运用视觉基础模型分割「任意」点云

孔令东

2023/07/20

【OpenMMLab 社区开放麦】是由 OpenMMLab 发起，面向所有社区成员的社区分享直播活动，每周四晚八点准点播出。旨在搭建一个知识分享的舞台，在这里，社区里的每个人都能拿起话筒分享你的知识和见解。我们一直认为，分享与交流能更好地促进知识的传播；平等与共建能更好地维持社区的氛围。

Segment Any Point Cloud Sequences by Distilling Vision Foundation Models

Youquan Liu[1,*], Lingdong Kong[1,2,*], Jun Cen[1,3], Runnan Chen[1,4]

Wenwei Zhang[1,5], Liang Pan[5], Kai Chen[1], Ziwei Liu[5]

[1]Shanghai AI Lab [2]NUS [3]HKUST [4]HKU [5]S-Lab, NTU

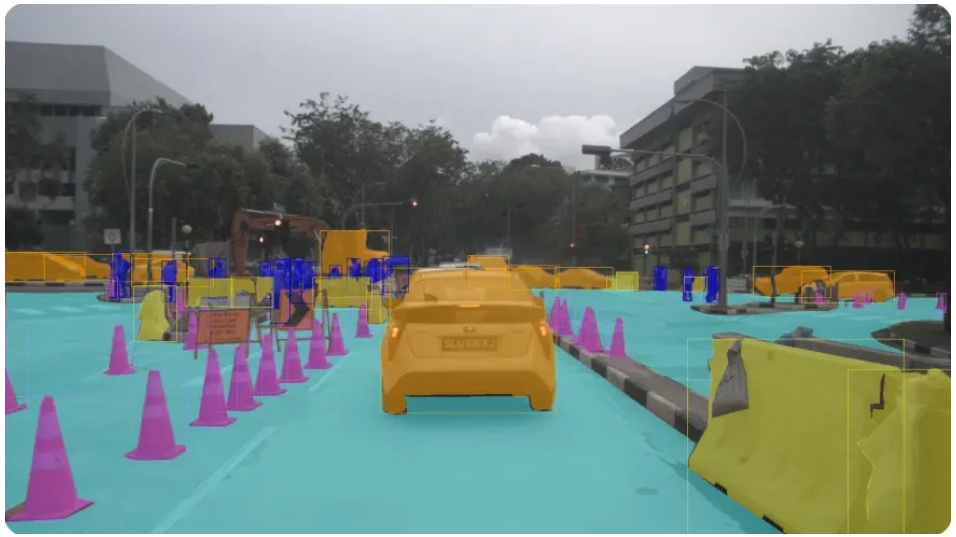
Topic

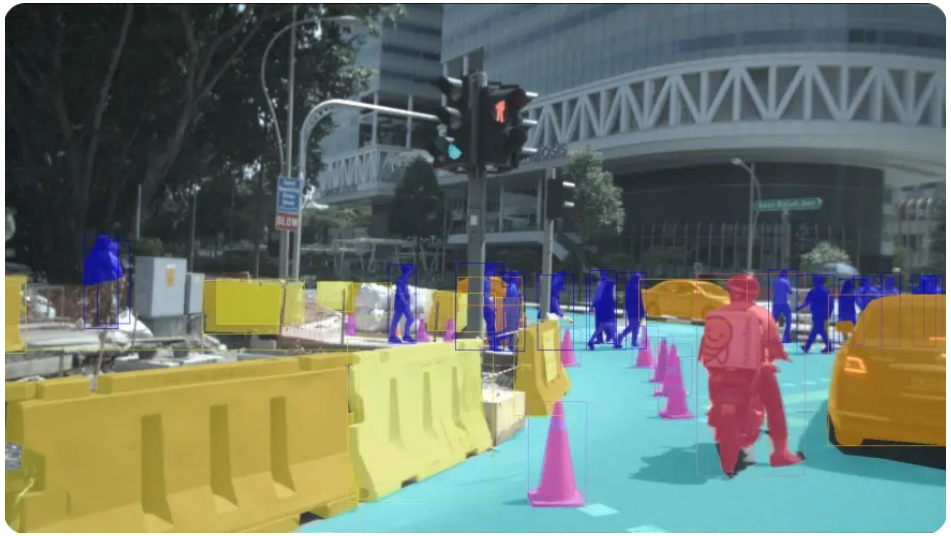
1. Automotive Point Cloud Segmentation
   自动驾驶场景中的点云分割
2. Point Cloud Segmentation with MMDetection3D
   高效点云分割工具箱
3. Introduction to Segment Anything
   简介「分割一切」模型
4. Seal: Segmentation Any Point Cloud Sequences
   运用视觉基础模型分割「任意」点云序列

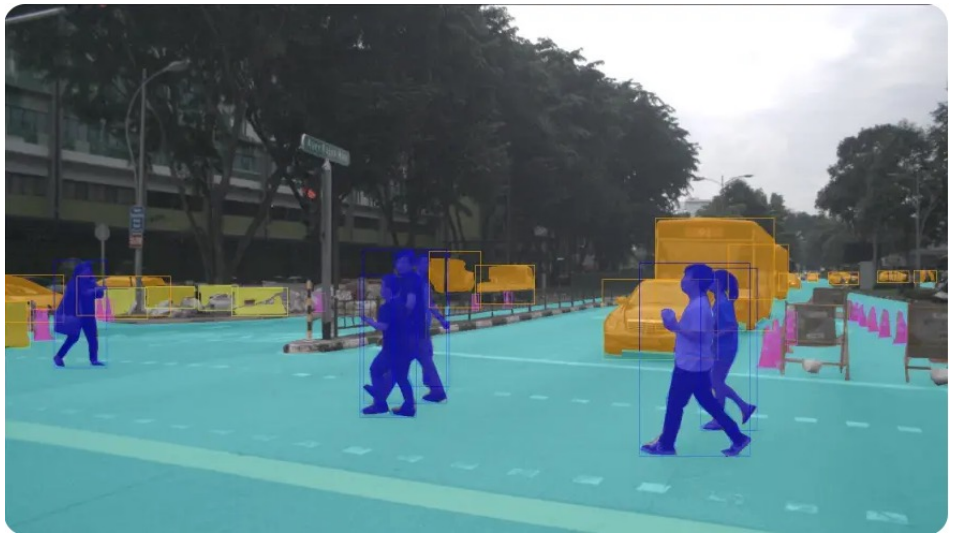
Topic

1. Automotive Point Cloud Segmentation
   自动驾驶场景中的点云分割
2. Point Cloud Segmentation with MMDetection3D
   高效点云分割工具箱
3. Introduction to Segment Anything
   简介「分割一切」模型
4. Seal: Segmentation Any Point Cloud Sequences
   运用视觉基础模型分割「任意」点云序列

Visual Perception System







H. Caesar, et al. "nuScenes: A multimodal dataset for autonomous driving," CVPR, 2020.

Visual Perception System

Novel Class

Novel Instance

OpenCC: https://github.com/Charmve/OpenCC

Visual Perception System

Here's a visual example of the **real-world challenges** autonomous driving perception systems face

H. Caesar, et al. "nuScenes: A multimodal dataset for autonomous driving," CVPR, 2020.

Visual Perception System

Here's a visual example of the **real-world challenges** autonomous driving perception systems face

- A single camera obviously cannot capture all **360 degrees** around a vehicle

H. Caesar, et al. "nuScenes: A multimodal dataset for autonomous driving," CVPR, 2020.

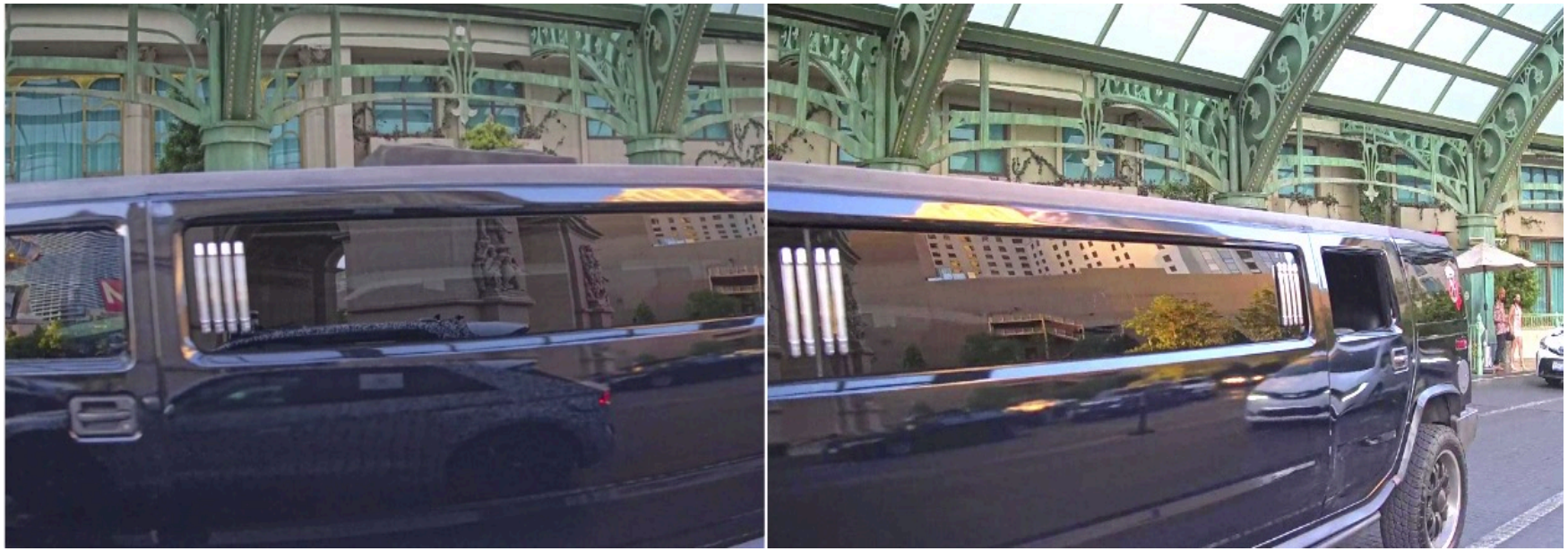
Visual Perception System

Here's a visual example of the **real-world challenges** autonomous driving perception systems face

- A single camera obviously cannot capture all **360 degrees** around a vehicle
- Just by looking at this image and lacking any additional context, it is **unclear** to the onboard computer what this object is. It could be a **barricade**, a **bus**, a **train** - any number of things

H. Caesar, et al. "nuScenes: A multimodal dataset for autonomous driving," CVPR, 2020.

Visual Perception System



- A single camera obviously cannot capture all **360 degrees** around a vehicle
- Just by looking at this image and lacking any additional context, it is **unclear** to the onboard computer what this object is. It could be a **barricade**, a **bus**, a **train** - any number of things
- At this point, our system doesn't have enough context to either **classify the object** or **define the object's boundaries** – two essential perception tasks

H. Caesar, et al. "nuScenes: A multimodal dataset for autonomous driving," CVPR, 2020.

Visual Perception System

- With these **three** camera views, the vehicle's perception network now has all the information and context needed to classify the object, as well as bound it properly

H. Caesar, et al. "nuScenes: A multimodal dataset for autonomous driving," CVPR, 2020.

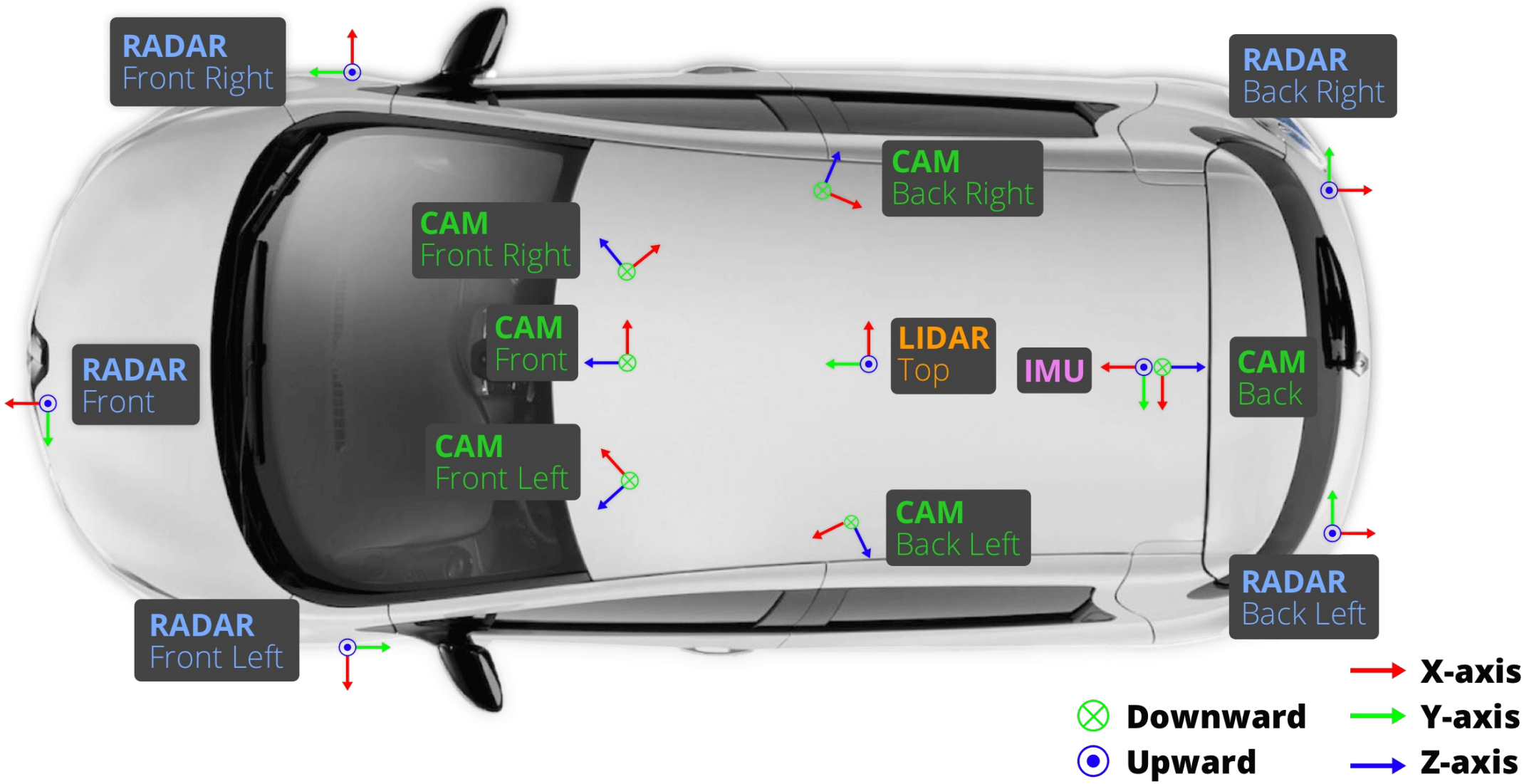
Visual Perception System

- With these **three** camera views, the vehicle's perception network now has all the information and context needed to classify the object, as well as bound it properly

- Solution:
  - Instead of relying solely on camera, seeking more modalities from other sensors
  - RADAR, **LiDAR**, IMU, etc.

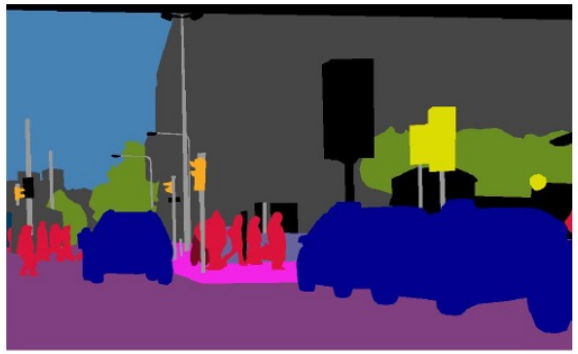
H. Caesar, et al. "nuScenes: A multimodal dataset for autonomous driving," CVPR, 2020.

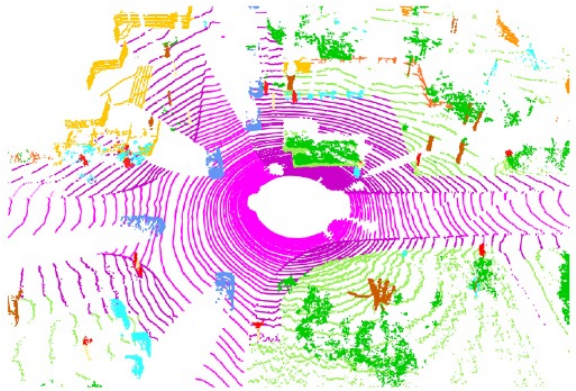
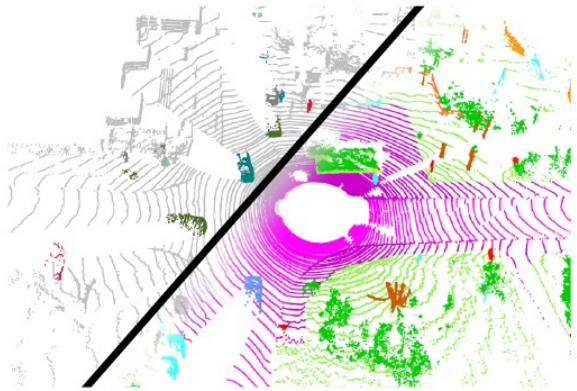
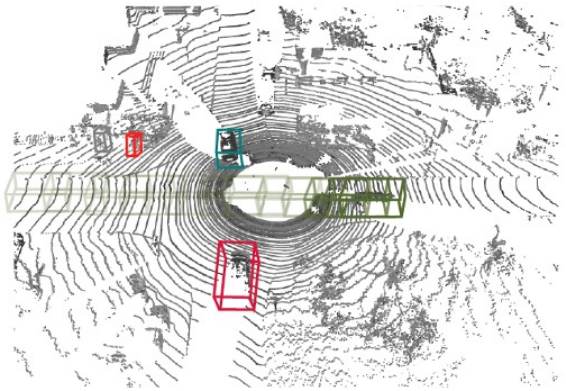
Autonomous Driving Perception System



H. Caesar, et al. "nuScenes: A multimodal dataset for autonomous driving," CVPR, 2020.

2D Perception vs. 3D Perception


From left to right:

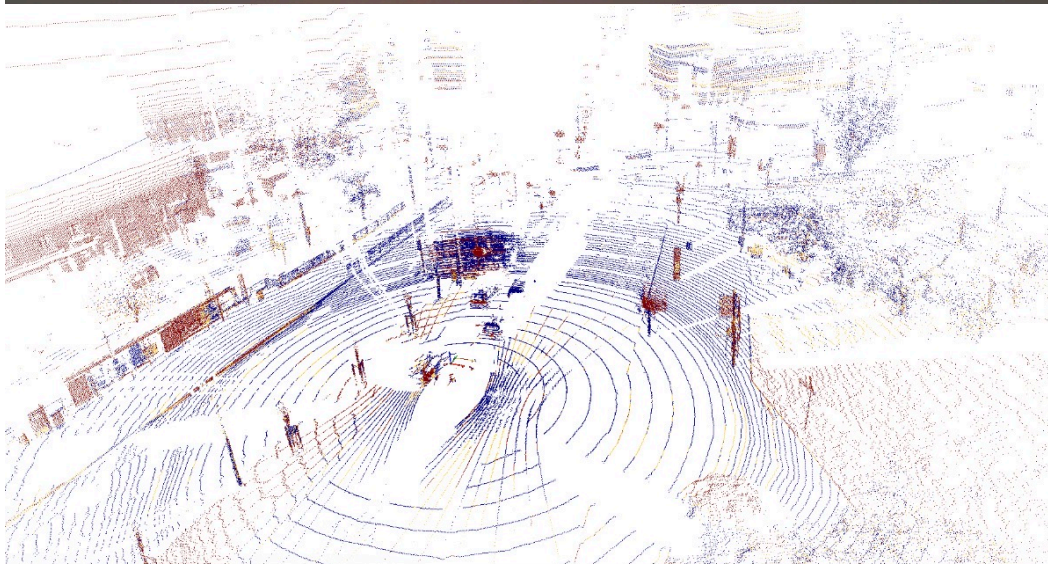
- LiDAR semantic segmentation
- LiDAR panoptic segmentation
- 3D object detection
- 4D panoptic segmentation

Why LiDAR sensors?

