Review

2D Human Pose Estimation with Deep Learning: A Review

Zheyu Zhang and Seong-Yoon Shin *

Posted Date: 13 June 2025

doi: 10.20944/preprints202506.1093.v1

Keywords: 2D human pose estimation; deep learning; top-down methods; bottom-up methods; benchmark datasets; evaluation metrics; keypoint detection

 doi:10.20944/preprints202506.1093.v1

Review

2D Human Pose Estimation with Deep Learning: A Review

Zheyu Zhang and Seong-Yoon Shin *

Department of Computer Science and Information Engineering, Kunsan National University, Gunsan 54150, Republic of Korea

* Correspondence: s3397220@kunsan.ac.kr; Tel.: +82-63-469-4860

Abstract: Two-dimensional human pose estimation (2D HPE) has become a fundamental task in computer vision, driven by growing demands in intelligent surveillance, sports analytics, and healthcare. The rapid advancement of deep learning has led to the development of numerous methods. However, the resulting diversity in research directions and model architectures has made systematic assessment and comparison difficult. This review presents a comprehensive overview of recent advances in 2D HPE, focusing on method classification, technical evolution, and performance evaluation. We classify mainstream approaches by task type (single-person vs. multi-person), output strategy (regression vs. heatmap), and architectural design (top-down vs. bottom-up) and analyze their respective strengths, limitations, and application scenarios. Additionally, we summarize commonly used evaluation metrics and benchmark datasets such as MPII, COCO, LSP, OCHuman, and CrowdPose. A major contribution of this review is the detailed comparison of the top six models on each benchmark, highlighting their network architectures, input resolutions, evaluation results, and key innovations. In light of current challenges, we also outline future research directions, including model compression, occlusion handling, and cross-domain generalization. This review serves as a valuable reference for researchers seeking both foundational insights and practical guidance in 2D human pose estimation.

Keywords: 2D human pose estimation; deep learning; top-down methods; bottom-up methods; benchmark datasets; evaluation metrics; keypoint detection;

1. Introduction

1.1. Background

The continuous advancement of computer vision and deep learning has made machine understanding of human behavior a critical research objective. Human pose estimation (HPE) is a central task in this domain, aiming to identify human joint positions and reconstruct skeletal structures from static images or video sequences. Two-dimensional human pose estimation (2D HPE) focuses on detecting keypoints within the image plane. Although it does not recover three-dimensional information, 2D HPE remains challenging, especially under uncontrolled conditions such as occlusion, multi-person interactions, and background clutter. Compared to 3D approaches, 2D HPE requires less computational power, is more data-accessible, and is easier to deploy on consumer-grade hardware, making it well-suited for both academic research and practical applications.

Due to its efficiency and practicality, 2D HPE has been widely integrated into numerous downstream tasks as a foundational component. For instance, in video surveillance, pose estimation enables abnormal behavior detection and crowd dynamics modeling. In sports analytics, it contributes to motion decomposition and technique refinement. In intelligent healthcare, it facilitates posture assessment and rehabilitation monitoring [1]. Moreover, the rising demand for contactless

 doi:10.20944/preprints202506.1093.v1

interaction and remote healthcare underscores the need for accurate and real-time pose estimation systems.

In recent years, deep learning has propelled substantial advances in human pose estimation (HPE) by overcoming the constraints of handcrafted-feature methods. Deep-learning architectures, including convolutional neural networks (CNNs) and graph convolutional networks (GCNs), can extract hierarchical image features. This capability enables accurate and robust joint localization under complex conditions [2,3]. Consequently, 2D HPE performance in multi-person scenes—characterized by high spatial and semantic ambiguity—has improved markedly.

Moreover, deep learning approaches have demonstrated high adaptability in addressing occlusion and crowded scenes—two persistent challenges in human pose estimation. Traditional models often struggle when keypoints are partially occluded or spatially overlapping. In contrast, deep models can leverage contextual cues and employ data augmentation techniques to improve robustness. Recent approaches have utilized generative adversarial networks (GANs) and self-supervised learning strategies to synthesize occlusion-aware training data, thereby enhancing model generalization and stability [4]. These techniques not only improve the accuracy of single-person pose estimation but also advance multi-person pose estimation and interaction understanding.

Despite substantial progress, challenges remain regarding model interpretability and computational efficiency. In safety-critical domains such as medical diagnostics and public surveillance, model transparency is indispensable. Consequently, developing trustworthy, lightweight deep-learning models that deliver reliable predictions under resource constraints has become a primary research objective. Current efforts concentrate on compact architectures that preserve high accuracy while reducing latency and memory consumption, thereby enabling deployment in diverse real-world scenarios.

In summary, deep learning has fundamentally transformed 2D human pose estimation by improving its accuracy and robustness and broadening its applicability across diverse domains. As research advances, 2D HPE is poised to play an increasingly pivotal role in bridging human behavior analysis and intelligent visual perception.

1.2. Research Status

Early approaches to 2D human pose estimation (2D HPE) originated from graphical model-based methods, including the Pictorial Structures Model [5] and the Deformable Part Model [6]. These methods represent the human body as a structured graph of joints and limbs, performing pose inference by modeling spatial dependencies among joints. Although interpretable, these methods rely heavily on handcrafted features and simple geometric representations, limiting their robustness in complex scenes and varied pose configurations. With the advent of deep learning and data-driven paradigms, 2D HPE research has shifted from feature-engineered models to deep neural network–based approaches trained on large-scale datasets. Contemporary mainstream methods are broadly categorized into two paradigms: top-down and bottom-up approaches.

Top-down approaches typically decompose the pose estimation process into two sequential stages. In the first stage, a human detector is used to localize all person instances in the image. In the second stage, keypoints are detected independently within each bounding box. These methods provide high localization accuracy and perform particularly well in scenes with few individuals and minimal occlusion.

Early representative methods, such as OpenPose [7] and Mask R-CNN [8], introduced structured outputs and established end-to-end learning pipelines for pose estimation. Subsequently, SimpleBaseline [9] simplified the architecture by employing a residual network for feature extraction and directly regressing keypoint heatmaps, achieving consistent performance across multiple benchmarks. HRNet [10] further advanced the field by maintaining high-resolution feature representations throughout the network, thereby mitigating information loss typical of conventional architectures and enhancing multi-scale pose representations. However, the computational cost of top-down methods scales linearly with the number of detected individuals, which limits their

 doi:10.20944/preprints202506.1093.v1

efficiency. Moreover, these methods exhibit reduced robustness in crowded scenes and under severe occlusion, with performance heavily dependent on the accuracy of the human detection stage.

Bottom-up approaches detect all keypoints across the entire image in a single forward pass and subsequently group them into individual instances using spatial or semantic cues. By bypassing the person detection stage, these methods provide more predictable computational complexity and are better suited for images containing an unknown number of individuals or severe occlusion. Representative approaches include Part Affinity Fields (PAFs) [7] and Associative Embedding (AE) [11]. PAFs learn directional vector fields to connect specific joints, enabling the model to infer part-to-person associations. In contrast, AE generates embedding vectors for each keypoint and groups them based on feature similarity. These strategies enhance the flexibility and robustness of multi-person pose estimation, particularly in crowded scenes.

In recent years, HigherHRNet [12] has advanced bottom-up approaches by incorporating pyramid features while preserving low-level image details, thereby significantly improving the detection of small-scale human instances. However, bottom-up approaches still encounter challenges such as inaccurate keypoint assignments and ambiguous part associations, especially in scenes with similar poses or frequent human interactions. In such scenarios, constructing accurate and consistent pose representations remains challenging.

To address the limitations of traditional architectures, researchers have proposed several architectural innovations. One research direction introduces Graph Neural Networks (GNNs) to explicitly model keypoint relationships, thereby enforcing structural consistency constraints [13]. Another line of research, inspired by the success of Transformers in natural language processing, replaces traditional convolutional backbones with self-attention mechanisms to capture global dependencies across the image. Methods such as TokenPose [14] and HRFormer [15] leverage sequence modeling to better capture long-range joint dependencies.

In addition, lightweight architectures have garnered growing attention, particularly for deployment on resource-constrained platforms such as mobile and wearable devices, where model size and inference speed are critical. Architectures such as Lite-HRNet [16] and MobilePose [17] aim to balance accuracy and efficiency, thereby contributing to the diversification of pose estimation approaches.

1.3. Purpose and Structure of the Paper

This review provides a systematic analysis and methodological comparison of two-dimensional human pose estimation (2D HPE). In response to rapid technical evolution and the growing diversity of approaches, we categorize and analyze the principal task paradigms. These paradigms include single-person versus multi-person estimation, regression-based versus heatmap-based methods, and top-down versus bottom-up architectures. This classification helps readers understand the field's developmental trajectory and current design trends. To facilitate model selection and performance evaluation, we compile and compare six representative models that consistently rank among the top performers on authoritative benchmark datasets. The comparison covers network architecture, input resolution, evaluation metrics, and key technical innovations, enabling readers to assess each model's strengths and limitations. In addition, we summarize widely used evaluation metrics and highlight the design characteristics of public datasets commonly adopted for 2D HPE. By examining current technical challenges, we outline several potential directions for future research. The remainder of the paper is organized as follows: Section 2 reviews core methodologies; Section 3 introduces evaluation metrics and datasets; Section 4 presents a comparative analysis of top-performing models; and Section 5 discusses future development trends and concludes the paper.

2. 2D Human Pose Estimation

Two-dimensional (2D) human pose estimation aims to automatically localize anatomical keypoints from images or video frames. These keypoints typically include the head, torso, arms, and legs, which together form a skeletal representation used for motion analysis and behavioral

 doi:10.20944/preprints202506.1093.v1

understanding. Early methods relied on handcrafted features and traditional machine learning techniques, but exhibited limited generalization capability. With the advent of deep learning, convolutional neural networks (CNNs) have become the dominant paradigm for pose estimation. Their ability to automatically learn hierarchical features has significantly improved the accuracy and robustness of pose estimation models. This section reviews representative approaches in 2D human pose estimation and outlines key advancements and technical developments in the field. A taxonomy of existing methods is illustrated in Figure 1.

