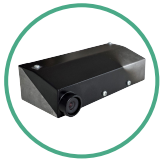
Foundation models in the automotive industry

Andrei BURSUC
valeo.ai
17 June 2024

Valeo’s history in ADAS

1991 — Ultrasonic Parking Aid

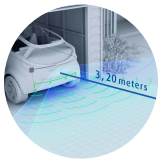
2004 — Lane-Departure Warning



2006 — Multi-beam Radar

2007 — Semi-automatic Park Assist

2013 — Object detection in Rear Camera



2016 — Remote Park Assist

2020 — Vision-only L2 System

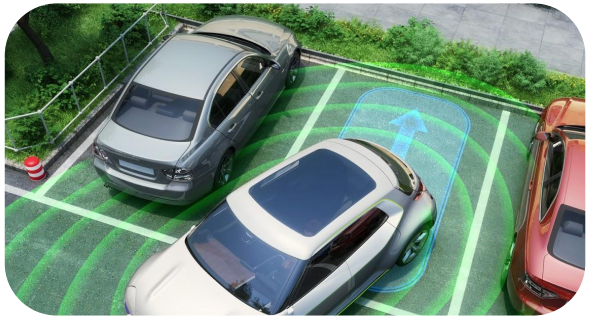
2022 — SCALA 2 for 1st consumer L3

2023 — Complete L2 ADAS System

2025 — ADAS Domain Controller

Covering all ADAS segments

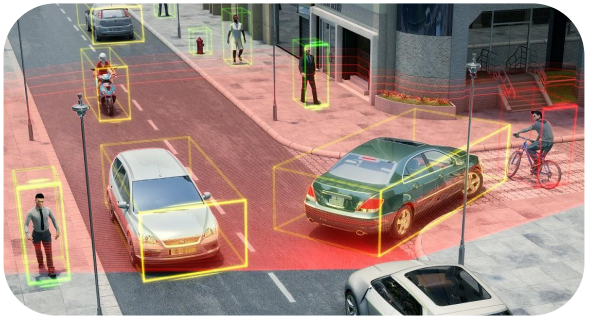
One-stop solutions for
PARKING ASSISTANCE



AI perception algorithms

Next generation sensors, cameras and processing units

One-stop solutions for
SAFETY & ASSISTED DRIVING



Scalable solutions from smart front camera to centralized architecture

Valeo algorithms for applications

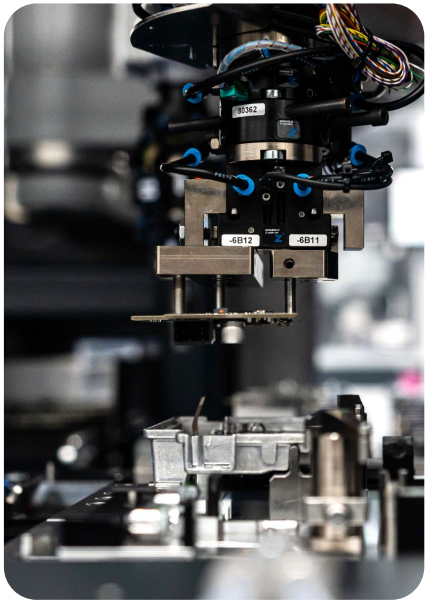
Perception & Functions for
AUTOMATED DRIVING

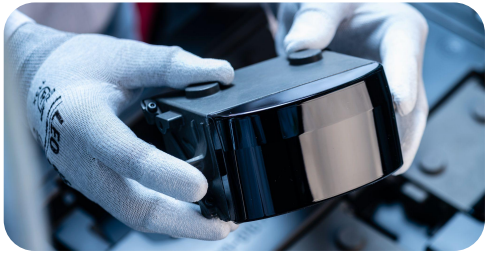
Modules & back-up functions

High performance sensors, cameras and processing units

Expertise for solutions ranging from individual components to complete turnkey systems

1.5+ Billion sensors shipped in 30 years



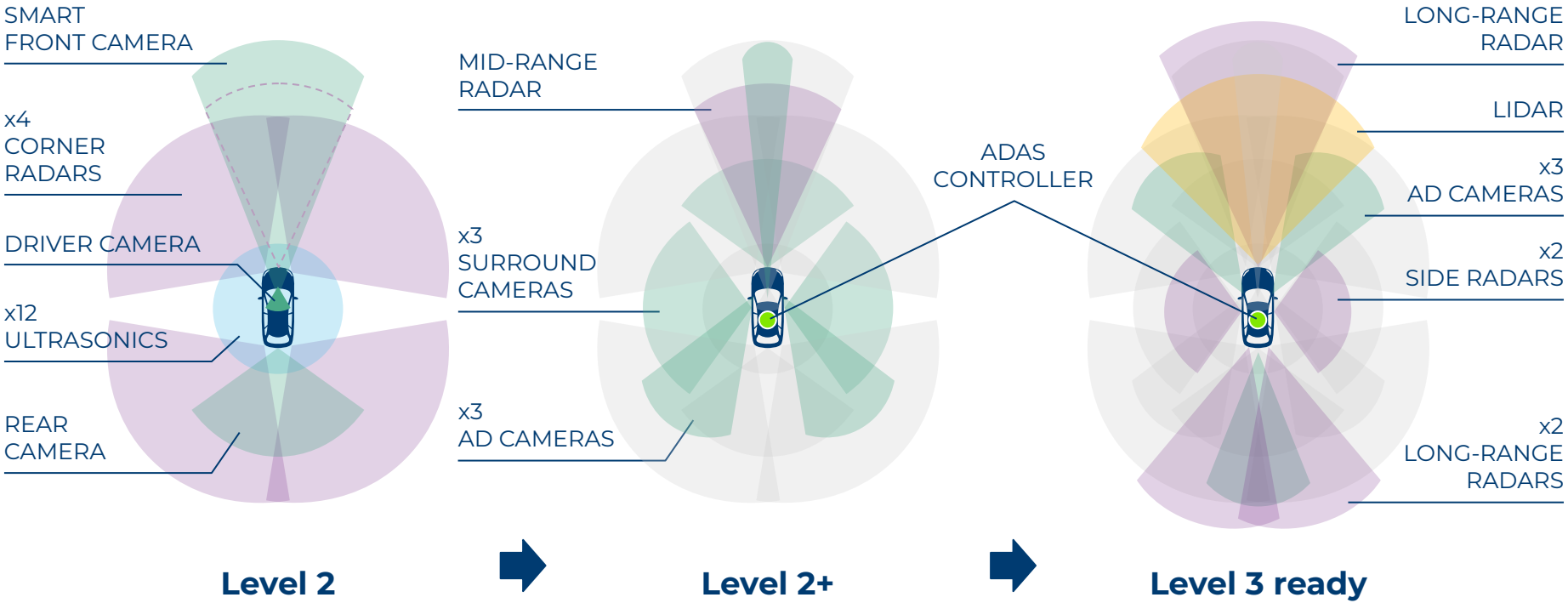


Another **1.5+ billion sensors** to be shipped in the next 5 years

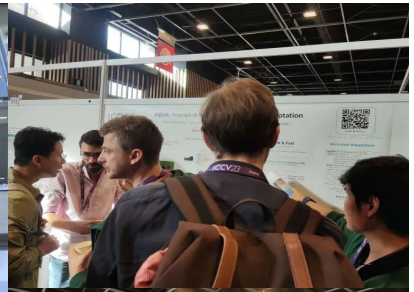
Valeo sensor suite

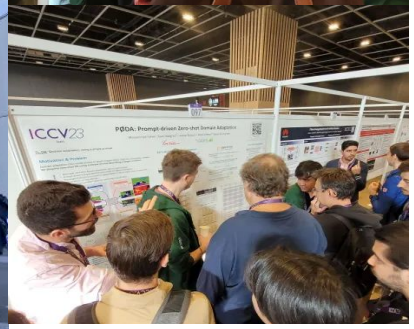
ULS	RADARS		CAMERAS		LIDARS	
Ultrasonic sensors	Near field radars	Mid range radars	Surround view cameras	Long range cameras	Near field lidars	SCALA

Scalable system architecture



valeo.ai





- ~25 researchers & PhDs
- Dedicated to open research
- 10s of academic collabs across France and Europe
- Offices: Paris, Prague
- Topics: perception, data efficiency, forecasting, reliability, explainability

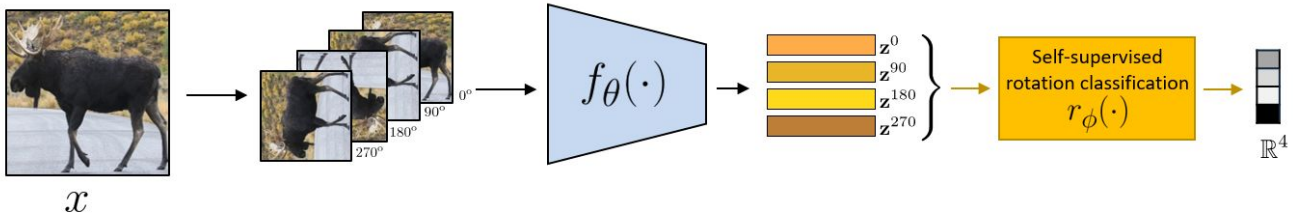
https://valeoai.github.io

Foundation models: train once, use many times

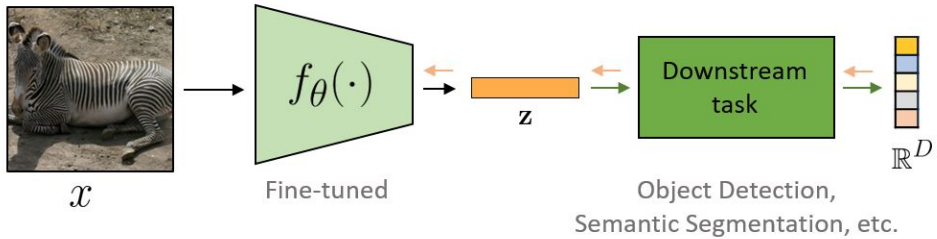
“A foundation model is any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks.” (Bommasani et al., 2021)

Recap - self-supervised learning pipeline

Stage 1: Pretrain network on pretext task (without human labels)



Stage 2: Fine-tune network for new task with fewer labels



Why would we want to do self-supervised learning?

For driving scenes:

Data is enormous,

Scalability of recent models (ChatGPT et al.),

When possible, manual annotation is difficult,

Annotation is impossible for some important tasks

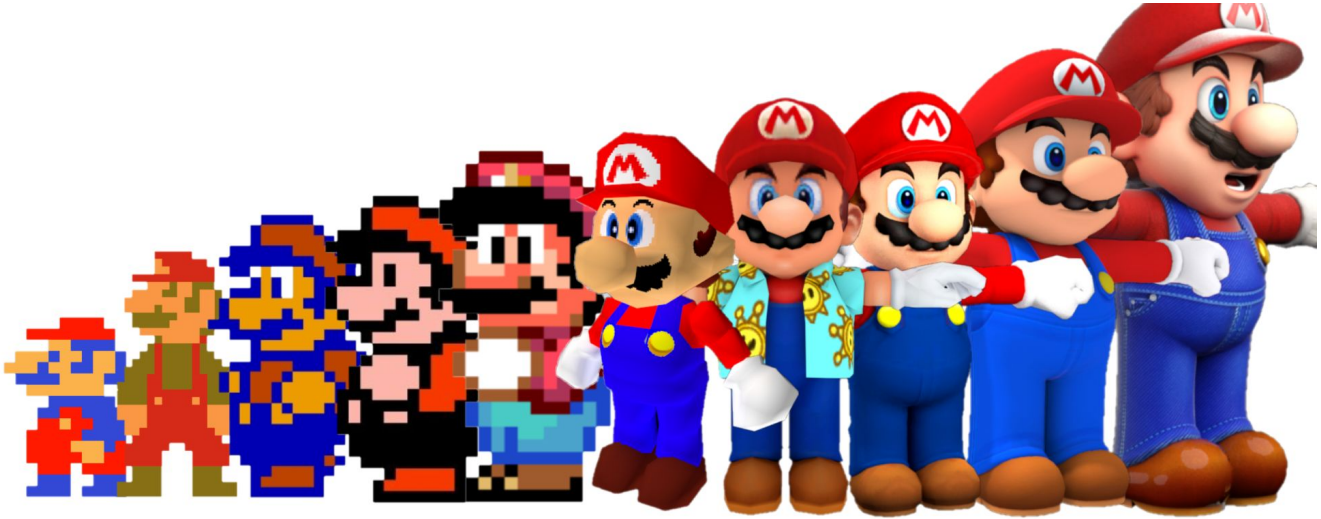
Training datasets are evolving continuously

Difficult to keep the pace with an ever changing world

Men's fashion trends 1980-1989

- Data distributions shift all the time, e.g., fashion trends, vehicle types
- Infeasible to launch large annotation campaigns each time

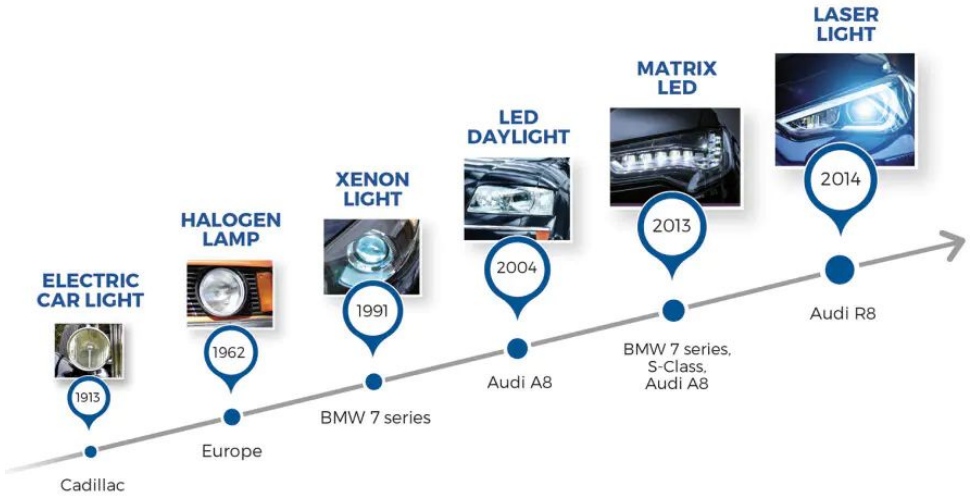
Difficult to keep the pace with an ever changing world



Super Mario from 1981 to 2017

- Sensors specs are frequently upgraded
- Infeasible to launch large annotation campaigns each time

Difficult to keep the pace with an ever changing world

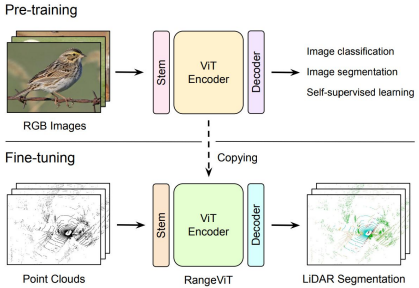


Halogen vs. LED

- Sensors specs are frequently upgraded
- Headlamps change the appearance of the scenes and of the vehicles to detect

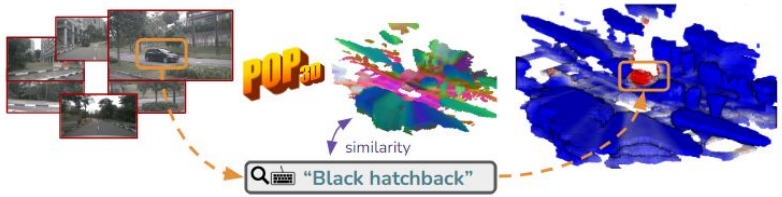
New use-cases possible with recent foundation models

Stage 2: Can also be distillation, auto-labeling, data mining, active learning, model initialization, etc.



