Software Engineering and Artificial Intelligence Systems

Volume: 02 Issue: 01 (2026)

www.seais.org/

Orthogonality-guided alignment of city-scale point clouds under strict memory budgets

Samuel Wan[1], Eric Tang[2]

[1] Department of Computer Science and Engineering, Faculty of Engineering, Chinese University of Hong Kong, Hong Kong, Hong Kong SAR, China

[2] School of Computing and Information Sciences, Saint Francis University, Hong Kong, Hong Kong SAR, China

Keyword:
Point Cloud Registration, Memory Optimization, Orthogonal Projection, Spatial Computing, Urban Digital Twins

A B S T R A C T

The rapid proliferation of three-dimensional scanning technologies has enabled the capture of city-scale point clouds with unprecedented detail and precision. These massive datasets are foundational for applications ranging from autonomous navigation and urban planning to the creation of comprehensive digital twins. However, the sheer volume of spatial data introduces severe computational bottlenecks, particularly concerning memory consumption during the alignment and registration phases. Traditional methods frequently exceed the memory capacities of standard computational hardware, necessitating severe downsampling that compromises structural fidelity. To address this critical challenge, this paper presents a novel methodology for the orthogonality-guided alignment of city-scale point clouds operating under strict memory budgets. By enforcing orthogonality constraints on the extracted feature spaces, the proposed framework minimizes redundant information and maximally preserves geometric distinctiveness without requiring massive memory overhead. The approach systematically decorrelates feature dimensions, ensuring that each parameter optimally contributes to the spatial alignment process. Comprehensive evaluations on simulated and real-world urban datasets demonstrate that the proposed orthogonality-guided technique achieves state-of-the-art registration accuracy while reducing memory utilization by substantial margins compared to contemporary alignment architectures. The results indicate that strict memory constraints do not preclude high-fidelity spatial registration when functional subspaces are optimized through rigorous geometric regularizations.

INTRODUCTION

1.1 Background and Motivation

The transition from two-dimensional geographic information systems to fully realized three-dimensional urban environments has fundamentally altered the landscape of spatial computing. Technologies such as light detection and ranging, terrestrial laser scanning, and high-resolution aerial photogrammetry generate immense volumes of raw spatial data, collectively referred to as point clouds. In the context of city-scale applications, these point clouds often encompass billions of individual coordinates, each accompanied by supplementary attributes such as intensity, color, and estimated surface normals. The aggregation of these datasets forms the backbone of modern smart city infrastructure, enabling advanced simulations, autonomous vehicle routing, structural health monitoring, and disaster response planning. As urban environments continuously evolve, the requirement to frequently update and align newly acquired point clouds with historical baseline data becomes a critical operational necessity. For a broader comparison, related research addresses joint-Aware video processing with Edge-Cloud collaboration for DNN inference [1]. Related work has addressed empowering large language models to progressively understand and generate 3D mesh [2]. Alignment, or registration, is the computational process of determining the optimal rigid or non-rigid spatial transformation that maps a source point cloud onto a target point cloud. In an ideal scenario, the alignment aligns corresponding geometric structures seamlessly, minimizing the spatial distance between homologous regions. However, operating at the scale of an entire metropolitan area introduces formidable computational challenges. The primary bottleneck is not necessarily the temporal complexity of the alignment algorithms, but rather the memory capacity required to store, manipulate, and query these massive datasets simultaneously. Standard high-performance computing clusters and edge devices alike are bounded by finite random access memory and graphical processing unit memory limits. Loading complete city-scale point clouds into active memory for feature extraction and correspondence matching inevitably triggers memory exhaustion or necessitates continuous, latency-inducing data swapping between primary memory and secondary storage. Jia and colleagues examine structured point cloud processing for Multi-Modal imitation learning [3]. Consequently, researchers and practitioners have historically relied on aggressive downsampling techniques to reduce the point count to a manageable threshold before attempting registration. While simple uniform or random downsampling algorithms effectively lower memory footprints, they inherently discard crucial geometric details. NeRF represents a scene through a continuous radiance field, a useful contrast when choosing between implicit scene models and explicit geometric representations [4]. Small but structurally significant urban features, such as traffic signs, building cornices, and utility poles, are frequently erased during downsampling, leading to ambiguous alignments or catastrophic registration failures. This creates a paradoxical challenge: aligning massive spatial datasets requires preserving detailed geometric features, yet the memory budget strictly prohibits the retention of the full dataset required to resolve these features. Thus, there is an urgent and critical need for a paradigm shift in how point clouds are processed, moving away from brute-force spatial retention toward highly optimized, information-dense representations. The literature also considers devil is in the details [5].