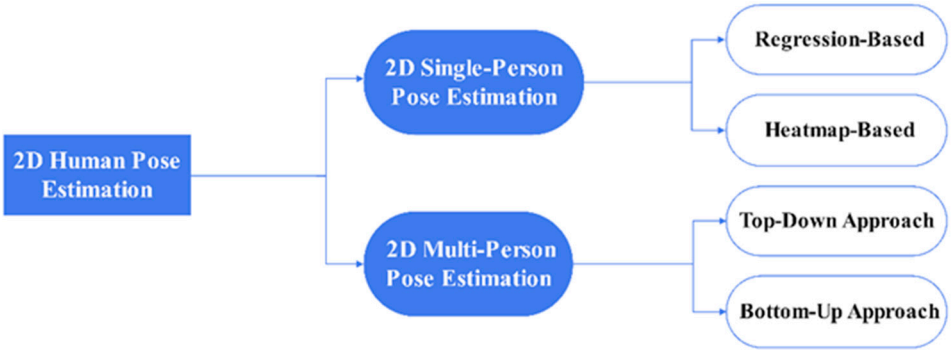

Figure 1. Taxonomy of two-dimensional human pose estimation methods.

2.1. 2D Single-Person Pose Estimation

For single-person scenarios, two-dimensional human pose estimation (2D HPE) commonly adopts either regression-based or heatmap-based approaches. Regression-based approaches formulate keypoint localization as an end-to-end regression task that maps image features directly to absolute coordinates or relative offsets of body joints. In contrast, heatmap-based methods cast keypoint detection as a probabilistic density estimation task, in which each joint is represented by a heatmap peaking at its most probable location. Figure 2 illustrates the workflows of (a) regression-based and (b) heatmap-based approaches.

Each approach has distinct advantages and limitations, making them suitable for different application scenarios. Regression-based methods predict continuous joint coordinates directly, avoiding quantization errors and offering high computational efficiency—making them particularly well-suited for real-time applications on mobile or embedded systems. However, they often struggle to preserve spatial relationships and are more susceptible to occlusion.

In contrast, heatmap-based methods localize joints by generating pixel-level probability maps, which more effectively capture spatial dependencies and exhibit greater robustness to occlusion. However, these methods require high-resolution heatmaps, which significantly increase computational cost. Moreover, the discretization of joint locations on heatmaps can introduce quantization errors. Despite these drawbacks, heatmap-based approaches remain the most widely adopted strategy in 2D human pose estimation. Representative methods from both paradigms will be discussed in the subsequent sections.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 June 2025 doi:10.20944/preprints202506.1093.v1

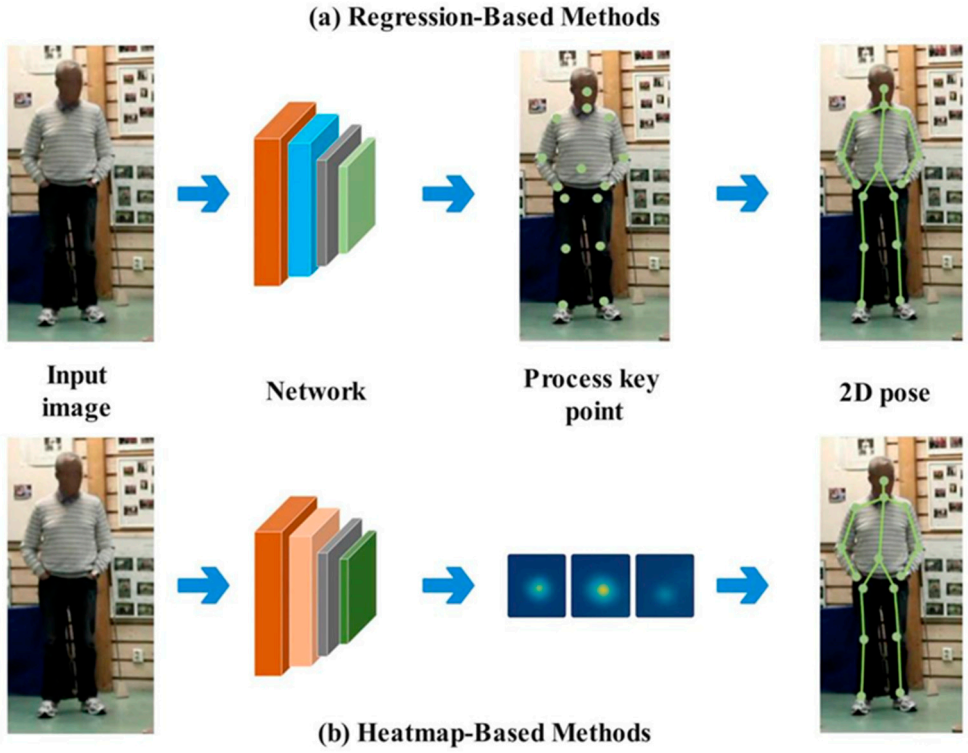

Figure 2. Methods for 2D single-person pose estimation: (a) regression-based approach and (b) heatmap-based approach.

2.1.1. Regression-Based Pose Estimation

Single-person 2D human pose estimation focuses on localizing body joints in images containing a single subject. Traditional approaches primarily relied on hand-crafted features and shallow architectures, which limited their representational capacity. A seminal work by Toshev et al. [18] introduced deep neural networks (DNNs) into pose estimation by formulating the task as direct regression to joint coordinates. The DeepPose framework is illustrated in Figure 3.

This pioneering approach introduced a cascaded regression framework that not only improved estimation accuracy but also effectively incorporated contextual information. It marked the first successful application of deep learning to human pose estimation and has since been widely adopted and extended, establishing itself as a foundational paradigm in the field.

Figure 3. DeepPose framework: a DNN-based pose regressor and optimizer.

Carreira et al. [19] proposed Iterative Error Feedback (IEF), a novel approach designed to improve the accuracy of human keypoint localization in complex scenarios. Compared to DeepPose, IEF incorporates several key enhancements. First, IEF employs a feedback mechanism that iteratively refines initial predictions, rather than estimating all keypoints in a single forward pass. This iterative correction process enhances overall localization accuracy. Second, it introduces bounded correction

 doi:10.20944/preprints202506.1093.v1

steps, allowing the model to make moderate and stable updates. In addition, IEF more effectively captures structural dependencies within the output space, helping to reduce the complexity of direct coordinate regression. Finally, the introduction of the Fixed Path Consolidation (FPC) strategy improves training efficiency by accelerating model convergence.

Sun et al. [20] proposed a simple yet effective integral operation that bridges and unifies heatmap-based representation with joint coordinate regression. This approach addresses key limitations of conventional heatmap-based methods, including non-differentiable post-processing and quantization errors. By integrating heatmap outputs using a differentiable operation, the method preserves the spatial richness of heatmaps while yielding continuous joint coordinates. As a result, it is considered an advanced variant of regression-based approaches. The fully differentiable nature of the integral operation enables end-to-end optimization. Moreover, the method is applicable to both 2D and 3D pose estimation tasks and supports joint training across them, substantially improving 3D pose estimation performance. The proposed integral regression paradigm introduced a new perspective for human pose estimation and has since been widely adopted and extended in subsequent studies.

Moon et al. [21] proposed PoseFix, a model-agnostic human pose refinement network designed to improve the accuracy of predictions generated by diverse mainstream pose estimation methods. Unlike conventional refinement strategies that require multi-stage end-to-end training and tight integration with specific architectures, PoseFix operates independently of any specific pose estimation model or its implementation. This design offers high flexibility and ease of use, enabling PoseFix to function as a post-processing module for improving a broad range of pose estimation outputs. Although originally developed as a post-hoc refinement tool, PoseFix introduced a novel regression-based refinement module that eliminates the need for heatmaps. This heatmap-free regression strategy has inspired follow-up studies in the area of regression-based pose optimization.

2.1.2. Heatmap -Based Pose Estimation

Heatmap-based methods have emerged as the dominant paradigm in two-dimensional (2D) human pose estimation. Tompson et al. [22] proposed a pioneering framework that jointly trains a deep convolutional neural network (ConvNet) with a Markov Random Field (MRF)-based graphical model. In this framework, the ConvNet extracts local image features and produces heatmap representations for body parts, while the MRF models spatial dependencies among joints. Joint optimization of these two components significantly enhances pose estimation performance, particularly in scenarios with complex joint interactions and occlusion. This work was among the first to integrate convolutional neural networks with heatmap-based representations and introduced a multi-scale feature fusion scheme that laid the foundation for numerous subsequent studies.

Wei et al. [23] introduced Convolutional Pose Machines (CPM), a cascaded architecture that progressively refines heatmap predictions through multiple processing stages. A key innovation of CPM is the use of intermediate supervision at each stage, which facilitates gradient flow and guides the learning process. By integrating convolutional neural networks with a pose machine framework, CPM implicitly captures long-range dependencies via stage-wise refinement. This multi-stage design, together with intermediate supervision, markedly enhances the accuracy and robustness of human pose estimation, particularly in challenging scenes with complex poses and occlusion.

Newell et al. [24] introduced the Stacked Hourglass convolutional network, a groundbreaking architecture for human pose estimation. This architecture captures features at multiple scales while preserving spatial information at each resolution via skip connections. By stacking multiple hourglass modules in an end-to-end fashion, the network performs repeated bottom-up and top-down inference, enabling iterative refinement of predictions. Intermediate supervision is applied at each stage to facilitate optimization. This architecture effectively models spatial relationships among body joints and has become a foundational benchmark for numerous subsequent studies in 2D human pose estimation. To address the challenge of scale variations across different human body parts, Yang et al. [25] proposed the Pyramid Residual Module (PRM). This module enables deep convolutional

 doi:10.20944/preprints202506.1093.v1

neural networks (DCNNs) to extract multi-scale features, thereby improving robustness to scale variations in human pose estimation.

Chu et al. [26] proposed an end-to-end human pose estimation framework that integrates multi-context attention mechanisms with enhanced residual units (HRUs). T The model generates attention maps enriched with multi-resolution and multi-level semantic representations using a stacked hourglass network. To refine the attention maps, a Conditional Random Field (CRF) is employed to capture spatial dependencies among adjacent keypoints. This work was among the first to introduce attention mechanisms into heatmap-based pose estimation, improving spatial attention precision and enhancing joint localization accuracy. This approach provides an efficient and accurate solution for human pose estimation tasks.

Xiao et al. [9] proposed a simple yet effective architecture that combines a ResNet backbone with sequential upsampling and convolutional layers. This design significantly reduces architectural complexity while maintaining strong performance, thereby facilitating easier analysis and comparison across methods. Despite its simplicity, the deconvolution-based head, coupled with the ResNet backbone, achieves performance comparable to that of more complex contemporary models. This work demonstrated that a straightforward architecture can be both effective and practical, thereby promoting accessibility and real-world applicability in human pose estimation research.

To address the limitations of conventional networks that downsample and then upsample feature representations, Wang et al. [10] proposed the High-Resolution Network (HRNet). Unlike traditional architectures, HRNet maintains high-resolution representations throughout the entire network by connecting multiple parallel convolutional streams operating at different resolutions. By repeatedly exchanging information across these multi-resolution branches, HRNet effectively enhances both spatial accuracy and semantic representation. This design significantly improves pose estimation performance, particularly in scenarios requiring fine-grained localization.

