Future Does Matter: Boosting 3D Object Detection with Temporal Motion Estimation in Point Cloud Sequences

Rui Yu[1], Runkai Zhao[2], Cong Nie[3], Heng Wang[2], Huaichen Yan[1], Meng Wang[1]

BACKGROUND

3D LiDAR object detector plays an important role in autonomous driving, it identifies object information within a 3D road scene represented. Although discrete LiDAR points reflect accurate spatial positioning of surrounding driving scenes, they are insufficient to fully describe traffic objects due to data sparsity, particularly at far distances. Moreover, the LiDAR sensor captures partial view information of a scene from a single-frame perspective, leading to incomplete information collection of the visible objects. These inherent limitations of LiDAR result in inconsistent point distribution for the same object across a driving sequence. Hence, a dynamic object may be represented with varying densities of point clouds in different frames, which intro duces ambiguity in accurately determining the true shape for a 3D detector.

MOTIVATION

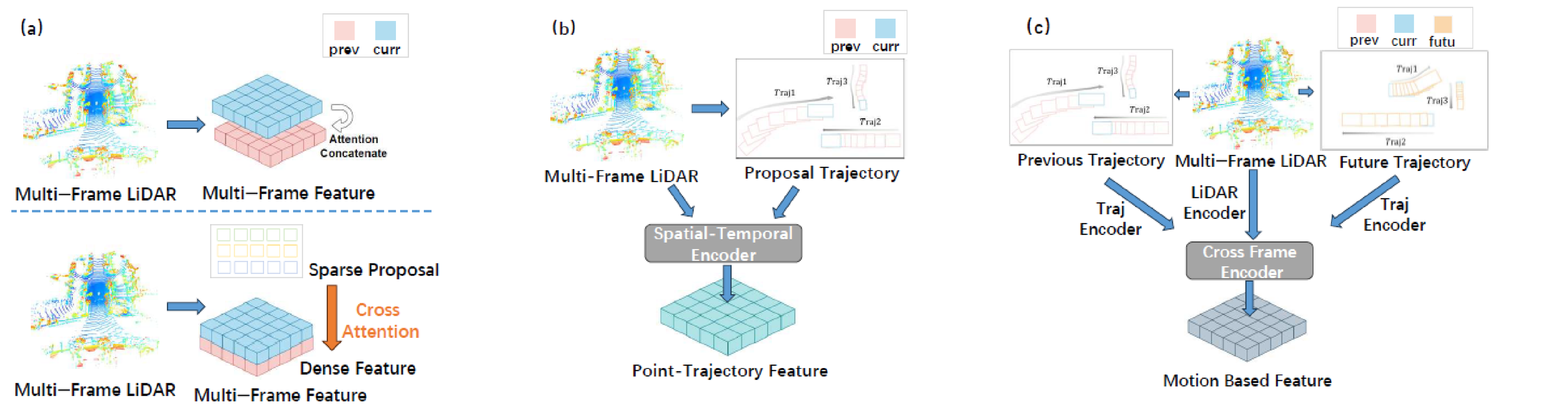


Figure 1: Different from the global bird's eye view (BEV) Neighbor Feature Fusion Method (a) and Trajectory-based Method (b), our proposed *LiSTM* count for the role of the future states

