Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ30650

Analysis of 3D Perception Models Based on the Mamba Architecture

Zihao Li

School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, Singapore
lizi0038@e.ntu.edu.sg

Abstract. Object detection and object tracking constitute core tasks in computer vision, aimed at identifying and localizing objects belonging to predefined categories within a scene. In recent years, the advent of the Mamba architecture has attained a significant milestone in deep learning. By harnessing State Space Models (SSMs), Mamba achieves linear computational complexity and superior long-range dependency modeling, in contrast to the quadratic complexity inherent in traditional Transformer architectures. Consequently, a growing body of researchers are applying Mamba to the domain of three-dimensional (3D) point clouds to improve processing efficiency. However, owing to the intrinsic sparsity, irregularity, and unstructured characteristics of point cloud data, the direct application of 1D sequential models to 3D spatial data confronts substantial challenges, particularly regarding data serialization and local feature preservation. To help researchers gain a comprehensive understanding of the current status and latest advancements in this field, this paper systematically reviews the recent research progress in 3D point cloud algorithms based on the Mamba architecture. Furthermore, this survey analyzes existing limitations regarding geometric information loss and interpretability. It concludes by delineating potential future research directions, such as learnable serialization strategies and hybrid architectures, aiming to provide a foundational reference for developing next-generation, efficient 3D perception systems.

Keywords: Mamba Architecture, 3D Point Clouds, Object, DetectionState, Space Models

1. Introduction

Object detection has long held a pivotal role in computer vision. As a cornerstone of image understanding, it finds pivotal applications in domains including robotics, unmanned aerial vehicles (UAVs), and autonomous driving. Concurrently, the rapid advancement of point cloud technology has spurred the proposition and practical implementation of numerous effective methods.

For instance, BEVFusion [1] fuses LiDAR and camera data by projecting point clouds onto a 2D plane via voxelization and mapping camera data to 2D using depth prediction, ultimately fusing these representations to obtain information-rich planar data. VoxelNeXt [2], a sparse-voxel-based network, performs voxel normalization to extract non-empty sets, which are then fed directly into a convolutional neural network (CNN) for feature extraction and classification. PointPillars [3]

DOI: 10.54254/2755-2721/2026.TJ30650

realized the transformation of point clouds into pseudo-images by partitioning the point cloud and mapping it into a compact 2D feature representation.

In contemporary research, the Transformer architecture has become a focal point for its superior global context modeling and cross-modal interaction capabilities. However, despite these advantages, it encounters prominent bottlenecks in processing large-scale 3D data, specifically concerning its quadratic computational complexity and deficiencies in geometric awareness. Addressing these fundamental limitations, the Mamba architecture has emerged as a revolutionary alternative for 3D point cloud processing. Unlike Transformers, Mamba is capable of capturing long-range dependencies between points with linear complexity through a selective state space model and a dynamic propagation mechanism, yielding substantially reduced computational and memory overheads. Consequently, this architecture is being increasingly adopted in 3D vision tasks, with recent models such as SegMamba, 3DET-Mamba, and 3DSS-Mamba demonstrating its potential to achieve an optimal balance between detection accuracy and inference efficiency.

To provide a systematic understanding of this emerging field, this paper investigates the application of the Mamba architecture in 3D point cloud perception. We first analyze the limitations of Transformer-based methods and the motivation for adopting State Space Models. Subsequently, we comprehensively review existing Mamba-based algorithms, categorizing them based on their serialization strategies and feature interaction mechanisms. Finally, we discuss unresolved challenges and outline potential future research directions. This survey aims to serve as a vital reference for researchers, facilitating the development of more efficient and robust 3D perception systems.

2. Literature review

2.1. Transformer-based 3D perception

The Transformer architecture has recently become a focal point in 3D perception due to its superior global context modeling capabilities. Researchers have proposed various methods to adapt Transformers for point cloud processing, addressing challenges such as data sparsity and computational cost.

CenterFormer [4] introduced a center-point-based Transformer network. It first adopts a standard voxel-based encoder to represent the point cloud as Bird's-Eye-View (BEV) features. A multi-scale center proposal network then predicts center locations via heatmaps. These center features serve as query embeddings for a decoder, which utilizes deformable cross-attention layers to aggregate neighborhood features for the final prediction. While this approach reduces the computational overhead, the cross-attention overhead remains substantial when a large number of center points exist, making it difficult to maintain high frame rates in inference-sensitive scenarios. Furthermore, Transformers are inherently better suited for sequential and 2D features, resulting in a potential loss of 3D geometric relationships and enduring challenges in addressing point cloud alignment.

