Beyond One Shot, Beyond One Perspective: Cross-View and Long-Horizon Distillation for Better LiDAR Representations

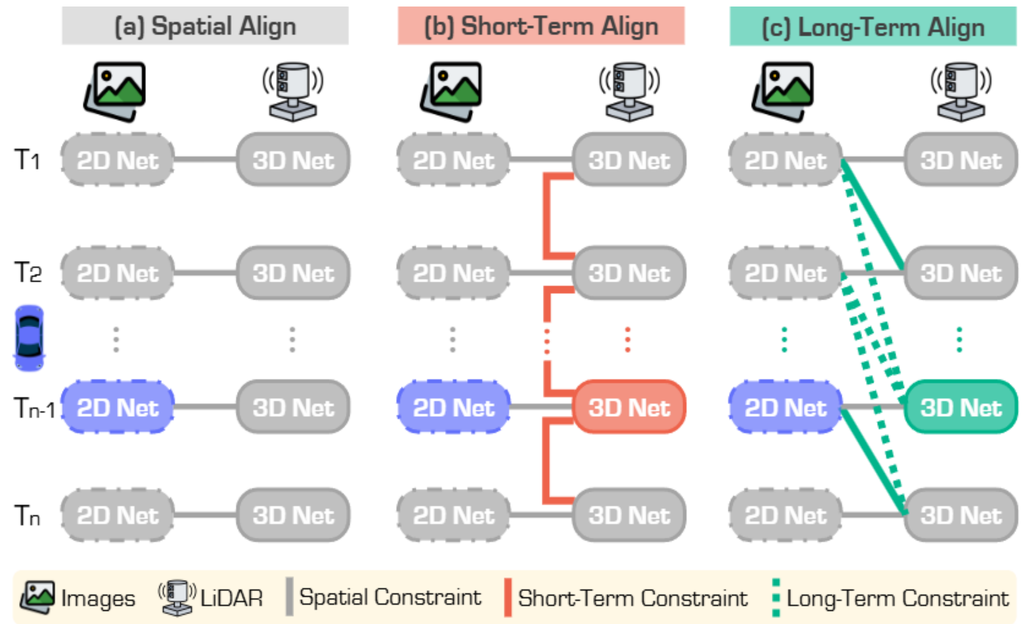
Xiang Xu   Lingdong Kong   Song Wang   Chuanwei Zhou   Qingshan Liu

ICCV HONOLULU HAWAII OCT 19-23, 2025

Motivation & Contribution

Overview of Approach

- **LiMA** is a novel long-term image-to-**Li**DAR **M**emory **A**ggregation framework, which explicitly captures the **longer range temporal correlations** to enhance LiDAR representation learning.
- **LiMA** designs an efficient **memory banking** structure to preserve historical 2D features, which enhances the **temporal consistency**, improving representations of LiDAR data, ultimately enables a more **effective and efficient** pretraining.

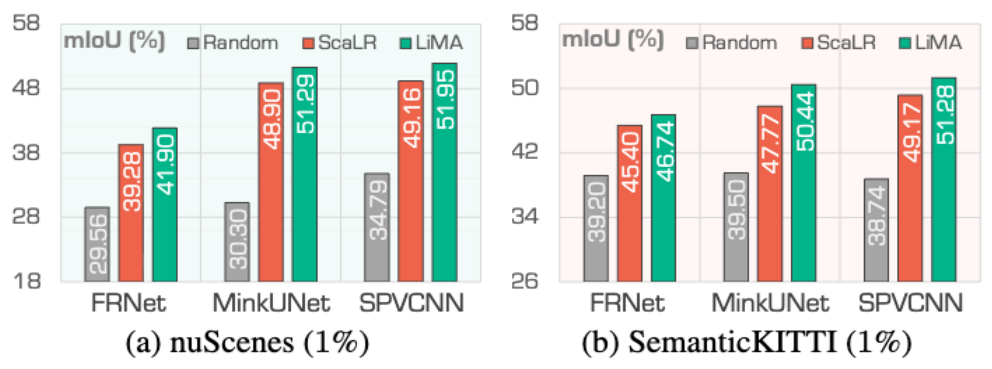


Motivation & Observation

- **Spatial Alignment** establishes accurate correspondences between **LiDAR features** and **image features** in the spatial domain, which often tends to disregard the temporal dynamics.
- **Short-Term** pretraining methods achieves temporal coherence by propagating LiDAR features **frame-by-frame**, ensuring consistent representations across neighboring frames, but are often limited in capturing the long-horizon dependencies across scans.
- **LiMA** take advantages of **Long-Term** image sequences to enable better LiDAR representation learning, thereby facilitating a more comprehensive understanding of **long-range dependencies** and complex motion pattern.