Zhang et al. [27] were the first to systematically investigate the substantial impact of heatmap coordinate decoding on pose estimation performance. Their work addressed a critical gap in the literature by identifying and analyzing design flaws in the coordinate encoding and decoding processes of heatmap-based methods, which had previously received little attention.

To address the quantization errors and information loss inherent in conventional decoding strategies, Zhang et al. proposed DARK (Distribution-Aware Coordinate Representation of Keypoints), a novel representation that models the underlying distribution of keypoint locations. DARK requires no modification to existing network architectures and can be seamlessly integrated into mainstream models such as HRNet and SimpleBaseline. Experimental results demonstrate that DARK significantly improves the average precision of keypoint detection.

To address data processing bias in human pose estimation, Huang et al. [28] proposed the Unbiased Data Processing (UDP) strategy. They systematically analyzed two major sources of error commonly observed in mainstream methods: misalignment during image flipping and statistical bias in the coordinate encoding–decoding process. UDP addresses these issues by replacing pixel-based measurements with unit-length scaling in continuous space, thereby enabling precise alignment of flipped predictions during inference. This enhances the generalization capability of pose estimation models. As a model-agnostic strategy, UDP can be readily integrated into a wide range of state-of-the-art pose estimation frameworks.

Li et al. [29] introduced a cascaded-Transformer architecture for regression-based human pose estimation. The model exploits the Transformer's encoder–decoder structure to unify person detection with keypoint localization. Unlike heatmap-based approaches, it removes complex post-processing and hand-crafted heuristics, enabling fully end-to-end training and inference. Moreover, leveraging the Transformer for keypoint prediction yields interpretable joint-association modeling, offering explicit insight into keypoint relationships.

To address the limitations of traditional heatmap-based methods in modeling global structural information, Wang et al. [30] proposed a novel two-stage human pose estimation framework, termed Graph-PCNN. The framework adopts a coarse-to-fine pipeline: the first stage performs coarse

 doi:10.20944/preprints202506.1093.v1

keypoint localization, followed by a refinement stage that improves prediction precision. A core component of the framework is the Graph Pose Refinement (GPR) module, which models structural relationships among keypoints via graph-based message passing and feature aggregation. By leveraging these dependencies, GPR effectively refines coarse predictions from the initial stage, resulting in significantly improved keypoint localization accuracy.

2.2. 2D Multi-Person Pose Estimation

Multi-person pose estimation extends the single-person setting but introduces substantially greater challenges. In addition to localizing the keypoints of each individual, the task must address keypoint grouping, inter-person occlusion, and body overlap—factors that significantly increase computational complexity and recognition difficulty.

Based on their processing pipelines, existing multi-person pose estimation methods are broadly classified into two paradigms: top-down and bottom-up approaches. Top-down approaches (e.g., Figure 4a) first detect all individuals in the image, followed by single-person pose estimation within each bounding box. These methods typically provide high accuracy; however, their computational cost increases linearly with the number of people, making them inefficient in crowded scenes. In contrast, bottom-up approaches (e.g., Figure 4b) bypass the person detection stage and detect all keypoints across the entire image in a single pass. The detected keypoints are subsequently grouped into individual poses using graph-based models or associative embedding techniques. Bottom-up methods offer greater parallelism and scalability, making them well-suited for large-scale scenarios. However, under heavy occlusion or in the presence of similar body postures, keypoint association becomes ambiguous and error-prone.

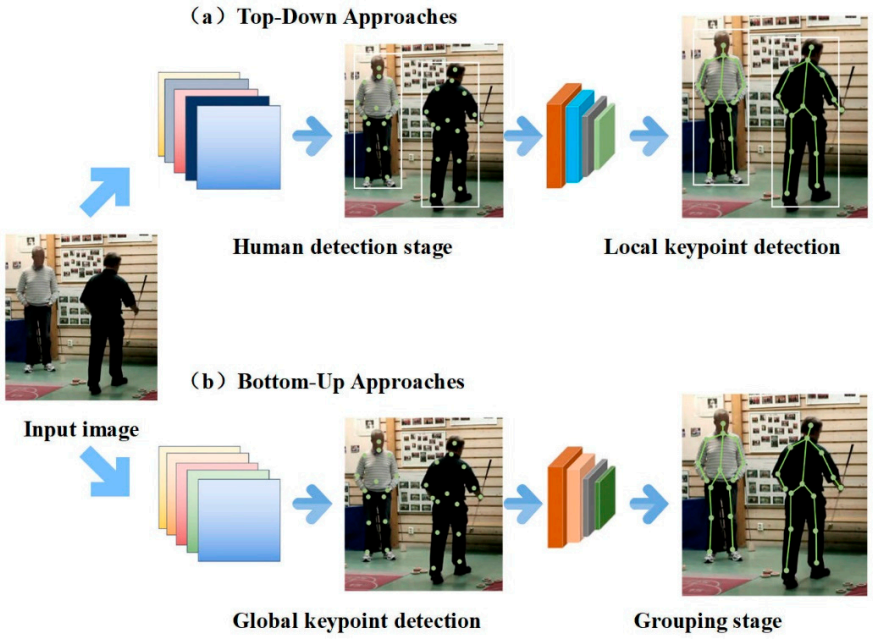

Figure 4. Approaches to 2D multi-person pose estimation: (a) top-down and (b) bottom-up.

2.2.1. Top-Down Approach of 2D Multi-Person Pose Estimation

He et al. [8] proposed Mask R-CNN, an extension of the Faster R-CNN framework developed for efficient object detection and instance segmentation. In addition to the classification and bounding box regression branches in Faster R-CNN, Mask R-CNN introduces a parallel branch that predicts a pixel-level segmentation mask for each Region of Interest (RoI). This design allows the model to perform accurate instance segmentation while preserving object classification and localization capabilities. A key innovation is the replacement of RoIPool with RoIAlign, which eliminates

 doi:10.20944/preprints202506.1093.v1

quantization errors and ensures precise alignment between extracted features and the original image at the pixel level. This substantially improves segmentation quality. Despite its additional functionality, Mask R-CNN remains computationally efficient, introducing only a marginal increase in inference time compared to Faster R-CNN. It is also straightforward to train and generalize. By integrating keypoint prediction into a general object detection framework, Mask R-CNN has become a standard top-down baseline for multi-person pose estimation.

Cao et al. [7] proposed an efficient multi-person 2D human pose estimation method based on Part Affinity Fields (PAFs). PAFs encode a set of 2D vector fields that capture the position and orientation of limbs by modeling pairwise associations between body keypoints. This representation allows the method to robustly handle complex scenarios such as occlusion, limb overlap, and inter-person interactions. After detecting all keypoints, a greedy graph-based parsing algorithm is applied to associate them based on PAFs, thereby assembling complete poses for multiple individuals. Unlike traditional top-down approaches that rely on person detection followed by single-person pose estimation, this method avoids errors caused by detection failures and decouples inference time from the number of people in the scene, leading to higher computational efficiency. The use of PAFs for keypoint grouping is a key innovation that laid the foundation for early integrations of top-down and bottom-up paradigms in multi-person pose estimation.

Kocabas et al. [31] introduced MultiPoseNet, a multi-task framework for efficient multi-person pose estimation. The central innovation of this framework is the Pose Residual Network (PRN), which is designed to accurately assign detected keypoints to their corresponding human instances. PRN utilizes a residual multilayer perceptron that jointly considers the configuration of all keypoints, enabling keypoint-to-person association to be performed in a single forward pass. This design significantly improves association accuracy, especially in challenging scenarios with substantial person overlap.

To address the heavy model dependency and architectural complexity often associated with human pose estimation, Moon et al. [21] introduced PoseFix, a model-agnostic pose refinement network. Unlike conventional end-to-end approaches that tightly couple pose estimation with specific architectures, PoseFix decouples the refinement stage and operates as an independent post-processing module. It can be flexibly integrated into various top-down frameworks to enhance keypoint localization accuracy, regardless of the underlying backbone architecture.

Khirodkar et al. [32] introduced the Multi-Instance Pose Network (MIPNet) for robust multi-person pose estimation. A core component of MIPNet is the Multi-Instance Modulation Block (MIMB), which adaptively modulates channel-wise features for individual instances. By enabling instance-specific feature transformation, MIMB enhances the network's ability to differentiate and predict multiple human poses simultaneously. This design significantly improves performance in crowded and occluded scenes, where conventional methods often fail.

Li et al. [14] introduced TokenPose, a novel human pose estimation method based on token representations. By introducing keypoint tokens, TokenPose effectively unifies the modeling of visual appearance and structural constraints within a Transformer-based framework. This approach offers a lightweight, efficient, and interpretable solution for pose estimation, and opens new directions for applying Transformer architectures to keypoint detection in computer vision.

To balance accuracy and computational efficiency, Yu et al. [16] introduced a high-resolution lightweight network specifically designed for human pose estimation. The proposed Lite-HRNet combines the multi-resolution and high-fidelity representation capabilities of HRNet with the lightweight design principles of ShuffleNet. This hybrid architecture maintains high-resolution feature representations while substantially reducing computational cost. Lite-HRNet contributes to the advancement of efficient human pose estimation models, particularly for deployment on resource-constrained devices.

Xu et al. [33] introduced ViTPose, a human pose estimation model built entirely on a pure Vision Transformer architecture. The model adopts a minimalist design that eliminates traditional convolutional components, demonstrating that top-down approaches based solely on Transformer

 doi:10.20944/preprints202506.1093.v1

architectures can achieve high-precision keypoint detection. ViTPose establishes a new performance benchmark for Transformer-based human pose estimation and offers valuable insights for future research in this domain.

2.2.2. Bottom-Up Approach of 2D Multi-Person Pose Estimation

Bottom-up methods originated from graphical model formulations, where spatial relationships between keypoints were optimized using graph-based structures. With the advancement of deep learning, these approaches have evolved into end-to-end frameworks centered on Part Affinity Fields (PAFs) and heatmaps for simultaneous keypoint detection and association. Since the introduction of OpenPose in 2017—the first framework to achieve efficient multi-person keypoint detection [7]—subsequent research has advanced along three primary dimensions: detection accuracy, inference speed, and system scalability.

