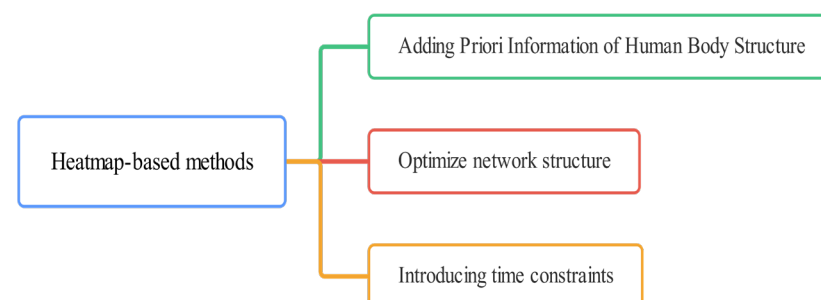


Figure 10. Classification of single-person estimation algorithm based on heatmap detection (Note: The solid lines indicate the hierarchical classification relationships among the heatmap-based methods. The different colors of the boxes represent distinct hierarchical levels—from the root category to specific algorithmic approaches—and are used solely for visual clarity).

Furthermore, as depicted in the architectural pipeline in Figure 11, a critical component of these models is the ‘Resolution Recovery’ module during the model testing phase. Deep convolutional networks intrinsically downsample input images to abstract

high-level semantic representations, which inevitably degrades the fine-grained spatial resolution required for precise joint localization. To mitigate this spatial information loss, the resolution recovery stage typically employs symmetric decoder structures—such as transposed convolutions or cascaded bilinear upsampling layers—to progressively project the low-resolution feature maps back to the original input scale. This spatial restoration is indispensable for refining heatmaps to the pixel level, thereby guaranteeing high-fidelity, sub-pixel accuracy in the final pose estimation.

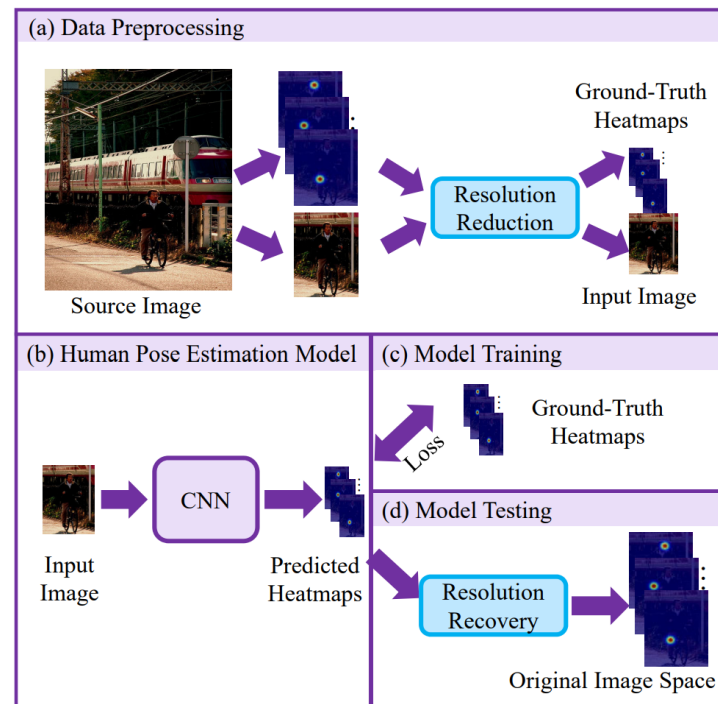


Figure 11. The flow of the human body pose estimation method based on Heatmap regression (Note: The arrows indicate the data flow, including downsampling, upsampling, and multi-scale feature fusion paths. The differently colored rectangular blocks represent feature maps at varying spatial resolutions).

Ning et al. modeled and jointly trained graph structures and neural networks for joint points [65]. In this model, CNN adopts a multi-scale image mechanism when extracting image features, so that the construction of heatmaps can combine global information and local details of skeletal joints, which ensures the accuracy of subsequent coordinate localization. Based on these heatmaps, the authors derived the probability distribution of joints and employed Markov Random Fields (MRFs) to model adjacent pairs of joints, filtering out inaccurate probability predictions from the original output. The network modifies the heatmap of each keypoint using a prior probability heatmap. This modification allows for the generation of a more precise distribution of skeletal joint points by obtaining the joint probability distribution of adjacent joint points. However, the construction of the Markov random field model requires the utilization of many component detectors, which makes the network structure complex, and the training scale increased. To address this, Szegedy et al. proposed an end-to-end deep neural network that uses a tree structure to model and jointly train human joints [66]. Converting the graph structure into a tree structure results in the network model compression and the decrease in the complexity of the model. Nevertheless, there are still a large number of parameters in the tree-structure-based Heatmap Net.

Given the limitations associated with explicitly adding structural priors, current research increasingly favors implicit learning paradigms to capture the structural context

of human poses. This difference in the way of obtaining structural information of nodes has also become a marker for distinguishing between explicit and implicit types.

Implicit learning methods effectively avoid the architectural complexities required for modeling explicit detector interdependencies. The main idea of this method is to increase the receptive field, i.e., the size of the pixel points on the output feature map of the network corresponding to the original input image area [67]. The larger the receptive field value, the larger the area of the pixel corresponding to the original input image. It implies that each pixel within the output feature map encompasses a more comprehensive and higher-level semantic description of features. By expanding the receptive field of the network, the model not only acquires richer semantic information regarding high-level anatomical structures but also facilitates the learning of connectivity features among spatially distant joints. Expanding the perceptual domain of the heatmap can be achieved by means of increasing the number of convolutional layers, enlarging the size of the pooling layer, and scaling up the size of the convolutional kernel. However, each of these methods has inherent design limitations. For example, increasing the number of convolutional layers usually leads to the issue of gradient disappearance. Meanwhile, increasing the size of the pooling layer usually leads to a sacrifice of processing accuracy. Furthermore, the increase in the convolutional kernel size results in growth in the network's parameter count and greater computational resource demands. Therefore, modern researchers mitigate the side effects of receptive field expansion by optimizing the network architecture itself, as seen in classic models like Convolutional Pose Machines (CPMs) and Stacked Hourglass Networks (SHNs).

- Optimize Network Structure

To recover spatial resolution lost during feature downsampling and to expand receptive fields, researchers have extensively optimized network architectures. Wei et al. [68] proposed Convolutional Pose Machines (CPMs), leveraging sequential convolutional layers to implicitly learn spatial dependencies between parts. Newell et al. [69] introduced the seminal Stacked Hourglass Networks (SHNs), which utilized symmetric encoder–decoder structures and multi-scale feature fusion to capture specific positional characteristics at various scales.

Subsequent innovations built heavily upon these foundations. Chu et al. [70] integrated global and limb-part attention mechanisms via Hourglass Residual Units (HRUs). Yang et al. [71] designed Pyramidal Residual Modules to enhance scale invariance, while Ke et al. [72] proposed the Multi-Scale Structure-Aware (MSSA) network specifically designed for complex scenes. Further structural optimizations included hierarchical body part representations by Tang et al. [73], lightweight designs via Fast Pose Distillation (FPD) [74], and the introduction of Generative Adversarial Networks (GANs) by Chou et al. [75] and Chen et al. [76] to help explicitly infer occluded joints. Cai, Y et al. [77] further refined this trajectory by merging hourglass stacks with U-Net structures to minimize identity connections and maximize performance. Chen et al. proposed a structure-aware convolutional network, which contains a pose generator, a pose discriminator, and a confidence discriminator to integrate priori information about the human structure [78]. The pose generator is designed by an hourglass network to predict the joint heatmap and the occlusion heatmap. The pose discriminator is used to determine the plausibility of body morphology, and the confidence discriminator displays the predicted confidence score. Iqbal, U et al. introduced a hybrid network that combines an hourglass stacking network and a U-Net structure to minimize the number of identity connections in the network with the same parameter budget and improve the performance [79]. As illustrated in Figure 12, each hourglass stack in this hybrid network consists of an encoder that progressively downsamples the feature map and a decoder that restores its spatial resolution.

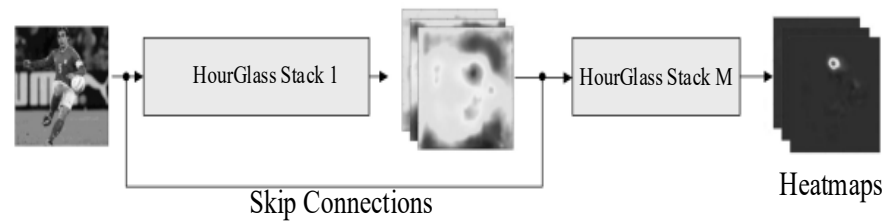


Figure 12. The hybrid network of Bulat et al. (Note: The arrows indicate the direction of data flow, including skip connections and the sequential processing through the downsampling and upsampling stages. The rectangular blocks represent feature maps at varying spatial resolutions, and the different colors are used solely to visually distinguish distinct network modules and intermediate supervision stages).

- **Introducing Time Constraints**

For monocular video sequences, integrating temporal dynamics significantly enhances spatial predictions. Huang, S et al. [80] and Osokin et al. [81] pioneered the use of optical flow as a supervisory signal to align heatmaps across adjacent frames. Later, recurrent architectures utilizing LSTM units [82] and time-delayed Generative Pose Estimators (GPEs) [83] were introduced to enforce temporal geometric consistency. Artacho et al. [84] further combined multi-scale spatial pooling with temporal fields of view to accurately predict keypoints in dynamic sequences.

Although these extensive architectural optimizations pushed heatmap-based single-person estimation to near saturation, their heavy reliance on local convolutional operations limited their ability to capture long-range global dependencies in severely occluded scenarios. This inherent limitation ultimately paved the way for the subsequent rise of global attention mechanisms in complex, multi-person environments.

Table 4 summarizes the comparisons on single-person pose estimation methods.

Table 4. Comparison of posture estimation methods of single person.

Category	Sub-Category	Advantages	Disadvantages	Scope of Application	Representatives
Coordinate Net	Multi-stage direct regression	Simple model structure with high time efficiency	Poor learning structure and information ability, low accuracy	Simple single pose full body	[46]
	Multi-stage stepwise regression	Improved the accuracy of the regression	Poor learning structure and information ability, more influenced by the initial posture	Simple single pose full body	[61,62]
Heatmap Net	Adding prior information of human body structure	Using the structural priorities to improve the accuracy rate, has clear joint relationships	Network structure is complex, the figure model structure is single fixed	Single pose half body full body	[49,50,64,74]
	Optimize network structure	The accuracy rate of the model prediction has been improved	Huge number of network participants, time efficiency needs to be improved	Single, multiple pose full body	[48,58,60,65–70,75,76]
	Introducing time constraints	It can be extended to video attitude estimation and can use video front and back frames to solve the occlusion problem	Large computation and susceptible to gradient disappearance	Single pose monocular video	[71–75]

3.2. Multi-Person Pose Estimation

When an image contains only a single target, 2D human skeletal keypoints can be estimated using the aforementioned single-person pose estimation methods. But in practical application scenarios, there is often more than one target in an image. In order to prevent allocating detected joint points to different targets, it is essential to employ the multi-person pose estimation approach. 2D multi-person pose estimation exhibits greater challenges compared with single-person pose estimation tasks. It is usually attributed to the presence of unknown individuals in the image, along with the need to detect and group them accurately. Multi-person pose estimation can be mainly classified into the two-stage approach [85,86] and the single-stage one [87,88], according to different algorithm steps.

