2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

MOSFormer: A Transformer-based Multi-Modal Fusion Network for Moving Object Segmentation

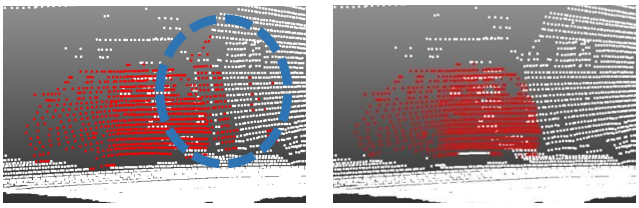
Zike Cheng[1] Hengwang Zhao[1] Qiyuan Shen[1] Weihao Yan[1] Chunxiang Wang[1] Ming Yang[1,*]

***Abstract*— 3D moving object segmentation (MOS) is vital for autonomous systems, providing essential information for downstream tasks like mapping and localization. However, current MOS methods face challenges due to the limitation of existing datasets, which are sparse in moving objects and limited in scene diversity. Meanwhile, the prevalent methods are projection-based, struggling with the challenge of blurred boundaries. To tackle the dataset issue, we introduce a nuScenes-based MOS dataset, which provides richer scenes and more dynamic instances. To alleviate the boundary blurring issue and further improve accuracy and generalizability, we propose a dual-branch multimodal fusion MOS network, MOSFormer. The Transformer structure is incorporated to extract spatio-temporal information better, while image semantic information is utilized to refine the boundaries of moving objects. Finally, experiments on two datasets show that our method achieves state-of-the-art performance, and a mapping experiment with our method confirms its effectiveness in downstream tasks such as mapping and localization.**

I. Introduction

Autonomous systems are affected by the presence of numerous moving objects in traffic scenarios, such as vehicles, pedestrians, and bicycles. These moving objects lead to ghosting phenomena during the mapping process and affect localization accuracy. To mitigate the adverse effects of moving objects, the moving object segmentation (MOS) task emphasizes segmenting points in motion, which utilizes point cloud sequences to infer the dynamic attributes of objects. Real-time and reliable MOS results provide important information for downstream tasks, such as path planning [1], mapping [2], and localization [3], which are crucial for autonomous systems.

Current works in 3D MOS field mainly fall into two categories. One is geometry-based methods [2], [3], which utilize the relationship between point clouds and maps to divide them into static and dynamic points. However, these methods rely on the maintenance of maps, and segmentation errors can occur if moving objects are not correctly filtered out from the maps. In recent years, learning-based methods such as LMNet [4], 4DMOS [5], RVMOS [6], and MotionSeg3D [7] have been proposed. Most of these methods project point clouds into range images, learning motion information from multiple frames and utilizing networks to infer MOS results. Due to the limited resolution of range images, multiple points can be projected onto a single pixel, leading to a many-to-one problem that causes boundary blurring when projecting the image back into 3D space. Fig. 1 illustrates an example of a blurred boundary in the prediction result. Although methods like MotionSeg3D [7] attempt to use a two-stage network architecture to improve the back-projection process, this issue has not been adequately addressed due to the lack of texture information in point clouds.



(a) Boundary Blurring (b) Ground Truth

Fig. 1. An example of boundary blurring. The left shows the prediction from MotionSeg3D, and the right shows the ground truth. Boundary blurring can lead to predicted boundaries expanding outward or contracting inward.

Additionally, most current learning-based MOS works are trained on the SemanticKITTI MOS [4] and its supplementary dataset, KITTI-Road MOS [7]. Both are annotated based on the original sequences in KITTI. Due to the limitations of the original data, these datasets suffer from restricted scenes and sparse moving objects. This imbalance between dynamic and static samples compromises the ability of datasets to evaluate the performance of MOS methods.

To address the boundary blurring issue, we design a Semantic-guided Label Voting (SLV) module in back-projection process. Compensating for the sparsity and lack of texture information in LiDAR, we utilize dense image semantic information to generate more refined and reliable segmentation boundaries. To enhance the capability of capturing motion information across frames, we also introduce the 3D Multi-head Self-Attention (MSA) module in our Transformer-based network. Moreover, regarding the limitations of current MOS datasets, we propose nuScenes for MOS dataset, offering richer scenes and more diverse moving objects than existing MOS datasets.

