

Article

Graph-Based Pattern Restoration for Occlusion-Robust Human Pose Estimation in Crowded Scenes

Mansoor Iqbal ¹, Syed Zarak Shah ² and Zahid Ullah ^{3,*}¹ Big Earth Data Analytics Department, Eratosthenes Center of Excellence, Limassol 3012, Cyprus² Department of Computational Sciences & Engineering, Harrisburg University of Science & Technology, 326 Market St, Harrisburg, PA 17101, USA³ Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, 20133 Milano, Italy

* Correspondence: zahid.ullah@polimi.it

Abstract

Human pose estimation is a core computer vision task with broad applications, yet its performance degrades significantly in crowded scenes and under heavy occlusion due to missing or unreliable visual evidence. To address this limitation, this work reformulates occluded pose estimation as a structured pattern restoration problem and proposes a graph-based framework that models the human body as a relational skeletal graph. Starting from noisy or incomplete keypoint detections, the proposed method employs a graph neural network to propagate contextual information from visible joints to occluded ones through iterative message passing. Geometry-aware constraints on bone lengths and joint angles are integrated to enforce anatomical plausibility, while an occlusion-aware prediction mechanism distinguishes visible from missing joints during inference. Experiments on COCO-Keypoints, CrowdPose, and OCHuman demonstrate consistent improvements over strong baselines, particularly under moderate and severe occlusion, confirming the effectiveness of structural reasoning for robust pose estimation in real-world environments. These results confirm that explicit structural reasoning enables more accurate, stable, and reliable human pose estimation in real-world, occlusion-heavy environments.

Keywords: human pose estimation; occlusion handling; graph neural networks; structural reasoning; crowded scenes; pattern restoration

1. Introduction

Human pose estimation is a fundamental problem in computer vision and serves as a critical building block for numerous applications, including human–computer interaction, intelligent surveillance, sports analytics, healthcare monitoring, and autonomous systems [1]. The task aims to localize anatomical landmarks, commonly referred to as keypoints, and to infer the underlying skeletal configuration of the human body from visual observations [2]. Over the past decade, substantial progress has been achieved through advances in convolutional neural networks (CNNs), high-resolution feature representations [3], attention-based architectures [4], and transformer-driven pose models [5]. Large-scale benchmarks such as COCO, MPII, and CrowdPose have further accelerated development by enabling standardized evaluation across diverse scenarios [6]. Despite these advances, pose estimation systems continue to struggle in real-world settings where visual evidence is incomplete, noisy, or highly ambiguous.

Occlusion and crowded scenes remain the most persistent and practically significant challenges faced by pose estimation systems. In unconstrained environments, such as



Academic Editors: Chih-Lung Lin,
Bor-Jiunn Hwang, Shaou-Gang Miaou
and Chi-Hung Chuang

Received: 30 December 2025

Revised: 30 January 2026

Accepted: 5 February 2026

Published: 10 February 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.

human gatherings, sports events, complex indoor scenes, etc., human bodies often overlap, self-occlude, or are partially occluded by surrounding objects [7]. Such pathologies lead to the interruption of visual consistencies and the loss of reliability of visual cues based on appearance, often losing joints, misassociation of limbs, or anatomically inconsistent skeletal structures [8]. Existing pose estimation methods, primarily based on heatmap regression or direct coordinate prediction, are highly dependent on full and localized visual evidence. In cases where this evidence cannot be found or is distorted, these models have brittle behavior, which shows a fundamental drawback of pure appearance-based pattern recognition in a partially visible world [9].

Conventional top-down and bottom-up pipelines attempt to mitigate occlusion effects through elaborate detection schemes, multi-person association heuristics, attention mechanisms, or heatmap post-processing. However, such approaches mostly act on the evidence of pixels at the local level and may therefore be prone to fail if large parts of the body are not observed. Other work introduced multi-view setups, or temporal smoothing, but such solutions typically require additional sensors, synchronized viewpoints, or video sequences, limiting their applicability in standard monocular image settings [8]. Consequently, there remains a clear need for a robust and generalizable framework that can recover human pose structure without relying on auxiliary modalities directly from the degraded visual inputs.

This work addresses the issue of degradation due to occlusion by reformulating the task of human pose estimation as a pattern-structured restoration problem. The human body, by its very nature, creates a kinematic graph where spatial and functional correlations of joints occur through well-defined anatomical constraints. These structural dependencies offer a powerful inductive prior that can be used to infer missing or unreliable keypoints, extending the inductive power beyond that provided by local appearance cues alone. It is based on this observation that we propose a pattern restoration framework based on graphs, where the human skeleton is modeled as a relational graph and message-passing operations are used to disseminate contextual information between connected joints. This design enables the recovery of occluded or ambiguous keypoints by leveraging the global pose structure and is not solely based on local visual evidence.

To further ensure coherent predictions, a structural consistency loss is introduced that regularizes bone lengths and joint angles, thereby avoiding physically implausible predictions in cases with severe occlusion. Further, a mechanism is introduced to model joint visibility for occlusion awareness and to modulate the influence of unreliable visual features within the feature extraction mechanism, thereby explicitly modeling joint visibility. Together, these components form a unified pipeline capable of producing complete and structurally consistent pose estimates from crowded and partially observed scenes.

The main contributions of this work are summarized as follows:

- Introducing a structural pattern restoration formulation for occlusion-robust human pose estimation, treating missing keypoints as a graph-based completion problem.
- Proposing a graph neural network architecture that performs message passing over the skeletal topology to infer occluded or uncertain joints.
- Designing a structural consistency loss that enforces anatomical plausibility through bone-length and joint-angle regularization.
- Integrating an occlusion-aware feature extraction mechanism that explicitly accounts for joint visibility and suppresses unreliable visual cues.
- Conducting extensive evaluations on COCO-Keypoints, CrowdPose, and OCHuman, demonstrating significant performance gains over strong baselines, particularly under moderate and heavy occlusion.

The remainder of this paper is organized as follows. Section 2 reviews related work in human pose estimation. Section 3 formalizes the occlusion-aware pose restoration problem

and defines the corresponding learning objective. Section 4 presents the proposed framework, detailing the backbone, graph-based reasoning module, and structural consistency constraints. Section 5 describes the experimental protocol, including datasets, evaluation metrics, and the implementation environment. Section 6 analyzes the quantitative results and discusses limitations and practical implications. Finally, Section 7 concludes the paper and outlines future work.

2. Related Work

Human pose estimation has advanced substantially with the adoption of deep convolutional and attention-based architectures [10–12]. A major milestone was the introduction of High-Resolution Networks (HRNet), which preserve high-resolution representations throughout the network, rather than relying on low-to-high upsampling pipelines [13]. This design enhances spatial precision and achieves robust keypoint localization performance on benchmarks such as COCO and MPII. Building on this paradigm, HigherHRNet extends HRNet to bottom-up multi-person pose estimation by constructing multi-scale feature pyramids and aggregating predictions across resolutions, improving robustness to scale variation and small or overlapping persons [14]. More recently, transformer-based pose estimators (e.g., ViTPose) have demonstrated competitive accuracy by leveraging global self-attention to capture long-range dependencies between joints [10]. Despite these gains, many state-of-the-art models remain strongly dependent on reliable visual evidence and often degrade when key body parts are partially or fully occluded.

Occlusion and crowding remain persistent challenges in real-world scenarios. In dense environments, overlapping people and self-occlusion frequently induce joint swapping, missing limbs, fragmented skeletons, and ambiguous association between persons [15,16]. To better quantify these failure modes, specialized datasets such as CrowdPose and OCHuman were introduced [17,18]. CrowdPose targets dense human–human interactions and provides a dedicated benchmark for crowded multi-person inference, while OCHuman filters COCO instances to retain heavily occluded poses, thereby exposing the limitations of conventional top-down and bottom-up pipelines under extreme partial visibility [17,19]. In response, several occlusion-aware strategies have been explored. Data augmentation methods simulate occlusion during training by masking limbs or joints to encourage robustness [20], while architectural adaptations introduce specialized heads or losses to suppress false activations caused by overlapping people. For example, CONet augments HRNet with an occlusion-aware output formulation and an interference-point loss to reduce crowd-induced confusion [21]. Although such approaches deliver measurable improvements, they primarily operate by refining detection and heatmap responses and do not explicitly represent the missing pose structure as a recoverable entity.

To move beyond purely appearance-driven inference, recent work has explored structured reasoning and visibility modeling [22]. Graph-based methods incorporate skeletal connectivity to refine initial predictions, with OPEC-Net introducing an image-guided progressive Graph Convolutional Network (GCN) that leverages pose structure to infer invisible joints more robustly [23]. However, these approaches typically treat graph reasoning as a refinement stage rather than framing occlusion handling as a first-class pattern restoration problem on a relational graph with missing nodes. In parallel, visibility-aware learning has gained attention: rethinking visibility studies show that naively supervising occluded joints can degrade estimation quality and propose visibility-aware mechanisms that modulate unreliable cues during inference [24]. Other pipelines for crowded scenes combine proposal-conditioned reasoning or association strategies to improve multi-person disambiguation in cluttered environments [6], yet they still rely predominantly on visible evidence and may fail when large body regions are unobserved.

Structural constraints such as bone-length consistency, joint-angle feasibility, and symmetry priors are well established in 3D pose estimation and are occasionally used as auxiliary regularizers in 2D settings [25]. Nevertheless, structural constraints serve as central mechanisms for completing missing 2D keypoints from monocular images, and most 2D estimators remain vulnerable under heavy occlusion, often producing anatomically implausible or fragmented poses when visual evidence collapses. Overall, the literature offers partial solutions, high-resolution backbones, occlusion-aware heads, augmentation, graph refinement, and visibility modeling, but no widely adopted framework unifies visibility-aware reasoning, graph-based message passing, and geometry-aware constraints to explicitly treat occlusion as a structured pattern completion problem [26]. This gap motivates our proposed framework, which models the human skeleton as a relational graph, treats occluded joints as latent nodes, and jointly leverages learned graph reasoning and anatomical priors to restore structurally consistent poses under severe occlusion. This work concentrates on single-frame monocular 2D pose estimation in the presence of occlusion and crowding. The proposed design is intentionally task-specific and has utilized skeletal topology as an inductive bias, rather than aiming for cross-domain generalization.

3. Problem Formulation

Human pose estimation in real-world environments frequently operates under partial visibility caused by crowding, self-occlusion, and environmental obstacles. Such conditions are common in applications such as public space monitoring, sports analytics, industrial safety, and human-robot interaction. In such situations, a pose estimation system has to make anatomically valid predictions on body configurations from incomplete visual evidence, where some keypoints are completely unobservable. This leads to one of the most fundamental problems: the problem of mapping from image observations to body joints becomes underconstrained, often resulting in unstable or statistically implausible pose predictions.