Cao et al. [7] first introduced Part Affinity Fields (PAFs), which model the spatial connections between body parts as two-dimensional vector fields. This formulation established the core architecture of bottom-up pose estimation methods. PAFs effectively addressed the challenge of assembling individual keypoints into complete human skeletons, thereby enabling multi-person pose estimation without the need for explicit person detection. This pioneering work led to the development of the OpenPose framework, which has since become a widely adopted benchmark in the field. In follow-up work, Cao et al. [34] extended OpenPose to support additional skeletal models and enhanced the keypoint decoding strategy.

To address the keypoint association errors in crowded scenes observed in OpenPose, Papandreou et al. [35] introduced PersonLab. PersonLab is a box-free, bottom-up framework for multi-person pose estimation and instance segmentation. Built on a fully convolutional network, PersonLab performs keypoint detection and instance grouping simultaneously in a single forward pass. The framework introduces geometric embeddings for each keypoint, enabling efficient and accurate instance association without relying on bounding boxes. This approach improves the reliability of keypoint grouping, especially in densely populated and highly occluded scenes.

Li et al. [36] introduced a bounding-box-constrained strategy to enhance keypoint grouping accuracy in bottom-up pose estimation. The proposed method employs a multi-stage residual network that jointly predicts keypoint heatmaps and directional fields, enabling full-image pose estimation in a single forward pass. During the association stage, bounding boxes are used to constrain the connection range, reducing the likelihood of incorrect keypoint associations between individuals. In addition, non-maximum suppression and broken-link recovery techniques are incorporated to further improve overall accuracy and robustness.

Jin et al. [37] conducted the first systematic comparison of bottom-up and top-down approaches using the PoseTrack dataset. The study revealed that bottom-up methods exhibit superior generalization performance in crowded and occluded scenarios. Based on this observation, they proposed a hybrid strategy that combines the strengths of both paradigms to improve overall pose estimation accuracy.

Li et al. [38] revisited bottom-up multi-person pose estimation and proposed a more intuitive and efficient framework. The proposed method focuses on accurately detecting the keypoints of all individuals in an image, resulting in substantial performance improvements.

Geng et al. [39] argued that accurate keypoint regression requires learning feature representations that are specifically focused on keypoint regions. Based on this insight, they proposed a novel bottom-up framework named Decoupled Keypoint Regression (DEKR). DEKR employs adaptive convolutions and a multi-branch architecture to decouple the regression of individual keypoints, thereby improving localization precision.

Shi et al. [40] proposed PETR, the first end-to-end Transformer-based framework for multi-person human pose estimation. PETR formulates pose estimation as a hierarchical set prediction task, effectively eliminating the dependence on handcrafted components. The framework incorporates a pose decoder and a joint decoder, which are designed to model inter-person and inter-joint

 doi:10.20944/preprints202506.1093.v1

relationships, respectively. By leveraging attention mechanisms to adaptively focus on relevant features, PETR effectively mitigates the issue of keypoint feature misalignment. Benefiting from the strong modeling capacity of Transformers and an end-to-end design, PETR streamlines the multi-person pose estimation pipeline and provides valuable insights for future research.

3. Evaluation Metrics and Benchmark Datasets

To ensure fair comparisons across different models, it is essential to establish standardized evaluation metrics and utilize high-quality benchmark datasets. Evaluation metrics quantify the accuracy of keypoint predictions and serve as the foundation for assessing model performance. However, inconsistencies in evaluation protocols can lead to incomparable results, thereby hindering fair comparisons across studies. Dataset quality directly influences training effectiveness and the generalization capability of pose estimation models. Factors such as annotation precision, sample diversity, and scene complexity significantly impact model performance. Therefore, developing well-annotated datasets and unified evaluation frameworks is fundamental to advancing the field of 2D human pose estimation.

This chapter presents a systematic overview of widely used evaluation standards in 2D human pose estimation, along with a categorized analysis of key performance metrics. It also introduces widely adopted benchmark datasets, such as MPII, with a focus on their annotation protocols, dataset scale, pose categories, and application scenarios. By summarizing mainstream evaluation frameworks and data resources, this chapter aims to offer a clear reference for fair model comparisons, reproducible experiments, and methodological advancements in the field.

3.1. Evaluation Metrics

Compared to image classification, human pose estimation is a structured regression task. Its evaluation requires not only quantifying the accuracy of individual keypoints but also assessing the plausibility of the overall body configuration. A model may localize most keypoints near their ground-truth positions; however, if the predicted pose violates human kinematic constraints, it should not be regarded as a reliable solution. Therefore, evaluation standards in pose estimation serve not only as benchmarks for comparing model performance, but also as guidelines for algorithm refinement and optimization.

In the early stages of pose estimation research, simple error-based metrics, such as End-Point Error (EPE), were commonly used to evaluate model performance [41]. These metrics primarily relied on the Euclidean distance between predicted and ground-truth keypoints. However, as application scenarios became more complex, more expressive and robust evaluation metrics were gradually introduced. These advanced metrics consider multiple factors, including localization error, keypoint importance, object scale, and keypoint visibility. These metrics have since become widely adopted as standard benchmarks in the field. Tables 1 and 2 summarize and compare commonly used evaluation metrics, highlighting their key characteristics and applicability.

Table 1. Comparative analysis of evaluation metrics in 2D human pose estimation.

Metric Name	Full Name	Normalized	Structure Considered	Confidence Considered	Applicable Datasets
PCP[42]	Percentage of Correct Parts	√	√	×	LSP, LSP-Extended
PCK[43]	Percentage of Correct Keypoints	√	×	×	MPII, LSP, AI Challenger
PCKh[44]	PCK with Head-normalized	√	×	×	MPII
AUC[45]	Area Under Curve	√	×	×	AI Challenger
OKS[46]	Object Keypoint Similarity	√	√	√	COCO、PoseTrack

 doi:10.20944/preprints202506.1093.v1

AP@OKS[46]	Average Precision based on OKS	√	√	√	COCO, CrowdPose
AR@OKS[46]	Average Recall based on OKS	√	√	√	COCO, CrowdPose
mAP@OKS[46]	Mean Average Precision based on OKS	√	√	√	COCO, PoseTrack, OCHuman
IoU[47]	Intersection over Union	√	×	×	COCO

Table 2. Summary of evaluation metrics: computation principles, advantages, and limitations.

Metric Name	Computation Principle	Advantages	Limitations
PCP[42]	A body part is considered correct if both predicted endpoints fall within a tolerance distance of the ground truth endpoints.	Captures structural correctness; accounts for limb connectivity.	Sensitive to limb length; ineffective for small parts; rarely used in modern benchmarks.
PCK[43]	A keypoint prediction is correct if it lies within α × reference length of the ground truth point.	Simple and intuitive; widely adopted in early pose estimation research.	Threshold-dependent; not scale-invariant; ignores keypoint visibility.
PCKh[44]	Similar to PCK but uses head segment length as the normalization factor	Better normalization for human scale; standard for the MPII dataset.	Depends on accurate head annotation; not suitable for multi-person scenarios.
AUC[45]	Computes the area under the PCK curve across varying thresholds.	Aggregates performance across multiple thresholds; reduces threshold sensitivity.	Less interpretable; inherits limitations of PCK.
OKS[46]	Measures keypoint similarity using a Gaussian penalty based on Euclidean distance, normalized by object scale and keypoint-specific constants.	Scale-invariant; considers keypoint visibility and relative importance.	Sensitive to hyperparameters (e.g., σ); requires manual calibration; complex to compute.
AP@OKS[46]	Computes average precision across multiple OKS thresholds (0.50 to 0.95) using confidence-ranked predictions.	Official COCO benchmark; evaluates both localization and detection confidence.	High computational cost; strongly dependent on confidence ranking; penalizes slight deviations.
AR@OKS[46]	Computes average recall across a fixed OKS threshold under varying detection limits (e.g., maxDets=20).	Measures detection completeness; useful in dense or occluded scenes.	Does not reflect precision; prone to false positives; affected by maxDets setting.
mAP@OKS[46]	Mean of AP@OKS scores across 10 OKS thresholds (0.50–0.95, step size 0.05).	Standard benchmark for multi-person pose estimation; balances precision and recall across scales and occlusion.	Sensitive to confidence calibration; penalizes invisible or slightly off-keypoints; computationally intensive.
IoU[47]	Ratio of overlap area to union area between predicted and ground truth bounding boxes.	Intuitive geometric measure; standard in object detection.	Inapplicable to keypoint evaluation; insensitive to pose structure or semantics.

 doi:10.20944/preprints202506.1093.v1

3.2. Benchmark Datasets

3.2.1. Max Planck Institute for Informatics (MPII)

The MPII dataset [44] is a widely recognized benchmark in human pose estimation and is extensively used to evaluate the performance of state-of-the-art algorithms. It consists of approximately 25,000 images covering over 40,000 human instances, each annotated with precise keypoint locations. All images were systematically collected based on a predefined taxonomy of daily human activities, encompassing 410 distinct activity categories. Each image is labeled with an activity category and extracted from YouTube videos, with adjacent unlabeled frames included to support research on temporal modeling and contextual understanding. The test set provides richer annotations, including body part occlusion states and 3D orientation information for the torso and head, thereby enabling advanced tasks beyond 2D pose estimation. Figure 5 displays representative examples from the MPII Human Pose dataset.

Figure 5. Example images from the MPII Human Pose Dataset.

3.2.2. Common Objects in Context (COCO)

The COCO dataset [46], released by Microsoft Research in 2014, is a large-scale visual benchmark covering image recognition, instance segmentation, and image captioning. It provides annotations for 80 object categories—including the critical "person" class—and contains diverse visual content with fine-grained labels. To support the growing research on human pose estimation, COCO was extended in 2016 with the Keypoints subset, specifically curated for 2D keypoint detection. Since its release, this subset has become one of the most widely used benchmarks for evaluating pose-estimation algorithms. For the keypoint task, COCO supplies annotations solely for objects labeled as "person". Each labeled individual is annotated with 17 anatomical keypoints, including the nose, eyes, ears, shoulders, elbows, wrists, hips, knees, and ankles. Every keypoint is specified by 2D coordinates (x, y) and a visibility flag indicating whether it is visible or occluded. The dataset offers more than 200,000 fully annotated person instances spanning a broad spectrum of real-world scenes—such as street photography, sports, and indoor environments. Owing to its diversity and complexity, COCO has become one of the most influential and widely adopted benchmarks in 2D human pose estimation. Figure 6 displays representative examples from the COCO dataset..

 doi:10.20944/preprints202506.1093.v1

Figure 6. Example images from the COCO Dataset.

