4D Contrastive Superflows are Dense 3D Representation Learners

Xiang Xu[1,*] Lingdong Kong[2,*] Hui Shuai[3] Wenwei Zhang[4] Liang Pan[4] Kai Chen[4] Ziwei Liu[5] QingShan Liu[3]

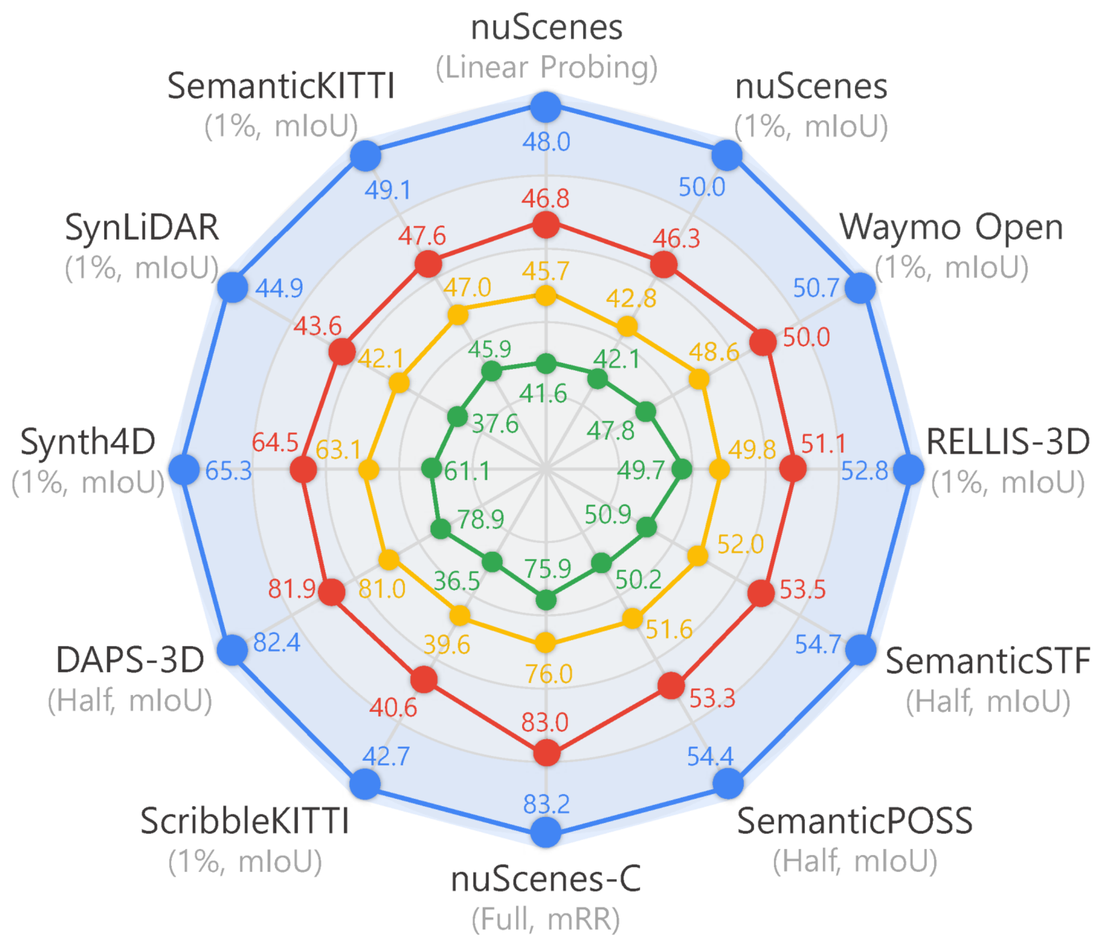
上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