3.2.1. Two-Stage Approach

The two-stage approach in multi-person pose estimation is categorized into the top-down method, the bottom-up method, and a hybrid of the first two methods based on the sequence of detection and group steps. The specific classification of multi-person pose estimation by the two-stage approach is shown in Figure 13. Table 5 presents a classification of the two-stage approach for multi-person pose estimation.

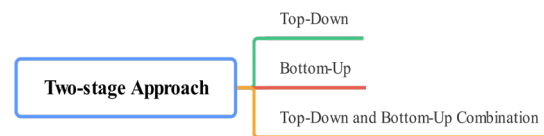


Figure 13. Classification of the two-stage approach for multi-person pose estimation (Note: The different colors used for the bounding boxes and estimated skeletal keypoints are solely for the visual distinction of different human instances and body parts).

Table 5. Comparison of two-stage multi-person pose estimation algorithms.

Category	Advantages	Disadvantages
Top-down	False detection and redundant detection can be reduced by improving the body detector, higher accuracy of node positioning	Needs to perform human detection for each person in the image, which takes up more memory, resulting in less time efficiency
Bottom-up	Less influenced by increasing number of people, time does not increase linearly with growing number of people	The influence of occlusion is large, and it is easy to lead to misconnection of nodes in complex scenes, and the accuracy of the algorithm is low
Top-down and Bottom-up Combination	Combines the advantages of both models with higher accuracy and reduced impact of increasing numbers	Detection performance still needs to be improved and remains vulnerable to changes in human scale

- **Top-down Method**

The approach of multi-person pose estimation is usually divided into two steps in the top-down method. Firstly, each target person in the image is detected and framed out through the target detection algorithm. Subsequently, a single-person pose estimation operation is performed for each extracted target person. The specific processing workflow of the top-down paradigm is demonstrated in Figure 14.

For the initial human detection phase, researchers typically rely on either high-precision two-stage detectors, such as the Region-based Convolutional Neural Network (R-CNN) family, or end-to-end one-stage methods like YOLO and Single Shot MultiBox Detector (SSD), which offer significantly faster inference speeds [89,90]. Despite their robust precision, two-stage detectors are not immune to errors; redundant region proposals and slight bounding box deviations inevitably propagate inaccuracies to the subsequent pose

estimation stage. Consequently, isolating the human subject accurately becomes the fundamental prerequisite of the top-down pipeline. Due to its superior localization capabilities, the R-CNN series—particularly Faster R-CNN—remains the predominant choice for the target detection module in many multi-person pose estimation frameworks [91].



Figure 14. Workflow of human pose estimation through top-down method (Note: The different colors used in the final parsing results are applied solely for the visual distinction of different human instances and body parts).

To systematically understand the evolution of top-down architectures, we categorize the existing literature into four distinct structural trajectories based on the specific bottlenecks they address.

The first trajectory focuses on bounding box optimization and non-maximal suppression (NMS) strategies. Severe overlap among individuals in crowded images often leads to redundant candidate boxes. Early milestone frameworks, such as Mask R-CNN [92], CPN [93], RMPE [94], and G-RMI [95], primarily addressed this by refining region proposal extractions to ensure that each target is detected only once. For instance, Mask R-CNN utilizes Faster R-CNN to extract Regions of Interest (ROIs), filtering redundancies via standard NMS, and subsequently binarizes spatial feature maps to represent joints via single-point masks. However, standard NMS rigidly discards boxes exceeding an Intersection over Union (IoU) threshold, which severely degrades recall rates when people overlap. To mitigate this missing detection issue, CPN adopts soft NMS [96], which gradually decays the confidence scores of overlapping boxes rather than abruptly discarding them. CPN further enhances precision by dividing its pose estimator into a global network for structural localization and an optimization network for fine-grained refinement. Similarly, RMPE tackles proposal redundancy by introducing p-Pose NMS alongside a Pose Distance (PD) metric [97], filtering out spatially similar pose predictions based on Euclidean distances. To handle misaligned cropping, RMPE strategically incorporates a Symmetric Spatial Transformer Network (SSTN) to align and normalize the human bounding boxes before feeding them into a spatial deconvolution network for heatmap generation.

The second trajectory addresses robustness in heavily crowded scenes and advanced node association. As benchmark scenarios grew more complex, simple bounding box filtering proved insufficient for extreme occlusions. To resolve pose estimation failures in dense crowds [98], Li et al. proposed CrowdPose. Rather than aggressively suppressing joints that seem unrelated to a specific detection box, this approach innovatively categorizes joints within the region as “target joints” and those belonging to adjacent people as “interference joints.” By applying distinct loss penalties to these categories, CrowdPose achieves robust multimodal predictions even when the initial human bounding boxes are imprecise. Additionally, dispensing with conventional NMS, the authors developed a Person-Joint Graph to structurally organize the final keypoints. By representing linkages through global maximum node association, this graph-based strategy offers significantly greater resilience against the anatomical conflicts inherent in crowded environments.

The third trajectory centers on CNN architectural efficiency and multi-scale feature fusion. Top-down pipelines inherently suffer from heavy computational overhead, as the

pose estimator must run independently for every detected person. To alleviate this, Huang et al. introduced LKConvPose [99], a highly efficient CNN-based hybrid architecture. It replaces traditional stacked small convolutions with large kernel convolutions (e.g., 17×17) to dramatically expand the effective receptive field, while operating parallel small kernel branches (e.g., 3×3) to retain intricate local textures. Utilizing the MLP-Mixer and the SKNet attention mechanism, LKConvPose adaptively fuses features across scales—leveraging large-scale blocks to capture global kinematic relationships and small-scale blocks to extract connecting edge details. This hybrid approach secures high keypoint accuracy while keeping computational costs remarkably low.

The fourth and most recent trajectory reflects a profound paradigm shift toward Vision Foundation Models (VFMs) and advanced feature alignment. Despite the impressive accuracy of classical CNN methods, their linear scaling of computational complexity relative to the crowd size remains a rigid bottleneck. Driven by the necessity to overcome this limitation and improve out-of-distribution (OOD) generalization, current research decisively favors massive data scaling. For example, ViTPose++ [100] demonstrates that a plain Vision Transformer (ViT), equipped merely with masked image pre-training, can achieve unprecedented robustness across highly diverse and uncommon body poses. To counterbalance the deployment inefficiencies typically associated with such massive architectures, researchers have developed two-stage knowledge distillation frameworks like DWPose [101], which effectively compress comprehensive pose knowledge from heavy VFMs into agile, lightweight student networks. Furthermore, Xu et al. [102] significantly advanced the top-down pipeline by proposing a highly efficient feature alignment mechanism. By seamlessly integrating dynamic attention modules, their approach explicitly corrects the cascading errors typically triggered by misaligned upstream bounding boxes, ultimately securing an optimal trade-off between rapid inference speed and high-fidelity pose accuracy in complex real-world settings.

In summary, the top-down paradigm excels at isolating subjects to predict highly accurate joint coordinates, provided that reliable human bounding boxes are extracted. Through meticulously designed filtering rules and alignment mechanisms, these networks effectively minimize both redundant detections and omissions. Nevertheless, its foundational architecture dictates that as the crowd density increases, the proportional surge in network forward passes inevitably remains the primary challenge to computational efficiency.

- Bottom-up Method

Unlike the top-down method, the bottom-to-up method first detects the skeletal articulation points of all people in the image and then assigns them to each person and splices them by means of clustering. Therefore, how to cluster the skeletal articulation points is the key for the bottom-up method. Typical techniques for clustering skeletal joint points encompass Semantic Part Segmentation [103], Part Affinity Fields [104], DeeperCut [105], etc. Furthermore, typical keypoint configurations and sample annotations for the MPII and MSCOCO datasets under unconstrained environments are displayed in Figure 15.

The keypoint detection approach in Semantic Part Segmentation resembles the Full Convolutional Network (FCN) method employed in Mask R-CNN, with additional utilization of a mask-based instance segmentation algorithm for clustering skeletal joints. The semantic part segmentation algorithm divides the human body mask into six detailed sub-masks and uses them as the truth labels. After acquiring the node's location, the network model undergoes the process of training with the ground truth labels to generate the semantic part distribution map. Finally, by constructing a conditional random field to combine skeletal joint locations with limb information, the network model is capable of inferring the human body region to which the skeletal joints belong and guiding the network to more accurately associate the joints belonging to the same person during the

clustering process. Part Affinity Fields (PAFs) provide a richer semantic representation for ground truth annotations. Rather than simply marking keypoint locations, PAFs encode the 2D vector field of each pixel, capturing both the position and orientation of limbs connecting two associated joints [103]. Through single-person pose estimation, the PAFs network acquires all skeletal joints in the image. Once this step is completed, the PAFs network proceeds to acquire the heatmaps for all skeletal joints in the image and derive the part affinity field for each limb with the assistance of truth labels. Subsequently, the joints obtained are clustered. The clustering of skeletal joint points based on PAFs involves three main steps. Firstly, the joint points are clustered into a bipartite graph, and the pixel directions are integrated to determine the weights of the edges. Secondly, the weights are maximized to establish connections between the joint points. Finally, the torso connections are established to complete the process of clustering.



Figure 15. Workflow chat of bottom–up method (Note: The dots represent the annotated human body keypoints. The different colors are used solely to visually distinguish different body parts and individuals, rather than representing any specific quantitative values).

DeeperCut employs a clustering approach which is similar to the graph optimization method. The ground truth labeling utilized during the process of model training is a ternary binary variable that indicates the corresponding relationship among skeletal joint points. Specifically, these variables encode the body part class of the joint, the identity of the individual, and whether two distinct joints share the same identity. In the initial stage, DeeperCut uses a CNN to extract candidate joints and build a dense graph. Subsequently, it calculates the inter-node correlation to determine the association between different skeletal joint points to determine whether they come from the same person. Finally, techniques such as NMS are used to eliminate redundant edges, thus completing the whole process of pose estimation. The computational complexity of DeeperCut’s dense graph optimization method is significantly higher than that of the PAF bipartite matching approach. However, this complexity is mitigated by compressing the number of node sets.

Cheng et al. made an improvement on the High-Resolution Network (HRNet) called HigherHRNet by means of adding an efficient deconvolution module to the backbone of HRNet [104]. They adopted a multi-resolution training and heatmap aggregation strategy, which allows for predicting scale-aware heatmaps. It is proven to be able to greatly improve the efficiency and the accuracy of the keypoint prediction, which is superior to HRNet. Additionally, they utilize an associated embedding label algorithm to group the keypoints.

Brasó et al. employed a novel attention mechanism to group the keypoints in a manner distinct from prior clustering methods [99]. Moreover, by leveraging the transformer technology, the authors encoded the interconnections between keypoints and centroids at a specific time point. This encoding yields contextually enriched embeddings.

