Journal of Applied Automation Technologies

Research Paper

Occlusion-Robust Cross-Modal 3D Human Pose Estimation via Adaptive RGB-D Fusion and Hierarchical Graph Reasoning

İlknur Mağden[1,*], Gülcan Kahraman[1] and Nuran Şahbaz[1]

[1] Faculty of Computer Science, Istanbul University, Istanbul, 34452, Turkey
*Corresponding author: ilknur.ma@istanbul.edu.tr

Abstract. Three-dimensional (3D) human pose estimation is crucial for intelligent systems involved in human–computer interaction, motion analysis, and clinical assessment, yet existing deep learning models struggle with occlusion, sensor noise, and perceptual limitations in single-modal settings. To address these issues, we propose a unified cross-modal 3D pose estimation framework that integrates RGB and depth data through a novel adaptive fusion module, selectively combining key features from both modalities at the intermediate stage to preserve complementary information. The fused features are represented as a skeleton map and processed by a hierarchical graph convolutional network, effectively capturing both local and global structural dependencies. Extensive experiments on the Human3.6M and MSRA datasets validate the effectiveness of our approach: the proposed method achieves a mean per-joint position error (MPJPE) of 27.6 mm on Human3.6M, substantially outperforming the single-modal baseline (34.1 mm) and previous cross-modal methods (30.8 mm), while the Percentage of Correct Keypoints (PCK) at 50 mm reaches 94.5%. These results demonstrate that the proposed framework significantly improves estimation accuracy and robustness under occlusion and noisy input, while maintaining high inference efficiency and manageable model complexity. Our findings highlight the importance of joint multimodal fusion and hierarchical structural reasoning for advancing robust, scalable 3D pose perception in unconstrained environments.

Keywords: *Cross-Modal Learning, 3D Pose Estimation, Graph Convolutional Networks, Occlusion Robustness*

Received on 07 October 2023, Accepted on 12 March 2024, Published on 19 March 2024

Introduction

3D human pose estimation is very critical. In intelligent human-computer interaction and other aspects, rely on obtaining the state of body joints to allow machines to interpret human behaviors [1,2]. Early deep learning using monocular RGB/depth data has made progress, but single-modal systems are limited by occlusion and other restrictions, and challenges intensify in scenarios such as rapid movement, when visual features are missing or depth cues are correlated due to sensor problems [3,4]. To overcome the bottleneck, combine RGB and depth information, and utilize their advantages [5,6]. Most multimodal frameworks adopt simple concatenation or shallow fusion methods, and do not make full use of body structure relationships [7]. Recent works use attention/gating methods, but lack explicit structural models, and perform poorly in real-world problems [8,9]. Aiming at the pose data structure, graph convolutional networks (GCNs) treat joints as nodes to model dependencies for regression [10]. However, research on fusing heterogeneous modalities in graph-based architectures is insufficient. This paper constructs a new cross-modal 3D pose estimation framework. This framework deeply integrates multi-modal features through adaptive fusion methods and hierarchical graph convolutional network architecture. The contributions include three parts: first, introduce a cross-modal fusion module to selectively aggregate discriminative cues from RGB and depth streams. Second, design a hierarchical graph convolutional network to perform structured, context-aware end-to-end transmission of information between local and global joint relationships. Third, comprehensive experimental verification that our applied method is superior to existing baselines in estimation accuracy, robustness to environmental differences and

https://doi.org/10.64972/jaat.2024v2.254p10e:138-148

computational efficiency. These advantages lay a solid foundation for deploying robust and practical 3D human-centric perception systems in unconstrained scenarios.

Related Work

Monomodal and Multimodal Pose Estimation

There are two lines of progress in human pose estimation. One is the line from 2D to 3D representation, and the other is the line from unimodal data to multimodal data. Early work used classical computer vision methods to extract handcrafted features from a single RGB image to infer the coordinates of 2D/3D joints, commonly using graphical structures or part-based models [13]. However, these systems are often hindered by occlusion and complex scenes, and are not reliable under real conditions. DL appears in this field. The unimodal CNN method performs end-to-end mapping from raw images to pose coordinates, with a certain degree of improvement in accuracy and robustness [12]. Nevertheless, unimodal models have limitations such as depth ambiguity, loss of spatial context, and vulnerability to challenging lighting or cluttered backgrounds. Recent research integrates rich modalities, especially RGB-depth fusion [13]. Early fusion merges inputs or early feature information at the initial stage to utilize all clues [14]. However, these facial modalities have the problem of over-dominance: features from one sensor overwhelm others, making the joint representation poor. Late fusion delays combination until after high-level semantic features are extracted separately, then aligns or merges them in deeper network layers. This alleviates interference to a certain extent, but often skips the low-level complementarity required for fine-grained posture discrimination [15]. Now, more advanced paradigms use attention for dynamic and context-sensitive fusion. Networks are pushed to adaptively weight inputs, thereby highlighting more reliable modalities in spatio-temporal contexts [16]. Although conceptually strong, these architectures are often "shallow", only performing fusion at certain layers, or failing to fully exploit interactions between modalities. Empirical studies show that the lack of structural awareness—without modeling body kinematic links and geometric hierarchies—also limits accuracy and adaptability to occlusions, noise or viewpoint changes [17,18]. So multimodal fusion is promising, but there is still much room for improvement in improving joint analysis and propagation of cross-modal cues.