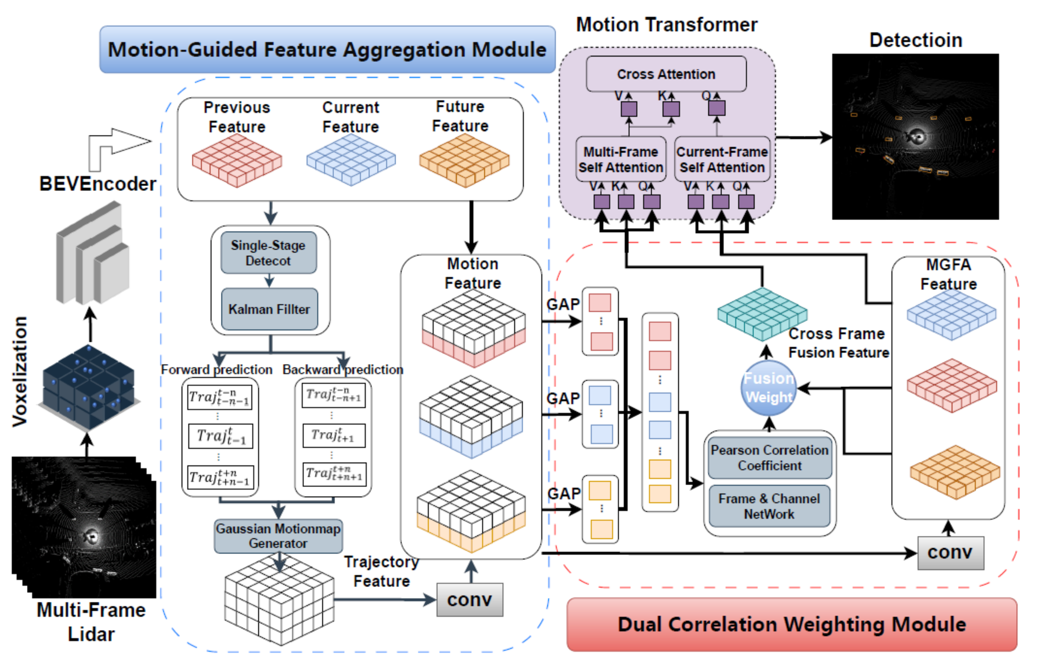


Figure 2: Overview of proposed framework LiSTM

- The first module employs a single-stage detector combined with tracking prediction to produce trajectories and then enhances the spatial representation with a Motion-Guided Feature Aggregation Module.
- The second module is used for cross-frame feature extraction by the proposed Dual Correlation Weighting Module and Motion Transformer.

Motion-Guided Feature Aggregation

- Motion-Guided Feature Aggregation (MGFA) is proposed to utilize the object trajectory from previous and future motion states to model spatial-temporal correlations into gaussian heatmap over a driving sequence. This motion-based heatmap then guides the temporal feature fusion, enriching the proposed object features.

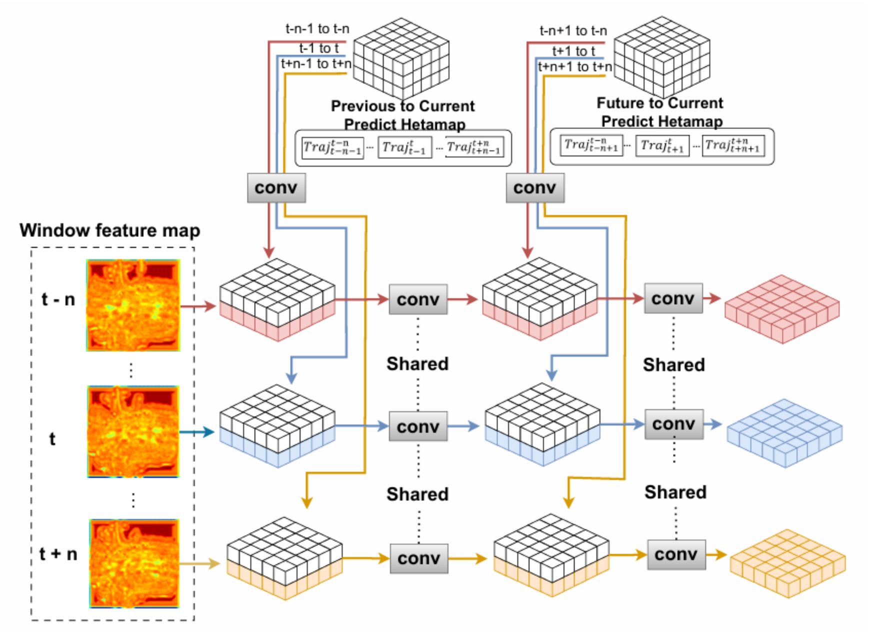


Figure 3: Motion Guided Feature Aggregation.

Dual Correlation Weighting Module

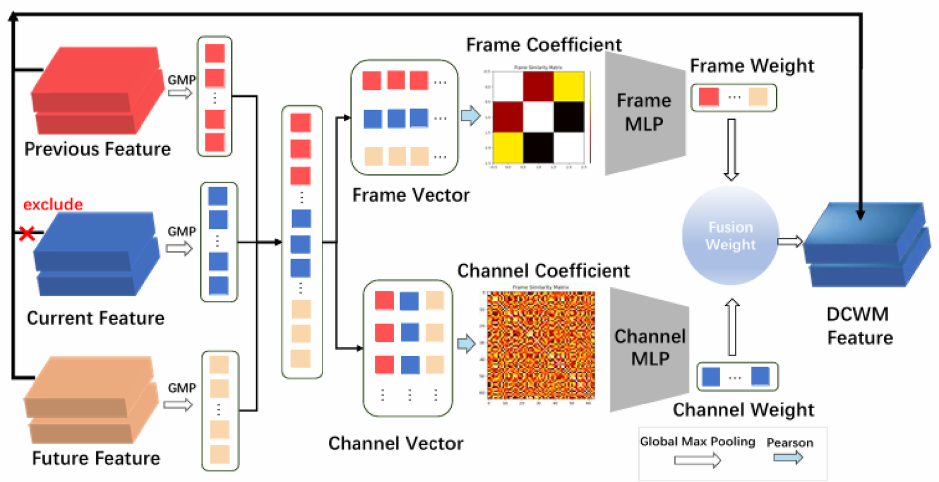


Figure 4: Dual Correlation Weighting Module.