A monocular RGB image is considered $I \in \mathbb{R}^{H \times W \times 3}$ containing a single person. The human pose is represented by a set of N anatomical keypoints

$$k = \{k_1, k_2, \dots, k_N\}, \quad k_i \in \mathbb{R}^2, \quad (1)$$

where each k_i denotes the 2D image coordinates of joint i . Occlusion is modelled using a visibility indicator vector

$$v = \{v_1, v_2, \dots, v_N\}, \quad v_i \in \{0, 1\}, \quad (2)$$

where $v_i = 1$ indicates that joint i is visible, and $v_i = 0$ denotes that it is occluded or visually unreliable. The observed joint location $\tilde{k}_i$ is therefore defined as

$$\tilde{k}_i = \begin{cases} k_i + \epsilon_i, & v_i = 1, \\ \emptyset, & v_i = 0, \end{cases} \quad (3)$$

with ϵ_i representing localization noise introduced by imaging conditions or estimation errors.

For visible joints, the observation likelihood can be approximated by a unimodal distribution,

$$p(\tilde{k}_i | I) = \mathcal{N}(\mu_i(I), \Sigma_i(I)), \quad (4)$$

whereas for occluded joints ($v_i = 0$), this likelihood is undefined due to the absence of reliable pixel evidence. When multiple joints are simultaneously occluded—as is common in crowded scenes—the pose estimation problem becomes severely under-determined.

To regularize this ill-posed setting, we exploit the inherent structure of the human body. The skeleton is modeled as an undirected kinematic graph

$$G = (V, E), \quad (5)$$

where each node $i \in V$ corresponds to a joint and each edge $(i, j) \in E$ represents an anatomically valid connection (e.g., shoulder–elbow and hip–knee). This graph encodes strong biomechanical priors that constrain feasible human poses. The skeletal graph topology is determined based on standard anatomical connectivity, providing a partially visible biomechanical prior. The graph structure is not dynamic, but the proposed framework enables adaptive relational reasoning through learned edge embeddings and visibility-sensitive message passage and allows dynamically modulating inter-joint influence across a range of occlusion patterns.

Specifically, connected joints obey stable bone-length relationships

$$\|k_i - k_j\|_2 \approx d_{ij}^{\text{ref}}, \quad (i, j) \in E, \quad (6)$$

where d_{ij}^{ref} denotes a reference bone length normalized across subjects. In addition, joint articulation is constrained by feasible angle ranges. In this work, these feasibility ranges are applied as soft regularizers to stabilize reconstruction under occlusion and are calibrated to typical physiological motion; consequently, rare non-standard articulations (e.g., injury- or pathology-induced deviations) are not explicitly modeled and may require relaxed or adaptive angle priors in domain-specific deployments. Let θ_{ijk} denote the angle at joint j formed by joints (i, j, k) ; then physical plausibility requires

$$\theta_{ijk} \in [\theta_{ijk}^{\min}, \theta_{ijk}^{\max}]. \quad (7)$$

Let $\mathcal{V} = \{i \mid v_i = 1\}$ and $\mathcal{O} = \{i \mid v_i = 0\}$ denote the sets of visible and occluded joints, respectively. Given the observed joints $\tilde{\mathbf{k}}_{\mathcal{V}}$, the objective is to recover a complete pose $\mathbf{k}$ that is consistent with both visual evidence and anatomical structure.

Under extreme visibility collapse, the reconstruction is underdetermined: several anatomically possible poses can fit the same sparse visual information. These kinematic constraints are not semantically correct regularizers but plausibility regularizers, even in the case of incomplete observations.

Joint-angle feasibility is defined using anatomical degrees of freedom (DoF) to ensure biomechanically plausible pose reconstruction under occlusion. Hinge joints (e.g., elbow, knee) exhibit one dominant rotational DoF, ball-and-socket joints (e.g., shoulder, hip) allow multi-axis rotation, and wrist/ankle joints follow constrained two-DoF motion. The DoF categories used in this work are summarized in Table 1 and guide the feasible angle intervals $[\theta_{ijk}^{\min}, \theta_{ijk}^{\max}]$ employed in $\mathcal{L}_{\text{angle}}$.

This can be expressed as the optimisation problem

$$\mathbf{k}^* = \arg \min_{\mathbf{k}} \left[\mathcal{L}_{\text{vis}}(\mathbf{k}_{\mathcal{V}}, \tilde{\mathbf{k}}_{\mathcal{V}}) + \lambda_1 \mathcal{L}_{\text{bone}}(\mathbf{k}, G) + \lambda_2 \mathcal{L}_{\text{angle}}(\mathbf{k}, G) \right], \quad (8)$$

where the loss terms are defined as

$$\mathcal{L}_{\text{vis}} = \sum_{i \in \mathcal{V}} \|k_i - \tilde{k}_i\|_2^2, \quad (9)$$

$$\mathcal{L}_{\text{bone}} = \sum_{(i,j) \in E} \left(\|k_i - k_j\|_2 - d_{ij}^{\text{ref}} \right)^2, \quad (10)$$

and

$$\mathcal{L}_{\text{angle}} = \sum_{(i,j,k) \in \mathcal{T}} \psi(\theta_{ijk}), \quad (11)$$

with $\mathcal{T}$ denoting valid kinematic triplets and $\psi(\cdot)$ penalizing angles outside the feasible range.

Table 1. Anatomical joint categories and corresponding degrees of freedom (DoF) used to define kinematic feasibility constraints.

Joint Category	Example Joints	Degrees of Freedom
Hinge joints	Elbow, Knee	1 (Flexion–Extension)
Ball-and-socket joints	Shoulder, Hip	3 (Multi-axis rotation)
Limited-rotation joints	Wrist, Ankle	2 (Constrained rotation)
Axial joints	Neck, Upper torso	2 (Restricted rotation)

Note: DoF values in Table 1 denote simplified rotational DoF used as kinematic priors for 2D feasibility constraints (not full anatomical/clinical joint DoF).

This formulation captures a practical and well-defined real-world problem: restoring complete and anatomically consistent human poses from incomplete visual observations in crowded or occluded scenes. Explicitly incorporating structural priors for inferring missing keypoints in cases where visual evidence is insufficient is a principled approach to support the graph-based restoration approach proposed in the next section.

4. Proposed Methodology

An occlusion-robust human pose estimation framework is proposed that formulates pose inference under partial visibility as a structured pattern restoration problem over a relational human-body graph. Unlike conventional heatmap-based methods that deteriorate when visual evidence is missing, the proposed approach integrates occlusion-aware feature extraction, graph-based relational reasoning, and geometric consistency enforcement within a unified architecture. The framework progressively refines incomplete or noisy pose predictions by leveraging anatomical structure and visibility cues, enabling reliable reconstruction even in crowded or severely occluded scenes. An overview of the full pipeline is shown in Figure 1.

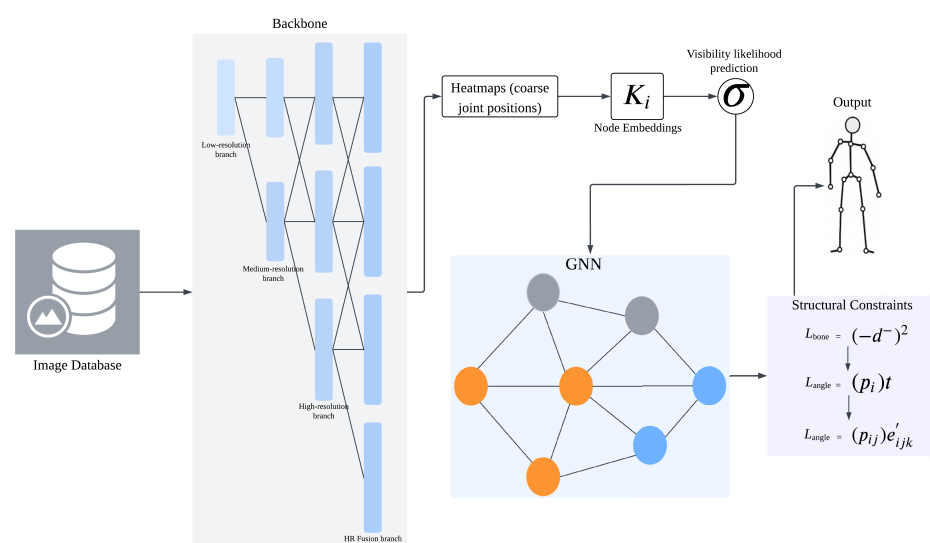


Figure 1. Proposed occlusion-robust pose restoration pipeline. The backbone extracts high-resolution pose features and predicts visibility, which form node embeddings used by a GNN for structural reasoning. The geometric consistency module ensures anatomically plausible reconstruction.

4.1. Occlusion-Aware Backbone and Initial Pose Estimation

Given an input image, we employ a high-resolution backbone (HRNet-W48 by default) to extract multi-scale feature representations while preserving spatial precision, as shown in Figure 2. The backbone outputs coarse joint heatmaps $\{H_i\}$, from which initial joint estimates $\tilde{k}_i$ are obtained, along with ROI-pooled local features f_i . In parallel, a lightweight visibility head predicts a joint-wise visibility likelihood $\hat{v}_i \in [0, 1]$, indicating the reliability of visual evidence for each keypoint. This visibility-aware design allows the model to explicitly distinguish between reliable and occluded joints early in the pipeline.

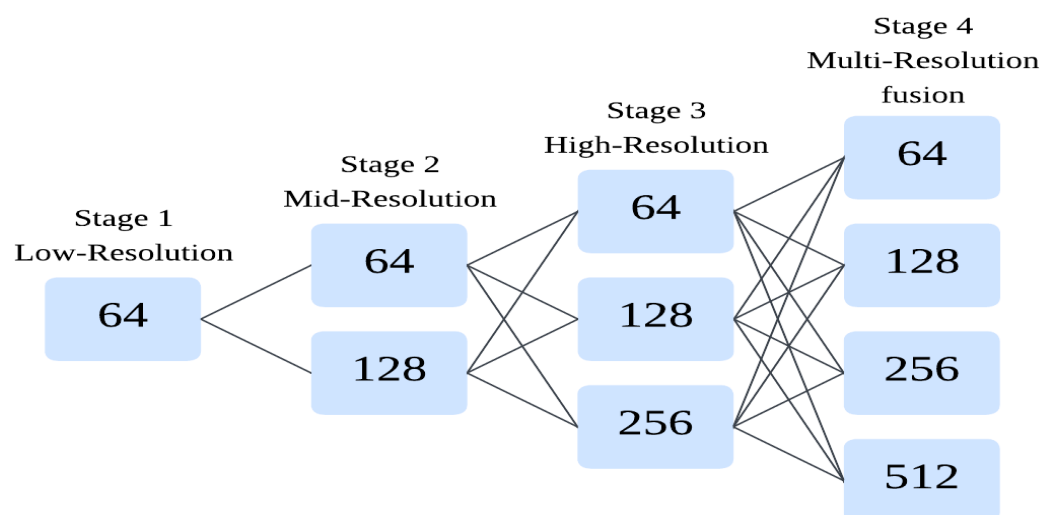


Figure 2. Detailed HRNet backbone used in proposed framework, showing multi-resolution branches and cross-stage fusion that preserve high-resolution features for heatmap and visibility prediction.

