Segment Any Point Cloud Sequences by Distilling Vision Foundation Models

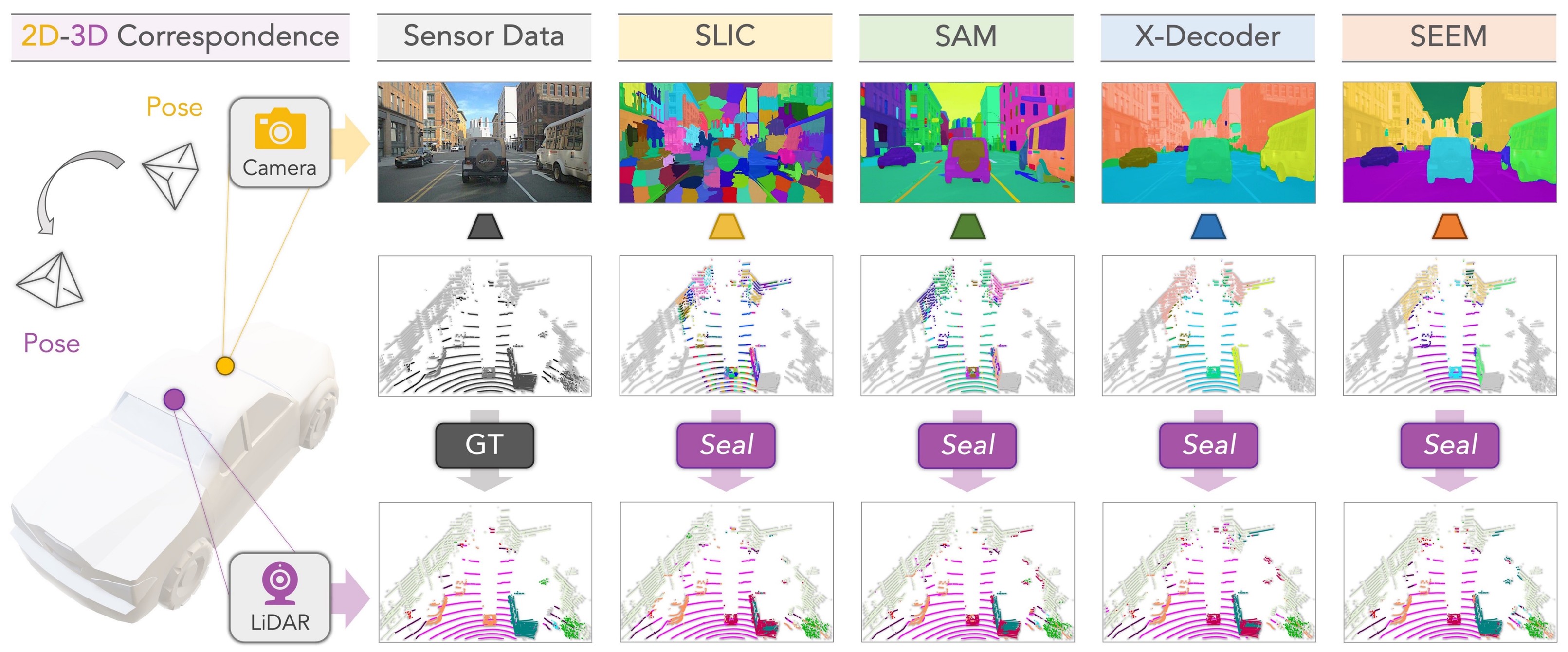
Youquan Liu*, Lingdong Kong*,
Jun Cen, Runnan Chen, Wenwei Zhang,
Liang Pan, Kai Chen, Ziwei Liu