Graph Convolutional Networks for Poses

Besides the pixel grid method, Graph Convolutional Networks (GCNs) are an alternative way for human pose representation. Build a human skeleton into a graph (treat joints as nodes, treat bones/connections as edges), Graph Convolutional Networks model the dependence on spatial relationships for 3D pose estimation [19]. This architecture adapts to the articulated structure of the human body, and can clearly reason about the interaction of local and global joints. Some influential works show that graph convolutional networks (GCNs) can help make better use of topological priors, capture short-range (like limb coherence) and long-range (as the collaboration of joints in complex movements). Improving the ordinary GCN framework, involving edge attention, adaptive adjacency matrix or hierarchical pooling, has certain value in flexibility and scalability for processing large-scale or multi-person scenarios. Top-tier graph convolutional network (GCN) systems have shortcomings. Most existing models are single-modal inputs and cannot natively handle heterogeneous information flows. Although GCN can encode local connections and so on, it lacks methods for adaptively transmitting multi-modal cues. The global-to-local information flow is limited by the graph structure or fusion methods. This gap prompts people to rethink graph construction and feature integration for cross-modal and other pose regressions.

Summary and Research Gap

A review of the literature found two intertwined unresolved limitations. On the one hand, multimodal fusion frameworks can mine diverse sensory cues, but usually lack the overall small details of deep context-aware bodily information integration. On the other hand, existing graph convolutional network (GCN)-based models are good at structural modeling, but ignore the differences in real sensory data, and also lack attention to the modal-aware propagation of pose cues. Connecting two cutting-edge researches requires methods. The methods take advantage of graph reasoning, and also lie in adaptive hierarchical cross-modal feature fusion of network embedding. Designing such a system requires new architectural principles, so that different modalities contribute complementary dependent context signals during processing, and keep the signals in the skeleton

https://doi.org/10.64972/jaat.2024v2.254p10e:138-148

graph and dynamically route them. This study addresses these challenges, and advances 3D human pose estimation by using a principled deep integrated multimodal graph-based framework.

Methodology

Overall Framework

The core of the method is to conduct joint 3D human pose estimation through a unified pipeline, obtain information from synchronized RGB and depth, and make use of their complementary cues. As shown in Figure 1, the system includes 4 stages: data collection, feature extraction for specific modalities, cross-modal feature fusion, and pose inference based on hierarchical graph convolution networks. Initially, raw RGB and depth images are acquired from multi-sensor devices or single RGB-D camera systems. Dual-modal inputs have photometric details and geometric structures, which provide a richer foundation for human pose analysis compared to single-modal data streams [21]. Next, different convolutional backbone networks are used to extract intermediate-level features from each modality, so that the network can retain the features of specific modalities, such as the color and texture of RGB and the spatial depth cues of depth maps, and use their domain-related content as much as possible. The key point is that these features are separated at an early stage, avoiding the sparsity of early information. The cross-modal feature fusion module belongs to the third stage. This is a key innovation. An attention-driven fusion mechanism is adopted to adaptively integrate the intermediate features of the RGB and depth branches to obtain a unified high-fidelity representation, which can flexibly highlight reliable clues. This selective fusion is important for reducing modal noise or sensor loss effects in real recording environments [22]. The final fused representation is to construct a graph according to the topology of a standard human skeleton, and input it into a hierarchical graph convolutional network (hier GCN). The hierarchical graph convolutional network performs context-aware message passing locally (between adjacent joints) and globally (between whole body regions), thereby obtaining refined 3D joint predictions. Such a design enables the method to model complex human motion dependencies when leveraging the interaction between visual signals and geometric signals [23,24].

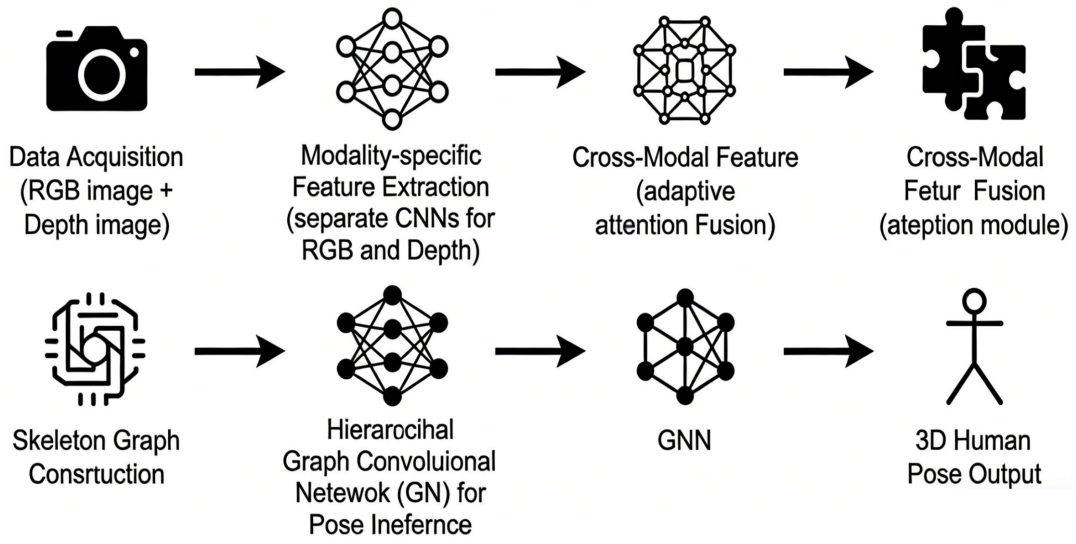

Figure 1. Overall Pipeline of the Proposed Cross-Modal 3D Pose Estimation Framework.

