LiMoE: Mixture of LiDAR Representation Learners From Automotive Scenes

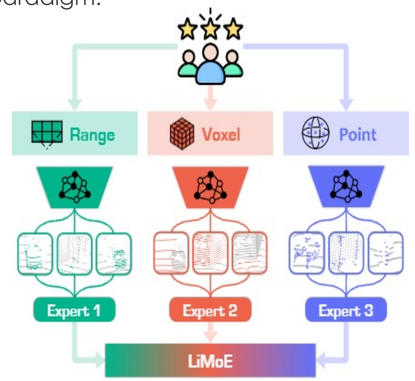
Xiang Xu* Lingdong Kong* Hui Shuai Liang Pan Ziwei Liu Qingshan Liu

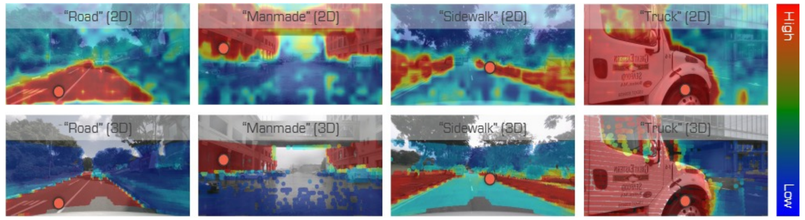
CVPR Nashville JUNE 11-15, 2025

Motivation & Contribution

Overview of Approach

- **LiMoE** is a new LiDAR-based cross-sensor representation learning framework that aims to integrate multiple LiDAR representations dynamically through the **MoE** paradigm.
- **LiMoE** integrates three popular LiDAR representations: **range**, **voxel**, and **point**, into a unified framework.
- With **MoE** design to combine multiple LiDAR representations, **LiMoE** enables use to capture potential complementary cue from each representation with specific task objectives.
- Compared to existing single-representation methods, we achieves large improvements across **11** LiDAR datasets.



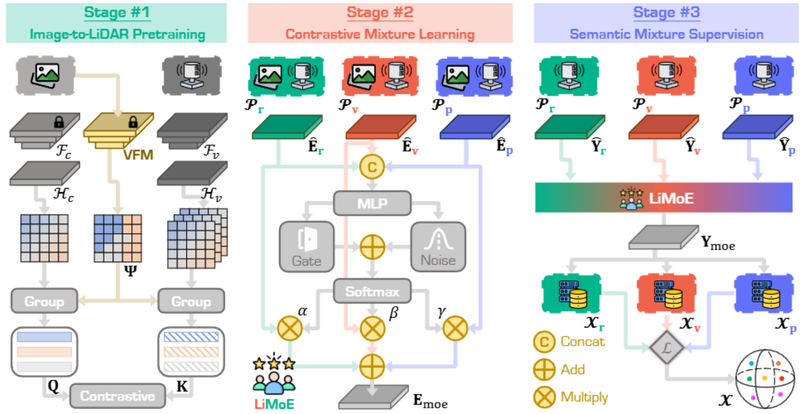


- Similarity map between a query point and the pretrained 2D image backbone as well as other LiDAR points during the **CML** stage demonstrate that **LiMoE** enhances semantic features during pretraining, even without semantic supervision.

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

Design & Methodology

Three-Stage Training Paradigm



- **Image-to-LiDAR Pretraining** leverages previous LiDAR-based cross-sensor data pretraining framework to transfer prior knowledge from images to point clouds across different LiDAR representations, providing semantic-rich initial parameters for each representation network.
- **Contrastive Mixture Learning (CML)** fuses the pretrained features across various LiDAR representations from stage #1 into a unified representation via a Mixture of Expert (MoE) layer. With the tailored contrastive learning objective, CML encourages the MoE to focus on data attributes from laser beams and attributes that play different roles from representations.
- **Semantic Mixture Supervision (SMS)** aims to enhance the downstream performance by combining semantic logits with the MoE layers. By supervising the semantic logits in such a way, the SMS module can encourage experts to focus on object attributes in the LiDAR scenes.

(a) Expert Activation Paths (Beam Number) (b) Expert Activation Paths (Distances to Ego)

- We observe distinct and complementary LiDAR attributes from our MoE modules that each LiDAR representation potentially focuses on.

