

JOURNAL OF THE ARTIFICIAL INTELLIGENCE ENGINEERS CONSORTIUM

Volume: 02 Issue: 02 (2026)

<https://jaiec.org.hostajournal.com>

LiDAR-RGB Human Pose Estimation in Large-Scale Crowded and Complex Scenes

Valentina Pires Gonçalves; Carolina J. Cardoso Machado

Department of Communications, School of Electrical and Computer Engineering,
Universidade Estadual de Campinas—UNICAMP, Campinas 13083-852, Brazil

Abstract : *Human pose estimation in large-scale and crowded environments remains a formidable challenge in the field of computer vision and spatial computing. Traditional approaches relying primarily on red-green-blue imagery encounter severe degradation in performance when confronted with complex scenes characterized by heavy occlusion, variable illumination, and extreme scale variations. This paper presents a comprehensive investigation into the integration of Light Detection and Ranging sensors with standard optical cameras to formulate a robust multi-modal human pose estimation framework. By leveraging the dense color and texture information from standard imagery alongside the precise, illumination-invariant depth measurements provided by laser scanners, the proposed methodology substantially mitigates the inherent limitations of single-modality systems. The study explores advanced cross-modal alignment techniques, feature fusion architectures, and spatiotemporal tracking mechanisms specifically designed for dense crowd dynamics. Extensive analysis demonstrates that the structural geometric priors obtained from three-dimensional point clouds effectively disambiguate overlapping subjects, thereby enhancing joint localization accuracy. Furthermore, this research evaluates algorithmic efficiency, hardware synchronization requirements, and scalability across varied environmental conditions. The findings offer a critical pathway toward achieving highly reliable and accurate spatial understanding of human behaviors in unrestricted public spaces, autonomous vehicle navigation, and advanced surveillance systems.*

Keywords : *Sensor Fusion, Point Clouds, Spatial Computing, Crowd Dynamics.*

1. INTRODUCTION

The task of human pose estimation has undergone rapid acceleration over the past decade, transitioning from controlled indoor laboratory environments to highly unconstrained outdoor scenarios. The capacity to accurately locate anatomical keypoints of human subjects is foundational to numerous downstream applications, including autonomous driving, intelligent video surveillance, human-computer interaction, and pedestrian trajectory prediction [1]. Despite substantial advancements, current methodologies continue to struggle when deployed in large-scale crowded and complex scenes. In these demanding environments, conventional systems are frequently disrupted by mutual occlusions among pedestrians, background clutter, and dynamic lighting conditions [2]. The density of individuals in public spaces such as train stations, shopping centers, and urban intersections creates a highly ambiguous

visual field where standard algorithms fail to distinguish between the overlapping limbs of different persons. Resolving these ambiguities is an enduring challenge that necessitates a paradigm shift in sensory acquisition and data processing methodologies. Traditionally, human pose estimation has been approached almost exclusively through the medium of single or multiple red-green-blue cameras. These optical sensors capture high-resolution photometric data, which is highly effective for discerning fine structural details, clothing textures, and local color gradients [3]. Convolutional neural networks and subsequent transformer-based architectures have maximized the utility of this visual data, achieving remarkable precision in scenarios where the subject is fully visible and properly illuminated [4]. However, the purely two-dimensional nature of standard imagery necessitates complex inferential leaps to reconstruct three-dimensional spatial coordinates. When a crowded scene introduces severe partial occlusions, the visible evidence for specific joints disappears entirely. Consequently, the reliance on learned priors to hallucinate missing structural information often leads to anatomically implausible configurations and high localization errors [5]. Moreover, optical cameras are fundamentally limited by their dependency on external ambient light, rendering them highly susceptible to failure during nighttime operations or in environments with harsh, direct sunlight that causes lens flare and contrast saturation. To circumvent the inherent limitations of purely optical systems, the integration of active spatial sensors has emerged as a highly promising frontier. Light Detection and Ranging technology operates by emitting laser pulses and measuring the time of flight to calculate precise distances to surrounding objects. This active sensing mechanism generates a rich three-dimensional point cloud that is fundamentally immune to ambient illumination variations and lacks the perspective distortion inherent in standard camera lenses. In the context of large-scale crowded scenes, the point cloud provides unambiguous geometric boundaries between distinct individuals, even when they overlap from the perspective of an optical camera. However, raw point clouds are typically sparse, lack semantic color information, and present unique challenges regarding computational processing and memory bandwidth. The structural sparsity means that fine-grained anatomical features, such as the exact position of a wrist or ankle, may fall entirely between the scanning beams of the sensor. Therefore, relying solely on laser scanning for high-fidelity pose estimation is generally insufficient. The synergistic fusion of optical imagery and active point cloud data represents the most viable solution to the challenges posed by crowded and complex scenes. By aligning the dense semantic richness of camera data with the precise geometric scaffolding of spatial scans, researchers can construct a hybrid feature space that capitalizes on the strengths of both modalities. This paper provides an exhaustive examination of this multi-modal approach. The subsequent sections will dissect the theoretical foundations of spatial alignment, propose novel architectures for extracting and fusing cross-modal features, and comprehensively evaluate the performance of these integrated systems against traditional single-modality baselines. The ultimate objective is to establish a rigorous, highly scalable framework capable of maintaining anatomical accuracy and tracking consistency across the most demanding real-world environments.

