Lightweight MLP-Based Operator for Real-Time Voxel-Based 3D Object Detection

George Albert Bitwire
Graduate School of Electronic and Electrical Engineering, Kyungpook National University
Daegu, Republic of Korea
bitwire@knu.ac.kr

Samuel Kakuba
Graduate School of Electronic and Electrical Engineering, Kyungpook National University
Daegu, Republic of Korea
2021327392@knu.ac.kr

Dong Seog Han*
Graduate School of Electronic and Electrical Engineering, Kyungpook National University
Daegu, Republic of Korea
dshan@knu.ac.kr

***Abstract*—Real-time 3D object detection is essential for autonomous driving, enabling rapid understanding of dynamic environments. While voxel-based methods excel due to their structured representation and high accuracy, they face computational bottlenecks stemming from the costliness of 3D convolutions and transformer attention, which hinders real-time deployment. Unlike linear recurrent neural network designs such as Mamba, we propose a lightweight MLP-based operator that replaces these components with parallelizable feed-forward blocks, enabling efficient, geometry-aware feature learning in sparse voxel backbones. On the KITTI benchmark, the proposed operator achieves accuracy within ≈1% of a Mamba-based baseline across AP_{3D}, AP_{BEV}, AP_{bbox}, and AP_{aos}, while reducing forward latency from 126.58ms to 123.99ms (∼2.0%) and increasing inference throughput from 15.80 to 16.13 samples/s (∼2.1%) under identical settings. These results demonstrate its suitability for scalable, real-time 3D perception in autonomous systems.**

***Index Terms*—3D Object Detection, Autonomous Driving, MLP, Voxel-based, KITTI.**

I. Introduction

Efficient and accurate 3D object detection is crucial for autonomous driving [1], where real-time environmental understanding directly impacts safety. Voxel-based methods dominate due to their ability to convert irregular point clouds into structured 3D grids for effective spatial reasoning. However, their reliance on computationally intensive 3D convolutions, transformer-based attention and linear RNNs limits scalability, particularly on resource-constrained platforms. Additionally, depth-guided attention mechanisms, recently applied in monocular 3D object detection such as MonoDGAE [2] integrate geometric priors and bilateral filtering to enhance robustness, yet still encounter computational bottlenecks in large-scale 3D perception tasks.

Recent efforts to improve efficiency have focused on redesigning 3D detection backbones. Transformer-based approaches, such as DSVT [3], leverage windowed or sparse attention to boost accuracy but remain hindered by the high latency of quadratic complexity. The LION [4] framework adopts linear group RNNs to deliver state-of-the-art results on benchmarks like Waymo, nuScenes, and KITTI, yet its sequential processing limits parallelization and increases training overhead. Alternatively, MLP-based architectures, exemplified by MLP-Mixer [5], have achieved competitive performance in vision tasks without relying on convolution or attention, inspiring the pursuit of lightweight MLP solutions for 3D perception.

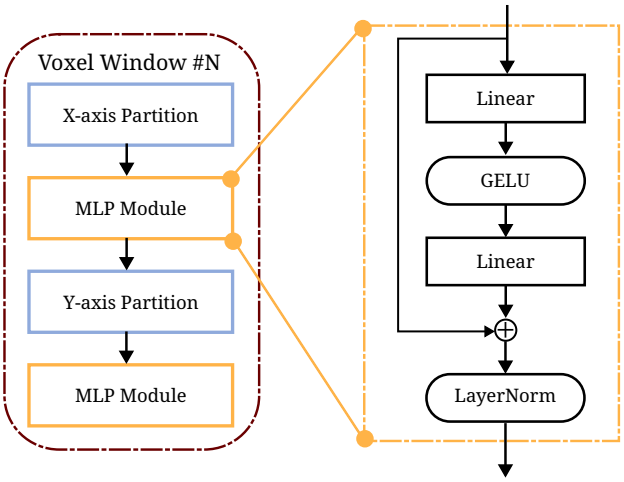


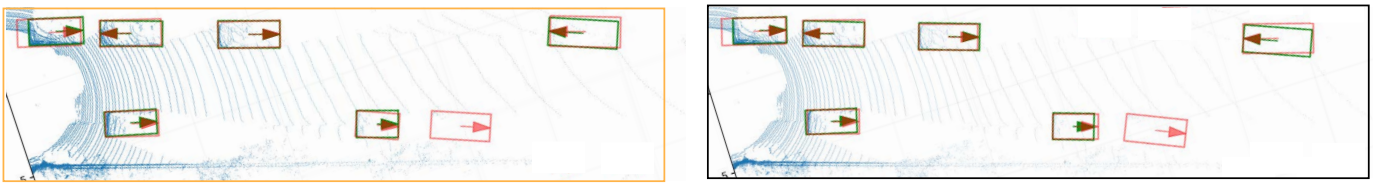
Fig. 1. The MLP-based operator: each voxel window is sequentially partitioned along the X and Y axes, with an MLP module applied after each partition. The MLP module consists of two linear layers with a GELU activation, residual connection, and LayerNorm.

We propose a lightweight MLP-based operator for voxel-based 3D object detection that replaces 3D convolutions, attention, and linear group RNNs. Integrated into a sparse hierarchical backbone, it enables fast, geometry-aware feature learning with an optimal balance of accuracy and efficiency.