In summary, the bottom–up method for multi-person pose estimation involves feature map extraction only once and utilizes clustering to assign each joint to a different person. The advantage of this estimation method is that the detection speed is faster, but the accuracy is lower. The algorithm faces challenges with clustering, especially when dealing with highly similar body joints under different lighting conditions, backgrounds, and

occlusions. Therefore, the improvement of clustering and increasing detection accuracy will be important areas of future research for the bottom-up method.

- Top-down and Bottom-up Combination

To address the respective limitations of isolated top-down and bottom-up paradigms, researchers have explored integrated approaches. In recent years, hybrid architectures combining both top-down and bottom-up models for multi-person pose estimation have emerged [78].

One primary strategy focuses on incorporating explicit feedback loops. For example, Hu and Ramanan proposed a hierarchical Gaussian model to incorporate top-down feedback into bottom-up Convolutional Neural Networks [103]. It is based on multi-scale predictions, which introduces state-of-the-art bottom-up baselines and continuously improves these methods through top-down feedback, particularly during occluded periods when bottom-up evidence may be ambiguous. This approach is highly effective in handling occlusion and other similar scenarios.

Another strategy involves sequential inference and temporal integration. Tang et al. developed a framework with a bottom-up inference followed by a top-down refinement based on a compositional model of the human body [67]. Furthermore, Li et al. used LSTM and combined bottom-up heatmaps with human detection to solve the occlusion and detection bias problems [105]. However, the approach remains essentially bottom-up, as a bottom-up network is utilized and only detection bounding boxes are added as the top-down information for joint grouping. Therefore, it is still susceptible to human scale variations.

To further overcome the cascading errors and scale vulnerabilities inherent in these early hybrid architectures, the latest trajectory has shifted toward dynamic feature alignment and end-to-end models. For instance, Xu et al. [100] introduced a highly efficient feature alignment framework that seamlessly bridges the gap between instance detection and spatial localization. By integrating dynamic attention mechanisms, this approach explicitly mitigates the misalignment errors typically found in cascaded hybrid pipelines, achieving an optimal balance between inference speed and pose accuracy in complex environments.

Table 5 presents the comparison of three algorithms for two-stage multi-person pose estimation.

3.2.2. One-Stage Approach

To simplify the two-stage approach for multi-person pose estimation and improve its efficiency, Papandreou et al. introduced the PersonLab model [91]. This model learns to detect individual keypoints and predict their relative displacements. Unlike prior methods, PersonLab employs a box-free bottom-up approach, which not only streamlines the estimation pipeline but also outperformed all existing bottom-up systems in keypoint localization accuracy. Different from PersonLab, Nie et al. proposed a Single-stage multi-person Pose Machine (SPM), as shown in Figure 16, which simplifies the process of human segmentation and keypoint localization [92]. SPM was developed based on SPR, which diminishes the resolution of the feature map to learn abstract semantic representations. They first introduced a new Structured Pose Representation (SPR) to encode the positions of body joints via the displacement of human joints relative to the root joint, which is utilized to determine the position of the human body. Subsequently, the upsampled feature map is used for refining high-level semantic information with low-level spatial information during the process of keypoint localization. Furthermore, an offset regression branch was added to extend the hourglass network to estimate the offset of human joints. It brings in the achievement of efficient single-stage multi-person pose estimation. However, its performance is still less accurate than advanced bottom-up methods. Therefore, Shi et al.

proposed a direct and simple single-stage multi-person pose estimation framework, namely, InsPose, by using an instance-aware dynamic network to adaptively adjust the network parameters for each instance to enhance the network's ability to recognize various poses and adaptability [106]. Tian et al. proposed a new framework, i.e., PoseDet, which can improve the speed of both locating and associating body joints [107]. They introduced a joint-aware pose embedding to represent human instances based on keypoint positions. Compared with SPM, the approach of combining joint detection and association in the same pipeline is proven to be a faster single-stage solution.

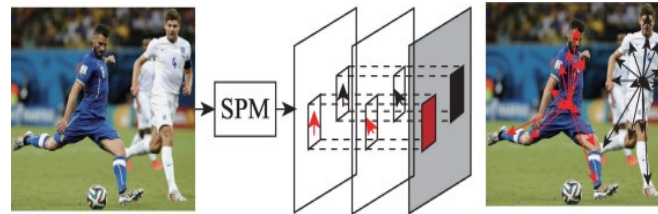


Figure 16. SPM single-phase attitude estimation scheme (Note: The input image is divided into spatial regions at different pyramid levels, and features are extracted from each region to form a hierarchical representation. The red blocks indicate selected local feature regions, while the final image shows the visualized matching or feature correspondence results. The colors and markings are used solely for visual illustration, rather than denoting specific quantitative values).

3.2.3. End-to-End Framework

Yang et al. proposed a novel end-to-end framework with Explicit box Detection for multi-person pose estimation (ED-Pose), which unifies contextual learning between human-level (global) and keypoint-level (local) information [105]. Unlike previous single-stage approaches, ED-Pose reconceptualizes this task as two explicit box detection processes with unified representation and regression supervision. Firstly, a human detection decoder is introduced from the encoded tokens to extract global features. It can provide a good initialization for the later keypoint detection and make the training process converge quickly. Secondly, in order to introduce contextual information near the keypoint, pose estimation is considered as a keypoint box detection problem to learn the box position and content of each keypoint. The person-to-keypoint detection decoder employs an interactive learning strategy between human and keypoint features to further enhance global and local feature aggregation. Overall, the ED-Pose concept is simple and does not require post-processing and intensive heatmap supervision. It exhibits higher effectiveness and efficiency compared with both two-stage and single-stage approaches.

Liu et al. simplified ED-Pose by simply treating the K keypoint pose estimation as predicting a set of $N \cdot K$ keypoint positions, each from a keypoint query, as well as representing each pose with an instance query to score the N pose predictions [108]. They introduced a simple modification to the decoder's self-attention mechanism. They replaced the single self-attention for all $N \cdot (K + 1)$ queries with two subsequent grouped self-attentions. The resulting decoder eliminates cross-instance interactions among different queries, thereby simplifying optimization and improving performance. However, despite their efficiency, existing single-stage deterministic regression methods and end-to-end keypoint regression models are often prone to missed or false detections in crowded or occluded scenes due to their inability to reason about pose ambiguity. To address these challenges, Tan et al. converted a single-stage, end-to-end keypoint regression model into a diffusion-based sampling process [106]. A generative approach was used to process fuzzy poses, i.e., sampling from image conditional pose distributions characterized by a diffusion probability model. Specifically, they extracted initial pose markers from the images and progressively refined noisy candidate poses by means of interacting with the initial markers through an

attention layer. The DiffusionRegPose proposed significantly improves pose accuracy in crowded scenes.

3.2.4. Transformer-Based

Most of the existing CNN-based methods perform well in visual rendering, but the CNN model is usually regarded as “black boxes”, and it is difficult to explain how the relationships between keypoints are captured within the model, so they lack the ability to explicitly learn the constraints between keypoints, which restricts their flexibility in processing different input images. Yang et al. proposed TransPose, a Transformer-based model for human pose estimation that is lighter and faster than mainstream CNN architecture [107], which employs CNN to extract low-level features, then captures long-distance dependencies between keypoints through Transformer’s self-attention mechanism, and interprets the model’s predictions through attention scores. Moreover, the revealed dependencies are image-specific and fine-grained, which can also demonstrate how the model handles special cases such as occlusion.

However, TransPose is not completely free from the constraints of CNN architectures. Thus, Yang et al. proposes a human pose estimation method called TokenPose, which is based on the transformer structure and learns the constraints between visual cues and keypoints simultaneously by representing keypoints as “tokens” [109]. TokenPose divides the image into patches, each of which is converted into a visual token. Each keypoint is represented by a learnable token, which interacts with the visual token through a self-attention mechanism while learning appearance cues and constraint relationships. Next, a multi-layer Transformer encoder is used to capture the constraints between the keypoints through the self-attention mechanism. Finally, the keypoint tokens are projected into heatmaps for keypoint localization using an MLP head. To further optimize the performance of TokenPose, Chen et al. innovatively proposed an SDPose model by introducing the MCT module and the self-distillation method into TokenPose, which significantly improves the performance while keeping the small model lightweight [110]. SDPose enhances feature representation during training with the Multi-Cycle Transformer (MCT) module and compresses knowledge from multiple cycles into a single forward propagation through self-distillation. During inference, the model outputs results with only a single forward propagation, without relying on external human detectors, and predicts keypoint locations in an end-to-end manner.

3.3. Comparison of Test Results of Classical Algorithms on Mainstream Datasets

3.3.1. Single-Person Pose Estimation Algorithm

Table 6 exhibits the respective experimental results of the classical single-person pose estimation algorithm on datasets MPII and LSP. The evaluation metrics are PCKh@0.5 with a tolerance of 0.5 for dataset MPII, and PCK@0.2 with a tolerance of 0.2 for dataset LSP.

Table 6. Test result of single-human pose estimation strategies on LSP and MPII (Red: best; Blue: second best; Orange: third best).

Method	Dataset	
	MPII (PCKh@0.5)	LSP (PCK@0.2)
DeepPose [46]	-	61.0
CNN + MRF [33]	82.0	-
CPM [65]	88.5	87.9

Table 6. Cont.

Method	Dataset	
	MPII (PCKh@0.5)	LSP (PCK@0.2)
SHN [60]	90.9	-
HRU [71]	91.5	92.6
PRM [58]	92.0	93.9
Jain, A et al. [68]	91.8	94.0
Adversarial Posenet [67]	92.1	93.1
MSSA [73]	92.1	-
DLCM [67]	92.3	95.1
Li, S et al. [49]	-	93.9
Jain, A et al. [69]	91.9	-
Bulat et al. [68]	94.1	94.8

3.3.2. Multi-Person Pose Estimation Algorithm

The experimental results of the multi-person pose estimation algorithm for AP on dataset MSCOCO and the prediction times are listed as shown in Table 7.

Table 7. Experimental results of multi-person pose estimation algorithms on MSCOCO (Red: best; Blue: second best; Orange: third best).

Method	AP	AP ⁵⁰	AP ⁷⁵	AP ^M	AP ^L	Time(s)
Mask R-CNN [92]	62.7	87.0	68.4	57.4	71.1	0.2
RMPE [94]	61.8	83.7	69.8	58.6	67.6	0.4
G-RMI [95]	64.9	85.5	71.3	62.3	70.0	-
CPN [93]	73.0	91.7	80.9	69.5	78.1	-
LKConvPose [99]	75	92.6	82.7	72.0	79.6	-
OpenPose [97]	61.8	84.9	67.5	57.1	68.2	0.6
Associative Embedding [83]	65.5	86.8	72.3	60.6	72.6	0.25
PersonLab [91]	68.7	89.0	75.4	64.1	75.5	0.464
PifPaf [2]	66.7	-	-	62.4	72.9	0.24
HigherHRnet [104]	70.5	89.3	75.4	64.1	75.5	-
CenterGroup [102]	71.4	90.4	78.1	67.2	77.5	-
ED-Pose [105]	71.6	89.6	78.1	65.9	79.8	-
Group Pose [109]	72.0	89.4	79.1	66.8	79.7	-
DiffusionRegPose [110]	72.5	89.8	79.5	66.8	80.5	-
TransPose [111]	74.2	89.6	80.8	70.6	81	-
TokenPose [112]	73.2	89.5	80.2	70.1	79.8	-
SDPose [113]	73.7	89.6	80.4	70.3	80.5	-

As illustrated in Figure 17, the past decade has seen major milestones in deep learning-based 2D human pose estimation, laying the groundwork for addressing current challenges and exploring future directions.

