A Graph Optimization Framework for Occlusion-Aware Human Pose Estimation

Alfaz Uddin Emon | C213057[1], Shirajul Islam Shakur | C213040[1], Md. Mohaiminul Islam | C213067[1]

[a]*Department of Computer Science,, Chittagong, Bangladesh*

Abstract

Human pose estimation is widely used in applications ranging from sports analysis to healthcare, where understanding body movement is critical. However, this becomes difficult when the subject is partially occluded, leading to missing or noisy pose data. This paper proposes a graph-based optimization method using the G2O framework to improve pose estimation under occlusion. The method applies bone length, joint angle, and temporal continuity constraints to refine 2D keypoints extracted using OpenPose. Results show improved anatomical plausibility and smooth motion continuity, making it suitable for real-world applications such as physical therapy, elder care, and human-computer interaction.

Keywords: Pose Estimation, Graph Optimization, Occlusion, G2O, Human Motion, Biomechanics

1. Introduction

Human pose estimation has emerged as a crucial research area due to its diverse applications in fields like healthcare, sports science, and human-computer interaction. Understanding body movements plays a vital role in these domains, particularly when working with elderly individuals, athletes, or patients undergoing physical therapy. The process entails identifying and mapping keypoints such as joints and limbs of the human body from a single image or sequence of images in 2D or 3D space.

Challenges arise when body parts are occluded or the background becomes obscured, often resulting in incomplete or inaccurate estimations. Earlier methods for pose estimation relied on hand-engineered features and rule-based models, which often struggle in complex environments involving occlusion, poor lighting conditions, or unusual body postures.

The advent of deep learning has significantly improved this field. Techniques incorporating Convolutional Neural Networks (CNNs) have boosted the precision of human keypoint detection and mapping. Models such as OpenPose, MobileNetV2, ResNet50, VGG16, and VGG19 are among the top-performing frameworks, excelling in scenarios where the human figure is clear and unobstructed.

Despite these advancements, these models still face difficulties under occlusion, struggling to maintain high accuracy when parts of the body are not visible. Estimating poses in such conditions is particularly complex, as the system must predict missing keypoints based on visible ones and prior anatomical knowledge. The task becomes even harder in real-world settings where occlusions occur unpredictably, either from external objects or the subject's movements. This issue is critical in health monitoring and rehabilitation—especially for elderly patients—where accurate motion tracking is essential for proper evaluation. Therefore, a robust solution is necessary to effectively address these challenges.Estimating human pose in situations where parts of the body are partially or fully occluded remains a significant challenge. Although modern deep learning models perform well in controlled and unobstructed environments, their effectiveness drops notably when key body regions are hidden. This decline in accuracy poses a serious issue in domains like healthcare, where precise movement tracking is essential for effective diagnosis, therapy, and patient observation. The motivation behind this work is the increased demand in many movement-based industries for accurate human pose estimation. Be it for physical therapy, training in sports, or elder care, gait analysis, among other types of body movement tracking, is priceless. While current models have improved significantly, their performance still has its shortcomings when dealing with occluded or partly visible parts of the body, which introduces errors in pose estimation and subsequent analyses. A more adaptive and robust model that could perform well under challenging conditions would fill an important gap in the current ecology of pose estimation technologies. Traditional methods also rely on neural network based architectures which also proves to be extremely costly. And thus, reducing that cost by relying on other methods is also an important aspect of our work. The objective of this study is to implement a methodology that will improve human pose estimation under occlusion through graph optimization techniques. Specifically, in this work, we intend to: Use a kinematic model of the human body represented with the help of a graph, where vertices represent joint keypoints and edges are defined based on the anatomical connections between those keypoints. The G2O framework is then used to enforce constraints on bone lengths and joint angles, ensuring it is biomechanically plausible and maintains temporal continuity to make sure movements don't seem out of place in pose estimation.