Voxel Set Transformer (VSA) [5] feeds voxel features into a feature extraction network and subsequently into a VSA module. This module innovatively circumvents direct self-attention through the introduction of a latent space. It substitutes self-attention with a dual cross-attention mechanism to reduce the quadratic complexity and adaptively build global dependencies. This structure is then stacked to achieve multi-level interaction for object prediction. However, critical issues remain: voxelization inherently causes information loss, the latent space functions as a "black box" that hinders the visualization of the model's attention regions, and contextual communication between distinct voxels remains limited.

DOI: 10.54254/2755-2721/2026.TJ30650

Transfusion [6] is a Transformer-based multi-modal fusion architecture for LiDAR and camera data. It proposes a "soft-association" mechanism to learn the intricate relationships between point cloud and image features. The Transformer module first predicts initial object boxes based on LiDAR features. These initial proposals are then refined using image data before being converted into 3D bounding boxes. The main drawbacks are that this model is prone to introducing noise, demands large-scale datasets to achieve stable performance, and its computational cost remains relatively high.

2.2. Problem definition and key challenges

Despite these advancements, applying Transformers to 3D perception is fraught with challenges, primarily stemming from the unique properties of point cloud data.

The core challenge in adapting Transformers for 3D perception originates from the inherent contradiction between the unstructured characteristics of point clouds and the sequential design of the architecture. Point cloud data is typically massive, sparse, and highly irregular, representing a significant departure from the dense, grid-like structure of 2D images. Since the Transformer architecture is primarily tailored for modeling sequential or 2D planar features, it is inherently less proficient in modeling 3D spatial relationships. Consequently, directly applying it to 3D data often results in weaker geometric modeling capabilities, as the standard architecture struggles to effectively encode the complex topology and neighborhood information inherent in volumetric data.

Beyond these structural discrepancies, computational limitations and data inhomogeneities impose significant bottlenecks. The quadratic complexity of standard self-attention mechanisms renders it computationally intractable given the vast number of points or voxels in a typical 3D scene, limiting the model's scalability. This issue is exacerbated by the significant variation in point cloud density; as the distance from the sensor increases, the point density decreases drastically, making stable feature capture and precise position estimation extremely difficult. Even with algorithms designed to reduce computational load, processing such non-uniform data remains highly resource-intensive.

Furthermore, robust 3D perception necessitates balancing subtle trade-offs in feature aggregation and multi-modal fusion. A critical dilemma lies in balancing the receptive field: focusing on excessively small local regions fails to capture sufficient semantic context, whereas attending to extensive global regions tends to introduce noise and ambiguity. This complexity is further compounded in multi-modal contexts involving LiDAR and camera data. "Hard" association methods based on rigid matching are typically error-prone owing to calibration inconsistencies, while "soft" feature-level fusion strategies risk introducing significant noise, thereby hindering the overall detection accuracy.

2.3. Emergence of Mamba architecture

To address the limitations of Transformers, Mamba-based 3D detection algorithms have been proposed. Mamba, derived from Structured State Space Models (SSMs) [7], was originally designed to mitigate the limitations of Transformers in modeling ultra-long sequences.

The core idea of an SSM is to model sequence dependencies using the state transition equation of a continuous-time system, where each time step depends only on the previous state and the current input. Mamba [8] builds on this foundation by proposing a dynamic, input-adaptive selection mechanism to decide which information to retain or discard.

DOI: 10.54254/2755-2721/2026.TJ30650

It achieves linear complexity O_N by utilizing parallel scanning and linear recurrence, enabling long-sequence dependency modeling without constructing an explicit attention matrix. Considering the spatial continuity of point clouds, they can be transformed into a spatial sequence through scanning, allowing Mamba to model spatial features more naturally. This linear complexity substantially reduces the memory overhead. Mamba also incorporates gating mechanisms to autonomously decide the update intensity, enhancing robustness to noise. Furthermore, its inherent temporal memory makes it a more stable and effective backbone for multi-frame fusion or sequential detection tasks.

3. Analysize of Mamba

3.1. Review of Mamba-based methods

The introduction of Mamba [8] has catalyzed a new wave of research in 3D perception, seeking to exploit its linear complexity and global modeling capabilities while tackling the unique challenges posed by point cloud data.

VoxelMamba [9] processes voxelized point clouds by initially adopting a Hilbert curve to preserve the spatial locality of voxels. Non-empty voxels are encoded and ordered based on the Hilbert curve prior to being fed into the SSM. This method proposes positional vectors as an "implicit window," abolishing the dependence on explicit window partitioning. By employing bidirectional modeling (a forward SSM for high-resolution features and a backward SSM for global context), VoxelMamba enhances its receptive field and spatial awareness, all while supporting real-time inference.