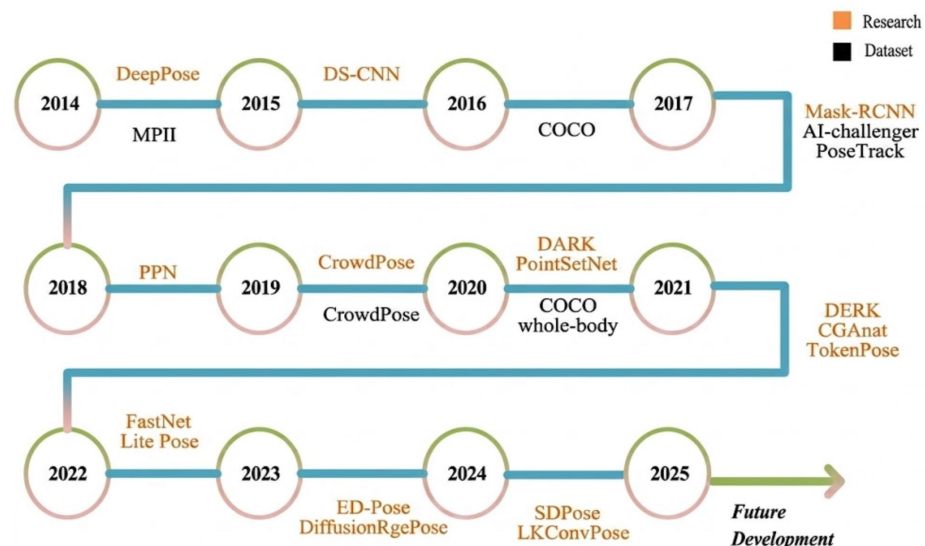


Figure 17. Major milestones in the last decade of deep learning-based 2D human pose estimation due to its unique innovation and outstanding performance (DeepPose [46], DS-CNN [114], PPN [115], DARK [116], PointSetNet [117], COCO whole-body [115], DERK [118], CGAnet [119], TokenPose [112], FastNet [120], Lite Pose [121], ED-Pose [105], and DiffusionRgePose [110]).

4. Current Challenges and Future Directions

4.1. Challenges and Solutions

Numerous obstacles remain to be tackled in the field of 2D human pose estimation, as indicated by the summary in Table 8, which outlines the challenges and emerging trends in its development. Generally speaking, there are three main challenges faced by 2D human pose estimation based on Computer Vision.

4.1.1. Complex Environmental Factors

Interactions between humans and other entities are ubiquitous in real-life scenarios, but they often result in occlusion, which can be classified into three categories based on the occlusion source, i.e., self-occlusion, background occlusion, and object occlusion. Due to the flexibility and variability of human movements, self-occlusion occurs frequently (as shown in Figure 18). In addition, background occlusion and object occlusion are also difficult to avoid due to cluttered backgrounds and crowded environments (as shown in Figure 19).

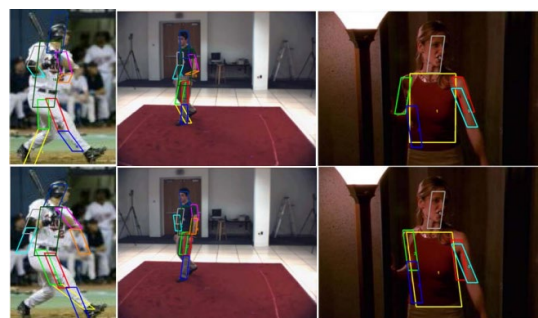


Figure 18. Human self-occlusion (**bottom**) and its correction (**top**) [68,74] (Note: The different colors used for bounding boxes and skeleton keypoints are solely for visually distinguishing different individuals and body parts, and do not represent specific quantitative values).

In addition, unreasonable shooting angles, overexposure or underexposure, and cast shadows can also easily lead to occlusion. In the case of shooting scenes and characters without any constraints, several phenomena mentioned above inevitably occur. Occlusion can

easily result in false positives and redundant bounding boxes, while keypoint grouping is often compromised by overlapping joints; both of these issues severely degrade subsequent pose estimation. Most of the existing pose estimation algorithms require the human body to be unobstructed or slightly self-obstructed in order to capture itself well. To address this, datasets such as OCHuman provide images with explicit occlusion annotations [121]. In terms of algorithms' optimization, some methods improve the accuracy and the robustness of detection by means of learning explicit or implicit spatial models, but they produce overfitting phenomena [65,122]. There are also some methods that combine human priors and a data-driven approach to solve such problems. However, these methods do not take into account the fact that a priori knowledge of the human body may not be sufficiently comprehensive or too simplified to cover all the possible postures and movements, resulting in an inability to accurately predict postures in some cases. Consequently, the results still lack robustness [123–126].



Figure 19. Crowding usually leads to frequent object occlusion and self-occlusion (Note: The different colors of the bounding boxes and skeletal keypoints are solely for the visual distinction of different instances and body parts).

Table 8. Taxonomy of key methodologies in 2D human pose estimation.

Single/Multi	Category	Sub-Category	Methods	Key Strategies	Shortcomings
Single-person pose estimation	Coordinate Net		DeepPose [46]	Convolutional network with multiple iterations of direct coordinate regression	Complex Environmental Factors
			IEF [61]	Multi-stage stepwise regression	Imbalance in the Number of Human Postures
			CPR [62]	Exploiting inter-joint dependencies	Lacks Robustness
	Adding prior information of human body structure		Li, S et al. [49]	Graph neural network	Lacks Robustness
			Szegedy et al. [64]	Tree-structured models	Lacks Robustness
			CPM [65], SHN [60]	Cascade Module Feature Fusion	Ignoring Timeliness Requirements
	Heatmap Net	Optimize network structure	Yang et al. [58]	Design Pyramid Residuals Module	Ignoring Timeliness Requirements
			MSSA [73]	Multi-scale structural perceptual neural networks	Ignoring Timeliness Requirements
			DLCM [67]	Design of the body part hierarchy representation for intermediate supervision	Ignoring Timeliness Requirements
			Jain, A et al. [68], Adversarial PoseNet [67]	Introducing Generative Adversarial Networks	Imbalance in the Number of Human Postures
			Chen et al. [69]	Proposed structure-aware convolutional networks	Ignoring Timeliness Requirements
			Bulat et al. [68]	Combining Stacked Hourglass Network and U-Net Network	Ignoring Timeliness Requirements
			FPD [70], Simple Baselines [55]	Lightweight network structure	Ignoring Timeliness Requirements
		Introducing time constraints	Jain et al. [69]	Use RGB images and motion features as input	Complex Environmental Factors
			Pfister et al. [72]	Use of optical flow diagrams as supervisory information	Complex Environmental Factors
			Luo et al. [70], UniPose [75]	Adjacent frame processing using LSTM network	Complex Environmental Factors
			GPE [74]	Time Delay	Complex Environmental Factors

Table 8. Cont.

Single/Multi	Category	Sub-Category	Methods	Key Strategies	Shortcomings
Multi-person pose estimation	Two-stage	Top-down	Mask R-CNN [92]	Example of detection frame + mask	Imbalance in the Number of Human Postures
			CPN [93]	Feature Pyramid Network for Feature Fusion	Ignoring Timeliness Requirements
			RMPE [94]	Spatial transformation sample correction	Ignoring Timeliness Requirements
			HRNet [58]	High-resolution maintenance	Ignoring Timeliness Requirements
			G-RMI [95]	Offset-assisted positioning	Ignoring Timeliness Requirements
			CrowdedPose [95]	Introduction of interference joints to improve the accuracy of node positioning	Ignoring Timeliness Requirements
			LKConvPose [99]	Combines large kernel convolution and multi-scale feature fusion mechanisms	Ignoring Timeliness Requirements
		Bottom-up	DeepCut [126], DeeperCut [99]	Figure optimization strategy	Complex Environmental Factors
			OpenPose [97], PifPaf [2]	Complex field vector representation	Complex Environmental Factors
			Associative Embedding [83]	Associative Coding Clustering	Complex Environmental Factors
			HigherHRNet [104]	Uses multi-resolution training and heatmap aggregation	Complex Environmental Factors
			CenterGroup [102]	Introduction of attention mechanism for keypoint grouping	Complex Environmental Factors
		Top-down and Bottom-up Combination	Hu and Ramanan [103]	Stratified Corrected Gaussian Model	Lacks Robustness
			Li et al. [104]	Reuse prediction results from previous frames	Lacks Robustness
			Tang et al. [67]	A novel network with a hierarchical compositional architecture	Lacks Robustness
	One-Stage		SPM [85]	Graded structured gestalt representation	Complex Environmental Factors
			InsPose [108]	Adaptively adjusts the network parameters of each instance using instance-aware dynamic networks	Complex Environmental Factors
			PoseDet [106]	Propose node-aware pose embedding to represent objects based on the location of keypoints	Complex Environmental Factors
	End-to-End Framework		ED-Pose [105]	Reconceptualization as two explicit box detection processes with unified representation and regression supervision	Ignoring Timeliness Requirements
			Group Pose [109]	A simple modification of the decoder's self-concern eliminates interaction between different queries across instance types	Ignoring Timeliness Requirements
			DiffusionRgePose [110]	Converting single-stage, end-to-end keypoint regression models to a diffusion-based sampling process	Ignoring Timeliness Requirements
	Transformer structure		TransPose [111]	A human pose estimation model by combining Transformer and CNNs	Ignoring Timeliness Requirements
			TokenPose [112]	Constraints between visual cues and keypoints are learnt simultaneously by representing keypoints as 'markers' based on a transformer structure	Ignoring Timeliness Requirements
			SDPose [113]	Introduction of the MCT module and self-distillation methods	Ignoring Timeliness Requirements

4.1.2. Timeliness Requirements

In practical applications of human pose estimation, achieving both high accuracy and real-time performance remains highly challenging. Compared with typical classification and detection tasks, human pose detection requires output feature maps with higher resolution. Therefore, high-precision pose estimation algorithms are attracting more and more attention from academia and industry. In recent years, there has been a consistent improvement in the detection accuracy of deep learning-based human pose estimation algorithms. However, it is achieved at the cost of complex network structures and a significant amount of time consumption. As illustrated in Figure 20, the inference latency of certain multi-person pose estimation algorithms surges noticeably as the number of individuals in the scene increases. It is unfriendly to time-sensitive applications, such as human–computer interaction, autonomous driving, etc. As the number of individuals in the image or video rises, there is a noticeable drop in the mAP for certain multi-person pose estimation algorithms (as depicted in Figure 20). To improve model accuracy, researchers frequently design overly complex networks, inevitably sacrificing real-time performance [127–129]. Although there have been a lot of studies exploring network compression and network

acceleration, none have specifically targeted human pose detection. Therefore, how to ensure the improvement of timeliness with high detection accuracy in the field of human pose estimation is still a problem. Table 9 shows the difficulty and development trend of 2D human pose estimation.