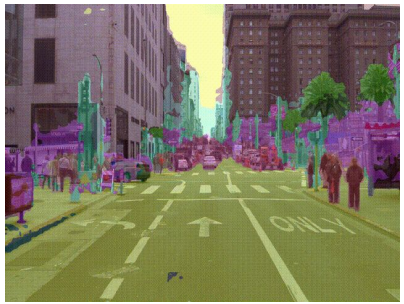
RangeViT [CVPR'23]

text-driven 3D retrieval from cameras

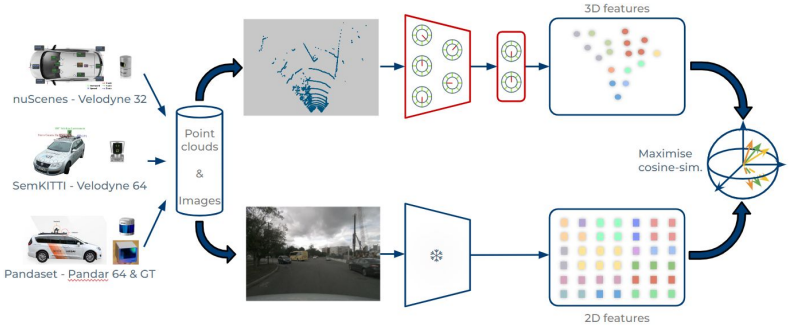


POP-3D [NeurIPS'23]

unsupervised semantic segmentation

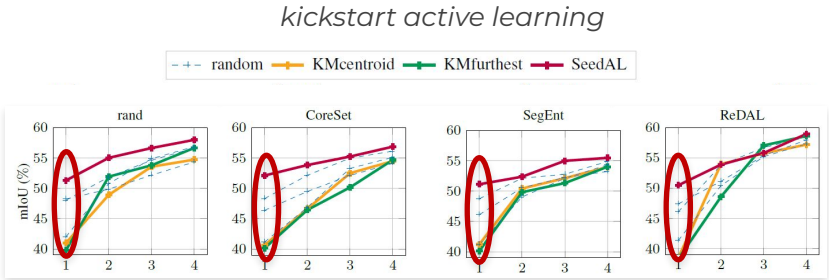


Drive&Segment [ECCV'22]



ScaLR [CVPR'24]

kickstart active learning



SeedAL [ICCV'23]

Modular pretraining

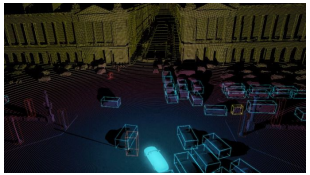


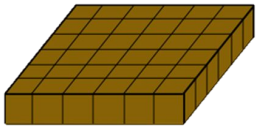
Image encoder

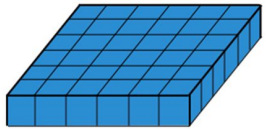
Lidar encoder

BEV projection

Flatten

BEV features

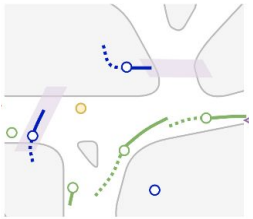


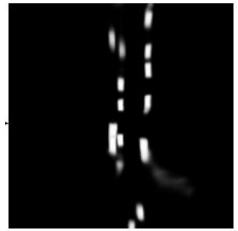


Downstream decoders

Perception + Forecasting

...



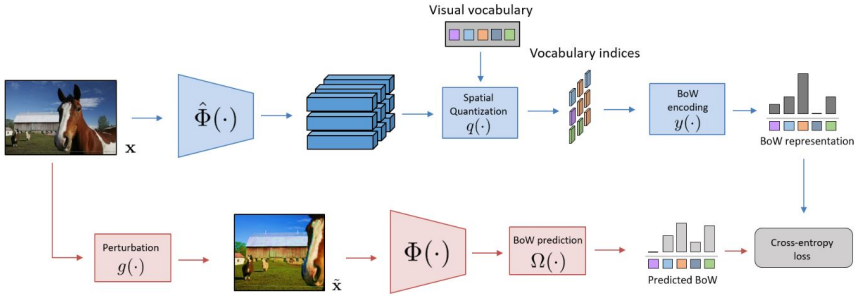


- Pretrain each branch individually to leverage knowledge from different distributions and data sources, sensors, sensor rigs
- Plug pretrained modules to downstream ones and continue training

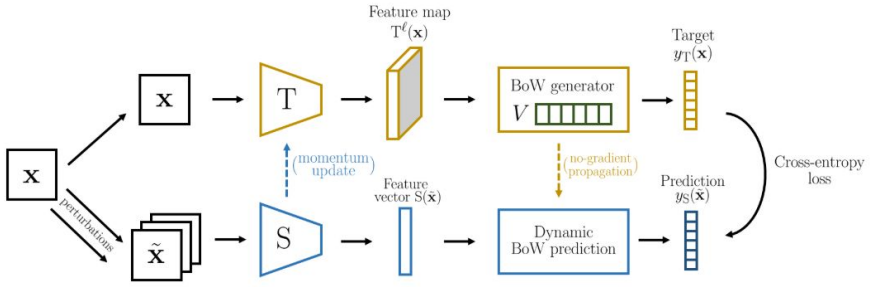
One model for all

Pretraining image encoders

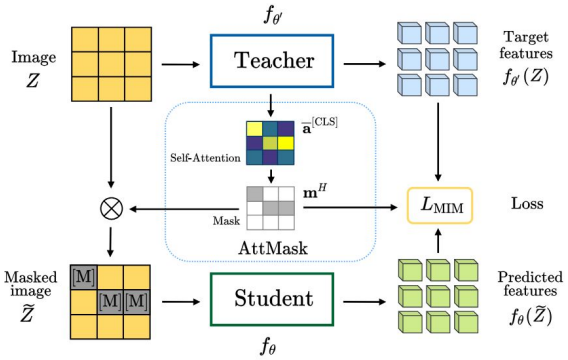
Pretraining image encoders



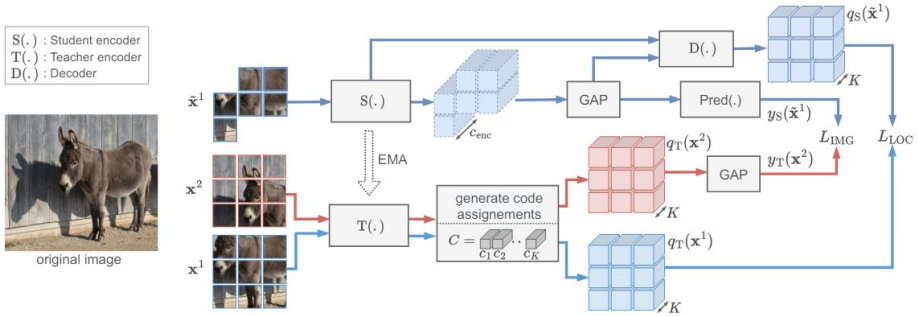
BowNet[CVPR'20]



OBoW[CVPR'21]



AttMask[ECCV'22]



MOCA[TMLR'24]

- Long-term exploration of (dense) SSL strategies with focus on 2D downstream performance and few-shot adaptation

Challenges in using self-supervised learning with autonomous driving data

$\sim$ *ImageNet*

$\sim$ *BDD100K*

Deng. et al., ImageNet: A large-scale hierarchical image database, CVPR 2009
Yu et al., BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning, CVPR 2020

$\sim$ *ImageNet*

$\sim$ *BDD100K*

Deng. et al., ImageNet: A large-scale hierarchical image database, CVPR 2009
Yu et al., BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning, CVPR 2020

$\sim$ *ImageNet*

$\sim$ *BDD100K*

Deng. et al., ImageNet: A large-scale hierarchical image database, CVPR 2009
Yu et al., BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning, CVPR 2020

$\sim$ *ImageNet*

$\sim$ *BDD100K*

Deng. et al., ImageNet: A large-scale hierarchical image database, CVPR 2009
Yu et al., BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning, CVPR 2020

$\sim$ *ImageNet*

$\sim$ *BDD100K*

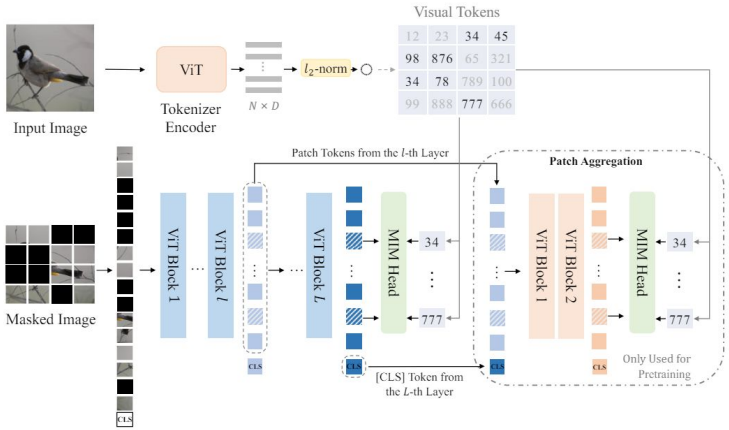
Deng. et al., ImageNet: A large-scale hierarchical image database, CVPR 2009
Yu et al., BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning, CVPR 2020

From the ImageNet perspective, AD data is “boring”.
Most SSL methods seem devised for well curated datasets.

Yu et al., BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning, CVPR 2020

Pretraining image encoders

- AD pretraining can struggle to match pretraining on large image collections (e.g,, ImageNet1K/21K, Instagram, etc.) (Chen et al., 2021)
- Recent SSL methods (BEITv2) distill knowledge from pretrained models (DINO, CLIP)
- Alternatively, finetuning large pretrained models is effective (Wei et al., 2024)



BEiT v2

Backbone	Fine-tune Method	Trainable Params*	mIoU Citys	mIoU BDD	mIoU Map	mIoU Avg.
CLIP [63] (ViT-Large)	Full	304.15M	51.3	47.6	54.3	51.1
	Freeze	0.00M	53.7	48.7	55.0	52.4
	Rein	2.99M	**57.1**	**54.7**	**60.5**	**57.4**
MAE [27] (Large)	Full	330.94M	53.7	**50.8**	58.1	54.2
	Freeze	0.00M	43.3	37.8	48.0	43.0
	Rein	2.99M	**55.0**	49.3	**58.6**	**54.3**
SAM [42] (Huge)	Full	632.18M	57.6	51.7	61.5	56.9
	Freeze	0.00M	57.0	47.1	58.4	54.2
	Rein	4.51M	**59.6**	**52.0**	**62.1**	**57.9**
EVA02 [18, 19] (Large)	Full	304.24M	62.1	56.2	64.6	60.9
	Freeze	0.00M	56.5	53.6	58.6	56.2
	Rein	2.99M	**65.3**	**60.5**	**64.9**	**63.6**
DINOV2 [58] (Large)	Full	304.20M	63.7	57.4	64.2	61.7
	Freeze	0.00M	63.3	56.1	63.9	61.1
	Rein	2.99M	**66.4**	**60.4**	**66.1**	**64.3**

Generalization for semantic segmentation

K. Chen et al., MultiSiam: Self-supervised Multi-instance Siamese Representation Learning for Autonomous Driving, ICCV 2021
Z. Peng et al., BEiT v2: Masked Image Modeling with Vector-Quantized Visual Tokenizers, arXiv 2022
Z. Wei et al., Stronger, Fewer, & Superior: Harnessing Vision Foundation Models for Domain Generalized Sem. Seg., CVPR 2024

One model for all

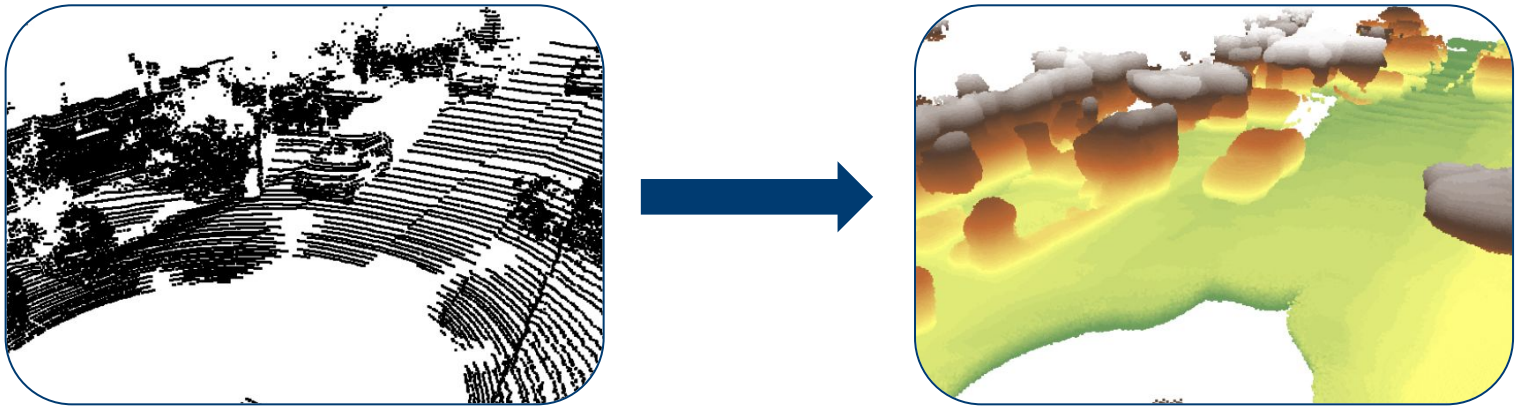
Pretraining Lidar encoders

The self-supervised 2-stage paradigm (pretrain + finetune) can be extend to other sensor modalities with some adequate pretext tasks