- This motion-based heatmap then guides the temporal feature fusion, enriching the proposed object features. Moreover, we design a Dual Correlation Weighting Module (DCWM) that effectively facilitates the interaction be tween past and prospective frames through scene- and channel-wise feature abstraction.
- In the end, a cascade cross-attention-based decoder is employed to refine the 3D prediction.

RESULTS

Qualitative Comparisons

- LiSTM achieves an impressive improvement of over 8% compared to single-stage models like CenterPoint [1], while also outperforming two-stage models such as PVRCNN[2]
- When com pared to two-stage models MPPNet [3] and MSF [4], LiSTM demonstrates clear advance ments in vehicle and cyclist detection which is attributed to motion-based feature integration.
- On the nuScenes dataset, LiSTM outperforms the benchmarks, improving NDS and mAP by 2-3% compared to CenterPoint [1]. Meanwhile, LiSTM boost in ATE and ASE

Model	Frames	Vehicle (AP/APH)↑ L1	Vehicle (AP/APH)↑ L2	Pedestrian (AP/APH)↑ L1	Pedestrian (AP/APH)↑ L2	Cyclist (AP/APH)↑ L1	Cyclist (AP/APH)↑ L2
PointPillar [11]	1	66.94 / 66.36	58.96 / 58.43	63.35 / 45.22	55.21 / 39.32	55.06 / 52.55	52.97 / 50.55
VoxelNet [37]	1	68.73 / 67.31	60.11 / 59.97	69.65 / 57.38	60.19 / 53.67	62.31 / 59.85	60.34 / 55.89
PillarNet [19]	1	66.29 / 65.63	59.03 / 58.43	70.35 / 64.24	64.24 / 55.75	65.43 / 63.93	63.53 / 62.08
Second [31]	1	68.95 / 68.33	61.81 / 61.24	65.59 / 54.80	57.85 / 48.16	61.14 / 59.50	56.84 / 55.26
CenterPoint [34]	1	67.87 / 67.27	59.96 / 59.43	69.31 / 62.55	61.17 / 55.06	64.28 / 63.05	61.86 / 60.68
PartA2 [21]	1	65.52 / 64.85	57.32 / 56.63	54.83 / 37.72	46.85 / 32.19	54.29 / 48.75	52.21 / 46.89
PVRCNN [22]	1	71.11 / 70.32	62.60 / 61.88	63.63 / 32.77	54.88 / 28.26	59.49 / 34.14	57.22 / 32.83
VoxelRCNN [6]	1	71.51 / 70.98	63.75 / 63.26	65.95 / 65.99	65.47 / 60.86	70.11 / 68.71	67.98 / 66.63
CenterPoint [34]	4	71.27 / 70.73	63.59 / 63.09	73.91 / 70.45	66.28 / 60.10	63.78 / 62.98	61.59 / 60.82
CenterPoint [34]	16	72.53 / 71.31	64.18 / 64.21	74.05 / 71.17	66.17 / 61.03	64.05/ 64.54	62.31 / 61.77
MPPNet [3]	4	74.24 / 73.55	66.29 / 65.38	76.94 / 72.29	68.63 / 66.16	67.34/ 66.67	65.12 / 64.48
MSF [8]	4	74.37 / 73.97	66.35 / 65.85	**78.16 / 74.91**	**70.27 / 67.21**	67.89/ 67.14	65.58 / 64.89
LiSTM	3	**74.83 / 74.32**	**66.85 / 66.17**	75.89 / 69.72	66.83 / 63.43	**70.84 / 69.75**	**68.23 / 69.12**

Table 1: Quantative comparisons on 20% Sequence Waymo validation set.

Model	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
PointPillar [11]	58.62	45.27	0.3353	0.259	03286	0.2784	0.2002
Second [31]	62.31	50.8	0.3140	0.2554	**0.2785**	0.2587	0.2019
CenterPoint [34]	66.29	58.77	0.2919	0.2566	0.3692	**0.2081**	**0.1837**
VoxelNext [4]	67.09	60.55	0.3023	0.2526	0.3701	0.2087	0.1851
LiSTM	**68.32**	**63.77**	**0.2895**	**0.2479**	0.3182	0.2472	0.1850

Table 2: Quantative comparisons on nuScenes validation set.