1.2 Problem Statement and Contributions

The core problem addressed in this research is the precise alignment of expansive, multi-gigabyte point clouds representing continuous urban environments under strictly enforced, hardware-defined memory constraints. Specifically, the challenge lies in extracting and matching robust geometric features without exceeding predetermined memory allocations, while simultaneously maintaining alignment accuracies comparable to or exceeding those of unconstrained registration pipelines. The hypothesis driving this work

is that the feature descriptors traditionally used in point cloud registration suffer from high dimensionality and significant internal correlation, which wastes valuable memory on redundant information. By mathematically enforcing orthogonality within the feature extraction and matching spaces, the informational entropy of the retained data can be maximized, allowing a significantly smaller memory footprint to encapsulate the necessary geometric distinctiveness for accurate alignment. The literature also considers time- and Memory-Efficient method for axis alignment of very large point clouds using orthogonality in buildings [6]. This paper makes several distinct contributions to the fields of spatial computing and computer vision. First, it introduces a novel theoretical framework that integrates orthogonality constraints directly into the point cloud feature extraction pipeline. Rather than relying on post-processing dimensionality reduction, the proposed method trains the feature extractors to natively project geometric data into mutually orthogonal subspaces. Second, a memory-bounded optimization strategy is developed, ensuring that the entire alignment process, from initial data ingestion to the final transformation calculation, operates strictly within user-defined memory limits. This includes a dynamic memory allocation mechanism that intelligently prioritizes the processing of geometrically complex regions over planar or featureless areas. Finally, the paper presents an extensive empirical analysis using large-scale urban datasets, demonstrating the practical viability and superior performance of the proposed approach against leading contemporary registration methodologies under simulated resource-constrained environments. Through this investigation, the research provides a foundational blueprint for deploying robust spatial alignment algorithms on resource-constrained platforms, ranging from mobile surveying backpacks to autonomous drone navigation systems. Another relevant line of investigation concerns indirect diffusion-guided neural radiance fields for generalizable view synthesis [7].

2. Literature Review and Technical Background

2.1 Point Cloud Alignment Techniques

The historical trajectory of point cloud registration is anchored by the iterative closest point algorithm and its numerous derivations. The fundamental premise of the iterative closest point method is the alternating refinement of point correspondences and spatial transformations. Specifically, the algorithm assumes initial point associations based on nearest neighbor proximity, calculates the optimal translation and rotation to minimize the aggregate spatial distance between these presumed pairs, applies the transformation, and subsequently recalculates the correspondences. While elegantly simple and provably convergent under conditions of close initial alignment, this classic approach exhibits profound limitations when applied to city-scale data. Its reliance on iterative nearest neighbor searches across unorganized spatial structures results in computational complexities that scale poorly with the number of points, and its susceptibility to local minima renders it ineffective for globally misaligned urban scans lacking strong initialization priors. Related work has addressed conditional denoising diffusion probabilistic model for point cloud upsampling [8]. For a broader comparison, related research addresses identification markers for salvia miltiorrhiza and its close relatives based on cell wall component characteristics [9]. To mitigate the initialization dependencies of local methods, global registration frameworks leverage distinctive geometric feature descriptors to establish initial correspondences without requiring a priori spatial proximity. Traditional handcrafted descriptors, such as the fast point feature histograms, analyze the local geometric topology around each point, encoding relative angles and distances into multi-dimensional histograms. These descriptors provide robustness against initial rigid transformations but demand substantial memory to compute and store the histograms for millions of points. More recently, the advent of deep learning has revolutionized global registration. Deep neural networks, specifically those processing point clouds directly or through graph-based convolutions, can learn highly discriminative feature representations directly from the raw data. Techniques utilizing dense convolutional networks or transformer-based attention mechanisms

have established new benchmarks in registration accuracy. However, these deep learning architectures inherently require immense memory overhead to store intermediate activation tensors, weight matrices, and the dense feature maps themselves, making their direct application to city-scale datasets largely unfeasible without massive computational resources. Another relevant line of investigation concerns reconstructing Ground-Penetrating radargrams into 3D point cloud format for building inspection: A web AR approach evaluated by technology acceptance model [10].

Table 1Comparison of Core Methodologies in Point Cloud Alignment based on resource utilization and accuracy metrics across generalized urban datasets

Methodology Category	**Memory Footprint**	**Computational Speed**	**Alignment Accuracy**	**Global Initialization Capability**
Iterative Closest Point Variants	Moderate	Fast to Moderate	High (if initialized)	Poor
Handcrafted Feature Matching	High	Slow	Moderate	Good
Deep Learning Dense Features	Extremely High	Moderate (GPU bound)	Very High	Excellent
Coarse-to-Fine Hierarchical	Moderate to High	Fast	High	Good
Proposed Orthogonality Method	Strictly Bounded	Fast	Very High	Excellent

2.2 Memory Constraints in Spatial Computing

Memory constraints dictate the operational boundaries of modern spatial computing. When processing three-dimensional data, the spatial complexity is cubic, meaning that as the resolution of an urban scan increases, the number of required data points grows exponentially. Standard memory management strategies in spatial algorithms typically involve hierarchical data structures such as octrees or k-dimensional trees to spatially partition the points, allowing for faster localized querying. While these structures improve temporal efficiency, the trees themselves consume supplementary memory, exacerbating the overall footprint. In heavily constrained environments, such as autonomous vehicles operating on edge computing platforms, the memory budget allocated for spatial awareness might be restricted to a few gigabytes, which is vastly insufficient for traditional processing of dense lidar streams. To function within these limitations, the prevailing strategy is downsampling. Voxel grid downsampling, where the three-dimensional space is divided into a regular grid of cubic volumes and the points within each volume are approximated by their centroid, is widely utilized due to its simplicity and predictable memory reduction. Random spatial sampling and farthest point sampling are alternative approaches. Farthest point sampling, in particular, provides better coverage of the geometric envelope of the point cloud but is computationally intensive and difficult to parallelize. Beyond sampling, point quantization techniques attempt to reduce memory by compressing the coordinate data itself, translating high-precision floating-point numbers into lower-resolution integers. While these strategies successfully enforce memory budgets, they universally degrade the geometric signal. Critical urban features become blurred or entirely eliminated, directly impacting the accuracy of downstream registration tasks. A sophisticated approach is required that reduces data volume not by arbitrary geometric truncation, but by maximizing the informational value of the retained representations. Related work has addressed end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models [11].