Compatible with Backbones

LiMA is integrated with different LiDAR representations, of which demonstrating strong **robustness** and **flexibility** in practice.

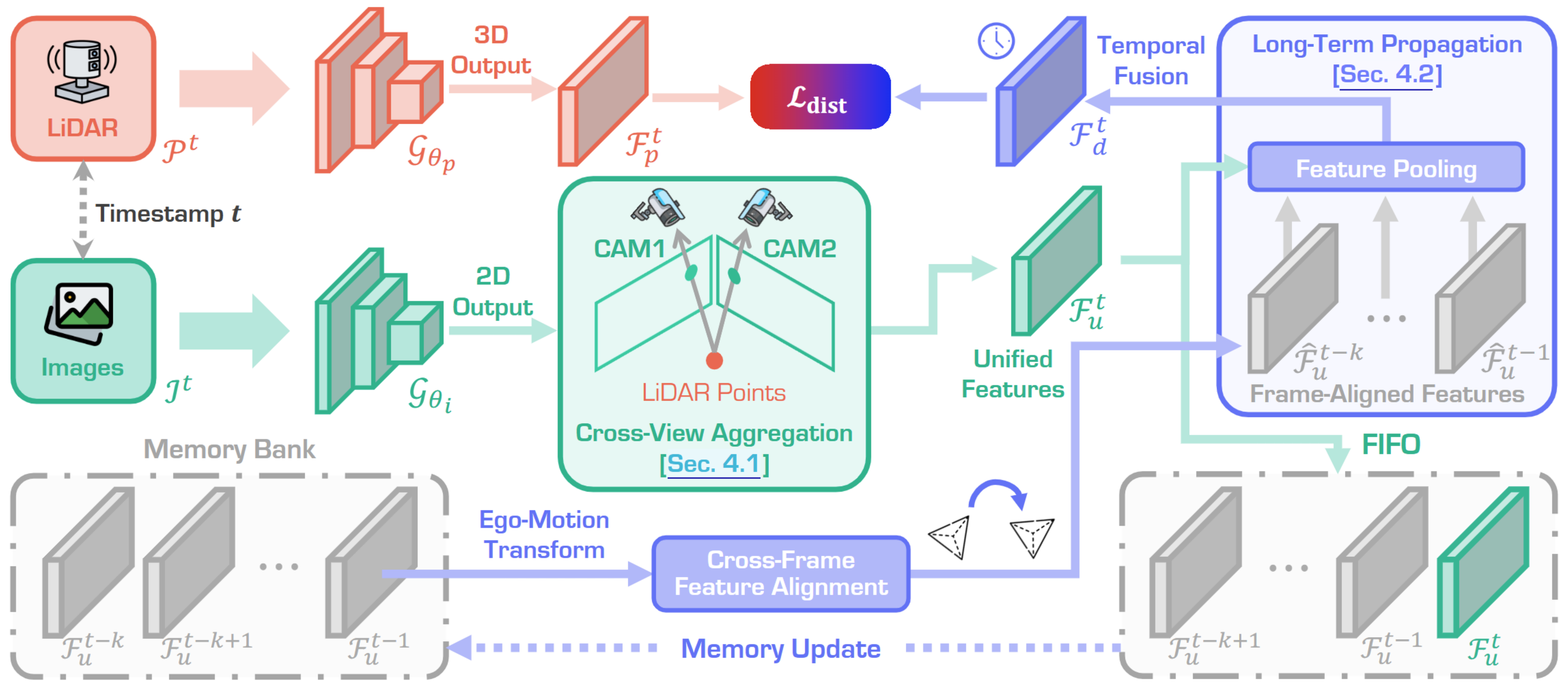


(a) nuScenes (1%)   (b) SemanticKITTI (1%)