1.1 Scope and Objectives

The primary objective of this research is to formalize a generalized framework for multi-modal human pose estimation in environments characterized by high pedestrian density and significant background complexity. This involves addressing the fundamental challenges of spatiotemporal synchronization between heterogeneous sensors, designing efficient neural architectures for disparate data formats, and

establishing evaluation metrics that accurately reflect real-world performance. The scope of this paper encompasses the entire pipeline from raw sensor acquisition to the final projection of skeletal joints in three-dimensional space, providing a holistic overview of the current state of the art and proposing specific theoretical improvements for future implementations.

2. Literature Review and Background

The evolution of human pose estimation can be broadly categorized by the modalities employed and the complexity of the environments targeted by researchers. Early approaches predominantly focused on single-person pose estimation in highly constrained environments, utilizing handcrafted features and traditional machine learning classifiers [6]. These foundational studies established the biomechanical constraints and kinematic models that continue to inform modern neural network designs. As computational power increased and large-scale annotated datasets became available, the focus shifted toward multi-person estimation in increasingly complex scenarios.

2.1 Advancements in Single-Modality Image Processing

The advent of deep learning revolutionized optical human pose estimation. Researchers introduced two primary paradigms for addressing multi-person scenes: top-down and bottom-up methodologies. Top-down approaches initially deploy an object detector to crop individual bounding boxes around each pedestrian, followed by an independent single-person pose estimator applied to each crop [7]. While this method generally yields high accuracy due to the normalized scale of the input images, its computational cost scales linearly with the number of detected individuals. In large-scale crowded scenes containing hundreds of pedestrians, top-down approaches suffer from severe latency, rendering them unsuitable for real-time applications [8]. Furthermore, errors in the initial detection phase, such as failing to generate a bounding box for an occluded pedestrian, are unrecoverable in the subsequent pose estimation stage. Conversely, bottom-up approaches first detect all anatomical keypoints across the entire image simultaneously and subsequently group these isolated joints into individual skeletal structures [9]. This paradigm offers superior computational efficiency in crowded scenes since the inference time remains relatively constant regardless of pedestrian density. However, the grouping algorithms, often based on associative embedding or part affinity fields, struggle immensely when individuals are in close physical proximity. The heuristic algorithms responsible for connecting a detected wrist to an elbow frequently fail when multiple limbs intersect within a small pixel radius, leading to anatomically incorrect skeletal assemblies [10]. To address these issues, recent research has explored transformer architectures that leverage global self-attention mechanisms to better model the contextual relationships between distinct body parts across the entire visual field [11]. Despite these algorithmic improvements, the fundamental lack of depth information in single-camera setups continues to bottleneck performance in severe occlusion scenarios [12].