In summary, the contributions of this paper are as follows:

- We introduce MOSFormer, a dual-branch multi-modal fusion MOS network that addresses boundary blurring with the 3D MSA module for temporal information extraction and the SLV module for boundary refinement.
- To improve the capability of dataset to evaluate MOS methods, we propose the nuScenes for MOS dataset, offering richer scenes, more dynamic instances, and

This work was supported by the National Natural Science Foundation of China (U22A20100/62373250/62173228).

[1]Department of Automation, Shanghai Jiao Tong University, Shanghai, 200240, China; also with Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, 200240, China.
*Corresponding author: Ming Yang (e-mail: mingyang@sjtu.edu.cn).

more comprehensive sensors. The dataset will be open-sourced as planned.
- Our method achieves state-of-the-art performance on two datasets and demonstrates its practical value through a qualitative mapping experiment.

II. Related Works

A. LiDAR Semantic Segmentation

LiDAR semantic segmentation methods can be categorized into three approaches based on the representation of point clouds: 1) *Point-based* methods [8], [9], inspired by the approach of extracting point-wise features in PointNet [8] and PointNet++ [9], processes points based on neighborhood relations. These methods usually require excessive computational resources. 2) *Voxel-based* methods [10], [11] divide space into voxels and use sparse 3D convolutions to extract voxel features, thus balancing accuracy and speed. 3) *View-based*, projects point clouds into range views [12], [13] or bird-eye views [14], which are also known as projection-based methods. Benefiting from advancements in 2D image domain [15], [16], view-based methods introduce Transformer structure and show significant potential.

3D semantic segmentation tasks and MOS tasks are closely related, both involving point-level segmentation of point clouds. However, MOS tasks place greater emphasis on the dynamic information from instances across consecutive frames, while semantic segmentation focuses exclusively on the semantic categories of objects within a single frame.

B. 3D Moving Object Segmentation

Initially, geometry-based methods [2], [3] were proposed, which segmented dynamic and static points through the consistency of points when matching LiDAR scans with offline maps. The M-detector [17] method later improved on this by comparing the occlusion relationships of points over time. However, both methods depend on offline maps and are prone to ghosting phenomena from moving objects.

In recent years, numerous learning-based methods [4]–[7], [18] have been proposed. LMNet [4] utilizes LiDAR range images and residual images to segment moving objects, while 4DMOS [5] employs sparse 4D convolutions to extract spatial and temporal features jointly. MotionSeg3D [7] has designed a dual-branch network to leverage spatial and temporal information better in sequential LiDAR scans. RVMOS [6] integrates the motion and semantic features of point clouds for improved segmentation of moving objects.

Notably, to balance network performance and efficiency, current methods typically use range view as an intermediate representation for feature extraction. When projecting point clouds into range images, limited by the resolution, multiple points often exist within a single pixel. This many-to-one problem leads to ambiguities when back-projecting the image to 3D space, which restricts accuracy improvements.

Existing methods mainly focus on LiDAR and underutilize low-cost cameras, which are common on autonomous vehicles. Research in 3D semantic segmentation [19], [20] has shown that images can provide dense semantic information, compensating the sparsity of point clouds and offering potential benefits for the 3D MOS field.

C. 3D MOS Datasets

Currently, open-source MOS datasets are limited, with two widely used datasets as follows:

SemanticKITTI MOS Dataset [4]: A large LiDAR-based MOS benchmark based on SemanticKITTI [21], which includes 11 training sequences and 11 testing sequences, totaling 43,552 annotated frames. The dataset suffers from an overly sparse presence of dynamic instances, with 25.77% of the frames containing more than 100 dynamic points.

KITTI-Road MOS Dataset [7]: An MOS dataset based on the KITTI road data, includes 12 sequences, 5,794 annotated frames, with over 60% containing more than 100 dynamic points. However, due to its limited annotated data, it generally serves as a supplementary dataset to the SemanticKITTI MOS Dataset.