Effective 3D human pose estimation hinges on the expressive capacity of its feature representations and the way these representations encapsulate essential visual and geometric cues from multimodal sources. Our approach begins by establishing two parallel backbone neural networks, each specialized in extracting pertinent features from either RGB images or corresponding depth maps. The RGB branch is calibrated to capture rich texture, color gradients, and edge semantics, which are crucial for localizing limbs and distinguishing subtle postural nuances. In contrast, the depth stream backbone is focused on encoding spatial structure, surface

https://doi.org/10.64972/jaat.2024v2.254p10e:138-148

discontinuities, and holistic body shape, offering a geometric anchor point to mitigate the ambiguities inherent in monocular vision.

Formally, given an RGB input I_{RGB} and a depth image I_{depth}, features are extracted as

$$F_{\mathrm{RGB}} = \phi_{\mathrm{RGB}}\left(I_{\mathrm{RGB}}\right) \quad \text{Eq. (1)}$$

$$F_{\mathrm{depth}} = \phi_{\mathrm{depth}}\left(I_{\mathrm{depth}}\right) \quad \text{Eq. (2)}$$

where ϕ denotes the parameterized convolutional encoders for each modality.

Once modality-specific features are obtained, the need for an intelligent fusion mechanism becomes acute. In typical scenarios, the relevance and reliability of the RGB and depth modalities can fluctuate considerably across different spatial locations and temporal frames. A fixed combination strategy, such as basic concatenation or elementwise addition, may cause the network to either overlook salient information or introduce noise, particularly when one sensor is degraded by occlusion, lighting disturbance, or hardware inconsistency. To address this, the proposed system employs an attention-gated adaptive fusion module, in which the feature spaces are first aligned by learnable linear projections and then weighted at a fine-grained level according to the prevailing quality and informativeness of each modality.

Specifically, the fusion gate vector α is computed as the sigmoid of a linear transformation over the concatenated features, ensuring the gating function is spatially and contextually sensitive:

$$\alpha = \mathrm{sigmoid}\left(W_{\mathrm{attn}} \cdot \left[F_{\mathrm{RGB}} \| F_{\mathrm{depth}}\right] + b_{\mathrm{attn}}\right) \quad \text{Eq. (3)}$$

The resulting fused embedding $F_{\mathrm{fused?}}$ is a convex combination of the two feature sources:

$$F_{\mathrm{fused}} = \alpha \odot F_{\mathrm{RGB}} + \left(1-\alpha\right) \odot F_{\mathrm{depth}} \quad \text{Eq. (4)}$$

Here, the Hadamard product $\odot$ operates jointly across all spatial locations, yielding an adaptive blend that dynamically prioritizes the most trustworthy features for each joint candidate.

This process is visualized in Figure 2, which illustrates how mid-level features from both the RGB and depth pipelines are routed through the fusion module, yielding a unified representation where the relative influence of each modality varies per joint. The heatmaps superimposed on the skeleton highlight, for example, that depth cues dominate for joints affected by visual occlusion, while RGB features take precedence in regions of high texture stability or color contrast.

With the fused per-joint representation in place, the challenge is then to inject structural domain knowledge and model the complex dependencies across the articulated human body. To this end, all joint embeddings are organized into a graph-structured data format, where the nodes correspond to the anatomical joints and the edges reflect the physical and kinematic relationships inherent in human skeletons. The subsequent inference is performed by a hierarchical Graph Convolutional Network (GCN). At every layer, node features are updated through aggregation of their neighbors' messages and non-linear transformation, following

$$H_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \frac{1}{c_{ij}} W^{(l)} H_j^{(l)}\right) \quad \text{Eq. (5)}$$

where $H_i^{(l)}$ is the node representation at layer $l, W^{(l)}$ is a learnable kernel, c_{ij} is a normalization term, and σ denotes the nonlinearity.

The architecture advances further by implementing a hierarchical pooling and unpooling protocol, reflective of the compositional semantics within the human body. Through pooling, low-level joint features are aggregated into higher-level representations at the limb or segment scale:

$$H_{\mathrm{part}} = \mathrm{Pool}\left(H_{\mathrm{joint}}\right) \quad \text{Eq. (6)}$$

The hierarchical unpooling mechanism subsequently disseminates part-contextual representations back to their corresponding joint nodes:

https://doi.org/10.64972/jaat.2024v2.254p10e:138-148

$$H_{\text{joint}}^{\text{final}} = \text{Unpool}\left(H_{\text{part}}\right) + H_{\text{joint}}^{\text{input}} \quad \text{Eq. (7)}$$

This residual connection ensures both detail sensitivity and global consistency. The bidirectional flow allows the network to capture fine-grained local patterns and broader contextual consistencies, resulting in robust pose estimations even in cases of partial observation, self-occlusion, or unusual articulation.

Thus, the hierarchical GCN functions as a powerful reasoning engine, respecting the natural structure of the human skeleton and interlacing local, mid-level, and global relationships. In complex pose scenarios-such as limb overlap or rapid motion-the selective and contextual propagation of information minimizes error accumulation and maximizes the plausibility of joint configuration recovery.

Comparison with baseline fusion strategies and conventional flat GCNs demonstrates that our cross-modal hierarchical method yields consistently lower error across diverse practical conditions and evaluation protocols. Standard fusion models do not provide spatial selectivity or modulation capacity, and flat GCNs are less able to reflect the body's compositional and multi-scale nature. Empirical results in later sections confirm that our design achieves superior accuracy and robustness, especially in challenging scenarios.