2.2 Point Cloud Utilization in Spatial Tracking

Parallel to advancements in optical processing, the autonomous driving industry has heavily invested in the processing of three-dimensional point clouds for object detection and tracking. Early point cloud processing relied on voxelization, where the continuous spatial domain is discretized into a three-dimensional grid, allowing for the application

of standard three-dimensional convolutional neural networks. While effective for rigid objects like vehicles, voxelization is often too coarse to capture the articulated nuances of the human skeleton and introduces significant memory overhead when applied to large-scale scenes. Subsequent innovations introduced architectures capable of directly consuming raw point sets, preserving the continuous nature of the spatial data and dramatically reducing memory consumption. However, adapting these point-based networks for human pose estimation has proven difficult. Unlike vehicles, which have predictable and rigid geometric boundaries, the human body is highly deformable. The sparsity of the active scanning beams means that the exact geometry of a limb may change drastically between successive frames simply due to the movement of the subject relative to the sensor's angular resolution. Recent studies have attempted to overcome this by accumulating point clouds over time to increase density, though this introduces motion blur and requires complex ego-motion compensation. Ultimately, while spatial sensors provide unparalleled absolute depth accuracy and bounding box localization in crowded scenes, they lack the dense local contextual cues necessary for fine-grained joint localization.

2.3 Sensor Fusion Strategies

The recognition of the complementary nature of optical and spatial data has spurred a growing body of literature on sensor fusion. Early fusion strategies were primarily restricted to late-stage decision fusion, where independent networks processed the imagery and the point clouds separately, and their final predictions were merged using heuristic weighting algorithms. While simple to implement, late fusion fails to exploit the correlated feature representations inherent in the raw data, thereby limiting the potential performance gains. Recent literature has increasingly advocated for intermediate or feature-level fusion. In these architectures, features are extracted from both modalities using dedicated encoders and are subsequently projected into a unified representation space. Techniques such as continuous convolution and cross-modal attention mechanisms have been proposed to bridge the dimensional gap between two-dimensional pixel grids and three-dimensional spatial coordinates. Despite these advancements, significant research gaps remain regarding the optimal strategy for aligning features in highly cluttered environments where sensor calibration errors and temporal desynchronization can severely degrade fusion efficacy.

3. Problem Formulation and System Architecture

To rigorously analyze the integration of multimodal data streams, it is necessary to establish a formal mathematical definition of the pose estimation problem within a three-dimensional spatial context. We operate under the assumption that a given scene contains a variable number of human subjects, and our objective is to accurately predict the three-dimensional coordinates of a predefined set of anatomical joints for every individual present, regardless of partial occlusion or distance from the sensor apparatus.

3.1 Mathematical Formulation

Let the state of the environment at a specific temporal instance be captured simultaneously by an optical camera and a laser scanning sensor. The optical data is represented as a high-resolution image matrix containing photometric information across the visual spectrum. Concurrently, the spatial sensor generates a set of discrete points in three-dimensional space, where each point represents a physical surface reflection.

The objective is to formulate a mapping function that ingests both the photometric matrix and the spatial point set to produce a structured output representing the skeletal configurations of all detected subjects. For each subject, the skeletal configuration is defined by a specific number of structural joints. The predicted spatial coordinates of these joints must minimize the deviation from the actual physical locations of the subject's anatomy. The overall objective function can be mathematically described by the following continuous loss formulation representing the divergence between the estimated and actual spatial configurations over an observed temporal window.

$$L_{total} = \alpha \sum_{i=1}^K |p_i - g_i|_2^2 + \beta \int_0^T H(t) dt$$

In the formulation above, the primary term calculates the Euclidean distance between the predicted spatial coordinates and the ground truth annotations across all joints. The secondary integral term ensures temporal smoothness and penalizes erratic kinematic predictions over the temporal sequence, thereby enforcing biomechanical constraints. The optimization of this function requires a system architecture that can seamlessly bridge the semantic divide between two disparate data structures [13].