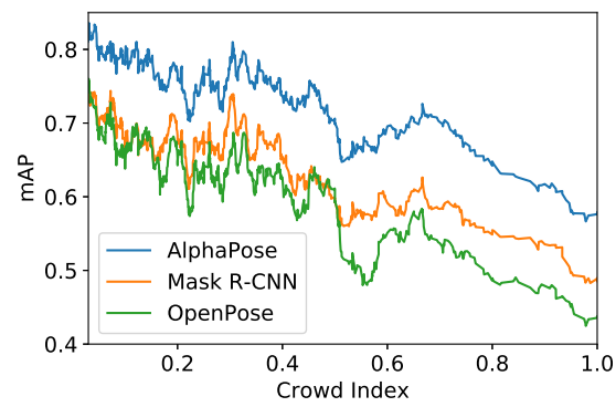


Figure 20. As the crowding index increases, the prediction effect of the multi-person pose estimation algorithm decreases significantly [95].

Table 9. Difficulty and development trend of 2D human pose estimation.

Challenges	Specific Difficulties	Technical Limitations	Research Trends and Solutions
Complex environmental factors	Shading, light changes, cast shadows	Limited ability of network to extract key features, insufficient ability to filter environmental noise	Occlusion repair, precise human pose target detection frame, combined human a priori information and data-driven
Timeliness requirements	High timeliness requirements, complex network structure, numerous network parameters	High-resolution output feature maps lead to complex network structures and higher time costs	Simplify network structure, guarantee accuracy while using lightweight network model
Imbalance in the number of human postures	Lack of complex postures such as falling and overturning	Limited to simple postures such as upright, poor robustness to complex postures	Focus on the collection of complex human postures, study of multi-frame continuous postures

4.1.3. Imbalance in the Number of Human Postures

The progress of Artificial Intelligence (AI) is closely intertwined with the parallel advancement of datasets and algorithms. Data serves as the basis of algorithm performance, and a high-quality dataset is able to make the model's effect greatly improved. However, the following problems still exist in terms of datasets.

Although multiple human pose estimation datasets (e.g., COCO, MPII, and FLIC) are available for various application scenarios, they primarily consist of images featuring simple poses, such as standing or sitting (as illustrated in Figure 21). However, these datasets lack more complex postures. Complex postures usually refer to poses in which the relative positions and angles between the joints of the human body show large changes, such as blocked joints, overlapping body parts, and non-linear forms of body posture, such as bending and twisting [36].

At the same time, these complex postures often have constraints such as self-occlusion. In addition, the imbalance in the number of human poses and the occurrence of occlusion in the datasets lead to a decrease in the accuracy of models when predicting these postures. However, there is still a lack of datasets that use imbalanced data to accurately detect complex postures. Therefore, the expansion and enrichment of existing datasets, as well as the exploration of more effective training methods, continue to be the focus of research in the field of 2D human pose estimation.



Figure 21. The existing datasets are insufficient for gestures [46,54] (Note: the colored skeletal keypoints are generated automatically by the algorithm solely for the visual distinction of different body parts and individuals. They do not represent specific categories or quantitative values).

4.2. Future Directions

Addressing the intertwined challenges of severe occlusion, computational latency, and data imbalance requires more than incremental algorithmic tweaks. Moving forward, the developmental trajectory of 2D human pose estimation can be evaluated along two distinct dimensions: resolving near-term deployment bottlenecks and pioneering long-term structural paradigms.

In the near term, research efforts are primarily driven by the friction of real-world deployment. This necessitates the development of lightweight architectures optimized for edge computing and explicit spatial–temporal mechanisms to repair transient occlusions. Conversely, breaking the theoretical ceiling in the long term demands a fundamental departure from heuristic pipelines and purely supervised 2D annotations. The field’s ultimate progression relies on unifying multi-task learning frameworks, leveraging zero-shot generative data augmentation, and incorporating cross-modal 3D priors to resolve the inherent depth ambiguities of 2D planar images. Details regarding future development are illustrated in Figure 22.

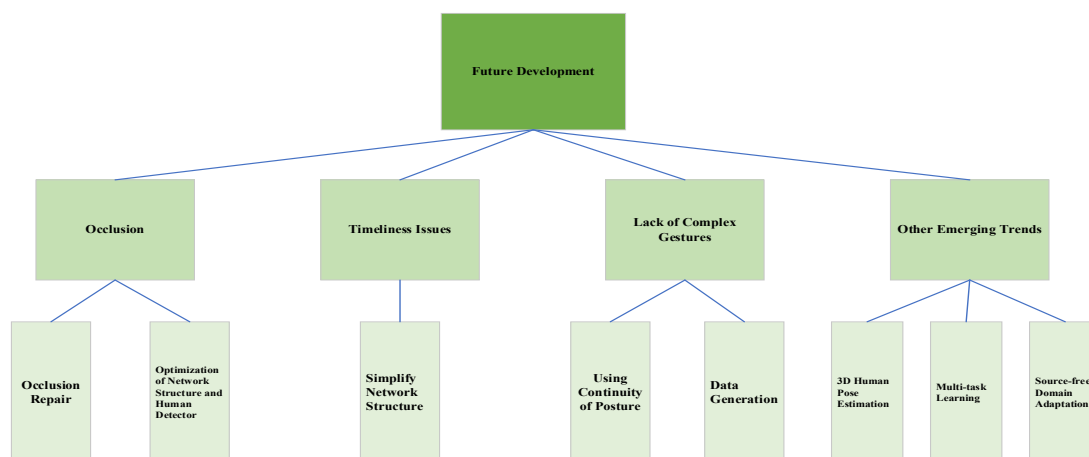


Figure 22. Future development.

4.2.1. Occlusion

• Occlusion Repair

For 2D human pose evaluation in complex scenes with occlusion, restoration of the masked part can be attempted, thus converting a complex scene into a simple unobscured scene. Therefore, the combination with the existing posture estimation methods contributes

to generating better results. For example, Ke et al. proposed a unified multi-task framework for joint video object mask completion and object appearance recovery [130]. Zhou et al. improved an algorithm for restoring anomalous joints, which recovers occluded keypoints based on human scale and motion characteristics [131].

- Optimization of Network Structure and Human Detector

To tackle occlusion problems in single-person pose estimation, one possible strategy is to enhance the network structure by expanding the receptive field of convolutional neural networks. For example, Chu et al. designed a novel Hourglass Residual Unit (HRU) to increase the network's receptive field, achieving promising results in handling occlusion [65]. In multi-person pose estimation, in order to enhance the accuracy of human pose estimation, the focus of the top-down algorithm can be put on the human detector, namely, how to minimize the error and redundancy of human detection frames. The bottom-up algorithm can improve on its most difficult step, the keypoint association problem. For example, Kocabas et al. proposed a novel keypoint assignment network, the Pose Residual Network (PRN), which can effectively reduce the impact on reducing the estimation results due to occlusion as well as the dense environment [127].

4.2.2. Timeliness Issues

- Simplify Network Structure

In order to obtain high-resolution output feature maps, the network models constructed are also more complex, and time consumption rises with the increase in accuracy. Therefore, exploring a lightweight network while ensuring high accuracy is an inevitable trend in the field of human-computer interaction. Human pose estimation algorithms generally comprise three stages: feature extraction, keypoint detection, and keypoint association. To achieve lightweight models, researchers can focus on the keypoint detection stage, utilizing techniques such as bounding box refinement, hierarchical keypoint grouping, and feature fusion to balance model compression and accuracy. For example, Lightweight OpenPose improves the detection speed by means of simplifying the node detection phase at the expense of accuracy. Fast Pose Distillation (FPD) simplifies this process through knowledge distillation of the Hourglass model, effectively halving the model size and significantly reducing time complexity. SPM is able to optimize the hierarchical grouping of keypoints [70,85,98].

4.2.3. Lack of Complex Gestures

- Using Continuity of Posture

Human motion, regarded as “flow”, is an important feature for human pose prediction models. By means of examining the pose variations in numerous consecutive frames before and after a complex pose, the analysis of the human motion continuity can significantly enhance the accuracy of prediction. Utilizing optical flow algorithms to capture temporal object information across consecutive images effectively improves human pose estimation. For example, Romero et al. proposed FlowCap, in which they used only optical flow to estimate 2D human poses from videos and achieved some success [132].

- Data Generation

Whether in 2D or 3D, data generation is a very important research direction. Taking the 2D dataset COCO as an example, although it contains more than 60,000 images, most of the poses captured in the dataset are just regular standing, walking, etc. In addition, the data collection process is costly and labor-intensive. Given these two points, the development of models for unusually complex poses (such as falling or climbing) faces serious challenges. To address this challenge, generating complex data through data augmentation methods

can effectively enhance the performance of algorithms. Despite the fact that the current quality of data generated by Generative Adversarial Networks (GANs) and similar methods is not high, it is adequate for human pose estimation. That is because highly detailed and realistic features are not required. Instead, diverse foregrounds (such as individuals with different clothing) and poses are necessary, which can be achieved via Multi-Agent Diverse GAN [133] (MADGAN), which consists of multiple generators and a discriminator, where the discriminator determines whether the input sample is a real sample or not; if it is a generated sample, MADGAN determines what the generator is. It helps to improve the diversity of generated samples through competition and cooperation between multiple generators, such as PacGAN [134] (multiple samples of the same category are ‘packaged’ together and fed into the discriminator, which ensures that the input samples are varied each time, thus improving the diversity of generated samples), or synthetic data, etc.

4.2.4. Other Emerging Trends

- 3D Human Pose Estimation

Many studies on 3D human pose estimation are built upon the foundation of 2D human pose estimation based on deep learning. Additionally, the technique of human keypoint localization is indispensable for reconstructing the 3D human form. Moreover, some of the current 3D human pose estimation continues to adopt the network architecture and research concepts derived from 2D human pose estimation. For example, Wang et al. proposed a self-supervised correction mechanism, which can effectively utilize a large amount of 2D data annotations to assist 3D pose prediction [135]. Despite the superior detection performance demonstrated by 3D human pose estimation over 2D human pose estimation, research on 3D human pose estimation is more challenging due to the limited corresponding training datasets, difficulties in data acquisition, and immature research techniques. Consequently, the existing research on 3D human pose estimation methods heavily relies on the well-developed techniques of 2D human pose estimation. During the process of training, deep convolutional neural networks are capable of learning network parameters by means of utilizing feedback information, typically depending on accurately annotated data as the training data to acquire more precise feedback information. As for 3D human pose estimation, the problem of data scarcity caused by the high cost of data collection and annotation leads to the use of weakly supervised or unsupervised learning methods by most researchers to train the network. Compared with strong supervision, the detection accuracy of these two learning methods is still relatively low. However, as research based on deep learning continues to evolve, weakly supervised learning methods can achieve comparable or even better performance than traditional strongly supervised learning in some cases [136]. Therefore, for 3D human pose estimation, how to improve and innovate the algorithms to fully utilize the existing lack of labeled data is also a serious problem that needs to be solved.

- Multi-task Learning

Joint training in 2D and 3D human pose estimations is able to improve the overall results. Additionally, it is possible to combine human pose estimation with other tasks related to the human body for data annotation and training. Examples include human segmentation and human parsing. The annotation of human segmentation involves the creation of polygons and sampling along the contours of these polygons, which can seamlessly integrate with these tasks. Joint training of multiple tasks related to the human body is meaningful for a comprehensive understanding of pedestrians and can also improve the accuracy of each individual task. However, there is a risk of increasing the data annotation costs. One alternative is to engage in cross-dataset joint training, wherein one dataset

solely consists of skeleton annotations, while the other solely consists of segmentation annotations. That is also a common issue in industry.