Automotive Lidar Self-Supervision by Occupancy Estimation

Intuition

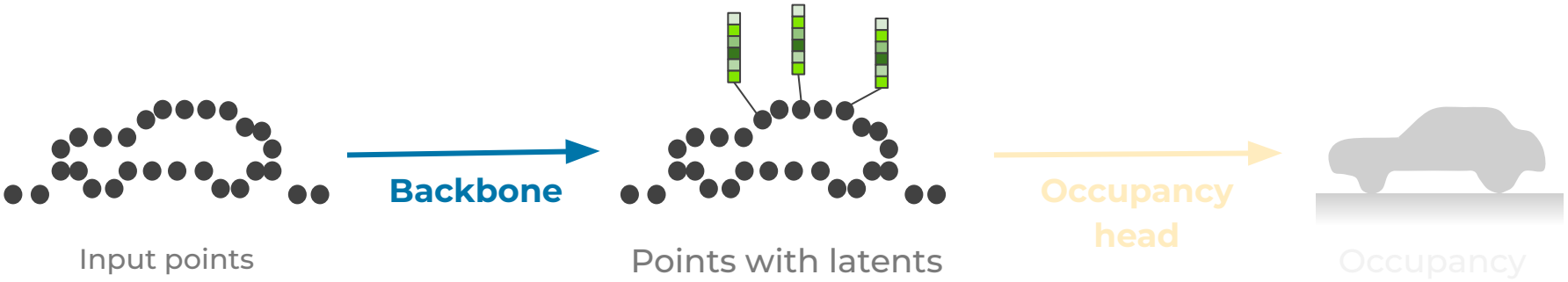
- Reconstruction information (geometry) requires semantic knowledge
- Point cloud ~ masked surface



A. Boulch et al., ALSO: Automotive Lidar Self-Supervision by Occupancy Estimation, CVPR 2023

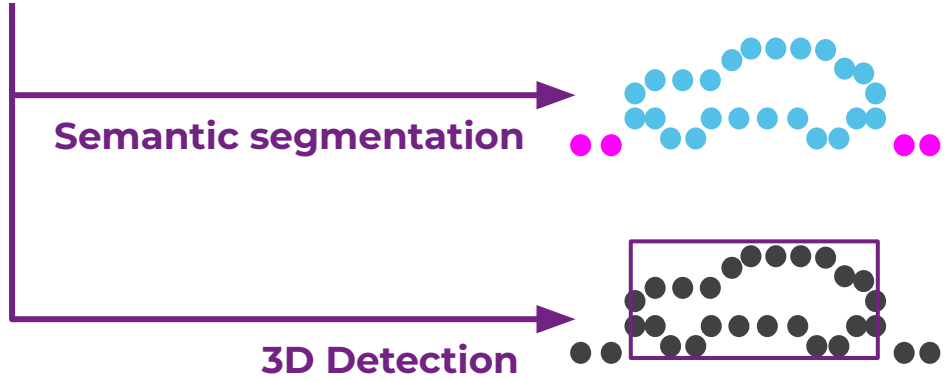
Automotive Lidar Self-Supervision by Occupancy Estimation

Overview



Pretext task: scene reconstruction
- Implicit representation of surface (latent vectors)
- Occupancy decoding

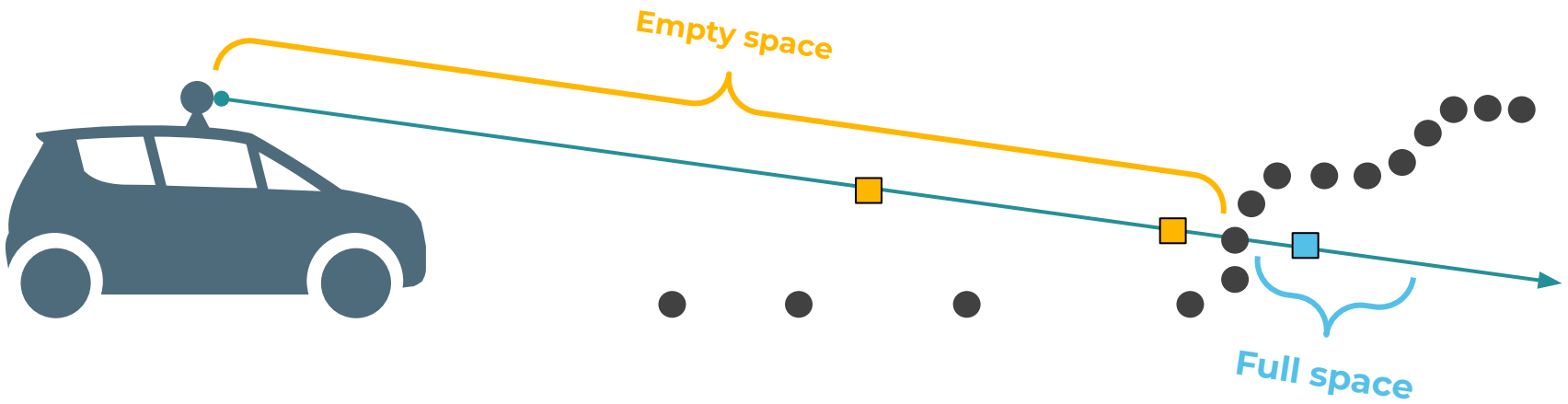
Downstream
Semantic segmentation and Detection



A. Boulch et al., ALSO: Automotive Lidar Self-Supervision by Occupancy Estimation, CVPR 2023

Automotive Lidar Self-Supervision by Occupancy Estimation

Pretext task



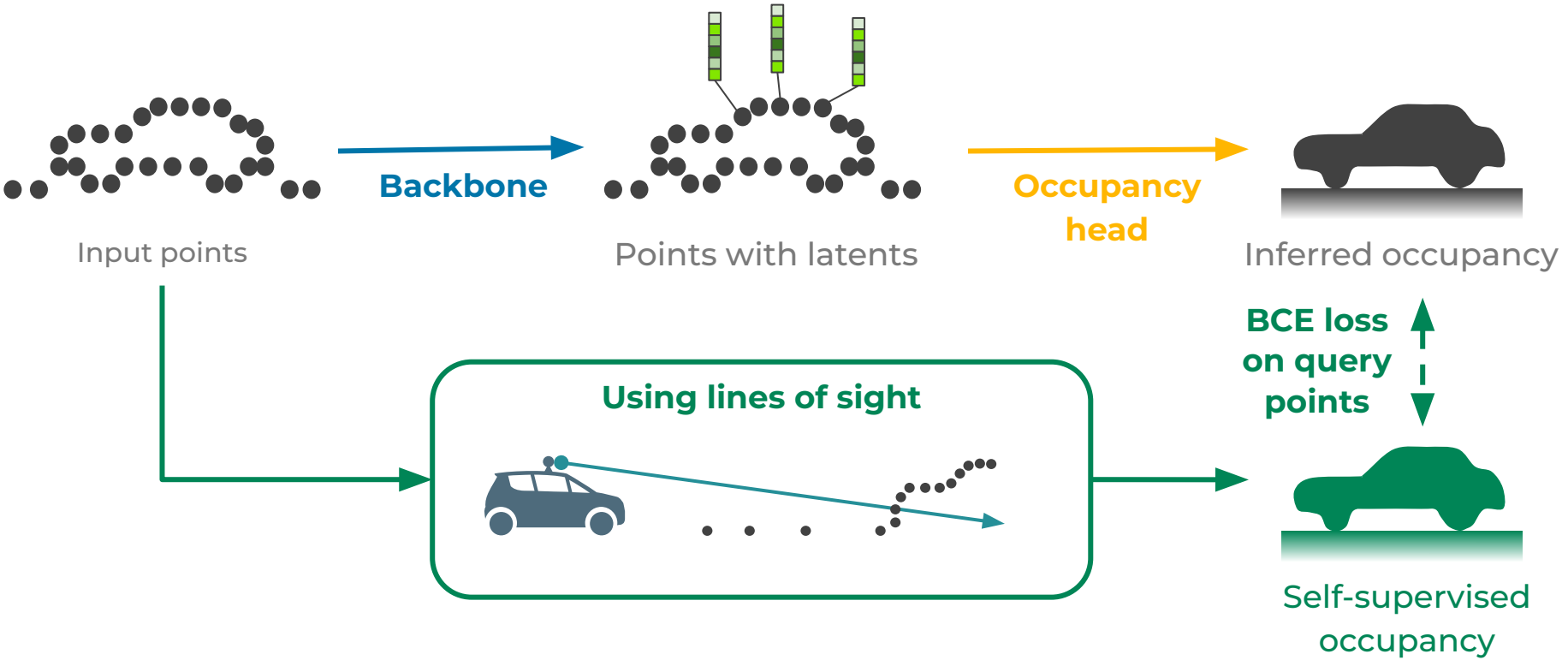
Along lidar lines of sight

- **Empty queries:** from sensor to observed point
- **Full queries:** just behind the point (max distance δ = 0.1 m)