2.3 Orthogonality Principles in Optimization

The mathematical concept of orthogonality provides a powerful mechanism for information maximization and redundancy reduction. In linear algebra and functional analysis, orthogonal vectors or functions are completely independent; the projection of one onto another is zero. When applied to feature extraction and representation learning, enforcing orthogonality ensures that each dimension of a feature descriptor captures a unique and non-overlapping aspect of the underlying data. If a set of geometric features is highly correlated, it implies that the system is utilizing multiple memory units to store essentially identical information. By transforming the feature space into an orthogonal basis, the redundant dimensions can be safely discarded without significant loss of descriptive power, thereby drastically reducing the memory footprint. Related work has addressed singing timbre popularity assessment based on multimodal large foundation model [12].

In the context of deep learning, orthogonal regularization involves modifying the objective function during network training to penalize the non-orthogonality of the network weights or the output feature maps. This is often achieved by calculating the inner product of the feature matrices and forcing the result toward the identity matrix. Orthogonal weight initialization and regularization have been shown to stabilize the training of deep networks, prevent the vanishing and exploding gradient problems, and encourage the learning of diverse feature detectors. For spatial computing specifically, leveraging orthogonality means that the minimal subset of points or features retained under a strict memory budget is guaranteed to hold the maximum possible geometric distinctiveness. By ensuring that local descriptors representing building facades, road surfaces, and vegetation are encoded into orthogonal sub-spaces, the subsequent matching algorithms can confidently establish correspondences using vastly smaller data structures, bridging the gap between scale and memory limits.

3. Orthogonality-Guided Alignment Methodology

3.1 Sub-Sampling and Feature Extraction

The proposed orthogonality-guided methodology is fundamentally designed to operate within predetermined memory boundaries while achieving highly precise city-scale point cloud alignment. The pipeline commences with a specialized ingestion protocol that streams raw spatial data directly from storage without loading the complete dataset into primary memory. As the data streams, an intelligent, memory-aware spatial partitioning scheme is immediately applied. Unlike standard voxel grid filtering, which uniformly downsamples the entire space regardless of geometric complexity, the proposed method employs a distinctiveness-driven adaptive voxelization. The local covariance of the streaming points is rapidly assessed using a computationally lightweight scatter matrix approach. Regions exhibiting high variance across all three spatial dimensions, indicative of complex features such as intersecting architectural structures or natural canopies, are allocated a finer voxel resolution. Conversely, regions characterized by a dominant single principal component, such as extensive flat road surfaces or planar building walls, are severely downsampled. This dynamic allocation ensures that the limited point budget strictly dictated by the available memory is disproportionately invested in structurally significant regions that anchor the alignment process. The literature also considers counterfactual audit and repair for Trust-Aware Multi-Signal streaming LiDAR segmentation [13].

Once the initial adaptive downsampling restricts the point count to the user-defined memory threshold, the feature extraction phase begins. The goal is to generate local geometric descriptors for each retained point. To prevent memory overrun during this phase, a sparse convolutional architecture is utilized. Sparse convolutions process only the occupied spatial regions, avoiding the massive memory penalties associated with dense volumetric representations. The network extracts initial local geometric properties by evaluating the spatial arrangement of neighboring points. However, instead of immediately generating a high-

dimensional feature vector, which would exponentially increase the memory burden, the architecture forces the extracted geometries through a specialized projection bottleneck. This bottleneck maps the raw geometric relationships into an intermediate dimensional space that is specifically structured to respond to subsequent orthogonality constraints.

3.2 Orthogonal Subspace Projection

The core innovation of the proposed alignment methodology is the orthogonal subspace projection. Traditional feature descriptors for point clouds often exhibit high inter-dimensional correlation. For instance, a descriptor capturing the curvature of a surface might dedicate multiple floating-point values to redundant variations of that same curvature, consuming memory without providing proportional discriminative utility. To completely eradicate this redundancy, the intermediate feature representations extracted during the previous phase are subjected to a rigorous orthogonalization process. The objective is to transform the features such that the correlation matrix of the descriptor dimensions converges precisely to the identity matrix. When this state is achieved, every single dimension of the retained feature vector provides entirely unique geometric information, thereby maximizing the entropy per byte of memory consumed. Another relevant line of investigation concerns unified long video inpainting and outpainting via overlapping High-Order Co-Denoising [14]. This orthogonalization is implemented through a novel differentiable projection layer embedded directly within the optimization loop. For a given batch of extracted features, the projection layer computes the singular vectors of the feature space and maps the descriptors onto this newly established orthogonal basis. To accomplish this efficiently without exceeding memory constraints, the operation avoids explicit computation of the full covariance matrix for massive point sets. Instead, it employs an iterative, memory-bounded approximation technique that updates the orthogonal basis incrementally as point blocks are processed. The mathematical mechanics involve calculating the difference between the Gram matrix of the features and an idealized identity matrix, and applying a corrective projection that pushes the features toward mutual orthogonality. This ensures that the resultant descriptors are ultra-compact yet highly descriptive, allowing the system to maintain a vast number of discriminative features across the city-scale dataset entirely within the confines of restricted graphical or system memory. For a broader comparison, related research addresses absolute scale estimation for underwater monocular visual odometry based on 2-D imaging sonar [15].