- Accurate depth sensing
- Robust at low-light conditions
- 3D positional information
- …

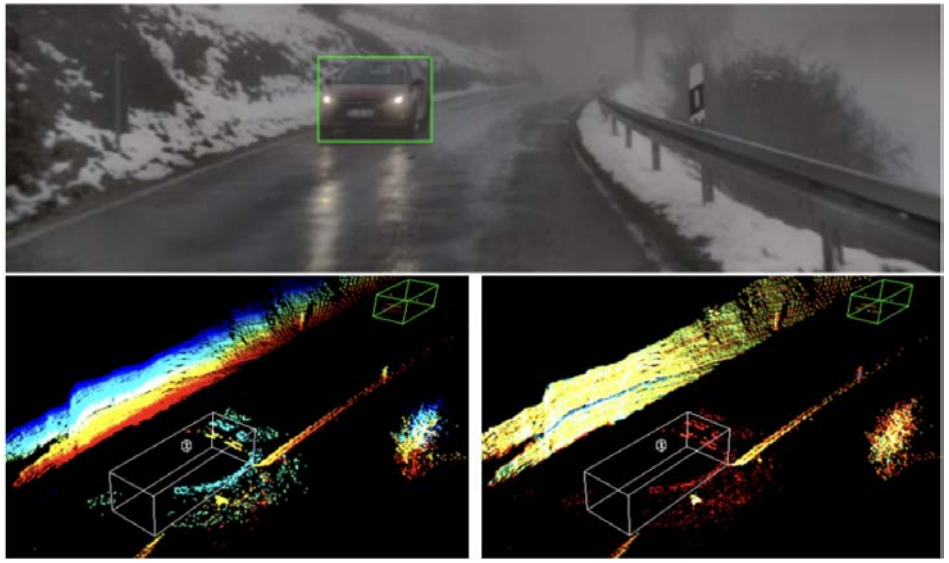
M. Aygun, et al. "4D panoptic LiDAR segmentation," CVPR, 2021.

2D Perception vs. 3D Perception

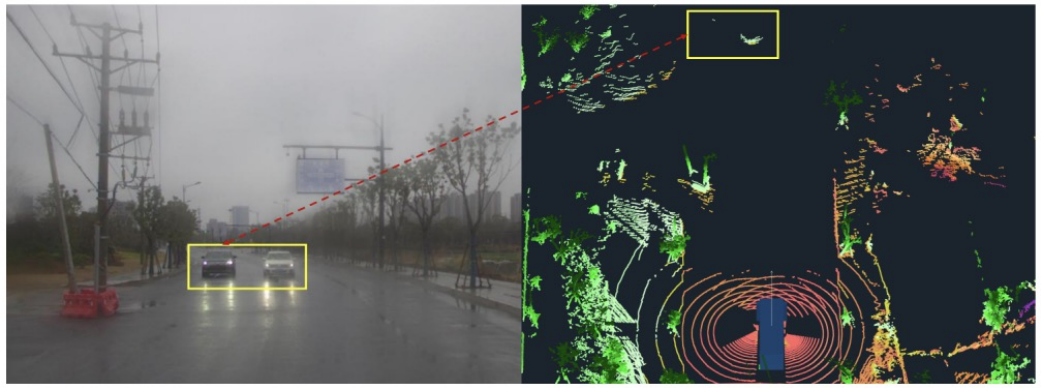


Source: https://zod.zenseact.com

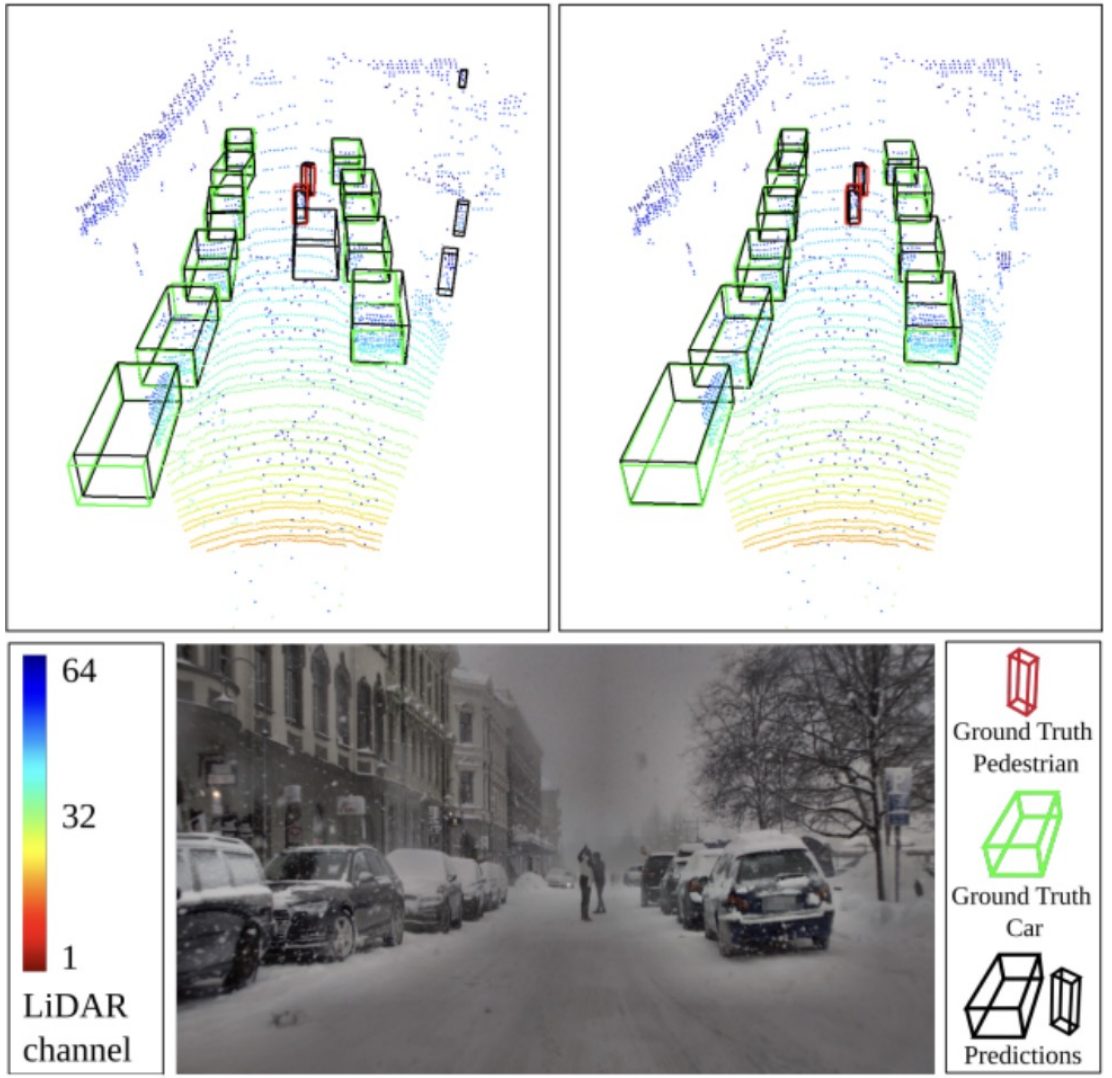
2D Perception vs. 3D Perception



Hahner, et al. Fog. [ICCV 2021]

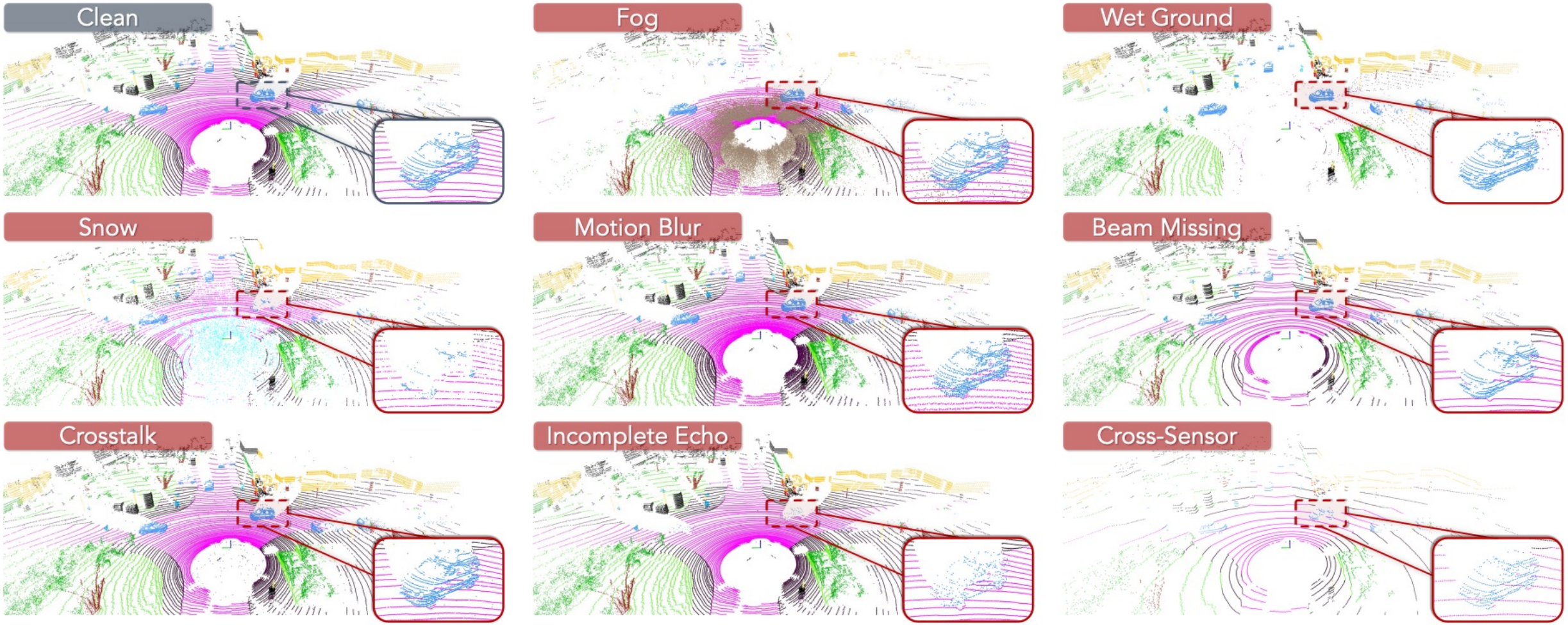


Yu, et al. Incomplete Echo. [arXiv 2022]



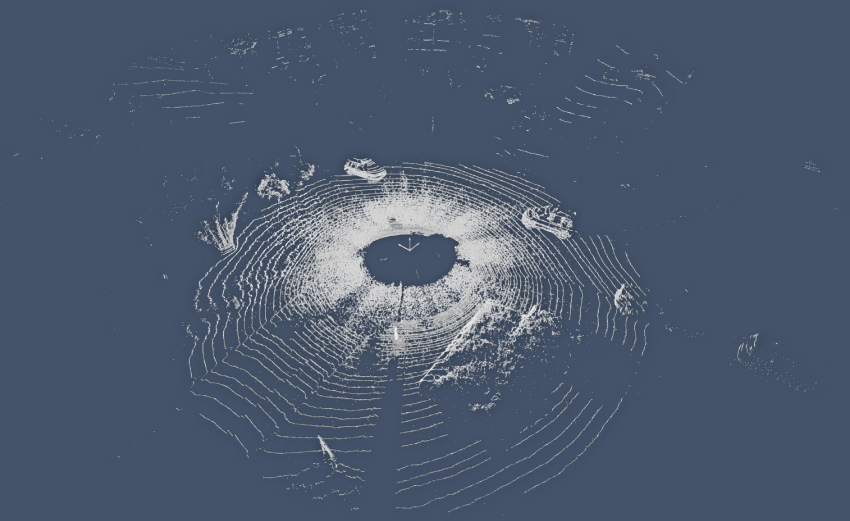
Hahner, et al. Snow. [CVPR 2022]

Robust 3D Perception



L. Kong, et al. "Robo3D: Towards robust and reliable 3D perception against corruptions," ICCV, 2023.

Robo3D



TL;DR

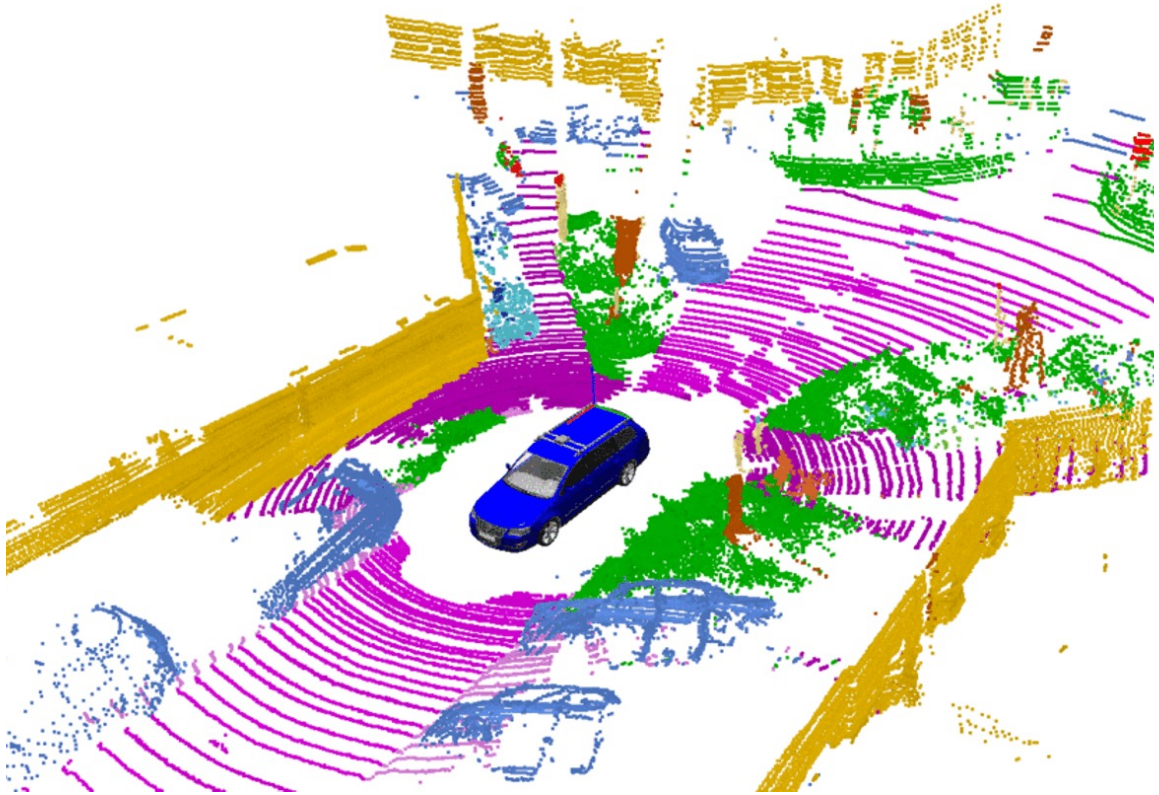
- We introduce **Robo3D**, the first systematically-designed robustness evaluation suite for LiDAR-based 3D perception under corruptions and sensor failure
- We benchmark 34 perception models for LiDAR-based **semantic segmentation** and **object detection** tasks, on their robustness against corruptions.
- Based on our observations, we draw in-depth discussions on the receipt of designing **robust** and **reliable** 3D perception models.

Point Cloud Segmentation

Indoor

Dai, et al. ScanNet V2. [CVPR 2017]

Outdoor



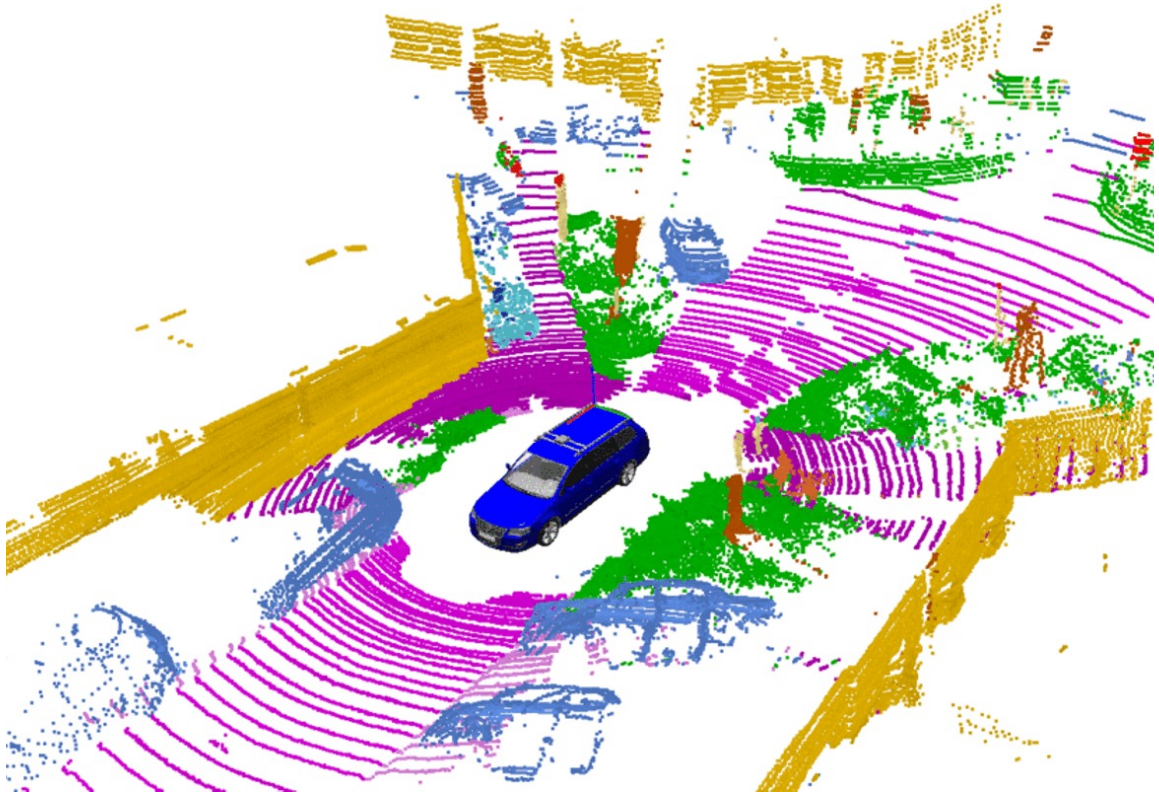
Behley, et al. SemanticKITTI. [ICCV 2019]

Point Cloud Segmentation

Main Features

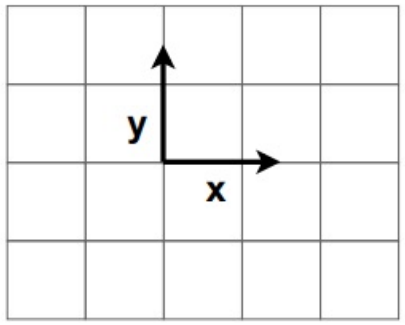
- Autonomous driving perception
- Larger in scale (>120k points per scan)
- Sparser, more diverse
- Rich semantic categories
- Different sensor configurations (32, 64, 128)
- Real-world distribution
- Standard benchmarks: nuScenes, SemanticKITTI, Waymo Open, etc.

Outdoor

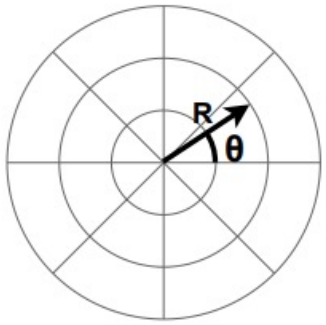


Behley, et al. SemanticKITTI. [ICCV 2019]

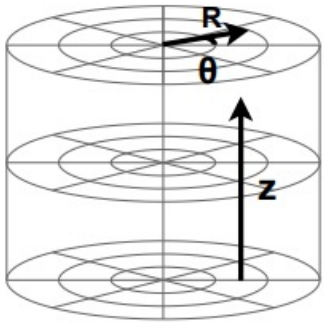
Point Cloud Representation



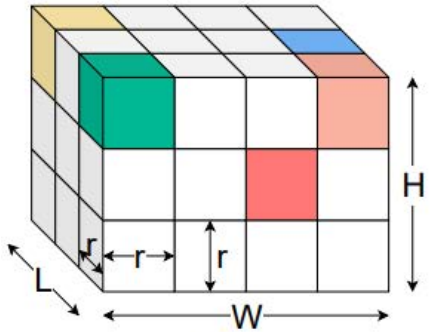
Cartesian Coordinates



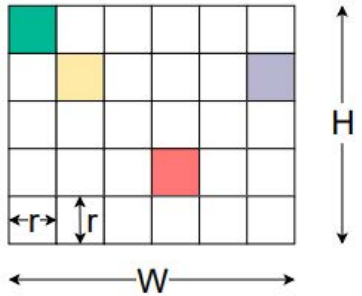
Polar Coordinates



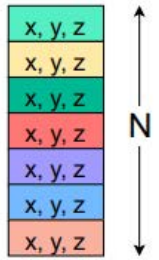
Cylinder Coordinates



3D Voxels



2D Projection



1D Bag-of-Points

Representation:

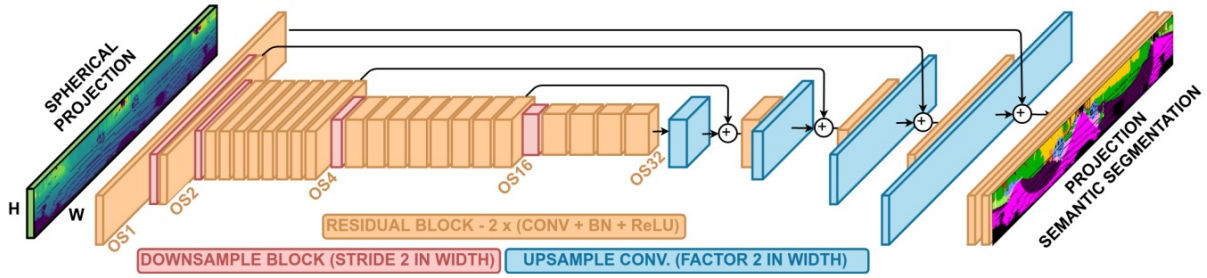
- **2D:** range view, bird's eye view
- **3D:** cubic voxel, cylinder voxel

Operator:

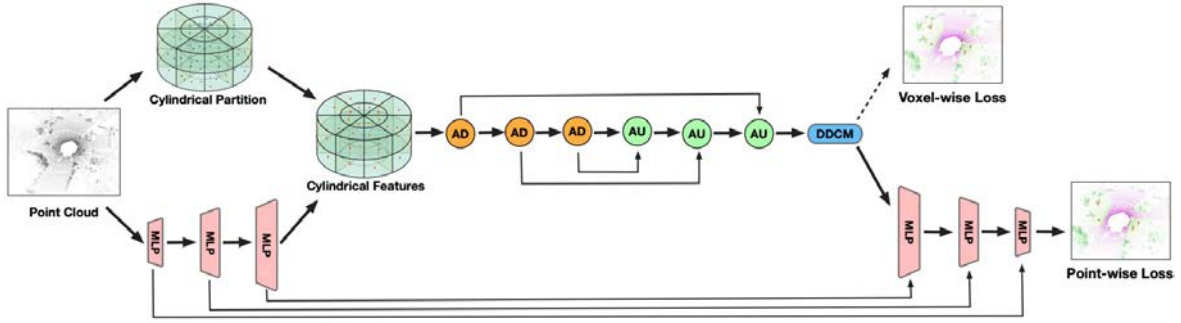
- **3D:** Conv3d, SparseConv, etc.
- **2D:** Conv2d, Linear, etc.
- **1D:** Conv1d, Linear, etc.