Overall, current datasets have the following issues: 1) lack of scenario diversity, which is reflected in the limited number of data sequences; 2) imbalance between dynamic and static samples, with moving objects being scarce in the scenes.

III. NuScenes for MOS Dataset

To address the shortcomings of the current MOS datasets with limited scenes and sparse moving objects, we propose a new MOS dataset: nuScenes for MOS.

1) 3D Annotation of Moving Objects:

The nuScenes dataset [22] is a large-scale dataset for autonomous driving research, including high-resolution images, LiDAR data, GPS, and IMU data. The instance-level 3D object detection annotations allow for tracking the motion of each instance and creating a nuScenes-based MOS dataset.

Following SemanticKITTI MOS annotation rules [4], nuScenes for MOS includes point-level annotations, classifying points into static, moving, or ignore. Instances are tracked to determine velocity and due to the jitter in bounding box annotations, a speed threshold of 2 km/h is set to categorize them as dynamic or static. Point clouds are then annotated based on instance bounding boxes and categories. Instances that appear only once are categorized as ignore.

2) Comparison with Existing MOS Datasets:

Table I compares our proposed nuScenes for MOS with existing MOS datasets (SemanticKITTI MOS Dataset [4], KITTI-Road MOS Dataset [7], and a combination of both). The nuScenes for MOS has the following advantages:

- **More Diverse Scenes:** NuScenes for MOS contains 1000 data sequences, each with a duration of 20 seconds. Compared to the existing KITTI-based datasets, it offers more diverse scenes and environments, presenting a more challenging dataset for MOS tasks.
- **Richer Moving Instances:** NuScenes for MOS includes 25.4k moving instances within 40k annotated frames, while Combined KITTI dataset contains about 1k moving instances. Following the definition of dynamic frames in [7], the proportion of frames with more

TABLE I
MOS DATASETS COMPARISON

	SemanticKITTI MOS [4]	KITTI-Road MOS [7]	Combined KITTI MOS	nuScenes for MOS
Sensors	Lidar*1 Camera*2			**Lidar*1 Camera*6**
Scenes	22	12	34	**1000**
Annotation for training	23201	5794	28995	**40k**
Moving instances	660	484	1144	**25.4k**
Proportion of dynamic frames	25.77%	>60%	>30%	**58%**

than 100 dynamic points in nuScenes for MOS is significantly higher than in the Combined KITTI dataset, which helps alleviate the issue of sample imbalance.

- **More Comprehensive Sensors:** NuScenes data is collected using a platform with a top-view LiDAR and 6 surround-view cameras. Compared to the front-view images in KITTI-based datasets, nuScenes for MOS is better suited for multi-modal fusion.

IV. Method

A. MOSFormer

1) Overview:

We propose a dual-branch network for segmenting moving objects, utilizing LiDAR and image information to achieve accurate 3D segmentation. Fig. 2 shows the specific network structure. This network comprises three parts: the LiDAR branch, the Camera branch, and the back-projection module.

LiDAR Branch: The input point clouds are first projected into range images according to equations in [13], and residual images are obtained using the pose information of the vehicle (which can be acquired from the dataset or calculated directly from IMU data and wheel speed sensor data). The residual images and range images are then concatenated and fed into the Sequential Transformer Module (STM) to obtain MOS results in range view.

Camera Branch: To optimize the efficiency of the camera branch, the images are fed into the existing BiSeNetv2 [23] to obtain semantic segmentation results, which can balance speed and accuracy.

Back-projection Module: To address potential ambiguities when back-projecting the range image into 3D space, we designed a Semantic-guided Label Vote (SLV) operation. This operation integrates image-level semantics with MOS results in range view to assign dynamic or static labels to unknown points, ultimately yielding 3D MOS results.

2) Sequential Transformer Module:

Inspired by [24], we design the Sequential Transformer Module (STM) to process sequential frames of residual and range images. Similar to [13], we adopt a pyramid structure to extract information at different scales. This pyramid structure comprises four stages, corresponding to down-sampling ratios of 1, 2, 4, and 8. As illustrated in Fig. 3, the structure of Transformer Block at each stage comprises two components: a 3D Multi-head Self-Attention (MSA) module and a Feed-Forward Network (FFN) module.