4.2. Node Embedding Construction

To enable structured reasoning, each joint is represented as a node embedding that integrates positional, visual, and uncertainty information. Specifically, for joint i , we construct the initial node embedding as

$$h_i^{(0)} = [\tilde{k}_i \parallel \hat{v}_i \parallel f_i] \in \mathbb{R}^d, \quad (12)$$

where $\tilde{k}_i$ denotes the coarse joint location, $\hat{v}_i$ the visibility likelihood, and f_i the corresponding local appearance descriptor. For occluded joints, positional entries are zero-filled while appearance features are retained, explicitly encoding uncertainty into the graph representation. The embedding construction process is illustrated in Figure 3.

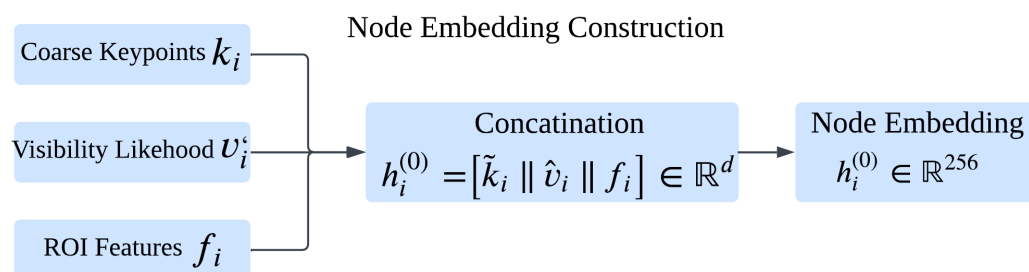


Figure 3. Construction of node embeddings from coarse keypoints $\tilde{k}_i$, visibility likelihoods $\hat{v}_i$, and ROI features f_i . These components are concatenated to form the initial joint embedding $h_i^{(0)}$ used by the graph reasoning module.

4.3. Graph-Based Structural Reasoning

The human body is modeled as a graph $G = (V, E)$, where nodes correspond to joints and edges encode anatomical connectivity. Graph reasoning is performed via a multi-layer graph neural network that iteratively propagates contextual information across connected joints. At iteration t , each node aggregates messages from its neighbors:

$$m_i^{(t)} = \sum_{j:(i,j) \in E} \phi_m(h_j^{(t)}, e_{ij}), \quad (13)$$

where ϕ_m is a learned message function and e_{ij} denotes a trainable edge embedding.

Unlike generic multi-branch fusion, graph propagation is aligned with skeletal topology and explicitly encodes joint-to-joint dependencies required for structured completion.

The primary fusion mechanism is message passing, since pose restoration is relational: occluded joints must be inferred from anatomically connected evidence. Unlike generic multi-branch fusion, graph propagation is aligned with skeletal topology and explicitly represents joint-to-joint dependencies to complete the structure.

Node states are updated through a visibility-aware update function:

$$h_i^{(t+1)} = \phi_u(h_i^{(t)}, m_i^{(t)}, \hat{v}_i). \quad (14)$$

This design enables inference of occluded joints by leveraging information from visible, structurally stable joints. Three message-passing layers are employed, yielding refined embeddings that encode globally consistent pose structure. The reasoning process is visualized in Figure 4.

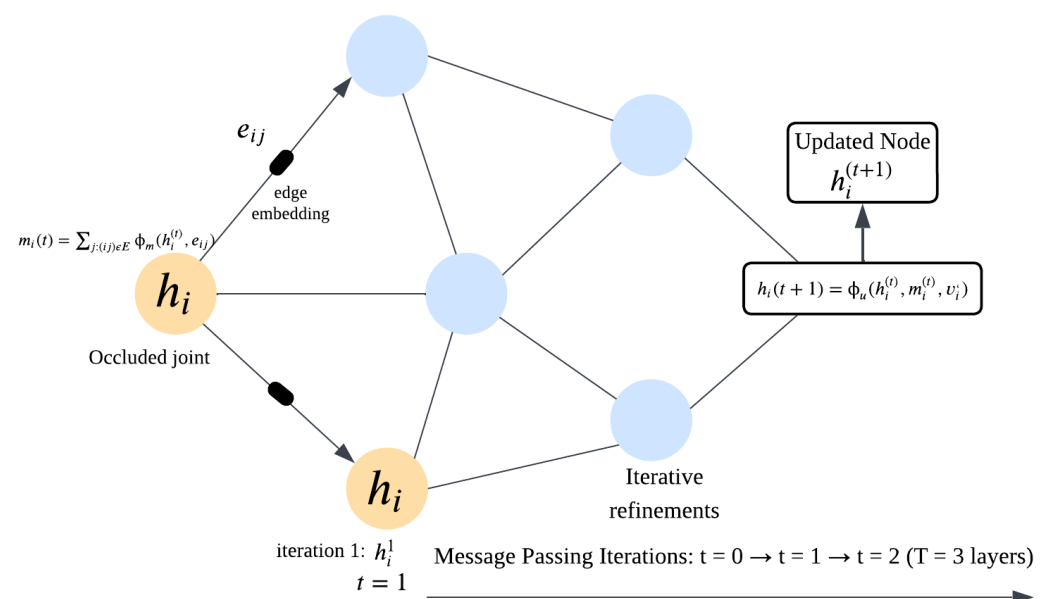


Figure 4. Graph message passing and visibility-aware node updates used to refine occluded joints, illustrating the iterative GNN reasoning process adopted in the proposed framework.

4.4. Geometric Consistency and Training Objective

Final 2D joint coordinates are obtained via a linear regression head applied to the refined embeddings. To ensure anatomical plausibility, the model is trained with a composite loss that combines visible-joint alignment, bone-length consistency, and joint-angle feasibility constraints:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{vis}} + \lambda_1 \mathcal{L}_{\text{bone}} + \lambda_2 \mathcal{L}_{\text{angle}} + \lambda_3 \mathcal{L}_{\text{GNN}}. \quad (15)$$

These constraints regularize the reconstruction toward biomechanically valid poses and prevent physically implausible configurations, particularly when large body parts are occluded. The integration of structural constraints is illustrated in Figure 5.

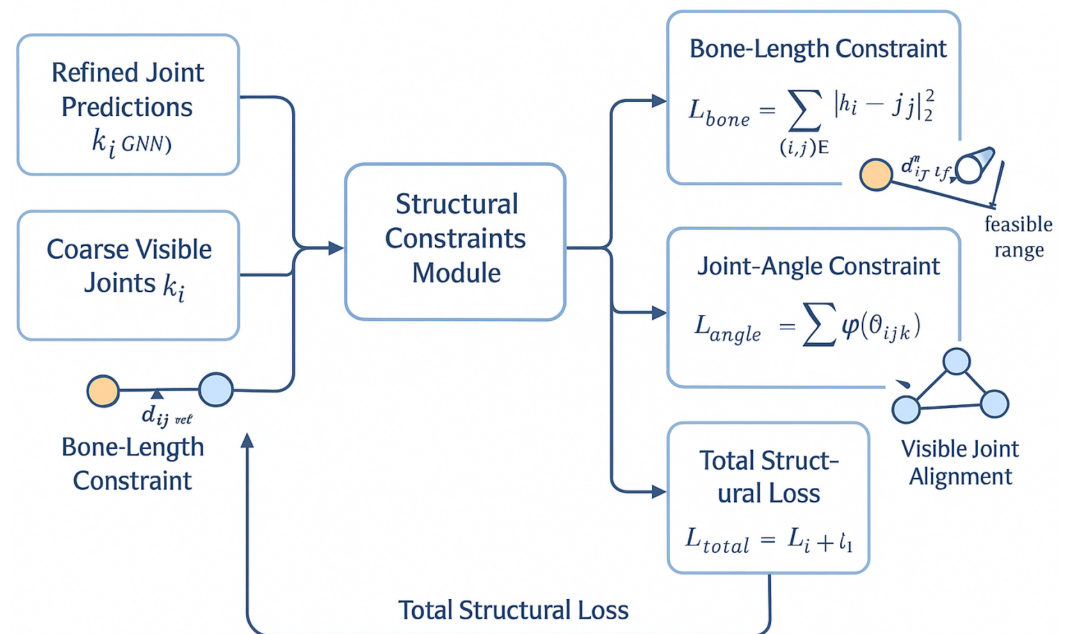


Figure 5. The refined GNN keypoints and visible-joint evidence are regularized using bone-length deviation, joint-angle feasibility, and visible-keypoint alignment constraints, which are aggregated into the total structural loss enforcing bio-mechanically valid pose reconstruction.

4.5. Architecture Details

All architectural choices and training hyperparameters are summarized in Table 2. For clarity and reproducibility, Algorithm 1 summarizes the end-to-end pose restoration procedure used in this work. Given an input image I and the skeletal graph $G = (V, E)$, the pipeline first extracts backbone feature maps F , joint heatmaps H , ROI features f_i , and visibility scores $\hat{v}_i$ (Step 1). Coarse joint locations $\tilde{k}_i$ are then obtained from the heatmaps for all joints $i \in V$ (Step 2). Each joint is initialized as a node embedding $h_i^{(0)}$ by combining $\tilde{k}_i$, $\hat{v}_i$, and f_i , thereby encoding both spatial evidence and occlusion uncertainty (Step 3).

Table 2. Summary of architectural and optimization details for the proposed pose restoration system.

Component	Specification
Backbone	HRNet-W48/ViTPose-S; 64/128/256/512 channels; heatmap size 64×48
Visibility Head	MLP: $256 \rightarrow 128 \rightarrow 1$; ReLU, Sigmoid; ROI features: 512-d
Node Embedding	256-d vector: $[\tilde{k}_i, \hat{v}_i, f_i]$
GNN Layers	3 layers, 256 units each; residual connections
Message Function	2-layer MLP, 256 units, GELU
Update Function	Gated fusion with GELU
Edge Embeddings	32-dimensional learned vectors
Regression Head	Linear: $256 \rightarrow 2$
Loss Weights	$\lambda_1 = 0.4, \lambda_2 = 0.2, \lambda_3 = 1.0$
Optimizer	Adam, LR 10^{-4} , cosine decay, warm-up of 5 epochs
Training	220 epochs, batch size 64, FP16 mixed precision
Total Parameters	73M (HRNet), 92M (ViTPose)

Algorithm 1 provides a procedural summary of the full inference and training pipeline.

