The 2nd Place Solution from the 3D Semantic Segmentation Track in the 2024 Waymo Open Dataset Challenge

Qing Wu
Marvell Technology
qing20210419@gmail.com

Abstract

3D semantic segmentation is one of the most crucial tasks in driving perception. The ability of a learning-based model to accurately perceive dense 3D surroundings often ensures the safe operation of autonomous vehicles. However, existing LiDAR-based 3D semantic segmentation databases consist of sequentially acquired LiDAR scans that are long-tailed and lack training diversity. In this report, we introduce MixSeg3D, a sophisticated combination of the strong point cloud segmentation model with advanced 3D data mixing strategies. Specifically, our approach integrates the MinkUNet family with LaserMix and PolarMix, two scene-scale data augmentation methods that blend LiDAR point clouds along the ego-scene's inclination and azimuth directions. Through empirical experiments, we demonstrate the superiority of MixSeg3D over the baseline and prior arts. Our team achieved 2nd place in the 3D semantic segmentation track of the 2024 Waymo Open Dataset Challenge.

1. Introduction

The rapid advancement of autonomous driving technology has underscored the critical importance of accurate and efficient perception systems [10]. Among the various tasks involved in driving perception, 3D semantic segmentation stands out as a pivotal component [2, 9, 20]. It involves the classification of each point in a LiDAR point cloud into distinct semantic categories, such as vehicles, pedestrians, and road surfaces, thereby enabling the autonomous vehicle to understand and navigate its environment safely [3].

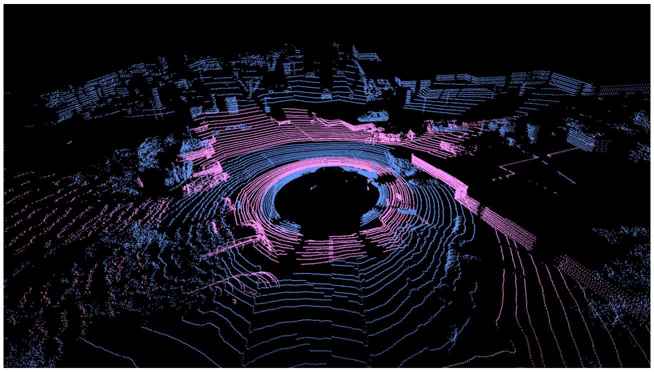
LiDAR sensors have become a cornerstone in autonomous driving due to their ability to provide high-resolution 3D information about the surrounding environment [4]. Unlike cameras, which capture 2D images, LiDAR sensors emit laser pulses and measure the time it takes for the pulses to return after hitting an object [8, 19]. This process generates detailed 3D point clouds that represent the spatial structure of the environment [1, 5, 13, 17, 24].

3D semantic segmentation of LiDAR point clouds is crucial for various autonomous driving tasks, including object detection, scene understanding, and navigation [16, 26]. The challenge lies in accurately identifying and categorizing each point in the cloud, which requires sophisticated algorithms capable of handling large-scale data and diverse environmental conditions. Traditional methods often struggle with the complexity and variability of real-world driving scenarios, highlighting the need for advanced approaches [12].

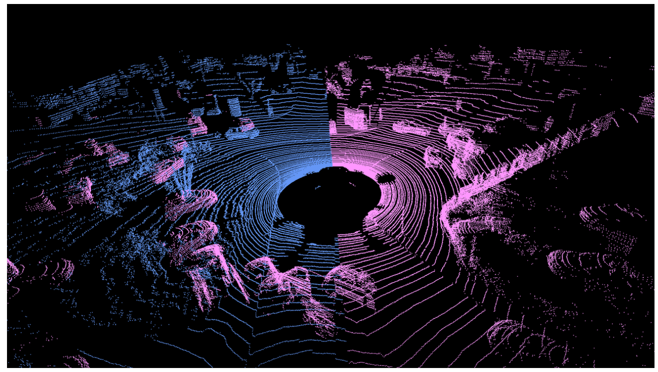
The 2024 Waymo Open Dataset Challenge [20] provides a competitive platform for exploring and advancing solutions in this domain. The challenge focuses on segmenting LiDAR point clouds into semantic classes, aiming to push the boundaries of current methodologies and foster innovation. This competition attracts top research teams worldwide, encouraging the development of novel approaches and the exchange of cutting-edge ideas.