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation & Contribution

TL;DR

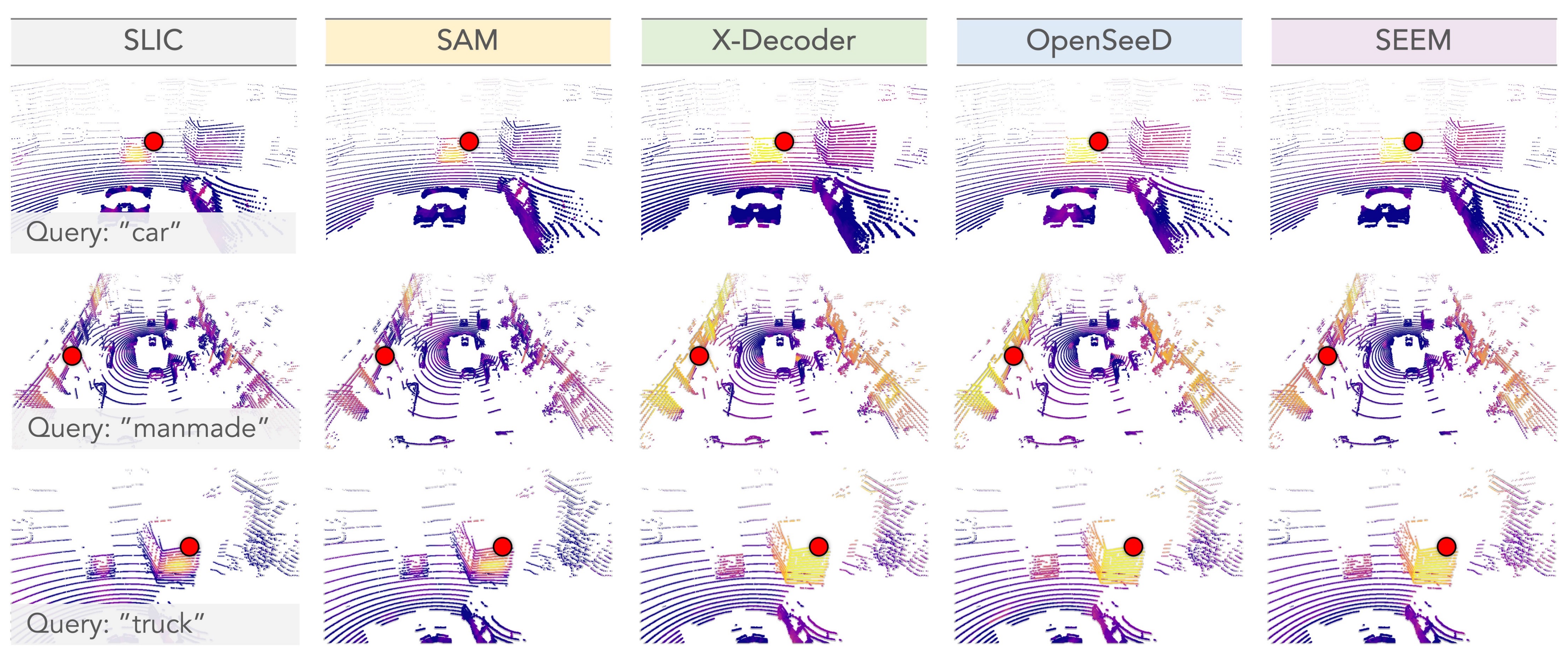
❖ We introduce Seal, a novel framework tailored to harness vision foundation models (VFMs) to segment diverse automotive point cloud sequences.



❖ Seal has three appealing properties: i) Scalability for not needing either 2D or 3D annotations during pretraining; ii) Consistency for aligning between LiDAR and camera via cross-modal contrastive learning; iii) Generalizability for exhibiting effectiveness across a wide range of point cloud datasets.

Knowledge Transfer from VFMs

❖ VFMs can generate superpixels from the camera views and provide off-the-shelf semantic coherence for distinct objects & backgrounds in the 3D scene.