2.Previous Work

Human pose estimation has become a highly sought-after topic in computer vision due to its broad range of applications, including healthcare, sports analytics, and virtual or augmented reality. Despite these advancements, occlusion remains one of the most significant challenges in this field. State-of-the-art models often struggle to accurately identify key points when parts of the body are hidden, and even when they succeed, the process tends to be computationally intensive. Numerous research efforts have explored different strategies to overcome this issue, from real-time pose estimation frameworks to a variety of optimization techniques. This section provides a brief overview of the contributions and limitations found in previous studies conducted by other researchers.

OpenPose: Multi-Person Pose Estimation Framework

Cao et al. introduced OpenPose, a real-time multi-person pose estimation framework capable of detecting human keypoints and associating them to form a skeleton for each individual in a frame. The core innovation of OpenPose lies in the use of Part Affinity Fields (PAFs), which model spatial relationships between body parts and enable precise skeletal estimations, even when multiple people are present in a scene. OpenPose has shown substantial success in achieving real-time processing speeds and has found wide application in domains such as sports, healthcare, and entertainment. Its open-source nature has encouraged further research and development. However, its reliance on visual data alone leads to inaccuracies in cases of occlusion, and it does not incorporate prior anatomical knowledge that could aid in resolving such challenges. In this thesis, OpenPose is used to extract keypoints from the recorded dataset. The proposed methodology builds upon these keypoints, refining occluded joints using graph-based optimization techniques.

G2O-Pose: Real-Time Monocular 3D Human Pose Estimation Based on General Graph Optimization

un et al. (2022) presented a lightweight and efficient method for real-time 3D human pose estimation from monocular images. Their approach models the human skeleton as a graph and applies general graph optimization under spatial and temporal constraints. The method addresses the problem of depth ambiguity by reconstructing 3D poses from 2D keypoints over time. Key contributions of this study include a low-error method for recovering 3D bone proportions from 2D data, a 95.4% accuracy rate in classifying human body orientation using 2D joints, and the use of a heuristic algorithm to correct reverse joint rotations. While this method is efficient, its accuracy decreases in the presence of significant depth changes across frames, and its performance is highly dependent on the accuracy of the initial 2D keypoint detection. This study informs our thesis by demonstrating how graph optimization can effectively address depth ambiguity and supports our decision to use G2O for refining pose estimations

G2O-Based Optimization for Monocular Human Pose Estimation

Zhou et al. (2022) employed the General Graph Optimization (G2O) framework to address challenges in monocular pose estimation. Their work represents the human body as a kinematic graph, where vertices correspond to joints and edges denote bones, capturing the skeletal structure effectively. The study focuses on projecting 2D keypoints onto 3D space while enforcing human anatomical constraints such as consistent bone lengths and realistic joint angles. This not only ensures anatomical plausibility but also supports real-time applications with high accuracy. Our thesis builds upon Zhou et al.'s findings by applying similar constraints within the 2D domain. In addition, the proposed method introduces temporal continuity constraints, which are essential for handling sequential data such as video frames.

Pose Estimation Under Occlusion: Reasoning-Based Approaches

Sun et al. (2024) addressed the challenge of occlusion in human pose estimation by introducing reasoning mechanisms capable of predicting occluded keypoints using spatial and temporal patterns from visible joints. Their occlusion-aware model demonstrated improved performance on occlusion-rich datasets such as COCO and MPII, validating the effectiveness of this approach. However, the model's performance is limited by its dependency on the presence of adjacent visible keypoints, making it less reliable under severe occlusion. Additionally, it does not incorporate biomechanical plausibility, which limits the anatomical realism of inferred poses. This thesis extends the reasoning-based approach by integrating it with graph-based optimization and biomechanical constraints to enhance keypoint refinement under occlusion.