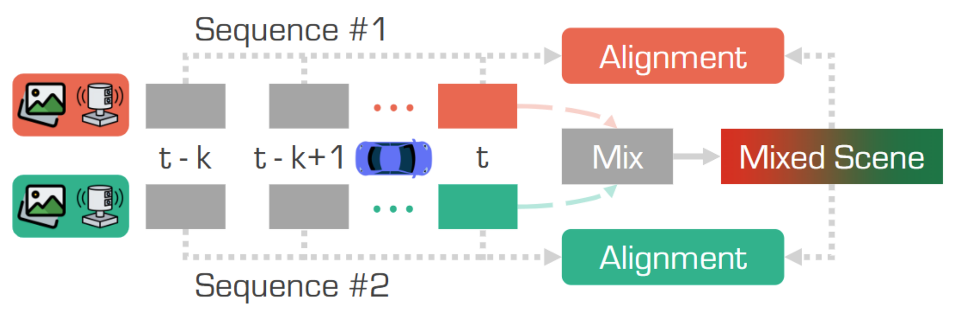
南京航空航天大学 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS
NUS National University of Singapore | Computing
OpenMMLab
南京郵電大學 Nanjing University of Posts and Telecommunications
浙江大學 ZHEJIANG UNIVERSITY

Design & Methodology

Image-to-LiDAR Memory Aggregation

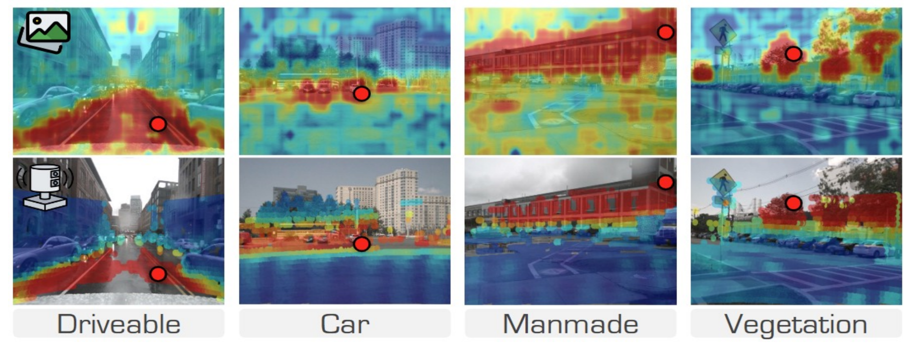


- **Cross-View Aggregation** enhances **spatial coherence** of LiDAR features by unifying multi-view image representations, mitigating optimization conflicts, and ensuring stable and efficient training.
- **Long-Term Feature Propagation** serves as a **cornerstone** of the framework, enabling the model to capture temporal dynamics efficiently. By integrating **motion-aware contexture** information across frames, **LiMA** enhances the spatial-temporal consistency of LiDAR feature learning, leading to better representations.
- **Memory Bank** stores unified image features from history frames, enabling the **temporal feature propagation and fusion**, reducing redundant computations and improving resource utilization.



Cross-Sequence Memory Bank serves as a **mixed pretraining strategy**, designed to bridge gaps across scans, improving the model generalizability.

Cosine similarity between query point (marked as **red** dot) and: (1) image features, and (2) LiDAR features projected onto images, showing **semantic coherence**.



Experiments & Observations

Comparative & Ablation Study

- **LiMA** achieves significant improvements across various datasets with the integration of **rich contextual temporal information**.

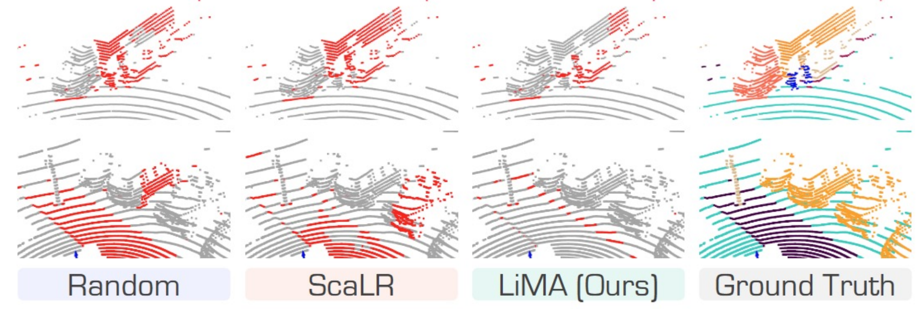
Tab. Compare with state-of-the-art LiDAR Pretraining methods