3D Multi-head Self-Attention: The mechanism of 2D Multi-head Self-Attention has been proven effective for extracting image features [15]. We extend this mechanism to 3D, allowing it to process sequential input information. For the input tokens $H \times W \times T$, given a window size of $P \times P \times D$, the tokens are divided into $H/P \times W/P \times T/D$ 3D windows, within which multi-head self-attention is performed. This adaptation enables the model to capture spatial-temporal relationships within the data, enhancing its ability to understand and segment moving objects.

Feed-Forward Network: The FFN consists of two Multi-Layer Perceptrons (MLPs) layers, with a GELU activation layer in between:

$$FFN(I) = Linear(GELU(Linear(I))) \oplus I \tag{2}$$

where I represents the input to FFN, $\oplus$ denotes the residual connection and *Linear* denotes MLP layer.

3) Semantic-guided Label Vote:

In LiDAR branch and Camera branch, we obtain MOS results in range view and image semantic segmentation results, respectively. The original point cloud is projected onto the semantic images to assign semantic labels, creating a 3D semantic point cloud. Simultaneously, the MOS results in range view are back-projected into 3D space to obtain a coarse 3D MOS result. Due to the many-to-one problem of range view projection, the dynamic-static state of some points is uncertain and thus requires refinement. We concatenate the semantic labels of points with the MOS labels and input them into PointMLP [25] to obtain the final refined 3D MOS results. The structure is shown in Fig. 4.

V. Experiment

A. Experiment Setups:

1) Datasets:

We evaluate our proposed method on two MOS datasets and compare it with the current state-of-the-art (SOTA) methods. The details of the datasets are as follows:

KITTI MOS Dataset combines two existing KITTI-based MOS Dataset [4], [7]. We adopt the dataset division in [7], which includes 34 sequences, with 16 sequences (22,035 frames) for training, 7 sequences (6,960 frames) for validation, and 11 sequences (20,351 frames) for testing.

nuScenes for MOS Dataset is our proposed MOS dataset, containing 1,000 sequences (40,000 frames). We divide the dataset into three parts: sequences 000-599 for training, 600-699 for validation, and 700-999 for testing.

2) Implementation Details:

All experiments are conducted with NVIDIA RTX 3090. The input sequence comprises four frames, with a time interval of 0.2s in KITTI MOS Dataset and 0.5s in the nuScenes for MOS Dataset. The AdamW optimizer and a warmup