II. Method

The proposed MLP-based operator is designed as a lightweight alternative to conventional 3D convolutions, attention mechanisms, and linear group RNNs in voxel-based 3D object detection. Its architecture emphasizes parallelizable spatial modeling while preserving geometric awareness.

Given a sparse voxel tensor $X \in \mathbb{R}^{N \times C}$, where N is the number of non-empty voxels and C is the feature dimension, the input is first divided into fixed-size voxel windows. Each window is processed sequentially along two orthogonal spatial axes to capture directional context. The process begins with an



(a) **(b)**

Fig. 2. Comparison of detection results in BEV for the same scene: **(a)** - MLP-based operator, **(b)** - Mamba-based operator. Green boxes denote ground truth, red boxes denote predictions, with arrow direction indicating estimated heading.

TABLE I
COMPARISON OF AP_{3D}, AP_{BEV}, AP_{bbox}, AND AP_{aos} ON THE KITTI VALIDATION SET (AP R40) FOR CAR, PEDESTRIAN, AND CYCLIST CLASSES.

Class	Method	AP_{3D}			AP_{BEV}			AP_{bbox}			AP_{aos}		
		Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
Car (IoU=0.7)	Mamba-based	87.44	78.60	75.99	91.18	87.63	86.95	95.72	93.71	91.47	95.70	93.58	91.26
	MLP-based (Ours)	87.02	77.74	75.12	91.01	87.27	86.51	95.44	91.91	90.97	95.43	91.76	90.72
Pedestrian (IoU=0.5)	Mamba-based	63.53	57.28	51.21	69.07	62.87	57.64	78.44	74.05	69.87	73.56	68.26	64.02
	MLP-based (Ours)	63.44	55.57	49.56	67.93	60.54	55.05	78.33	72.99	68.59	73.22	67.30	62.87
Cyclist (IoU=0.5)	Mamba-based	82.35	65.24	61.11	86.41	70.15	66.04	91.15	74.70	71.54	90.93	73.55	70.48
	MLP-based (Ours)	84.16	66.18	62.10	88.99	70.59	66.22	91.71	73.72	70.63	91.50	73.13	69.99

X-axis partition, grouping features along the forward direction, followed by an MLP module to model intra-axis dependencies. The resulting features are then partitioned along the Y-axis to aggregate lateral context, after which a second MLP module is applied.

As illustrated in Fig. 1, each MLP module applies a Linear layer, GELU activation, and another Linear layer, adds the result to the original input via a residual connection, and uses Layer Normalization for training stability. Formally, the transformation is expressed as

$$F_{\text{mlp}} = \text{LayerNorm}\left(X + \text{Linear}\left(\text{GELU}\left(\text{Linear}(X)\right)\right)\right) \quad (1)$$

The MLP-based operator replaces heavier layers in a sparse voxel backbone, using windowed feed-forward processing to expand the receptive field efficiently. This design preserves geometric cues, reduces latency, and maintains competitive detection performance.

III. RESULTS AND DISCUSSION

Table I and Fig. 2 compare the proposed MLP-based operator with the Mamba-based baseline on KITTI (AP R40). Across all classes and IoU thresholds, the MLP-based design matches the baseline within $\approx$1% for most AP_{3D}, AP_{BEV}, AP_{bbox}, and AP_{aos} scores.

For Cars at IoU=0.7, AP_{3D} is 87.02%/77.74%/75.12% (Easy/Mod./Hard), with similar trends across other metrics. Pedestrian and Cyclist results show minor drops in some settings but also gains, e.g., Cyclist AP_{BEV} at IoU=0.5 improves by +2.58%.

BEV visualizations show both methods yield similarly accurate boxes and orientations, confirming that the lightweight MLP blocks maintain accuracy while greatly reducing training and inference time for real-time 3D detection.

IV. CONCLUSION

We presented a lightweight MLP-based operator for voxel-based 3D object detection that replaces computationally expensive 3D convolutions, attention and sequential modules with parallelizable feed-forward blocks. Integrated into a sparse hierarchical backbone, the operator achieves accuracy comparable to a Mamba-based baseline on the KITTI benchmark while significantly reducing training and inference time. These results highlight its potential as an efficient backbone component for real-time autonomous driving systems. Future work will explore integrating gating mechanisms into the MLP-based operator for adaptive computation, scaling to larger datasets, and extending to multi-modal 3D perception.

V. ACKNOWLEDGMENT

This study is the result of the research performance of "Defense SMEs Competency Enhancement Program" (NO.DC2023CS) project supported by "Korea Research Institute for defense Technology planning and advancement"

REFERENCES

[1] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The international journal of robotics research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[2] G. A. Bitwire, S. Kakuba, D. W. Cha, and D. S. Han, "Monodgae: depth-guided attention and bilateral filtering for robust monocular 3d object detection," *Artificial Life and Robotics*, pp. 1–12, 2025.

[3] H. Wang, C. Shi, S. Shi, M. Lei, S. Wang, D. He, B. Schiele, and L. Wang, "Dsvt: Dynamic sparse voxel transformer with rotated sets," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 520–13 529.

[4] Z. Liu, J. Hou, X. Wang, X. Ye, J. Wang, H. Zhao, and X. Bai, "Lion: Linear group rnn for 3d object detection in point clouds," *Advances in Neural Information Processing Systems*, vol. 37, pp. 13 601–13 626, 2024.

[5] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit *et al.*, "Mlp-mixer: An all-mlp architecture for vision," *Advances in neural information processing systems*, vol. 34, pp. 24 261–24 272, 2021.