Method	Backbone (2D)	Backbone (3D)	Frames	nuScenes LP	nuScenes 1%	nuScenes 5%	nuScenes 10%	nuScenes 25%	nuScenes Full	KITTI 1%	Waymo 1%
Random	-	-	-	8.10	30.30	47.84	56.15	65.48	74.66	39.50	39.41
SLidR [68]	ResNet-50 [23]	MinkUNet-34 [12]	1	38.80	38.30	52.49	59.84	66.91	74.79	44.60	47.12
TriCC [60]			2	38.00	41.20	54.10	60.40	67.60	75.60	45.90	-
Seal [49]			2	44.95	45.84	55.64	62.97	68.41	75.60	46.63	49.34
CSC [6]			1	46.00	47.00	57.00	63.30	68.60	75.70	47.20	-
HVDistill [104]			1	39.50	42.70	56.60	62.90	69.30	76.60	49.70	-
Seal [49]	ViT-S [14]	MinkUNet-34 [12]	2	45.16	44.27	55.13	62.46	67.64	75.58	46.51	48.67
SuperFlow [96]			3	46.44	47.81	59.44	64.47	69.20	76.54	47.97	49.94
ScaLR [63]			1	49.66	45.89	56.52	61.07	65.79	73.39	46.06	47.67
LiMA			6	**54.76**	**48.75**	**60.83**	**65.41**	**69.31**	**76.94**	**49.28**	**50.23**
Seal [49]	ViT-B [14]	MinkUNet-34 [12]	2	46.59	45.98	57.15	62.79	68.18	75.41	47.24	48.91
SuperFlow [96]			3	47.66	48.09	59.66	64.52	69.79	76.57	48.40	50.20
ScaLR [63]			1	51.90	48.90	57.69	62.88	66.85	74.15	47.77	49.38
LiMA			6	**56.65**	**51.29**	**61.11**	**65.62**	**70.43**	**76.91**	**50.44**	**51.35**
Seal [49]	ViT-L [14]	MinkUNet-34 [12]	2	46.81	46.27	58.14	63.27	68.67	75.66	47.55	50.02
SuperFlow [96]			3	48.01	49.95	60.72	65.09	70.01	77.19	49.07	50.67
ScaLR [63]			1	51.77	49.13	58.36	62.75	66.80	74.16	48.64	49.72
LiMA			6	**56.67**	**53.22**	**62.46**	**66.00**	**70.59**	**77.23**	**52.29**	**51.19**

Tab. Pretraining efficiency

Method	Frames	Training Time (Hours)	Memory (GB)	nuScenes LP	nuScenes 1%	KITTI 1%
ScaLR [63]	1	~ 10.1	12.43	51.90	48.90	47.77
LiMA (Ours)	2	~ 14.7	18.23	53.34	49.14	48.27
	3	~ 15.3	20.67	54.52	49.75	48.92
	4	~ 16.1	23.19	55.65	50.29	49.44
	5	~ 17.0	26.59	56.03	50.95	50.32
	6	~ 17.9	29.07	**56.65**	**51.29**	50.44
	7	~ 18.7	33.55	55.37	50.91	51.00
	8	~ 19.5	36.27	54.97	49.36	**51.78**
Seal [49]	2	~ 27.3	20.92	46.59	45.98	47.24
SuperFlow [96]	3	~ 30.7	23.65	47.66	48.09	48.40

- Compared to **spatial alignment**, **LiMA** increases the number of propagated frames generally improves performance.
- Compared to those temporal methods, **LiMA** achieves higher efficiency and **more effective memory usage**, which can be credited to the memory bank.
- **LiMA** obtains performance with a **fewer mis-classification** and better localization of dynamics under limited annotations.
- **LiMA** leverages the long-term contexts to **better align spatial-temporal cues**, enabling much better performances in scenes where 3D objects move rapidly.



- We hope this work can **pave the foundation** for future work in 3D representation learning.