- Source-free Domain Adaptation

Domain adaptation methods for 2D human pose estimation typically require continuous access to source data during adaptation, which can be challenging due to privacy, memory, or computational limitations [137]. To address this limitation, future research could be focused on passive domain adaptation tasks for pose estimation, where the source model uses only unlabeled target data to be adapted to the new target domain. While recent advances have introduced passive methods for classification tasks, extending them to regression-based pose estimation is not easy.

To systematically summarize the architectural evolution discussed above, Table 10 compares the main families of 2D human pose estimation methods. By evaluating each paradigm across key criteria—such as computational complexity, occlusion robustness, and core limitations—this table serves as a direct reference for selecting appropriate algorithms based on real-world deployment constraints.

Table 10. Comparison of 2D human pose estimation paradigms.

Method Family	Representative Models	Core Strengths	Key Limitations	Occlusion Robustness	Computational Complexity	Suitable Scenarios
Coordinate-based (Single-person)	DeepPose, IEF, CPR	Intuitive; low memory footprint; fast inference.	Destroys spatial inductive bias; highly non-linear mapping.	Low	Low	Simple single-person tracking on edge devices.
Heatmap-based (Single-person)	CPM, Stacked Hourglass, HRNet	Preserves spatial context; high sub-pixel localization precision.	Bounded by quantization errors; heavier memory overhead.	Medium	Medium	High-precision single-person tasks.
Top-Down Two-Stage (Multi-person)	Mask R-CNN, RMPE, CPN	Scale-invariant instances via cropping; superior individual accuracy.	Highly dependent on upstream detector; latency scales with crowd.	Medium–High	$O(n)$ (Scales with person count)	High-accuracy multi-person analysis (non-real-time).
Bottom-Up Two-Stage (Multi-person)	OpenPose, HigherHRNet	Identity-agnostic extraction; fast multi-person inference.	Relies on fragile, non-differentiable heuristic grouping.	Low–Medium	Near $O(1)$ (Constant time)	Real-time tracking in moderately crowded scenes.
End-to-End and Transformer (Multi-person)	ED-Pose, TransPose, TokenPose	Fully differentiable global reasoning; eliminates heuristic glue.	High training cost; requires massive datasets.	High	High (Self-attention overhead)	Complex interactions with severe occlusions.

5. Conclusions

As a popular research topic in Computer Vision in recent years, 2D human pose estimation has been widely applied in fields such as motion recognition, intelligent surveillance, and human–computer interaction. In this survey, we conducted a comprehensive and systematic review on 2D human pose estimation based on deep learning, covering the development background, datasets and evaluation metrics, algorithm analysis, challenges, and future trends. Through detailed comparative analysis of the advantages and disadvantages of various models, it can be concluded that the field’s progression is fundamentally driven by the trade-off between spatial localization precision and algorithmic scalability. While traditional CNN-based algorithms have their applicable environments, their reliance on heuristic post-processing limits their robustness. Nonetheless, with rapid advancements in deep learning—particularly the emergence of fully differentiable Transformer architectures—the precision and global reasoning capabilities of pose estimation models have significantly improved. Nevertheless, critical challenges remain, most notably, severe occlusion and high inference latency. To solve these problems, future research directions will be shifted to lightweight network optimization for near-term edge deployment, as well

as zero-shot data generation and cross-modal 3D priors for long-term breakthroughs. To conclude, it is obvious that ongoing technology and endeavors will broaden the horizons of human pose estimation, gradually overcoming its inherent architectural challenges.

Author Contributions: Y.Z. (Yujie Zhang): Writing and Investigation; D.L.: Conceptualization and Theory analysis; Y.Y.: Writing and Revising; L.Z.: Writing and Original draft preparation; S.G.: Writing and Investigation; Y.Z. (Yufei Zhao): Writing and Investigation. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China under Grant No. 62461041, Natural Science Foundation of Jiangxi Province under Grant No. 20242BAB25068, and China Scholarship Council under Grant No. 202106825021.

Institutional Review Board Statement: Not applicable. This study is a survey article and did not involve any new experiments on humans or animals.

Informed Consent Statement: Not applicable. This study did not involve human participants or identifiable personal data.

Data Availability Statement: Some or all data, models, or code generated or used during the study are available from the corresponding author by request.

Acknowledgments: The authors would like to thank Deyu.Lin for their valuable technical support and assistance in the experiments.

Conflicts of Interest: All authors declare that there are no conflicts of interest.

References

1. Souto, H.; Musse, S. Automatic detection of 2D human postures based on single images. In Proceedings of the 2011 24th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI'11), Alagoas, Brazil, 28–31 August 2011; pp. 48–55.
2. Kreiss, S.; Bertoni, L.; Alahi, A. Pifpaf: Composite fields for human pose estimation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'19), Long Beach, CA, USA, 15–20 June 2019; pp. 11977–11986.
3. Insafutdinov, E.; Andrilukam, M.; Pishchulin, L.; Tang, S.; Levinkov, E.; Andres, B.; Schiele, B. ArtTrack: Articulated multi-person tracking in the wild. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17), Honolulu, HI, USA, 21–26 July 2017; pp. 1293–1301.
4. Fischler, M.A.; Elschlager, R.A. The representation and matching of pictorial structures. *IEEE Trans. Comput.* **1973**, *22*, 67–92. [[CrossRef](#)]
5. Yang, Y.; Ramanan, D. Articulated pose estimation with flexible mixtures-of-parts. In Proceedings of the 2011 IEEE conference on computer vision and pattern recognition (CVPR'11), Colorado Springs, CO, USA, 20–25 June 2011; pp. 1385–1392.
6. Johnson, S.; Everingham, M. Learning effective human pose estimation from inaccurate annotation. In Proceedings of the 2011 IEEE conference on computer vision and pattern recognition (CVPR'11), Colorado Springs, CO, USA, 20–25 June 2011; pp. 1465–1472.
7. Yu, Z.; Li, Y.; Liu, Y.; Liu, T.; Fu, Y. Synpose: A large-scale and densely annotated synthetic dataset for human pose estimation in classroom. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'22), Singapore, 23–27 May 2022; pp. 3428–3432.
8. An, W.; Yu, S.; Makihara, Y.; Wu, X.; Xu, C.; Yu, Y.; Liao, R.; Yagi, Y. Performance evaluation of model-based gait on multi-view very large population database with pose sequences. *IEEE Trans. Biom. Behav. Identity Sci.* **2020**, *2*, 421–430.
9. Lowe, D.G. Object recognition from local scale-invariant features. In Proceedings of the 1999 International Conference on Computer Vision (ICCV'99), Kerkyra, Greece, 20–27 September 1999; pp. 1150–1157.
10. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), San Diego, CA, USA, 20–26 June 2005; pp. 886–893.
11. 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](#)]
12. Zhang, C.; Jiang, X.; Zhao, Y. Efficient instantaneous channel propagation modeling for aeronautical communications systems with compressed sensing. *IEEE Trans. Antennas Propag.* **2022**, *70*, 1211–1220.
13. Zhao, Y.; Zhang, C. Orbital angular momentum beamforming for index modulation with partial arc reception. *Electron. Lett.* **2019**, *55*, 1271–1273. [[CrossRef](#)]

14. Ismail, A.M.; Zhao, Y.; Wang, Z.; Guan, Y.L.; Yuen, C. Visually steered reconfigurable intelligent surface-assisted mobile communications. *IEEE Antennas Wirel. Propag. Lett.* **2025**, *24*, 4497–4501. [\[CrossRef\]](#)
15. Hinton, G.; Deng, L.; Yu, D.; Dahl, G.E.; Mohamed, A.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T.N.; et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.* **2012**, *29*, 82–97. [\[CrossRef\]](#)
16. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [\[CrossRef\]](#)
17. Huang, Z.; Liu, Y.; Fang, Y.; Horn, B.K.P. Video-based fall detection for seniors with human pose estimation. In Proceedings of the 2018 4th international conference on Universal Village (UV'18), Boston, MA, USA, 12–14 October 2018; pp. 1–4.
18. Perez-Sala, X.; Escalera, S.; Angulo, C.; González, J. A survey on model based approaches for 2D and 3D visual human pose recovery. *Sensors* **2014**, *14*, 4189–4210. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Chen, Y.; Tian, Y.; He, M. Monocular human pose estimation: A survey of deep learning-based methods. *Comput. Vis. Image Underst.* **2020**, *192*, 102897. [\[CrossRef\]](#)
20. Wang, C.; Zhang, F.; Ge, S.S. A comprehensive survey on 2D multi-person pose estimation methods. *Eng. Appl. Artif. Intel.* **2021**, *102*, 104260.
21. Murphy-Chutorian, E.; Trivedi, M.M. Head pose estimation in computer vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *31*, 607–626. [\[CrossRef\]](#)
22. Saroja, M.N.; Baskaran, K.R.; Priyanka, P. Human pose estimation approaches for human activity recognition. In Proceedings of the 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA'21), Coimbatore, India, 8–9 October 2021; pp. 1–4.
23. Mitra, S.; Acharya, T. Gesture recognition: A survey. *IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.)* **2007**, *37*, 311–324. [\[CrossRef\]](#)
24. Liu, Z.; Zhu, J.; Bu, J.; Chen, C. A survey of human pose estimation: The body parts parsing based methods. *J. Vis. Commun. Image R* **2015**, *32*, 10–19. [\[CrossRef\]](#)
25. Zhang, H.; Lei, Q.; Zhong, B.; Du, J.; Peng, J. A survey on human pose estimation. *Intell. Autom. Soft Comput.* **2015**, *22*, 483–489. [\[CrossRef\]](#)
26. Gong, W.; Zhang, X.; González, J.; Sobral, A.; Bouwmans, T.; Tu, C.; Zahzah, E. Human pose estimation from monocular images: A comprehensive survey. *Sensors* **2016**, *16*, 1966. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Sun, J.; Chen, X.; Lu, Y.; Cao, J. 2D human pose estimation from monocular images: A survey. In Proceedings of the 2020 IEEE 3rd International Conference on Computer and Communication Engineering Technology (CCET'20), Beijing, China, 21–23 August 2020; pp. 111–121.
28. Zheng, C.; Wu, W.; Chen, C.; Yang, T.; Zhu, S.; Shen, J. Deep learning-based human pose estimation: A survey. *ACM Comput. Surv.* **2023**, *56*, 1–37. [\[CrossRef\]](#)
29. 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 (CVPR'13), Portland, OR, USA, 23–28 June 2013; pp. 3674–3681.
30. Lin, T.; 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 13th European on Computer Vision (ECCV'14), Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
31. Johnson, S.; Everingham, M. Clustered pose and nonlinear appearance models for human pose estimation. In Proceedings of the British Machine Vision Conference (BMVC'10), Aberystwyth, UK, 31 August–3 September 2010; pp. 1–11.
32. 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 the IEEE Conference on computer Vision and Pattern Recognition (CVPR'14), Columbus, OH, USA, 23–28 June 2014; pp. 3686–3693.
33. Wu, J.; Zheng, H.; Zhao, B.; Li, Y.; Yan, B.; Liang, R.; Wang, W.; Zhou, S.; Lin, G.; Fu, Y.; et al. AI challenger: A large-scale dataset for going deeper in image understanding. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME'19), Shanghai, China, 8–12 July 2019; pp. 1–11.
34. Zhang, W.; Zhu, M.; Derpanis, K.G. From actemes to action: A strongly supervised representation for detailed action understanding. In Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV'13), Sydney, Australia, 1–8 December 2013; pp. 2248–2255.
35. Jhuang, H.; Gall, J.; Zuffi, S.; Schmid, C.; Black, M.J. Towards understanding action recognition. In Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV'13), Sydney, Australia, 1–8 December 2013; pp. 3192–3199.
36. Andriluka, M.; Iqbal, U.; Milan, A.; Insafutdinov, E.; Pishchulin, L.; Gall, J.; Schiele, B. Posetrack: A benchmark for human pose estimation and tracking. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'18), Salt Lake City, UT, USA, 18–22 June 2018; pp. 5167–5176.