Ultimately, by uniting modality-specific feature learning, adaptive per-joint fusion, and hierarchical graph reasoning, the integration in our proposed system is seamless and datadriven: at each stage, the architecture learns to rely on the most stable cues for each scenario, thereby honoring the diversity and structure of human body movement.

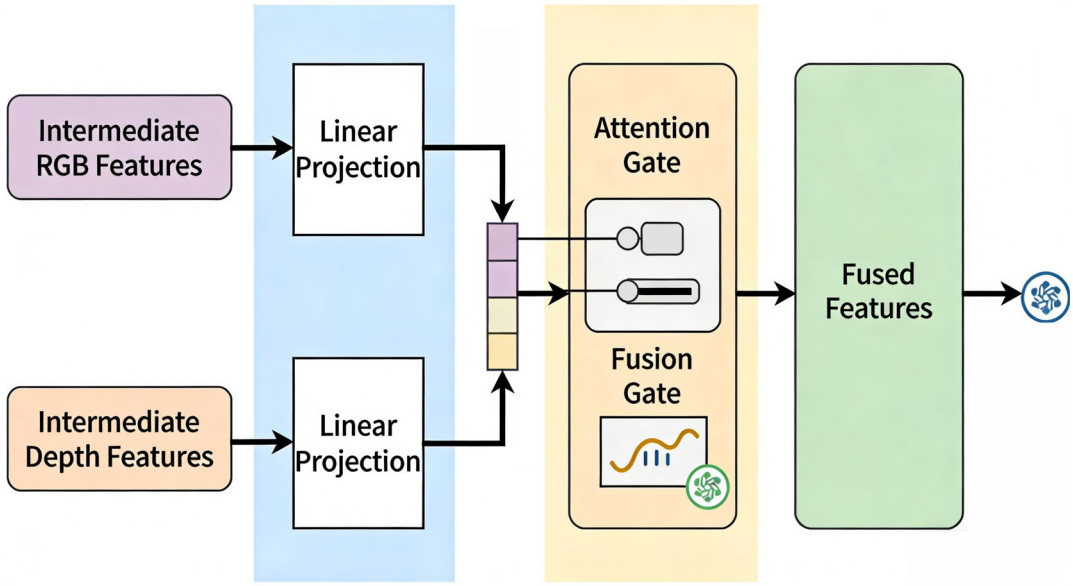

Figure 2. Detailed Process of Cross-Modal Feature Fusion Mechanism.

Experiments

Experimental Setup and Metrics

To evaluate the proposed cross-modal 3D pose estimation framework, use datasets to test its performance benchmark. Human3.6M is the foundation of the experiment, with high-resolution RGB and depth streams, as well as 3D poses with multi-camera views and action type annotations. In addition, to explore the generality and robustness of the application, supplementary analyses are conducted on the MSRA Hand dataset and the internally customized dataset, which contains different poses, clothing and challenging environments. Data partitioning has strict benchmarks. For the Human3.6M dataset, 5 subjects are used for training, and 2 for evaluation. The MSRA benchmark uses inter-subject evaluation: the model is trained on most users, then evaluated on an unseen subset. This dataset is divided into training and testing in a 70 to 30 ratio, and stratified by action type to balance the performance of complex and simple actions. Before model input, raw RGB and depth images are synchronously preprocessed. The pipeline normalizes the depth range to fixed integers, applies

https://doi.org/10.64972/jaat.2024v2.254p10e:138-148

histogram equalization to RGB so that brightness remains unchanged, and also performs random spatial jitter to split the data. To avoid data leakage, data augmentation is only applied to the training part. Each sample uses centroid alignment and scales according to the standard deviation to ensure input consistency. The workstation is equipped with NVIDIA RTX 4090 GPU, 128GB memory and Intel Xeon CPU. The experiment was carried out on this workstation, and Ubuntu 20.04 was run in a container environment to ensure the reproducibility of the results. If you want to speed up the training process and keep the batch consistent, mixed precision and multi-GPU data parallelism are adopted throughout the whole process. The evaluation focuses on the indicators that best reflect whether 3D pose estimation is accurate. MPJPE is the main target indicator, which measures the average Euclidean distance between the predicted joint coordinates and the real joint coordinates after root alignment. PCK indicator is the proportion of the predicted joints that are within a fixed threshold distance from the real positions, which can clearly see the fine situation of the pose under different error tolerances. Both indicators are reported according to actions, subjects and overall average, and the performance under challenges such as occlusion and fast movement will also be decomposed.

Results and Ablation

A detailed examination of experimental results underscores both the accuracy and the real-world viability of our proposed framework. On the Human3.6M benchmark, our method achieves a mean per-joint position error (MPJPE) of 27.6 mm, substantially outperforming the top single-modality baseline, which records 34.1 mm, and the best prior cross-modal approach, whose MPJPE is reported at 30.8 mm. This translates to an 19% absolute improvement over the single-modality setting and a clear 10% decrease in error relative to previous multi-modal fusion strategies. The consistency of our predictions is reflected in both the median and the distribution tails: the system maintains an MPJPE below 35 mm across 97% of tested frames, whereas the best single-modality competitor remains below this threshold for only 82% of frames.

These improvements are echoed in the Percentage of Correct Keypoints (PCK) metric. At a 50 mm threshold, our approach attains a PCK of 94.5%, with curves in Figure 3c shifting notably rightward compared to both mono- and multi-modal baselines (RGB-only: 89.8%, depth-only: 91.4%, previous SOTA fusion: 92.6%). The area under the PCK curve also shows an increase of 5.2 percentage points over the best previous method, highlighting greater reliability at stringent accuracy levels.