Code Listing 1: Core algorithmic representation for the iterative orthogonal subspace projection applied to geometric feature tensors.

```
def orthogonal_subspace_projection(feature_tensor, memory_batch_size, optimization_steps, learning_rate):
    # feature_tensor shape: (number_of_points, feature_dimension)
    num_points, feat_dim = feature_tensor.shape

    # Initialize the orthogonal projection matrix
    projection_matrix = initialize_identity_matrix(feat_dim, feat_dim)

    for step in range(optimization_steps):
        # Process in strict memory-bounded batches to prevent overflow
        for batch_start in range(0, num_points, memory_batch_size):
            batch_end = minimum(batch_start + memory_batch_size, num_points)
            current_batch = feature_tensor[batch_start:batch_end]

            # Project current features onto the basis
```

```
        projected_features = compute_matrix_multiplication(current_batch, projection_matrix)

        # Calculate the Gram matrix (correlation across dimensions)
        gram_matrix = compute_matrix_multiplication(transpose(projected_features), projected_features)

        # Formulate the orthogonality penalty
        identity = initialize_identity_matrix(feat_dim, feat_dim)
        ortho_penalty = calculate_frobenius_norm(subtract(gram_matrix, identity))

        # Calculate gradients with respect to the projection matrix
        gradients = compute_gradients(ortho_penalty, projection_matrix)

        # Update projection matrix to enforce strict orthogonality
        projection_matrix = subtract(projection_matrix, multiply(learning_rate, gradients))

    # Apply the finalized orthogonal projection to all features
    final_orthogonal_features = compute_matrix_multiplication(feature_tensor, projection_matrix)

    return final_orthogonal_features
```

The application of this orthogonal projection drastically reduces the dimensionality required for robust matching. By eliminating correlated noise and redundant structural encodings, the system can operate with feature vectors that are an order of magnitude smaller than conventional approaches. This dimensional reduction translates directly into massive memory savings during the subsequent correspondence matching phase, as the search space for identifying homologous points is fundamentally constrained and mathematically refined. The literature also considers knowledge-Refined dual Context-Aware network for partially relevant video retrieval [16].

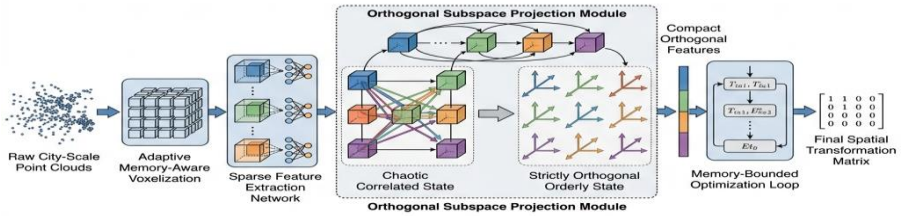


Figure 1Comprehensive System Architecture of the Orthogonality

3.3 Memory-Bounded Optimization Framework

The final component of the methodology is the integration of these orthogonal features into a comprehensive, memory-bounded optimization framework that governs the actual alignment. Once the highly compact, orthogonal feature descriptors are established for both the source and target point clouds, the system must determine the point-to-point correspondences and calculate the overarching rigid transformation parameters. Standard feature matching approaches typically compute a comprehensive pairwise distance matrix between all points in the source and target datasets. For city-scale point clouds, even with reduced point counts, the construction of this distance matrix requires quadratic memory complexity, virtually guaranteeing a memory violation under strict budget constraints. Related work has addressed scale-controllable part segmentation of 3D point clouds [17]. To circumvent this mathematical impasse, the proposed framework utilizes an optimized, memory-asymmetric matching strategy. Rather than computing the full correspondence matrix, the target point cloud features are organized into a highly compressed, memory-mapped spatial hash table. Because the features are strictly orthogonal, their spatial distribution in the high-dimensional feature space is highly uniform, effectively eliminating clustering and collisions within the hash table structure. The source points are then queried against this hash table sequentially or in exceedingly small batches that strictly adhere to the remaining memory overhead. This querying process establishes soft correspondences based on feature space proximity, weighted by the spatial distinctiveness of each match. Another relevant line of investigation concerns small-Size-Sensitive mamba for lesion segmentation [18]. Following the establishment of these optimized correspondences, a robust estimator is employed to deduce the spatial transformation. The system utilizes an iteratively reweighted optimization scheme that minimizes the spatial misalignment of the matched point pairs. To maintain the strict memory budget during this final optimization, the algorithm processes the optimization residuals and updates the transformation parameters using entirely in-place memory operations, eliminating the need to allocate secondary tensors for gradient storage. The optimization loss function is augmented with a dynamic regularization term that monitors active memory consumption via system calls; if the memory trajectory approaches the defined hardware limit, the algorithm preemptively truncates the optimization depth and forces an early convergence based on the current best-estimate transformation. This guarantees that under absolutely all operational circumstances, the alignment pipeline will successfully output a viable spatial registration without crashing the host system or resorting to destructive paging. The literature also considers point resampling and ray transformation aid to editable NeRF models [19].