37. Park, S.; Lee, S.; Lee, S.H. Enhanced prediction model for human activity using an end-to-end approach. *IEEE Internet Thing J.* **2023**, *10*, 6031–6041.
38. Lin, J.; Zeng, A.; Wang, H.; Zhang, L.; Li, Y. UBody: A Million-Scale Dataset for Whole-Body Human Pose Estimation. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–21 June 2023; pp. 2131–2141.
39. Ju, X.; Zeng, A.; Wang, J.; Xu, Q.; Zhang, L. Human-Art: A Versatile Human-Centric Dataset for Artworks. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–21 June 2023; pp. 618–629.
40. Ferrari, V.; Marin-Jimenez, M.; Zisserman, A. Progressive search space reduction for human pose estimation. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
41. Yang, Y.; Ramanan, D. Articulated human detection with flexible mixtures of parts. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 2878–2890. [[CrossRef](#)]
42. Zhao, L.; Xu, J.; Gong, C.; Yang, J.; Zuo, W.; Gao, X. Learning to acquire the quality of human pose estimation. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 1555–1568.
43. Ning, G.; Zhang, Z.; He, Z. Knowledge-guided deep fractal neural networks for human pose estimation. *IEEE Trans. Multimed.* **2018**, *20*, 1246–1259.
44. Pfister, T.; Simonyan, K.; Charles, J.; Zisserman, A. Deep convolutional neural networks for efficient pose estimation in gesture videos. In Proceedings of the 12th Asian Conference on Computer Vision (ACCV'14), Singapore, 1–5 November 2014; pp. 538–552.
45. Luvizon, D.C.; Hedi, T.; David, P. Human pose regression by combining indirect part detection and contextual information. In Proceedings of the 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)/Symposium on Virtual and Augmented Reality (SVR'19), Rio de Janeiro, Brazil, 28–31 October 2019; pp. 15–22.
46. 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 (CVPR'14), Columbus, OH, USA, 23–28 June 2014; pp. 1653–1660.
47. Zhang, F.; Zhu, X.; Ye, M. Fast human pose estimation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'19), Long Beach, CA, USA, 16–20 June 2019; pp. 3512–3521.
48. Das, A.; Chakraborty, A.; Roy-Chowdhury, A.K. Consistent re-identification in a camera network. In Proceedings of the 13th European Conference on Computer Vision (ECCV'14), Zurich, Switzerland, 6–12 September 2014; pp. 330–345.
49. Li, S.; Liu, Z.; Chan, A. Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'14), Columbus, OH, USA, 23–28 June 2014; pp. 488–496.
50. Lifshitz, I.; Fetaya, E.; Ullman, S. Human pose estimation using deep consensus voting. In Proceedings of the 14th European Conference on Computer Vision (ECCV'16), Amsterdam, The Netherlands, 8–16 October 2016; pp. 246–260.
51. Chen, X.; Yuille, A.L. Articulated pose estimation by a graphical model with image dependent pairwise relations. In Proceedings of the 28th Conference on Neural Information Processing Systems (NIPS'14), Montreal, QC, Canada, 8–13 December 2014; pp. 1736–1744.
52. Varamesh, A.; Tuytelaars, T. Mixture dense regression for object detection and human pose estimation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20), Seattle, WA, USA, 14–19 June 2020; pp. 13086–13095.
53. Xiao, B.; Wu, H.; Wei, Y. Simple baselines for human pose estimation and tracking. In Proceedings of the 15th European Conference on Computer Vision (ECCV'18), Munich, Germany, 8–14 September 2018; pp. 472–487.
54. Belagiannis, V.; Zisserman, A. Recurrent human pose estimation. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG'2017), Washington, DC, USA, 30 May –3 June 2017; pp. 468–475.
55. Yang, W.; Li, S.; Ouyang, W.; Li, H.; Wang, X. Learning feature pyramids for human pose estimation. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV'17), Venice, Italy, 22–29 October 2017; pp. 1281–1290.
56. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'19), Long Beach, CA, USA, 16–20 June 2019; pp. 5686–5696.
57. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In Proceedings of the 14th European Conference on Computer Vision (ECCV'16), Amsterdam, The Netherlands, 8–16 October 2016; pp. 483–499.
58. Carreira, J.; Agrawal, P.; Fragkiadaki, K.; Malik, J. Human pose estimation with iterative error feedback. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16), Las Vegas, NV, USA, 27–30 June 2016; pp. 4733–4742.
59. Sun, X.; Shang, J.; Liang, S.; Wei, Y. Compositional human pose regression. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV'17), Venice, Italy, 22–29 October 2017; pp. 2621–2630.

60. Ke, L.; Chang, M.; Qi, H.; Lyu, S. Multi-scale structure aware network for human pose estimation. In Proceedings of the 15th European Conference on Computer Vision (ECCV'18), Munich, Germany, 8–14 September 2018; pp. 731–746.
61. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 3349–3364. [[CrossRef](#)] [[PubMed](#)]
62. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the 31st Conference on Artificial Intelligence (AAAI'17), San Francisco, CA, USA, 4–9 February 2017; pp. 4278–4284.
63. Wei, S.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. Convolutional pose machines. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16), Las Vegas, NV, USA, 27–30 June 2016; pp. 4724–4732.
64. Chu, X.; Yang, W.; Ouyang, W.; Ma, C.; Yuille, A.L.; Wang, X. Multi-context attention for human pose estimation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17), Honolulu, HI, USA, 21–26 July 2017; pp. 1831–1840.
65. Tang, W.; Yu, P.; Wu, Y. Deeply learned compositional models for human pose estimation. In Proceedings of the 15th European Conference on Computer Vision (ECCV'18), Munich, Germany, 8–14 September 2018; pp. 190–206.
66. Chou, C.; Chien, J.; Chen, H. Self adversarial training for human pose estimation. In Proceedings of the 10th Asia-Pacific-Signal-and-Information-Processing-Association Annual Summit and Conference (APSIPA ASC'18), Honolulu, HI, USA, 12–15 November 2018; pp. 17–30.
67. Chen, Y.; Shen, C.; Wei, X.; Liu, L.; Yang, J. Adversarial posenet: A structureaware convolutional network for human pose estimation. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV'17), Venice, Italy, 22–29 October 2017; pp. 1212–1221.
68. Bulat, A.; Kossai, J.; Tzimiropoulos, G.; Pantic, M. Toward fast and accurate human pose estimation via softgated skip connections. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG'20), Buenos Aires, Argentina, 16–20 November 2020; pp. 8–15.
69. Jain, A.; Tompson, J.; LeCun, Y.; Bregler, C. Modeep: A deep learning framework using motion features for human pose estimation. In Proceedings of the 12th Asian Conference on Computer Vision (ACCV'14), Singapore, 1–5 November 2014; pp. 302–315.
70. Pfister, T.; Charles, J.; Zisserman, A. Flowing convnets for human pose estimation in videos. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV'15), Santiago, Chile, 7–13 December 2015; pp. 1913–1921.
71. Luo, Y.; Ren, J.; Wang, Z.; Sun, W.; Pan, J.; Liu, J.; Pang, J.; Lin, L. Lstm pose machines. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'18), Salt Lake City, UT, USA, 18–22 June 2018; pp. 5207–5215.
72. Vidanapathirana, M.; Sudasingha, I.; Vidanapathirana, J.; Kanchana, P.; Perera, I. Tracking and frame-rate enhancement for real-time 2D human pose estimation. *Vis. Comput.* **2019**, *36*, 1501–1519. [[CrossRef](#)]
73. Artacho, B.; Savakis, A. UniPose, unified human pose estimation in single images and videos. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20), Seattle, WA, USA, 14–19 June 2020; pp. 7035–7044.
74. Yang, W.; Ouyang, W.; Li, H.; Wang, X. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16), Las Vegas, NV, USA, 27–30 June 2016; pp. 3073–3082.
75. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
76. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
77. 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. In Proceedings of the 16th European Conference on Computer Vision (ECCV'20), Glasgow, UK, 23–28 August 2020; pp. 455–472.
78. Zhang, F.; Zhu, X.; Dai, H.; Ye, M.; Zhu, C. Distribution-aware coordinate representation for human pose estimation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20), Seattle, WA, USA, 14–19 June 2020; pp. 7091–7100.
79. Iqbal, U.; Gall, J. Multi-person pose estimation with local joint-to-person associations. In Proceedings of the 14th European Conference on Computer Vision (ECCV'14), Amsterdam, The Netherlands, 8–16 October 2016; pp. 627–642.
80. Huang, S.; Gong, M.; Tao, D. A coarse-fine network for keypoint localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV'17), Venice, Italy, 22–29 October 2017; pp. 3047–3056.
81. Shi, D.; Wei, X.; Yu, X.; Tan, W.; Ren, Y.; Pu, S. Inpose: Instance-aware networks for single-stage multi-person pose estimation. In Proceedings of the 29th ACM International Conference on Multimedia (ACM Multimedia'21), Chengdu, China, 17–21 October 2021; pp. 3079–3087.