3.2.3. Leeds Sports Pose (LSP) and LSP Extended (LSPE)

The LSP dataset [48], released by the University of Leeds in 2010, is one of the earliest and most representative benchmark datasets in human pose estimation. It is primarily designed for single-person 2D pose estimation and features large motion variations, complex backgrounds, and diverse body configurations. The original dataset includes 2,000 images, equally split into 1,000 for training and 1,000 for testing. The images were primarily sourced from Flickr and depict a variety of sports activities—such as football, gymnastics, athletics, and skateboarding—spanning a wide range of dynamic and challenging poses. Each image is annotated with 14 keypoints corresponding to the major joints of the human body. Notably, left and right joints are consistently labeled from a person-centric perspective. A subsequent extension of the dataset introduced 10,000 additional annotated training images, further enriching the dataset and enabling more robust model development. The LSP dataset has played a pivotal role in advancing early pose estimation algorithms, serving as a foundational benchmark for evaluating models under highly variable and unconstrained conditions. Figure 7 displays representative examples from the LSP dataset.

Figure 7. Example images from the LSP Dataset.

3.2.4. Occluded Human Dataset (OCHuman)

The OCHuman dataset [49] is a large-scale image benchmark specifically designed to address the challenges of human pose estimation under severe occlusion. Released by MEGVII Research in 2018, the dataset was constructed by filtering human instances from the COCO dataset with an occlusion ratio exceeding 60%. OCHuman comprises approximately 6,700 images and over 11,000 human instances with significant occlusion. The dataset is divided into training, validation, and test sets, with 1,000 images each in the validation and test sets. Each instance is annotated with bounding boxes, human keypoints, and segmentation masks, offering comprehensive supervision for a variety of vision tasks. Notably, each keypoint is labeled with a visibility flag indicating whether it is visible or occluded. Owing to its emphasis on severe occlusion and fine-grained annotations, OCHuman is regarded as one of the most challenging and complex benchmarks in human-centric vision research. Figure 8 displays representative examples from the OCHuman dataset.

 doi:10.20944/preprints202506.1093.v1

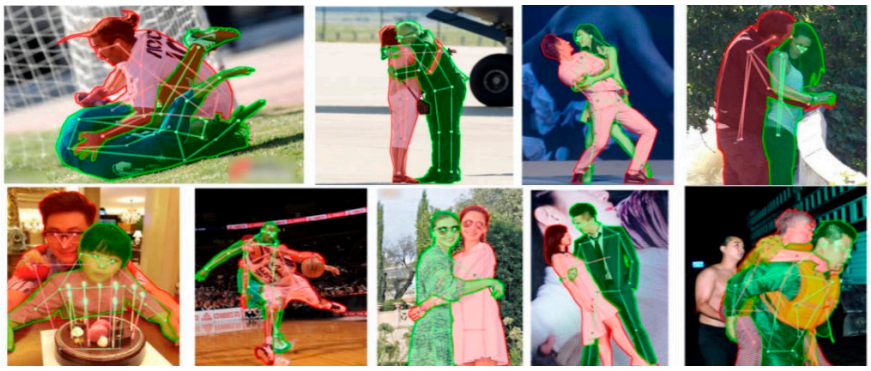

Figure 8. Example images from the OCHuman Dataset.

3.2.5. CrowdPose Dataset

The CrowdPose dataset [50] is a benchmark specifically designed to evaluate the performance of human pose estimation algorithms in crowded multi-person scenarios. Released by Megvii Research in 2019, it comprises approximately 20,000 images and over 32,000 human instances annotated with keypoints. Each person is annotated with 17 keypoints in accordance with the COCO annotation format, covering major joints from head to foot. To assess algorithm robustness under varying levels of occlusion and crowd density, the dataset introduces a novel evaluation metric called the Crowd Index. The Crowd Index quantifies crowding and occlusion levels for each image and is used to divide the dataset into three subsets: Easy, Medium, and Hard. Each subset contains an equal number of images, ensuring balanced and fair evaluation across different levels of scene complexity. Figure 9 displays representative examples from the CrowdPose dataset.

Figure 9. Example images from the CrowdPose Dataset.

4. Top-Performing Models on Benchmark Datasets

The preceding chapter surveyed the leading benchmark datasets for 2D human pose estimation together with their evaluation protocols. In this chapter, we analyze the top-performing models on these benchmarks, with emphasis on three core aspects: network architecture, training strategy, and key innovations. These models define the current research frontier and embody the latest design trends and technical advances in the field. For each dataset, we detail the top-ranked model and summarize the next five leading entries in a concise table. This comparative overview gives readers a comprehensive view of the most influential 2-D pose-estimation methods and provides deeper insight into prevailing technical paradigms. To ensure a fair comparison, all discussed models were trained exclusively on the official training splits of their respective benchmarks, with no external pose-estimation data. This constraint removes confounding effects caused by training-data variability and enables a clearer assessment of each model's architectural merits.

 doi:10.20944/preprints202506.1093.v1

4.1. Top-Performing Models on MPII Dataset

On the MPII dataset, the Pose Contrastive Transformer (PCT) achieved a state-of-the-art accuracy of 94.3% under the PCKh@0.5 metric, representing the highest performance reported on this benchmark to date [51]. PCT employs a two-stage network architecture. In the encoding stage, a learnable encoder transforms the input pose into a set of tokens, each encoding a semantically meaningful substructure of the human body. These tokens are discretized using a shared codebook, generating class indices corresponding to distinct structural components. In the prediction stage, a classifier predicts the class index of each token based on features extracted from the input image. These predicted indices are subsequently decoded by a pretrained decoder to reconstruct the final human keypoint coordinates.

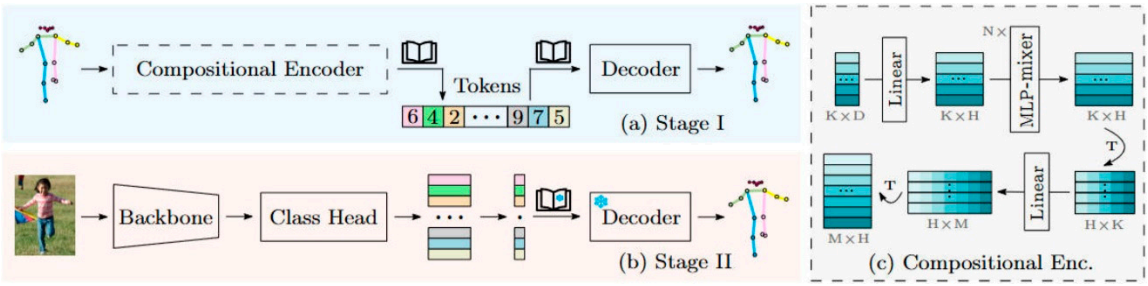

Figure 10. The two-stage pipeline of PCT: Stage I discretizes poses into token indices via a compositional encoder and codebook; Stage II classifies these tokens and decodes them to recover the final pose.

The training process similarly comprises two stages. In the first stage, the encoder and codebook are jointly optimized to minimize the reconstruction error of the quantized pose representation, ensuring that each token effectively captures the structural characteristics of its corresponding body part. In the second stage, the classifier is trained to predict token class indices directly from the input image. A pretrained decoder then translates the predicted indices into final keypoint coordinates. This two-stage training paradigm enables the model to learn structured and semantically coherent pose representations, thereby enhancing accuracy and robustness, particularly in complex scenes with severe occlusions.

The core innovation of PCT lies in decomposing the holistic human pose into a set of structured and interpretable subcomponents, each encoded as a discrete token. Unlike conventional methods that predict each joint independently, PCT explicitly models inter-joint dependencies through a compositional representation. The use of a shared codebook enables a compact and discrete representation of human pose, reducing representational complexity while preserving rich structural semantics. This design enhances the model's capacity to reason about complex joint relationships and improves robustness under occlusion. Furthermore, the encoder, codebook, and decoder are trained in an end-to-end manner, enabling unified learning of structured pose representations.

Table 3. Comparative performance of top-ranked models (2nd to 6th) on the MPII dataset.

Model	Year	Backbone	PCKh@0.5	Input size	Characteristic
PCT[51]	2023	swin-base	93.8%	256 × 256	PCT enhances dependency modeling and occlusion-robust inference by decomposing human pose into structured, discrete symbolic subcomponents.
4xRSN-50[52]	2020	ResNet-50	93.0%	256×192	4×RSN-50 enhances keypoint localization accuracy and efficiency by stacking Residual Steps Blocks (RSBs) to integrate fine-grained local features.

 doi:10.20944/preprints202506.1093.v1

UniPose[53]	2020	ResNet-101	92.7%	1280×720	By integrating the ResNet architecture with the WASP multi-scale module, the model achieves efficient and accurate single-stage pose estimation and bounding box detection.
MSPN[54]	2019	ResNet-50	92.6%	384×288	MSPN improves both accuracy and efficiency by optimizing single-stage modules, feature aggregation mechanisms, and supervision strategies.
Spatial Context[55]	2019	ResNet-50	92.5%	256×256	The Spatial Context model integrates multi-stage prediction with joint graph structures to accurately capture spatial relationships between human joints.

4.2. Top-Performing Models on COCO Test-Dev Dataset

On the COCO test-dev dataset, ViTPose (ensemble) achieved an average precision (AP) of 81.1, setting a new state-of-the-art on this benchmark [33]. ViTPose employs a pure, non-hierarchical Vision Transformer (ViT) as its backbone to extract high-dimensional feature representations from input images. Specifically, the input image is divided into fixed-size patches and processed by a standard ViT encoder for feature extraction. The backbone adopts a unified and streamlined design that eliminates convolutional modules and complex components, resulting in high scalability and architectural simplicity. In the decoding stage, ViTPose utilizes a lightweight head to transform Transformer-derived features into keypoint heatmaps. This decoder is both structurally simple and computationally efficient, allowing compatibility with inputs of varying resolutions. This flexibility enables multi-task joint training and inference, improving both the practicality and scalability of the system.

ViTPose's training strategy fully exploits the transferability of large-scale pretrained models. The Vision Transformer (ViT) backbone is initially pretrained on general-purpose datasets (e.g., ImageNet) and subsequently fine-tuned on task-specific datasets such as MS COCO, resulting in substantial improvements in keypoint detection accuracy. ViTPose also supports flexible input resolutions, where higher-resolution inputs generally lead to better performance. To reduce training costs, the framework allows partial freezing of Transformer layers, fine-tuning only essential components. This enables efficient adaptation to resource-constrained environments without significant degradation in performance.