Since the LiDAR scans are often sequentially acquired, the consecutive point clouds share overlapped class and spatial distributions, which might lead to sub-par performance. In this report, we present our approach, **MixSeg3D**, which combines a strong point cloud segmentation model with advanced 3D data mixing strategies. Our method integrates the MinkUNet family [6] with LaserMix [15] and PolarMix [25], two novel scene-scale data augmentation techniques. MinkUNet [6], known for its efficient sparse convolutional operations, provides a robust backbone for our segmentation framework. LaserMix [15] and PolarMix [25] blend LiDAR point clouds along the ego-scene's inclination and azimuth directions, respectively, enhancing the LiDAR segmentation model's ability to generalize across different environments [11, 14].

By incorporating LaserMix [15] and PolarMix [25], we introduce additional variations in the training data, allowing the MinkUNet model to learn from a wider range of scenarios. This approach improves segmentation accuracy and enhances the model's robustness to domain shifts and out-of-distribution data. Our empirical studies demonstrate that **MixSeg3D** outperforms the baseline and prior meth-



(a) LaserMix



(b) PolarMix

Figure 1. Illustrative examples of LiDAR scene augmentations using the (a) LaserMix [15] and (b) PolarMix [25] strategies.

ods, achieving superior segmentation accuracy. We achieved 69.83% mIoU on the leaderboard, culminated in securing 2nd place in the 3D semantic segmentation track of the 2024 Waymo Open Dataset Challenge.

2. Approach

Recent years have seen significant progress in the field of 3D semantic segmentation, driven by the development of powerful deep learning models and the availability of large annotated datasets. Neural networks, particularly convolutional neural networks and their 3D variants [21–23, 27], have demonstrated remarkable capabilities in learning rich representations from point cloud data. Architectures such as MinkUNet [6] have set new benchmarks in 3D semantic segmentation by leveraging hierarchical feature extraction and efficient spatial representations.

2.1. Revisiting MinkUNet

MinkUNet [6] is a widely recognized architecture for 3D semantic segmentation due to its efficient sparse convolutional operations and robust hierarchical feature extraction. The architecture employs sparse tensors to handle the high dimensionality and sparsity of LiDAR point clouds, making it computationally efficient while maintaining high accuracy.

MinkUNet's core design revolves around the use of Minkowski Engine, which facilitates the computation of sparse convolutions and offers significant improvements in processing speed and memory usage. This allows MinkUNet to scale effectively with large datasets and dense point clouds. The network architecture consists of an encoder-decoder structure, where the encoder extracts multi-scale features through a series of sparse convolutional layers, and the decoder reconstructs the segmented output using transposed sparse convolutions.

In our approach, we adopt the MinkUNet as the backbone due to its proven effectiveness in handling large-scale 3D point clouds. Different from the original MinkUNet configuration, we used a larger backbone, dubbed **MinkUNet-101**, to learn more informative features from the complex point clouds in the Waymo Open dataset [20]. The hierarchical feature extraction capability of MinkUNet allows it to capture fine-grained details and global context, which are crucial for accurate semantic segmentation. For additional details, kindly refer to Choy *et al*. [6].

2.2. MixSeg3D

To further enhance the performance of MinkUNet, we introduce MixSeg3D, a sophisticated combination of the MinkUNet backbone with advanced 3D data mixing strategies: LaserMix [15] and PolarMix [25]. These techniques are designed to augment the training data by blending LiDAR point clouds in innovative ways, thereby improving the model's generalization and robustness. Fig. 1 provides illustrative examples of conducting LaserMix and PolarMix on LiDAR point clouds, where the two colors represent points from two randomly sampled LiDAR scans.