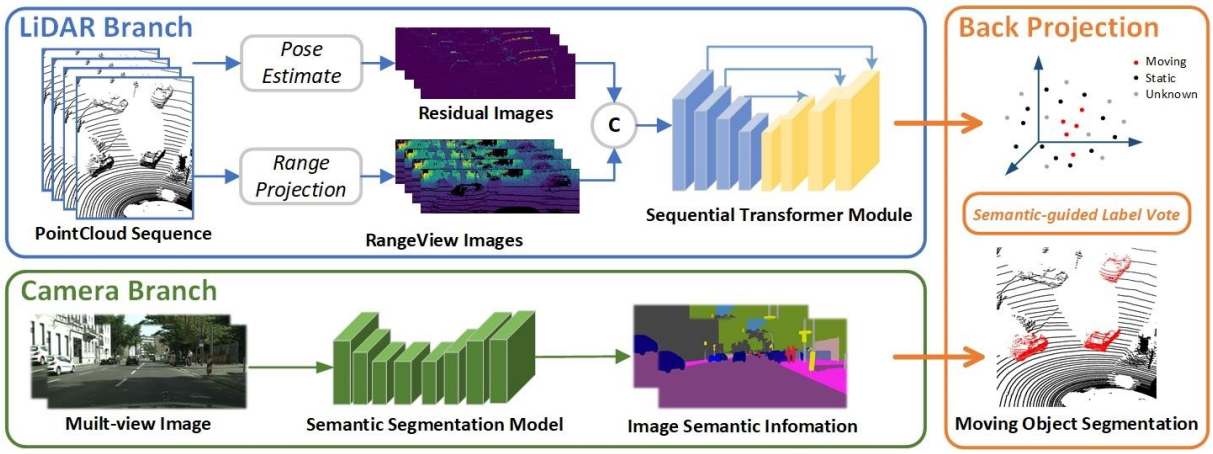


Fig. 2. A Network Overview of MOSFormer

learning rate policy are utilized. Consistent with [13], we employed cross-entropy dice loss, Lovasz-Softmax loss, and boundary loss [26] for supervision. And Intersection-over-Union (IoU) is used to evaluate MOS performance.

B. Experiment Results and Comparisons

We compare our method with SOTA approaches, and Table II shows the performance of different methods. In the experiments, our method is compared with several LiDAR-based moving object segmentation methods. Notably, all training and testing are conducted respectively on KITTI MOS and nuScenes for MOS.

LiMoSeg [27] and 4DMOS [5], which are voxel-based methods, achieve higher IoU on the KITTI MOS dataset but show a performance drop on the nuScenes for MOS dataset. This may be due to their poor adaptability to different scene environments, failing to handle the complex environments in the nuScenes dataset effectively.

LMNet [4], MotionSeg3D [7], RVMOS [6], and LiDAR-SGMOS [18] are range view-based methods that perform well on the KITTI MOS dataset but also experience a severe performance decrease on nuScenes for MOS dataset. RVMOS and LiDAR-SGMOS cannot be tested on nuScenes for MOS dataset due to the lack of open-source code.

In contrast, our proposed method shows a certain improvement on the KITTI MOS dataset and significantly outperforms on the nuScenes for MOS dataset. And more qualitative comparisons in Fig. 5 highlight our method's superior accuracy and stability, especially in handling boundary blurring issues. In qualitative and quantitative experiments, our method achieves robust and accurate prediction results in complex scenarios with refined boundaries.

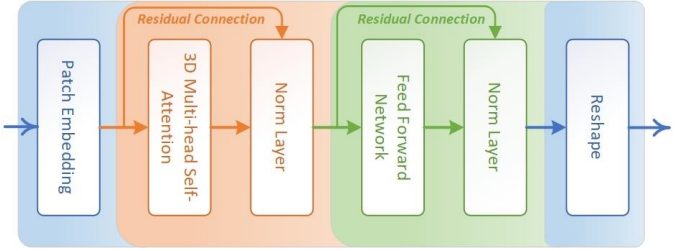


Fig. 3. Architecture of basic Transformer Block in Sequential Transformer Module. 3D Multi-head Self-Attention is introduced to fully extract spatio-temporal information.

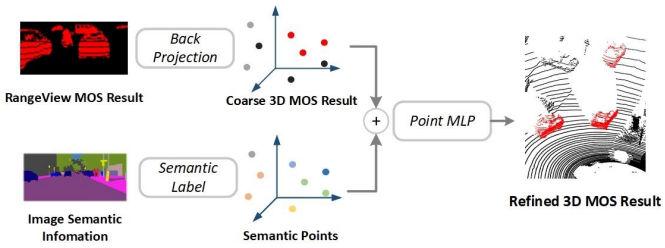


Fig. 4. Architecture of Semantic-guided Label Vote Module. PointMLP is used to fuse semantic points and coarse 3D MOS results to obtain refined MOS results.

C. Ablation Study

We conduct ablation studies on our contributions to verify their impact on accuracy improvement. All ablation experiments are based on the nuScenes for MOS dataset to better showcase the advantages of our method. Table III shows the results of the ablation experiment.

Transformer Structure: Compared to traditional CNN network structures, the Transformer structure has the advantage of global attention. Works in the 2D semantic segmentation field [15], [16] also prove the significant value of the Transformer structure in improving accuracy. Our test between using a Transformer structure and a CNN structure (SalsaNext [12]) shows that even a standard Transformer structure could enhance network performance and generalizability (with 5.4% improvement in accuracy).