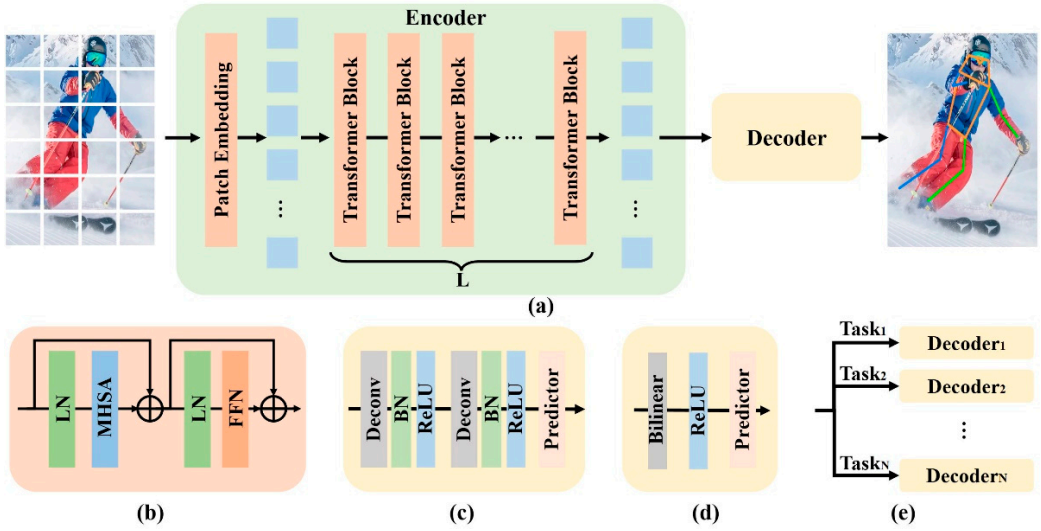

Figure 11. (a) The framework of ViTPose. (b) The transformer block. (c) The classic decoder. (d) The simple decoder. (e) The decoders for multiple datasets.

 doi:10.20944/preprints202506.1093.v1

The core innovation of ViTPose lies in its structurally minimalist yet highly effective design. By employing a pure Transformer-based backbone, ViTPose achieves high-precision pose estimation without relying on convolutional layers or complex architectural modules. The framework is inherently scalable, allowing flexible adjustments in network depth and width to accommodate diverse task requirements. ViTPose further improves generalization through joint training across multiple tasks and datasets. A particularly notable contribution is the introduction of a learnable knowledge token, which enables efficient knowledge transfer from large-scale to lightweight models. This significantly enhances the performance of compact models in real-world deployment scenarios.

Table 4. Comparative performance of top-ranked models (2nd to 6th) on the COCO test-dev dataset.

Model	Year	Backbone	AP	Input size	Characteristic
ViTPose[33]	2022	ViTAE-G	80.9	256 × 192	ViTPose is built on a streamlined ViT architecture, offering a compelling combination of high performance, scalability, and strong transferability.
UDP-Pose-PSA [56]	2021	HRNet-W48	79.5	384 × 288	UDP-Pose-PSA integrates HRNet-W48 with polarized self-attention to enhance long-range dependency modeling, thereby improving the accuracy and fine-grained performance of keypoint detection.
4xRSN-50 (ensemble)[52]	2020	ResNet-50	79.2	256×192	4×RSN-50 (ensemble) integrates four multi-branch Residual Steps Networks to effectively fuse intra-layer features.
CCM+ [57]	2020	HRNet-w48	78.9	384×288	CCM+ significantly improves keypoint detection accuracy through cascaded contextual fusion, sub-pixel localization, and an efficient training strategy.
UDP-Pose-PSA [56]	2021	HRNet-W48	78.9	256x192	UDP-Pose-PSA achieves an AP of 78.9 at an input resolution of 256×192, which is slightly lower than the 79.5 obtained at 384×288.

4.3. Top-Performing Models on LSP Dataset

On the LSP dataset, OmniPose achieved a PCK accuracy of 99.5%, approaching the upper bound of performance on this single-person pose benchmark [58]. OmniPose is based on a single-stage, end-to-end trainable framework that integrates an optimized HRNet backbone with the novel WASPv2 module. The input image is first processed by an enhanced HRNet, which extracts high-resolution, multi-scale feature representations. These features are then passed to the WASPv2 module for multi-scale fusion and decoding, generating Gaussian heatmaps for each joint to enable precise localization. To enhance localization accuracy, the HRNet backbone is refined by replacing conventional upsampling with a Gaussian-modulated deconvolution operation. This modification reduces pixel-level quantization errors and improves the precision of local keypoint regression.

OmniPose adopts a unified end-to-end training framework that eliminates the need for complex post-processing steps. The model is supervised with a Gaussian heatmap regression loss that facilitates accurate keypoint localization. To improve robustness and generalization, data augmentation techniques such as rotation, scaling, and flipping are applied during training. Additionally, the network supports multi-resolution training, allowing a flexible balance between accuracy and inference efficiency.

 doi:10.20944/preprints202506.1093.v1

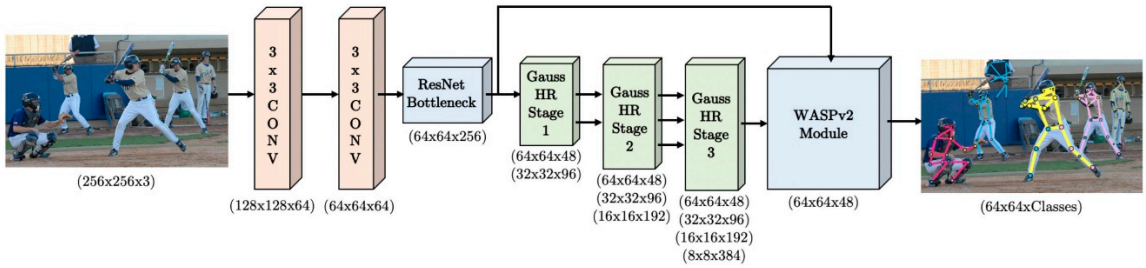

Figure 12. OmniPose Framework for Multi-Person Pose Estimation. The input image is processed by an enhanced HRNet and the WASPv2 module to generate heatmaps for each joint.

The core innovations of OmniPose reside in the design of the WASPv2 module and the Gaussian-modulated deconvolution mechanism. WASPv2 utilizes a hierarchical stack of dilated convolutions to significantly expand the receptive field and improve contextual feature representation. It directly generates keypoint heatmaps, eliminating the need for a separate decoding module and enhancing overall computational efficiency. Simultaneously, the Gaussian-modulated deconvolution replaces conventional upsampling, effectively reducing pixel quantization errors and enhancing keypoint localization accuracy. In parallel, the optimized HRNet backbone maintains high-resolution feature propagation throughout the network, reinforcing multi-scale feature preservation and fusion.

Table 5. Comparative performance of top-ranked models (2nd to 6th) on the LSP dataset.

Model	Year	Backbone	PCK	Input size	Characteristic
Soft-gated Skip Connections [59]	2020	HourGlassand U-Net	94.8%	256×256	A channel scaling factor is introduced to optimize feature fusion, enhancing both model accuracy and computational efficiency.
UniPose [60]	2020	ResNet-101	94.5%	1280×720	UniPose integrates a ResNet backbone with the WASP module to enable efficient end-to-end pose estimation and bounding box detection.
Residual Hourglass + ASR + AHO [61]	2018	Stacked Hourglass	94.5%	256×256	Residual Hourglass combines residual structures with adversarial enhancement to improve multi-scale feature modeling and the accuracy of keypoint localization.
SAHPE-Network [62]	2017	Stacked Hourglass Network	94%	256×256	This approach introduces a discriminator to learn structural constraints, thereby enhancing the accuracy and plausibility of pose estimation.
PRMs [25]	2017	stacked Hourglass network	93.9%	256×256	PRMs employs multi-branch convolutions to extract multi-scale features, enhancing the adaptability and accuracy of pose estimation under varying scales.

4.4. Top-Performing Models on OCHuman Dataset

On the OCHuman dataset, UniHCP achieved an average precision (AP) of 87.4, marking the current state-of-the-art performance on this benchmark [62]. UniHCP comprises three key components: a task-agnostic Transformer encoder, a task-adaptive decoder, and a task-guided interpreter. The encoder contains 12 Transformer layers with a hidden size of 768 and approximately 91.1 million parameters. It extracts general-purpose visual features from the input image. The decoder consists of 9 layers with a hidden size of 256 and approximately 14.5 million parameters, and is designed to retrieve task-specific features via learned queries. The task-guided interpreter, comprising approximately 3.5 million parameters, translates the query outputs into task-specific

 doi:10.20944/preprints202506.1093.v1

representations. This unified design allows for up to 99.97% parameter sharing across tasks, significantly improving knowledge transfer and enhancing generalization capabilities.

UniHCP employs a large-scale joint pretraining strategy across multiple tasks. It utilizes 33 human-centric datasets that cover five key areas: pose estimation, semantic part segmentation, pedestrian detection, person re-identification, and human attribute recognition. Through the use of unified task-specific queries and a shared interpreter, the model can simultaneously perform multiple tasks within a unified framework.

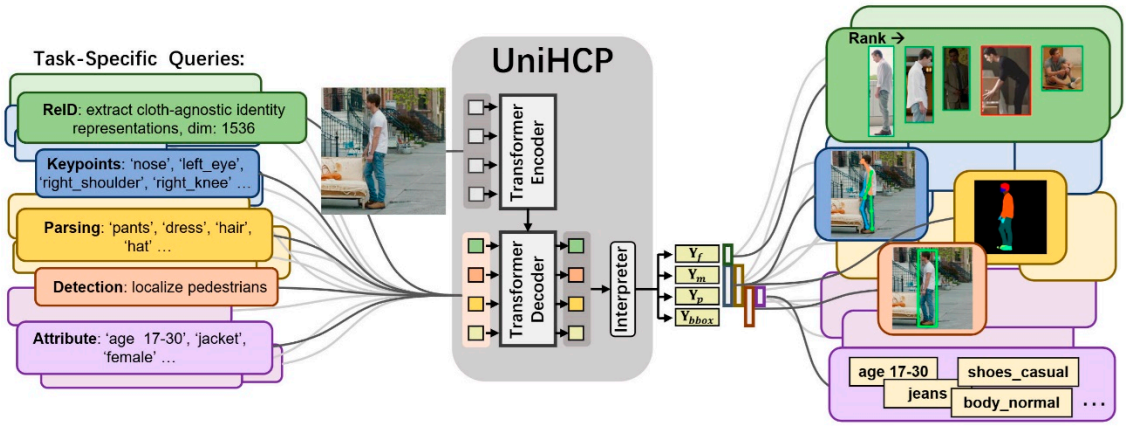

Figure 13. UniHCP handles diverse human-centric tasks uniformly by employing task-specific queries and a task-guided interpreter. All task predictions are generated in parallel through a unified encoder–decoder transformer framework.