Algorithm 1 Pose Restoration Pipeline

Require: Image I , human-body graph $G = (V, E)$, model parameters
Ensure: Refined keypoints $\{k_i^{\text{GNN}}\}_{i \in V}$

- 1: Extract feature maps F , heatmaps H , ROI features f_i , and visibility scores $\hat{v}_i$ from I
- 2: Compute coarse joint locations $\tilde{k}_i$ from H_i for all $i \in V$
- 3: Construct node embeddings $h_i^{(0)} \leftarrow [\tilde{k}_i \parallel \hat{v}_i \parallel f_i]$ for all $i \in V$
- 4: **for** $t = 0$ to $T - 1$ **do**
- 5: **for all** $i \in V$ **do**
- 6: $m_i^{(t)} \leftarrow \sum_{j: (i,j) \in E} \phi_m(h_j^{(t)}, e_{ij})$
- 7: $h_i^{(t+1)} \leftarrow \phi_u(h_i^{(t)}, m_i^{(t)}, \hat{v}_i)$
- 8: **end for**
- 9: **end for**
- 10: **for all** $i \in V$ **do**
- 11: $k_i^{\text{GNN}} \leftarrow W_k^\top h_i^{(T)} + b_k$
- 12: **end for**
- 13: Compute loss $\mathcal{L}_{\text{total}}$ using $\{k_i^{\text{GNN}}\}$, $\{\tilde{k}_i\}$ and geometric constraints
- 14: Update all model parameters using Adam optimiser
- 15: **return** $\{k_i^{\text{GNN}}\}_{i \in V}$

Pose completion is performed via T iterations of graph message passing (Steps 4–9). At each iteration, node i aggregates messages from its neighbors j through edge embeddings e_{ij} , forming $m_i^{(t)}$, and updates its hidden state using $\phi_u(\cdot)$, which integrates the previous state, aggregated structural context, and visibility cues. After T iterations, refined keypoints are obtained by projecting the final node states $h_i^{(T)}$ to coordinates, yielding $\{k_i^{\text{GNN}}\}_{i \in V}$ (Steps 10–12). During training, the model minimizes a total objective that combines keypoint supervision with geometric regularization (bone-length and joint-angle constraints), and the parameters are updated using the Adam algorithm (Steps 13–14).

Overall, the algorithm operationalizes our formulation of occlusion-robust pose estimation as a structured completion problem, producing anatomically consistent keypoints.

5. Performance Validation

This section outlines the datasets, evaluation metrics, and environmental setup and configurations used to evaluate the proposed method. We evaluate the framework on largely used benchmarks for human pose estimation, which differ in terms of occlusion, crowd density, and appearance diversity. All experiments were carried out under controlled and reproducible conditions to ensure a fair comparison with state-of-the-art approaches could be made.

5.1. Datasets

To evaluate the proposed occlusion-aware pose restoration framework comprehensively, experiments were conducted on four complementary benchmark datasets that collectively span a wide spectrum of scene complexity, human interaction density, and occlusion severity. This selection allows for the analysis of the robustness of the model for increasingly difficult visibility conditions systematically.

CrowdPose [17] specifically addresses crowded environments where there is significant human–human overlap and can therefore be used to measure how robust the algorithm is to mutual occlusion and identity ambiguity problems. COCO-Keypoints [27] is the most frequently used large-scale benchmark, which provides various indoor and outdoor scenes with moderate occlusions and large appearance variations. It is used for training as well

as general performance evaluation. In order to test performance under extreme occlusion, OCHuman [28] is used, which includes instances of people with heavy overlapping so that a majority of the joints are not visible. Finally, MPII Human Pose [29] has high-quality annotations in mostly single-person scenarios with significant self-occlusion due to articulations, which allows for the study of generalization and structural reasoning beyond crowd-heavy settings.

A brief comparison of datasets, the format of the annotation, and the characteristics of the occlusion is presented in Table 3. Understanding the distribution of occlusion levels is critical for interpreting performance trends. Figure 6 is used to visualize the distribution of the low-, moderate-, and high-occlusion instances across all four datasets. The figure shows a distinct development from COCO and MPII, which are dominated by low to moderate occlusion, to CrowdPose and OCHuman, which focus on dense interactions and severe visibility loss. This distribution is in line with our experimental design, which allows one to control stress and test the robustness of occlusion.

Table 3. Summary of datasets used for evaluation, highlighting scale, annotation details, and occlusion characteristics.

Dataset	Images	Keypoints	Occlusion Level	Description and Usage
CrowdPose [17]	20 k images (98 k persons)	14	High	Designed for dense human–human occlusion; critical for evaluating robustness in crowded scenes.
COCO-Keypoints [27]	~200 k persons	17	Moderate	Large-scale benchmark with diverse scenes and partial occlusion; used for training and general evaluation.
OCHuman [28]	8 k persons	17	Extreme	Focuses on heavily overlapping persons (>70% occluded); ideal for keypoint completion evaluation.
MPII [29]	25 k images	16	Low–Moderate	Mostly single-person images with frequent self-occlusion; used for cross-dataset generalisation.

While the occlusion statistics provide a quantitative analysis, visual examination offers additional insight into the types of problems represented in each dataset. Figure 7 shows representative samples showing differences in complexity of the scene, density of the crowd, and visibility. COCO has various scenes that contain partial occlusion, MPII focuses on self-occlusion because of articulation, CrowdPose has frequent multi-person overlap, and OCHuman has extreme scenarios where huge body parts are completely invisible. Together, these examples are great visual motivation for the need for explicit structural reasoning and keypoint completion mechanisms.

Overall, the combination of these datasets provides a gradual assessment range, from mild to extreme occlusion, allowing for a strict evaluation of the proposed model in terms of its capacity to detect visible joints, project occluded structures, and generalize across different real-life settings.

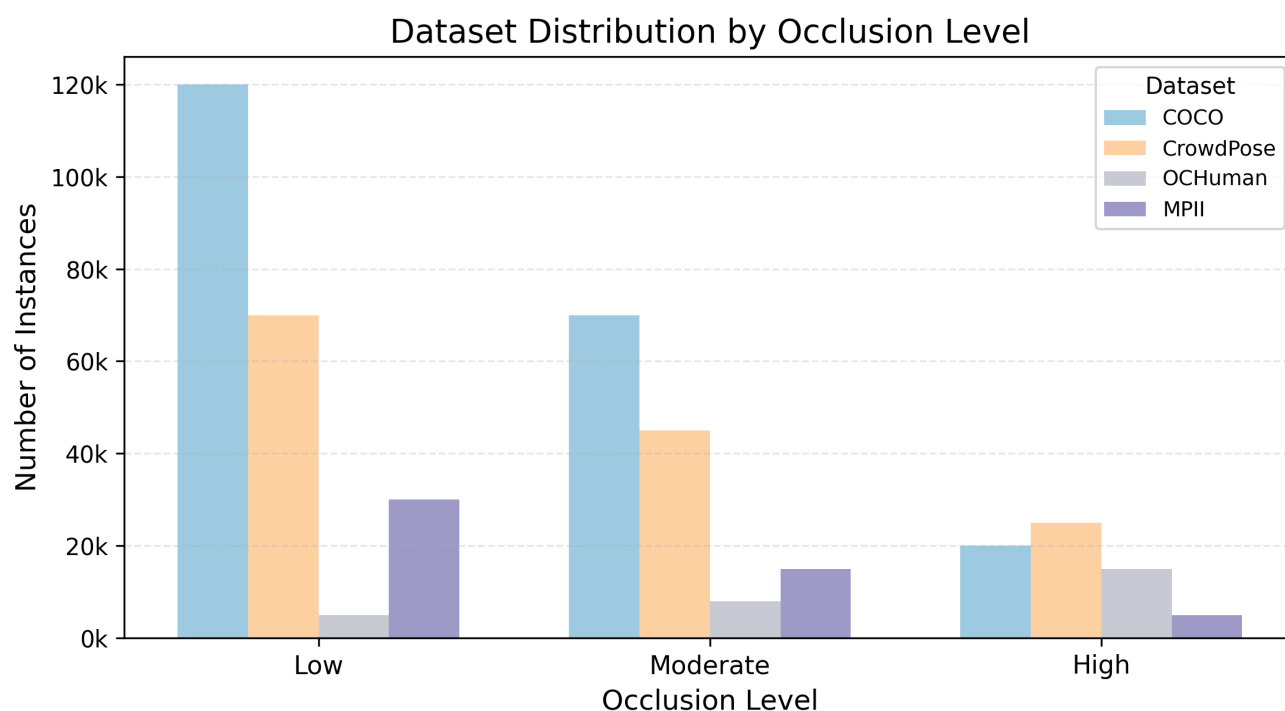


Figure 6. Distribution of occlusion levels across COCO, CrowdPose, OCHuman, and MPII. The datasets collectively form a continuum from low to extreme occlusion, enabling systematic evaluation of occlusion robustness.

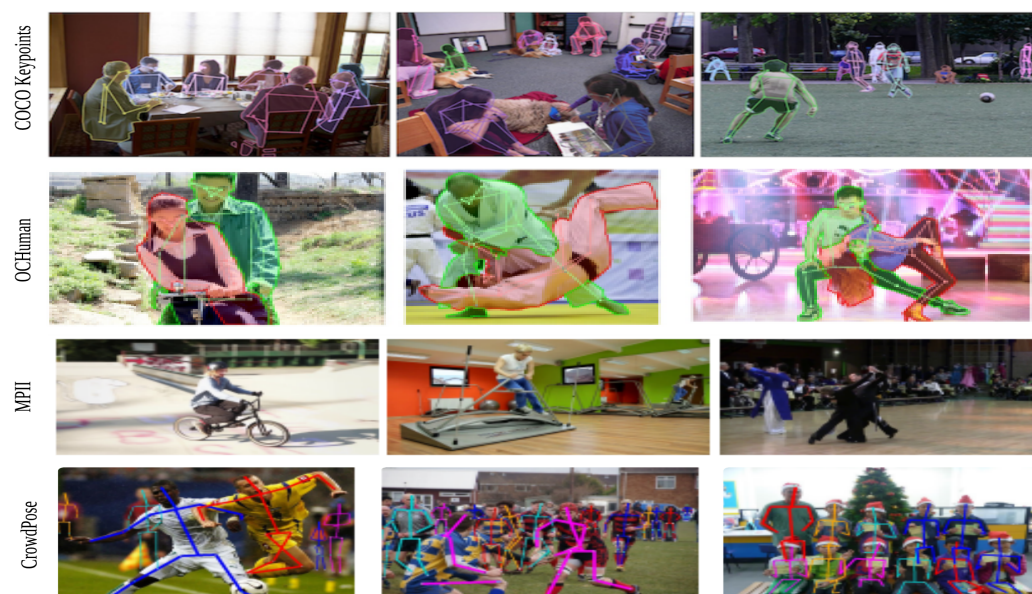


Figure 7. Representative samples of the evaluation data sets. COCO-Keypoints presents moderate occlusion across various scenarios, MPII focuses on articulation-based self-occlusions, CrowdPose involves dense human-to-human interactions, and OCHuman exhibits extreme overlap with severe visibility loss.