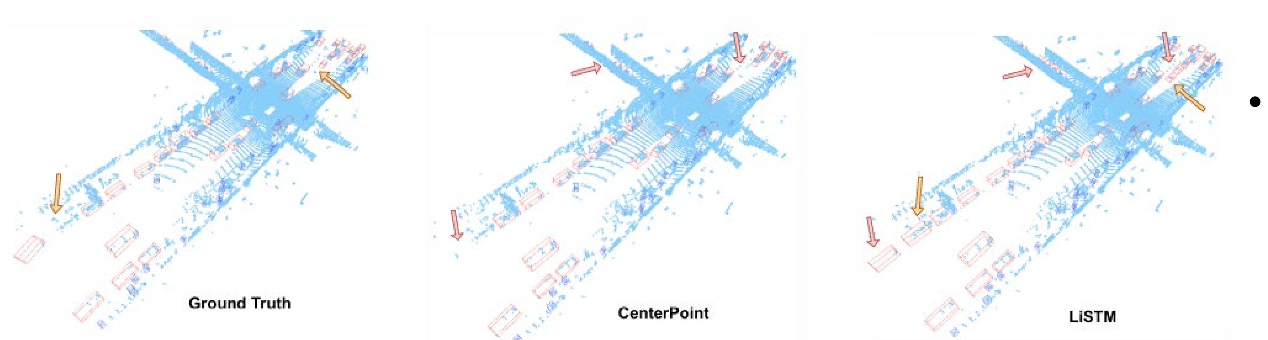
Computational Efficiency

Model	Model Parameter	Memory cost	FPS
CenterPoint [34]	7758811	2464 MiB	5.68 it/s
PV-RCNN++ [23]	13073505	3918 MiB	3.75 it/s
MSF [8]	15661651	6684 MiB	4.58 it/s
LiSTM	17592422	4400 MiB	5.26 it/s

Table 7: Computational efficiency

- Despite LiSTM having significantly larger model parameters, its actual FPS is comparable to that of CenterPoint [1].
- Moreover, LiSTM demonstrates a nearly 50% speed improvement over PV-RCNN[2] while consuming less memory and operating more efficiently than MSF [4].

Qualitative Visualization



- we compare the baseline with our module. LiSTM demonstrates superior capability, particularly highlighted by the pink arrows, in detecting cases that CenterPoint fails to identify due to distance and occlusion challenges.
- Additionally, LiSTM offers an increased number of positive samples with no annotations, as indicated by the yellow arrows.

Long Distance Perception

Model	25m away mAP↑ Vehicle	25m away mAP↑ Pedestrians	25m away mAP↑ Cyclist	50m-75m mAP↑ Vehicle	50m-75m mAP↑ Pedestrians	50m-75m mAP↑ Cyclist	75m away mAP↑ Vehicle	75m away mAP↑ Pedestrians	75m away mAP↑ Cyclist
CenterPoint [34]	58.80	63.12	61.37	41.82	54.00	50.50	11.46	16.30	14.82
LiSTM	**64.14**	**68.05**	**65.25**	**46.31**	**57.51**	**53.87**	**12.89**	**17.25**	**15.16**

Table 9: Long distance perception metric on the Waymo validation set.

- The LiSTM architecture leverages continuous frames and motion priors to enhance performance, particularly for long-range detection.
- In our evaluation with the Waymo dataset, we use three distance thresholds to metric.
- Results show that LiSTM outperforms the baseline by an average of 5 points in the 25m to 75m range.
- Even beyond this range, where the point cloud is mostly filtered out, LiSTM metrics remain somewhat elevated compared to the baseline.

Feature Fusion Strategies

Motion Feature	Cyl. L2 APH	Veh. L2 APH
pre2cur	66.51	65.72
fut2cur	66.47	65.78
cur2pre	66.13	65.7
cur2fut	66.21	65.72
cur2pre + cur2fut	66.57	65.83
pre2cur + fut2cur	**68.80**	**65.88**

CONCLUSION

Addressing the challenge of detecting sparse and occluded long-range LiDAR point clouds, we introduce LiSTM, a motion-based spatial-temporal fusion 3D point cloud detector. It leverages well-designed motion features and motion-guided feature fusion to enhance detection performance on Waymo and nuScenes datasets. In future work, we will focus on developing an end-to-end motion generator and exploring sparse feature representations.

REFERENCES

1 Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11784–11793, 2021.

2 Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni tion, pages 10529–10538, 2020.

3 Xuesong Chen, Shaoshuai Shi, Benjin Zhu, Ka Chun Cheung, Hang Xu, and Hong sheng Li. Mppnet: Multi-frame feature intertwining with proxy points for 3d tem poral object detection. In European Conference on Computer Vision, pages 680–697. Springer, 2022.

4 Chenhang He, Ruihuang Li, Yabin Zhang, Shuai Li, and Lei Zhang. Msf: Motion guided sequential fusion for efficient 3d object detection from point cloud sequences. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog nition, pages 5196–5205, 2023.