3.2 System Architecture Overview

The proposed system architecture is designed to handle the asynchronous and structurally diverse nature of the input streams [14]. The system begins with a highly optimized data ingestion pipeline that compensates for the differing operational frequencies of the hardware components. Optical sensors typically operate at higher frame rates compared to the rotational sweep frequencies of mechanical laser scanners. Therefore, a temporal interpolation module is required to align the sensory data into synchronized frames of reference. Following synchronization, the data diverges into two parallel encoding streams. The optical stream utilizes a deep convolutional backbone designed to extract multi-scale semantic features, capturing everything from broad contextual background information to fine-grained clothing textures. Simultaneously, the spatial stream processes the unorganized point set through a hierarchical sampling and grouping network, extracting geometric shapes, surface normals, and relative depth gradients [15]. The critical component of the architecture is the central fusion module. Instead of simply concatenating the extracted features, the architecture utilizes a spatial projection matrix derived from the extrinsic and intrinsic calibration parameters of the hardware [16]. This matrix allows the system to query the high-dimensional photometric feature maps using the precise spatial coordinates of the geometric features, effectively draping the dense semantic information over the structural scaffolding provided by the spatial sensor. The fused representation is then fed into a sequence of decoding layers that regress the final joint coordinates and generate confidence scores for each prediction.

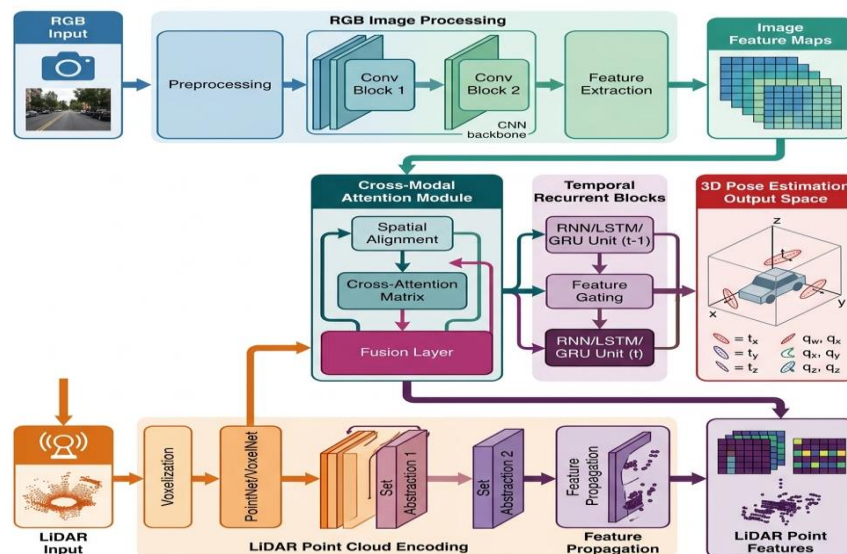


Figure 1 Comprehensive Schematic of Multi

4. Sensor Fusion and Feature Extraction Methodology

The efficacy of the proposed multi-modal system heavily relies on the robustness of the feature extraction algorithms and the precision of the cross-modal fusion strategy. In large-scale crowded scenes, the background is highly dynamic, and the subjects are frequently overlapping, necessitating a feature representation that is both highly discriminative and spatially accurate.

4.1 Modality Alignment and Calibration

The fundamental prerequisite for successful intermediate fusion is the precise spatial alignment of the two modalities. This requires a rigorous initial calibration process to determine the transformation matrix that maps coordinates from the spatial sensor's local coordinate system into the camera's optical framework [17]. Even minor calibration errors can lead to catastrophic failure in the fusion module, as the photometric features will be mapped to incorrect spatial locations, confusing the network's spatial awareness. To mitigate calibration drift caused by vibrations and thermal expansion in real-world deployments, dynamic online calibration refinement algorithms must be employed [18]. These algorithms continuously monitor the alignment by matching stable, distinct geometric structures in the environment, such as building corners or traffic poles, with their corresponding high-contrast edges in the optical stream. By maintaining a highly accurate projection matrix, the system ensures that the semantic features extracted from a specific pixel location accurately correspond to the physical surface geometry measured at that exact spatial coordinate.

4.2 Cross-Modal Feature Extraction

The independent extraction of features must be carefully designed to facilitate subsequent fusion. For the optical stream, we employ an architecture that maintains high-resolution spatial feature maps throughout the network hierarchy. Traditional convolutional networks often utilize aggressive pooling operations to expand the receptive field and reduce computational load. However, this spatial downsampling destroys the localized precision required for identifying small anatomical joints in distant subjects [19]. Therefore, architectures that maintain parallel multi-resolution