3D MSA: In our method, 3D MSA replaces the ordinary global self-attention mechanism [15], resulting in a 6.6% increase in accuracy. The results demonstrate that 3D MSA is more effective in extracting temporal information from input sequences and helping the network capture the motion information of objects.

TABLE II
MOS PERFORMANCE COMPARED WITH THE STATE-OF-THE-ART METHODS

Methods	Rep.	IoU[1]	IoU[2]	Latency
LiMoSeg	Voxel	52.6	41.7	**8ms**
4DMOS		72.9	63.6	110ms
LMNet	Range	64.3	53.3	35ms
MotionSeg3D		71.4	62.0	132ms
RVMOS		71.2	-	-
LiDAR-SGMOS		69.0	-	-
Ours(L)[3]	Range	73.2	67.8	83 ms
Ours(L+C)[3]		**73.9**	**68.7**	97 ms

[1] is tested on KITTI MOS val dataset
[2] is tested on nuScenes for MOS test dataset
[3] **L** indicates LiDAR input, and **C** indicates image input

SLV: Our method introduces a camera branch and the SLV module to address boundary blurring issues during the back-projection process. Utilizing semantic information, the network can correct misclassified points. The ablation study shows a modest accuracy improvement (0.9%) with SLV, but this is less significant than expected. Considering the lack of ground truth for image semantic segmentation in nuScenes, we attribute this to the use of a pre-trained semantic segmentation model that was not fine-tuned on nuScenes. This leads to ambiguities in image semantic segmentation that could confuse the judgment of the network, preventing the SLV module from achieving its expected effect. Additionally, our fusion module employs a late fusion approach that only partially integrates the information from both modalities, which will be the focus of our future research.

D. Efficiency

The efficiency of our method is analyzed on the NVIDIA 3090 platform, with results shown in Table I. Despite using a more complex Transformer structure, our method still satisfies the real-time requirement of 10Hz, aligning with the LiDAR sampling frequency. Considering that we only utilize the most basic Transformer structure, there is still room for improvement in network efficiency, which will be the direction of our future work.

E. Mapping with Proposed MOS

The MOS task was initially proposed to filter out moving objects in scenes to eliminate the interference of moving objects on mapping and localization tasks [2], [3]. Therefore, to demonstrate the effectiveness of MOSFormer, we conduct qualitative experiments on its impact on mapping results.

We collect data in a campus environment with complex moving objects using a data collection vehicle equipped with a Hesai 40-Beam LiDAR, four fisheye cameras, and GPS/IMU sensors. Moving objects in the scene include vehicles, bicycles, and pedestrians. Fig. 6 shows the qualitative results of the mapping experiments. The upper image displays the mapping result using LIO-SAM [28] directly. As demonstrated in the mapping results within the red box, severe ghosting phenomena caused by moving objects significantly reduce map quality. The lower image shows the result after filtering out dynamic points with our method, where ghosting is eliminated and map quality is significantly improved. The experiment indicates that our proposed MOS method can significantly improve the quality of point cloud maps, providing assistance for downstream tasks.

TABLE III
ABLATION STUDY EVALUATED ON NUSCENES FOR MOS

CNN	Transformer	+3D MSA	+SLV	IoU (%)
✓				55.8
✓			✓	56.0
	✓			61.2
	✓		✓	62.5
	✓	✓		67.8
	✓	✓	✓	**68.7**

VI. Conclusion

In this paper, we attempt to address the challenges the MOS task faces. To enhance the evaluation capability of the dataset for MOS methods, we introduce the nuScenes for MOS dataset with diverse scenes and moving objects. We propose MOSFormer, a new MOS network that mitigates boundary blurring in projection-based methods through a dual-branch structure and Semantic-guided Label Voting. Utilizing the 3D MSA module, our network achieves state-of-the-art results on both KITTI MOS and nuScenes for MOS datasets. Additionally, a qualitative mapping experiment validates the value of our method for downstream tasks. Future efforts will focus on improving real-time performance and enhancing multi-modal fusion for practical application.

References

[1] X. Shi and X. Li, "Speed planning for an autonomous vehicle with conflict moving objects," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, 2019, pp. 413–418.