M. Uecker, et al. "Analyzing deep learning representations of point clouds for real-time in-vehicle LiDAR perception," arXiv, 2022.

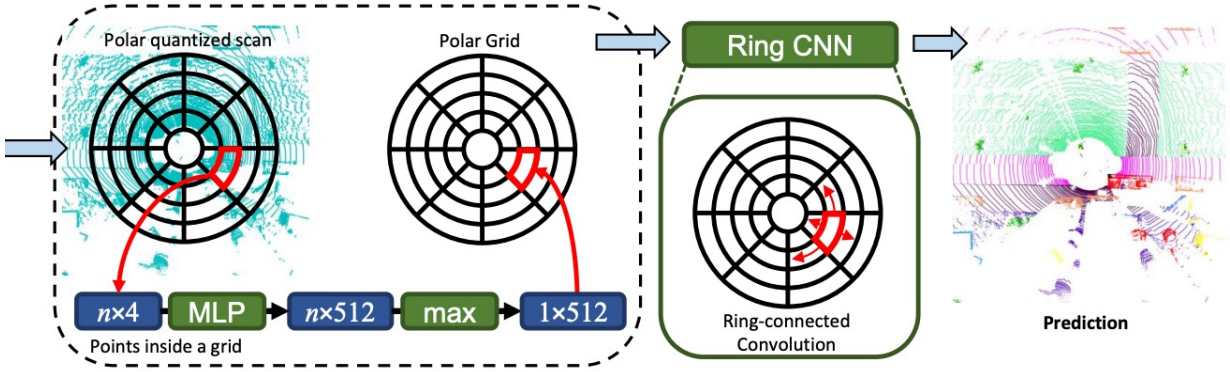
Segmentation Model



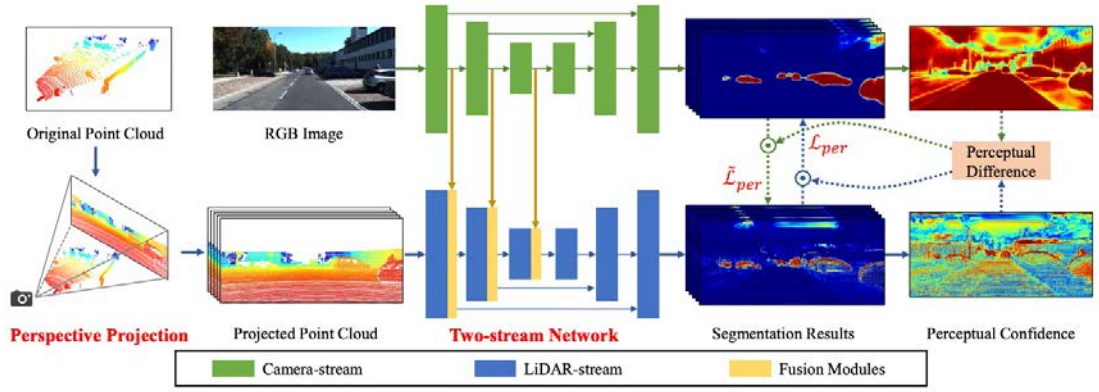
Milioto, et al. RangeNet++. [IROS 2019]



Zhu, et al. Cylinder3D. [CVPR 2021]



Zhang, et al. PolarNet. [CVPR 2020]



Zhang, et al. PMF. [ICCV 2021]

Topic

1. Automotive Point Cloud Segmentation
   自动驾驶场景中的点云分割
2. Point Cloud Segmentation with MMDetection3D
   高效点云分割工具箱
3. Introduction to Segment Anything
   简介「分割一切」模型
4. Seal: Segmentation Any Point Cloud Sequences
   运用视觉基础模型「分割一切点云序列」

MMDetection3D

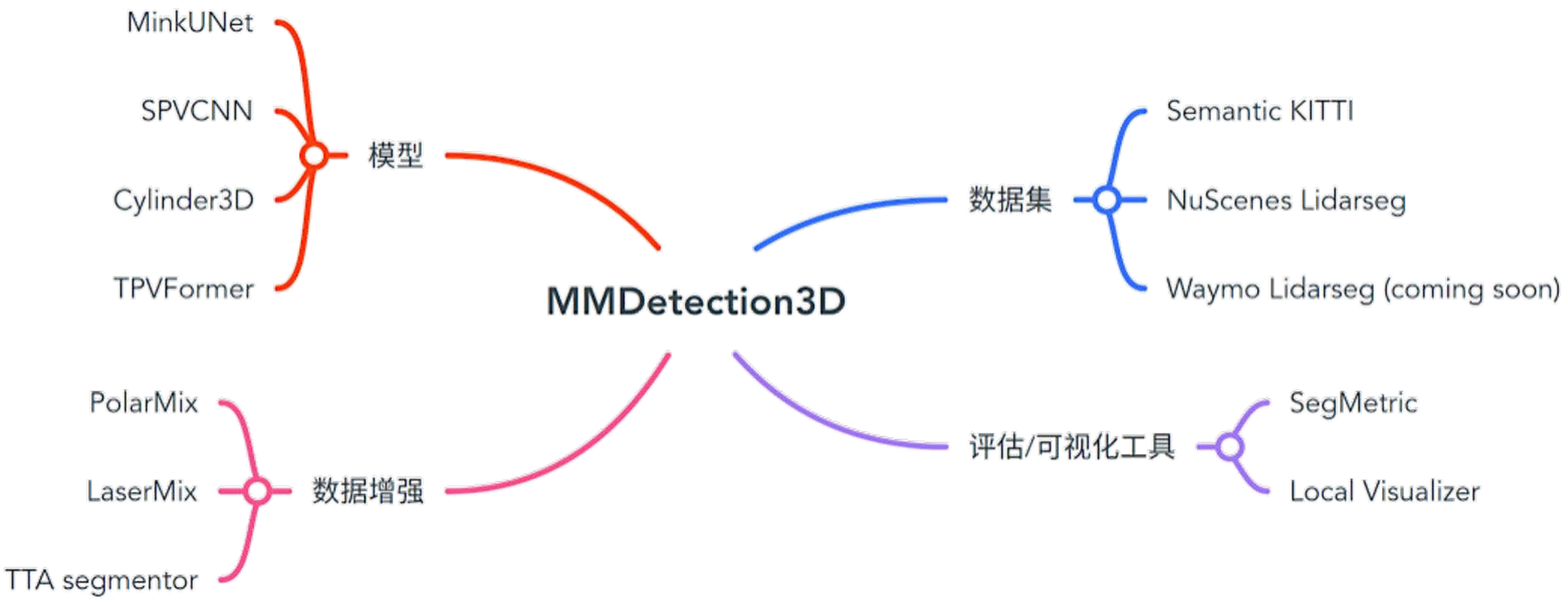
- MMDetection3D is an open-source toolbox based on PyTorch, towards the next-generation platform for general 3D detection and **[NEW] 3D Semantic Segmentation**.

3D Object Detection	Monocular 3D Object Detection	Multi-modal 3D Object Detection	3D Semantic Segmentation

• **Outdoor**
  • SECOND (Sensor'2018)
  • PointPillars (CVPR'2019)
  • SSN (ECCV'2020)
  • 3DSSD (CVPR'2020)
  • SA-SSD (CVPR'2020)
  • PointRCNN (CVPR'2019)
  • Part-A2 (TPAMI'2020)
	• **Outdoor**
	  • ImVoxelNet (WACV'2022)
	  • SMOKE (CVPRW'2020)
	  • FCOS3D (ICCVW'2021)
	  • PGD (CoRL'2021)
	  • MonoFlex (CVPR'2021)
	• **Indoor**
	  • ImVoxelNet (WACV'2022)
		• **Outdoor**
		  • MVXNet (ICRA'2019)
		• **Indoor**
		  • ImVoteNet (CVPR'2020)
			• **Outdoor**
			  • MinkUNet (CVPR'2019)
			  • SPVCNN (ECCV'2020)
			  • Cylinder3D (CVPR'2021)
			• **Indoor**
			  • PointNet++ (NeurIPS'2017)
			  • PAConv (CVPR'2021)
			  • DGCNN (TOG'2019)

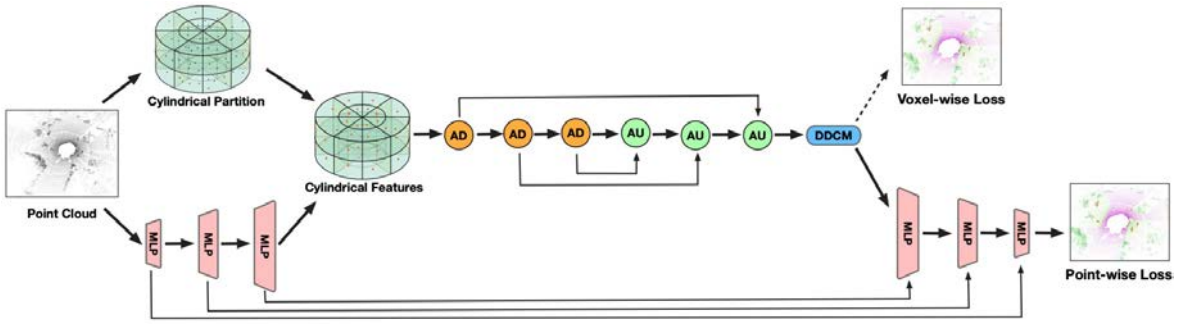
The MMDetection3D Codebase: https://github.com/open-mmlab/mmdetection3d

MMDetection3D v1.1.1

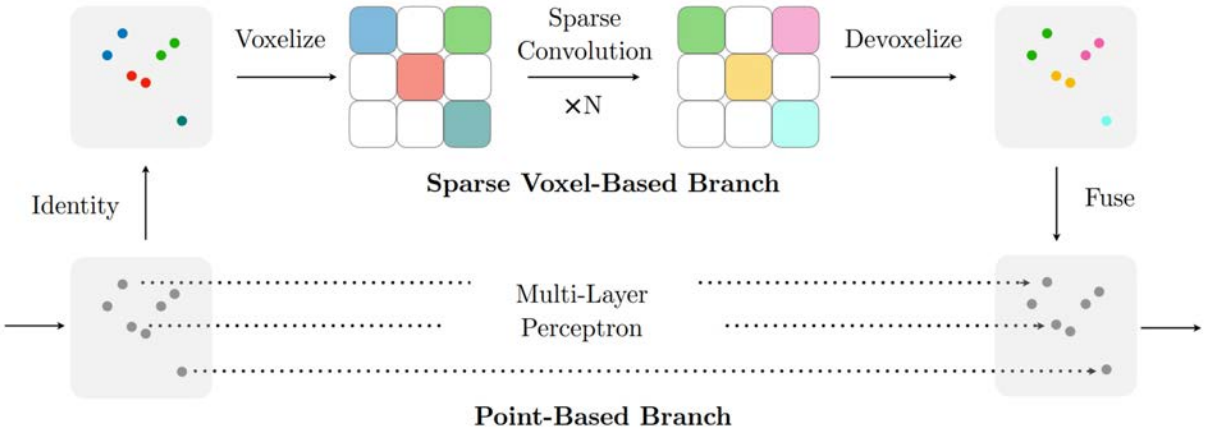


The MMDetection3D Codebase: https://github.com/open-mmlab/mmdetection3d

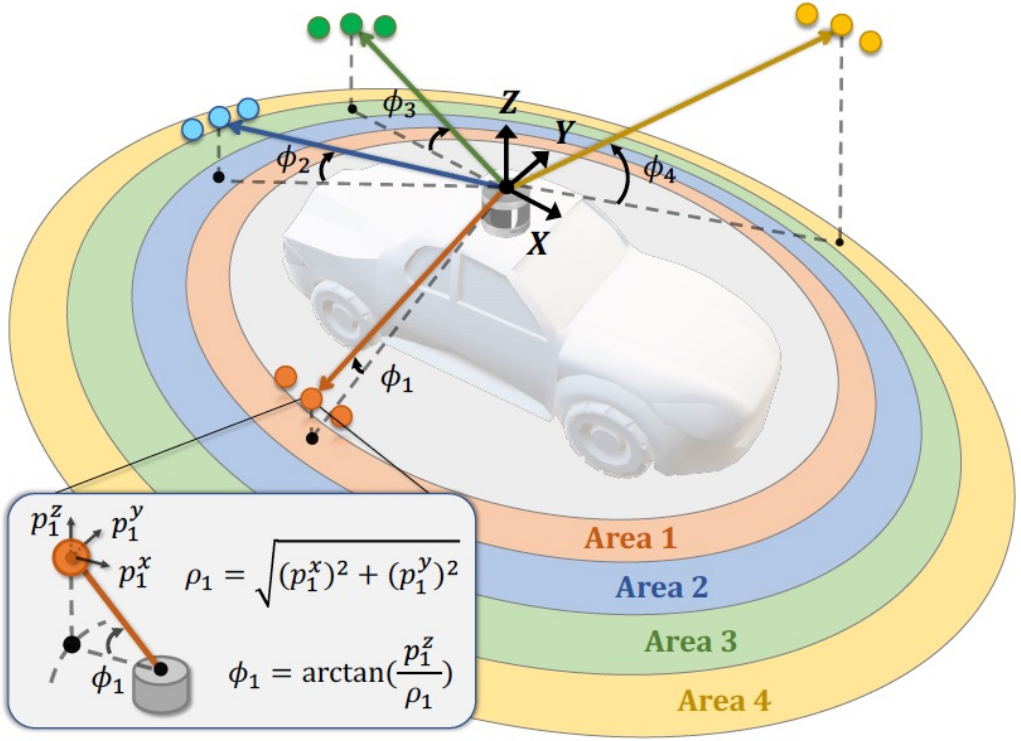
MMDetection3D v1.1.1



Zhu, et al. Cylinder3D. [CVPR 2021]



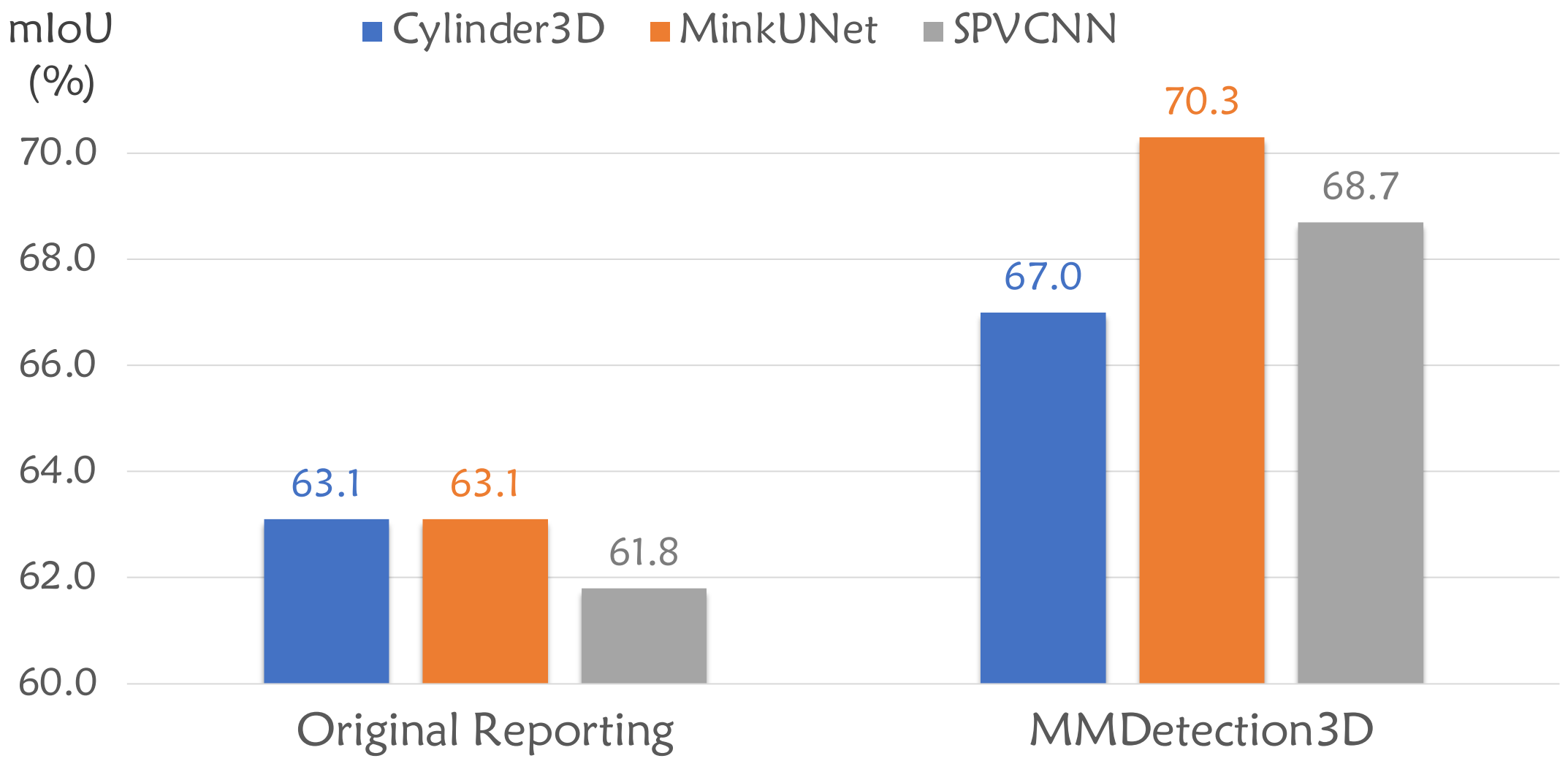
Tang, et al. SPVCNN. [ECCV 2020]



Kong, et al. LaserMix. [CVPR 2023]

The MMDetection3D Codebase: https://github.com/open-mmlab/mmdetection3d

MMDetection3D v1.1.1



The MMDetection3D Codebase: https://github.com/open-mmlab/mmdetection3d

MMDetection3D v1.1.1

Method	Backend	Lr schd	Amp	Laser-Polar Mix	Mem (GB)	Training Time (hours)	FPS	mIoU	Download
MinkUNet18-W16	torchsparse	15e	✓	×	3.4	-	-	60.3	model | log
MinkUNet18-W20	torchsparse	15e	✓	×	3.7	-	-	61.6	model | log
MinkUNet18-W32	torchsparse	15e	✓	×	4.9	-	-	63.1	model | log
MinkUNet34-W32	minkowski engine	3x	×	✓	11.5	6.5	12.2	69.2	model | log
MinkUNet34-W32	spconv	3x	✓	✓	6.7	2	14.6*	68.3	model | log
MinkUNet34-W32	spconv	3x	×	✓	10.5	6	14.5	69.3	model | log
MinkUNet34-W32	torchsparse	3x	✓	✓	6.6	3	12.8	69.3	model | log
MinkUNet34-W32	torchsparse	3x	×	✓	11.8	5.5	15.9	68.7	model | log
MinkUNet34v2-W32	torchsparse	3x	✓	✓	8.9	-	-	70.3	model | log