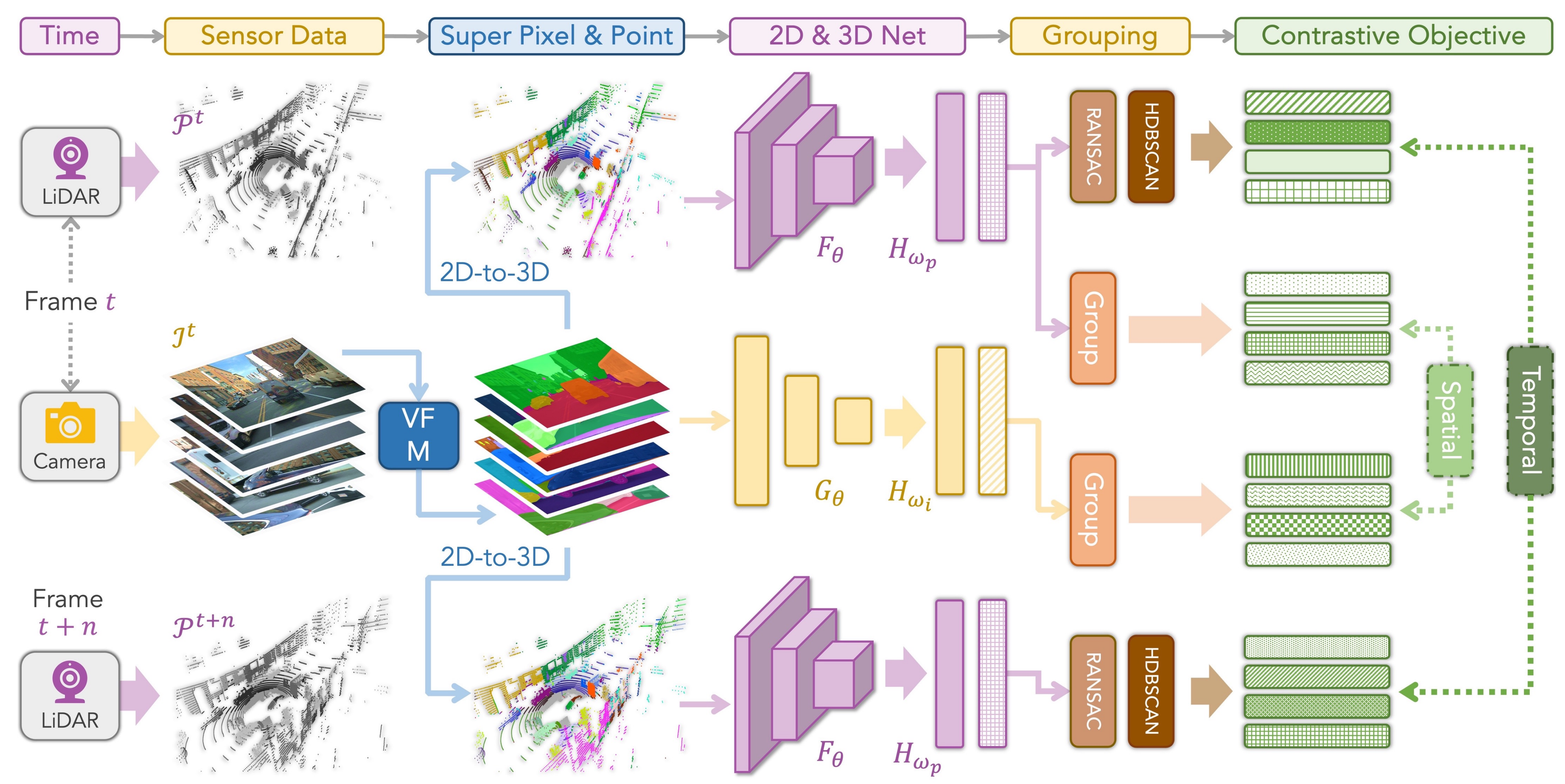
❖ Compared to prior works, our VFM-assisted contrastive learning: i) mitigates the severe self-conflict problem; ii) forms a more coherent optimization landscape, yielding a faster convergence rate; iii) reduces the number of superpixels generated, which extenuates the overhead during pretraining.

Methodology

The Seal Framework

❖ We generate, for each {LiDAR, camera} pair $\{\mathcal{P}^t, \mathcal{I}^t\}$ at timestamp t and another LiDAR frame $\mathcal{P}^{t+n}$ at timestamp $t+n$, the semantic superpixels and superpoints by VFMs. Two pertaining objectives are then leveraged.

❖ We aim to encourage i) spatial contrastive between paired LiDAR-camera features for cross-sensor learning; ii) temporal consistency between point segments at two different timestamps for semantic view regularization.



❖ These two regularization objectives are complementary to each other; a combination of both introduces strong consistency during the pretraining.