UniHCP is the first model to introduce a unified architecture that efficiently integrates and handles multiple representative human-centric perception tasks. Its query-driven output mechanism replaces conventional task-specific heads, significantly reducing task-specific parameters and enhancing parameter sharing and feature reuse. Furthermore, the model leverages large-scale multi-task joint pretraining to enable extensive cross-task knowledge transfer and feature sharing. Owing to its efficient weight-sharing strategy, UniHCP achieves high parameter and data efficiency, thereby improving its transferability and adaptability to a wide range of human-centric tasks.

Table 6. Comparative performance of top-ranked models (2nd to 6th) on the OCHuman dataset.

Model	Year	Backbone	Test AP	Input size	Characteristic
BUCTD [63]	2023	CID-W32	47.2	256×192	BUCTD leverages bottom-up pose estimation as a conditional input to enhance accuracy and robustness in crowded and occluded scenarios.
HQNet [64]	2024	ViT-L	45.6	512×512	HQNet enables unified modeling of multi-task human understanding by sharing a Transformer decoder and a human query mechanism.
CID [32]	2022	HRNet-W48	45.0	512x512	CID introduces a context-instance decoupling mechanism that employs spatial and channel attention to extract instance-aware features for keypoint estimation.
MIPNet [62]	2021	HRNet-W48	42.5	256×256	MIPNet introduces a Multi-Instance Modulation Block (MIMB) that enables efficient prediction of multiple pose instances within a single detection box.
HQNet [64]	2024	ResNet-50	40.0	512×512	Built on a ResNet-50 backbone, the model offers strong spatial locality modeling, high

 doi:10.20944/preprints202506.1093.v1

	computational efficiency, and a lightweight structure.

4.5. Top-Performing Models on CrowdPose Dataset

On the CrowdPose dataset, BUCTD achieved an AP score of 78.5, marking the highest reported performance on this benchmark to date [63]. BUCTD (Bottom-Up Conditioned Top-Down) combines the strengths of bottom-up (BU) and top-down (TD) pose estimation paradigms to enhance overall accuracy and robustness. The framework operates in two main stages. In the first stage, a bottom-up model performs initial person detection and pose estimation. The resulting bottom-up predictions are then used as conditional inputs for a top-down refinement network in the second stage. To improve refinement, the top-down network incorporates a Conditional Attention Module (CoAM) that effectively fuses conditional pose information with image features. This design enhances the model's capability to identify and recover occluded or overlapping keypoints, especially in crowded or heavily occluded scenes.

BUCTD adopts a staged training pipeline. The bottom-up (BU) model is initially pre-trained to produce reliable pose predictions, which serve as conditional inputs for training the top-down refinement network. To improve the model's tolerance to prediction noise and enhance robustness, two conditional input sampling strategies are employed during training. The first strategy involves empirical sampling based on actual outputs from the BU model. The second draws inspiration from PoseFix [21] and introduces synthetic perturbations to mimic common prediction errors. This hybrid sampling strategy not only diversifies the training data but also substantially enhances the model's generalization under real-world uncertainty.

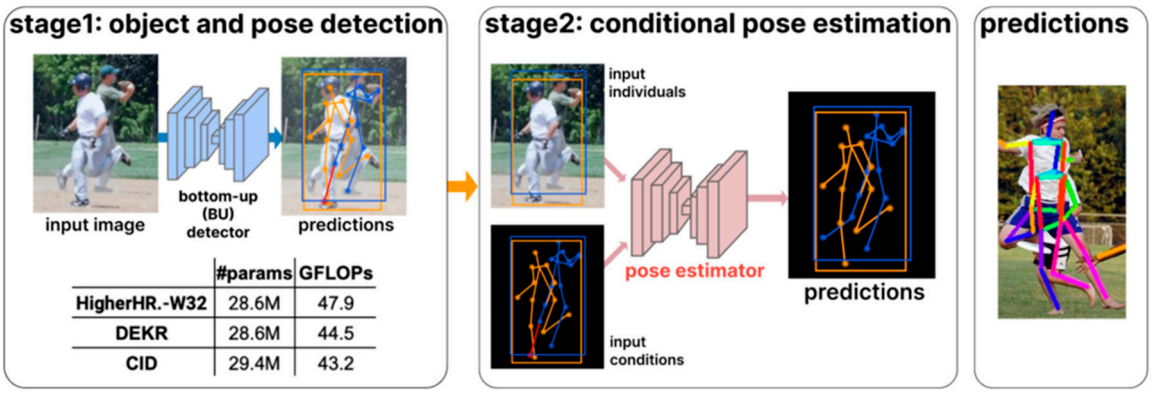

Figure 14. Overview of the BUCTD framework. A hybrid approach that incorporates bottom-up pose estimation as conditional input to enhance multi-person pose estimation under crowded and occluded scenarios.

This design is particularly effective in crowded and complex environments. The core innovation of BUCTD lies in leveraging bottom-up pose predictions as explicit guidance to inform the top-down pose estimation process. The Conditional Attention Module (CoAM) plays a central role by fusing conditional pose cues with image features, thereby significantly improving the model's capability to recover occluded keypoints. Furthermore, the integration of empirical and synthetic sampling strategies enables the model to effectively handle real-world prediction noise, while maintaining stable and accurate performance across diverse scenarios.

Table 7. Comparative performance of top-ranked models (2nd to 6th) on the CrowdPose dataset.

Model	Year	Backbone	AP	Input size	Characteristic
ViTPose [33]	2022	ViTAE-G	78.3	256×192	ViTPose is built on a streamlined ViT architecture, offering a compelling combination

 doi:10.20944/preprints202506.1093.v1

					of high performance, scalability, and strong transferability.
BUCTD [63]	2023	HRNet-W48	76.7	384×288	The BUCTD model integrates conditional inputs inspired by the PETR framework.
SwinV2-L 1K-MIM [66]	2022	Swin Transformer	75.5	384×384	By employing self-supervised Masked Image Modeling (MIM) pretraining on the ImageNet-1K dataset, the model is able to learn rich and generalizable representations.
SwinV2-B 1K-MIM [66]	2022	Swin Transformer	74.9	224×224	Self-supervised pretraining with Masked Image Modeling (MIM) significantly enhances the model's performance on geometric and motion-related tasks.
BUCTD [63]	2023	ResNet-50	72.9	384×288	The BUCTD model uses the CoAM-W48 configuration without employing any sampling strategy.

5. Conclusions and Future Perspectives

This review provides a systematic summary of recent advances in deep learning-based 2D human pose estimation (2D HPE). It covers fundamental methodologies, representative models, evaluation metrics, and widely adopted benchmark datasets. We categorize and analyze approaches for both single-person and multi-person pose estimation. Special attention is given to the evolution of methodologies—from early handcrafted feature-based techniques to deep neural network architectures, and more recently, to attention-based mechanisms and Transformer frameworks. Heatmap-based methods remain the dominant paradigm because of their high spatial localization accuracy. Meanwhile, regression-based and hybrid approaches are continuously evolving, especially in terms of lightweight design and end-to-end training efficiency..

In multi-person scenarios, top-down and bottom-up paradigms each offer distinct advantages. Top-down methods perform well in relatively clean scenes. In contrast, bottom-up approaches offer greater scalability and robustness in crowded or occluded environments. Hybrid strategies that integrate both paradigms—such as BUCTD [63] and PoseFix [21]—have demonstrated promising potential. These approaches aim to balance accuracy and computational efficiency in complex scenarios. Furthermore, the emergence of lightweight architectures and Transformer-based models has created new opportunities. These advances support real-time applications and enable efficient deployment on edge devices.

Despite substantial progress in 2D human pose estimation, several critical challenges persist. First, existing models exhibit limited robustness in crowded or heavily occluded scenes. This limitation hinders accurate keypoint detection and association across multiple individuals. Second, most methods rely heavily on curated and standardized datasets. As a result, their ability to generalize to complex, unconstrained real-world environments is limited. Third, although Transformer-based architectures have achieved impressive performance, their high computational cost remains a barrier to real-time inference and deployment on edge devices. Lastly, the lack of large-scale, high-quality annotated pose datasets continues to be a major bottleneck to further advancement in this field. To overcome these challenges, future research may focus on several promising directions.

1. Develop unified end-to-end frameworks that integrate object detection, keypoint localization, and post-processing. These frameworks should emphasize the use of Transformer architectures to capture long-range spatial and semantic dependencies;
2. Improve the robustness of keypoint prediction under occlusion. This can be achieved by incorporating generative models and self-supervised learning techniques;
3. Design compact models that are efficient during inference. These models should be suitable for deployment on resource-constrained platforms such as mobile phones and wearable devices;

 doi:10.20944/preprints202506.1093.v1

4. Enhance model adaptability across different datasets, environments, and population groups. This can be achieved by leveraging cross-domain generalization techniques such as domain adaptation and transfer learning;
5. Build more diverse and densely annotated datasets. Additionally, refine evaluation protocols to enable the extension of pose estimation to high-level tasks, such as action recognition and semantic behavior analysis;

In summary, deep learning-based 2D human pose estimation has developed into a rapidly evolving and interdisciplinary research domain. This review serves as a comprehensive reference for researchers and practitioners. It systematically categorizes existing approaches, compares representative models, and analyzes widely adopted evaluation metrics and benchmark datasets. It helps newcomers gain a quick understanding of the current state of the field. Meanwhile, it provides experienced scholars with structured insights into emerging trends and potential research directions.