5.2. Metrics

To evaluate the proposed framework, three established metrics were used to evaluate localization accuracy, structural consistency, and robustness under occlusion.

5.2.1. Average Precision (AP)

Following the COCO keypoint evaluation protocol, AP is computed using the Object Keypoint Similarity (OKS) and averaged over multiple OKS thresholds $\tau \in \{0.50, 0.55, \dots, 0.95\}$:

$$\text{AP} = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \text{AP}_{\tau}, \quad \mathcal{T} = \{0.50, 0.55, \dots, 0.95\}, \quad (16)$$

where AP_{τ} denotes the area under the precision–recall curve when a prediction is considered correct if $\text{OKS} \geq \tau$.

5.2.2. Object Keypoint Similarity (OKS)

OKS measures the similarity between predicted and ground-truth keypoints, normalized by object scale and per-joint labeling variance:

$$\text{OKS} = \frac{\sum_i \exp\left(-\frac{\|k_i - k_i^{gt}\|_2^2}{2s^2\kappa_i^2}\right) \cdot \delta(v_i = 1)}{\sum_i \delta(v_i = 1)}, \quad (17)$$

where s is the object scale, κ_i is a per-joint constant, and $\delta(v_i = 1)$ masks out occluded joints. This is consistent with our visibility model: occluded keypoints ($v_i = 0$) are excluded from similarity computation, reflecting the lack of visual evidence.

5.2.3. Percentage of Correct Keypoints (PCK)

PCK quantifies the percentage of predicted joints that fall within a normalized radius around the ground-truth position:

$$\text{PCK} = \frac{\#\{i : \|k_i - k_i^{gt}\|_2 < \alpha L\}}{N}, \quad (18)$$

where L is a normalization length (typically head size or torso length), and α is set to 0.2. The term $\#$ represents the cardinality of the set, i.e., the number of keypoints that are at the Euclidean distance to the ground-truth location under the normalized threshold of αL , and N is the number of keypoints that are considered. PCK provides an interpretable spatial accuracy measure and complements OKS in scenarios with severe occlusion.

AP and OKS are reported for COCO, CrowdPose, and OCHuman, while PCK is additionally used for MPII following standard benchmarking practice.

5.3. Implementation Details

To ensure reproducibility and fair comparison, all experiments are conducted under a unified software and hardware configuration. The same training protocol, optimization strategy, and random seed initialization are consistently applied across the proposed method and all baseline models in the experimental evaluation. The high-resolution backbones are associated with the need for spatially specific keypoint localization, which adds more parameters than lightweight end-to-end models in other settings. The restoration module is compatible with lightweight backbones and graph pruning/compression, which will be explored to further improve efficiency.

5.3.1. Environmental Setup

All the experiments are implemented in Python 3.10 and are executed on a Windows 11 operating system. A dedicated conda virtual environment is created to manage dependencies and avoid version conflicts. The core deep learning framework is PyTorch 2.0, with CUDA 11.8 and cuDNN 8.x for GPU acceleration. Additional libraries

include torchvision for data handling, NumPy and SciPy for numerical operations, and OpenCV for image preprocessing, as shown in Figure 8. All random number generators are seeded so that the behavior is deterministic during training and evaluation.

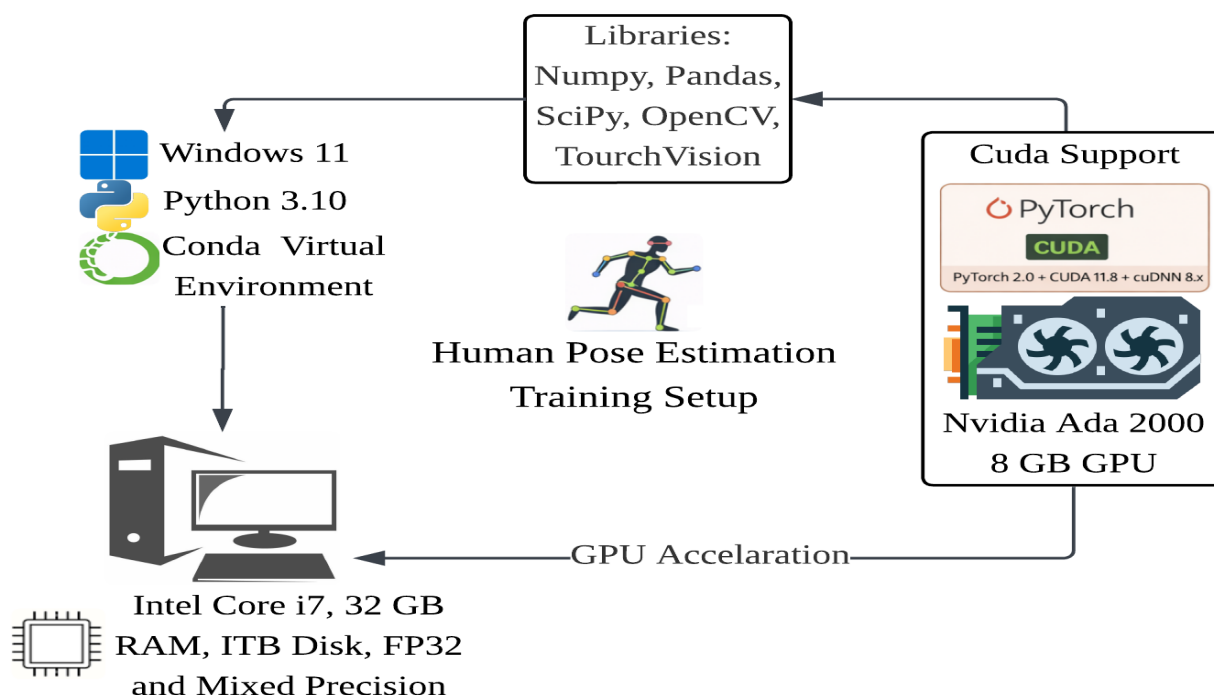


Figure 8. Experimental setup.

All models are trained and evaluated on a single workstation that has an NVIDIA RTX Ada 2000 GPU (8GB VRAM), an Intel Core i7, and 32GB system memory. Unless otherwise specified, training is done on one GPU. The implementation setup is summarized in Table 4. This configuration is indicative of a realistic research and deployment environment as opposed to a large-scale multi-GPU infrastructure.

Table 4. Experimental setup and implementation details used for all models.

Component	Specification
Operating System	Windows 11
Python Version	Python 3.10
Environment	Conda virtual environment
Deep Learning Framework	PyTorch 2.0, CUDA 11.8, cuDNN 8.x
Key Libraries	torchvision, NumPy, SciPy, OpenCV
GPU	NVIDIA RTX Ada 2000 (8 GB VRAM)
CPU	Intel Core i7 Ultra
System Memory	32 GB RAM
Training Epochs	220
Optimizer	Adam
Initial Learning Rate	1×10^{-4} (cosine decay, warm-up 5 epochs)
Batch Size	64
Input Resolution	256×192
Data Augmentation	Rotation, scaling, flipping, synthetic occlusion
Random Seed	Fixed for all experiments
Baseline Configuration	Official settings from original papers

5.3.2. Training Protocol

The backbones used in both HRNet-W48 and ViTPose-S are initialized by ImageNet weights. Models are trained for 220 epochs using the Adam optimizer with an initial learning rate of 1×10^{-4} , decayed via a cosine annealing schedule with a warm-up period of five epochs.

A batch size of 64 is used when training. Input images are resized to 256×192 resolution. Standard data augmentation techniques are used, such as random rotation ($\pm 40^\circ$), random scaling (0.7–1.35), horizontal flipping, and synthetic occlusion masking to enhance robustness against missing visual evidence.

5.3.3. Graph Configuration and Initialisation

The human-body graph follows standard skeletal connectivity with $N = 17$ joints for the COCO-based dataset and $N = 16$ joints for the MPII. The graph neural network consists of three message-passing layers and 256-dimensional hidden states and learned edge embeddings of 32 dimensions. Node embeddings are initialized using the concatenation of coarse heatmap-based joint estimates, predicted visibility likelihoods, and ROI-pooled appearance features. Occluded joints are explicitly represented by zero-initialized coordinates while retaining local appearance descriptors.

5.3.4. Loss Weights and Reproducibility

The composite training objective combines visible-joint alignment, bone-length consistency, and joint-angle feasibility losses, weighted by $\lambda_1 = 0.4$, $\lambda_2 = 0.2$, and $\lambda_3 = 1.0$, respectively. These values are selected via validation on the COCO split and are fixed across all experiments. All baseline models are trained or evaluated using their officially reported configurations and hyperparameters, with the same data splits and evaluation protocols, ensuring a fair and controlled comparison.

This implementation ensures a fair and rigorous comparison with existing state-of-the-art methods, demonstrating that the proposed approach remains robust across a wide range of occlusion severities.

6. Results and Discussion

This section evaluates the proposed framework in relation to the core problem formulated in Section 3, which is to recover complete and anatomically consistent human poses when visual evidence is partially or severely missing. In contrast to traditional methods for pose estimation, which primarily utilize visible image cues, the proposed method explicitly models visible image cues as occlusion through visibility prediction, relational graph reasoning, and structural constraints. Accordingly, it is necessary in experimental analysis to have not only detection accuracy but also robustness against various degrees of occlusion, as well as the structural plausibility of reconstructed poses and computational feasibility.