Major Features

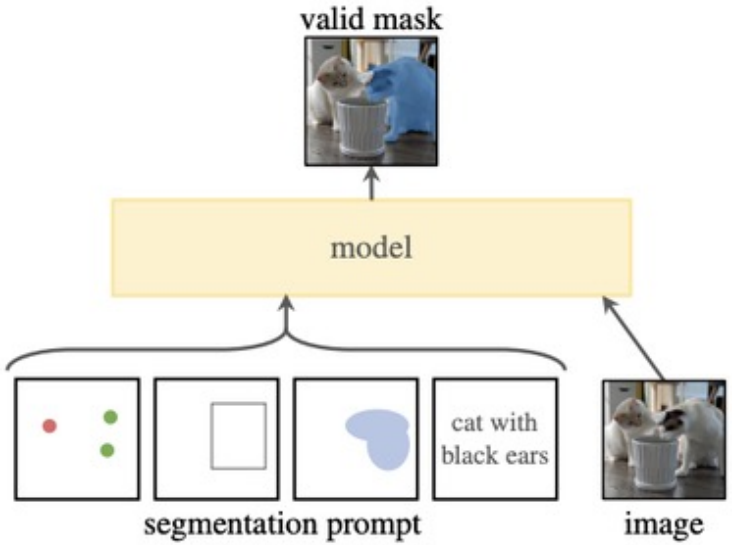
- Supports different sparse convolution backends
- Supports most recent 3D augmentation techniques
- Achieves SoTA 3D semantic segmentation performance

The MMDetection3D Codebase: https://github.com/open-mmlab/mmdetection3d

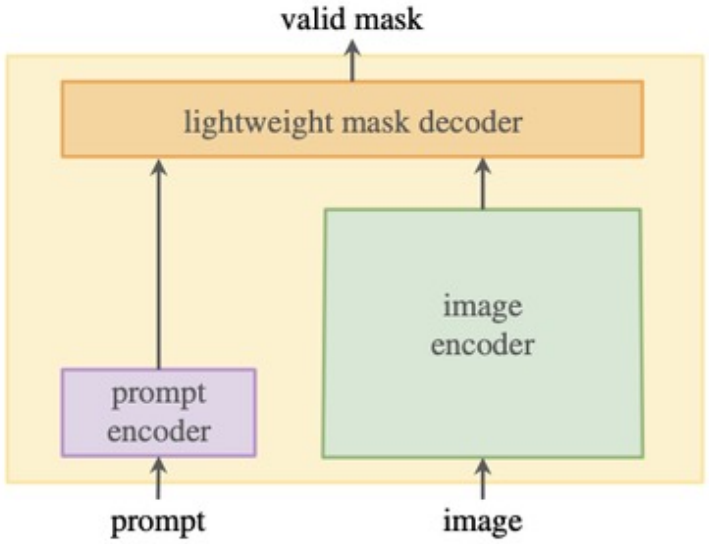
Topic

1. Automotive Point Cloud Segmentation
   自动驾驶场景中的点云分割
2. Point Cloud Segmentation with MMDetection3D
   高效点云分割工具箱
3. Introduction to Segment Anything
   简介「分割一切」模型
4. Seal: Segmentation Any Point Cloud Sequences
   运用视觉基础模型分割「任意」点云序列

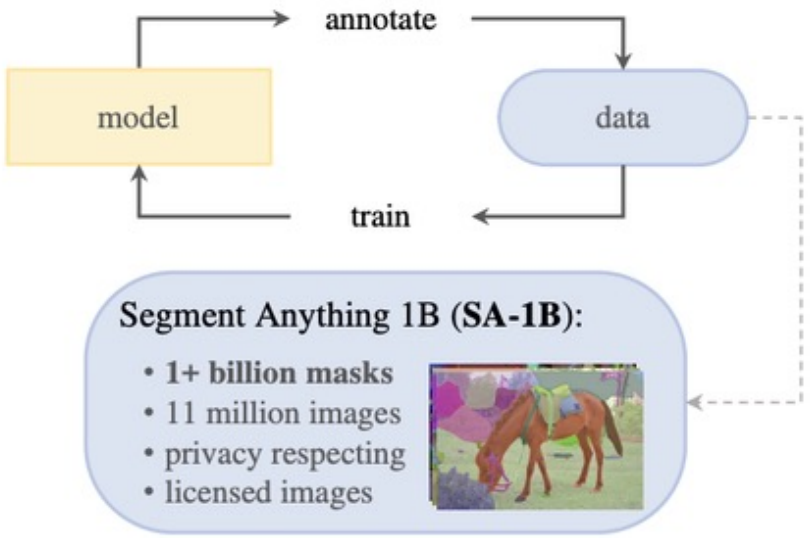
Segment Anything



(a) **Task**: promptable segmentation



(b) **Model**: Segment Anything Model (**SAM**)

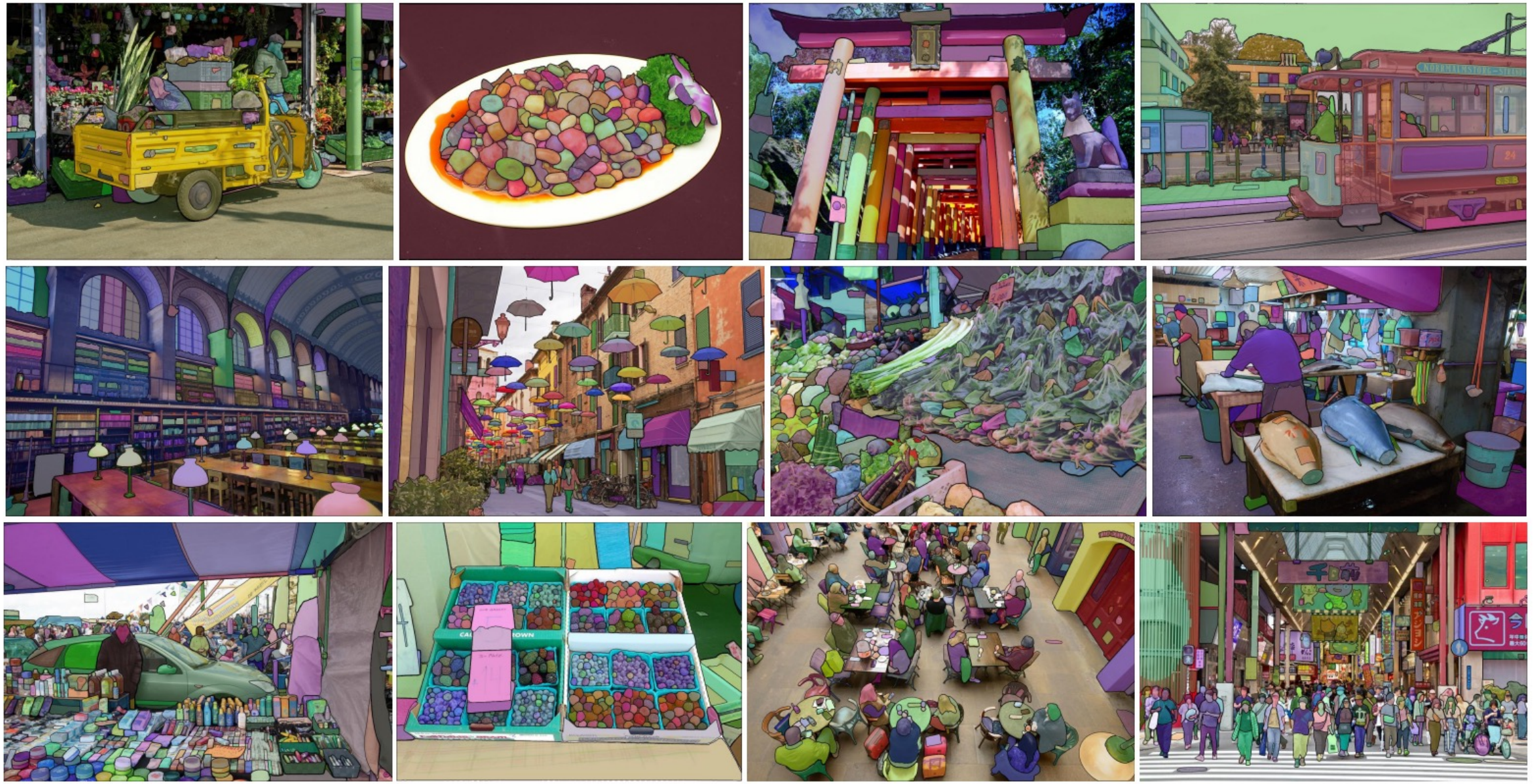


(c) **Data**: data engine (top) & dataset (bottom)

- **Training Set:** SA-1B, over 1B masks on 11M images
- **Model:** ViT-H SAM model
- **Demo:** https://segment-anything.com/demo

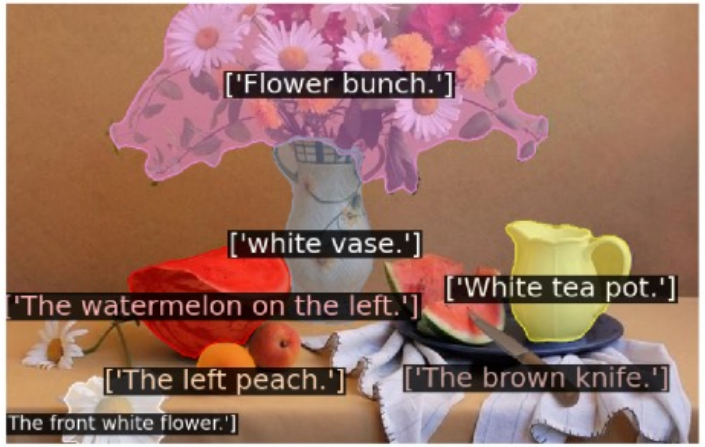
A. Kirillov, et al. "Segment Anything," arXiv, 2023.

Segment Anything



A. Kirillov, et al. "Segment Anything," arXiv, 2023.

X-Decoder



Referring Segmentation

Open-Vocabulary Semantic Segmentation

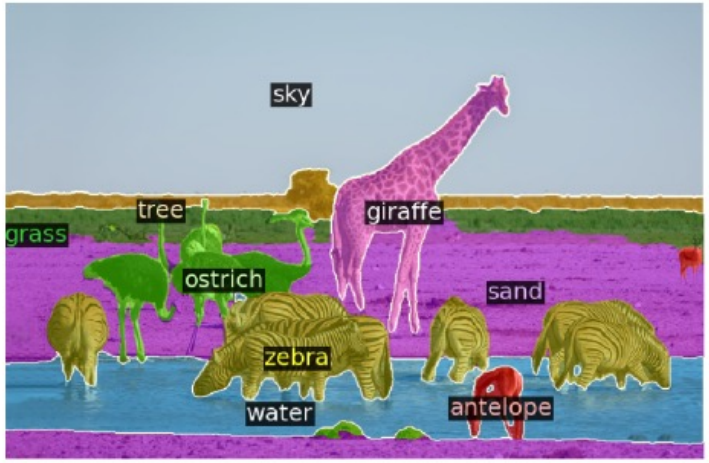


Image Captioning

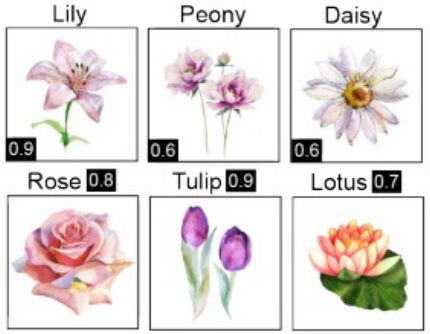


Image-Text Retrieval

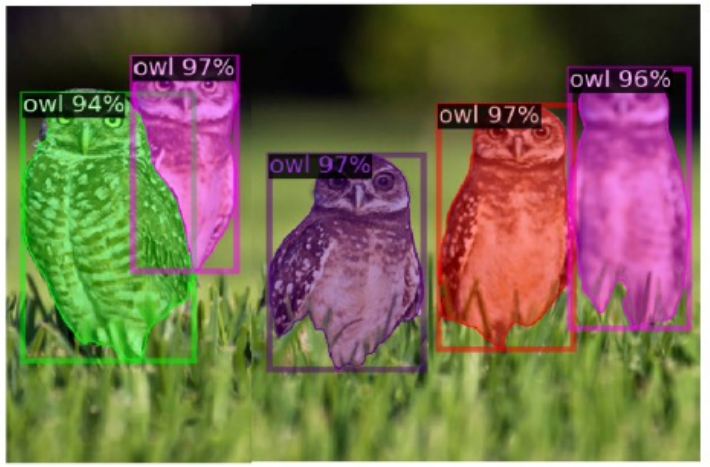
X-Decoder

Referring Captioning

Referring Image Editing

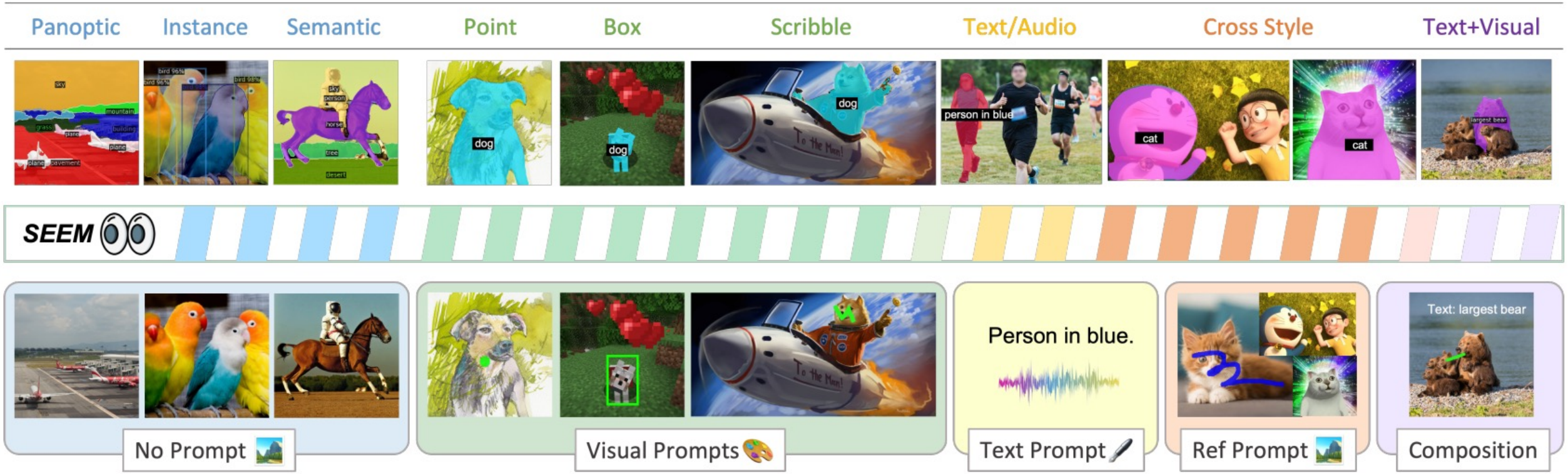
Open-Vocabulary Panoptic Segmentation

Open-Vocabulary Instance Segmentation



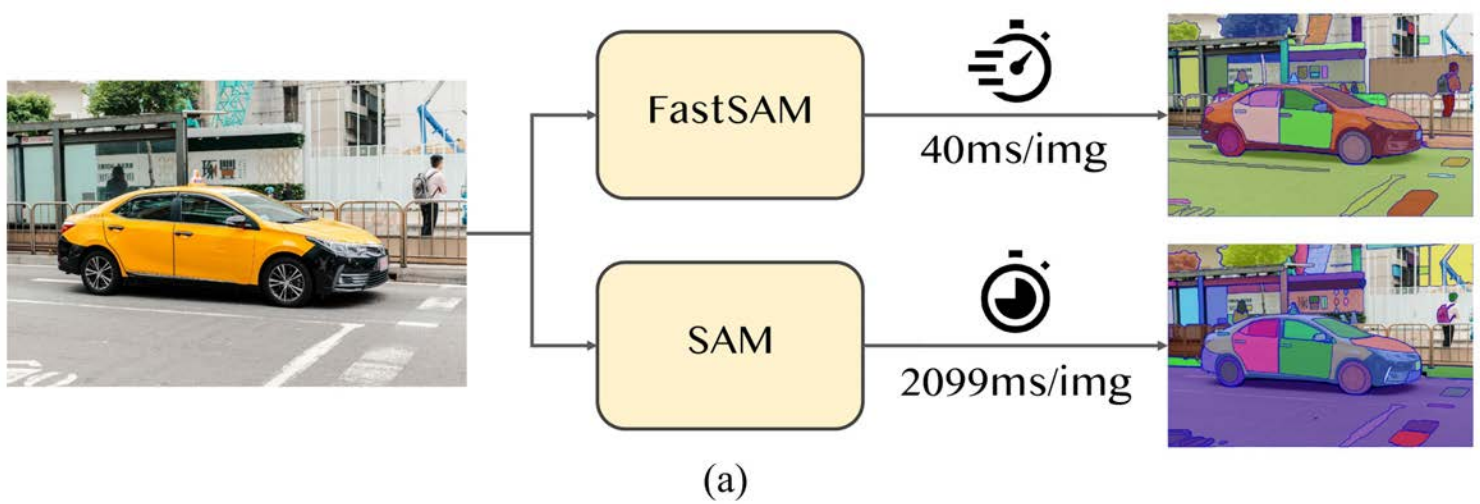
X. Zou, et al. "Generalized decoding for pixel, image, and language," CVPR, 2023.

SEEM

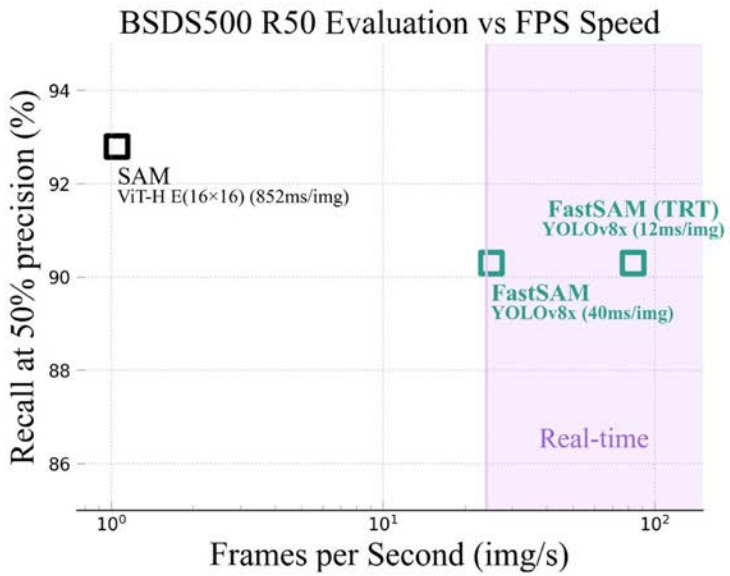


X. Zou, et al. "Segment everything everywhere all at once," arXiv, 2023.

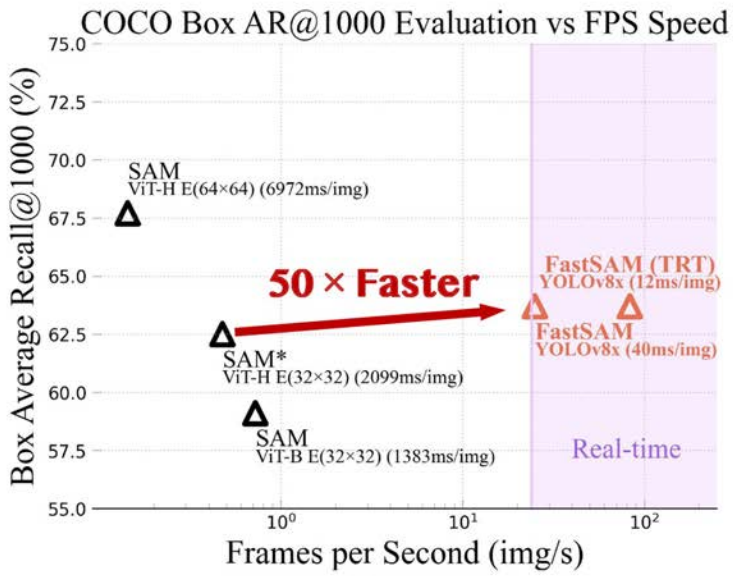
Fast-SAM



(a)