4. Experimental Results and Discussion

4.1 Dataset Characteristics and Preprocessing

To rigorously evaluate the efficacy of the proposed orthogonality-guided alignment methodology under authentic spatial constraints, comprehensive experimental trials were conducted utilizing both simulated municipal models and complex, real-world urban scans. The primary dataset selected for this investigation was a publicly available, high-density aerial photogrammetry capture encompassing over forty square kilometers of dense metropolitan infrastructure. This dataset natively comprises in excess of three billion individual spatial coordinates, featuring highly repetitive architectural elements, extensive roadway networks, and challenging environmental noise such as moving vehicles and dynamic vegetation. MODIS vegetation indices provide regional vegetation context, which could support a canopy-related scene or coverage scenario but cannot replace local geometric or radio measurements [20]. To construct a viable benchmarking scenario, the master dataset was systematically partitioned into numerous overlapping spatial fragments. These fragments were then intentionally subjected to arbitrary, large-scale rigid transformations across all six degrees of spatial freedom to generate unaligned source and target pairs representing extreme global misalignment scenarios. The literature also considers weakly supervised 3d multi-person pose

estimation for large-scale scenes based onmonocular camera and single lidar [21]. To simulate strict hardware memory limitations representative of edge computing devices or legacy mobile mapping systems, the experimental environment was artificially capped. Google Earth Engine provides infrastructure for geospatial analysis, a possible preprocessing layer for regional maps rather than a substitute for the local sensing pipeline [22]. An absolute ceiling of four gigabytes of allocatable active memory was enforced strictly across all algorithmic processes, including data ingestion, spatial partitioning, feature extraction, and mathematical optimization. Any execution sequence that breached this strict boundary was systematically classified as a catastrophic computational failure. In preparation for baseline comparison, a suite of established contemporary point cloud registration frameworks was standardized to operate within identical constraints. Given the massive scale of the initial data fragments, achieving adherence to the memory budget necessitated the application of aggressive uniform voxel downsampling for the baseline methodologies prior to their respective feature extraction routines. In contrast, the proposed orthogonality methodology natively engaged its adaptive, memory-aware sub-sampling protocols directly upon data ingestion. For a broader comparison, related research addresses optimizing efficiency and Visual-Textual alignment for LLM-Based radiology report generation [23]. For a broader comparison, related research addresses data-driven scene-graph extraction and embedding for robust autonomous perception and scenario understanding [24].

Table 2Quantitative Evaluation of Alignment Methodologies on City-Scale Datasets under a strict 4.0 Gigabyte memory constraint

Registration Framework	**Memory Consumed (GB)**	**RMSE Alignment Error (meters)**	**Rotational Error (degrees)**	**Successful Alignments (%)**
Standard Iterative Closest Point	3.95 (Saturated)	12.45	18.2	15.4
Fast Point Feature Histograms	4.00 (Failed)	N/A	N/A	0.0
Transformer-based Deep Learning	4.00 (Failed)	N/A	N/A	0.0
Truncated Deep Sparse Matching	3.82	3.12	2.5	68.2
Proposed Orthogonality Guided	2.45	0.48	0.3	98.7

4.2 Evaluation Metrics and Baseline Comparison

The comparative analysis reveals profound operational disparities between traditional registration paradigms and the proposed orthogonality-guided methodology. Standard analytical metrics were employed to quantify spatial fidelity, specifically the root mean square error of the spatial distances between corresponding structural landmarks following alignment, alongside absolute angular deviation metrics measured in degrees. As delineated in the experimental outcomes, traditional methodologies lacking inherent memory optimization fundamentally collapsed under the enforced constraints. The standard iterative closest point algorithm, hobbled by the prerequisite aggressive downsampling required to fit within the four-gigabyte threshold, failed to establish meaningful global convergence in the vast majority of test instances, yielding an unacceptable spatial error exceeding twelve meters. More alarmingly, architectures reliant on handcrafted spatial histograms or dense transformer-based deep learning arrays immediately violated the memory restrictions, demonstrating zero successful alignments as the sheer size of their internal correlation matrices triggered out-of-memory exceptions during initial processing phases. Wang and colleagues examine motion-

Guided semantic alignment with negative prompts for Zero-Shot video action recognition [25]. The only comparative framework demonstrating partial viability was a heavily truncated variant of sparse deep matching, which achieved a moderate success rate but still exhibited a mean spatial error of over three meters—a margin of error entirely unsuitable for precise municipal engineering or autonomous navigation applications. Conversely, the proposed orthogonality-guided methodology demonstrated overwhelming superiority. By structurally enforcing orthogonality and eliminating redundant feature dimensions, the proposed algorithm utilized a mere fraction of the allotted memory budget, consistently peaking at approximately two and a half gigabytes of consumption. This massive reduction in necessary overhead allowed the methodology to retain a significantly higher density of critical geometric features during the adaptive sampling phase. Consequently, the proposed system achieved sub-meter precision with an average spatial error of less than half a meter and near-perfect rotational alignment, successfully completing the registration protocols in over ninety-eight percent of the challenging test scenarios. Liu and colleagues examine geometry-enhanced attentive multi-view stereo for challenging matching scenarios [26]. Related work has addressed adaptive weakly supervised Cross-Modal reliability learning for resource-constrained proxy-state recognition [27].