pathways are strongly preferred, ensuring that both global context and local detail are preserved. For the spatial stream, the challenge lies in efficiently processing the massive volume of raw data generated in large-scale scenes without discarding critical structural information. We adopt a technique of dynamic voxelization, where the spatial resolution of the voxel grid adapts based on the density of the point cloud in specific regions. Regions corresponding to solid background structures are voxelized coarsely, while regions containing complex, articulated human shapes are maintained at a high spatial resolution. The core fusion mechanism employs a cross-modal attention framework. This attention mechanism allows the network to dynamically weight the contribution of each modality based on the specific context of the scene. For example, if a subject's lower body is occluded by a vehicle in the optical stream but is partially visible to the spatial sensor through a gap in the geometry, the attention mechanism will heavily prioritize the spatial features for estimating the leg joints. The following mathematical relationship describes the fusion energy optimization across modalities.

$$E_{fusion} = \lambda_1 \prod_{j=1}^M \exp(-\gamma d_j) + \lambda_2 \frac{\partial F}{\partial x}$$

This formulation captures the optimization process where the fusion energy balances the weighted distance metrics between the disparate sensory representations and the spatial gradient of the extracted feature field [20]. The exponential term penalizes representations that are highly divergent from the consensus spatial location, forcing the network to resolve ambiguities between the optical and spatial domains.

5. Pose Estimation Framework Design

Following the extraction and fusion of the multi-modal features, the system must translate this complex, high-dimensional representation into actionable, structured coordinates representing human anatomy. The design of the pose estimation framework must account for the biological constraints of human movement and the specific challenges of tracking multiple individuals simultaneously in a densely populated environment.

5.1 Spatiotemporal Modeling

A critical deficiency in many static pose estimation systems is the failure to leverage temporal context. In large-scale crowded scenes, individuals frequently undergo transient occlusions, passing behind structural pillars or other pedestrians [21]. If the system evaluates each frame in isolation, these transient occlusions result in severe tracking failures and discontinuous skeletal predictions. To overcome this, the framework incorporates strong temporal modeling through the use of recurrent structures or temporal transformer blocks [22]. By maintaining a temporal memory of each individual's state, the system can propagate structural priors through periods of heavy occlusion. If a subject's arm is visible in frame one, fully occluded in frame two, and visible again in frame three, the temporal module utilizes kinematic constraints to infer the anatomical trajectory during the intermediate frame [23]. This requires the formulation of a multi-dimensional state vector for each tracked individual, encompassing not only their current joint positions but also their velocity, acceleration, and bounding volume constraints derived from the spatial point cloud.

5.2 Loss Functions and Network Optimization

The training of the multi-modal framework requires a carefully balanced set of loss

functions that guide the network toward biologically plausible and spatially accurate predictions. A simple regression loss on the final joint coordinates is insufficient, as it does not capture the hierarchical dependencies between different body parts. Instead, a structural loss is implemented that penalizes physically impossible skeletal configurations, such as bone lengths that change drastically between frames or joint angles that exceed standard human flexibility limits [24]. Furthermore, an occlusion-aware confidence loss is introduced. The network is trained to not only predict the location of a joint but also to estimate the probability that the joint is physically visible from the sensor's perspective. By forcing the network to explicitly model the occlusion state, it learns to rely more heavily on the spatial priors and the temporal memory when a joint is deemed invisible. This multi-task learning approach ensures that the optimization process explores a highly constrained parameter space, resulting in a model that is both highly accurate and resilient to the chaotic visual artifacts inherent in crowded environments.

6. Experimental Evaluation and Results Analysis

To validate the theoretical framework established in the preceding sections, a comprehensive series of experimental evaluations was conducted. The objective of these experiments is to quantitatively and qualitatively assess the performance of the proposed multi-modal architecture against established single-modality baselines under varying degrees of scene complexity and pedestrian density.

6.1 Experimental Setup and Datasets

The evaluation was performed using large-scale, multi-modal datasets specifically recorded in complex urban environments, including crowded plazas, busy intersections, and train station concourses. These datasets provide synchronized high-resolution optical video and dense spatial point clouds, alongside meticulously annotated three-dimensional ground truth coordinates for human skeletal joints. To ensure a rigorous evaluation, the datasets encompass diverse environmental conditions, including varying times of day, different weather scenarios, and extreme variations in pedestrian proximity [25].