A. Boulch et al., ALSO: Automotive Lidar Self-Supervision by Occupancy Estimation, CVPR 2023

Automotive Lidar Self-Supervision by Occupancy Estimation

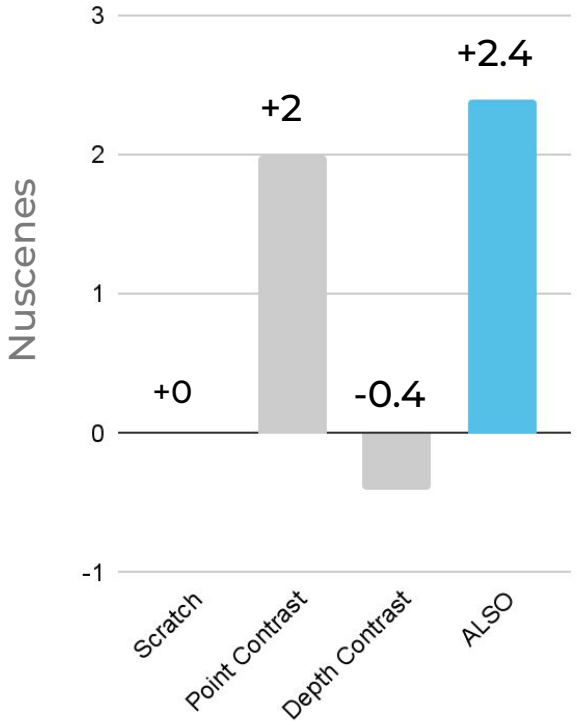
Pretext task

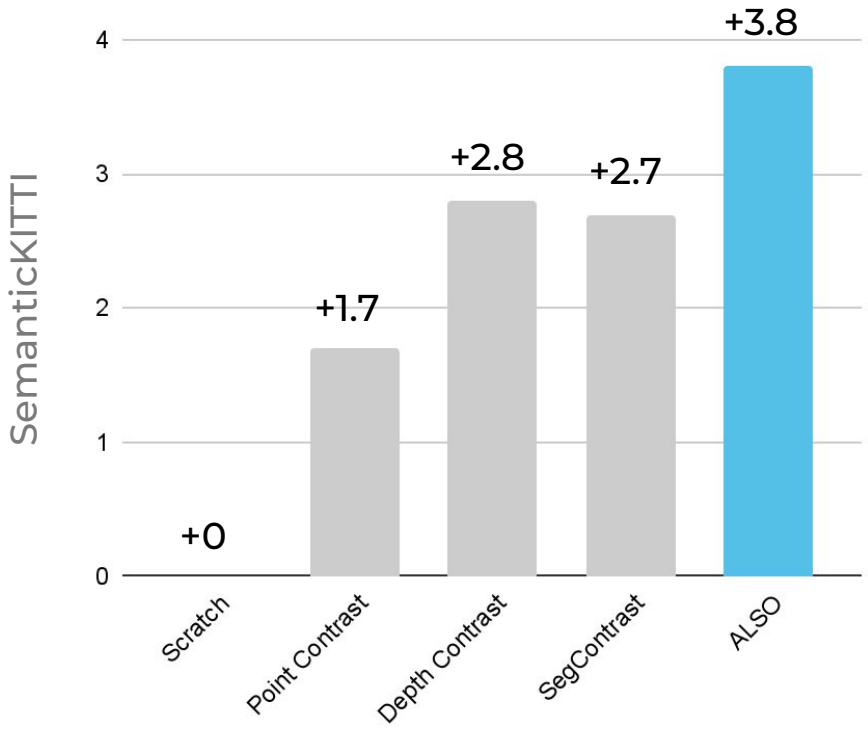


A. Boulch et al., ALSO: Automotive Lidar Self-Supervision by Occupancy Estimation, CVPR 2023

Downstream performance

1% annotated training data

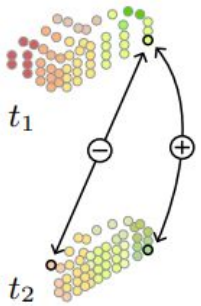




A. Boulch et al., ALSO: Automotive Lidar Self-Supervision by Occupancy Estimation, CVPR 2023

Contrastive self-supervised learning for point clouds

BEVContrast



PointContrast

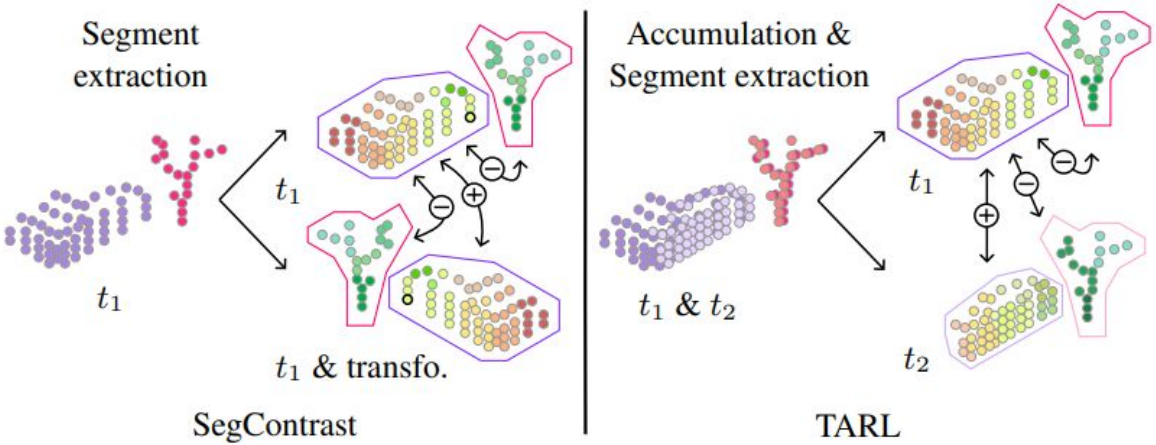
PointContrast

+++ Easy to implement
- - - Contrast in the same object

C. Sautier et al., BEVContrast: Self-Supervision in BEV Space for Automotive Lidar Point Clouds, 3DV 2024

Contrastive self-supervised learning for point clouds

BEVContrast



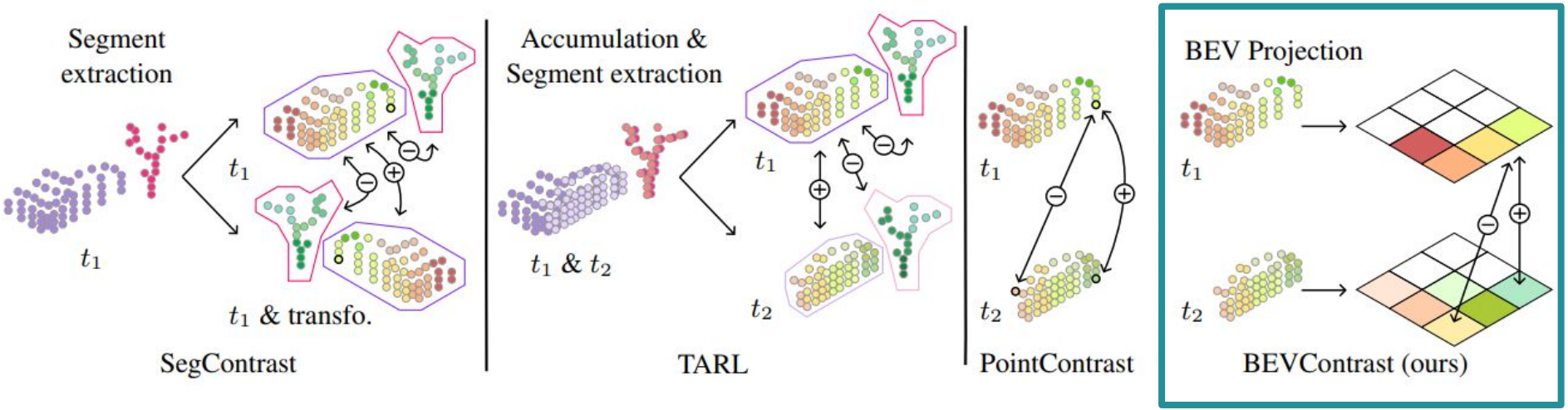
SegContrast / TARL

+++ Efficient thanks to object segmentation (temporal for TARL)
- - - Difficult to set up → rely on HDBScan (hyperparameters)

C. Sautier et al., BEVContrast: Self-Supervision in BEV Space for Automotive Lidar Point Clouds, 3DV 2024

Contrastive self-supervised learning for point clouds

BEVContrast



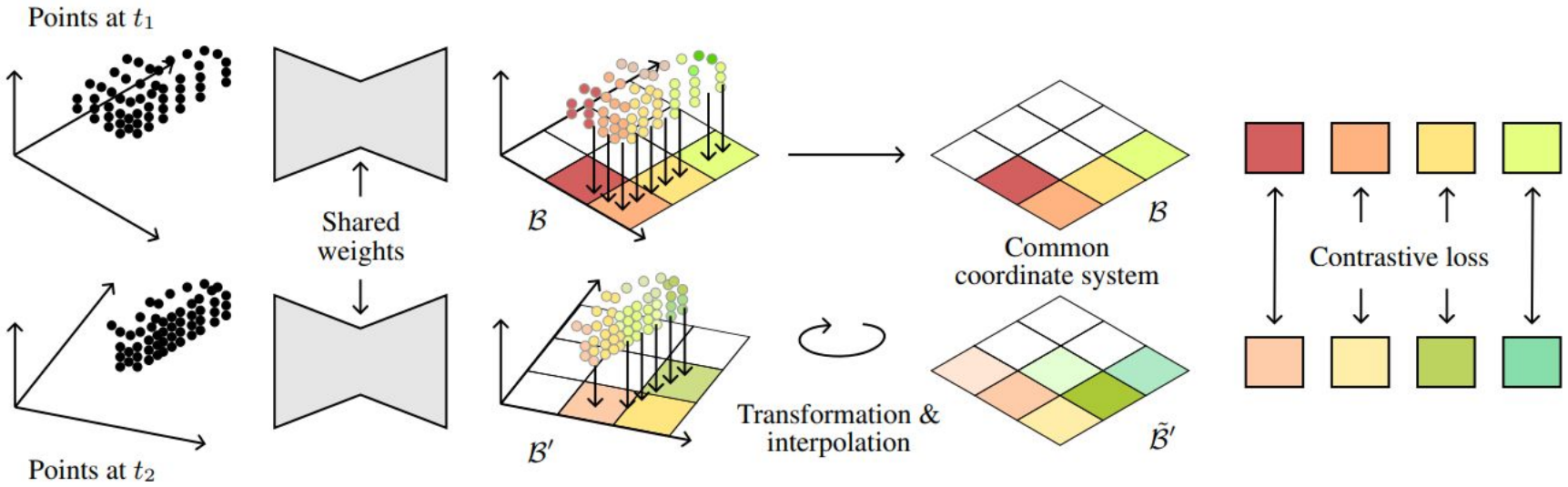
BEVContrast

+++ Simple → easy projection in BEV
+++ Object separation approximation with BEV cells

C. Sautier et al., BEVContrast: Self-Supervision in BEV Space for Automotive Lidar Point Clouds, 3DV 2024

Contrastive self-supervised learning for point clouds

BEVContrast



+++ Simple → easy projection in BEV
+++ Object separation approximation with BEV cells

C. Sautier et al., BEVContrast: Self-Supervision in BEV Space for Automotive Lidar Point Clouds, 3DV 2024

BEVContrast

Semantic segmentation: qualitative results

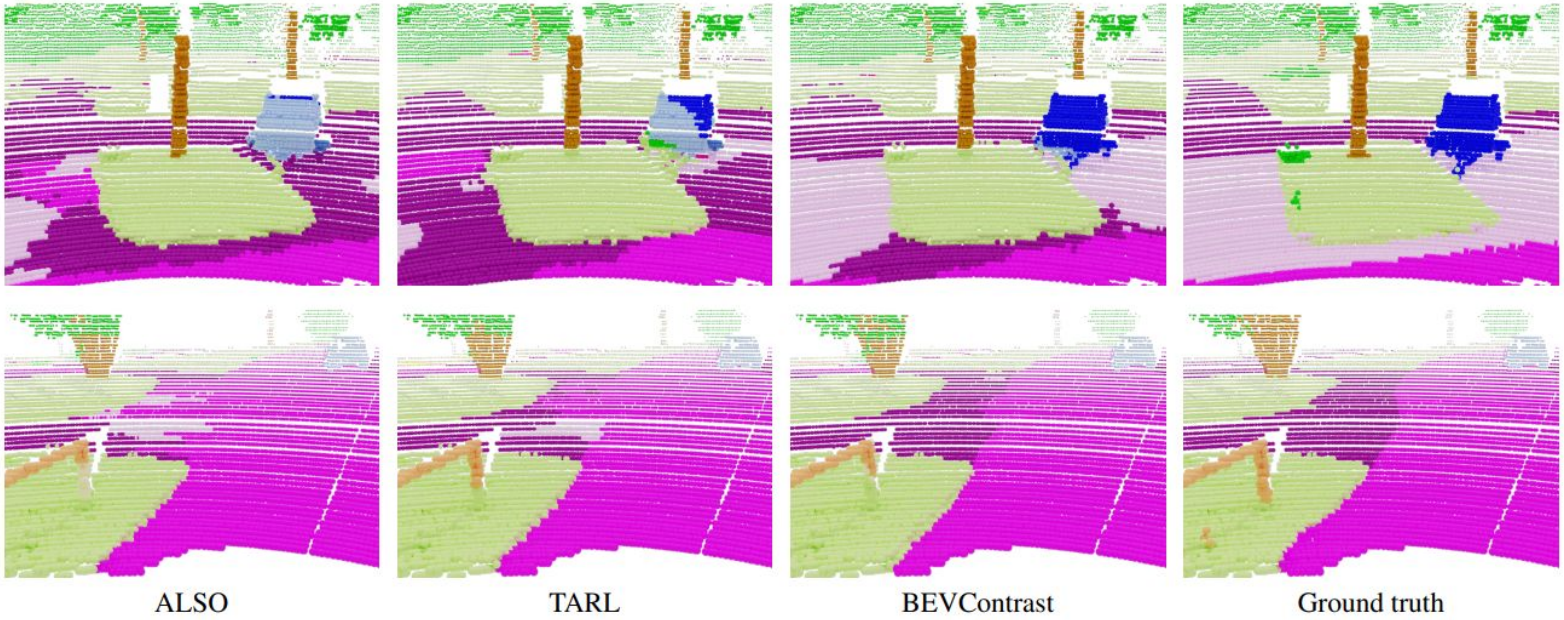


Figure 3. **Semantic segmentation visualizations on SemanticKITTI** after pre-training on the full training dataset and fine-tuning using 1% of the annotated data. car, other veh., sidewalk, drivable surf., parking, terrain, vegetation, trunk, fence

---

C. Sautier et al., BEVContrast: Self-Supervision in BEV Space for Automotive Lidar Point Clouds, 3DV 2024

BEVContrast

Semantic segmentation: quantitative results

Dataset	Method	0.1%		1%		10%		50%		100%	
nuScenes	No pre-training	21.6	±0.5	35.0	±0.3	57.3	±0.4	69.0	±0.2	71.2	±0.2
	PointContrast† [40]	27.1	±0.5	37.0	±0.5	58.9	±0.2	69.4	±0.3	71.1	±0.2
	DepthContrast†[46]	21.7	±0.3	34.6	±0.5	57.4	±0.5	69.2	±0.3	71.2	±0.2
	ALSO [3]	26.2	±0.5	37.4	±0.3	59.0	±0.4	69.8	±0.2	71.8	±0.2
	BEVContrast (ours)	26.6	±0.5	37.9	±0.4	59.0	±0.6	70.5	±0.2	72.2	±0.1
SemanticKITTI	No pre-training	30.0	±0.2	46.2	±0.6	57.6	±0.9	61.8	±0.4	62.7	±0.3
	PointContrast‡ [40]	32.4	±0.5	47.9	±0.5	59.7	±0.5	62.7	±0.3	63.4	±0.4
	SegContrast [29]	32.3	±0.3	48.9	±0.3	58.7	±0.5	62.1	±0.4	62.3	±0.4
	DepthContrast† [46]	32.5	±0.4	49.0	±0.4	60.3	±0.5	62.9	±0.5	63.9	±0.4
	STSSL [39]	32.0	±0.4	49.4	±1.1	60.0	±0.6	62.9	±0.7	63.3	±0.3
	ALSO [3]	35.0	±0.1	50.0	±0.4	60.5	±0.1	63.4	±0.5	63.6	±0.5
	TARL [30]	37.9	±0.4	52.5	±0.5	61.2	±0.3	63.4	±0.2	63.7	±0.3
	BEVContrast (ours)	39.7	±0.9	53.8	±1.0	61.4	±0.4	63.4	±0.6	64.1	±0.4

C. Sautier et al., BEVContrast: Self-Supervision in BEV Space for Automotive Lidar Point Clouds, 3DV 2024

One model from many for all

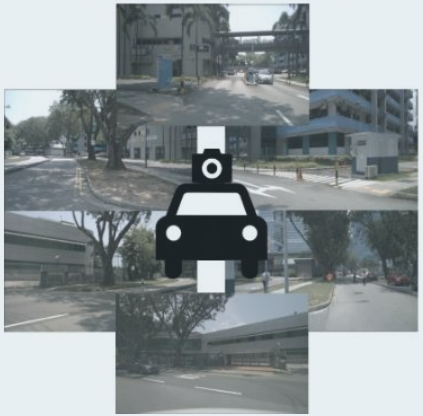
Multi-modal training

Lidar-to-Image models

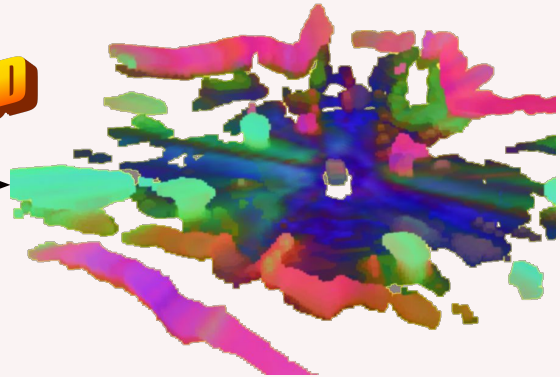
Open-Vocabulary 3D Occupancy Prediction from Images

Goals

INPUT: surround-view images

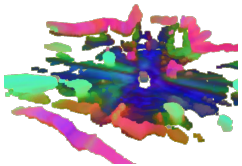


POP3D



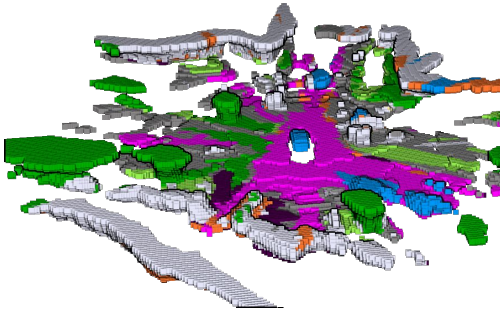
OUTPUT:
3D voxel field with:
- occupancy
- open-vocabulary features

TASK #1: zero-shot semantic occupancy segmentation



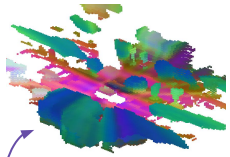
+ class names

road, car, terrain, vegetation, building, ...