Paper Code

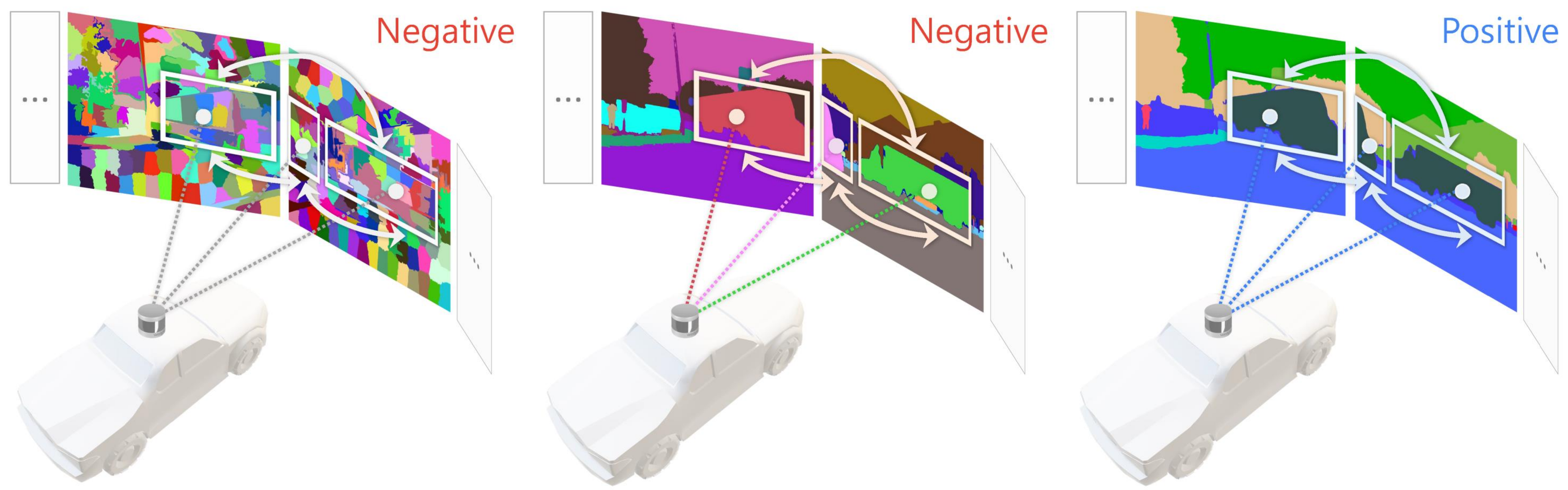
Motivation & Contribution

TL;DR

- We introduce **SuperFlow**, a novel framework designed to harness consecutive LiDAR-camera pairs for establishing spatiotemporal pretraining objectives.
- Our **SuperFlow** in corporates novel designs including view consistency alignment, dense-to-sparse regularization, and flow-based contrastive learning, which better encourages data representation learning effects between camera and LiDAR sensors across consecutive scans.
- Extensive comparative across **11** heterogeneous datasets validate the effectiveness and superiority of **SuperFlow**.



View Consistency Alignment



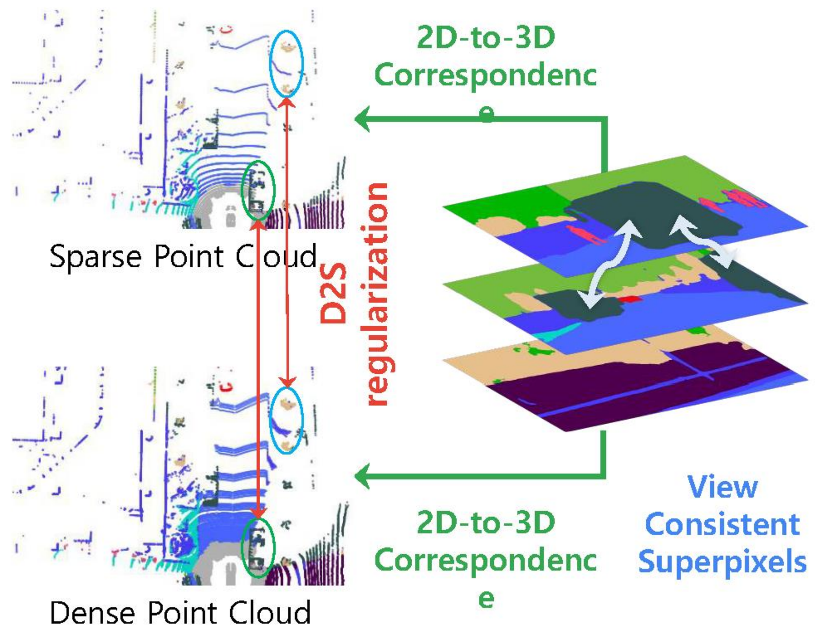
(a) Heuristic (b) Class Agnostic (c) View Consistent

- We employ CLIP's text encoder and fine-tune the last layer of the segmentation head from visual foundation models with predefined text prompts, which allows the segmentation head to generate language-guided semantic categories for each pixel.