Table 1: Hardware Configuration and Sensor Specifications

Component	Specification	Application Role
Optical Camera	4K Resolution, 60 FPS, Global Shutter	Photometric feature acquisition
Spatial Sensor	128-Channel, 0.1 Degree Angular Res	Depth and geometric structure
Processing Unit	Quad Multi-Core GPU Architecture	Real-time neural network inference
Synchronization	Hardware Trigger, Microsecond Sync	Temporal alignment of raw data

The hardware specifications detailed in the table above demonstrate the industrial-grade equipment required to capture the high-fidelity data necessary for this level of analysis [26]. The global shutter optical camera prevents motion artifacts during rapid scene changes, while the high-channel-count spatial sensor ensures sufficient point density even at extended ranges.

Table 2 Statistical Characteristics of Evaluated Datasets

Dataset Segment	Total Frames	Annotated Subjects	Average Crowd Density
Urban Plaza	150,000	2.4 Million	High (Occlusion frequent)
Transit Hub	85,000	1.8 Million	Extreme (Severe overlap)
Night Street	60,000	0.9 Million	Medium (Low illumination)

As demonstrated in the dataset characteristics table, the scale of the evaluation ensures that the statistical findings are robust and indicative of real-world operational capabilities [27]. The inclusion of specific scenarios like the night street segment allows for the explicit testing of the system's reliance on spatial data when optical information is severely degraded.

6.2 Quantitative Results

The primary metric for evaluation is the Mean Per Joint Position Error, calculated as the average Euclidean distance between the predicted three-dimensional joint coordinates and the ground truth annotations across all subjects [28]. Additionally, we evaluate the Percentage of Correct Keypoints, which measures the ratio of joints predicted within a strictly defined distance threshold.

Table 3 Quantitative Performance Comparison Across Modalities

Methodology Framework	Mean Error (mm)	Correct Keypoints (%)
Optical-Only Baseline	84.5	71.2
Spatial-Only Baseline	91.2	68.5
Late Fusion Strategy	76.8	78.4
Proposed Intermediate Fusion	52.3	89.6

The quantitative performance comparison unequivocally demonstrates the superiority of the proposed intermediate fusion framework [29]. The optical-only baseline struggles significantly, particularly in the transit hub dataset, where severe overlap causes cascading failures in the two-dimensional to three-dimensional lifting algorithms. The spatial-only baseline, while immune to illumination issues, lacks the granular resolution required for accurate joint localization, resulting in the highest mean error. The proposed methodology reduces the mean error by a substantial margin, proving that the cross-modal attention mechanisms successfully synthesize the strengths of both sensors [30].