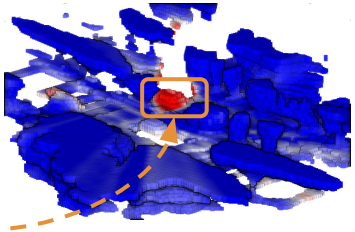
TASK #2: text-driven 3D retrieval from cameras

POP3D



similarity

"Black hatchback"

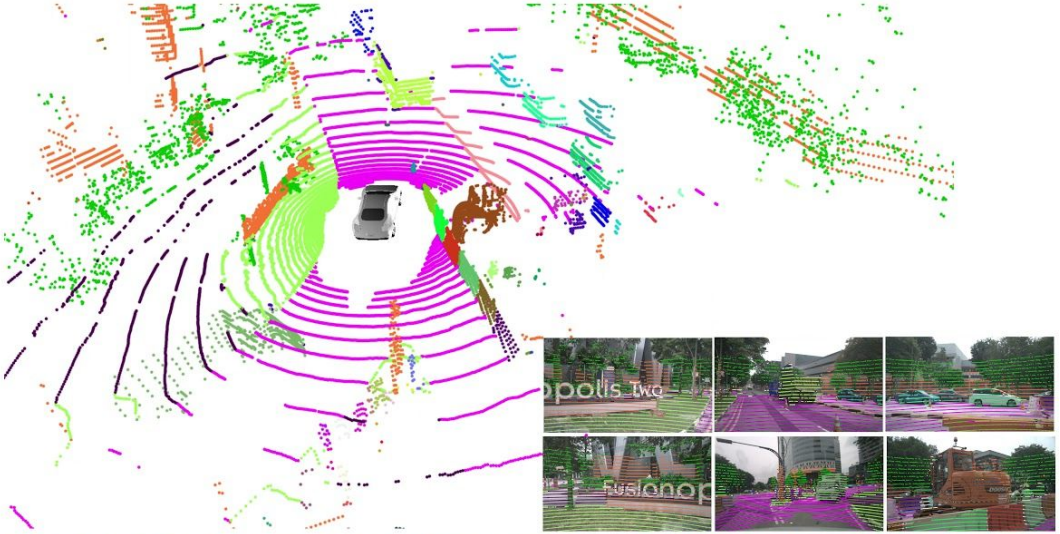


A. Vobecky et al., POP-3D: Open-Vocabulary 3D Occupancy Prediction from Images, NeurIPS 2023

Open-Vocabulary 3D Occupancy Prediction from Images

Motivation

- Only sparse LiDAR annotations available (costly and difficult to scale).
- Annotated labels for a predefined closed-vocabulary.



A. Vobecky et al., POP-3D: Open-Vocabulary 3D Occupancy Prediction from Images, NeurIPS 2023

Open-Vocabulary 3D Occupancy Prediction from Images

Overview

open-vocabulary 3D semantic occupancy prediction

Training:

unlabeled image-LiDAR data and a **pre-trained image-language model**

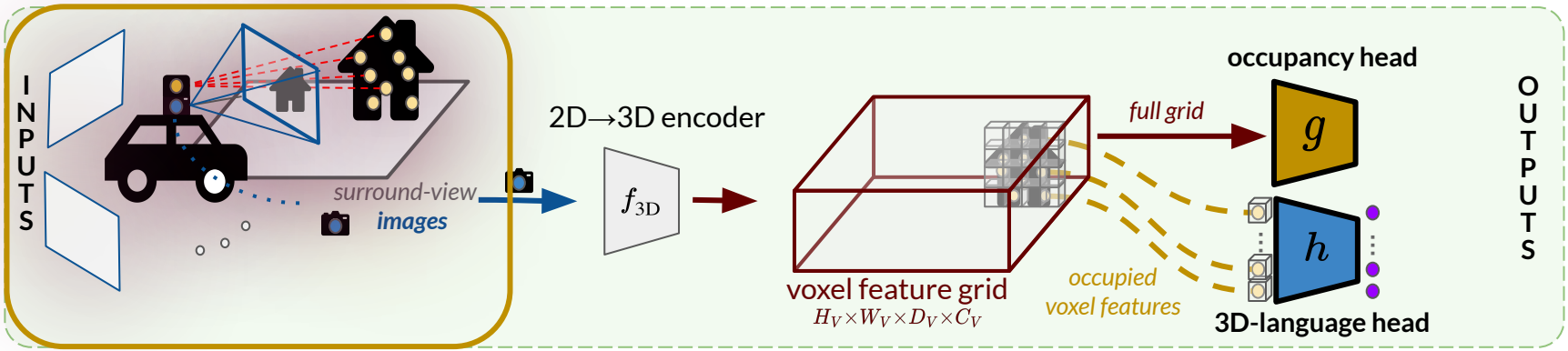
Inference:

only images + text queries

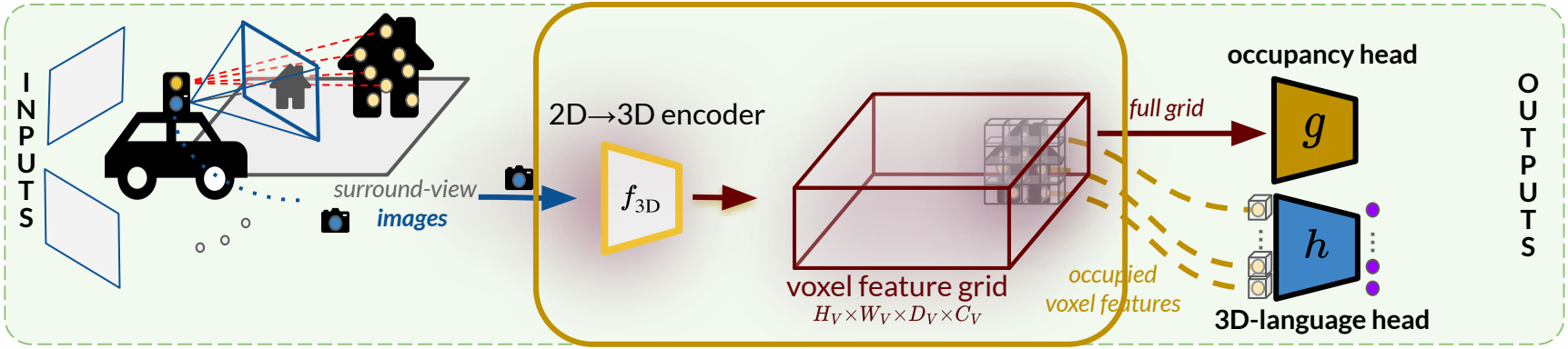
Example of text-driven retrieval

A. Vobecky et al., POP-3D: Open-Vocabulary 3D Occupancy Prediction from Images, NeurIPS 2023

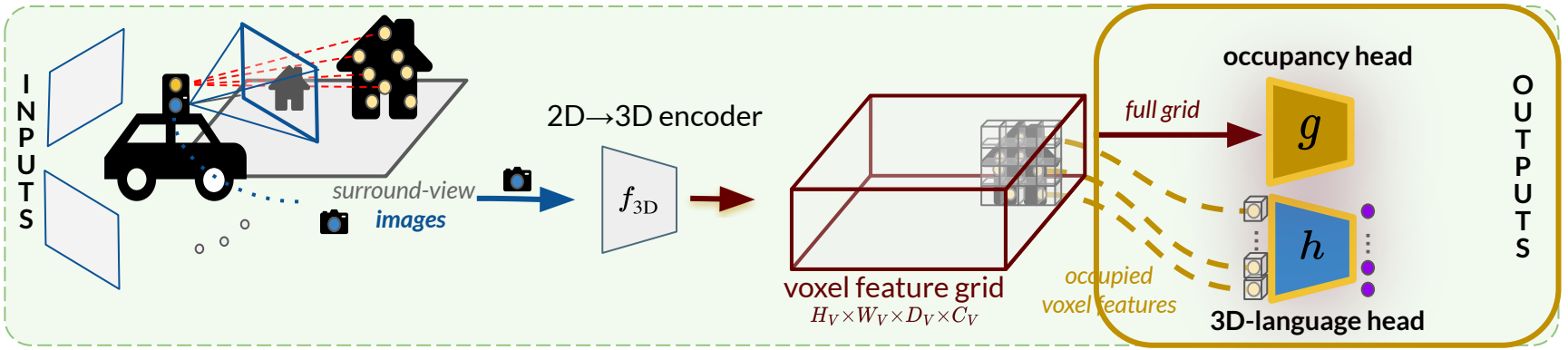
POP-3D Architecture



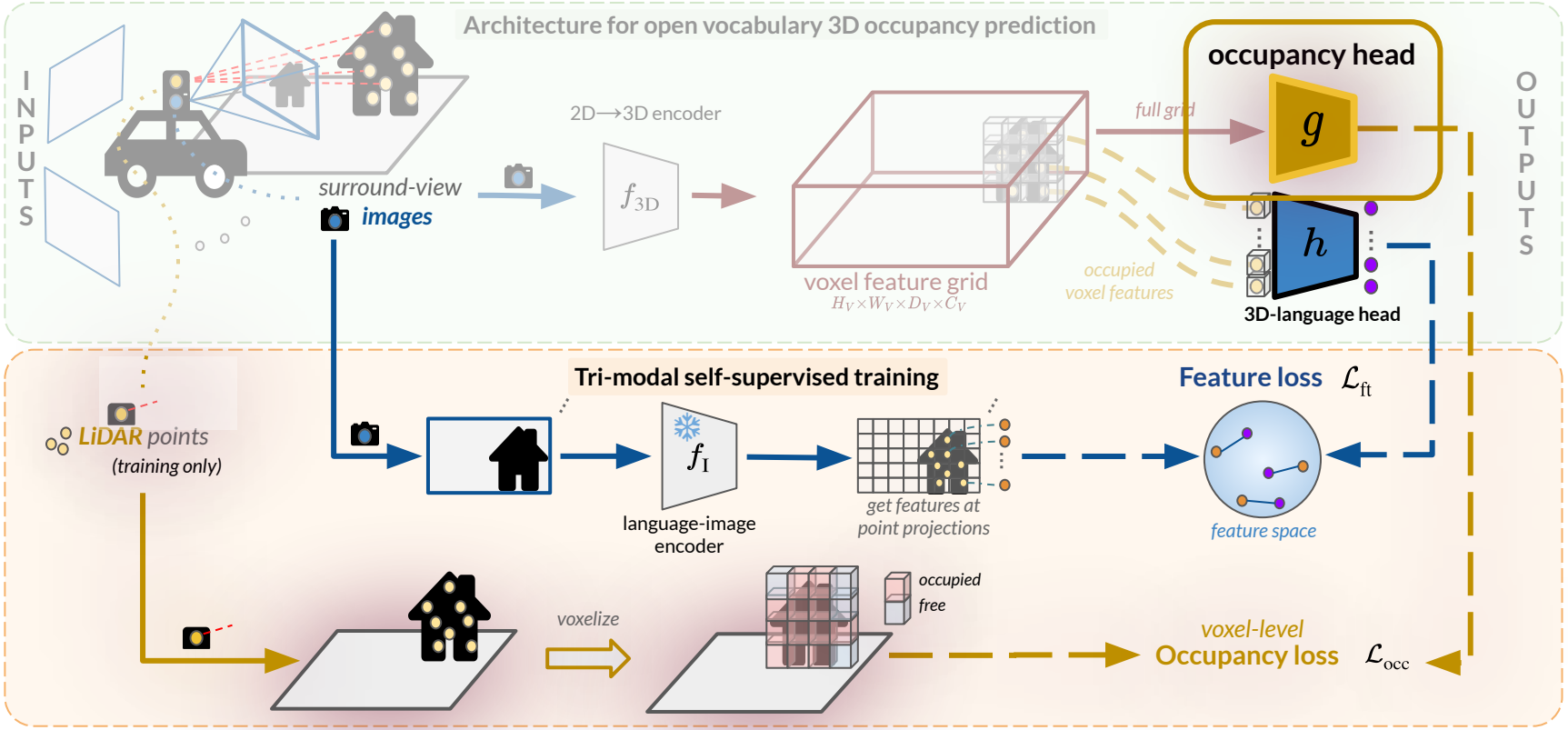
POP-3D Architecture



POP-3D Architecture

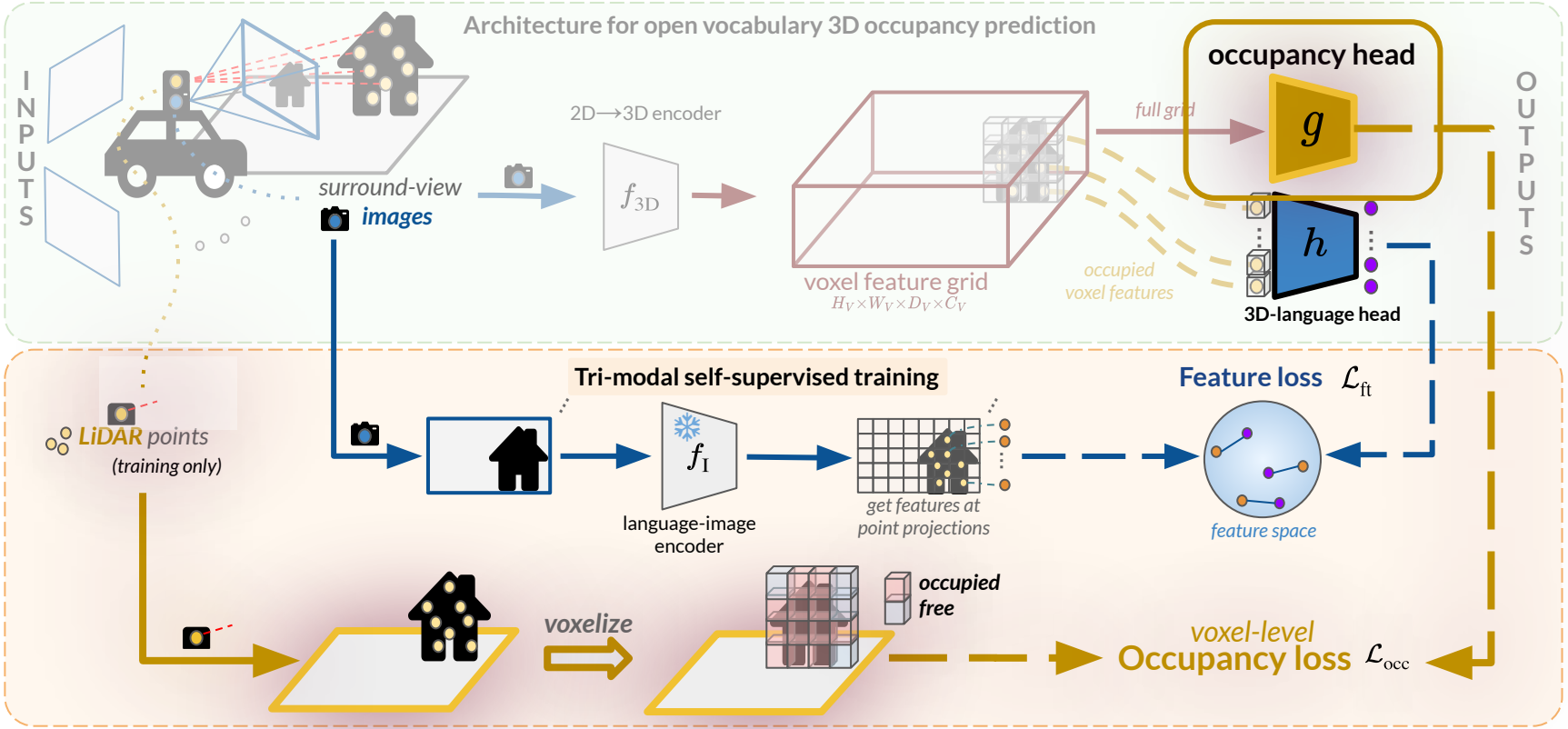


POP-3D Losses

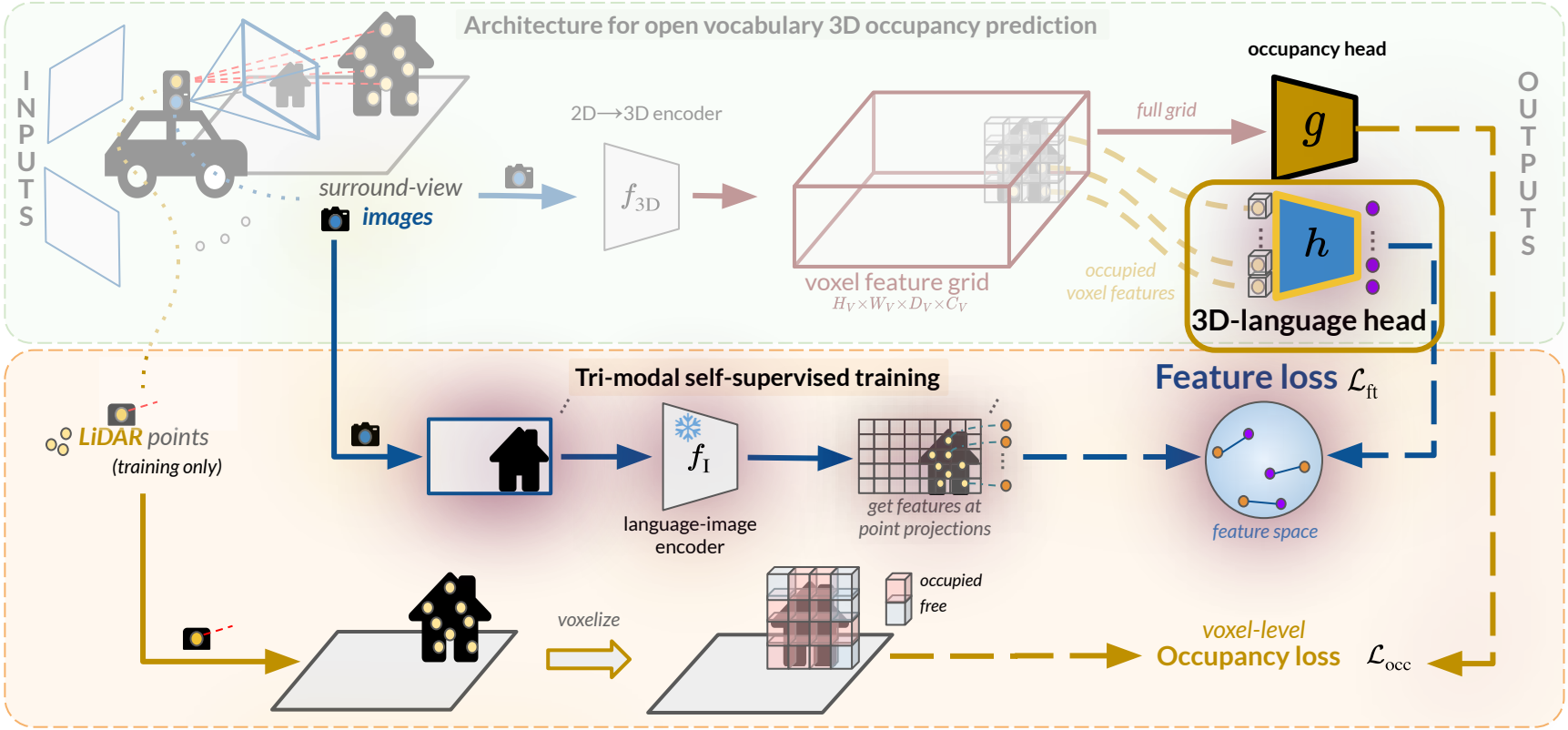


Feature loss: MSE loss **Occupancy loss**: Lovasz and CE losses

POP-3D Losses

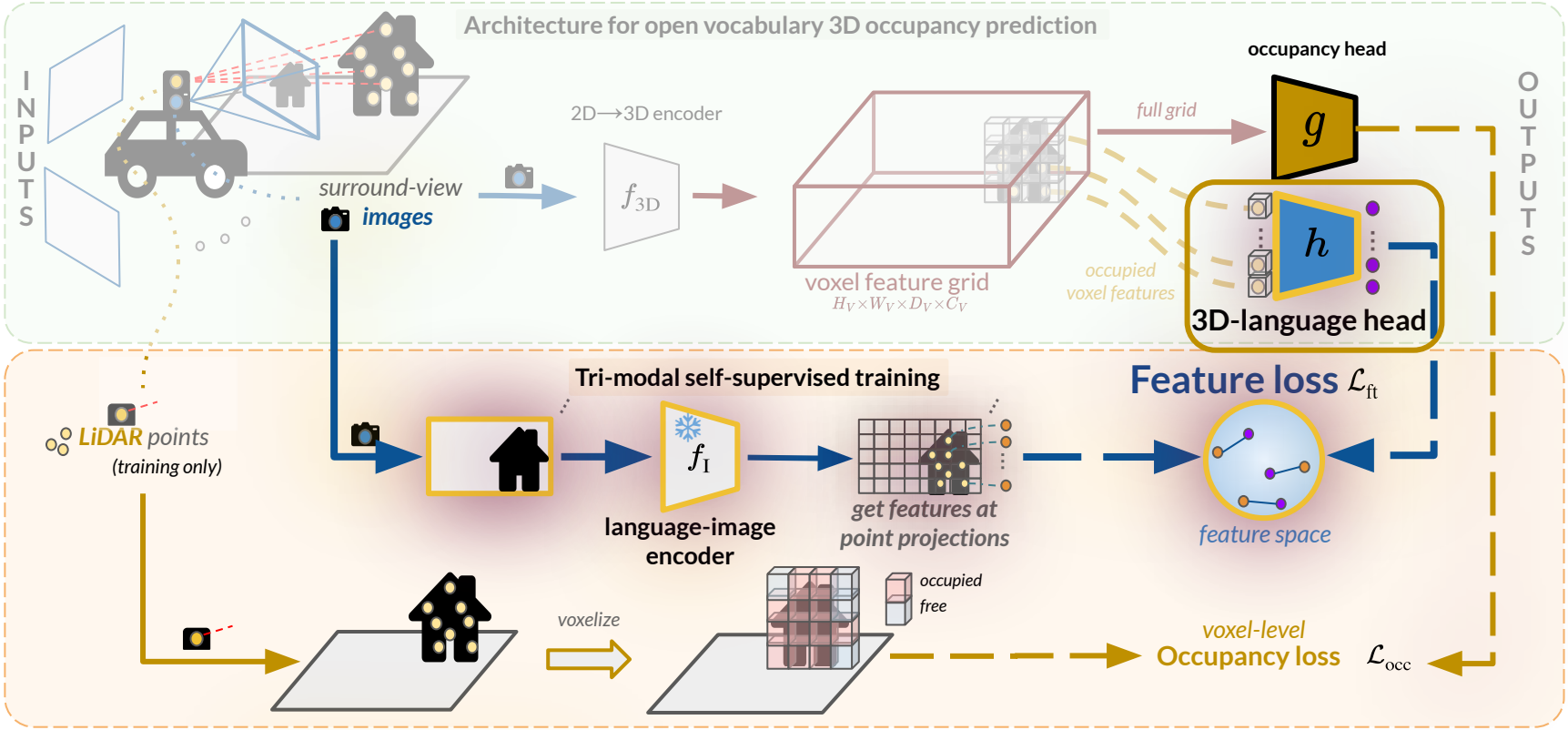


POP-3D Losses



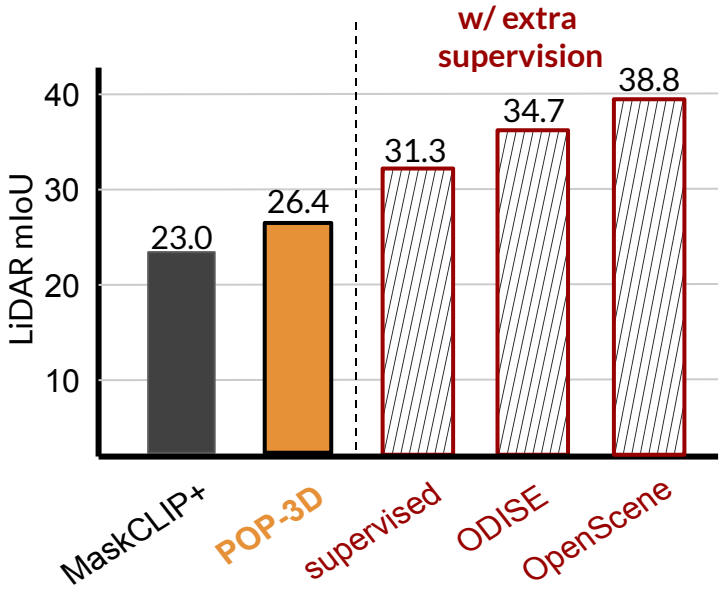
Feature loss: MSE loss **Occupancy loss**: Lovasz and CE losses

POP-3D Losses

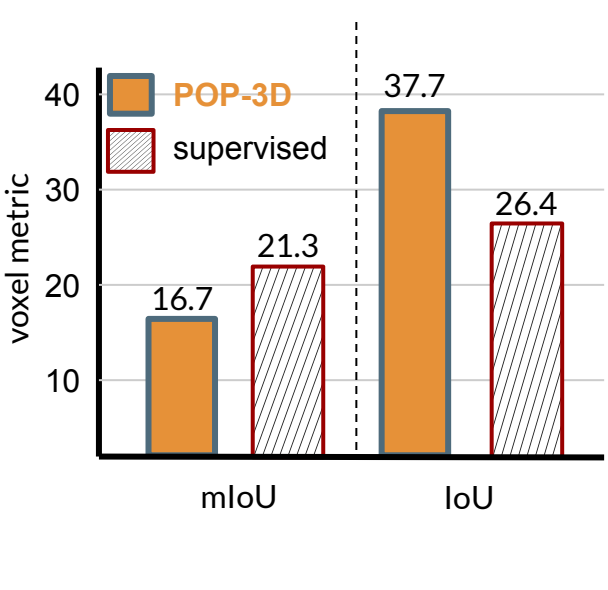


Feature loss: MSE loss **Occupancy loss**: Lovasz and CE losses

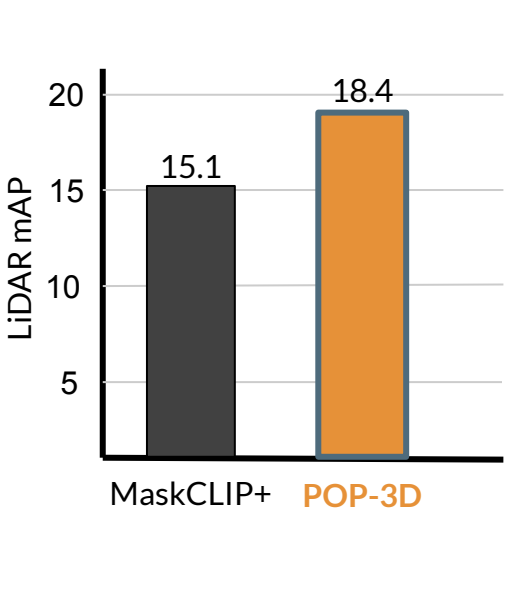
POP-3D Quantitative results (nuscenes)



LiDAR-based evaluation



3D occupancy prediction

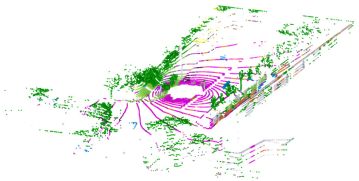


Retrieval (mAP)