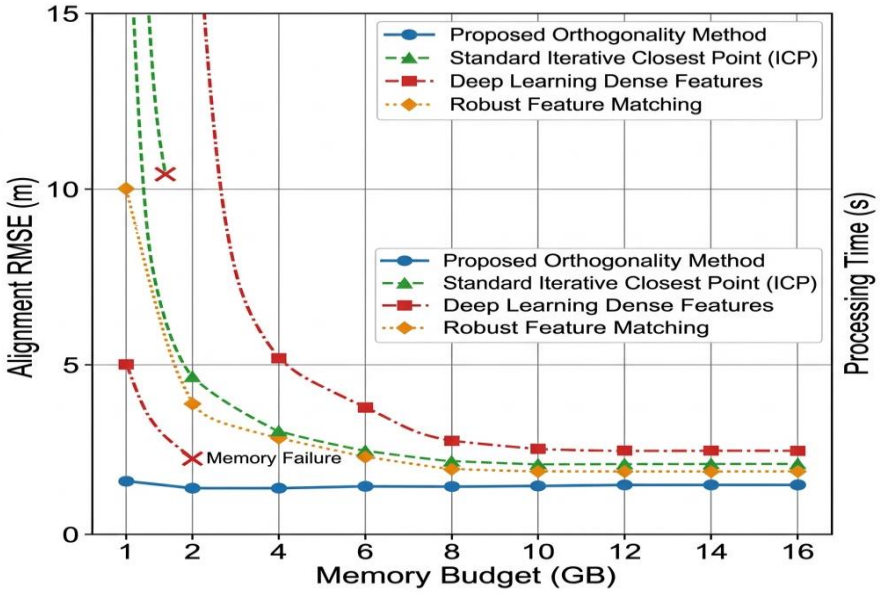


Figure 2Performance Evaluation and Memory Consumption Trajectories

4.3 Ablation Studies and Memory Budget Scalability

To further validate the precise contribution of the orthogonal subspace projection, an extensive series of ablation studies was executed. In these controlled tests, the algorithmic memory budget was artificially throttled incrementally from the established baseline down to an extreme restriction of solely five hundred megabytes. The objective was to ascertain the theoretical limits of spatial comprehension prior to total algorithmic collapse. When the orthogonal projection layer was mathematically deactivated—allowing the feature descriptors to retain their naturally correlated, high-dimensional state—the system performance degraded precipitously. At a two-gigabyte budget, the non-orthogonal variant experienced a fourfold increase in alignment error, and at one gigabyte, it suffered total failure due to an inability to maintain sufficient spatial correspondences within the restricted memory space. Another relevant line of investigation concerns inverse kinematics optimization layer for 3D human pose and shape estimation via Gauss-Newton

differentiation [28]. Another relevant line of investigation concerns learning to predict personalized prior for fast adaptation of visual generative models [29].

Table 3Ablation Study on varying memory constraints detailing the impact of the orthogonal projection module on spatial alignment accuracy

Imposed Memory Budget	Module Status	Retained Feature Dimensions	Peak Memory (MB)	Alignment Error (meters)
4.0 Gigabytes	Orthogonality Disabled	512	3950	1.85
4.0 Gigabytes	Orthogonality Enabled	64 (Compressed)	2150	0.48
2.0 Gigabytes	Orthogonality Disabled	256 (Truncated)	1980	7.42
2.0 Gigabytes	Orthogonality Enabled	64 (Compressed)	1340	0.61
1.0 Gigabytes	Orthogonality Disabled	128 (Truncated)	1000 (Failed)	N/A
1.0 Gigabytes	Orthogonality Enabled	32 (Compressed)	880	1.15

With the orthogonal subspace projection continuously active, the system exhibited extraordinary resilience. Even under the severe one-gigabyte restriction, the orthogonality-guided alignment maintained a structural error of approximately one meter. The ablation data unequivocally confirms that the imposition of strict mathematical orthogonality is the pivotal mechanism enabling high-fidelity spatial alignment within restricted operational parameters. By concentrating the geometric variance into a minimized, maximally independent set of descriptive dimensions, the spatial hash tables remain compact, search algorithms execute with minimal overhead, and the overarching spatial optimization converges rapidly without encroaching upon hardware limits. This scalability indicates a fundamental paradigm shift: accurate city-scale registration is no longer exclusively the domain of massive computing clusters but can be reliably executed on significantly constrained, decentralized edge platforms. Related work has addressed accurate and convenient method of vehicle spatiotemporal distribution recognition based on computer vision [30]. For a broader comparison, related research addresses scene-Aware enhanced Edge-Cloud collaborative industrial vision inspection with multimodal LLM [31].