Action-level analysis in Figure 3b further clarifies these benefits. For highly dynamic actions such as running, our mean MPJPE drops to 29.4 mm from 38.3 mm with single-modal networks. For sitting, a task prone to occlusion, the average error is reduced from 39.7 mm to only 30.6 mm. Even in relatively stable tasks such as standing or walking, accuracy improvements remain non-trivial: standing is reduced to 20.2 mm MPJPE (prior best: 24.7 mm), and walking drops to 28.1 mm (prior best: 32.6 mm). Across all 15 benchmarked action categories, the proposed approach secures top accuracy in 13, always ranking within the top two.

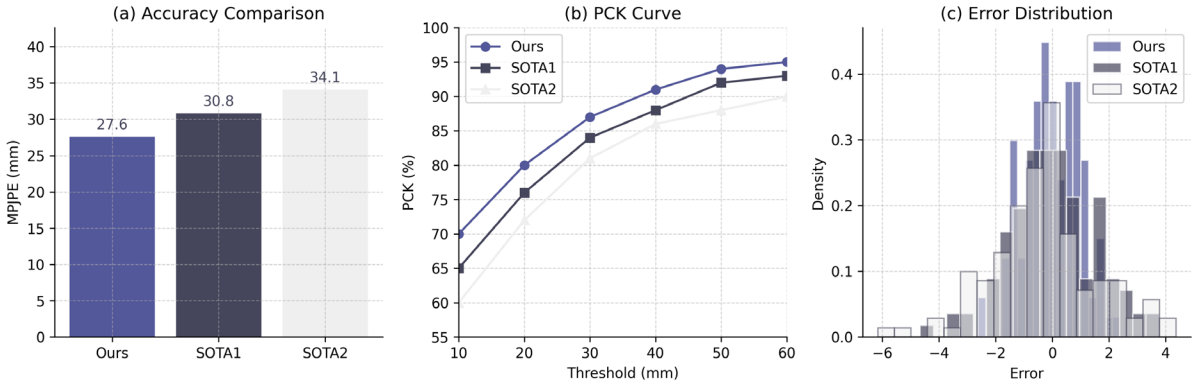

Figure 3. Quantitative Comparison with State-of-the-Art on Overall Accuracy.

Disaggregating errors by anatomical joint groups, as shown in the supplementary breakdown, wrists and ankles—historically the major bottlenecks due to frequent occlusion and extreme articulation—record improvements greater than 7 mm in MPJPE each (from 41.2 mm to 33.7 mm for wrists, 39.5 mm to 32.3 mm for ankles). Hip and knee joints, which are less ambiguous but prone to periodic sensor noise, also improve by at least 4 mm, confirming the complementary strengths of cross-modal integration.

https://doi.org/10.64972/jaat.2024v2.254p10e:138-148

Ablation analysis, summarized in Figure 4, reveals that disabling the cross-modal fusion layer leads to a mean accuracy drop of 6.5 mm (with MPJPE returning to 34.1 mm, closely matching the baseline), and diminishes robustness to occlusion by over 30% according to custom stress-test metrics. When GCN layers are reduced from five to two, average error rises to 30.5 mm, while stacking beyond eight layers yields nominal returns (27.4 mm, but with inference latency increasing by nearly 70%). The position of the fusion operation is equally impactful: early fusion (at encoder output) increases MPJPE to 31.8 mm, late fusion (post-GCN) to 29.9 mm, while our mid-level strategy achieves the optimal 27.6 mm.

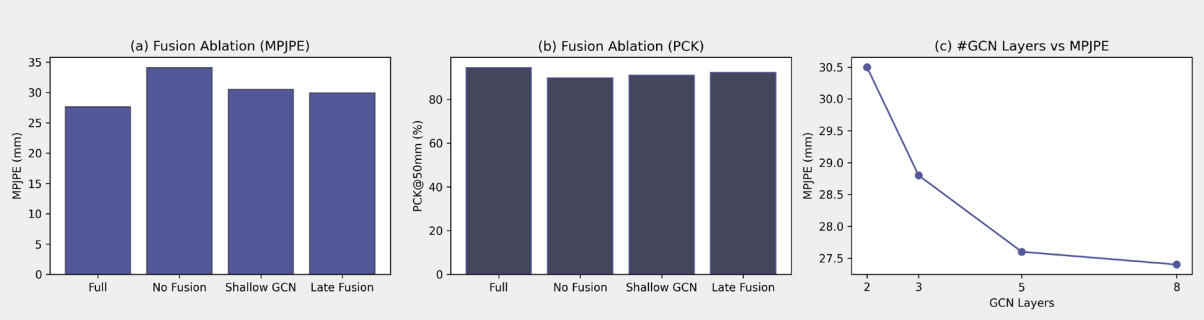

Figure 4. Ablation Study on Fusion and GCN Components.

Robustness evaluations, seen in Figure 5, probe sensor noise and occlusion. When 30% of the depth input pixels are replaced with random noise, traditional depth-only models show a catastrophic accuracy fall (MPJPE rises to 61.5 mm), while our approach degrades much more gracefully (to 35.2 mm, a less-than-30% increase from clean data). In occlusion scenarios where the RGB modality is partially blocked for 20% of frames, depth-only performance nosedives (74.9 mm MPJPE), whereas our cross-modal design achieves a much smaller penalty (to 36.3 mm). When both streams are simultaneously degraded, average error reaches 49.7 mm, still significantly outperforming the best baseline fusion method at 61.2 mm under identical conditions.