3DET-Mamba [10] proposes an end-to-end 3D object detection architecture. It initially extracts lightweight features, which are then processed by a lightweight Mamba backbone. To balance local and global information, it employs both Farthest Point Sampling (FPS) to capture a global field of view and Neighborhood Point Sampling (NPS) to maintain local continuity. These two sequences are fused via a gating mechanism. Finally, 'M' key points are selected to cover object regions, serving as queries for the decoder to adaptively focus on scene-specific areas. This approach avoids training instability, achieves a balance between local and global feature extraction, and enhances computational efficiency and geometric awareness by eliminating window slicing.

Serialized Point Mamba [11] presents a voxel-free approach. It projects 3D point clouds onto a grid but avoids fixed voxelization, using the projection as a local aggregation method instead. To preserve geometric information, it integrates positional encoding derived from Sparse Submanifold Convolutions. The method explores various sorting strategies to preserve different spatial orders. By integrating the selective Mamba mechanism, it enhances robustness. Each sequence undergoes normalization after feature extraction by the Mamba module.

UniMamba [12] is a voxel-based 3D detection backbone that innovatively and effectively combines a convolutional architecture with the Mamba architecture. Recognizing that SSMs, despite their proficiency in capturing global information, may disrupt local neighborhood information, UniMamba prepends a sparse convolutional layer before the SSM. This initial layer efficiently extracts neighborhood features and strengthens the spatial inductive bias. The features are subsequently serialized using a z-order curve and processed through dual modeling to generate the final feature representation.

DOI: 10.54254/2755-2721/2026.TJ30650

3.2. Summary and analysis

The Mamba architecture, as a selective state space model, replaces the attention mechanism of Transformers with state recursion and dynamic gating. Its core advantages reside in linear computational complexity O_N and efficient long-range global dependency modeling.

However, applying this architecture to 3D point clouds poses a prominent core challenge: point clouds are inherently sparse, and their spatial characteristics are prone to degradation during spatial-to-sequential transformation. Current Mamba-based methods primarily focus on resolving this challenge, specifically on how to preserve spatial properties and achieve better perceptual features during the space-to-sequence transformation.

To address the central challenge of applying Mamba to point clouds, prevailing strategies primarily focus on robust space-to-sequence conversion. This involves employing various encoding methods, such as Hilbert curves, Z-order curves, or sampling techniques like Farthest Point Sampling (FPS) and Neighborhood Point Sampling (NPS), to project 3D spatial data into a 1D sequence while mitigating the degradation of spatial locality. To further compensate for the inevitable information loss during serialization, researchers integrate window priors, infusing local contextual information either explicitly through window partitioning or implicitly via positional encodings.

Beyond serialization, optimizing the scanning mechanism and feature fusion is of paramount importance. Since a single-direction scan is often insufficient for comprehensive 3D spatial understanding, contemporary methods utilize multi-directional processing, deploying bidirectional or multi-directional State Space Models (SSMs) to ensure reliable feature extraction throughout the geometric structure. Furthermore, to efficiently fuse these heterogeneous information streams, models adopt gated fusion mechanisms. These mechanisms allow for the selective fusion of features derived from different scan directions, sampling resolutions (e.g., local versus global), or distinct modalities, ensuring that the most relevant information is retained for final prediction.

4. Discussion

4.1. Unresolved issues and challenges

While Mamba-based architectures exhibit substantial potential, several key challenges must be addressed to tap into their full potential in 3D point cloud processing.

Serialization strategies persist as a core bottleneck, as current methods each have inherent limitations. The Hilbert curve strategy, while effective at preserving locality, is complex to implement and highly sensitive to dimensionality. Conversely, the Z-order curve strategy is efficient but offers inferior neighborhood preservation compared to Hilbert. Sampling strategies like FPS and NPS are strongly reliant on the input data distribution and are thus prone to noise interference.

Furthermore, a fundamental challenge lies in Mamba's core architecture. Its primary strength is in modeling long-range, global dependencies, while its capability for extracting fine-grained local features remains comparatively weak. How to efficiently acquire fine-grained local features within the SSM framework remains an open research question.

Finally, Mamba-based models face distinct difficulties regarding interpretability, stability, and theoretical understanding. A primary concern is the lack of transparency in the mechanism, which makes it difficult to visualize or determine precisely which specific points the model focuses on, rendering the architecture somewhat of a "black box." In addition to this interpretability issue, the dynamic gating mechanism may demonstrate training instability during the training process.

DOI: 10.54254/2755-2721/2026.TJ30650

Furthermore, the inductive biases inherent within the coefficient space of the State Space Model (SSM) have not been fully elucidated, posing a latent challenge that complicates model design and optimization.

4.2. Future research directions

Building on the aforementioned challenges, future research in this field can pursue several promising avenues:

One primary avenue for future research involves the evolution of serialization techniques. Instead of relying on fixed, hand-crafted serialization curves like Hilbert or Z-order patterns, future work could prioritize the development of learnable encoding strategies. There is significant potential in training geometry-aware serialization policies—potentially utilizing Transformer encoders or other machine learning models—to adaptively convert 3D space into an optimal 1D sequence, thereby preserving complex spatial information more effectively than static methods.