General Graph Optimization Framework

Kuemmerle et al. introduced G2O, a versatile and efficient framework for solving graph-based optimization problems in computer vision and robotics. The framework supports a range of solvers and robust kernels, enabling it to handle noisy data and outliers effectively. Its modularity allows application across diverse problems, including Simultaneous Localization and Mapping (SLAM) and pose estimation. However, when applied to human pose estimation, G2O requires additional constraints to ensure biomechanical accuracy and cannot inherently model temporal sequences. This thesis customizes G2O by incorporating both anatomical and temporal constraints, enabling improved pose estimation performance under occlusion and in sequential data. [twocolumn]article amsmath,amssymb enumitem lipsum

3. Methods and Experiments

In this paper, we introduce AT-G2O (Anatomical and Temporal Graph Optimization for Obstacle-Affected Pose Estimation), a novel algorithm designed to refine 2D human pose estimations by integrating anatomical and temporal constraints within a graph optimization framework. While existing 2D pose estimation methods often suffer from inconsistencies due to occlusions, noise, or missing data, AT-G2O addresses these challenges by leveraging the structural relationships inherent in the human skeleton and the temporal continuity of motion sequences.

Algorithm 1: Anatomical and Temporal Graph Optimization for Obstacle-Affected Pose Estimation

```
Input: OpenPose keypoint JSON files for multiple persons and sequences (extracted from video).
Output: Optimized keypoint JSON files for each sequence (excluding the reference).
Step1: For Each Person Folder (p1, p2, p3)
1.1  Find all sequence folders (e.g., p11, p12, p13)
1.2  For each sequence folder:
       – Load all keypoint JSONs into a sequence array
       – Filter out low-confidence keypoints (set to NaN if confidence < 0.3)
       – Interpolate missing/low-confidence keypoints using linear interpolation
1.3  Set the first sequence (pX1) as reference:
       – Compute reference bone lengths (median length for each skeleton connection)
Step2: For Each Non-Reference Sequence
2.1  Initialize a g2o optimizer
2.2  For each frame:
       – Create a vertex for each keypoint (fixed if confidence > 0.5)
2.3  Add temporal edges:
       – For each keypoint, connect consecutive frames
       – Use reference sequence movement and confidence to set edge weights
2.4  Add skeleton (anatomical) constraints:
       – For each skeleton connection, add an edge to maintain reference bone length (soft constraint)
2.5  Run two-stage optimization:
       – Stage 1, Initial optimization (10 iterations)
       – Stage 2, Refinement (20 iterations)
2.6  Extract optimized keypoints from the optimizer
2.7  Apply temporal smoothing (Savitzky-Golay filter) to reduce jitter
2.8  Save optimized keypoints as new JSON files in an output folder
Step3: Repeat for all persons and sequences
```