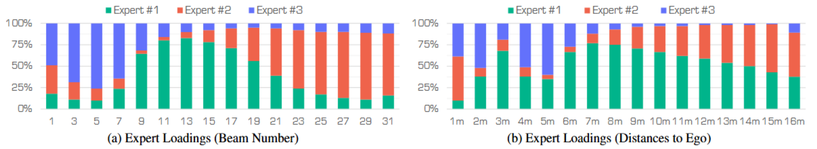
Experiments & Analysis

Comparative & Ablation Study

- **LiMoE** achieves significant improvements across various datasets with the suitable integrations of multiple LiDAR representations.

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

Method	Venue	Backbone (2D)	Backbone (3D)	Expert	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 [63]	CVPR'22	ResNet-50 [25]	MinkUNet-34 [14]	Single ○	38.80	38.30	52.49	59.84	66.91	74.79	44.60	47.12
TriCC [55]	CVPR'23			Single ○	38.00	41.20	54.10	60.40	67.60	75.60	45.90	–
Seal [43]	NeurIPS'23			Single ○	44.95	45.84	55.64	62.97	68.41	75.60	46.63	**49.34**
CSC [8]	CVPR'24			Single ○	**46.00**	**47.00**	**57.00**	**63.30**	68.60	75.70	47.20	–
HVDistill [85]	IJCV'24			Single ○	39.50	42.70	56.60	62.90	**69.30**	**76.60**	**49.70**	–
SLidR [63]	CVPR'22	ViT-S [53]	MinkUNet-34 [14]	Single ○	44.70	41.16	53.65	61.47	66.71	74.20	44.67	47.57
+ LiMoE	Ours			Multi ●	45.80	46.82	57.54	63.85	68.61	75.64	46.81	48.81
Seal [43]	NeurIPS'23			Single ○	45.16	44.27	55.13	62.46	67.64	75.58	46.51	48.67
SuperFlow [80]	ECCV'24			Single ○	46.44	47.81	59.44	64.47	69.20	76.54	47.97	49.94
+ LiMoE	Ours			Multi ●	**48.20**	**49.60**	**60.54**	**65.65**	**71.39**	**77.27**	**49.53**	**51.42**
SLidR [63]	CVPR'22	ViT-B [53]	MinkUNet-34 [14]	Single ○	45.35	41.64	55.83	62.68	67.61	74.98	45.50	48.32
+ LiMoE	Ours			Multi ●	46.56	46.89	58.09	63.87	69.02	75.87	47.96	49.50
Seal [43]	NeurIPS'23			Single ○	46.59	45.98	57.15	62.79	68.18	75.41	47.24	48.91
SuperFlow [80]	ECCV'24			Single ○	47.66	48.09	59.66	64.52	69.79	76.57	48.40	50.20
+ LiMoE	Ours			Multi ●	**49.07**	**50.23**	**61.51**	**66.17**	**71.56**	**77.81**	**50.30**	**51.77**
SLidR [63]	CVPR'22	ViT-L [53]	MinkUNet-34 [14]	Single ○	45.70	42.77	57.45	63.20	68.13	75.51	47.01	48.60
+ LiMoE	Ours			Multi ●	47.43	46.92	58.41	64.54	69.69	76.32	48.25	50.23
Seal [43]	NeurIPS'23			Single ○	46.81	46.27	58.14	63.27	68.67	75.66	47.55	50.02
SuperFlow [80]	ECCV'24			Single ○	48.01	49.95	60.72	65.09	70.01	77.19	49.07	50.67
+ LiMoE	Ours			Multi ●	**49.35**	**51.41**	**62.07**	**66.64**	**71.59**	**77.85**	**50.69**	**51.93**



(a) Expert Loadings (Beam Number) (b) Expert Loadings (Distances to Ego)

- In the **CML** stage, we observe that different representations focus on distinct LiDAR data attributes. The range images predominantly capture middle beams and distances, sparse voxels focus more on upper beams and longer distances, while raw LiDAR points can concentrate on lower beams and near distances around ego-car.
- In the **SMS** stage, we find that the range images are more sensitive to dynamic objects, voxels can highlight background objects, and raw LiDAR points capture more detailed objects.