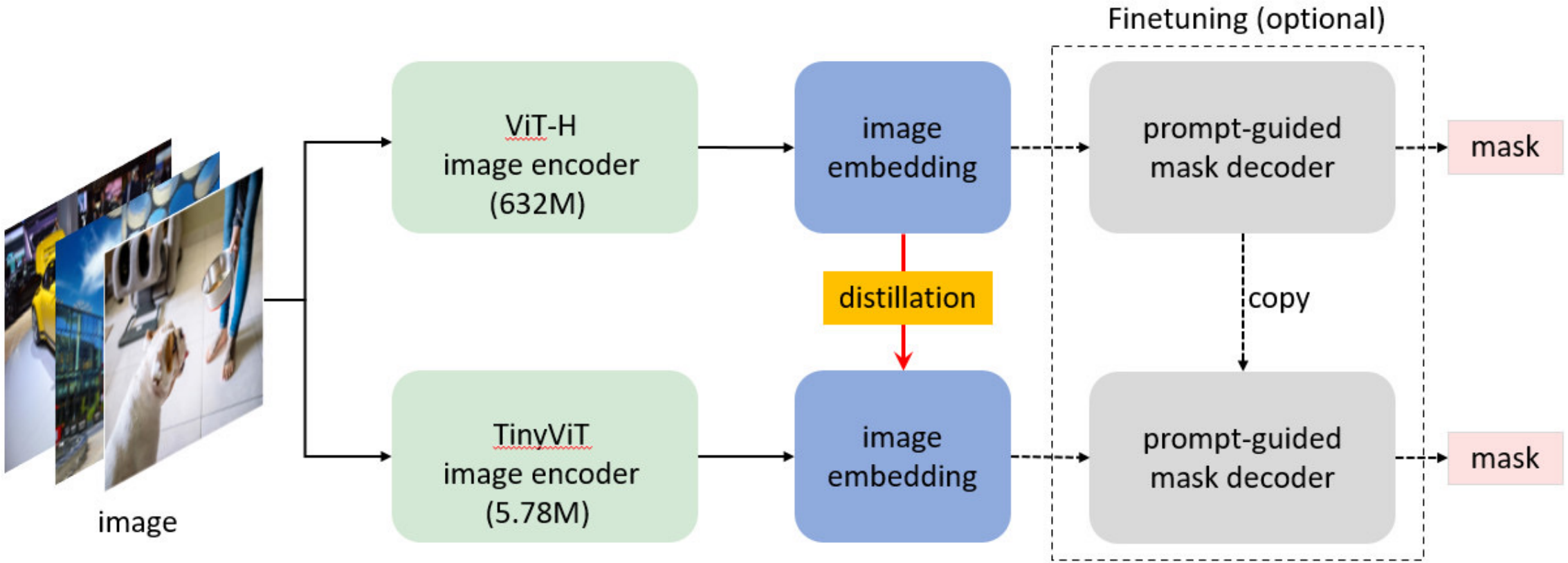
(b)



(c)

X. Zhao. "Fast segment anything," arXiv, 2023.

Faster Segment Anything



X. Zhao. "Faster segment anything: Towards lightweight SAM for mobile applications," arXiv, 2023.

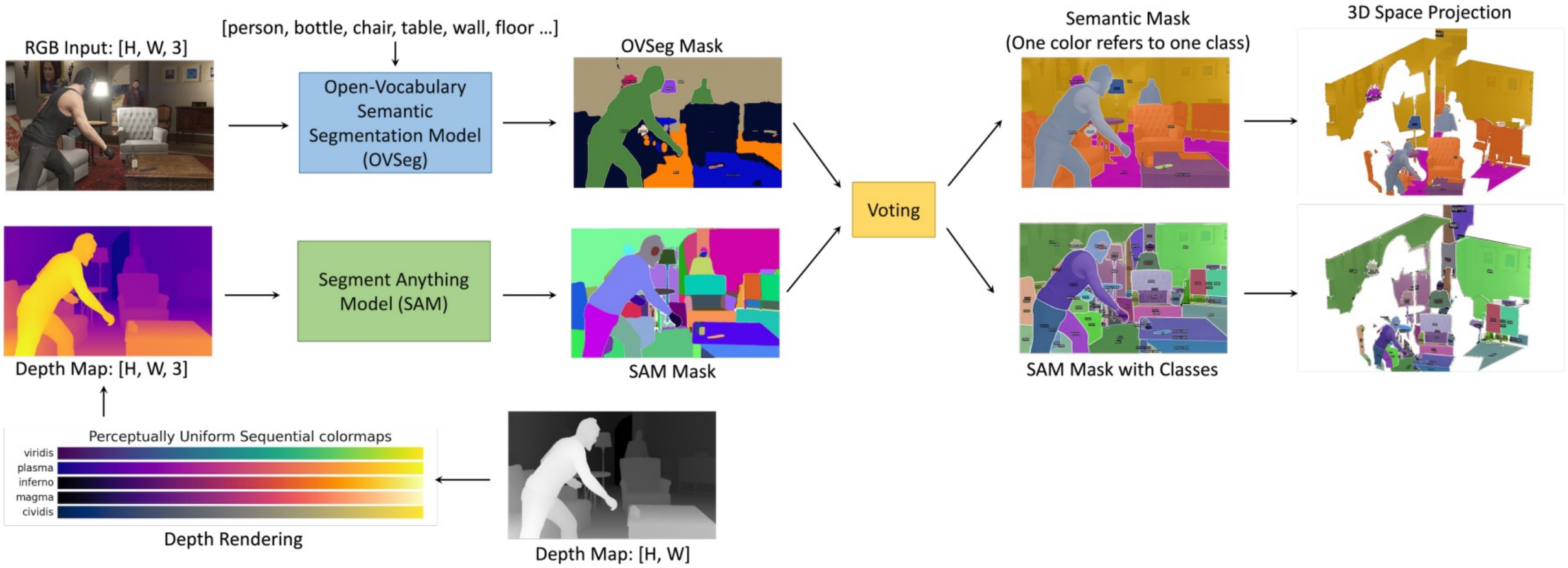
Semantic-SAM

F. Li. "Semantic-SAM: Segment and recognize anything at any granularity," arXiv, 2023.

DINO v2

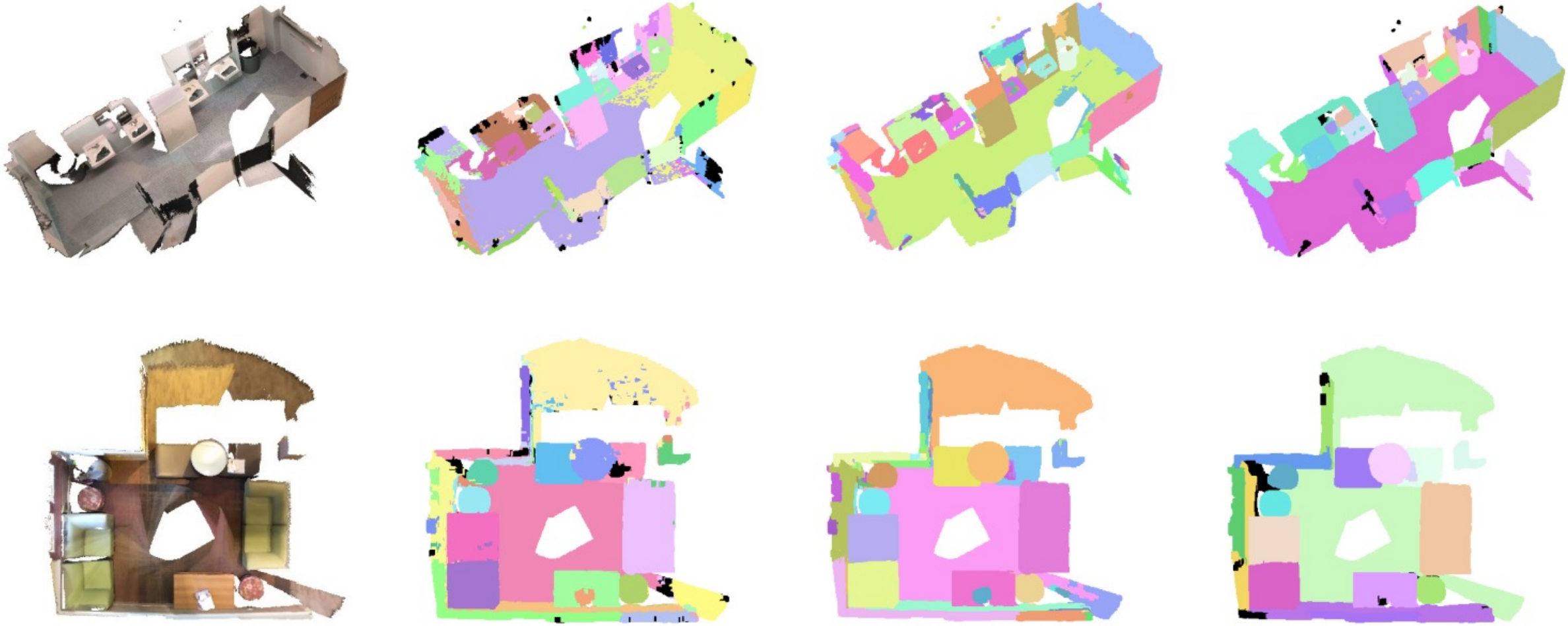
M. Oquab, et al. "DINOv2: Learning robust visual features without supervision," arXiv, 2023.

Segment Any RGB-D



J. Cen, et al. "SAD: Segment any RGBD," arXiv, 2023.

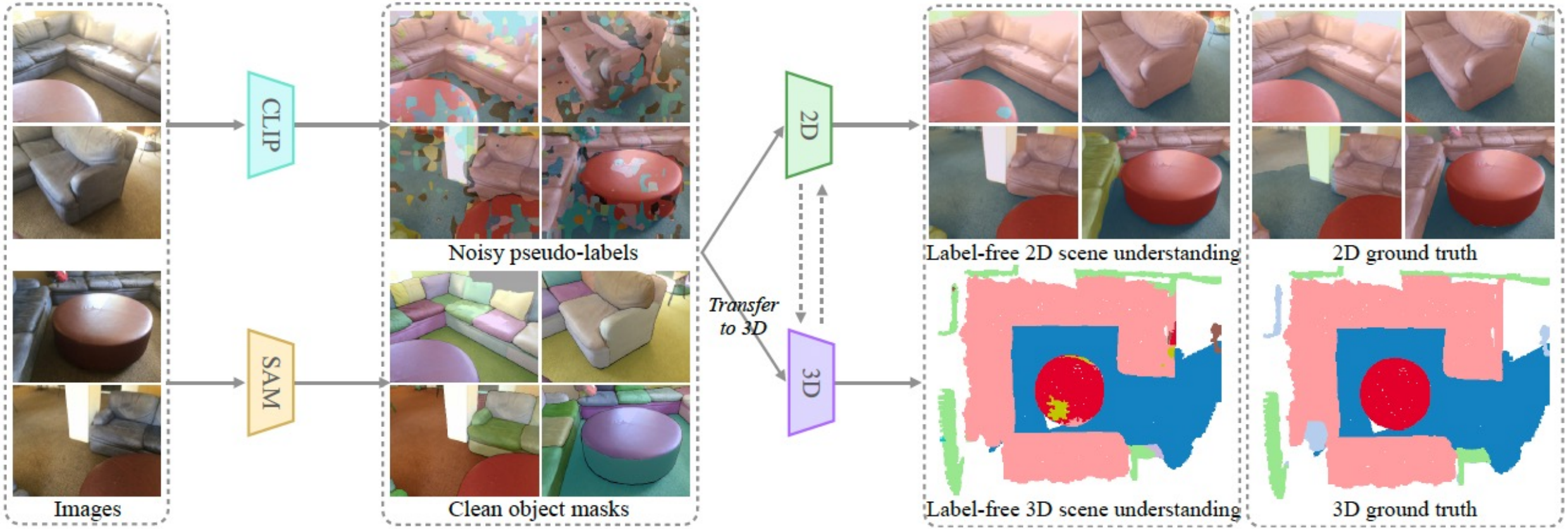
SAM3D



3D Scene | SAM3D | SAM3D (ensembled) | Ground-truth

Y. Yang, et al. "SAM3D: Segment anything in 3D scenes," arXiv, 2023.

CNS: Cross-Modality Noisy Supervision



R. Chen, et al. "Towards label-free scene understanding by vision foundation models," arXiv, 2023.

Topic

1. Automotive Point Cloud Segmentation
   自动驾驶场景中的点云分割
2. Point Cloud Segmentation with MMDetection3D
   高效点云分割工具箱
3. Introduction to Segment Anything
   简介「分割一切」模型
4. **Seal: Segmentation Any Point Cloud Sequences**
   运用视觉基础模型分割「任意」点云序列

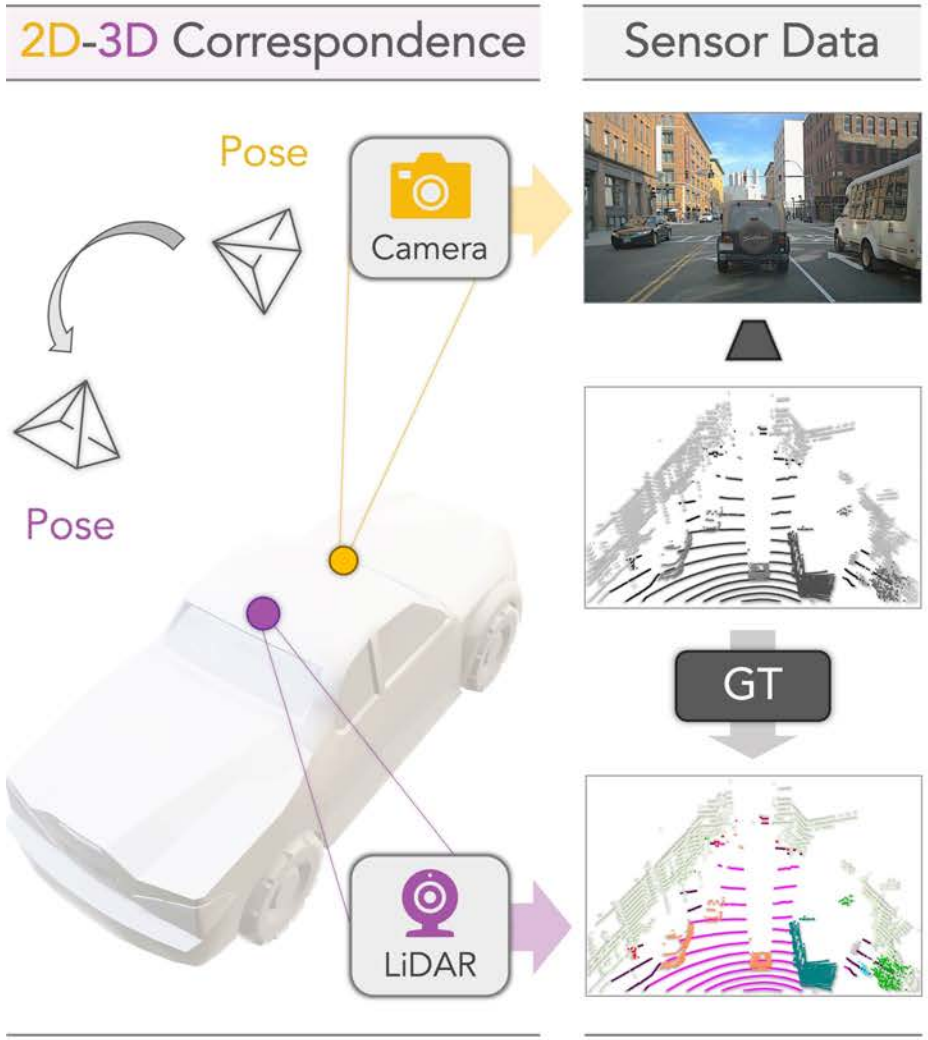
Seal

segment any point cloud

TL;DR

- **Seal** is a versatile self-supervised learning framework capable of segmenting any automotive point clouds.
- **Seal** leverages off-the-shelf knowledge from vision foundation models (VFMs) and encouraging spatial and temporal consistency from such knowledge during the representation learning stage.
- **Seal** enables knowledge transfer in an off-the-shelf manner to downstream tasks involving diverse point clouds, including those from real/synthetic, low/high-resolution, large/small-scale, and clean/corrupted datasets.

Seal: Challenge & Motivation



Challenge

- Camera views ≠ LiDAR views
- Automotive point clouds have unique label mappings
- LiDAR and cameras are not perfectly synchronized

We hope to have a framework that …

- Conducts self-supervised learning on automotive point clouds
- Enforces spatial and temporal consistency during representation learning
- Can be generalizable to diverse downstream tasks

Superpixel

What is Superpixel

- Grouped pixels of perceptually meaningful atomic regions, which can be used to replace the rigid structure of the pixel grid.

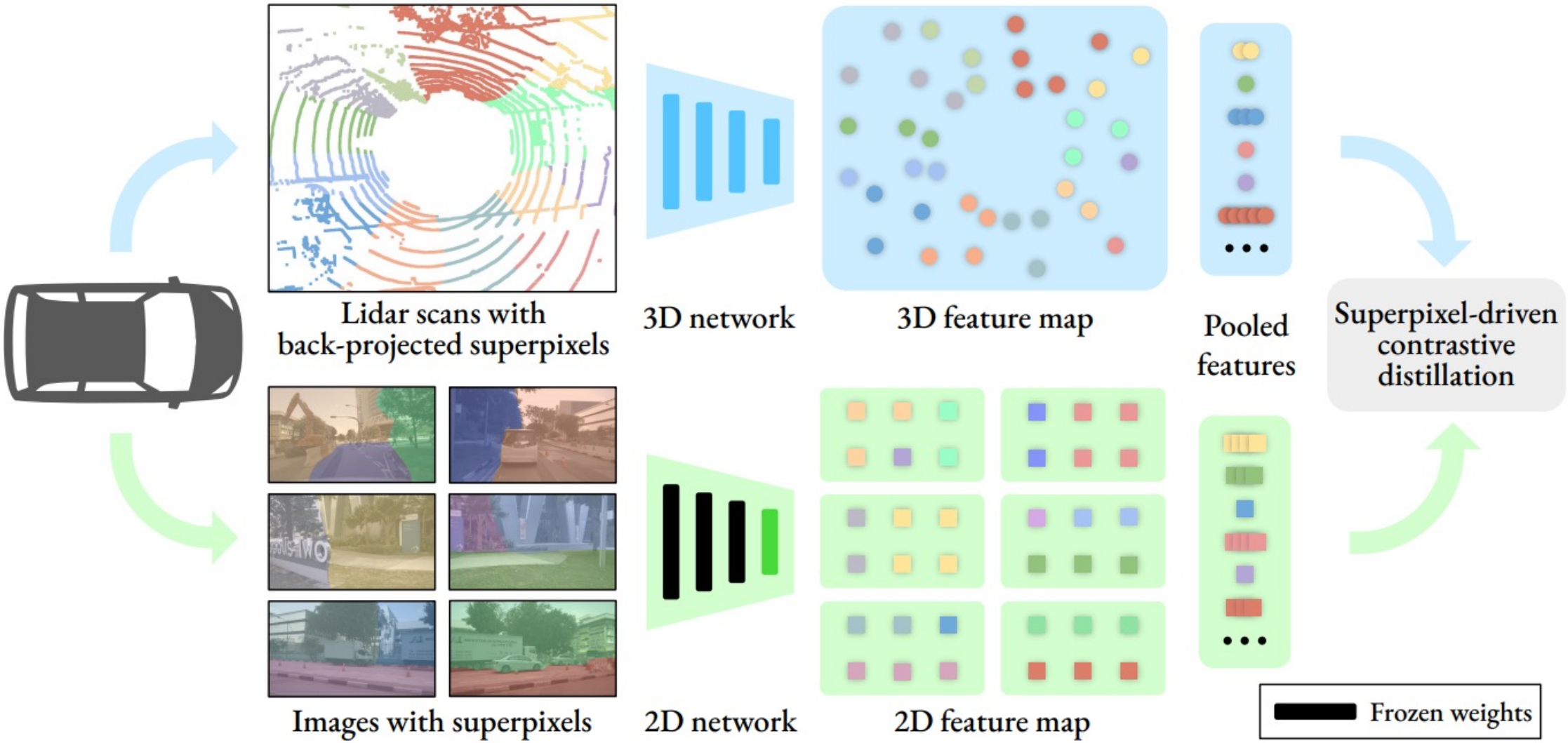
Key Features of Superpixel

- Captures image redundancy;
- Provides a convenient primitive from which to compute image features;
- Greatly reduces the complexity of subsequent image processing tasks.

Fig. 1: Images segmented using SLIC into superpixels of size 64, 256, and 1024 pixels (approximately).