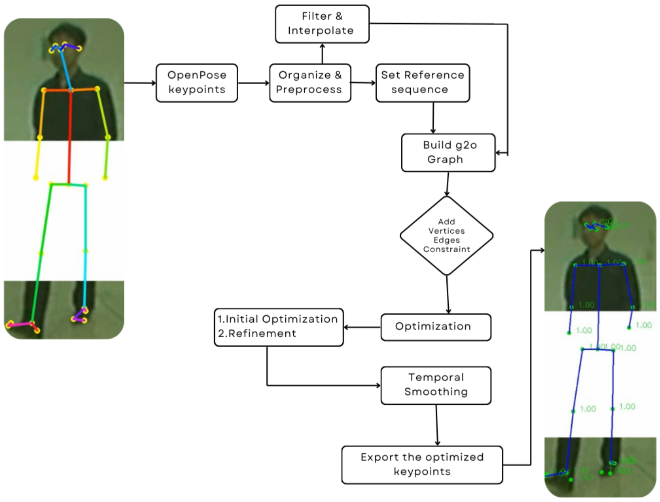

Figure 1: Flowchart of the AT-G2O Optimization Framework

Per-person keypoint preprocessing operates on OpenPose outputs to produce refined, temporally consistent keypoints. Each dataset contains subject folders $P = \{p_1, p_2, \ldots, p_n\}$, each with sequences $S = \{S_1, S_2, \ldots, S_m\}$ and frames $F = \{F_1, \ldots, F_T\}$. Each frame F_t contains J keypoints $k_{t,j} = (x_{t,j}, y_{t,j}, c_{t,j})$, forming a tensor $K^{(i)} \in \mathbb{R}^{T\times J\times 3}$. Low-confidence keypoints are filtered: if $c_{t,j} < \tau$ (with threshold $\tau = 0.3$), then $x_{t,j} = \text{NaN}$, $y_{t,j} = \text{NaN}$. Missing keypoints are interpolated linearly using valid neighbors in time, with fallback to zero-order hold or spline for edge frames. The output is $K^{(i)} = [\hat{x}_{t,j}, \hat{y}_{t,j}] \in \mathbb{R}^{T\times J\times 2}$, a dense, smoothed trajectory for each joint. Interpolation ensures temporal consistency; filtering ensures confidence-aware data quality. These keypoints are suitable for anatomical constraints and further optimization.

The first available sequence S_1 of each subject is set as reference to compute canonical bone lengths. Bones are defined as connections $\mathcal{B} = \{(j_a, j_b)\}$ between keypoints, e.g., (shoulder, elbow). Bone lengths per frame are computed as

$$L_t = \sqrt{(x_{t,a} - x_{t,b})^2 + (y_{t,a} - y_{t,b})^2}.$$

Reference bone lengths are set as median values across frames for stability.

Step 2: Optimization of Non-Reference Sequences

The optimizer minimizes pose graph error:

$$\min_{v} \sum_{i} \|f_i(v)\|^2,$$

where v are keypoint positions and f_i are constraint residuals. Each keypoint $V_{t,k} \in \mathbb{R}^2$ is a graph vertex; fix $V_{t,k}$ if $c_{t,k} > 0.5$. Temporal edges enforce consistency: residuals connect $V_{t,k}$ and $V_{t+1,k}$ to encourage smooth motion. Anatomical edges preserve distances using reference bone lengths between connected joints. Skeleton constraints enforce

$$\|V_{t,a} - V_{t,b}\| \approx L_{ab}^{ref}$$

for all $(j_a, j_b) \in \mathcal{B}$.

Two-stage optimization is performed: - Stage 1 (10 iterations) for coarse alignment, - Stage 2 (20 iterations) for fine tuning.

Optimized keypoints are

$$v_{t,k} = \arg\min_{v_{t,k}} (E_{\text{temporal}} + E_{\text{skeleton}}).$$

Savitzky–Golay filter is applied:

$$v_{t,k}^{smooth} = \text{SGFilter}(v_{t,k}, \text{window\_length} = 7, \text{polyorder} = 3).$$

Optimized and smoothed keypoints are saved per frame in JSON with positions and optionally updated confidence. The pipeline is repeated for all persons and sequences to ensure scalability and consistency in multi-subject datasets. Final output is high-quality, temporally consistent, anatomically plausible keypoints for downstream tasks.

*3.1 **Dataset***

The dataset used in this study consists of video recordings of participants walking forward, captured from three different camera angles. The dataset involves one occurrence of occlusion in each angle. Out of the 6 camera angles, occlusions occur for 15–25 frames in each sequence, caused by external objects.

Structure of the Dataset. The dataset is hierarchically organized as follows:

- Participant Folders**:** Each participant has a dedicated folder, e.g., `p1`. Each participant folder contains 6 subfolders corresponding to different camera views (`000 (ref)`, `000`, `018`, `036`, `162`, `180`) for the participant.

- Each sequence contains approximately 200 frames, where each frame corresponds to a JSON file with 2D pose keypoints extracted by OpenPose.

- Each frame is represented by a set of keypoints: Here, $N = 25$ is the total number of joints (as defined by the OpenPose body 25 model). (X_i, Y_i) are the 2D coordinates of joint i in image space. $C_i \in [0, 1]$ is the confidence score of joint i, indicating the reliability of the detection. t is the frame index, and T is the total number of frames in the sequence.

Challenges in the Dataset

1. Occlusion**:** Certain keypoints are either missing ($c_i = 0$) or have very low confidence ($c_i < 0.3$).
2. Missing Frames**:** Entire frames may lack valid keypoints due to severe occlusions.