[2] F. Pomerleau, P. Krüsi, F. Colas, P. Furgale, and R. Siegwart, "Long-term 3d map maintenance in dynamic environments," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 3712–3719.

[3] T. Fan, B. Shen, H. Chen, W. Zhang, and J. Pan, "Dynamicfilter: an online dynamic objects removal framework for highly dynamic environments," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 7988–7994.

[4] X. Chen, S. Li, B. Mersch, L. Wiesmann, J. Gall, J. Behley, and C. Stachniss, "Moving object segmentation in 3d lidar data: A learning-based approach exploiting sequential data," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 6529–6536, 2021.

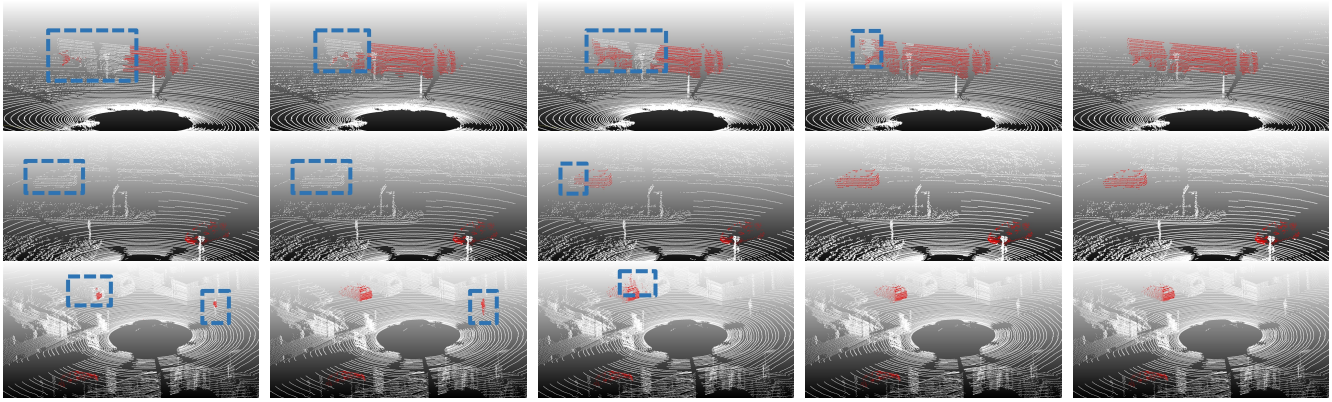
[5] B. Mersch, X. Chen, I. Vizzo, L. Nunes, J. Behley, and C. Stachniss, "Receding moving object segmentation in 3d lidar data using sparse 4d convolutions," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7503–7510, 2022.

[6] J. Kim, J. Woo, and S. Im, "Rvmos: Range-view moving object segmentation leveraged by semantic and motion features," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 8044–8051, 2022.

[7] J. Sun, Y. Dai, X. Zhang, J. Xu, R. Ai, W. Gu, and X. Chen, "Efficient spatial-temporal information fusion for lidar-based 3d moving object segmentation," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 11 456–11 463.

[8] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 77–85.

[9] C. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Neural Information Processing Systems*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:1745976



(a) LMNet (b) 4DMOS (c) MotionSeg3D (d) Ours (e) Ground Truth

Fig. 5. Qualitative experimental results of the SOTA methods. Red points represent dynamic points, white points represent static points, and blue rectangles highlight incorrect prediction results.

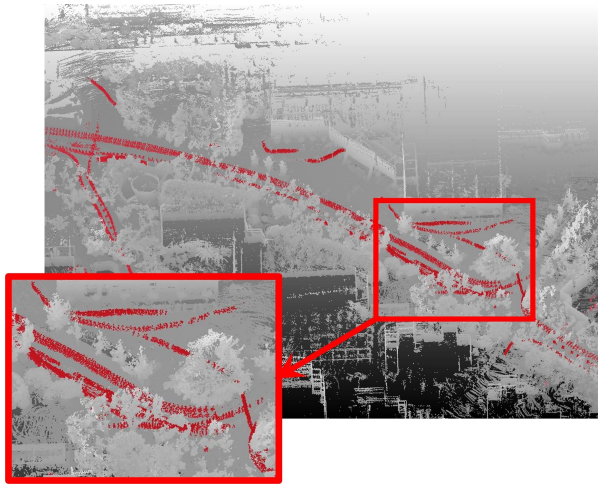


Fig. 6. Improvement in mapping with our method. With moving objects in red points filtered out, the ghosting phenomenon is greatly alleviated .