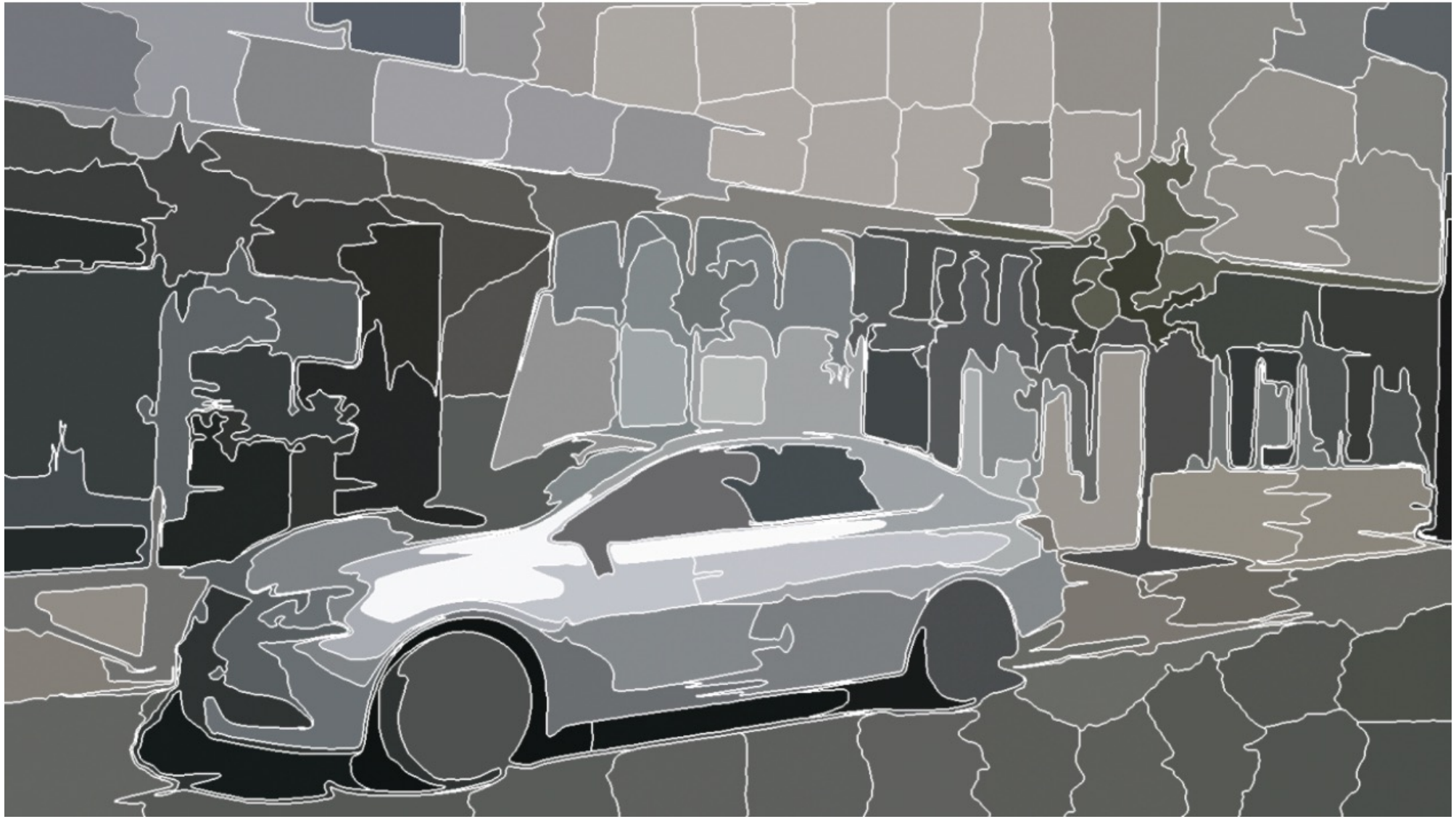
R. Achanta, et al. "SLIC superpixels compared to state-of-the-art superpixel methods," TPAMI, 2012.

SLidR: Superpixel-Driven LiDAR Representation



C. Sautier, et al. "Image-to-LiDAR self-supervised distillation for autonomous driving data," CVPR, 2022.

SLidR: Superpixel-Driven LiDAR Representation



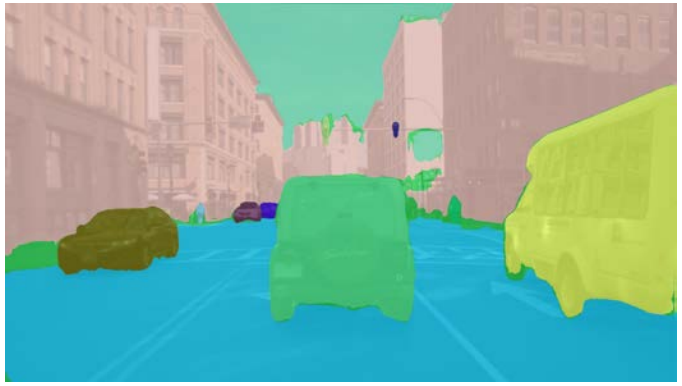
C. Sautier, et al. "Image-to-LiDAR self-supervised distillation for autonomous driving data," CVPR, 2022.

Semantic Superpixel

Camera View
(Front)

Superpixel
(SLIC)

Semantic Superpixel
(SAM)



Semantic Superpixel
(X-Decoder)

Semantic Superpixel
(OpenSeeD)

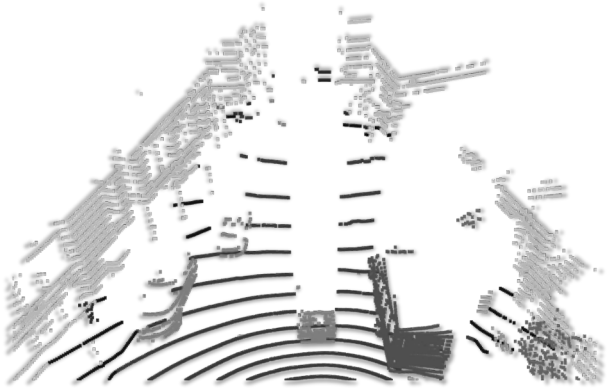
Semantic Superpixel
(SEEM)

Superpixel to Superpoint

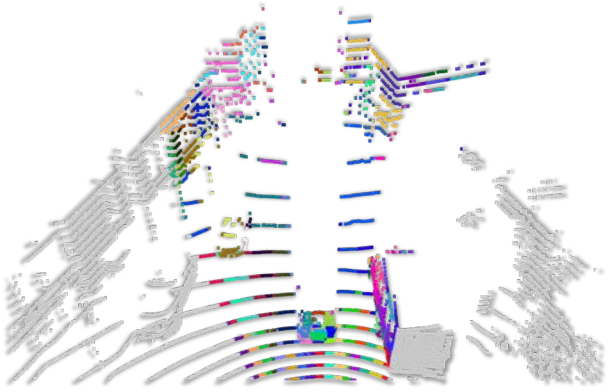
Camera View
(Front)

Superpixel
(SLIC)

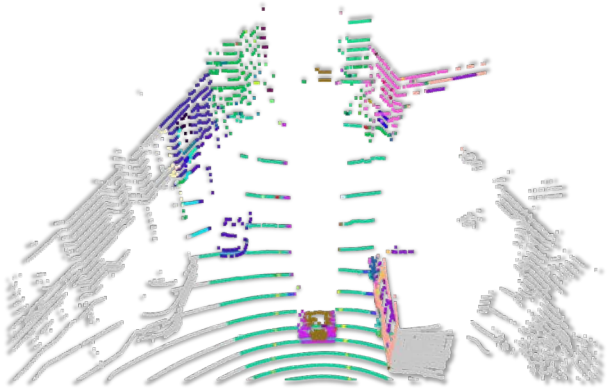
Semantic Superpixel
(SAM)



LiDAR View
(Front)

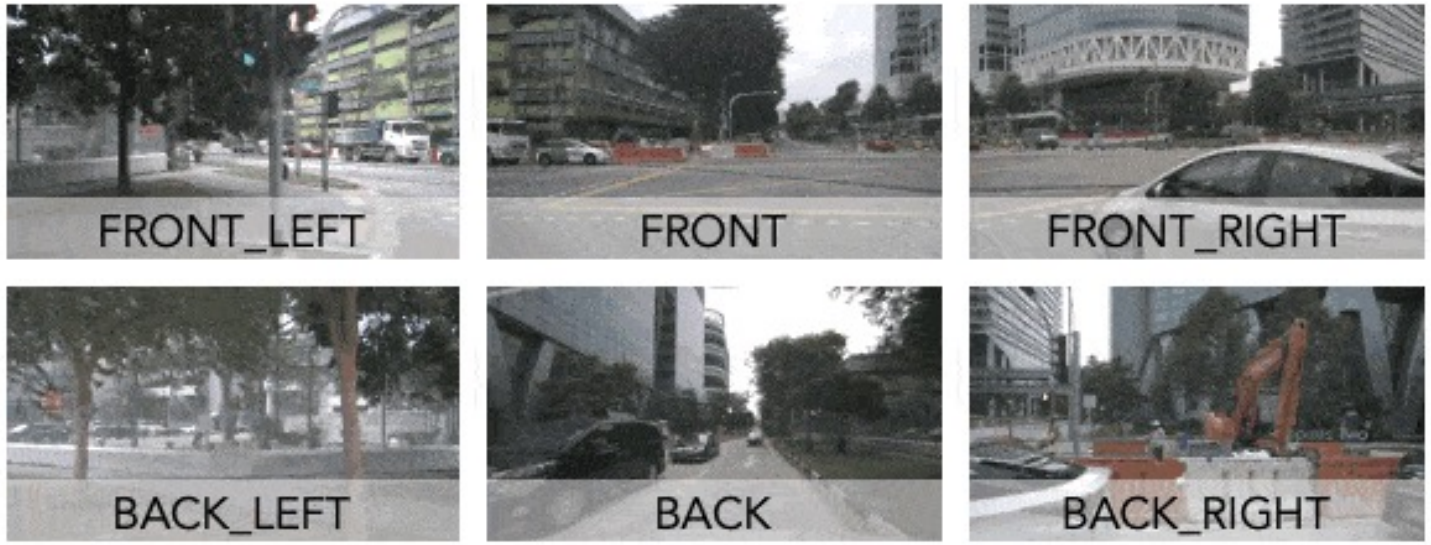


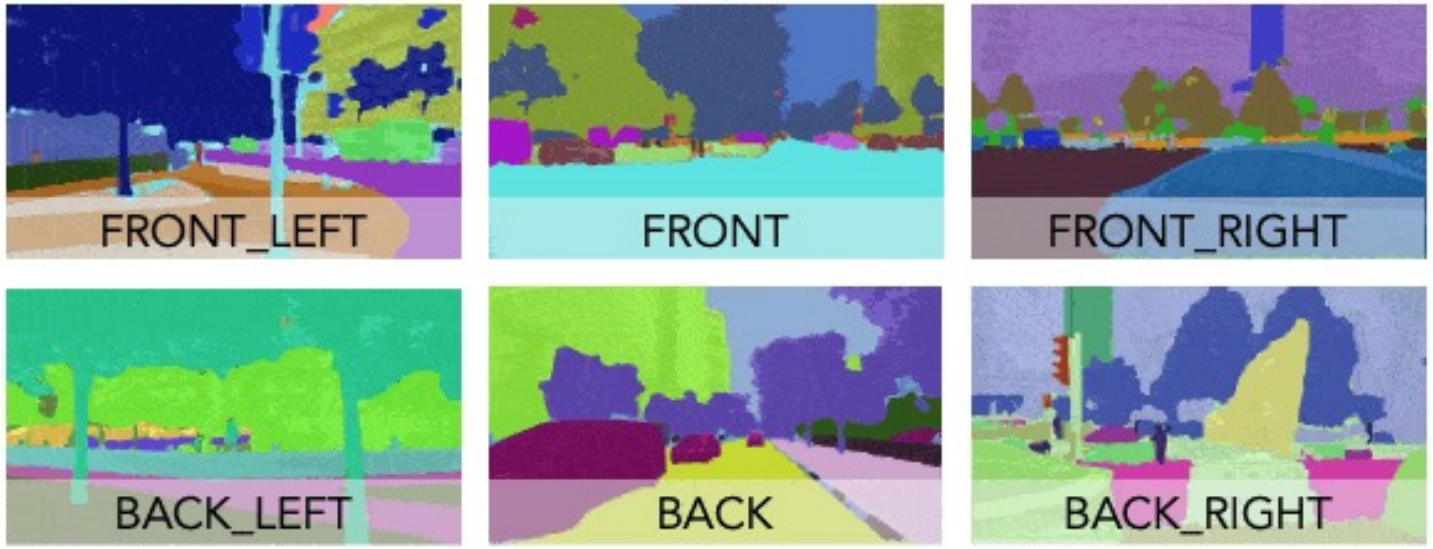
Superpoint
(SLIC)



Semantic Superpoint
(SAM)

Superpixel to Superpoint

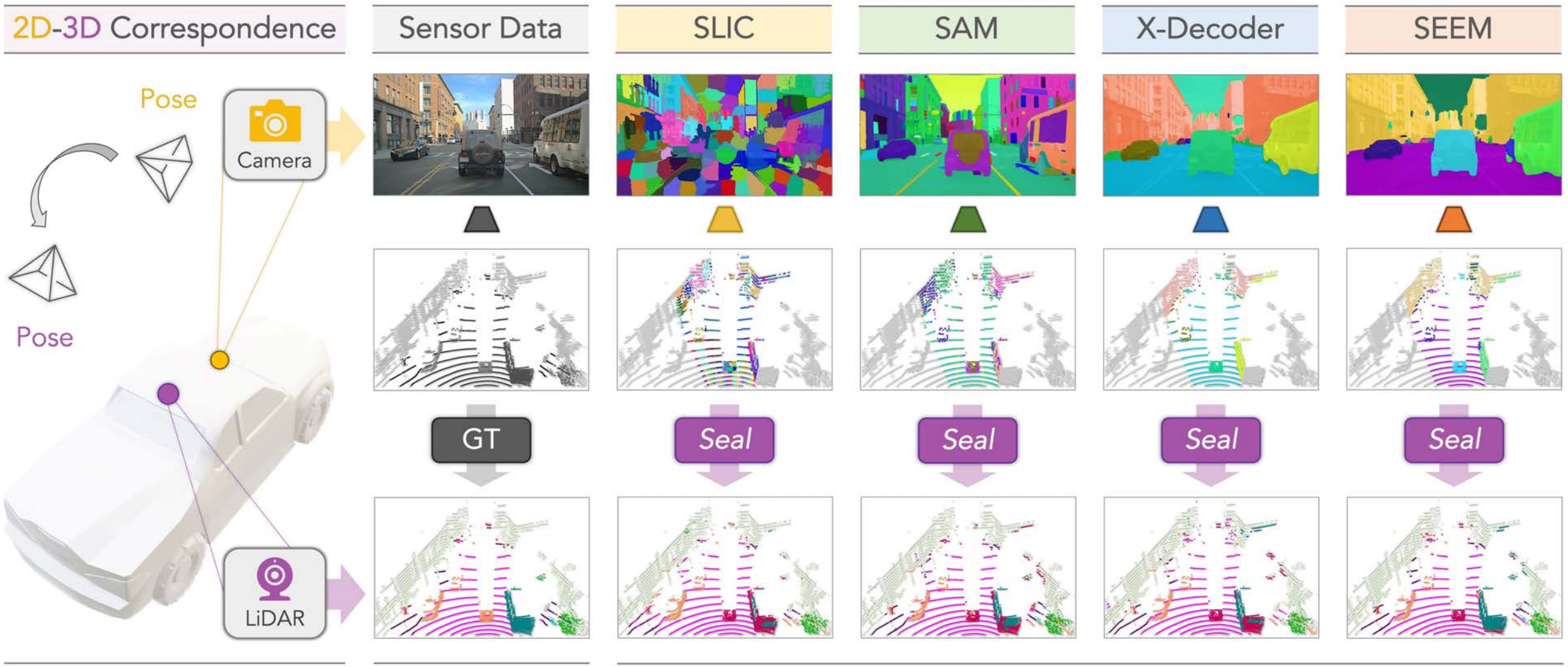




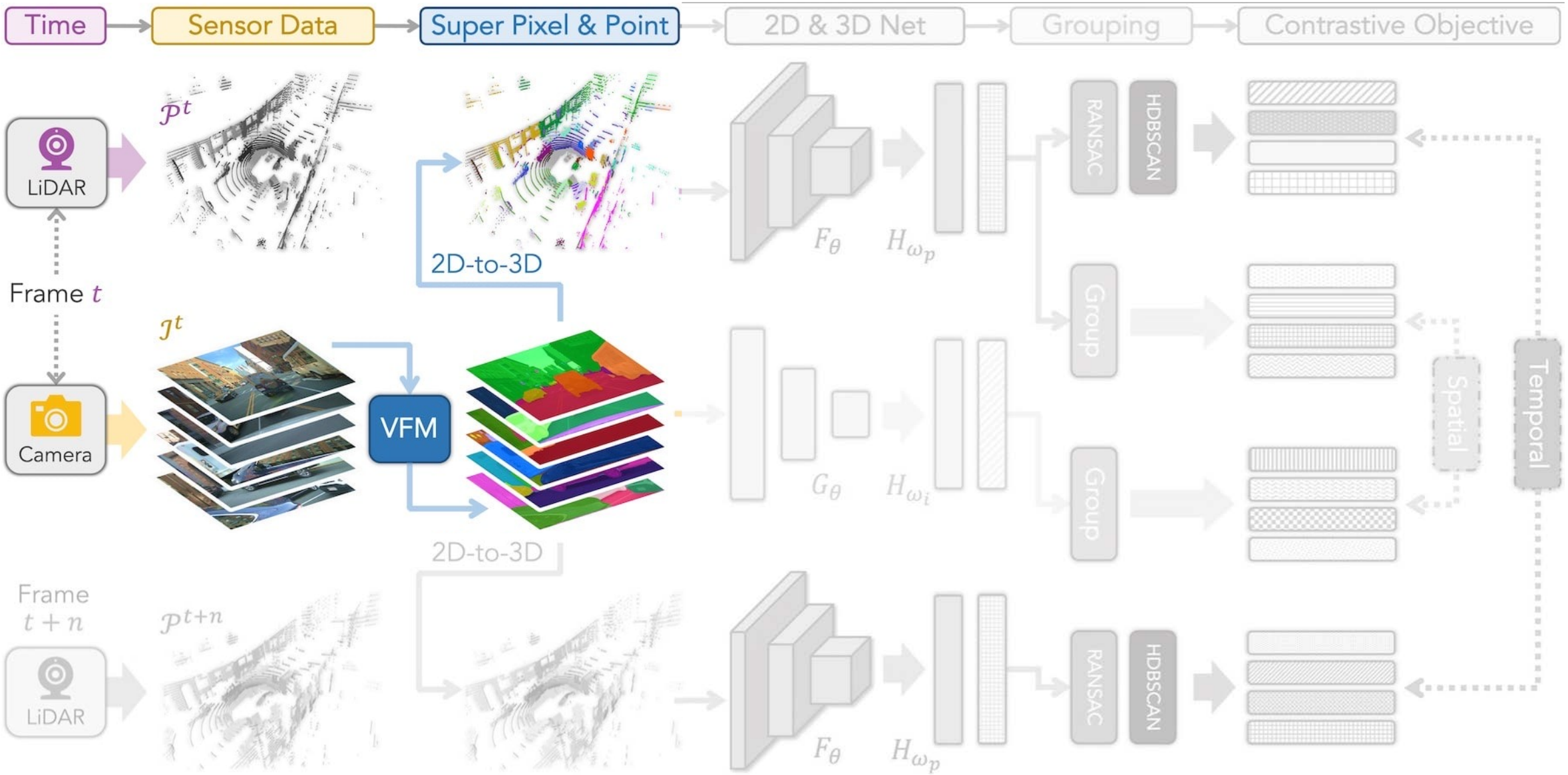
LIDAR (RAW UNLABELED POINT)

LIDAR (SEMANTIC SUPERPOINT)