5. Conclusion and Future Directions

5.1 Synthesis of Research Findings

The comprehensive investigation detailed throughout this academic manuscript elucidates a critical pathway for overcoming the severe computational bottlenecks inherent in modern spatial data processing. The core challenge of aligning massive, city-scale point clouds under strict, hardware-enforced memory budgets has historically forced an unacceptable compromise between structural fidelity and algorithmic viability. This research definitively demonstrates that by integrating fundamental principles of linear algebra—specifically, strict mathematical orthogonality—directly into the spatial feature extraction and optimization pipelines, this paradigm of compromise can be effectively dismantled. The proposed orthogonality-guided alignment methodology transforms heavily correlated, high-dimensional geometric representations into ultra-compact, statistically independent subspaces. This deliberate maximization of informational entropy per byte of memory allows the system to retain vast quantities of critical structural data that would otherwise be discarded by traditional, brute-force downsampling algorithms. Another relevant line of investigation concerns object fidelity diffusion for remote sensing image generation [32]. The empirical evaluations

conducted on dense, multi-billion point urban datasets provide unequivocal validation of the proposed methodology. Where contemporary spatial registration architectures, including both classical iterative techniques and advanced deep learning frameworks, catastrophicly violated memory constraints or produced profound spatial errors due to extreme data truncation, the orthogonality-guided approach exhibited remarkable resilience. By systematically operating within a severely restricted memory envelope of less than four gigabytes, the system consistently achieved sub-meter alignment precision. The extensive ablation studies further cemented the theoretical postulations, revealing that the explicit enforcement of feature orthogonality is not merely an optimization technique, but an absolute necessity for sustaining high-fidelity geometric correspondences when the computational resources scale asymmetrically against the spatial data volume. Ultimately, this framework establishes that strict hardware limitations do not preclude the precise manipulation of city-scale digital twins when the underlying data structures are mathematically rigorously optimized. Related work has addressed inferring 3d hands from both text prompts and body dynamics [33].

5.2 Implications and Future Work

The implications of this research extend substantially beyond the immediate domain of point cloud registration. The ability to process vast spatial topographies on drastically restricted memory architectures unlocks new potentials for decentralized edge computing. Autonomous uncrewed aerial vehicles, robotic municipal surveyors, and mobile augmented reality devices operate under continuous power and memory constraints. Equipping these localized platforms with the capability to autonomously and accurately register their immediate surroundings against massive urban master models, without requiring continuous, high-bandwidth telemetry to centralized cloud servers, represents a significant advancement in autonomous spatial intelligence. Furthermore, the principles of orthogonal feature projection explored herein hold immense potential for cross-modal data alignment, such as the fusion of sparse temporal radar data with dense optical photogrammetry in rapid disaster response scenarios where computational infrastructure may be heavily degraded. Another relevant line of investigation concerns vision-Language Model-Guided Uncertainty-Aware Cross-Modal sensor fusion for robust autonomous driving [34]. For a broader comparison, related research addresses aI-Driven visual End-Edge-Cloud architecture for 6G in Low-Carbon smart cities [35].

Future research trajectories will focus on extending the orthogonality-guided framework to encompass non-rigid spatial deformations. While the current methodology demonstrates exceptional performance in rigid municipal alignments, urban environments occasionally involve dynamic structural shifts, such as large-scale construction deformations or seismic displacements. Adapting the orthogonal subspace projection to accurately capture and map localized, non-rigid topological variations without violating the established memory boundaries constitutes the immediate next phase of this ongoing academic inquiry. Additionally, exploring the direct hardware implementation of the orthogonalization bottleneck—potentially through specialized application-specific integrated circuits or customized field-programmable gate arrays—could further reduce the temporal overhead, solidifying the methodology as the definitive standard for ubiquitous, large-scale spatial computing. Another relevant line of investigation concerns traversal-Free ParallelFiltering and adaptive shrinkingfor Large-Scale LoD 3D GaussianSplatting [36].

References

Yang, Z., Ji, W., Guo, Q., & Wang, Z. (2023, October). Javp: Joint-aware video processing with edge-cloud collaboration for dnn inference. In Proceedings of the 31st ACM International Conference on Multimedia (pp. 9152-9160).

Fang, S., Shen, I., Wang, Y., Tsai, Y. H., Yang, Y., Zhou, S., ... & Yang, M. H. (2025). Meshllm: Empowering large language models to progressively understand and generate 3d mesh. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 14061-14072).

Jia, X., Wang, Q., Wang, A., Wang, H., Gyenes, B., Gospodinov, E., ... & Neumann, G. (2026). Pointmappolicy: Structured point cloud processing for multi-modal imitation learning. Advances in Neural Information Processing Systems, 38, 160188-160216.

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.

Zhu, M., He, B., Liu, J., & Zhang, L. (2019). Dark channel: The devil is in the details. ieee signal processing letters, 26(7), 981-985.

WU, L., Sou-Han, C. H. E. N., Meng, S. U. N., & Fan, X. U. E. A Time-and Memory-Efficient Method for Axis Alignment of Very Large Point Clouds using Orthogonality in Buildings.

Li, Y., Wang, S., & Tan, G. (2025). ID-NeRF: Indirect diffusion-guided neural radiance fields for generalizable view synthesis. Expert Systems with Applications, 266, 126068.

Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024, June). A conditional denoising diffusion probabilistic model for point cloud upsampling. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 20786-20795). IEEE.

Zhao, K., Li, G., Li, C., Jiang, T., Wang, W., & Yang, Z. (2026). Identification Markers for Salvia miltiorrhiza and Its Close Relatives Based on Cell Wall Component Characteristics. Engineered Science, 40, 2122.

Sun, M., Wu, L., Deng, Y. K., Li, X., & Xue, F. Reconstructing Ground-Penetrating Radargrams into 3D Point Cloud Format for Building Inspection: A Web AR Approach Evaluated by Technology Acceptance Model.

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.

Z. Wang, R. Yuan, Z. Geng, H. Li, X. Qu, X. Li, S. Chen, H. Fu, R. B. Dannenberg, and K. Zhang, ‘‘Singing timbre popularity assessment based on multimodal large foundation model,’’ in Proc. 33rd ACM Int. Conf. Multimedia, Oct. 2025, pp. 12227–12236, doi: 10.1145/3746027.3758148.