Predict

Occupancy: ❌

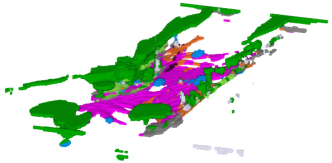
Semantics: ✅



Predict

Occupancy: ✅

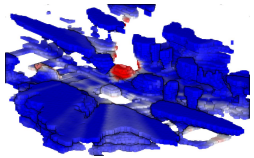
Semantics: ✅



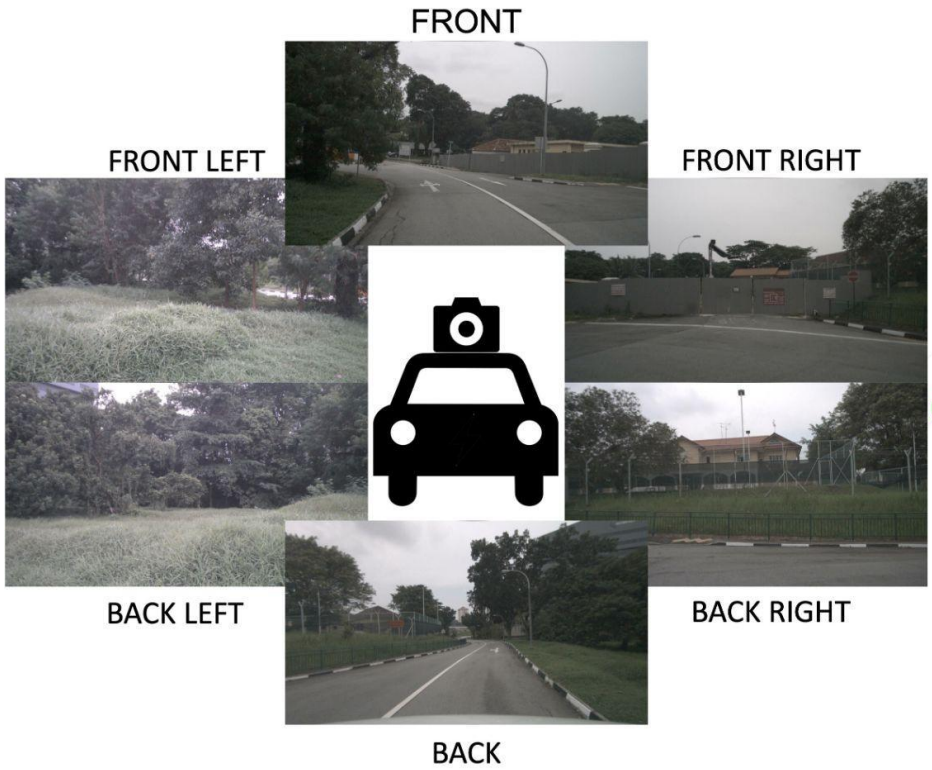
Predict

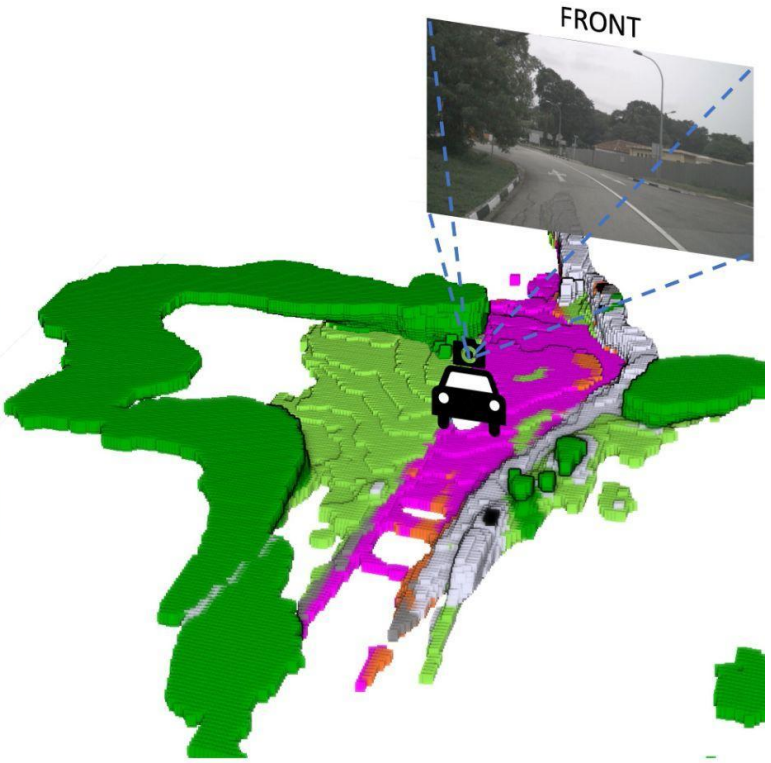
Occupancy: ❌

Semantics: ✅
(relevancy)



POP-3D Qualitative results: zero-shot semantic segmentation

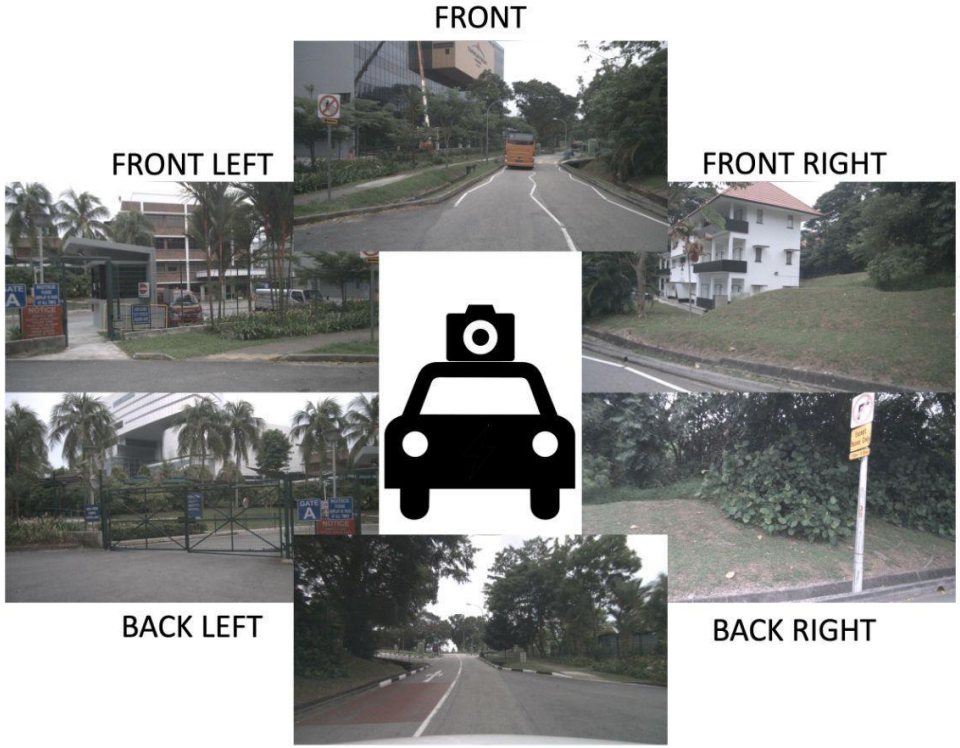


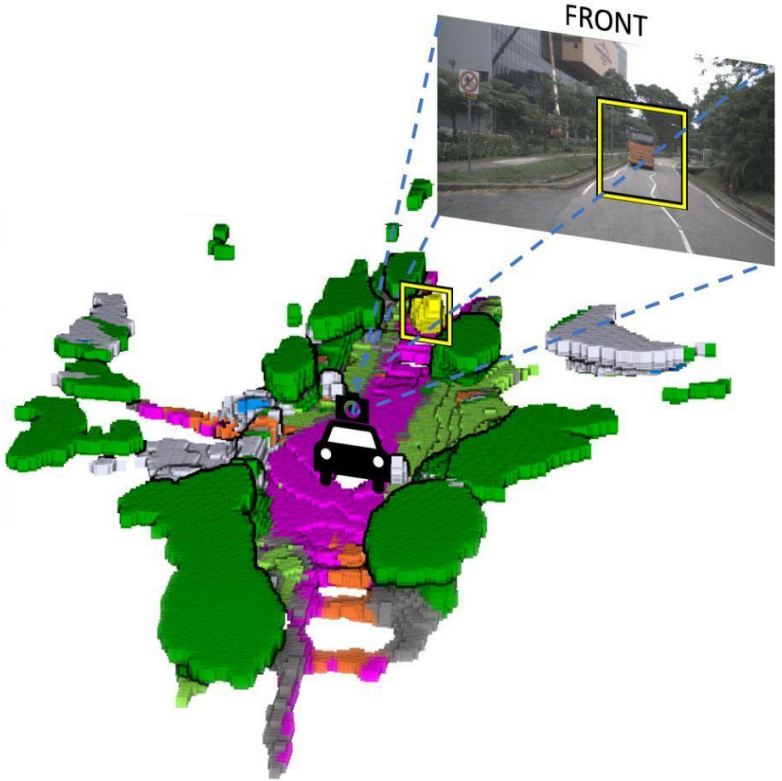


driveable surface | car | bus | truck | terrain | vegetation | sidewalk | other flat | pedestrian | bicycle

manmade | motorcycle | barrier | construction vehicle | trailer | traffic cone

POP-3D Qualitative results: zero-shot semantic segmentation

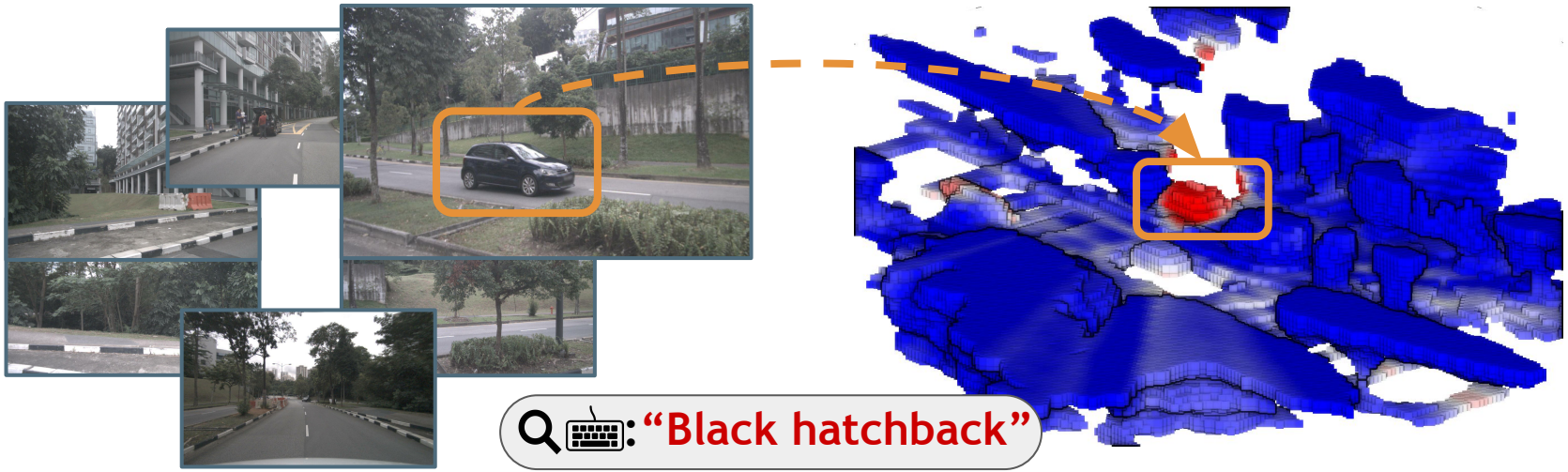




driveable surface | car | bus | truck | terrain | vegetation | sidewalk | other flat | pedestrian | bicycle

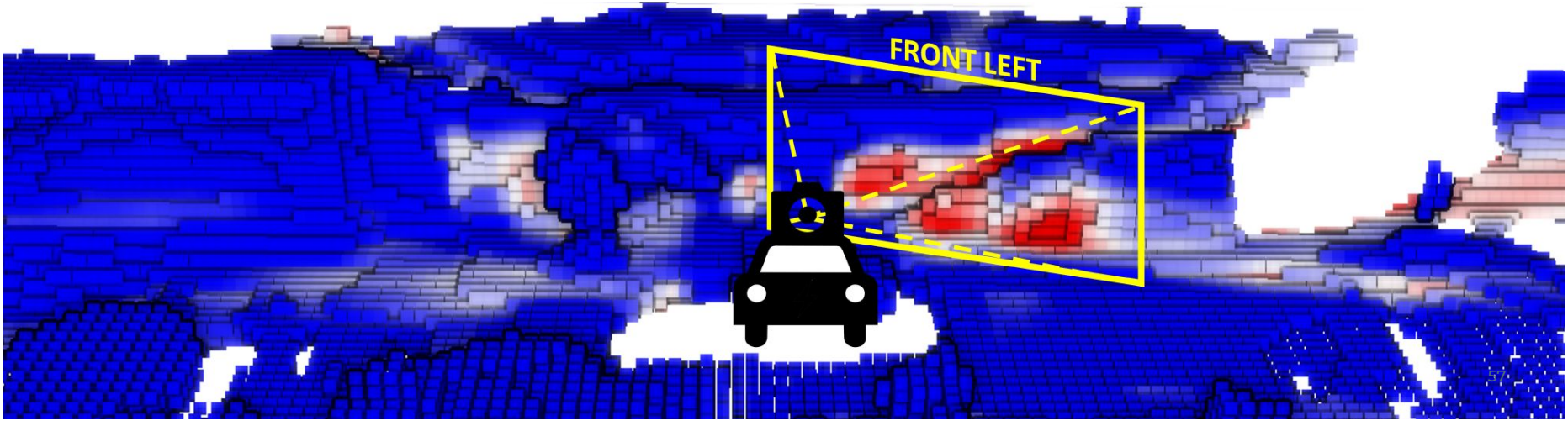
manmade | motorcycle | barrier | construction vehicle | trailer | traffic cone

POP-3D Qualitative results: retrieval



POP-3D Qualitative results: retrieval

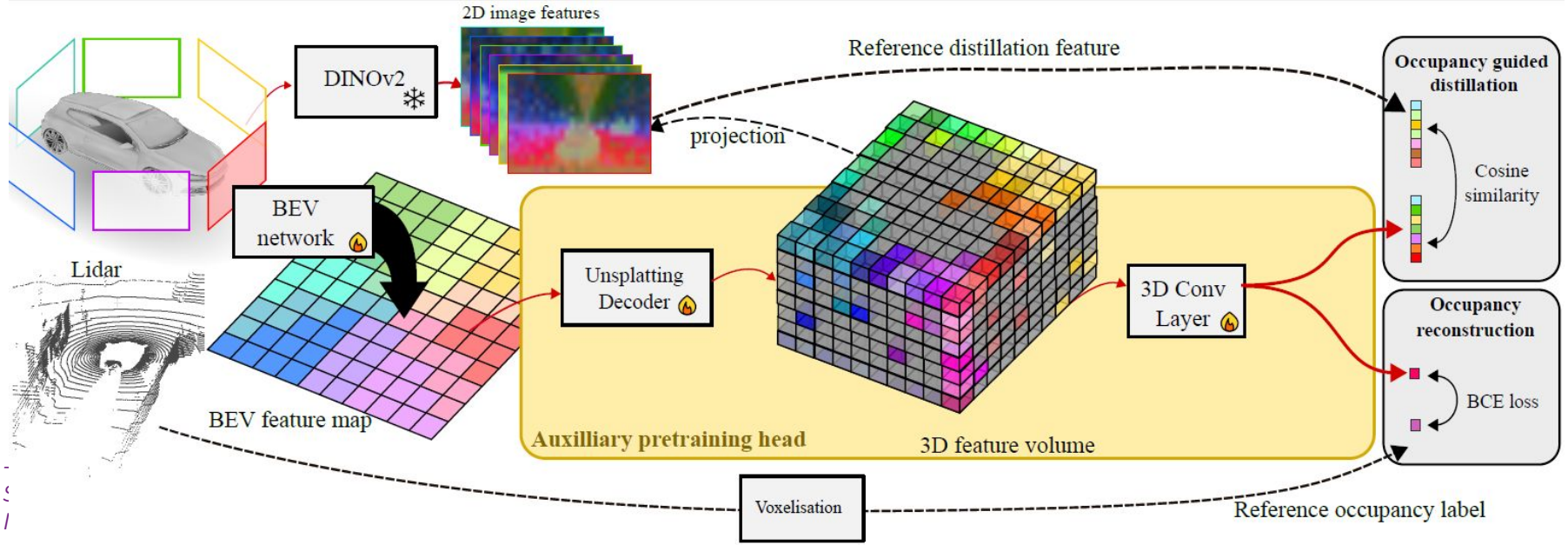
: “stairs”