LaserMix is a data augmentation technique that blends LiDAR point clouds along the ego-scene's inclination direction. By partitioning the point clouds and combining them at various angles, LaserMix generates diverse training samples that simulate different driving conditions and perspectives. This method helps the model learn to recognize objects and semantic classes from various inclinations, enhancing its ability to generalize to unseen data. We set the probability of conducting the LaserMix operation during training as p_1.

PolarMix complements LaserMix by blending LiDAR point clouds along the azimuth direction. This technique rotates the point clouds around the vertical axis, creating augmented samples that represent different orientations. PolarMix ensures that the model can handle variations in object orientation and spatial distribution, which are common in real-world driving scenarios. We set the probability of conducting the PolarMix operation during training as p_2.

Test Time Augmentation (TTA) is a technique used during

Table 1. The 3D semantic segmentation results from different participants in the 2024 Waymo Open Dataset Challenge.

Method	mIoU	Access Date
PTv3-EX	72.76	24/05/2024, 12 : 08 : 05
MixSeg3D	69.83	24/05/2024, 10 : 48 : 19
RangeFormer++	69.06	18/05/2024, 22 : 53 : 51
PointTransformer-LR	68.92	24/05/2024, 09 : 50 : 26
vFusedSeg3D	68.79	24/05/2024, 03 : 07 : 11
vSeg3D	68.06	15/05/2024, 22 : 19 : 17
SPVCNN++	68.05	10/05/2024, 03 : 40 : 55

model inference to enhance segmentation accuracy. This approach involves applying various transformations to the input data at test time and aggregating the predictions from multiple augmented versions of the data. In this competition, we adopted rotations, scaling, flipping, and shifting.

3. Experiments

3.1. Experimental Settings

Datasets. We follow the 2024 Waymo Open Dataset Challenge rules in preparing the training, validation, and test data. Specifically, the model was trained on the official *train* split of Waymo Open [20] and tested on the associated *val* and *test* splits. The number of samples from each of the three splits is ∼24,000, ∼7,000, and ∼3,000, respectively.

Implementation Details. Our model is implemented based on MMDetection3D [7]. Our model was trained with an effective batch size of 32 and a learning rate of 0.008 across 4 NVIDIA A100 GPUs. We adopted the AdamW optimizer and OneCycle learning rate scheduler for model optimization [18]. To enhance training data diversity, we used random rotations, scaling, flipping, and shifting, before the LaserMix and PolarMix operations. After that, we set the probabilities of conducting LaserMix and PolarMix, *i.e.*, p_1 and p_2, as 0.8 and 1.0, respectively. After training, we set a combo of 8 times of TTA during the evaluation. The model was trained for 80 epochs on the official training set and then tested on the corresponding validation and test sets.

Evaluation Protocol. Following the official evaluation protocol of the 2024 Waymo Open Dataset Challenge, we report the Intersection-over-Union (IoU) scores for each class and the mean IoU (mIoU) scores across all semantic classes.

3.2. Experimental Results

Tab. 1 presents the mIoU scores of different teams in this year's challenge. Our proposed MixSeg3D achieved a competitive mIoU score of 69.83%, securing second place among the participants. This performance is notable considering the diversity and complexity of the Waymo Open dataset, indicating the robustness and effectiveness of our approach.

Furthermore, Tab. 2 details the class-wise IoU scores for MixSeg3D. The results demonstrate that MixSeg3D performs particularly well in several key classes such as Car (95.6%), Pedestrian (91.5%), and Building (97.0%). These high IoU scores in critical classes underscore the model's ability to accurately segment important objects in the driving environment, which is essential for driving perception.

Despite the strong overall performance, there are areas where MixSeg3D shows room for improvement. For instance, the IoU scores for classes such as Motorcyclist (8.4%) and Traffic Light (32.2%) indicate challenges in segmenting smaller or less frequent objects. These observations highlight potential directions for future enhancements, such as incorporating more targeted data augmentation techniques or refining the model architecture to better capture these classes.