Concurrently, to address the inherent limitation of Mamba in local feature extraction, investigating hybrid architectures constitutes a promising avenue. The SSM backbone can be efficiently enhanced by lightweight convolutional layers or self-attention mechanisms specifically designed to process local neighborhoods. This hybrid design would allow the system to leverage Mamba for efficient global context modeling while utilizing a separate, specialized mechanism to capture fine-grained local details, ensuring a balance between macroscopic understanding and microscopic precision.

Finally, improving model interpretability remains a pivotal goal. The integration of traditional self-attention mechanisms, even in a limited or local capacity, could serve to improve the transparency of these models. By enabling the generation of explicit attention maps for local regions, researchers could gain clearer insights into the model's decision-making process and identify specific focus factors within the point cloud data, transcending the "black box" characteristic of existing implementations.

5. Conclusion

This study endeavors to offer a comprehensive review of the emerging applications of the Mamba architecture in 3D point cloud perception, with a specific focus on addressing the limitations of conventional Transformer-based approaches. The analysis reveals that Mamba models, by leveraging State Space Models (SSMs) and dynamic gating mechanisms, successfully mitigate the quadratic computational bottlenecks of self-attention while retaining superior long-range dependency modeling. Our findings suggest that through innovative serialization strategies and hybrid architectural configurations, Mamba achieves a compelling trade-off between detection accuracy and inference efficiency.

Theoretically, this research contributes to the existing body of knowledge by systematically categorizing the nascent field of Mamba-based 3D vision, filling a critical gap in the literature that has predominantly focused on 2D applications. Practically, this study carries substantial implications for autonomous driving and robotics, delineating feasible pathways for deploying high-performance perception models on resource-constrained hardware.

However, this review is constrained by the nascent stage of this specific domain. The available pool of Mamba-based 3D algorithms is currently smaller than that of established architectures, limiting the scope of extensive quantitative cross-comparisons. Furthermore, the analysis depends largely on recent developments that are still evolving. Future research should prioritize the

DOI: 10.54254/2755-2721/2026.TJ30650

development of learnable, geometry-aware serialization strategies and interpretable visualization methodologies to address current "black box" constraints. Overall, by shedding light on the paradigm shift toward linear-complexity modeling, this research paves the way for the development of next-generation, real-time, and robust 3D perception systems.

References

[1] Y. Liu, T. Tang, H. M. Y. T. L. A. M. and O. Beijbom, "BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation, " arXiv preprint arXiv: 2205.13542, 2022.
[2] Y. Chen, P. Wang, J. Yang, Q. Ma, and Y. Wang, "VoxelNeXt: Fully Sparse VBoxel-based 3D Object Detection and Tracking, " in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.
[3] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast Encoders for 3D Object Detection, " in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[4] Z. Zhou, X. Zhao, Y. Wang, P. Wang, and H. Foroosh, "CenterFormer: Center-based Transformer for 3D Object Detection, " in European Conference on Computer Vision (ECCV), 2022, pp. 487–504.
[5] C. He, H. Li, Y. Li, S. Wang, H. Wang, and L. Zhang, "Voxel Set Transformer: A Set-to-Set Approach to 3D Object Detection from Point Clouds, " in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11002–11011.
[6] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "TransFusion: Robust LiDAR-Camera Fusion for 3D Object Detection with Transformers, " in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1090–1099.
[7] A. Gu, K. Goel, and C. Ré, "Efficiently Modeling Long Sequences with Structured State Spaces, " arXiv preprint arXiv: 2111.00396, 2021.
[8] A. Gu and T. Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces, " arXiv preprint arXiv: 2312.00752, 2023.
[9] G. Zhang, L. Fan, C. He, Z. Lei, Z. Zhang, and L. Zhang, "Voxel Mamba: Group-Free State Space Models for Point Cloud based 3D Object Detection, " arXiv preprint arXiv: 2406.10700, 2024.
[10] C. Xia, Y. Zhao, L. An, G. Chen, Y. Wang, and T. Luan, "3DET-Mamba: State Space Model for End-to-End 3D Object Detection, " in Advances in Neural Information Processing Systems (NeurIPS), 2024.
[11] T. Wang, W. Wen, J. Zhai, K. Xu, and H. Luo, "Serialized Point Mamba: A Serialized Point Cloud Mamba Segmentation Model, " arXiv preprint arXiv: 2407.12319, 2024.
[12] X. Jin, Y. Li, J. Li, Z. Chen, J. Yang, and Y. Wang, "UniMamba: Unified Spatial-Channel Representation Learning with Group-Efficient Mamba for LiDAR-based 3D Object Detection, " arXiv preprint arXiv: 2503.12009, 2025.