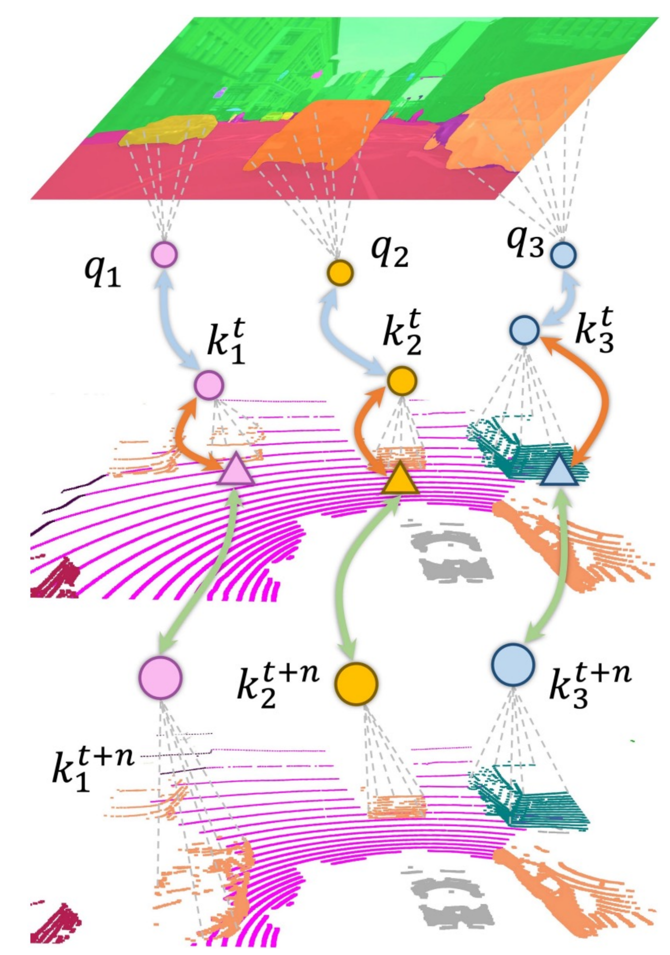
Spatial-Temporal Consistency

❖ Seal defines a suitable positive feature correspondence in contrastive learning via implicit geometric clustering.

❖ Such a design can mitigate the potential errors caused by inaccurate cross-sensor calibration and synchronization.

❖ Besides, point-to-segment regularization mechanism can serve to aggregate the spatial information thus yielding better-distinguishing instances in LiDAR scenes.

❖ Regularization at different levels enables Seal to be consistent and generalizable.



Experiments & Analyses

Comparative Study

❖ We verify the effectiveness of Seal across eleven point cloud datasets with various scales, modalities, sensor configurations, fidelities, and noise levels.

Method & Year	nuScenes LP	nuScenes 1%	nuScenes 5%	nuScenes 10%	nuScenes 25%	nuScenes Full	KITTI 1%	Waymo 1%	Synth4D 1%
Random	8.10	30.30	47.84	56.15	65.48	74.66	39.50	39.41	20.22
PointContrast [ECCV'20] [103]	21.90	32.50	-	-	-	-	41.10	-	-
DepthContrast [ICCV'21] [116]	22.10	31.70	-	-	-	-	41.50	-	-
PPKT [arXiv'21] [65]	35.90	37.80	53.74	60.25	67.14	74.52	44.00	47.60	61.10
SLidR [CVPR'22] [85]	38.80	38.30	52.49	59.84	66.91	74.79	44.60	47.12	63.10
ST-SLidR [CVPR'23] [66]	40.48	40.75	54.69	60.75	67.70	75.14	44.72	44.93	-
Seal (Ours)	**44.95**	**45.84**	**55.64**	**62.97**	**68.41**	**75.60**	**46.63**	**49.34**	**64.50**
Seal † (Ours)	-	48.41	57.84	65.52	70.80	77.13	-	-	-
Seal ‡ (Ours)	-	49.53	58.64	66.78	72.31	78.28	-	-	-

❖ Our approach constantly outperforms previous works on every setting by large margins, which strongly demonstrates the superiority and scalability.