Several preprocessing steps are applied to handle occlusion, noise, and missing data:

1. Confidence Filtering**:** Keypoints with $c_i < 0.3$ are marked as missing.
2. Interpolation of Missing Keypoints**:** Missing keypoints are interpolated using linear interpolation across valid frames.

*3.2 **Workflow Diagram of Proposed Methodology***

The proposed methodology is based on graph optimization, where the human body can be treated as a kinematic graph. It is also a 2D-keypoint refinement framework, which applies human anatomy and temporal consistency constraints to deal with occlusions and noise.

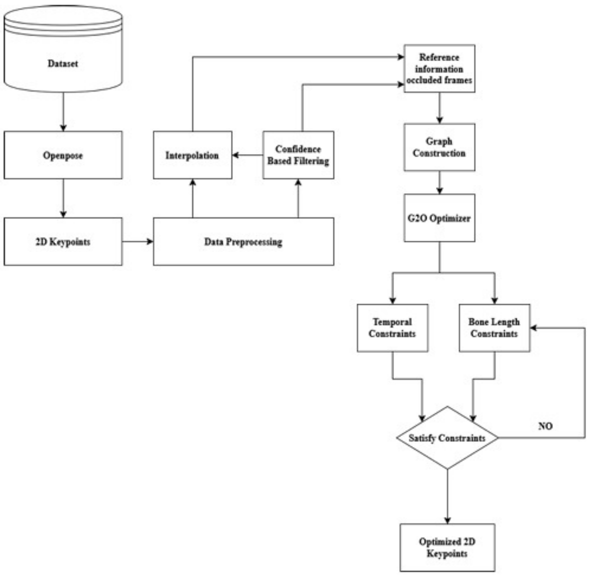

Figure 3.2: Workflow Diagram of Proposed Methodology

Reference Length Calculation: The reference sequence is used to compute bone lengths that serve as anatomical constraints during optimization. For each connected keypoint pair i and j, the reference length L_{ref} is calculated using the Euclidean distance:

$$L_{\text{ref}} = \|x_i - x_j\| = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2},$$

where $x_i = (x_i, y_i)$ and $x_j = (x_j, y_j)$ denote the 2D coordinates of joints i and j, respectively. **Kinematic Graph Representation:** The human pose is modeled as a graph $G = (V, E)$, where V is the set of vertices representing joint positions and E is the set of edges representing bones, i.e., connections between joints. Each vertex holds a 2D joint position (x_i, y_i) in a given frame, capturing the spatial configuration of the pose. **Objective Function for Optimization:** The overall objective function E_{total} to be minimized ensures anatomical consistency, temporal smoothness, and confidence-based refinement. It is formulated as:

$$E_{\text{total}} = \lambda_1 E_{\text{anatomy}} + \lambda_2 E_{\text{temporal}} + \lambda_3 E_{\text{confidence}},$$

where E_{anatomy} enforces realistic bone lengths, E_{temporal} ensures smooth motion between frames, and $E_{\text{confidence}}$ adapts the keypoints according to detection confidence. The coefficients λ_1, λ_2, and λ_3 balance these constraints according to their relative importance. **Optimization Constraints Using G2O:** Optimization is performed using the G2O framework for graph-based optimization. The graph comprises $T \times N$ vertices representing N joints over T frames. Edges encode anatomical constraints within frames and temporal constraints across frames. Initial vertex positions are obtained from OpenPose detections. **Vertex Initialization:** Each vertex stores a 2D joint position along with a confidence score. Vertices corresponding to joints detected with high confidence are fixed during optimization, while those with low confidence are allowed to be adjusted to better satisfy the overall constraints. **Bone Length Constraints:** Anatomical edges connect joints within the same frame and enforce that the bone lengths remain consistent with the reference bone lengths L_{ref}. This preserves anatomical plausibility by penalizing deviations from expected limb lengths. **Temporal Constraints:** Temporal edges connect the same joint across consecutive frames, promoting temporal smoothness in the estimated joint trajectories. This encourages physically plausible motion by discouraging sudden jumps or jitter. **Optimization Algorithm:** The Levenberg–Marquardt algorithm is employed to minimize the total energy function. This algorithm iteratively updates vertex positions using a combination of gradient descent and Gauss–Newton steps. Each iteration computes the Jacobian matrix of the error function and applies corrections to vertex positions to reduce the residuals. **Handling Outliers:** To increase robustness against outliers, a Huber kernel is incorporated into the optimization. The Huber loss function is defined as:

$$\rho(r) = \begin{cases} \frac{1}{2}r^2 & \textbf{if } |r| \leq \delta \\ \delta(|r| - \frac{1}{2}\delta) & \textbf{if } |r| > \delta \end{cases}$$

where δ is a threshold parameter controlling the trade-off between sensitivity and robustness. This loss reduces the influence of large residuals caused by noisy or incorrect keypoint detections. **Two-Stage Optimization Process:** The optimization proceeds in two stages. The initial stage performs coarse alignment, roughly adjusting the keypoints to satisfy the anatomical and temporal constraints. The refinement stage further fine-tunes the vertex positions to minimize motion inconsistencies while maintaining anatomical correctness. **Post-Optimization Step:** After optimization, the refined joint positions are extracted from the graph and can be used for downstream tasks such as visualization or higher-level analysis. This results in anatomically plausible and temporally consistent pose trajectories, improving accuracy especially in frames affected by occlusion or noise.

4.Results and Discussion

This chapter presents the results of the proposed G2O-based optimization framework applied to human pose estimation. The experimental results are evaluated both quantitatively and qualitatively to demonstrate the effectiveness of the optimization methodology in addressing occlusions and improving pose accuracy. Various metrics, including Average Confidence Scores, Missing Keypoint Ratios, and Temporal Consistency, are analyzed for both OpenPose (baseline) and the optimized G2O framework. Comparisons are conducted for non-occluded and occluded frame ranges, and the findings are summarized with visual and numerical evidence.

In this study on human pose estimation under occlusion, we evaluated the performance of the G2O-based optimization framework in comparison to OpenPose, a widely used pose detection method. The evaluation was conducted on sequences containing occlusions in multiple camera views. Key metrics such as Average Confidence Score, Missing Keypoint Ratio, and Temporal Consistency were used to assess the effectiveness of the approaches.

***4.1 Experimental Analysis*. :** OpenPose is able to perfectly detect the pose when there is no occlusion as shown below. Since we modify the dataset, we mainly depend on visual comparison, as there is no way to compare our results with ground truth or validation data.

***4.2 Performance Analysis* :.** This chart illustrates the Bone Length Consistency Score (%) comparison between OpenPose and G2O frameworks across different camera views (e.g., p12, p13, etc.).

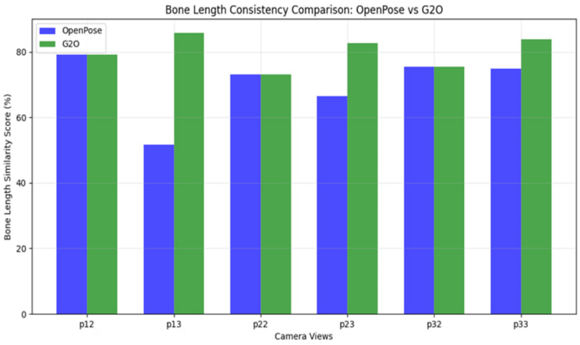

Figure 4.1: Bone Length Consistency Comparison

The missing keypoint ratios are visualized in the figure below, showcasing the effectiveness of G2O in recovering missing information during occluded frames. In Figure 4.2, after processing 5 consecutive frames, G2O significantly reduced the average number of missing keypoints across all six views. Occlusion-heavy views like p23 and p33 benefited the most, showing visible recovery in keypoint estimation. Temporal consistency across frames helped interpolate and refine previously