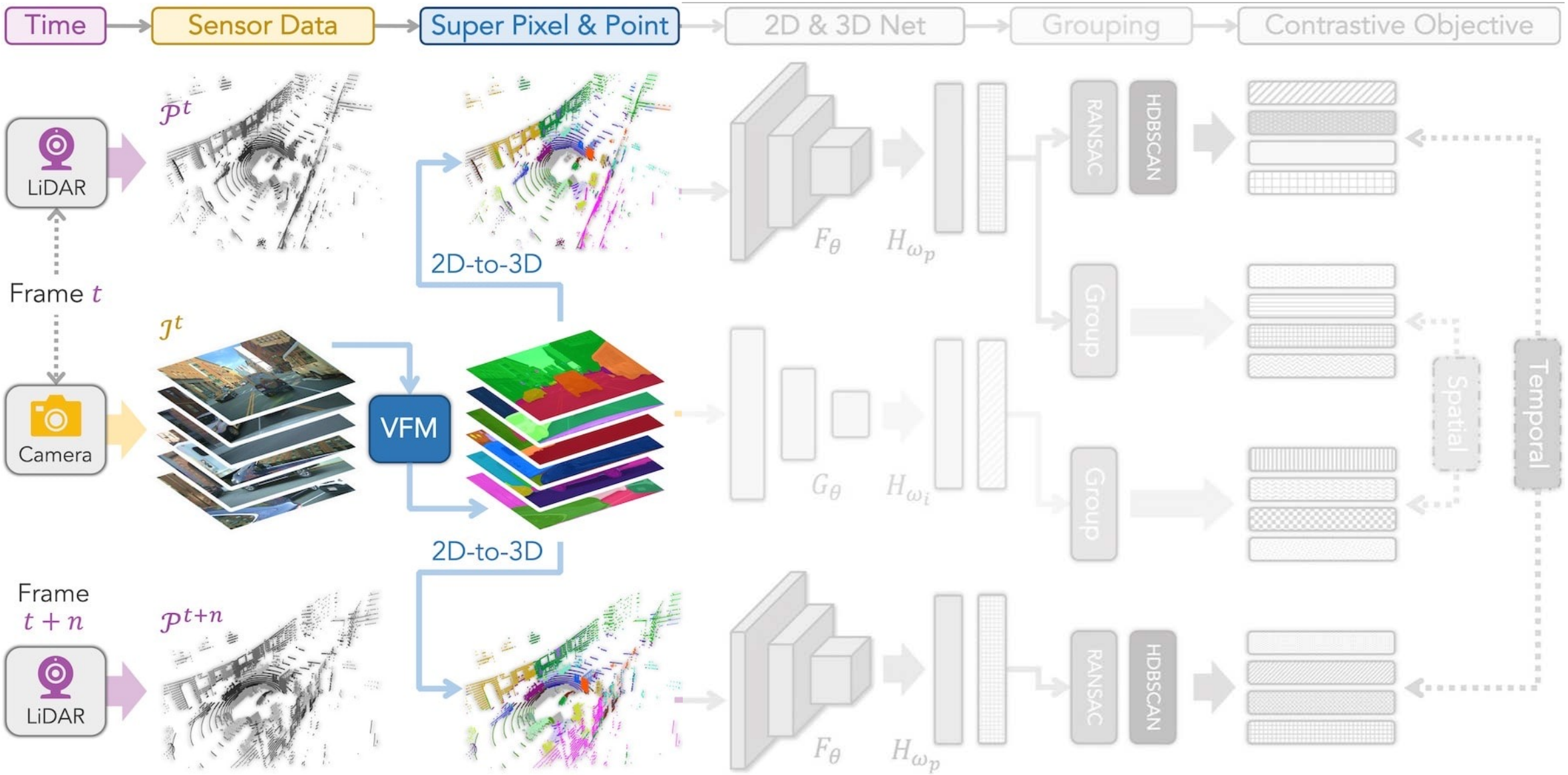
Seal: Vision Foundation Models



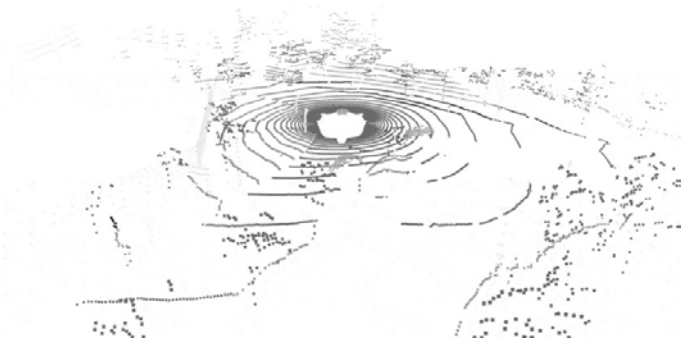
Seal: Framework

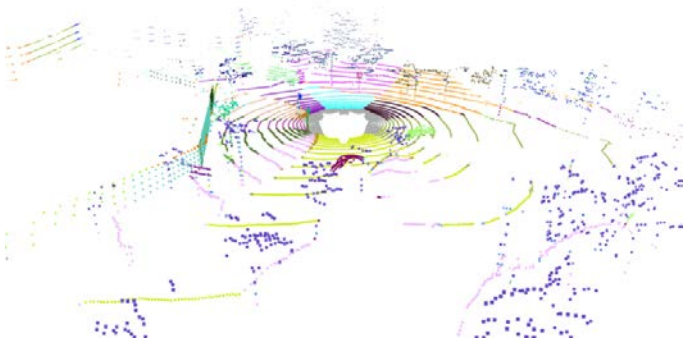


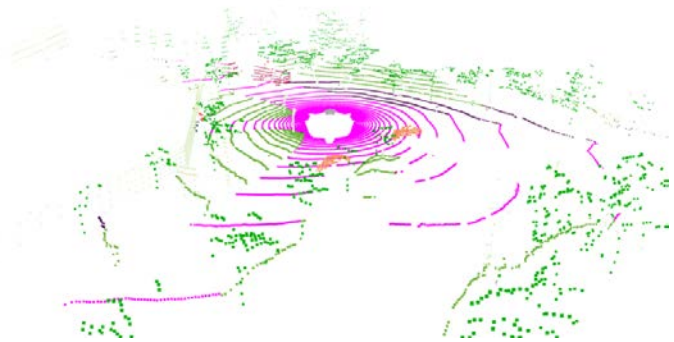
Seal: Framework

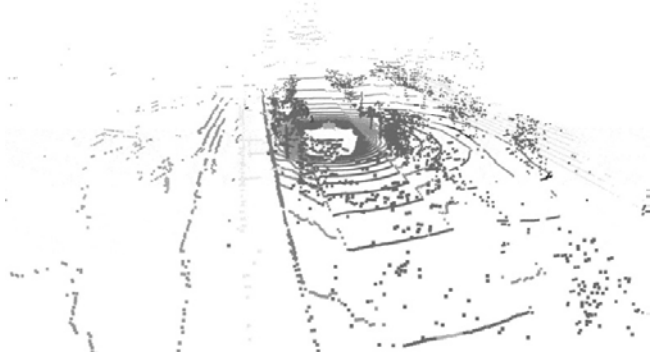


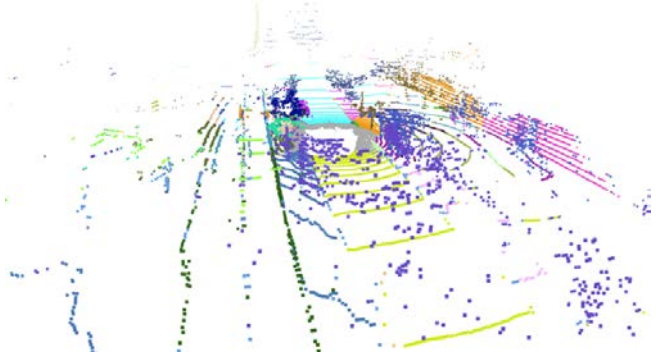
Seal: Semantic Superpoint

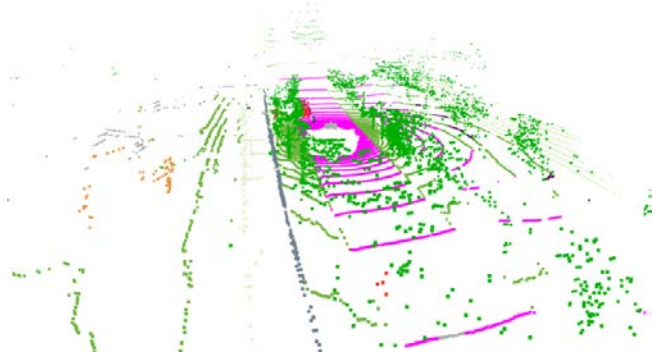










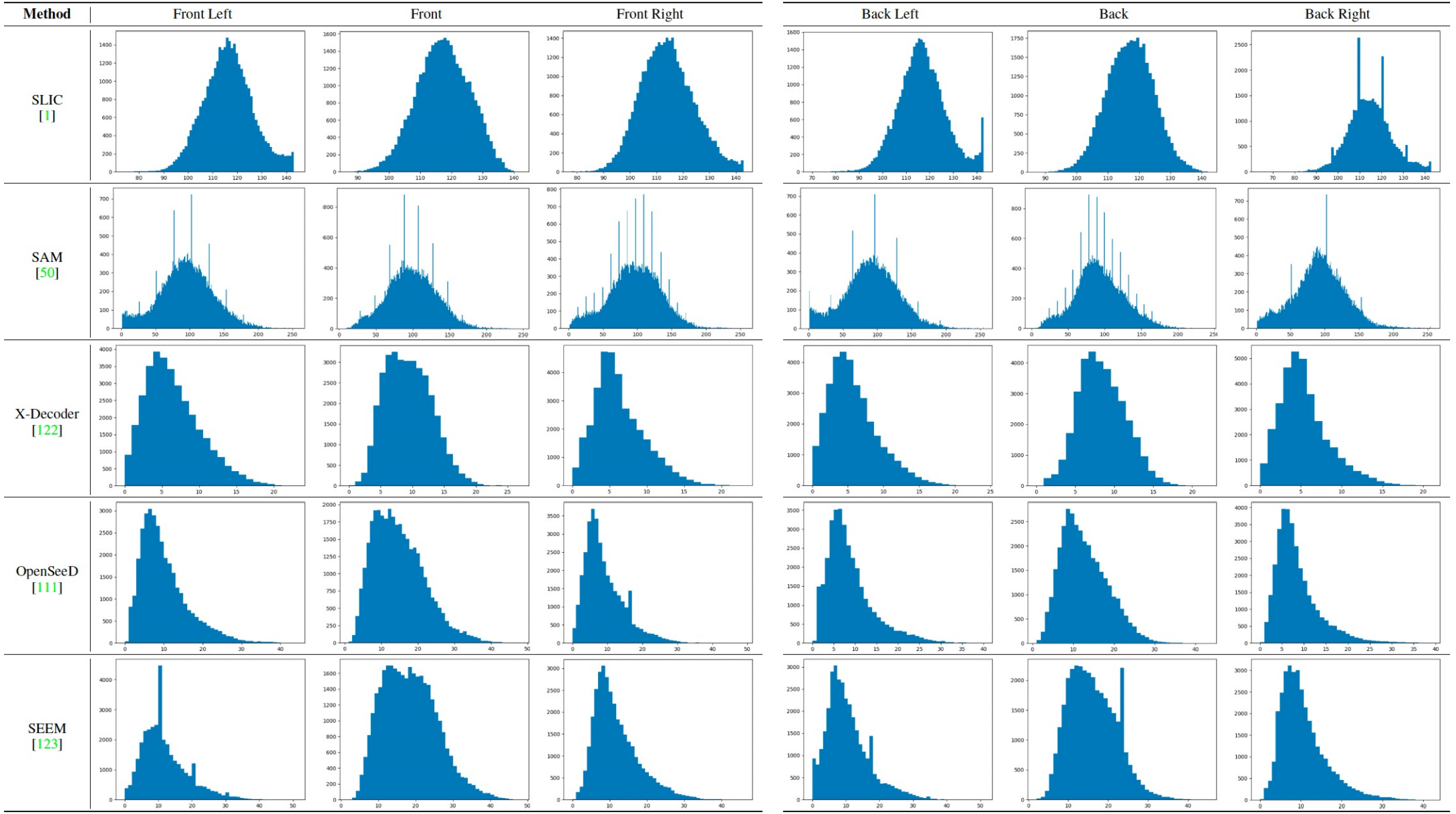


Raw Point Cloud

Semantic Superpoint

Ground-Truth

Seal: Semantic Superpixel



SLIC
(80~140)

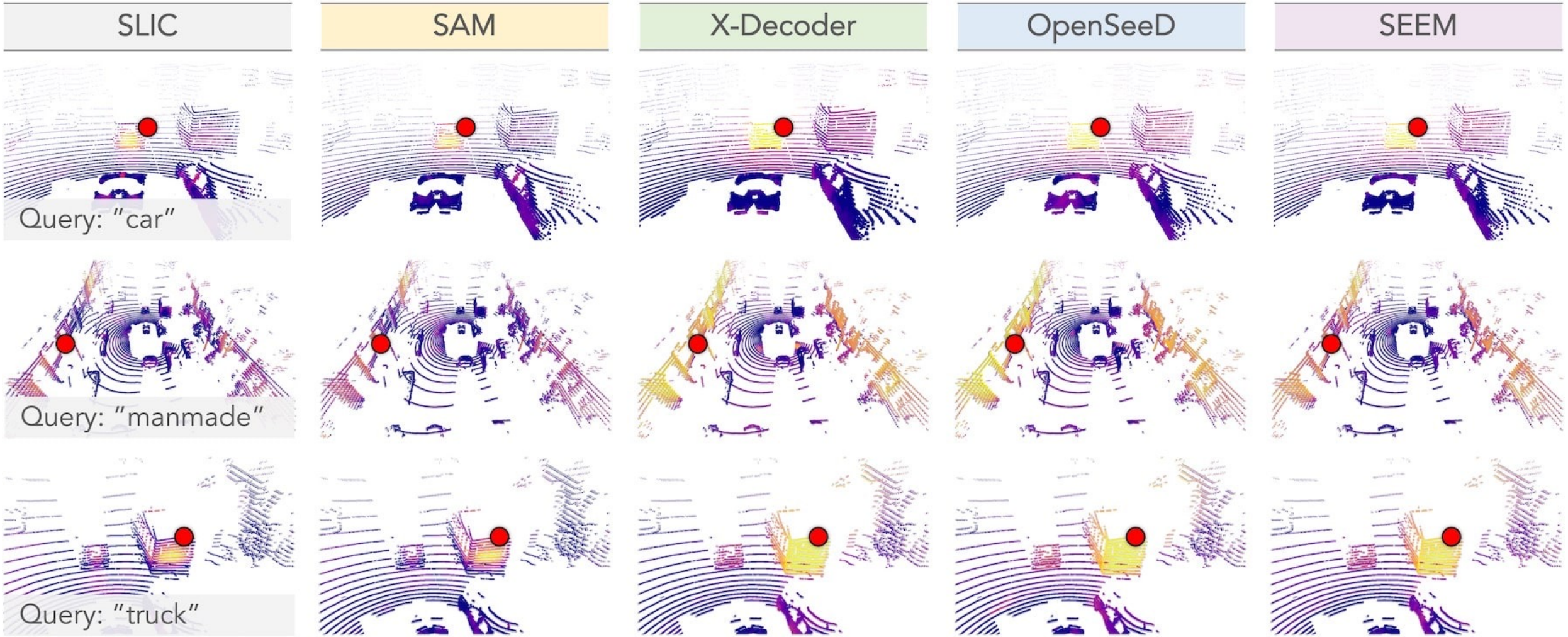
SAM
(5~250)

X-Decoder
(1~20)

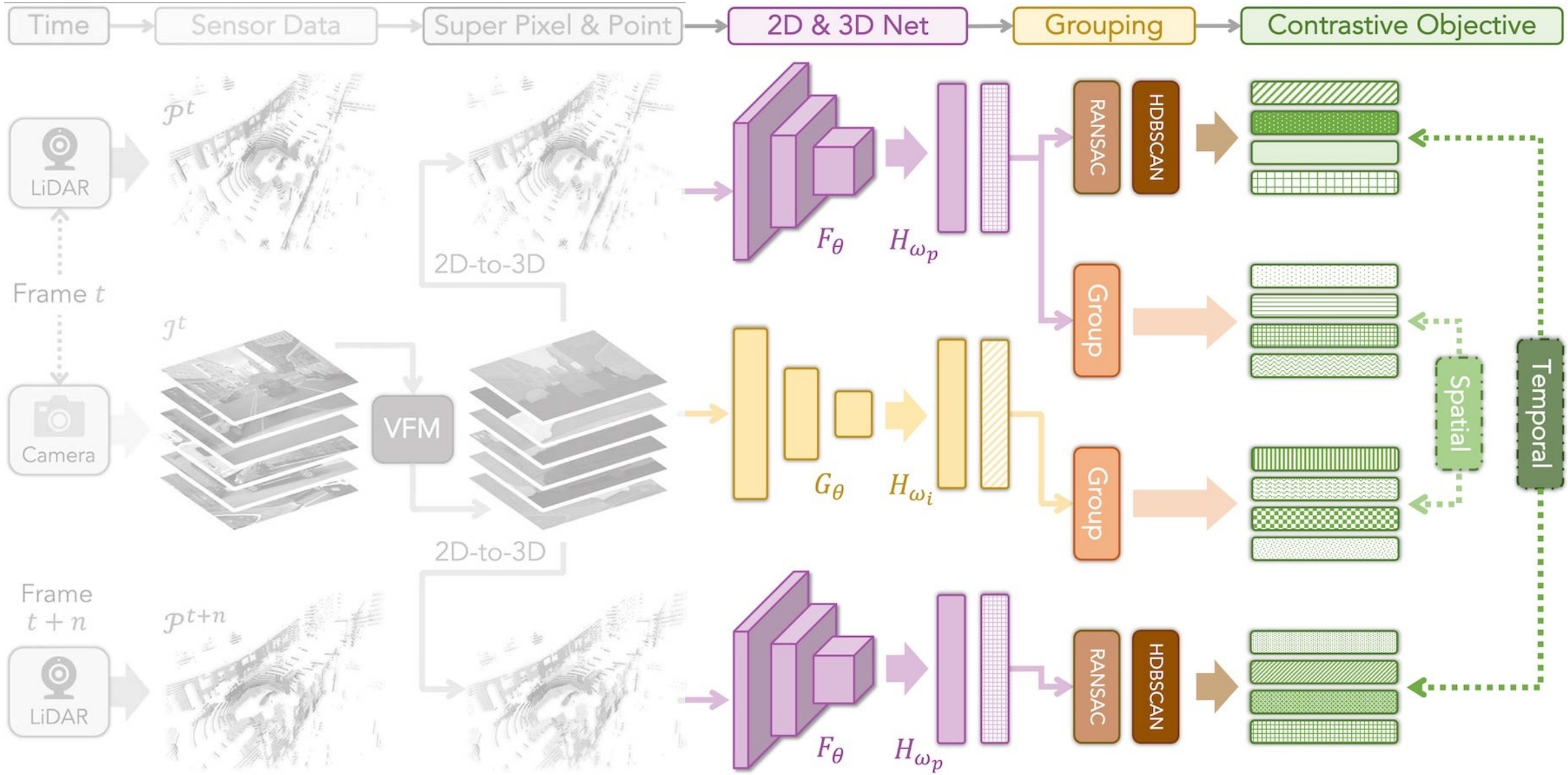
OpenSeeD
(1~40)

SEEM
(1~50)

Seal: Vision Foundation Models



Seal: Framework



Seal: Spatial & Temporal Consistency

Spatial Consistency

- Cross-modality contrastive learning:

$$\mathcal{L}^{vfm} = \mathcal{L}(\mathbf{Q}, \mathbf{K}) = -\frac{1}{M}\sum_{i=0}^{M}\log\left[\frac{e^{(\langle \mathbf{q}_i, \mathbf{k}_i\rangle/\tau)}}{\sum_{j\neq i} e^{(\langle \mathbf{q}_i, \mathbf{k}_j\rangle/\tau)} + e^{(\langle \mathbf{q}_i, \mathbf{k}_i\rangle/\tau)}}\right]$$

Temporal Consistency

- Superpoint temporal consistency:

$$\mathcal{L}^{t\to t+1} = -\frac{1}{M_k}\sum_{i=0}^{M_k}\log\left[\frac{e^{\left(\langle \mathbf{f}_i^t, \mathbf{f}_i^{t+1}\rangle/\tau\right)}}{\sum_{j\neq i} e^{\left(\langle \mathbf{f}_i^t, \mathbf{f}_j^{t+1}\rangle/\tau\right)} + e^{\left(\langle \mathbf{f}_i^t, \mathbf{f}_i^{t+1}\rangle/\tau\right)}}\right]$$

- Point-to-segment regularization:

$$\mathcal{L}^{p2s} = -\frac{1}{M_k}\sum_{i=0}^{M_k}\log\left[\frac{e^{\left(\langle \mathbf{f}_i^t, \mathbf{c}_i^t\rangle/\tau\right)}}{\sum_{j\neq i} e^{\left(\langle \mathbf{f}_i^t, \mathbf{c}_j^t\rangle/\tau\right)} + e^{\left(\langle \mathbf{f}_i^t, \mathbf{c}_i^t\rangle/\tau\right)}}\right]$$

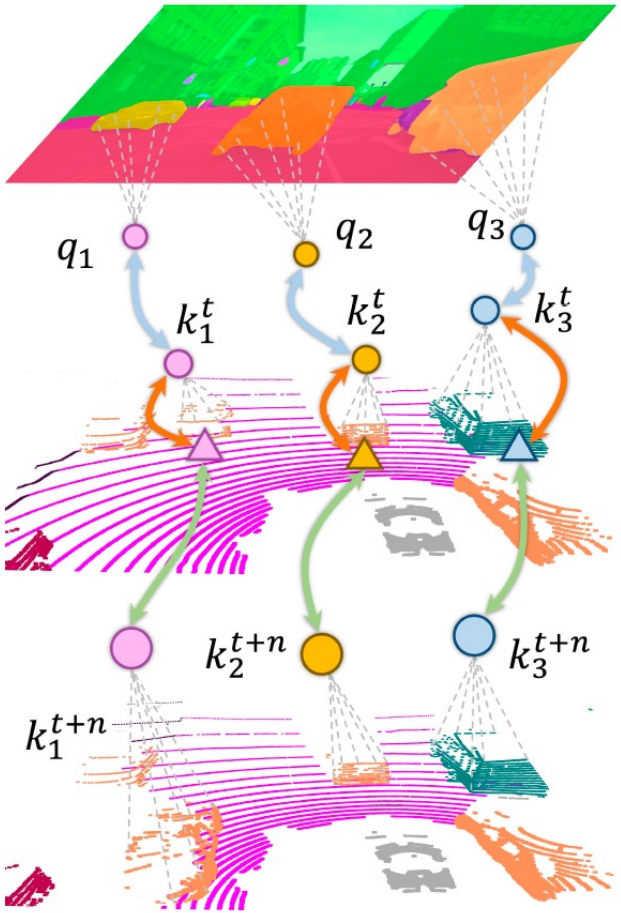


Figure 3: The positive feature correspondences in the contrastive learning objective in our contrastive learning framework. The *circles* and *triangles* represent the instance-level and the point-level features, respectively.