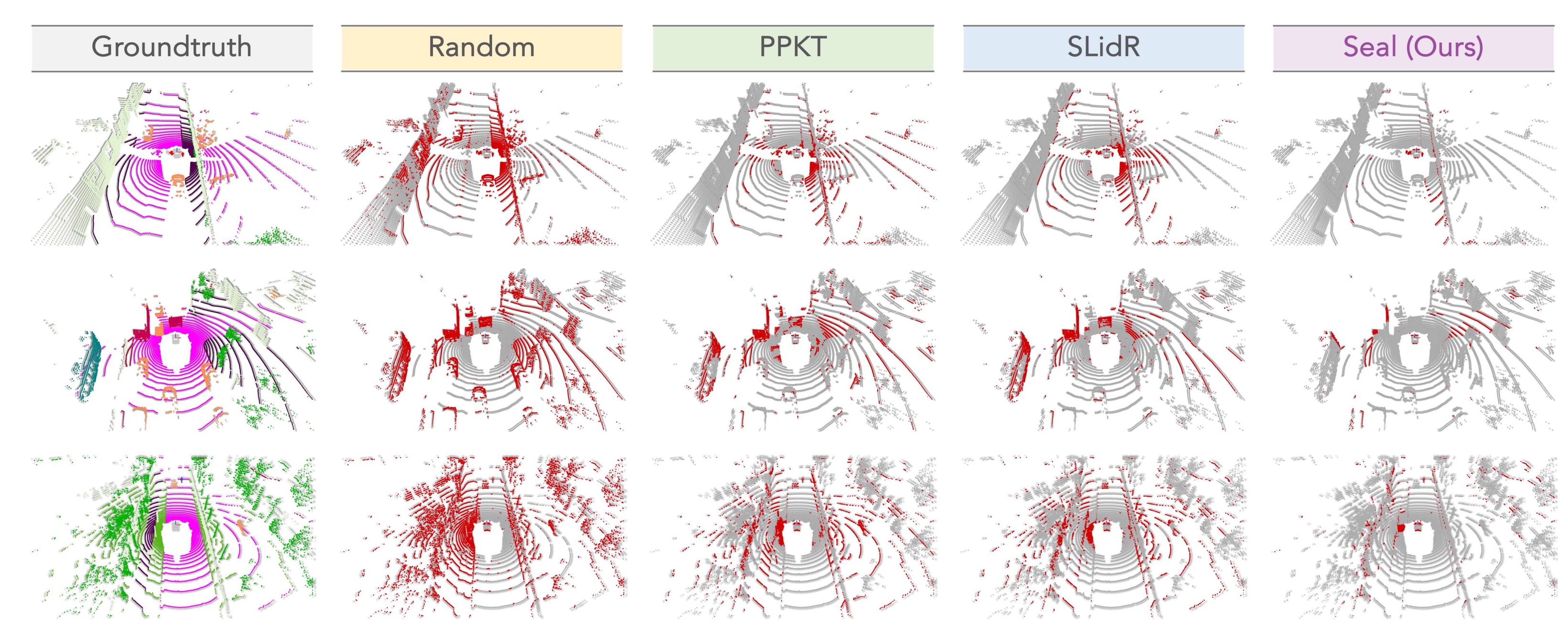
Method	ScribbleKITTI 1%	ScribbleKITTI 10%	RELLIS-3D 1%	RELLIS-3D 10%	SemanticPOSS Half	SemanticPOSS Full	SemanticSTF Half	SemanticSTF Full	SynLiDAR 1%	SynLiDAR 10%	DAPS-3D Half	DAPS-3D Full
Random	23.81	47.60	38.46	53.60	46.26	54.12	48.03	48.15	19.89	44.74	74.32	79.38
PPKT [65]	36.50	51.67	49.71	54.33	50.18	56.00	50.92	54.69	37.57	46.48	78.90	84.00
SLidR [85]	39.60	50.45	49.75	54.57	51.56	55.36	52.01	54.35	42.05	47.84	81.00	85.40
Seal (Ours)	**40.64**	**52.77**	**51.09**	**55.03**	**53.26**	**56.89**	**53.46**	**55.36**	**43.58**	**49.26**	**81.88**	**85.90**

Ablation Study

❖ The effectiveness of each component in Seal has been proven; with spatial & temporal contrastive, we can learn meaningful multi-modal representations.

#	C2L	VFM	STC	P2S	nuScenes LP	nuScenes 1%	nuScenes 5%	nuScenes 10%	nuScenes 25%	nuScenes Full	KITTI 1%	Waymo 1%
(1)	✓				38.80	38.30	52.49	59.84	66.91	74.79	44.60	47.12
(2)	✓		✓		40.45	41.62	54.67	60.48	67.61	75.30	45.38	48.08
(3)	✓	✓			43.00	44.02	53.03	60.84	67.38	75.21	45.72	48.75
(4)	✓	✓	✓		44.01	44.78	55.36	61.99	67.70	75.00	46.49	49.15
(5)	✓	✓		✓	43.35	44.25	53.69	61.11	67.42	75.44	46.07	48.82
(6)	✓	✓	✓	✓	**44.95**	**45.84**	**55.64**	**62.97**	**68.41**	**75.60**	**46.63**	**49.34**

❖ Qualitative results show that Seal can segment complex driving scenes in 3D.



OpenMMLab

NUS National University of Singapore

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

香 港 大 學 THE UNIVERSITY OF HONG KONG

Paper

Code