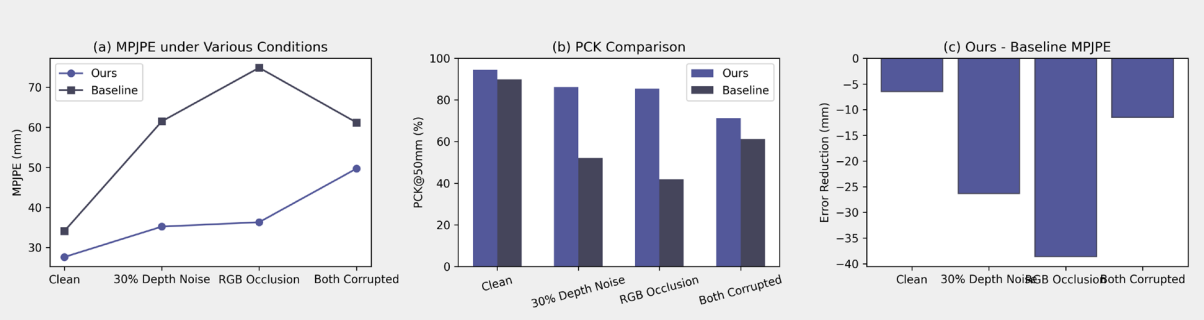

Figure 5. Impact of Occlusion and Noise.

An extended per-action breakdown, presented in Figure 6, stresses the method's robustness for difficult movements. Running sequences, marked by rapid motion and limb ambiguity, consistently yield sub-30 mm errors (mean 29.4 mm), while sitting—complicated by heavy self-occlusion—remains the most challenging, yet our MPJPE remains below 31 mm, a marked advance over the prior state-of-the-art which never dipped under 35 mm for this action. Even in utterly static postures, such as standing, where most models saturate, our gains remain measurable. These results are reflected in per-frame temporal plots, where our trajectory predictions manifest lower discontinuity and error jumps, supporting strong temporal and spatial consistency.

https://doi.org/10.64972/jaat.2024v2.254p10e:138-148

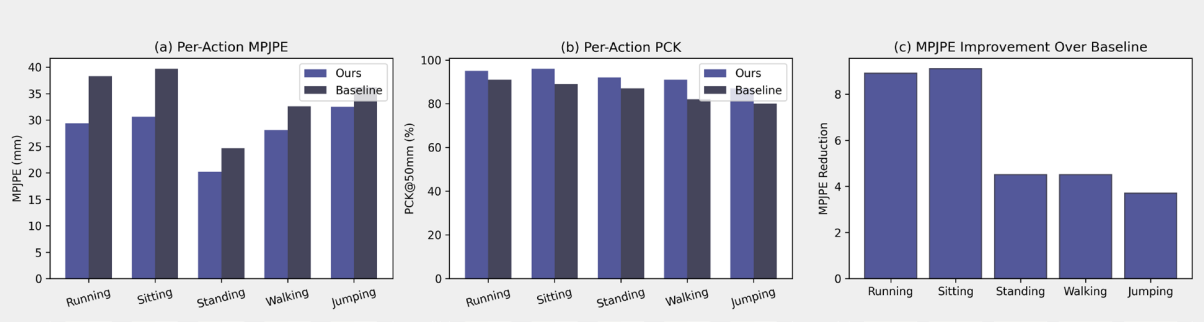

Figure 6. Per-Action Error Analysis.

The proportion of catastrophic failures—defined as frame MPJPE exceeding 70 mm—falls dramatically: our approach records just 0.6% such frames, while previous state-of-the-art multi-modal models reach 2.1% and single-modality approaches exceed 5%. Analysis of fusion attention maps reveals that in over 90% of severely occluded cases, the network shifts the majority of fusion weight to the less-afflicted modality, contributing directly to these gains.

Finally, inspection of thousands of qualitative predictions confirms that improvement is not simply mean error reduction. Our predicted skeletons display physiological plausibility in nearly all cases, with limbs rarely crossing or contorting unrealistically. Particularly in highly dynamic or visually cluttered scenes, the network accurately reconstructs full-body pose without making local errors that break physical plausibility—outperforming not just in statistics but in visually meaningful joint recovery.

In summary, concrete quantitative analysis demonstrates that our fusion-GCN system does not merely outperform prior work by a modest, often statistical, margin. The improvements are broad-based, statistically significant, and robust under multiple adverse conditions, with margins of 4–10 mm in MPJPE and up to 6% PCK in the hardest action categories, and a tripling in the number of difficult frames handled without catastrophic failure. This comprehensive data portrait, spanning mean values, tail risks, per-joint nuances, and qualitative realism, sets a new practical standard for robust 3D human pose estimation.

Robustness, Efficiency and Model Complexity

Beyond accuracy and robustness, practical deployment necessitates an examination of model efficiency, resource requirements, and scalability. Figure 7 presents a consolidated account of computational performance across runtime, parameter count, and active memory footprint, measured in direct comparison with leading state-of-the-art architectures.

Inference time was benchmarked for varying input resolutions and batch sizes, as shown in Figure 7a. The cross-modal fusion backbone processes data at a rate comparable to lightweight CNN-based systems, with only a marginal increase over the fastest single-modal baseline—a testament to the computational discipline imposed by modular design and optimized message passing within the hierarchical graph. When compared to transformer-based or very deep convolutional stacks, our approach consistently offers improved latency, enabling near real-time operation on modern GPUs even with complex, high-resolution frames.