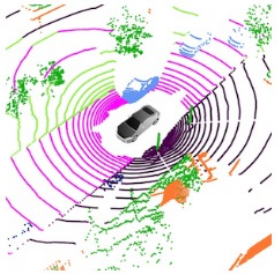
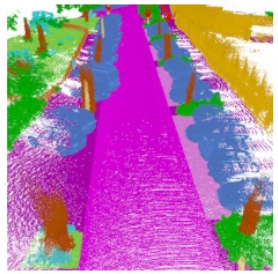
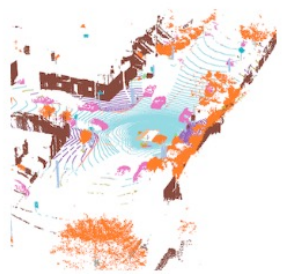
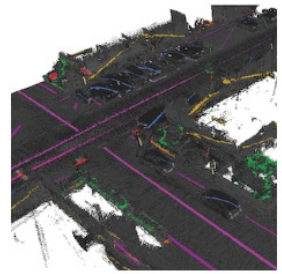
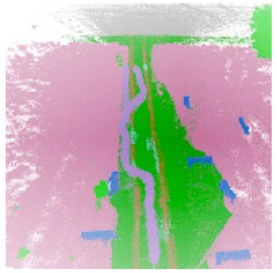
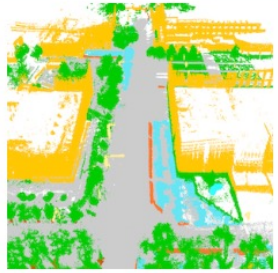
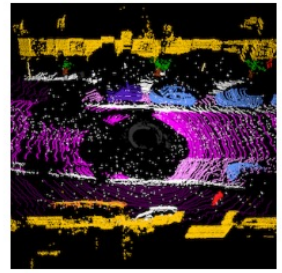
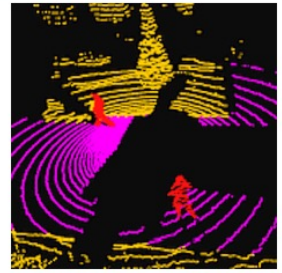
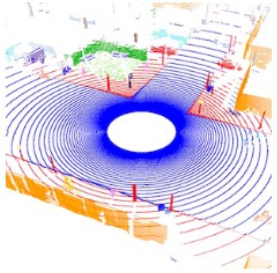
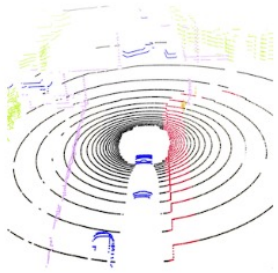
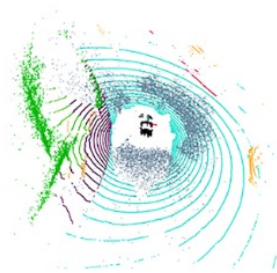
Seal: Experiments

Datasets

- Pretrained on nuScenes.
- Linear probing with frozen backbone.
- 20 downstream tasks on 11 point cloud datasets.

Backbones

- 2D: ResNet-50, pretrained with MoCoV2
- 3D: MinkUNet with cylinder voxels of size 0.1m as the input
- 2x A100 GPUs.

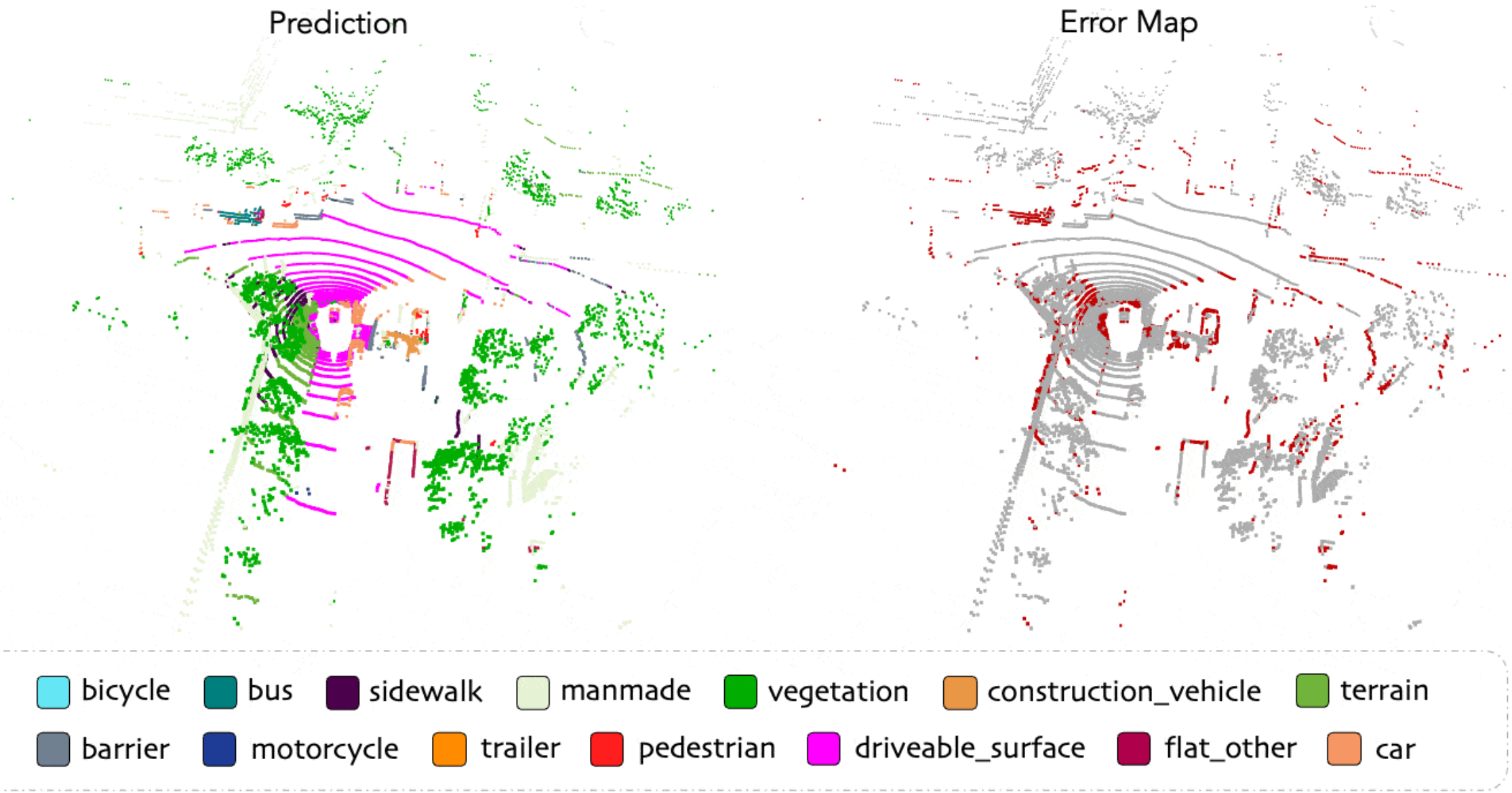
nuScenes	SemanticKITTI	Waymo Open	ScribbleKITTI
			
RELLIS-3D	**SemanticPOSS**	**SemanticSTF**	**DAPS-3D**
			
SynLiDAR	**Synth4D**	**nuScenes-C**	
			

Seal: Comparative Study

Table 1: Comparisons of different pretraining methods pretrained on *nuScenes* [26] and fine-tuned on *nuScenes* [7], *SemanticKITTI* [3], *Waymo Open* [88], and *Synth4D* [82]. **LP** denotes linear probing with frozen backbones. Symbol † denotes fine-tuning with the LaserMix augmentation [55]. Symbol ‡ denotes fine-tuning with semi-supervised learning. All mIoU scores are given in percentage (%).

Method & Year	nuScenes						KITTI	Waymo	Synth4D
	LP	1%	5%	10%	25%	Full	1%	1%	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	-	-	-

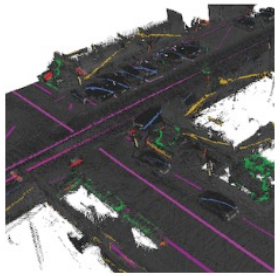
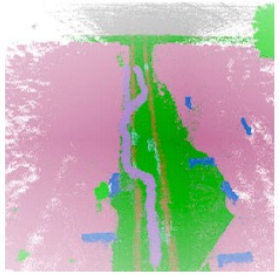
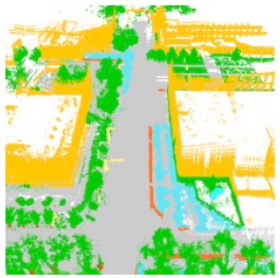
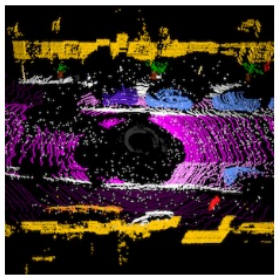
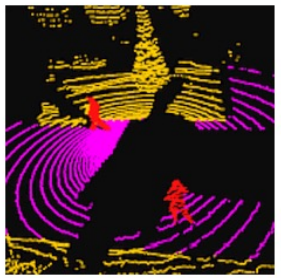
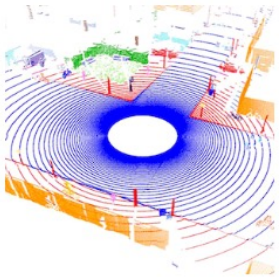
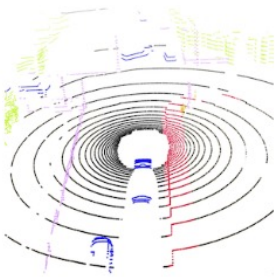
Seal: Linear Probing



Seal: Downstream Generalization

Table 2: Comparisons of different pretraining methods pretrained on *nuScenes* [26] and fine-tuned on different downstream point cloud datasets. All mIoU scores are given in percentage (%).

Method	ScribbleKITTI		RELLIS-3D		SemanticPOSS		SemanticSTF		SynLiDAR		DAPS-3D	
	1%	10%	1%	10%	Half	Full	Half	Full	1%	10%	Half	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**

ScribbleKITTI	RELLIS-3D	SemanticPOSS	SemanticSTF	DAPS-3D	SynLiDAR	Synth4D
						

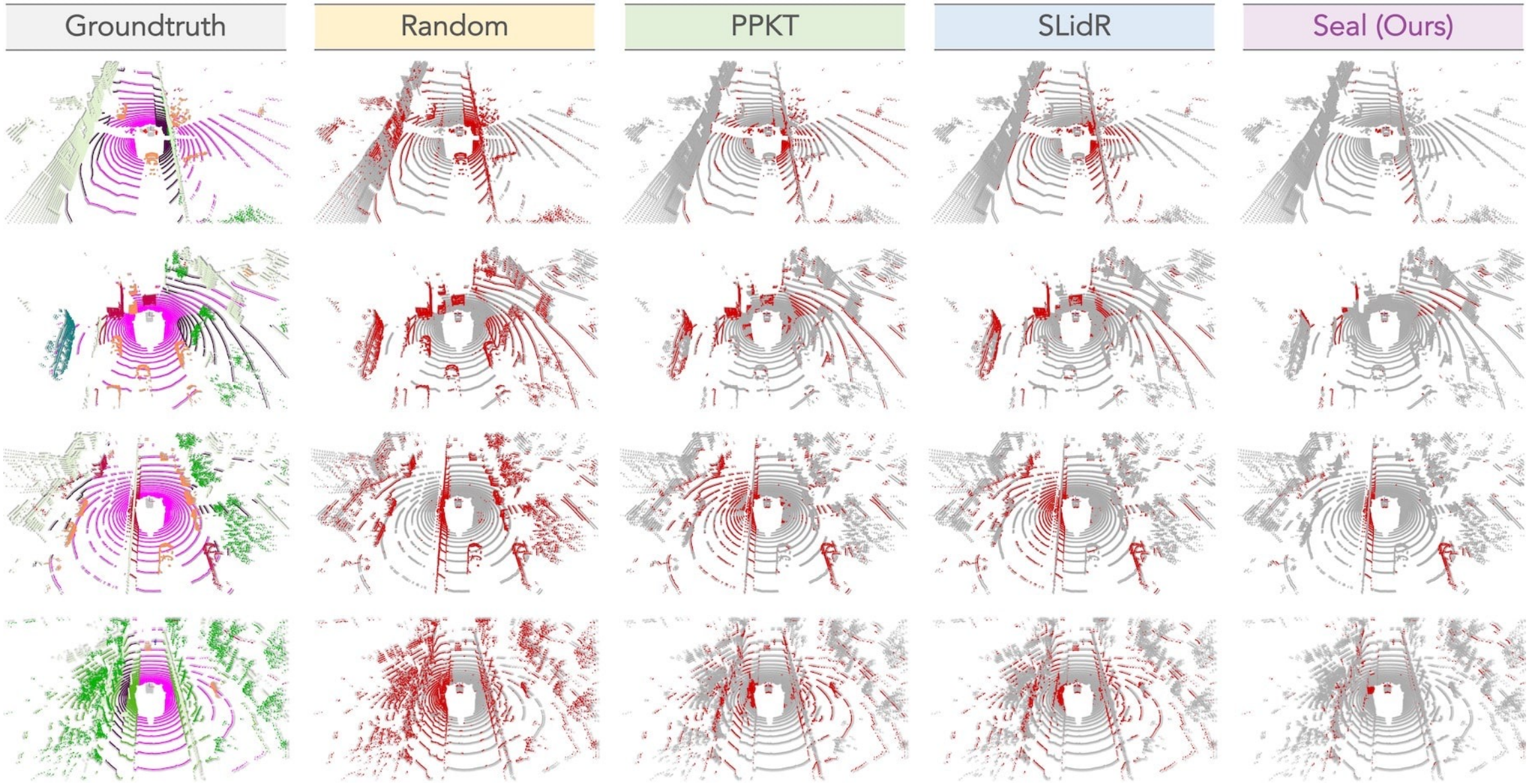
Seal: Robustness Probing

Table 3: Robustness evaluations under eight out-of-distribution corruptions in the *nuScenes-C* dataset from the Robo3D benchmark [53]. All mCE, mRR, and mIoU scores are given in percentage (%).

	Initial	Backbone	mCE ↓	mRR ↑	Fog	Wet	Snow	Move	Beam	Cross	Echo	Sensor
LP	PPKT [65]	MinkUNet	183.44	**78.15**	30.65	35.42	28.12	29.21	32.82	19.52	28.01	20.71
	SLidR [85]	MinkUNet	179.38	77.18	34.88	38.09	**32.64**	26.44	33.73	**20.81**	31.54	21.44
	Seal (Ours)	MinkUNet	**166.18**	75.38	**37.33**	**42.77**	29.93	**37.73**	**40.32**	20.31	**37.73**	**24.94**
Full	Random	PolarNet	115.09	76.34	58.23	69.91	64.82	44.60	61.91	40.77	53.64	42.01
	Random	CENet	112.79	76.04	67.01	69.87	61.64	58.31	49.97	**60.89**	53.31	24.78
	Random	WaffleIron	106.73	72.78	56.07	73.93	49.59	59.46	65.19	33.12	**61.51**	44.01
	Random	Cylinder3D	105.56	78.08	61.42	71.02	58.40	56.02	64.15	45.36	59.97	43.03
	Random	SPVCNN	106.65	74.70	59.01	72.46	41.08	58.36	65.36	36.83	62.29	**49.21**
	Random	MinkUNet	112.20	72.57	62.96	70.65	55.48	51.71	62.01	31.56	59.64	39.41
	PPKT [65]	MinkUNet	105.64	76.06	64.01	72.18	59.08	57.17	63.88	36.34	60.59	39.57
	SLidR [85]	MinkUNet	106.08	75.99	65.41	72.31	56.01	56.07	62.87	41.94	61.16	38.90
	Seal (Ours)	MinkUNet	**92.63**	**83.08**	**72.66**	**74.31**	**66.22**	**66.14**	**65.96**	57.44	59.87	39.85

L. Kong, et al. "Robo3D: Towards robust and reliable 3D perception under corruptions," arXiv, 2023.

Seal: Qualitative Assessment



Seal: Ablation Study

Table 4: Ablation study on pretraining frameworks (ours *vs.* SLidR [85]) and the knowledge transfer effects from different vision foundation models. All mIoU scores are given in percentage (%).

Method	Superpixel	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
SLidR	SLIC [1]	38.80	38.30	52.49	59.84	66.91	74.79	44.60	47.12	63.10
	SAM [50]	41.49	43.67	**55.97**	**61.74**	**68.85**	**75.40**	43.35	48.64	63.15
	X-Decoder [122]	41.71	43.02	54.24	61.32	67.35	75.11	45.70	48.73	63.21
	OpenSeeD [111]	42.61	43.82	54.17	61.03	67.30	74.85	**45.88**	48.64	**63.31**
	SEEM [123]	**43.00**	**44.02**	53.03	60.84	67.38	75.21	45.72	**48.75**	63.13
Seal	SLIC [1]	40.89	39.77	53.33	61.58	67.78	75.32	45.75	47.74	63.37
	SAM [50]	43.94	45.09	**56.95**	62.35	**69.08**	**75.92**	46.53	49.00	63.76
	X-Decoder [122]	42.64	44.31	55.18	62.03	68.24	75.56	46.02	49.11	64.21
	OpenSeeD [111]	44.67	44.74	55.13	62.36	69.00	75.64	46.13	48.98	64.29
	SEEM [123]	**44.95**	**45.84**	55.64	**62.97**	68.41	75.60	**46.63**	**49.34**	**64.50**

Seal: Ablation Study

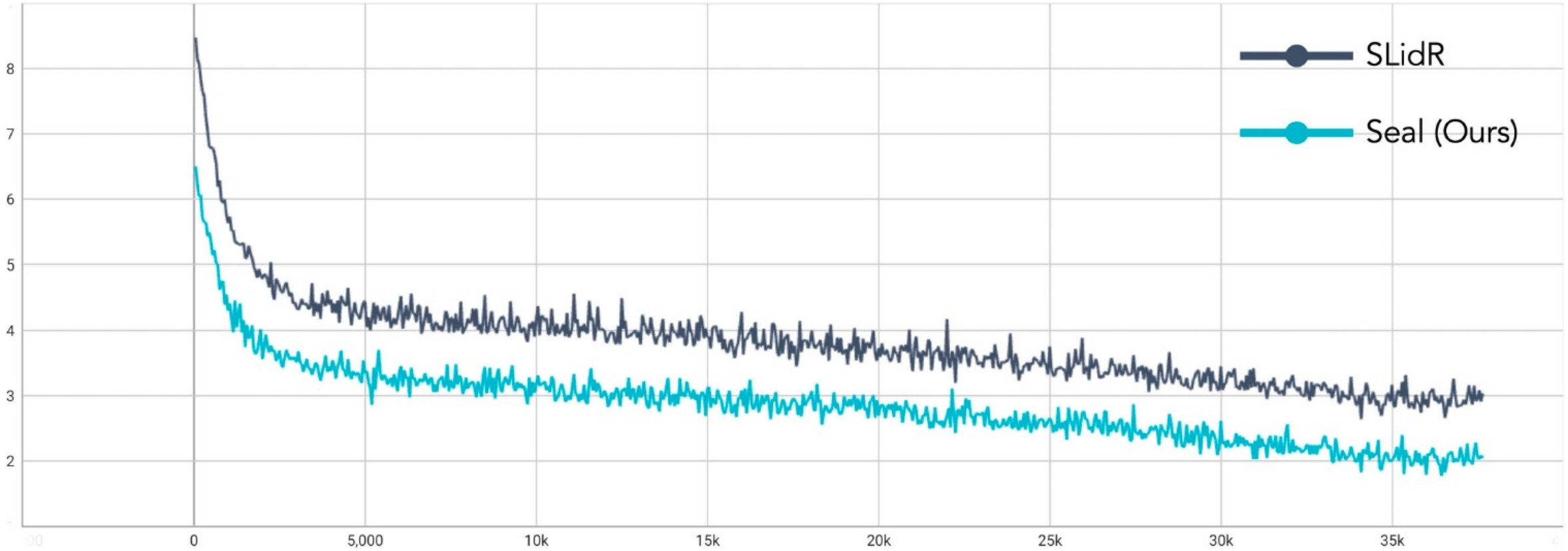


Figure 10: The convergence rate comparison between SLidR [85] and the proposed *Seal* framework.

Seal: Ablation Study

Table 5: Ablation study of each component pretrained on *nuScenes* [26] and fine-tuned on *nuScenes* [26], *SemanticKITTI* [3], and *Waymo Open* [88]. **C2L:** Camera-to-LiDAR distillation. **VFM:** Vision foundation models. **STC:** Superpoint temporal consistency. **P2S:** Point-to-segment regularization.

#	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**

Open-Source Resources

Seal

Paper

Code

MMDet3D

超级视客营第二期 再度起航

https://openmmlab.com/activity/codecamp

10 个课题方向、150+ 课题任务

基础架构、目标检测、3D目标检测

预训练+多模态、AIGC、部署

姿态估计、语义分割、动作识别

活动详情页 HOT

活动交流群

运用视觉基础模型分割「任意」点云

Q & A

OpenMMLab 公众号
回复“社区开放麦”即可获取课件

MMDetection3D 交流群
一起讨论技术流