6.1. Experimental Results

To eliminate citation-dependent ambiguity and ensure clear result provenance, all baseline numbers reported in this section are produced in this work by executing each method through its official public repository (or the official upstream framework explicitly used by the authors) and by following the repository-provided configuration for training and inference (model variant, input resolution, test-time settings, and dataset-specific preprocessing). For each baseline, evaluation is performed according to the standard protocol for the target benchmark. Official repository links are provided in the corresponding method papers and are also consolidated in our manuscript for traceability. To ensure a fair and consistent comparison, all baselines and the proposed method are executed in the

same software/hardware environment described in Section 5.3, and all random number generators are seeded to enable deterministic evaluation.

When a baseline provides multiple capacity variants (e.g., Small/Base/Large), the results are reported for the small variant evaluated in this study, and the corresponding parameter/FLOP counts are stated accordingly; larger variants typically yield higher accuracy at increased computational cost, so variant selection is kept consistent with the official repository's recommended setting for the target benchmark.

A diverse set of established baselines spanning top-down, bottom-up, transformer-based, visibility-aware, and occlusion-focused paradigms is considered: OpenPose [11], ViTPose-S [12], HRNet-W48 [13], HigherHRNet [14], CrowdPose baseline [17], Pose2Seg [18], Limb-Augment [20], OPEC-Net (GCN) [23], and Rethinking Visibility [24]. (Methods are included only when an official implementation is available to enable verifiable reproduction.) Table 5 summarizes the primary accuracy metrics across COCO, CrowdPose, OCHuman, and MPII.

Table 5. Comparison with state-of-the-art methods on COCO, CrowdPose, OCHuman, and MPII. All baseline results are reproduced in this work using official implementations under the unified experimental setup.

Method	COCO AP	CrowdPose AP	OCHuman AP	MPII PCK
OpenPose [11]	61.8	57.2	32.1	85.1
ViTPose-S [12]	78.0	73.9	46.5	93.1
HRNet-W48 [13]	75.5	70.6	43.7	92.3
HigherHRNet [14]	76.4	72.1	45.3	92.8
CrowdPose Baseline [17]	70.6	69.0	40.3	90.1
Pose2Seg [18]	73.1	69.8	41.2	90.6
Limb-Augment [20]	74.5	71.8	44.9	91.7
OPEC-Net (GCN) [23]	74.9	71.0	46.1	92.0
Rethinking Visibility [24]	77.2	73.1	47.8	93.0
Proposed (Ours)	79.6	76.8	50.7	94.2

The proposed framework achieves the strongest overall performance across all benchmarks. The largest margin is observed on OCHuman, which emphasizes severe occlusion: the proposed method improves upon ViTPose-S by +4.2 AP and upon Rethinking Visibility by +2.9 AP. These gains are consistent with the intended design objective—restoring missing joints via relational propagation and biomechanical regularization when direct visual evidence becomes unreliable. On MPII, which contains predominantly low-occlusion single-person poses, the proposed method also achieves the highest PCK (94.2), indicating that structural reasoning does not trade off accuracy under standard visibility.

6.1.1. Object Keypoint Similarity

Object Keypoint Similarity (OKS) normalizes localization error by object scale and keypoint-specific variance, providing a more sensitive measure of structural fidelity under occlusion. Table 6 reports OKS across COCO, CrowdPose, and OCHuman.

The OKS gains, particularly on OCHuman (+0.042 over ViTPose-S), indicate that the proposed method improves not only joint coverage but also the spatial coherence of reconstructed poses. Since OKS down-weights uncertain keypoints through per-joint variance, these improvements reflect more faithful structural restoration rather than overconfident predictions.

Table 6. OKS comparison across COCO, CrowdPose, and OCHuman. All baseline results are reproduced in this work using official implementations under the unified experimental setup.

Method	COCO OKS	CrowdPose OKS	OCHuman OKS
OpenPose [11]	0.782	0.741	0.512
ViTPose-S [12]	0.904	0.874	0.661
HRNet-W48 [13]	0.882	0.854	0.627
HigherHRNet [14]	0.891	0.866	0.648
CrowdPose Baseline [17]	0.861	0.842	0.588
Pose2Seg [18]	0.874	0.847	0.603
Limb-Augment [20]	0.884	0.859	0.640
OPEC-Net (GCN) [23]	0.887	0.861	0.654
Rethinking Visibility [24]	0.903	0.876	0.671
Proposed (Ours)	0.923	0.897	0.703

6.1.2. Occlusion-Stratified Performance and Structural Consistency

Figure 9 consolidates the behavior of the evaluated methods across occlusion-stratified AP, OKS, visibility prediction, and structural consistency. The proposed framework maintains the strongest performance under moderate and heavy occlusion and attains the lowest bone-length and joint-angle errors, supporting the claim that accuracy improvements coincide with improved anatomical consistency.

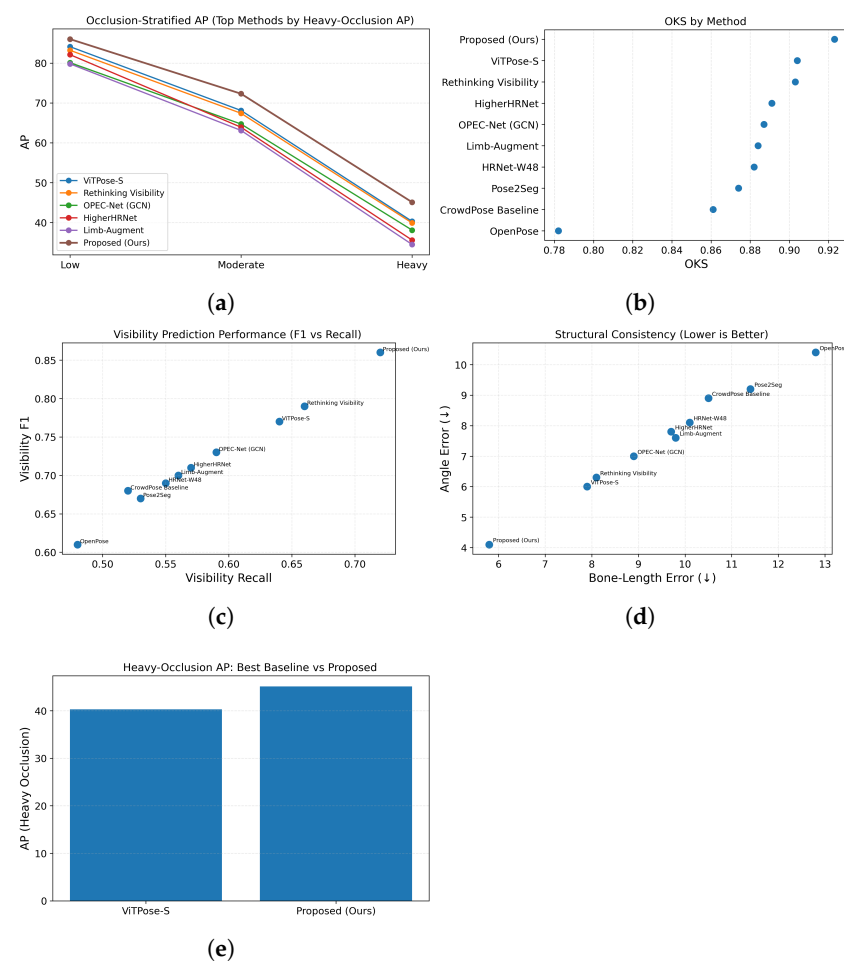


Figure 9. Integrated quantitative results. (a) AP under low/moderate/heavy occlusion; (b) OKS by method; (c) visibility prediction (F1 vs. recall); (d) structural consistency (bone-length and joint-angle error, lower is better); (e) heavy-occlusion AP comparison between the best baseline (ViTPose-S) and the proposed method.

6.1.3. Failure Modes and Behavior Analysis

Beyond aggregate metrics, Table 7 summarizes typical qualitative failure patterns observed for representative baselines and contrasts them with the behavior of the proposed framework.

Table 7. Behavioral analysis under common occlusion scenarios.

Occlusion Scenario	Typical Baseline Behavior	Proposed Model Behavior
Moderate limb occlusion	Mislocalized elbows or knees; inconsistent limb length	Coherent limb geometry via graph-based completion
Crowded scenes	Joint swapping between persons; broken skeletons	More stable person-wise poses via structural propagation
Severe lower-body occlusion	Missing or floating leg joints; implausible stance	Plausible lower-body estimates constrained by structure
Self-occlusion	Incorrect limb orientation; excessive joint bending	Angles constrained within feasible kinematic ranges
Low-occlusion scenes	High accuracy with minor joint jitter	Reduced jitter due to structural regularization

6.1.4. Runtime and Model Efficiency

Table 8 reports model size and runtime statistics of the proposed method and all baselines. To prevent the problems of citation-dependent ambiguity, all values in this table (Params, GFLOPs, and FPS) are generated in this study from running each of the baselines through its official public repository (or the official upstream framework explicitly used by the authors) under the unified environment described in Section 5.3. Unless a baseline repository specifies otherwise (e.g., fixed input resolution or preprocessing), inference is conducted with the same input size (256×192) and the same single-GPU measurement procedure as the proposed method. For methods that provide multiple capacities (e.g., ViTPose), the specific variant used in the current study (i.e., ViTPose-B) is reported, along with the number of parameters and the time scale relative to model size.

Table 8. Runtime and model efficiency comparison. All values (Params, GFLOPs, FPS) are *measured in this work* by executing each official implementation under the unified setup and the stated input resolution.

Method	Params (M)	GFLOPs	FPS
OpenPose [11]	26	34	48
ViTPose-B [12]	24	54	34
HRNet-W48 [13]	63	68	29
HigherHRNet [14]	67	73	26
CrowdPose Baseline [17]	42	51	36
Pose2Seg [18]	50	60	32
Limb-Augment [20]	48	59	33
OPEC-Net (GCN) [23]	52	62	30
Rethinking Visibility [24]	55	64	27
Proposed (Ours)	35	61	31

Note: When a repository enforces a different test resolution (e.g., 320×256 , 384×288 , or 368×654 for memory constraints), the enforced resolution is used and documented; otherwise, 256×192 is used for consistency.

These discrepancies between the GPLOP values listed here and those in the original papers are not unexpected and are mainly due to differences in input resolution and implementation-specific test pipelines. FLOPs are approximately linearly proportional to the number of input pixels and may also change with test-time augmentation, multi-scale inference, or post-processing. For instance, in some works we report FLOPs at 512×512

(e.g., certain settings of HigherHRNet), and in this study, FLOPs are obtained at 256×192 unless the official repository enforcement is applied. For the methods that don't report Params/GFLOPs in the papers (i.e., CrowdPose baseline, Pose2Seg, OPEC-Net, Rethinking Visibility), the numbers in Table 8 should be taken as this study's measurements from the actual code path executed, not as official author-reported numbers.