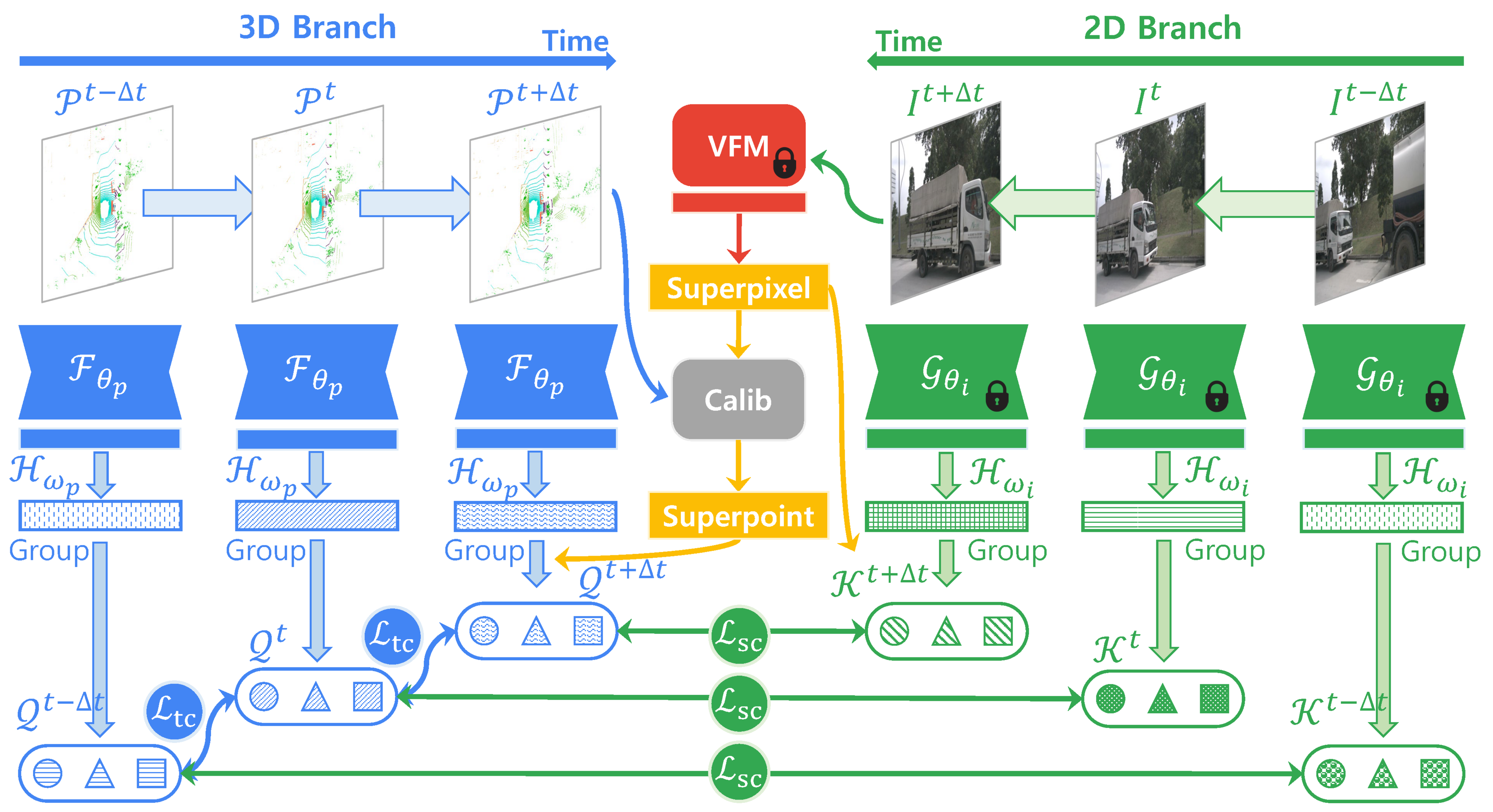
Methodology

D2S: Dense-to-Sparse Consistency Regularization

- Due to the natural of LiDAR scanning and data acquisition, different areas within the same scene can have significantly different point densities.
- We combine multi-sweep point clouds from consecutive frames to regularize the features of the key frame point cloud via semantic superpoints.