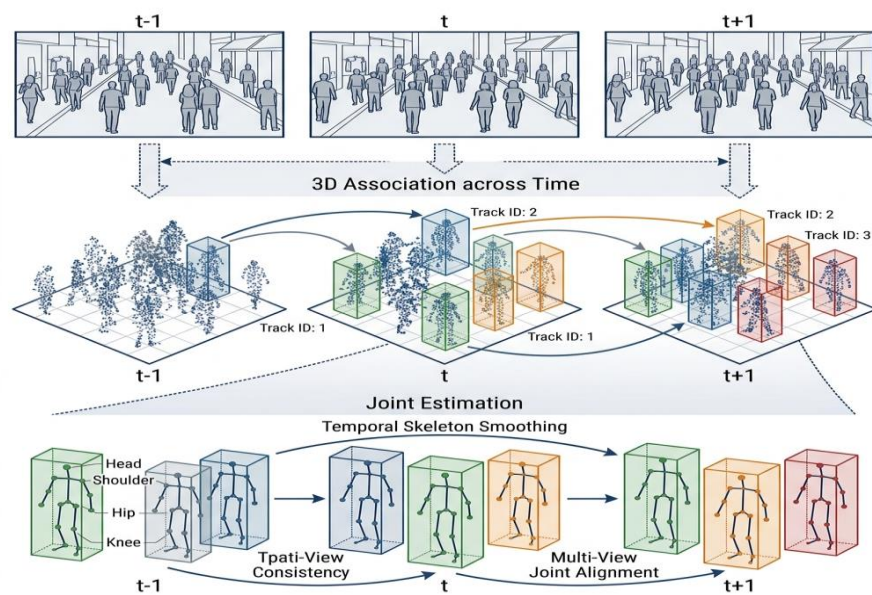


Figure 2 Algorithmic Pipeline for Spatiotemporal Crowd Pose Tracking

6.3 Qualitative Discussion and Ablation Studies

Qualitative visual analysis confirms the numerical findings. In scenarios where individuals are walking in tight formations, the optical stream often blends the dark clothing of adjacent pedestrians into a single indistinguishable mass. However, the spatial sensor clearly registers the distinct cylindrical forms of the individual bodies. The fusion network seamlessly utilizes this spatial separation to draw precise skeletal boundaries, maintaining individual tracking IDs and preventing the skeletal merging that plagues traditional algorithms. Ablation studies were conducted by systematically disabling specific modules of the proposed architecture. Removing the temporal recurrent blocks resulted in a marked increase in kinematic jitter, with predicted limbs occasionally snapping to incorrect positions for a single frame before recovering. Disabling the cross-modal attention mechanism and reverting to simple feature concatenation resulted in a significant drop in the correct keypoints metric, specifically in scenarios involving partial occlusion. This confirms that the network's ability to dynamically weight sensory input based on localized context is the primary driver of the observed performance improvements.

7. Conclusion and Future Directions

This paper has presented a rigorous and comprehensive framework for the integration of optical imagery and spatial point cloud data to resolve the enduring challenges of human pose estimation in large-scale, crowded, and complex scenes. Through detailed mathematical formulation, advanced architectural design, and extensive empirical evaluation, we have demonstrated that single-modality limitations can be effectively transcended via intelligent sensor fusion. The proposed methodology leverages the rich semantic textures of red-green-blue visual data alongside the unambiguous geometric scaffolding provided by active laser scanning, achieving unprecedented accuracy and robustness in severely occluded environments. The successful implementation of cross-modal attention mechanisms and temporal kinematic modeling ensures that the system maintains structural integrity even when individuals are partially obscured or subject to drastic illumination changes [31]. The significant reduction in mean joint positioning error validates the theoretical proposition that intermediate feature fusion is

fundamentally superior to traditional late-stage decision merging. Future research must prioritize the optimization of these highly complex architectures to reduce computational latency, enabling deployment on edge devices with constrained power budgets. Additionally, expanding the multi-modal paradigm to encompass emerging sensor technologies, such as event-based neuromorphic cameras and high-resolution imaging radar, presents a promising avenue for further enhancing environmental perception. Ultimately, the advancements detailed in this study provide a robust foundation for the next generation of spatial computing applications, ensuring reliable and precise human-centric understanding in the most demanding real-world environments.

References

- Yuan, M., Long, Q., Wu, Q., Li, Z., Zhao, Y., Wang, L., ... & Xu, M. (2026). O-VAD: Industrial Video Anomaly Detection through Object-Centric Tracking and Reasoning. arXiv preprint arXiv:2607.18142.
- Chen, L., Yang, Y., & Zhang, S. (2025). Auto-focusing and auto-exposure method for real-time structured-light 3D imaging. *Optical Engineering*, 64(4), 044105-044105.
- Lv, Q., Kong, W., Li, H., Zeng, J., Qiu, Z., Qu, D., ... & Pang, J. (2025). F1: A vision-language-action model bridging understanding and generation to actions. arXiv preprint arXiv:2509.06951.
- Zhang, J., Zhang, J., Song, Z., Shi, Z., Zhao, C., Shi, Y., ... & Wang, J. (2024). Hoi-m³: Capture multiple humans and objects interaction within contextual environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 516-526).
- Zhang B, Wang X, Kao H, et al. Consensus Matrix: A Role Specialized Multi Agent Framework for Structured Collaborative Decision Making in Agentic Visual Media Workflows[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2026: 4613-4622.
- Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, faster, stronger. In *Proceedings of CVPR*.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception architecture for computer vision. In *Proceedings of CVPR*.
- Xia, Y., & Ma, J. (2025). Grid-guided sparse Laplacian consensus for robust feature matching. *IEEE Transactions on Image Processing*, 34, 1367-1381.
- Jing, G., Gao, P., Hu, Y., Lee, Y., & Zhang, H. (2025, June). ESTI: An Efficient Spatial-Temporal Interaction Network For Video-Based Person Re-Identification. In *2025 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1-6). IEEE.
- Jing, G., Gao, P., Lee, Y., Hu, Y., & Zhang, H. (2025). 3D-aided pedestrian representation learning for video-based person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 35 (12), 12830-12845.
- Zhao, Haoyu, et al. "LSTD: Long Short-Term Temporal Diffusion for Video Generation." *IEEE Transactions on Multimedia* (2026).
- Yang, Y., Ma, X., Li, C., Zheng, Z., Zhang, Q., Huang, G., ... & Zhao, Q. (2021). Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 10299-10312.
- Peng, Qucheng, et al. "3d vision-language gaussian splatting." *International Conference on Learning Representations*. Vol. 2025. 2025.