To understand the efficacy of each of the backbones and data augmentation techniques applied, we conduct an ablation study and show the results in Tab. 3. We first study the model capacity. We found that larger backbone networks (*e.g.*, MinkUNet-50 and MinkUNet-101, which contain 31.9M and 70.8M parameters, respectively) tend to yield higher segmentation accuracy, as the larger sets of parameters often learn more meaningful representations. We then study the efficacy of LaserMix, PolarMix, and TTA. We found the best possible configuration by combining all of them together and setting proper hyperparameters.

It is worth mentioning that the LaserMix and PolarMix operations are very efficient in practice. During the model training, these two techniques manipulate the LiDAR point clouds "on-the-fly" at the point level, which can be regarded as basic tensor operations, such as adding, removing, replacing, *etc*. However, the time consumption of applying TTA is notably longer than plain evaluation. This is because the TTA operation requires multiple times of model inference, and, to a certain extent, is not very practical in terms of actual deployment.

Overall, the experimental results validate the effectiveness of MixSeg3D. The integration of MinkUNet with advanced data augmentation techniques, LaserMix and PolarMix, has proven to be a successful strategy for enhancing 3D semantic segmentation performance.

4. Conclusion

We presented MixSeg3D, a novel approach for 3D semantic segmentation that integrates the MinkUNet backbone with advanced data augmentation techniques, LaserMix and PolarMix. Our method significantly improves 3D segmentation accuracy and robustness in diverse driving environments, achieving a competitive mIoU score of 69.83%, securing second place. These results validate our approach and demonstrate its potential for advancing 3D semantic segmentation in autonomous driving.

Table 2. The class-wise 3D semantic segmentation results of MixSeg3D from the 2024 Waymo Open Dataset Challenge leaderboard.

Car	Truck	Bus	Other Vehicle	Motorcyclist	Bicyclist	Pedestrian	Sign	Traffic Light	Pole	Construction Cone	Bicycle	Motorcycle	Building	Vegetation	Tree Trunk	Curb	Road	Lane Marker	Other Ground	Walkable	Sidewalk
95.6	70.7	73.5	29.6	8.4	88.5	91.5	73.1	32.2	80.2	60.0	70.6	80.3	97.0	86.8	73.2	75.1	92.6	48.5	51.5	70.6	86.7

Table 3. Ablation study on different training and evaluation configurations in MixSeg3D. Results reported are from the official validation set of the 2024 Waymo Open Dataset Challenge.

Backbone	**Aug**	**TTA**	**mIoU**
MinkUNet-18	None	None	68.02
MinkUNet-34	None	None	70.18
MinkUNet-50	None	None	70.34
MinkUNet-101	None	None	70.98
MinkUNet-101	LaserMix	None	71.37
MinkUNet-101	PolarMix	None	71.31
MinkUNet-101	LaserMix + PolarMix	None	72.06
MinkUNet-101	LaserMix + PolarMix	3 times	72.41
MinkUNet-101	LaserMix + PolarMix	6 times	72.67
MinkUNet-101	LaserMix + PolarMix	8 times	74.03
MinkUNet-101	LaserMix + PolarMix	10 times	73.67

References

[1] Angelika Ando, Spyros Gidaris, Andrei Bursuc, Gilles Puy, Alexandre Boulch, and Renaud Marlet. Rangevit: Towards vision transformers for 3d semantic segmentation in autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5240–5250, 2023.

[2] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Juergen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *IEEE/CVF International Conference on Computer Vision*, pages 9297–9307, 2019.

[3] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Jürgen Gall, and Cyrill Stachniss. Towards 3d lidar-based semantic scene understanding of 3d point cloud sequences: The semantickitti dataset. *International Journal of Robotics Research*, 40:959–96, 2021.

[4] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020.

[5] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7030, 2023.