To further examine the trade-off between pose accuracy and computational efficiency, we visualize the trade-off between inference speed and performance across the datasets. Figure 10 uses a combined evaluation setup with reproduced accuracy measures (AP/PCK) and measured inference speed (FPS), enabling a direct and fair comparison of the accuracy-efficiency behavior of different methods.

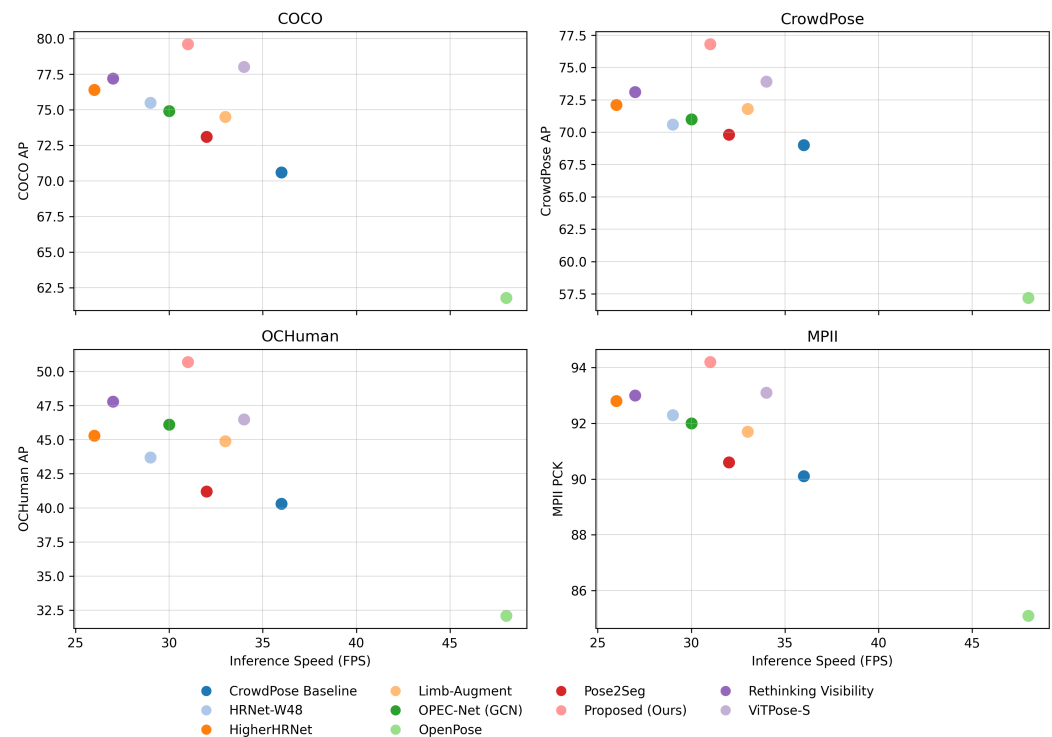


Figure 10. Accuracy vs speed trade-off For the benchmarks, COCO AP, CrowdPose AP, OCHuman AP, MPII PCK, and for the speed of inference (FPS). FPS is measured model by model under the unified setup and therefore is constant across subplots.

Inference speed is model-specific and therefore identical across all four subplots; only the accuracy metrics vary by dataset.

6.2. Ablation Studies

To assess the individual contributions of the proposed components, we conduct a series of controlled ablation experiments. To determine the contribution of each proposed component, a series of controlled ablation experiments is conducted. Specifically, we investigate the effect of (1) the GNN-based pattern restoration module, (2) the structural consistency loss terms that enforce anatomical plausibility, and (3) the choice of backbone feature extractor. Each version is tested on COCO, CrowdPose, and OCHuman in order to separate improvements in both normal and occlusion-heavy conditions. The results are summarized in Table 9.

Table 9. Ablation study on COCO, CrowdPose, and OCHuman. The GNN restoration module and structural consistency losses independently improve robustness, with the full model achieving the strongest performance across all settings.

Model Variant	COCO AP	OCHuman AP	CrowdPose AP
Baseline (Backbone only)	75.1	42.3	71.0
+ GNN Restoration	77.4	47.6	74.1
+ Structural Loss Only	76.3	45.1	72.9
+ GNN + Structural Loss (Ours)	79.6	50.7	76.8
Backbone: HRNet-W48 (Ours)	79.6	50.7	76.8
Backbone: ViTPose-S (Ours)	80.1	52.3	78.0

6.2.1. Effect of the GNN Restoration Module

Adding the GNN module results in significant improvements for all the datasets (COCO AP by +2.3 and +3.1 for CrowdPose AP). The gain is highest on OCHuman (+5.3 AP), where occluded joints must be inferred from the relational structure rather than from image evidence. This validates that by modeling the human body as a graph, the system can directly distribute contextual cues from visible joints to occluded joints, supporting the implementation of our pattern completion formulation.

6.2.2. Effect of Structural Consistency Loss

Incorporating bone-length and joint-angle constraints produces notable improvements under severe occlusion conditions, particularly on OCHuman (+2.8 AP), but yields moderate gains on COCO. These results show that geometry-aware regularization is needed in cases where there is low or ambiguous visual information. The constraints are to rule out anatomically implausible reconstructions and ensure that the restored poses are biomechanically feasible.

6.2.3. Complementarity of GNN and Structural Constraints

The best improvement is achieved by the full model, where both the GNN restoration module and the structural consistency losses are combined. The joint effect is more than the sum of individual contributions, and thus, the GNN is enhanced with the geometric priors, and at the same time, the structural constraints refine the latent predictions of the GNN. This synergy makes it important to take occlusion recovery as an inference problem in a systematic way and not as a feature enhancement step.

6.2.4. Impact of Backbone Variation

Receiving additional gains by replacing HRNet-W48 with ViTPose-S, OCHuman AP is improved from 50.7 to 52.3. This shows that the proposed graph reasoning pipeline is generalizable on any backbone architecture. CNN-based or transformer-based, and is enhanced by improved feature representations without having to change architectures. The method is therefore flexible and compatible with future advancements in the backbone.

To supplement the numerical results of ablation, The relative percentage improvement over the baseline of the backbone alone for each component is shown in Figure 11. The trend is consistent for all the datasets: The GNN-based restoration module gives the greatest individual boost, especially for OCHuman, where severe joint occlusion undermines heatmap-based inference. The structural consistency loss further contributes to gains by enforcing the plausibility of bone length and joint angles, which leads to anatomically inconsistent failures. When used together, these two modules have a synergistic effect, resulting in the greatest improvements across all evaluations. This visual evidence reinforces

the interpretation that occlusion-aware reasoning requires both relational message passing and biomechanical regularization to achieve stable pose reconstruction.

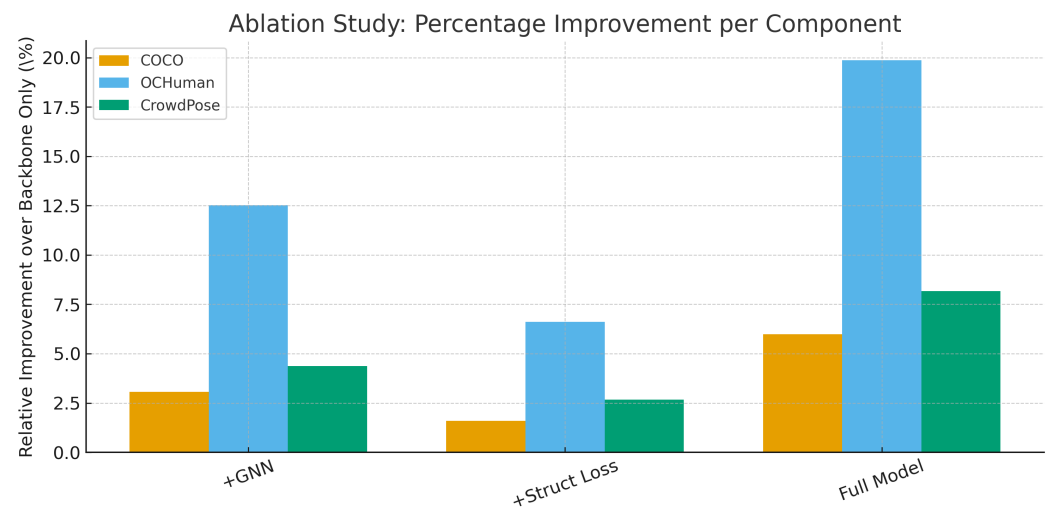


Figure 11. Relative improvement over the backbone-only baseline for each component on COCO, CrowdPose, and OCHuman. Bars show the percentage increase in AP when adding the GNN restoration module, the structural consistency loss, and the full model (GNN + structural loss). The largest relative gains appear on OCHuman, confirming that the proposed components are most beneficial under severe occlusion.

The results of ablation confirm that different parts of the framework have different roles and that each of them contributes to the overall framework: The backbone helps in extracting local appearance cues, the GNN helps in reconstructing missing structure through relational reasoning, and the structural losses help to enforce biomechanical plausibility. Their combination yields a robust and principled system that is able to reconstruct whole human poses even in the presence of severe occlusion.

Overall, the experimental results are consistent with the main hypothesis of this work that human pose estimation under occlusion is essentially a structured completion problem and not a visual detection problem. The proposed framework is more robust in extreme cases of occlusion and significantly more efficient in the case of full visibility due to the explicit modeling of joint visibility, relational information propagation on the basis of a skeletal graph, and imposing anatomical constraints. These results show that structural reasoning is not an additional refinement but a fundamental component that is needed for successful pose estimation in real-world crowded environments.

6.3. Limitations and Practical Implications

Although the proposed framework improves resilience to occlusion and crowding, certain limitations remain in regimes where visual evidence becomes insufficient to uniquely determine the pose. Under extreme visibility collapse—when only a small fraction of joints is observable—the reconstruction becomes under-constrained, and the model may produce anatomically plausible yet semantically incorrect configurations due to the lack of reliable visual grounding. Similarly, in highly dense crowds, severe interpersonal overlap can introduce ambiguous relational cues; while the proposed graph reasoning reduces joint swapping compared with appearance-only baselines, occasional identity ambiguities may still occur because the current formulation is person-centric and does not explicitly model cross-instance interactions. Moreover, the visibility likelihood $\hat{v}_i$ provides an implicit signal of reliability; the existing framework does not generate calibrated predictions of uncertainty (e.g., predictive variance) at restored keypoints, which may be required in safety-critical applications. These failure modes reflect fundamental ambiguity in monocular

pose estimation under severe occlusion rather than implementation errors and are common to structure-driven completion approaches. In practice, such cases primarily arise in situations such as lower-body occlusion in dense crowds or heavy object obstruction. Although infrequent in standard benchmarks, they remain intrinsically ambiguous because multiple anatomically valid poses can satisfy the same limited visual evidence, meaning that physically reasonable but semantically invalid reconstructions are better interpreted as manifestations of uncertainty than as model instability.