82. Kocabas, M.; Karagoz, S.; Akbas, E. Multiposenet: Fast multi-person pose estimation using pose residual network. In Proceedings of the 15th European Conference on Computer Vision (ECCV'18), Munich, Germany, 8–14 September 2018; pp. 417–433.
83. Newell, A.; Huang, Z.; Deng, J. Associative embedding: End-to-end learning for joint detection and grouping. In Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS'17), Long Beach, CA, USA, 4–9 December 2017; pp. 2278–2288.
84. Nie, X.; Feng, J.; Zhang, J.; Yan, S. Single-stage multi-person pose machines. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV'19), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6950–6959.
85. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17), Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
86. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the 14th European Conference on Computer Vision (ECCV'16), Amsterdam, The Netherlands, 8–16 October 2016; pp. 21–37.
87. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'14), Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
88. Xu, Y.; Zhang, J.; Zhang, Q.; Tao, D. ViTPose++: Vision Transformer Foundation Model for Generic Body Pose Estimation. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)* **2024**, *46*, 1212–1230.
89. Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; Sun, J. Cascaded pyramid network for multi-person pose estimation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'18), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7103–7112.
90. Yang, J.; Zeng, A.; Liu, S.; Li, F.; Zhang, R.; Zhang, L. Explicit box detection unifies end-to-end multi-person pose estimation. In Proceedings of the International Conference on Learning Representations (ICLR'23), Kigali, Rwanda, 1–5 May 2023; pp. 577–594.
91. Papandreou, G.; Zhu, T.; Chen, L.C.; Gidaris, S.; Tompson, J.; Murphy, K. PersonLab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In Proceedings of the European Conference on Computer Vision (ECCV'18), Munich, Germany, 8–14 September 2018; pp. 1–7.
92. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV'17), Venice, Italy, 22–29 October 2017; pp. 2961–2969.
93. Yang, Z.; Zeng, A.; Yuan, C.; Li, Y. Effective Whole-body Pose Estimation with Two-stage Distillation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 14850–14860.
94. Fang, H.; Xie, S.; Tai, Y.; Lu, C. Rmpe: Regional multi-person pose estimation. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV'17), Venice, Italy, 22–29 October 2017; pp. 2334–2343.
95. Hu, P.; Ramanan, D. Bottom-up and top-down reasoning with hierarchical rectified gaussians. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16), Las Vegas, NV, USA, 27–30 June 2016; pp. 5600–5609.
96. Bodla, N.; Singh, B.; Chellappa, R.; Davis, L.S. Soft-NMS-Improving object detection with one line of code. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV'17), Venice, Italy, 22–29 October 2017; pp. 5561–5569.
97. Osokin, D. Real-time 2D multi-person pose estimation on CPU: Lightweight openpose. In Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods (ICPRAM'19), Prague, Czech Republic, 19–21 February 2019; pp. 744–748.
98. Papandreou, G.; Zhu, T.; Kanazawa, N.; Toshev, A.; Tompson, J.; Bregler, C.; Murphy, K. Towards accurate multi-person pose estimation in the wild. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17), Honolulu, HI, USA, 21–26 July 2017; pp. 4903–4911.
99. Brasó, G.; Kister, N.; Leal-Taixé, L. The center of attention: Center-keypoint grouping via attention for multi-person pose estimation. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV'21), Montreal, QC, Canada, 11–17 October 2021; pp. 11853–11863.
100. 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'19), Long Beach, CA, USA, 16–20 June 2019; pp. 10863–10872.
101. Wang, D.; Zhang, S. Contextual Instance Decoupling for Robust Multi-Person Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 13134–13143.
102. Xia, F.; Wang, P.; Chen, X.; Yuille, A.L. Joint multi-person pose estimation and semantic part segmentation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17), Honolulu, HI, USA, 21–26 July 2017; pp. 6769–6778.
103. Cao, Z.; Simon, T.; Wei, S.; Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17), Honolulu, HI, USA, 21–26 July 2017; pp. 7291–7299.

104. Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T.S.; Zhang, L. HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20), Seattle, WA, USA, 14–19 June 2020; pp. 5385–5394.
105. Li, M.; Zhou, Z.; Liu, X. Multi-person pose estimation using bounding box constraint and LSTM. *IEEE Trans. Multimed.* **2019**, *21*, 2653–2663. [[CrossRef](#)]
106. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV'15), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
107. Tian, C.; Yu, R.; Zhao, X.; Xia, W.; Wang, H.; Yang, Y. Posedet: Fast multi-person pose estimation using pose embedding. In Proceedings of the 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG'21), Jodhpur, India, 15–18 December 2021; pp. 1–8.
108. Insafutdinov, E.; Pishchulin, L.; Andres, B.; Andriluka, M.; Schiele, B. Deepcrut: A deeper, stronger, and faster multi-person pose estimation model. In Proceedings of the 14th European Conference on Computer Vision (ECCV'16), Amsterdam, The Netherlands, 8–16 October 2016; pp. 34–50.
109. Liu, H.; Chen, Q.; Tan, Z.; Liu, 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. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 2146–2156.
110. Tan, D.; Chen, H.; Tian, W.; Xiong, L. DiffusionRegPose: Enhancing multi-person pose estimation using a diffusion-based end-to-end regression approach. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'24), Seattle, WA, USA, 17–21 June 2024; pp. 2230–2239.
111. Yang, S.; Quan, Z.; Nie, M.; Yang, W. TransPose: Keypoint localization via transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'2021), Montreal, QC, Canada, 11–17 October 2021; pp. 11802–11812.
112. 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 IEEE/CVF International Conference on Computer Vision (ICCV'21), Montreal, QC, Canada, 11–17 October 2021; pp. 11313–11322.
113. Chen, S.; Zhang, Y.; Huang, S.; Yi, R.; Fan, K.; Zhang, R.; Chen, P.; Wang, J.; Ding, S.; Ma, L. SDPose: Tokenized pose estimation via circulation-guide self-distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'2024), Seattle, WA, USA, 17–21 June 2024; pp. 1082–1090.
114. Fan, X.; Zheng, K.; Lin, Y.; Wang, S. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15), Boston, MA, USA, 7–12 June 2015; pp. 1347–1355.
115. Nie, X.; Feng, J.; Xing, J.; Yan, S. Pose partition networks for multi-person pose estimation. In Proceedings of the 15th European Conference on Computer Vision (ECCV'15), Munich, Germany, 8–14 September 2018; pp. 684–699.
116. Wei, F.; Sun, X.; Li, H.; Wang, J.; Lin, S. Point-set anchors for object detection, instance segmentation and pose estimation. In Proceedings of the 16th European Conference on Computer Vision (ECCV'15), Glasgow, UK, 23–28 August 2020; pp. 527–544.
117. Jin, S.; Xu, L.; Xu, J.; Wang, C.; Liu, W.; Qian, C.; Ouyang, W.; Luo, P. Whole-body human pose estimation in the wild. In Proceedings of the 16th European Conference on Computer Vision (ECCV'20), Glasgow, UK, 23–28 August 2020; pp. 196–214.
118. Geng, Z.; Sun, K.; Xiao, B.; Zhang, Z.; Wang, J. Bottom-up human pose estimation via disentangled keypoint regression. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21), Nashville, TN, USA, 20–25 June 2021; pp. 14671–14681.
119. Zhao, L.; Wen, J.; Wang, P.; Zheng, N. Context-guided adaptive network for efficient human pose estimation. In Proceedings of the 35th Conference on Artificial Intelligence (AAAI'21), Virtual Event, 2–9 February 2021; pp. 3492–3499.
120. Xiao, P.; Qin, Z.; Chen, D.; Zhang, N.; Ding, Y.; Deng, F.; Qin, Z.; Pang, M. Fastnet: A lightweight convolutional neural network for tumors fast identification in mobile computer-assisted devices. *IEEE Internet Things J.* **2023**, *10*, 9878–9891.
121. Wang, Y.; Li, M.; Cai, H.; Chen, W.; Han, S. Lite pose: Efficient architecture design for 2d human pose estimation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'22), New Orleans, LA, USA, 18–24 June 2022; pp. 13116–13126.
122. Zhang, S.; Li, R.; Dong, X.; Rosin, P.L.; Cai, Z.; Han, X.; Yang, D.; Huang, H.; Hu, S. Pose2seg: Detection free human instance segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'19), Long Beach, CA, USA, 16–20 June 2019; pp. 889–898.
123. Wang, J.; Long, X.; Gao, Y.; Ding, E.; Wen, S. Graph-pcnn: Two stage human pose estimation with graph pose refinement. In Proceedings of the 16th European Conference on Computer Vision (ECCV'20), Glasgow, UK, 23–28 August 2020; pp. 492–508.
124. Isack, H.; Haene, C.; Keskin, C.; Bouaziz, S.; Boykov, Y.; Izadi, S.; Khamis, S. Repose: Learning deep kinematic priors for fast human pose estimation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20), Seattle, WA, USA, 14–19 June 2020; pp. 1–9.

125. Zhang, K.; Yao, P.; Wu, R.; Yang, C.; Li, D.; Du, M.; Deng, K.; Liu, R.; Zheng, T. Learning positional priors for pretraining 2D pose estimators. In Proceedings of the 2nd International Workshop on Human-Centric Multimedia Analysis (HUMA'21), Virtual Event, China, 17 October 2021; pp. 3–11.
126. Su, Z.; Xu, L.; Zheng, Z.; Yu, T.; Liu, Y.; Fang, L. Robustfusion: Human volumetric capture with data-driven visual cues using a rgbd camera. In Proceedings of the 16th European Conference on Computer Vision (ECCV'20), Glasgow, UK, 23–28 August 2020; pp. 246–264.
127. Vu, H.T.; Wilkinson, R.H.; Lech, M.; Cheng, E. A hybrid neural network for graph-based human pose estimation from 2D images. *IEEE Access* **2020**, *8*, 52830–52840. [[CrossRef](#)]
128. Silva, L.J.S.; Silva, D.L.S.; Raposo, A.; Velho, L.; Lopes, H. Tensorpose: Real-time pose estimation for interactive applications. *Comput. Graph.* **2019**, *85*, 1–14. [[CrossRef](#)]
129. Yu, C.; Xiao, B.; Gao, C.; Yuan, L.; Zhang, L.; Sang, N.; Wang, J. Lite-HRNet: A lightweight high-resolution network. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21), Nashville, TN, USA, 20–25 June 2021; pp. 10435–10445.
130. Nie, X.; Li, Y.; Luo, L.; Zhang, N.; Feng, J. Dynamic kernel distillation for efficient pose estimation in videos. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV'19), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6941–6949.
131. Zhou, Y. Role of human body posture recognition method based on wireless network kinect in line dance aerobics and gymnastics training. *Wirel. Commun. Mob. Comput.* **2021**, *2021*, 9208891. [[CrossRef](#)]
132. Romero, J.; Loper, M.; Black, M.J. Flowcap: 2d human pose from optical flow. In Proceedings of the 37th German Conference on Pattern Recognition (GCPR'15), Aachen, Germany, 7–10 October 2015; pp. 412–423.
133. Ghosh, A.; Kulharia, V.; Namboodiri, V.P.; Torr, P.H.S.; Dokania, P.K. Multi-agent diverse generative adversarial networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'18), Salt Lake City, UT, USA, 18–22 June 2018; pp. 8513–8521.
134. Lin, Z.; Khetan, A.; Fanti, G.; Oh, S. PacGAN: The power of two samples in generative adversarial networks. In Proceedings of the 32th Annual Conference on Neural Information Processing Systems (NeurIPS'18), Montreal, QC, Canada, 3–8 December 2018; pp. 1498–1507.
135. Wang, K.; Lin, L.; Jiang, C.; Zheng, W.S. 3D human pose machines with self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 1069–1082. [[PubMed](#)]
136. Fries, J.A.; Steinberg, E.; Khattar, S.; Fleming, S.L.; Posada, J.; Callahan, A.; Shah, N.H. Ontology-driven weak supervision for clinical entity classification in electronic health records. *Nat. Commun.* **2021**, *12*, 2017. [[CrossRef](#)] [[PubMed](#)]
137. Raychaudhuri, D.S.; Ta, C.K.; Dutta, A.; Lal, R.; Roy-Chowdhury, A.K. Prior-guided source-free domain adaptation for human pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'23), Paris, France, 2–6 October 2023; pp. 14996–15006.

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.