[6] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019.

[7] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d, 2020.

[8] Bertrand Douillard, James Underwood, Noah Kuntz, Vsevolod Vlaskine, Alastair Quadros, Peter Morton, and Alon Frenkel. On the segmentation of 3d lidar point clouds. In *IEEE International Conference on Robotics and Automation*, pages 2798–2805, 2011.

[9] Whye Kit Fong, Rohit Mohan, Juana Valeria Hurtado, Lubing Zhou, Holger Caesar, Oscar Beijbom, and Abhinav Valada. Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking. *IEEE Robotics and Automation Letters*, 7(2):3795–3802, 2022.

[10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012.

[11] Martin Hahner, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Fog simulation on real lidar point clouds for 3d object detection in adverse weather. In *IEEE/CVF International Conference on Computer Vision*, pages 15283–15292, 2021.

[12] Qingyong Hu, Bo Yang, Sheikh Khalid, Wen Xiao, Niki Trigoni, and Andrew Markham. Towards semantic segmentation of urban-scale 3d point clouds: A dataset, benchmarks and challenges. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4977–4987, 2021.

[13] Lingdong Kong, Youquan Liu, Runnan Chen, Yuexin Ma, Xinge Zhu, Yikang Li, Yuenan Hou, Yu Qiao, and Ziwei Liu. Rethinking range view representation for lidar segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 228–240, 2023.

[14] Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robo3d: Towards robust and reliable 3d perception against corruptions. In *IEEE/CVF International Conference on Computer Vision*, pages 19994–20006, 2023.

[15] Lingdong Kong, Jiawei Ren, Liang Pan, and Ziwei Liu. Lasermix for semi-supervised lidar semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21705–21715, 2023.

[16] Youquan Liu, Runnan Chen, Xin Li, Lingdong Kong, Yuchen Yang, Zhaoyang Xia, Yeqi Bai, Xinge Zhu, Yuexin Ma, Yikang Li, Yu Qiao, and Yuenan Hou. Uniseg: A unified multi-modal lidar segmentation network and the openpcseg codebase. In *IEEE/CVF International Conference on Computer Vision*, pages 21662–21673, 2023.

[17] Youquan Liu, Lingdong Kong, Jun Cen, Runnan Chen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Segment any point cloud sequences by distilling vision foundation models. In *Advances in Neural Information Processing Systems*, 2023.

[18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[19] Andres Milioto, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. Rangenet++: Fast and accurate lidar semantic segmentation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4213–4220, 2019.

[20] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020.

[21] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching efficient 3d architectures with sparse point-voxel convolution. In *European Conference on Computer Vision*, pages 685–702, 2020.

[22] Haotian Tang, Zhijian Liu, Xiuyu Li, Yujun Lin, and Song Han. Torchsparse: Efficient point cloud inference engine. *Proceedings of Machine Learning and Systems*, 4:302–315, 2022.

[23] Haotian Tang, Shang Yang, Zhijian Liu, Ke Hong, Zhongming Yu, Xiuyu Li, Guohao Dai, Yu Wang, and Song Han. Torchsparse++: Efficient point cloud engine. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 202–209, 2023.

[24] Yuan Wang, Tianyue Shi, Peng Yun, Lei Tai, and Ming Liu. Pointseg: Real-time semantic segmentation based on 3d lidar point cloud. *arXiv preprint arXiv:1807.06288*, 2018.

[25] Aoran Xiao, Jiaxing Huang, Dayan Guan, Kaiwen Cui, Shijian Lu, and Ling Shao. Polarmix: A general data augmentation technique for lidar point clouds. In *Advances in Neural Information Processing Systems*, pages 11035–11048, 2022.

[26] Jianyun Xu, Ruixiang Zhang, Jian Dou, Yushi Zhu, Jie Sun, and Shiliang Pu. Rpvnet: A deep and efficient range-point-voxel fusion network for lidar point cloud segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 16024–16033, 2021.

[27] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.