The framework currently relies on a fixed skeletal topology; incorporating hierarchical or adaptive graph structures could further enhance robustness in dense scenes or unusual pose configurations.

Moreover, strong structural priors—critical for stabilizing completion—may over-regularize rare or unconventional articulations. Graph reasoning introduces moderate computational overhead, which is acceptable on GPUs but may limit deployment on resource-constrained edge devices without further optimization. Although the proposed method achieves real-time inference on a mid-tier GPU, deploying it on low-power edge devices requires additional optimization, such as model compression or a lightweight backbone design, which is left for future work.

The main failure mode is extreme visibility collapse, in which only a small fraction of joints can be observed. In such environments, monocular pose completion is inherently ambiguous, as several physically valid configurations can yield the same revised limited evidence, leading to anatomically plausible but semantically incorrect reconstructions. Such cases (i) do not occur (except in some special benchmark distributions) but may occur in real-life scenes with high object occlusion or lower-body occlusion in a dense crowd; and (ii) this work does not involve a specific cross-dataset transfer evaluation—training on one dataset and testing on another without fine-tuning—and is an essential direction to study and will be explored in the future.

Despite these constraints, the proposed framework remains practically relevant for occlusion-dominated applications such as crowd monitoring, sports analytics, industrial safety, and human–robot interaction. These limitations naturally motivate extensions toward cross-person relational reasoning, uncertainty-adaptive structural priors, and more computationally efficient graph inference for edge deployment. More broadly, while fusion strategies vary widely across deep learning domains, the present design is intentionally aligned with human skeletal topology and should be assessed within the monocular, occlusion-robust pose estimation setting.

7. Conclusions and Future Work

This paper introduces a graph-based pattern restoration framework for human pose estimation under occlusion, formulating pose recovery as a structured inference problem over the human skeletal graph. By combining visibility-aware modeling, relational message passing, and biomechanical consistency constraints, the proposed method reconstructs missing or unreliable keypoints in crowded and visually degraded scenes while maintaining anatomically plausible poses. Extensive experiments on COCO-Keypoints, CrowdPose, OCHuman, and MPII demonstrate consistent gains in AP/OKS/PCK, improved performance under heavy occlusion, and lower structural inconsistency, validating that explicit structural reasoning complements appearance-driven pose estimators.

Future work will extend the framework along targeted directions: (i) incorporating temporal information from video to stabilize pose completion under sustained occlusion and reduce identity ambiguity; (ii) exploring cross-instance relational modeling to better handle dense crowds; and (iii) improving deployment practicality via lightweight graph reasoning and adaptive/uncertainty-aware structural priors; and (iv) extend the framework

to include video with the addition of temporal continuity (e.g., graph-temporal message passing) to even lessen ambiguity in the case of extended occlusion. These efforts aim to enhance generalization and efficiency while preserving the core structural restoration principle established in this work.

Author Contributions: Conceptualization, M.I.; methodology, M.I. and Z.U.; software, M.I. and S.Z.S.; validation, Z.U., M.I. and S.Z.S.; formal analysis, Z.U.; investigation, M.I.; data curation, M.I. and S.Z.S.; writing—original draft preparation, M.I. and S.Z.S.; writing—review and editing, Z.U.; visualization, M.I. and S.Z.S.; supervision, Z.U. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets used in this study are available at [17,27–29]. The official github repos are available at [11–14,17,18,20,23,24,30,31].

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

Nomenclature

I	Input RGB image
$k_i \in \mathbb{R}^2$	2D coordinates of joint i
$\tilde{k}_i$	Coarse or initial estimate of joint i
v_i	Visibility indicator of joint i
$\hat{v}_i$	Predicted visibility likelihood of joint i
$G = (V, E)$	Human-body skeletal graph
h_i	Node embedding of joint i
e_{ij}	Edge embedding between joints i and j
$\mathcal{L}_{\text{bone}}$	Bone-length consistency loss
$\mathcal{L}_{\text{angle}}$	Joint-angle feasibility loss
$\mathcal{L}_{\text{vis}}$	Visible-joint alignment loss
AP	Average Precision
OKS	Object Keypoint Similarity
PCK	Percentage of Correct Keypoints
GNN	Graph Neural Network

References

1. Dong, C.; Du, G. An Enhanced Real-Time Human Pose Estimation Method Based on Modified YOLOv8 Framework. *Sci. Rep.* **2024**, *14*, 8012. [CrossRef] [PubMed]
2. Singh, P.; Yadav, I.; Agrawal, P.; Singh, V.P. Evaluating the Robustness of Human Pose Estimation Models: A Comparative Analysis. In Proceedings of the 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO), Noida, India, 14–15 March 2024; pp. 1–6.
3. Kim, G.; Kim, H.; Kong, K.; Song, J.W.; Kang, S.J. Human Body-Aware Feature Extractor Using Attachable Feature Corrector for Human Pose Estimation. *IEEE Trans. Multimed.* **2022**, *25*, 5789–5799. [CrossRef]
4. Wei, F.; Hu, X. A Lightweight Attention-Driven Distillation Model for Human Pose Estimation. *Pattern Recognit. Lett.* **2024**, *185*, 247–253. [CrossRef]
5. Ji, A.; Fan, H.; Xue, X. Vision-Based Body Pose Estimation of Excavator Using a Transformer-Based Deep-Learning Model. *J. Comput. Civ. Eng.* **2025**, *39*, 04024064. [CrossRef]
6. Zhou, M.; Stofl, L.; Mathis, M.W.; Mathis, A. Rethinking Pose Estimation in Crowds: Overcoming the Detection Information Bottleneck and Ambiguity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023*; pp. 14689–14699.
7. Badiola-Bengoia, A.; Mendez-Zorrilla, A. A Systematic Review of the Application of Camera-Based Human Pose Estimation in the Field of Sport and Physical Exercise. *Sensors* **2021**, *21*, 5996. [CrossRef] [PubMed]
8. Deng, H.; Bui, M.; Navab, N.; Guibas, L.; Ilıc, S.; Birdal, T. Deep Bingham Networks: Dealing with Uncertainty and Ambiguity in Pose Estimation. *Int. J. Comput. Vis.* **2022**, *130*, 1627–1654. [CrossRef]

9. Faisal, N.R.; Fathalla, A.; Elmanakhly, D.; Salah, A. Reported Challenges in Deep Learning-Based Human Pose Estimation: A Systematic Review. *IEEE Access* **2025**, *13*, 80520–80539. [CrossRef]
10. Zheng, C.; Wu, W.; Chen, C.; Yang, T.; Zhu, S.; Shen, J.; Kehtarnavaz, N.; Shah, M. Deep Learning-Based Human Pose Estimation: A Survey. *ACM Comput. Surv.* **2023**, *56*, 1–37. [CrossRef]
11. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.E.; Sheikh, Y. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 172–186. [CrossRef] [PubMed]
12. Xu, Y.; Zhang, J.; Zhang, Q.; Tao, D. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 38571–38584.
13. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019*; pp. 5693–5703.
14. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020*; pp. 5386–5395.
15. Zhang, Z.; Shin, S.Y. Two-Dimensional Human Pose Estimation with Deep Learning: A Review. *Appl. Sci.* **2025**, *15*, 7344. [CrossRef]
16. Wei, L.; Yu, X.; Liu, Z. Human Pose Estimation in Crowded Scenes Using Keypoint Likelihood Variance Reduction. *Displays* **2024**, *83*, 102675. [CrossRef]
17. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019*; pp. 10863–10872.
18. Golda, T.; Kalb, T.; Schumann, A.; Beyerer, J. Human Pose Estimation for Real-World Crowded Scenarios. In *Proceedings of the 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Taipei, Taiwan, 18–21 September 2019*; pp. 1–8.
19. Cheng, Y.; Yang, B.; Wang, B.; Yan, W.; Tan, R.T. Occlusion-Aware Networks for 3D Human Pose Estimation in Video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019*; pp. 723–732.
20. Han, G.; Song, C.; Wang, S.; Wang, H.; Chen, E.; Wang, G. Occluded Human Pose Estimation Based on Limb Joint Augmentation. *Neural Comput. Appl.* **2025**, *37*, 1241–1253. [CrossRef]
21. Bai, X.; Wei, X.; Wang, Z.; Zhang, M. CONet: Crowd and Occlusion-Aware Network for Occluded Human Pose Estimation. *Neural Netw.* **2024**, *172*, 106109. [CrossRef] [PubMed]
22. El Kaid, A.; Baïna, K. A Systematic Review of Recent Deep Learning Approaches for 3D Human Pose Estimation. *J. Imaging* **2023**, *9*, 275. [CrossRef] [PubMed]
23. Qiu, L.; Zhang, X.; Li, Y.; Li, G.; Wu, X.; Xiong, Z.; Han, X.; Cui, S. Peeking into Occluded Joints: A Novel Framework for Crowd Pose Estimation. In *Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020*; pp. 488–504.
24. Sun, P.; Gu, K.; Wang, Y.; Yang, L.; Yao, A. Rethinking Visibility in Human Pose Estimation: Occluded Pose Reasoning via Transformers. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2024*; pp. 5903–5912.
25. Liu, J.; Sun, W.; Yang, H.; Zeng, Z.; Liu, C.; Zheng, J.; Liu, X.; Rahmani, H.; Sebe, N.; Mian, A. Deep Learning-Based Object Pose Estimation: A Comprehensive Survey. *arXiv* **2024**, arXiv:2405.07801. [CrossRef]
26. Guan, J.; Hao, Y.; Wu, Q.; Li, S.; Fang, Y. A Survey of 6DoF Object Pose Estimation Methods for Different Application Scenarios. *Sensors* **2024**, *24*, 1076. [CrossRef] [PubMed]
27. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *Computer Vision—ECCV 2014; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2014; Volume 8693*, pp. 740–755.
28. Zhang, S.H.; Li, R.; Dong, X.; Rosin, P.; Cai, Z.; Han, X.; Yang, D.; Huang, H.; Hu, S.M. Pose2Seg: Detection-Free Human Instance Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019*; pp. 889–898.
29. Andriluka, M.; Pishchulin, L.; Gehler, P.; Schiele, B. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014*; pp. 3686–3693.
30. Available online: <https://github.com/leoxiaobin/deep-high-resolution-net.pytorch> (accessed on 26 January 2026).
31. Available online: <https://github.com/pengzhansun/Occluded-Pose-Reasoning> (accessed on 2 February 2026).

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.