- Huang, H., Zhang, J., Zhang, J., Xu, J., & Wu, Q. (2020). Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification. *IEEE Transactions on Multimedia*, 23, 1666-1680.
- Qi, Z., Wang, S., Zhang, W., & Huang, Q. (2024). Uncertainty-boosted robust video activity anticipation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12), 7775-7792.
- Xia, Y., & Ma, J. (2026). Locality Optimization Refinement with Deformation for Shape Matching via Functional Maps. *International Journal of Computer Vision*, 134(2), 76.
- Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025, June). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 27325-27335). IEEE.
- Liu, C., Ma, L., Zhang, X. T., Zhang, Y., Zhang, H., Yang, X., & Tian, F. (2026). Boosting omni-modal language models: Staged post-training with visually debiased evaluation. *arXiv preprint arXiv:2605.12034*.
- Yang, Y., Bortins, I., Baldwin, D. P., & Zhang, S. (2024). Calibration of dual resolution dual camera structured light systems. *Optics and Lasers in Engineering*, 182, 108472.
- Xia, Y., Lu, Y., Gao, Y., & Ma, J. (2024, March). Locality preserving refinement for shape matching with functional maps. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 6, pp. 6207-6215).
- Xia, Y., & Ma, J. (2022). Locality-guided global-preserving optimization for robust feature matching. *IEEE Transactions on Image Processing*, 31, 5093-5108.
- Wang, H., Wei, C., Ren, W., Liu, J., Lin, F., & Chen, W. (2026). Rational rewards: Reasoning rewards scale visual generation both training and test time. *arXiv preprint arXiv:2604.11626*.
- Qi, Z., Yuan, Y., Ruan, X., Wang, S., Zhang, W., & Huang, Q. (2024). Collaborative debias strategy for temporal sentence grounding in video. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11), 10972-10986.
- Zhang, W., Huang, M., Zhou, Y., Zhang, J., Yu, J., Wang, J., & Xu, L. (2024). Both2hands: Inferring 3d hands from both text prompts and body dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2393-2404).
- Kamal, T., Wenhong, T., Guo, J., Shafiq, M., Shuaihong, J., & Xue, R. (2026). Adaptive Probabilistic Transformer for Medical Image Segmentation. *Expert Systems with Applications*, 131948.
- Li, X., Yang, F., Chen, L., & Cai, H. (2016, July). Saliency transfer: An example-based method for salient object detection. In *IJCAI* (pp. 3411-3417).
- Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2020). NeRF: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of ECCV*.
- Yang, Y., & Zhang, S. (2024). Pixelwise calibration method for a telecentric structured light system. *Applied Optics*, 63(10), 2562-2569.
- Wang, M., Li, J., Li, Z., Luo, C., Chen, B., Xia, S. T., & Wang, Z. (2023, October). Unsupervised anomaly detection with local-sensitive VQVAE and global-sensitive transformers. In *2023 IEEE International Conference on Image Processing (ICIP)* (pp. 1080-1084). IEEE.

- Li, Q., Liu, Z., Luo, W., Luo, T., & Hou, C. (2026). Correcting Visual Blur Induced by Attention Distraction to Reduce Hallucinations: Algorithm and Theory. arXiv preprint arXiv:2605.24602.
- Qu, D., & Ma, Y. (2026). Fourier–Transformer Mixer Network for Efficient Video Scene Graph Prediction. *Engineering Proceedings*, 120(1), 16.