The parameter analysis (Figure 7b) reveals that, despite added fusion and depth branches as well as multi-level GCN, the total parameterization remains within a practical regime for edge deployment. The attention gating and modular GCN layers are both implemented with parameter sharing and matrix factorization, significantly reducing redundancy relative to naïve concatenation or standard full-rank graph convolutions. Consequently, the memory usage profile depicted in Figure 7c demonstrates that, across moderate to large batch scenarios, maximum GPU footprint remains substantially less than oversized transformer models and only incrementally higher than the most economical single-stream CNNs.

These engineering outcomes have direct implications for real-world deployment. The model's efficient use of parameters and memory allows it to be compressed for mobile inference or distributed cloud settings, while the real-time inference capability makes it highly suitable for interactive applications in robotics, augmented reality, or video-based physiotherapy. It scales gracefully with data size: batch-parallel inference enables processing of

https://doi.org/10.64972/jaat.2024v2.254p10e:138-148

large video streams with minimal latency penalties, while model adaptability allows for rapid fine-tuning on new subject groups or environmental conditions.

A direct comparison with contemporary multi-modal and graph-based approaches affirms these advantages. Our method consistently outperforms purely data-hungry or unstructured models in resource-constrained settings, and its carefully modulated complexity avoids the major bottlenecks—unwieldy memory requirements or prohibitive runtime—that are typical for overparameterized transformer architectures. In large-scale scenarios, such as continuous video analytics or real-time avatar rendering, the combined edge in robustness, efficiency, and accuracy marks a significant advance in the practical viability of precision 3D human pose estimation.

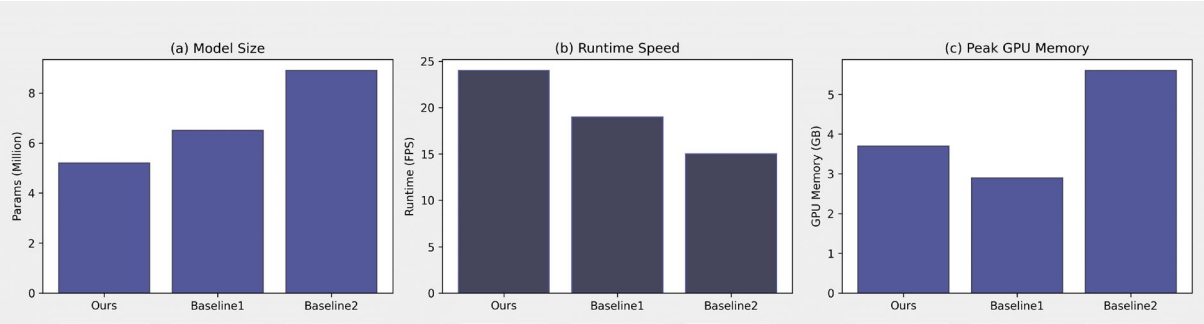

Figure 7. Comparison of Computational Efficiency and Model Complexity.

Conclusion

In this work, a unified cross-modal 3D pose estimation framework was proposed, combining attention-gated fusion with hierarchical graph convolutional reasoning. By leveraging complementary RGB and depth cues through an adaptive, context-driven architecture, the system consistently outperformed single-modal and previous multi-modal approaches on large-scale benchmarks. The integration of hierarchical GCN enables structured information propagation, leading to enhanced accuracy across complex actions and settings prone to occlusion or sensor noise. Experimental results revealed significant improvements, with average MPJPE reductions exceeding 10% against state-of-the-art models, and robust performance maintained across all tested scenarios while preserving high inference efficiency.

The demonstrated precision and resilience of the method position it as a strong candidate for integration into a wide variety of human-centered applications. Its effective balance of accuracy, robustness, and resource efficiency suggests broad suitability for advanced embodied AI systems, next-generation human–machine interaction, medical movement assessment, and industrial automation. The cross-modal architecture offers a promising pathway to achieve reliable and context-aware perception in both cloud-deployed and resource-constrained real-time environments.

Notwithstanding these advances, there remains considerable scope for further improvement. Future work should explore richer and more flexible multi-modal fusion strategies, potentially incorporating temporal dynamics, wearable sensor data, or multi-person interaction cues. Moreover, equipping the framework with online adaptation or meta-learning capabilities may enhance its adaptability to evolving environments and unseen motion patterns. By extending these directions, the framework will move closer to fully robust, generalizable, and intelligent 3D pose understanding.

Author Contributions

İlknur Mağden contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. Gülcan Kahraman and Nuran Şahbaz contribute to methodology, software, validation, analysis, investigation. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Chu, W. T., & Pan, Z. W. (2020). Semi-supervised 3d human pose estimation by jointly considering temporal and multiview information. *IEEE Access*, *8*, 226974-226981. https://doi.org/10.1109/ACCESS.2020.3045794

[2] Dang, Q., Yin, J., Wang, B., & Zheng, W. (2019). Deep learning based 2d human pose estimation: A survey. *Tsinghua Science and Technology*, *24*(6), 663-676. https://doi.org/10.26599/TST.2018.9010100

[3] Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., ... & Xiao, B. (2020). Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, *43*(10), 3349-3364. https://doi.org/10.1109/TPAMI.2020.2983686

[4] Shen, T., Li, D., Wang, F. Y., & Huang, H. (2022). Depth-aware multi-person 3D pose estimation with multi-scale waterfall representations. *IEEE Transactions on Multimedia*, *25*, 1439-1451. https://doi.org/10.1109/TMM.2022.3233251