FCL: Flow-Based Contrastive Learning



- **SuperFlow** takes multiple LiDAR-camera pairs from consecutive scans as input and establishes spatial consistency across sensors and temporal consistency across times.
- Spatial consistency module is designed to align 3D representation with 2D prior knowledge via contrastive learning loss.
- Temporal consistency module focuses on consistent dynamics via the semantic flow across different scenes.

Experiments & Analysis

In-Domain and Cross-Domain Benchmarks

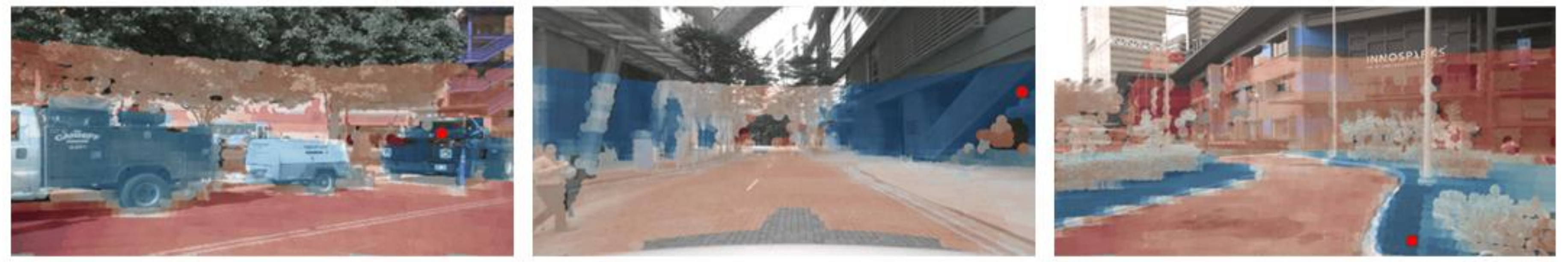
- In-domain and cross-domain downstream tasks verify the effectiveness of **SuperFlow**, and the larger-scale of 2D pretrained network also contribute to better representations.

Method	Venue	Distill	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
PPKT [63]	arXiv'21	ViT-S ○	38.60	40.60	52.06	59.99	65.76	73.97	43.25	47.44
SLidR [82]	CVPR'22	ViT-S ○	44.70	41.16	53.65	61.47	66.71	74.20	44.67	47.57
Seal [61]	NeurIPS'23	ViT-S ○	45.16	44.27	55.13	62.46	67.64	75.58	46.51	48.67
SuperFlow	**Ours**	ViT-S ●	46.44	47.81	59.44	64.47	69.20	76.54	47.97	49.94
PPKT [63]	arXiv'21	ViT-B ○	39.95	40.91	53.21	60.87	66.22	74.07	44.09	47.57
SLidR [82]	CVPR'22	ViT-B ○	45.35	41.64	55.83	62.68	67.61	74.98	45.50	48.32
Seal [61]	NeurIPS'23	ViT-B ○	46.59	45.98	57.15	62.79	68.18	75.41	47.24	48.91
SuperFlow	**Ours**	ViT-B ●	47.66	48.09	59.66	64.52	69.79	76.57	48.40	50.20
PPKT [63]	arXiv'21	ViT-L ○	41.57	42.05	55.75	61.26	66.88	74.33	45.87	47.82
SLidR [82]	CVPR'22	ViT-L ○	45.70	42.77	57.45	63.20	68.13	75.51	47.01	48.60
Seal [61]	NeurIPS'23	ViT-L ○	46.81	46.27	58.14	63.27	68.67	75.66	47.55	50.02
SuperFlow	**Ours**	ViT-L ●	48.01	49.95	60.72	65.09	70.01	77.19	49.07	50.67

Ablation Study

- Consistent improvements across varying datasets with the scale-up of the 3D network except for **MinkUNet-101** with a large set of trainable parameters that tends to be difficult to converge.

Backbone	Layer	nuScenes LP	nuScenes 1%	KITTI 1%	Waymo 1%
MinkUNet ○	18	47.20	47.70	48.04	49.24
MinkUNet ●	34	47.66	48.09	48.40	50.20
MinkUNet ○	50	54.11	52.86	49.22	51.20
MinkUNet ○	101	52.56	51.19	48.51	50.01



(a) "car" (3D) (b) "manmade" (3D) (c) "sidewalk" (3D)

(d) "car" (2D) (e) "manmade" (2D) (f) "sidewalk" (2D)