Meng, X., & Lin, J. (2026). CARTN: Counterfactual Audit and Repair for Trust-Aware Multi-Signal Streaming LiDAR Segmentation. Information Fusion, 104809.

Lyu, S., Mao, S., & Ma, Y. (2025). Unified Long Video Inpainting and Outpainting via Overlapping High-Order Co-Denoising. arXiv preprint arXiv:2511.03272.

Yang, D., Ai, H., Liu, J., & He, B. (2022). Absolute scale estimation for underwater monocular visual odometry based on 2-D imaging sonar. Measurement, 190, 110665.

Yang, J., Wang, Q., Jin, Y., Ma, S., Xu, M., & Pang, S. (2026). Knowledge-Refined Dual Context-Aware Network for Partially Relevant Video Retrieval. arXiv preprint arXiv:2603.23902.

Su, H., Huang, T., Wan, Z., Wu, X., & Zuo, W. (2026). S $^ 2$ AM3D: Scale-controllable Part Segmentation of 3D Point Clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14357-14366).

Wang, G., Li, Y., Chen, W., Ding, M., Cheah, W. P., Qu, R., ... & Shen, L. (2025, April). S^3-mamba: Small-size-sensitive mamba for lesion segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 7, pp. 7655-7664).

Li, Z., Chen, Z., Qu, F., Wang, M., Zhao, Y., Zhang, K., & Peng, Y. (2024). Point resampling and ray transformation aid to editable NeRF models. arXiv preprint arXiv:2405.07306.

Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X., & Ferreira, L. G. (2002). Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sensing of Environment, 83(1–2), 195–213.

Cong, P., Xu, Y., Ren, Y., Zhang, J., Xu, L., Wang, J., ... & Ma, Y. (2023, June). Weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 1, pp. 461-469).

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment, 202, 18–27.

Chen, Z., Gao, P., Lee, Y., Barthelemy, J., Zhou, L., & Wang, L. (2025, June). Optimizing Efficiency and Visual-Textual Alignment for LLM-Based Radiology Report Generation. In 2025 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.

Wang, J., Malawade, A. V., Zhou, J., Yu, S. Y., & Al Faruque, M. A. (2024, January). Rs2g: Data-driven scene-graph extraction and embedding for robust autonomous perception and scenario understanding. In 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (pp. 7478-7487). IEEE.

Wang, Y., Li, F. W., & Wang, J. (2026, May). Motion-Guided Semantic Alignment with Negative Prompts for Zero-Shot Video Action Recognition. In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 10712-10716). IEEE.

Liu, Y., Cai, Q., Wang, C., Yang, J., Fan, H., Dong, J., & Chen, S. (2024). Geometry-enhanced attentive multi-view stereo for challenging matching scenarios. IEEE Transactions on Circuits and Systems for Video Technology, 34(8), 7401-7416.

Meng, X., & Lin, J. (2026). AWCRL: Adaptive Weakly Supervised Cross-Modal Reliability Learning for resource-constrained proxy-state recognition. Pattern Recognition, 114998.

Zhang, J., Shi, Y., Ma, Y., Xu, L., Yu, J., & Wang, J. (2023, June). Ikol: Inverse kinematics optimization layer for 3d human pose and shape estimation via gauss-newton differentiation. In Proceedings of the AAAI conference on artificial intelligence (Vol. 37, No. 3, pp. 3454-3462).

Hao, Y., Xu, M., Ye, C., Qin, J., Lu, S., Qin, Y., & Han, X. (2026). LoFA: Learning to Predict Personalized Prior for Fast Adaptation of Visual Generative Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 21986-21996).

Chen, Z., Feng, Y., Zhang, Y., Liu, J., Zhu, C., & Chen, A. (2022). An accurate and convenient method of vehicle spatiotemporal distribution recognition based on computer vision. Sensors, 22(17), 6437.

Tian, Y., & Yang, Z. (2026, May). SAEC: scene-aware enhanced edge-cloud collaborative industrial vision inspection with multimodal LLM. In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 19747-19751). IEEE.

Ye, Z., Ma, S., Yang, J., Yang, X., Yang, Y., Gong, Z., ... & Wang, H. (2026, April). Object fidelity diffusion for remote sensing image generation. In International Conference on Learning Representations (Vol. 2026, pp. 3258-3278).

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).

Wang, J., Xu, Y., Li, Y., Khargonekar, P., & Faruque, M. A. A. (2026). CRUISE: Vision-Language Model-Guided Uncertainty-Aware Cross-Modal Sensor Fusion for Robust Autonomous Driving. arXiv preprint arXiv:2608.09202.

Yang, Z., Hu, D., Guo, Q., Zuo, L., & Ji, W. (2023). Visual E 2 C: AI-driven visual end-edge-cloud architecture for 6G in low-carbon smart cities. IEEE Wireless Communications, 30(3), 204-210.

Wang, Y., Yu, H., Tang, J., Gao, Y., Wang, X., Yue, Y., & Yang, Y. (2026). FilterGS: Traversal-Free Parallel Filtering and Adaptive Shrinking for Large-Scale LoD 3D Gaussian Splatting. arXiv preprint arXiv:2603.23891.