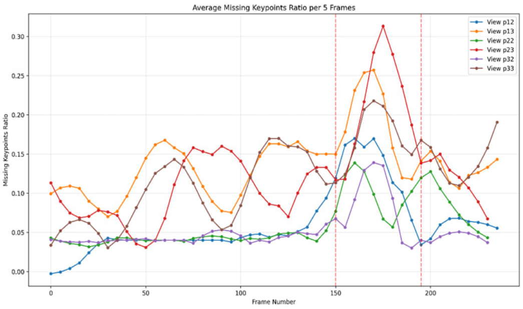

Figure 2: *

Figure 4.2: Missing Keypoint Ratio.

undetected joints. This multi-frame refinement confirms that G2O effectively leverages sequential data to improve pose reliability under occlusion.

5. Conclusion and Future Work

5.1 Conclusion. This study addressed the persistent challenge of human pose estimation under occlusion by leveraging a G2O-based optimization framework. The framework demonstrated significant improvements over OpenPose in scenarios involving occlusion, as evidenced by higher bone length similarity, lower missing keypoint ratios, and smoother temporal trajectories. By integrating anatomical and temporal constraints, the G2O framework refined keypoint predictions and maintained anatomical fidelity, even in the presence of significant occlusion. These findings underscore the potential of graph-based optimization in advancing the field of pose estimation. A major contribution of this work lies in the creation of a novel dataset tailored to the study's requirements. The dataset was recorded specifically for this research and includes three participants walking from three different camera views, with intentional occlusions introduced in two of these views for 15–25 frames. This dataset, processed through OpenPose for keypoint extraction, provided a robust testing ground for evaluating the G2O framework's performance. The outcomes highlight the framework's ability to address occlusion effectively, marking a significant step forward in pose estimation research.

5.2 Future Works. Future work will focus on overcoming the identified limitations and extending the applicability of the G2O framework. One promising direction is to optimize the computational efficiency of the framework, enabling real-time pose estimation without sacrificing accuracy. Expanding the dataset to include a wider range of activities, more participants, and diverse occlusion patterns would further validate the framework's robustness. Additionally, integrating the G2O framework with 3D pose estimation methods would allow for a deeper understanding of human motion and enhance its applicability in domains such as virtual reality and biomechanical analysis. Combining the G2O framework with state-of-the-art deep learning techniques, such as transformers, could also yield a hybrid approach that leverages the strengths of both optimization and learning-based methods. These advancements could significantly enhance the framework's performance and broaden its scope of application.

References

[1] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, Jan. 2021.

[2] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *International Conference on Learning Representations (ICLR)*, 2021.

[3] X. Zhou, M. Zhu, Z. Deng, and X. Yuan, "G2O-based optimization for monocular human pose estimation," *IEEE Transactions on Image Processing*, vol. 31, pp. 1235–1248, 2022.

[4] F. Wang, M. Cheng, X. Liu, and J. Luo, "Occlusion-aware human pose estimation using graph-based optimization," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 401–412.

[5] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, "G2O: A general framework for graph optimization," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2011, pp. 3607–3613.

[6] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3D human pose estimation," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 2659–2668.

[7] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[8] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 3686–3693.

[9] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4724–4732.

[10] D. Mehta et al., "Xnect: Real-time multi-person 3D motion capture with a single RGB camera," *ACM Transactions on Graphics*, vol. 39, no. 4, pp. 82:1–82:17, 2020.

[11] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei, "Towards 3D human pose estimation in the wild: A weakly-supervised approach," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017.

[12] C. Wang, Y. Wang, Z. Lin, A. L. Yuille, and W. Gao, "Detecting human-object interactions with part-aware attention," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018.

[13] S. Johnson and M. Everingham, "Learning effective human pose estimation from inaccurate annotation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1465–1472.

[14] M. J. Black and A. D. Jepson, "Eigentracking: Robust matching and tracking of articulated objects using a view-based representation," *International Journal of Computer Vision*, vol. 26, no. 1, pp. 63–84, 1998.

[15] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 90–126, 2006.