[10] X. Zhu, H. Zhou, T. Wang, F. Hong, W. Li, Y. Ma, H. Li, R. Yang, and D. Lin, "Cylindrical and asymmetrical 3d convolution networks for lidar-based perception," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6807–6822, 2022.

[11] X. Lai, Y. Chen, F. Lu, J. Liu, and J. Jia, "Spherical transformer for lidar-based 3d recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 17 545–17 555.

[12] E. E. Aksoy, S. Baci, and S. Cavdar, "Salsanet: Fast road and vehicle segmentation in lidar point clouds for autonomous driving," in *2020 IEEE Intelligent Vehicles Symposium (IV)*, 2020, pp. 926–932.

[13] L. Kong, Y. Liu, R. Chen, Y. Ma, X. Zhu, Y. Li, Y. Hou, Y. Qiao, and Z. Liu, "Rethinking range view representation for lidar segmentation," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 228–240.

[14] Y. Zhang, Z. Zhou, P. David, X. Yue, Z. Xi, B. Gong, and H. Foroosh, "Polarnet: An improved grid representation for online lidar point clouds semantic segmentation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9598–9607.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ArXiv*, vol. abs/2010.11929, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:225039882

[16] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9992–10 002.

[17] H. Wu, Y. Li, W. Xu, F. Kong, and F. Zhang, "Moving event detection from lidar point streams," *Nature Communications*, vol. 15, no. 1, p. 345, 2024.

[18] S. Gu, S. Yao, J. Yang, C. Xu, and H. Kong, "Lidar-sgmos: Semantics-guided moving object segmentation with 3d lidar," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 70–75.

[19] X. Yan, J. Gao, C. Zheng, C. Zheng, R. Zhang, S. Cui, and Z. Li, "2dpass: 2d priors assisted semantic segmentation on lidar point clouds," in *European Conference on Computer Vision*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:250425746

[20] Y.-C. Liu, R. Chen, X. Li, L. Kong, Y. Yang, Z. Xia, Y. Bai, X. Zhu, Y. Ma, Y. Li, Y. Qiao, and Y. Hou, "Uniseg: A unified multi-modal lidar segmentation network and the openpcseg codebase," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21 605–21 616, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:261696615

[21] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, J. Gall, and C. Stachniss, "Towards 3D LiDAR-based semantic scene understanding of 3D point cloud sequences: The SemanticKITTI Dataset," *The International Journal on Robotics Research*, vol. 40, no. 8-9, pp. 959–967, 2021.

[22] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 618–11 628.

[23] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation," *International Journal of Computer Vision*, vol. 129, pp. 3051 – 3068, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:214802217

[24] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 3192–3201.

[25] X. Ma, C. Qin, H. You, H. Ran, and Y. Fu, "Rethinking network design and local geometry in point cloud: A simple residual mlp framework," *arXiv preprint arXiv:2202.07123*, 2022.

[26] R. Razani, R. Cheng, E. Taghavi, and L. Bingbing, "Lite-hdseg: Lidar semantic segmentation using lite harmonic dense convolutions," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 9550–9556.

[27] S. Mohapatra, M. Hodaei, S. K. Yogamani, S. Milz, P. Mäder, H. Gotzig, M. Simon, and H. Rashed, "Limoseg: Real-time bird's eye view based lidar motion segmentation," *ArXiv*, vol. abs/2111.04875, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:243860931

[28] T. Shan, B. Englot, D. Meyers, W. Wang, C. Ratti, and D. Rus, "Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 5135–5142.