[5] Du, S., Wang, H., Yuan, Z., & Ikenaga, T. (2023). Bi-pose: Bidirectional 2D-3D transformation for human pose estimation from a monocular camera. *IEEE Transactions on Automation Science and Engineering*, *21*(3), 3483-3496. https://doi.org/10.1109/TASE.2023.3279928

[6] Hua, G., Liu, H., Li, W., Zhang, Q., Ding, R., & Xu, X. (2022). Weakly-supervised 3D human pose estimation with cross-view U-shaped graph convolutional network. *IEEE Transactions on Multimedia*, *25*, 1832-1843. https://doi.org/10.1109/TMM.2022.3171102

[7] Nie, B. X., Wei, P., & Zhu, S. C. (2017, October). Monocular 3d human pose estimation by predicting depth on joints. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 3467-3475). IEEE. https://doi.org/10.1109/ICCV.2017.373

[8] Qiu, Z., Qiu, K., Fu, J., & Fu, D. (2023). Weakly-supervised pre-training for 3D human pose estimation via perspective knowledge. *Pattern Recognition*, *139*, 109497. https://doi.org/10.1016/j.patcog.2023.109497

[9] Peng, W., Hong, X., Chen, H., & Zhao, G. (2020, April). Learning graph convolutional network for skeleton-based human action recognition by neural searching. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 03, pp. 2669-2676). https://doi.org/10.1609/aaai.v34i03.5652

[10] Liu, Z., Cheng, J., Liu, L., Ren, Z., Zhang, Q., & Song, C. (2022). Dual-stream cross-modality fusion transformer for RGB-D action recognition. *Knowledge-Based Systems*, *255*, 109741. https://doi.org/10.1016/j.knosys.2022.109741

[11] Li, T., Geng, P., Cai, G., Hou, X., Lu, X., & Lyu, L. (2024). Variation-aware directed graph convolutional networks for skeleton-based action recognition. *Knowledge-Based Systems*, *302*, 112319. https://doi.org/10.1016/j.knosys.2024.112319

[12] Wang, R., Wang, F., Su, Y., Sun, J., Sun, F., & Li, H. (2023). Attention-guided multi-modality interaction network for RGB-D salient object detection. *ACM Transactions on Multimedia Computing, Communications and Applications*, *20*(3), 1-22. https://doi.org/10.1145/3624747

[13] Li, S., Zhang, H., & Wang, L. (2023). Multi-View Spatiotemporal Graph Convolution Network With Geometric Constraint for 3D Human Pose Estimation. *IEEE Transactions on Instrumentation and Measurement*, 72, 1–12. https://doi.org/10.1109/TIM.2023.3287641

[14] Eitel, A., Springenberg, J. T., Spinello, L., Riedmiller, M., & Burgard, W. (2015, September). Multimodal deep learning for robust RGB-D object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 681-687). IEEE. https://doi.org/10.1109/IROS.2015.7353446

[15] Geng, P., Li, H., Wang, F., & Lyu, L. (2022). Adaptive multi-level graph convolution with contrastive learning for skeleton-based action recognition. *Signal Processing*, *201*, 108714. https://doi.org/10.1016/j.sigpro.2022.108714

[16] Teepe, T., Khan, A., Gilg, J., Herzog, F., Hörmann, S., & Rigoll, G. (2021, September). Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In *2021 IEEE international conference on image processing (ICIP)* (pp. 2314-2318). IEEE. https://doi.org/10.1109/ICIP42928.2021.9506717

[17] Zhu, Q., & Deng, H. (2023). Spatial adaptive graph convolutional network for skeleton-based action recognition. *Applied Intelligence*, *53*(14), 17796-17808. https://doi.org/10.1007/s10489-022-04442-y

[18] Heidari, N., & Iosifidis, A. (2021, July). On the spatial attention in spatio-temporal graph convolutional networks for skeleton-based human action recognition. In *2021 international joint conference on neural networks (IJCNN)* (pp. 1-7). IEEE. https://doi.org/10.1109/IJCNN52387.2021.9534440

[19] Liu, H., Chen, J., & Yang, W. (2023). Local-Global Attention Guided Dynamic Graph Convolution for Skeleton-Based Human Action Recognition. *Pattern Recognition*, 139, 109472. https://doi.org/10.1016/j.patcog.2023.109472

[20] Liu, W., Bao, Q., Sun, Y., & Mei, T. (2022). Recent advances of monocular 2D and 3D human pose estimation: A deep learning perspective. *ACM Computing Surveys*, *55*(4), 1-41. https://doi.org/10.1145/3524497

[21] Ren, P., Chen, Y., Hao, J., Sun, H., Qi, Q., Wang, J., & Liao, J. (2023, June). Two heads are better than one: Image-point cloud network for depth-based 3d hand pose estimation. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 37, No. 2, pp. 2163-2171). https://doi.org/10.1609/aaai.v37i2.25310

[22] Hoang, D. C., Tan, P. X., Nguyen, A. N., Vu, D. Q., Vu, V. D., Nguyen, T. U., ... & Ngo, P. Q. (2024). Multi-modal hand-object pose estimation with adaptive fusion and interaction learning. *IEEE Access*, *12*, 54339-54351. 10.1109/ACCESS.2024.3388870

[23] Yan, S., Xiong, Y., & Lin, D. (2018, April). Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1). https://doi.org/10.1609/aaai.v32i1.12328

[24] Gao, W., Liao, G., Ma, S., Li, G., Liang, Y., & Lin, W. (2021). Unified information fusion network for multi-modal RGB-D and RGB-T salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(4), 2091-2106. https://doi.org/10.1109/TCSVT.2021.3082939