Author Contributions:. Conceptualization, Z.Z. and S.-Y.S.; methodology, Z.Z. and S.-Y.S.; writing—original draft preparation, Z.Z. and S.-Y.S.; supervision, S.-Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: Please add: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

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

References

1. Sun, R., Lin, Z., Leng, S., Wang, A., & Zhao, L. (2025). An In-Depth Analysis of 2D and 3D Pose Estimation Techniques in Deep Learning: Methodologies and Advances. Electronics, 14(7), 1307.
2. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.
3. Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.
4. Tian, L., Wang, P., Liang, G., & Shen, C. (2021). An adversarial human pose estimation network injected with graph structure. Pattern Recognition, 115, 107863.
5. Felzenszwalb, P. F., & Huttenlocher, D. P. (2005). Pictorial structures for object recognition. International journal of computer vision, 61, 55-79.
6. Felzenszwalb, P., McAllester, D., & Ramanan, D. (2008, June). A discriminatively trained, multiscale, deformable part model. In 2008 IEEE conference on computer vision and pattern recognition (pp. 1-8). Ieee.
7. Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In Pro
8. ceedings of the IEEE conference on computer vision and pattern recognition (pp. 7291-7299).
9. He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).
10. Xiao, B., Wu, H., & Wei, Y. (2018). Simple baselines for human pose estimation and tracking. In Proceedings of the European conference on computer vision (ECCV) (pp. 466-481).
11. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., ... & Xiao, B. (2020). Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 43(10), 3349-3364.
12. Newell, A., Huang, Z., & Deng, J. (2017). Associative embedding: End-to-end learning for joint detection and grouping. Advances in neural information processing systems, 30.

 doi:10.20944/preprints202506.1093.v1

13. Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T. S., & Zhang, L. (2020). Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5386-5395).
14. Ma, X., Su, J., Wang, C., Ci, H., & Wang, Y. (2021). Context modeling in 3d human pose estimation: A unified perspective. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 6238-6247).
15. Li, Y., Zhang, S., Wang, Z., Yang, S., Yang, W., Xia, S. T., & Zhou, E. (2021). Tokenpose: Learning keypoint tokens for human pose estimation. In Proceedings of the IEEE/CVF International conference on computer vision (pp. 11313-11322).
16. Yuan, Y., Fu, R., Huang, L., Lin, W., Zhang, C., Chen, X., & Wang, J. (2021). Hrformer: High-resolution transformer for dense prediction. arXiv preprint arXiv:2110.09408.
17. Yu, C., Xiao, B., Gao, C., Yuan, L., Zhang, L., Sang, N., & Wang, J. (2021). Lite-hrnet: A lightweight high-resolution network. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10440-10450).
18. Hou, T., Ahmadyan, A., Zhang, L., Wei, J., & Grundmann, M. (2020). MobilePose: Real-time pose estimation for unseen objects with weak shape supervision. arXiv preprint arXiv:2003.03522.
19. Toshev, A., & Szegedy, C. (2014). Deeppose: Human pose estimation via deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1653-1660).
20. Carreira, J., Agrawal, P., Fragkiadaki, K., & Malik, J. (2016). Human pose estimation with iterative error feedback. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4733-4742).
21. Sun, X., Xiao, B., Wei, F., Liang, S., & Wei, Y. (2018). Integral human pose regression. In Proceedings of the European conference on computer vision (ECCV) (pp. 529-545).
22. Moon, G., Chang, J. Y., & Lee, K. M. (2019). Posefix: Model-agnostic general human pose refinement network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7773-7781).
23. Tompson, J. J., Jain, A., LeCun, Y., & Bregler, C. (2014). Joint training of a convolutional network and a graphical model for human pose estimation. Advances in neural information processing systems, 27.
24. Wei, S. E., Ramakrishna, V., Kanade, T., & Sheikh, Y. (2016). Convolutional pose machines. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 4724-4732).
25. Newell, A., Yang, K., & Deng, J. (2016). Stacked hourglass networks for human pose estimation. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14 (pp. 483-499). Springer International Publishing.
26. Yang, W., Li, S., Ouyang, W., Li, H., & Wang, X. (2017). Learning feature pyramids for human pose estimation. In proceedings of the IEEE international conference on computer vision (pp. 1281-1290).
27. Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A. L., & Wang, X. (2017). Multi-context attention for human pose estimation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1831-1840).
28. Zhang, F., Zhu, X., Dai, H., Ye, M., & Zhu, C. (2020). Distribution-aware coordinate representation for human pose estimation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 7093-7102).
29. Huang, J., Zhu, Z., Guo, F., & Huang, G. (2020). The devil is in the details: Delving into unbiased data processing for human pose estimation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5700-5709).
30. Li, K., Wang, S., Zhang, X., Xu, Y., Xu, W., & Tu, Z. (2021). Pose recognition with cascade transformers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 1944-1953).
31. Wang, J., Long, X., Gao, Y., Ding, E., & Wen, S. (2020). Graph-pcnn: Two stage human pose estimation with graph pose refinement. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16 (pp. 492-508). Springer International Publishing.
32. Kocabas, M., Karagoz, S., & Akbas, E. (2018). Multiposenet: Fast multi-person pose estimation using pose residual network. In Proceedings of the European conference on computer vision (ECCV) (pp. 417-433).

 doi:10.20944/preprints202506.1093.v1

33. Khirodkar, R., Chari, V., Agrawal, A., & Tyagi, A. (2021). Multi-instance pose networks: Rethinking top-down pose estimation. In Proceedings of the IEEE/CVF International conference on computer vision (pp. 3122-3131).
34. Xu, Y., Zhang, J., Zhang, Q., & Tao, D. (2022). Vitpose: Simple vision transformer baselines for human pose estimation. Advances in neural information processing systems, 35, 38571-38584.
35. Cao, Z., Hidalgo, G., Simon, T., Wei, S. E., & Sheikh, Y. (2019). Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE transactions on pattern analysis and machine intelligence, 43(1), 172-186.
36. Papandreou, G., Zhu, T., Chen, L. C., Gidaris, S., Tompson, J., & Murphy, K. (2018). 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) (pp. 269-286).
37. Li, M., Zhou, Z., Li, J., & Liu, X. (2018, August). Bottom-up pose estimation of multiple person with bounding box constraint. In 2018 24th international conference on pattern recognition (ICPR) (pp. 115-120). IEEE.
38. Jin, S., Ma, X., Han, Z., Wu, Y., Yang, W., Liu, W., ... & Ouyang, W. (2017, October). Towards multi-person pose tracking: Bottom-up and top-down methods. In ICCV posetrack workshop (Vol. 2, No. 3, p. 7).
39. Li, J., Su, W., & Wang, Z. (2020, April). Simple pose: Rethinking and improving a bottom-up approach for multi-person pose estimation. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 07, pp. 11354-11361).
40. Geng, Z., Sun, K., Xiao, B., Zhang, Z., & Wang, J. (2021). Bottom-up human pose estimation via disentangled keypoint regression. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 14676-14686).
41. Shi, D., Wei, X., Li, L., Ren, Y., & Tan, W. (2022). End-to-end multi-person pose estimation with transformers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11069-11078).
42. Barron, J. L., Fleet, D. J., & Beauchemin, S. S. (1994). Performance of optical flow techniques. International journal of computer vision, 12, 43-77.
43. Ferrari, V., Marin-Jimenez, M., & Zisserman, A. (2008, June). Progressive search space reduction for human pose estimation. In 2008 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-8). IEEE.
44. Yang, Y., & Ramanan, D. (2012). Articulated human detection with flexible mixtures of parts. IEEE transactions on pattern analysis and machine intelligence, 35(12), 2878-2890. Yang, Y., & Ramanan, D. (2012). Articulated human detection with flexible mixtures of parts. IEEE transactions on pattern analysis and machine intelligence, 35(12), 2878-2890.
45. Andriluka, M., Pishchulin, L., Gehler, P., & Schiele, B. (2014). 2d human pose estimation: New benchmark and state of the art analysis. In Proceedings of the IEEE Conference on computer Vision and Pattern Recognition (pp. 3686-3693).
46. Sun, X., Shang, J., Liang, S., & Wei, Y. (2017). Compositional human pose regression. In Proceedings of the IEEE international conference on computer vision (pp. 2602-2611).
47. Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13 (pp. 740-755). Springer International Publishing.
48. Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. International journal of computer vision, 88, 303-338.
49. Johnson, S., & Everingham, M. (2010, August). Clustered pose and nonlinear appearance models for human pose estimation. In bmvc (Vol. 2, No. 4, p. 5).
50. Zhang, S. H., Li, R., Dong, X., Rosin, P., Cai, Z., Han, X., ... & Hu, S. M. (2019). Pose2seg: Detection free human instance segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 889-898).
51. Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H. S., & Lu, C. (2019). Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10863-10872).

 doi:10.20944/preprints202506.1093.v1

52. Geng, Z., Wang, C., Wei, Y., Liu, Z., Li, H., & Hu, H. (2023). Human pose as compositional tokens. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 660-671).
53. Cai, Y., Wang, Z., Luo, Z., Yin, B., Du, A., Wang, H., ... & Sun, J. (2020, August). Learning delicate local representations for multi-person pose estimation. In European conference on computer vision (pp. 455-472). Cham: Springer International Publishing.
54. Artacho, B., & Savakis, A. (2020). Unipose: Unified human pose estimation in single images and videos. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 7035-7044).
55. Li, W., Wang, Z., Yin, B., Peng, Q., Du, Y., Xiao, T., ... & Sun, J. (2019). Rethinking on multi-stage networks for human pose estimation. arXiv preprint arXiv:1901.00148.
56. Zhang, H., Ouyang, H., Liu, S., Qi, X., Shen, X., Yang, R., & Jia, J. (2019). Human pose estimation with spatial contextual information. arXiv preprint arXiv:1901.01760.
57. Liu, H., Liu, F., Fan, X., & Huang, D. (2021). Polarized self-attention: Towards high-quality pixel-wise regression. arXiv preprint arXiv:2107.00782.
58. Zhang, J., Chen, Z., & Tao, D. (2021). Towards high performance human keypoint detection. International Journal of Computer Vision, 129(9), 2639-2662.
59. Artacho, B., & Savakis, A. (2021). Omnipose: A multi-scale framework for multi-person pose estimation. arXiv preprint arXiv:2103.10180.
60. Bulat, A., Kossaifi, J., Tzimiropoulos, G., & Pantic, M. (2020, November). Toward fast and accurate human pose estimation via soft-gated skip connections. In 2020 15th IEEE international conference on automatic face and gesture recognition (FG 2020) (pp. 8-15). IEEE.
61. Artacho, B., & Savakis, A. (2020). Unipose: Unified human pose estimation in single images and videos. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 7035-7044).
62. Peng, X., Tang, Z., Yang, F., Feris, R. S., & Metaxas, D. (2018). Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2226-2234).
63. Chou, C. J., Chien, J. T., & Chen, H. T. (2018, November). Self adversarial training for human pose estimation. In 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (pp. 17-30). IEEE.
64. Zhou, M., Stoffl, L., Mathis, M. W., & Mathis, A. (2023). Rethinking pose estimation in crowds: overcoming the detection information bottleneck and ambiguity. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 14689-14699).
65. Jin, S., Li, S., Li, T., Liu, W., Qian, C., & Luo, P. (2024, September). You only learn one query: learning unified human query for single-stage multi-person multi-task human-centric perception. In European Conference on Computer Vision (pp. 126-146). Cham: Springer Nature Switzerland.
66. Wang, D., & Zhang, S. (2022). Contextual instance decoupling for robust multi-person pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11060-11068).
67. Xie, Z., Geng, Z., Hu, J., Zhang, Z., Hu, H., & Cao, Y. (2023). Revealing the dark secrets of masked image modeling. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 14475-14485).