Self-supervised Occupancy Feature Prediction for Pretraining BEV

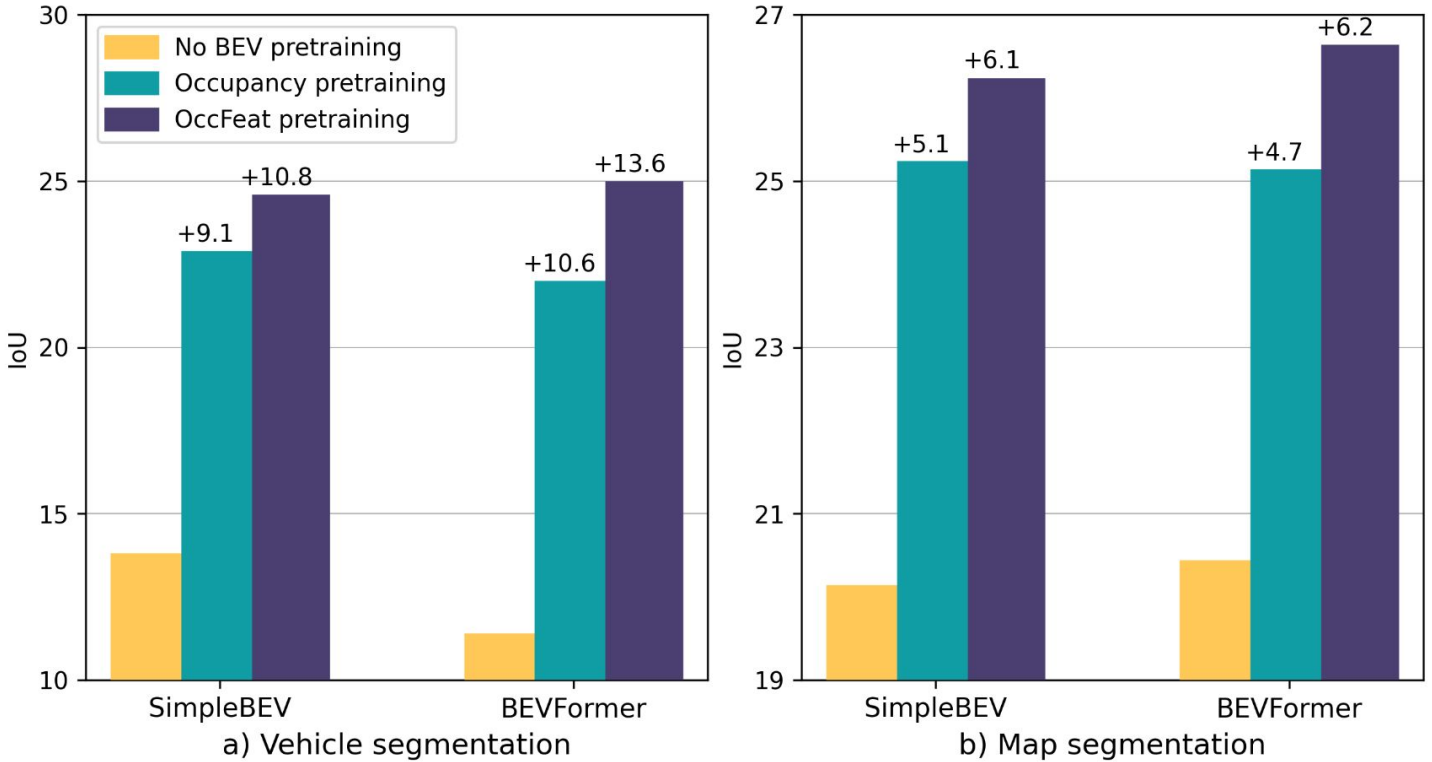
OccFeat

- Architecture-agnostic pretraining that teaches the model semantic-aware 3D geometry information
- Leveraging aligned Image & LiDAR data pretrained SSL image encoder (DINOv2)



Downstream performance - 1% labelled samples (nuscenes)

OccFeat

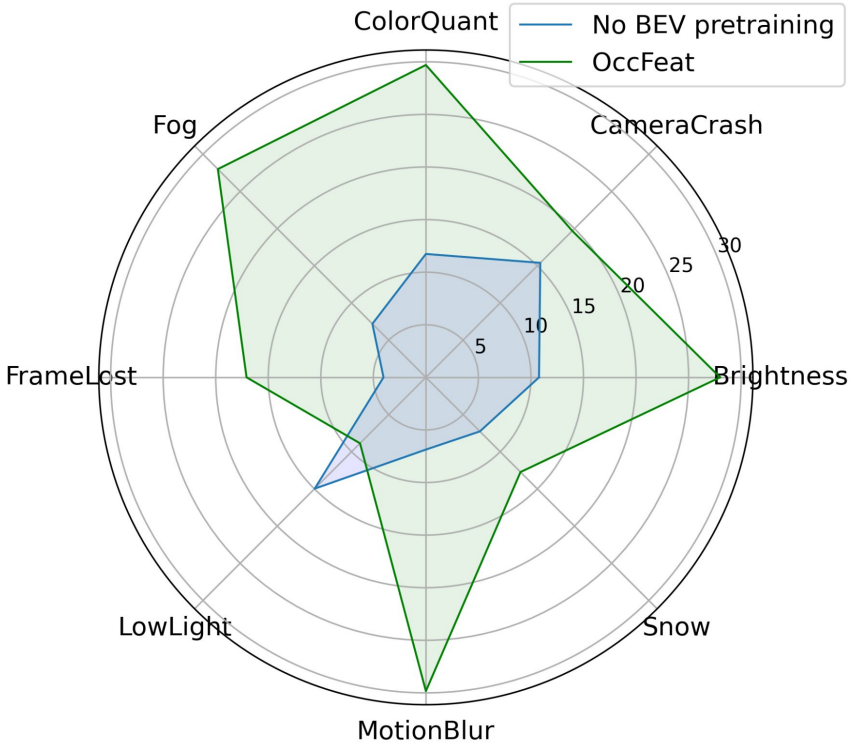


S. Sirko-Galouchenko et al., OccFeat: Self-supervised Occupancy Feature Prediction for Pretraining BEV Segmentation Networks, CVPR WAD 2024

Robustness to distribution shift

OccFeat

- Segmentation results on nuScenes-C dataset (8 corruptions on nuScenes validation set)
- OccFeat vs. no BEV pretraining



S. Sirko-Galouchenko et al., OccFeat: Self-supervised Occupancy Feature Prediction for Pretraining BEV Segmentation Networks, CVPR WAD 2024

Image-to-Lidar models

Image-to-Lidar distillation

How SoTA methods work

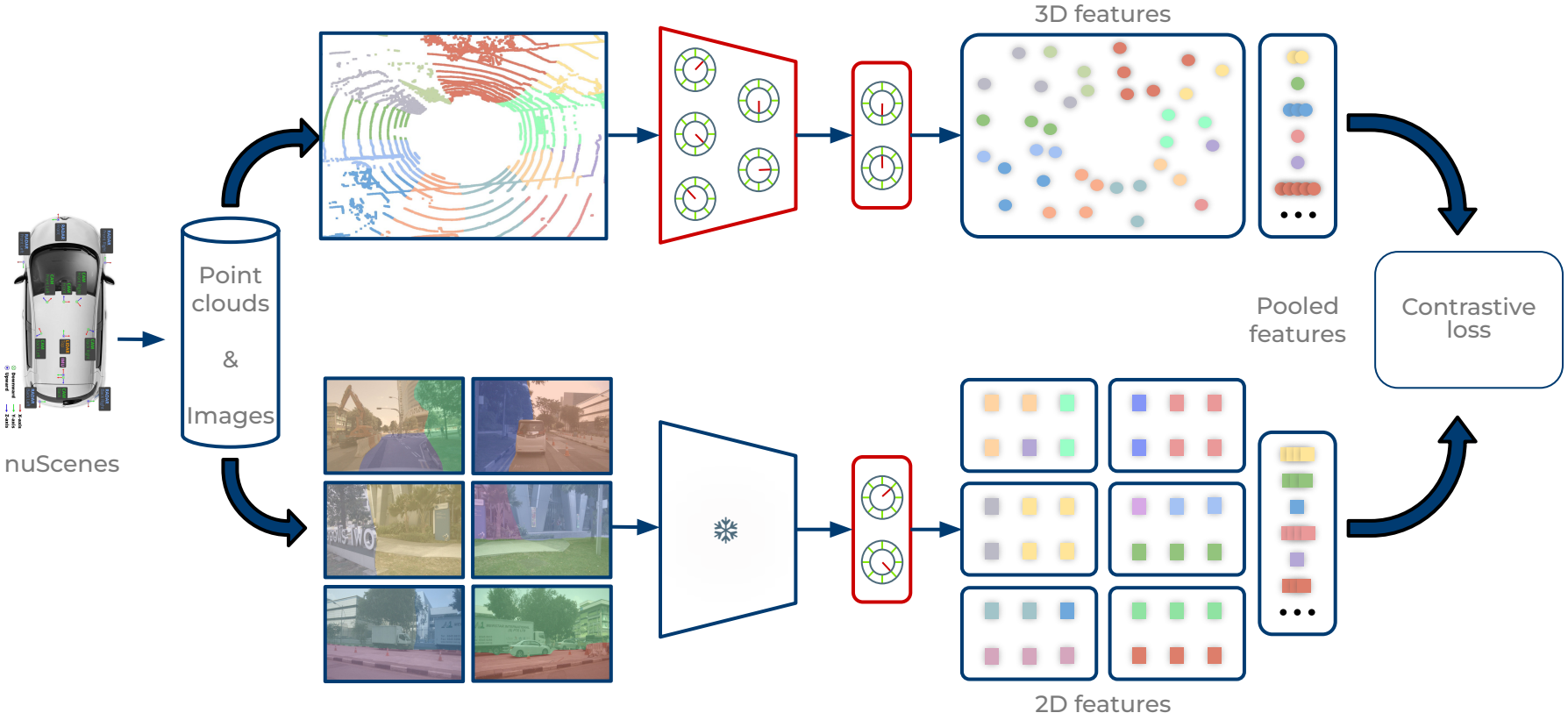
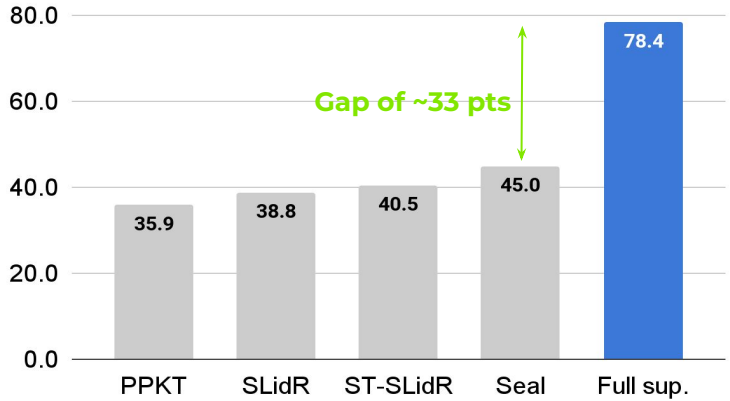
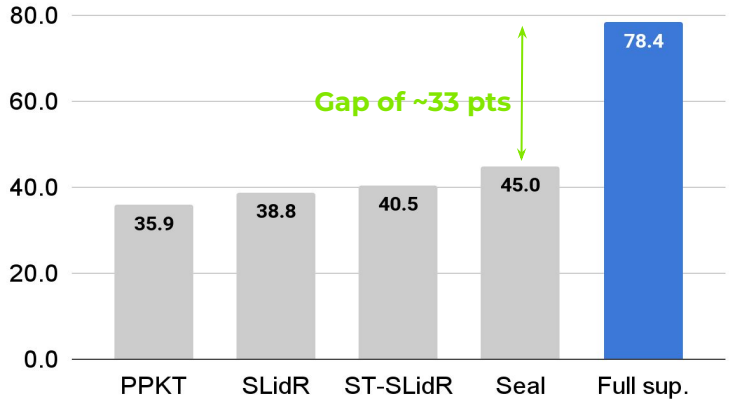


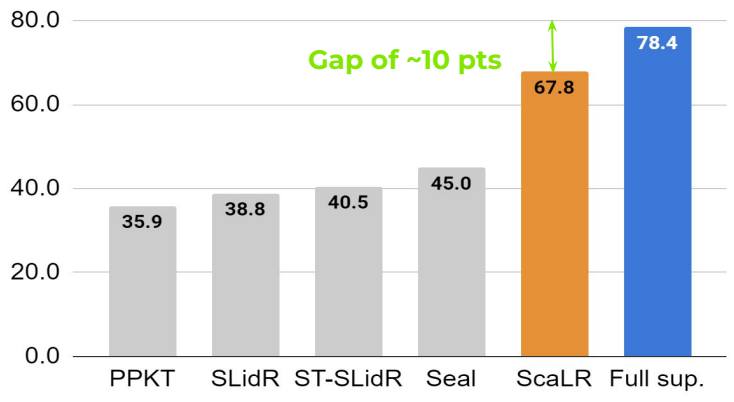
Image-to-Lidar distillation



Motivation:
Distilled 3D features << supervised 3D features

Image-to-Lidar distillation



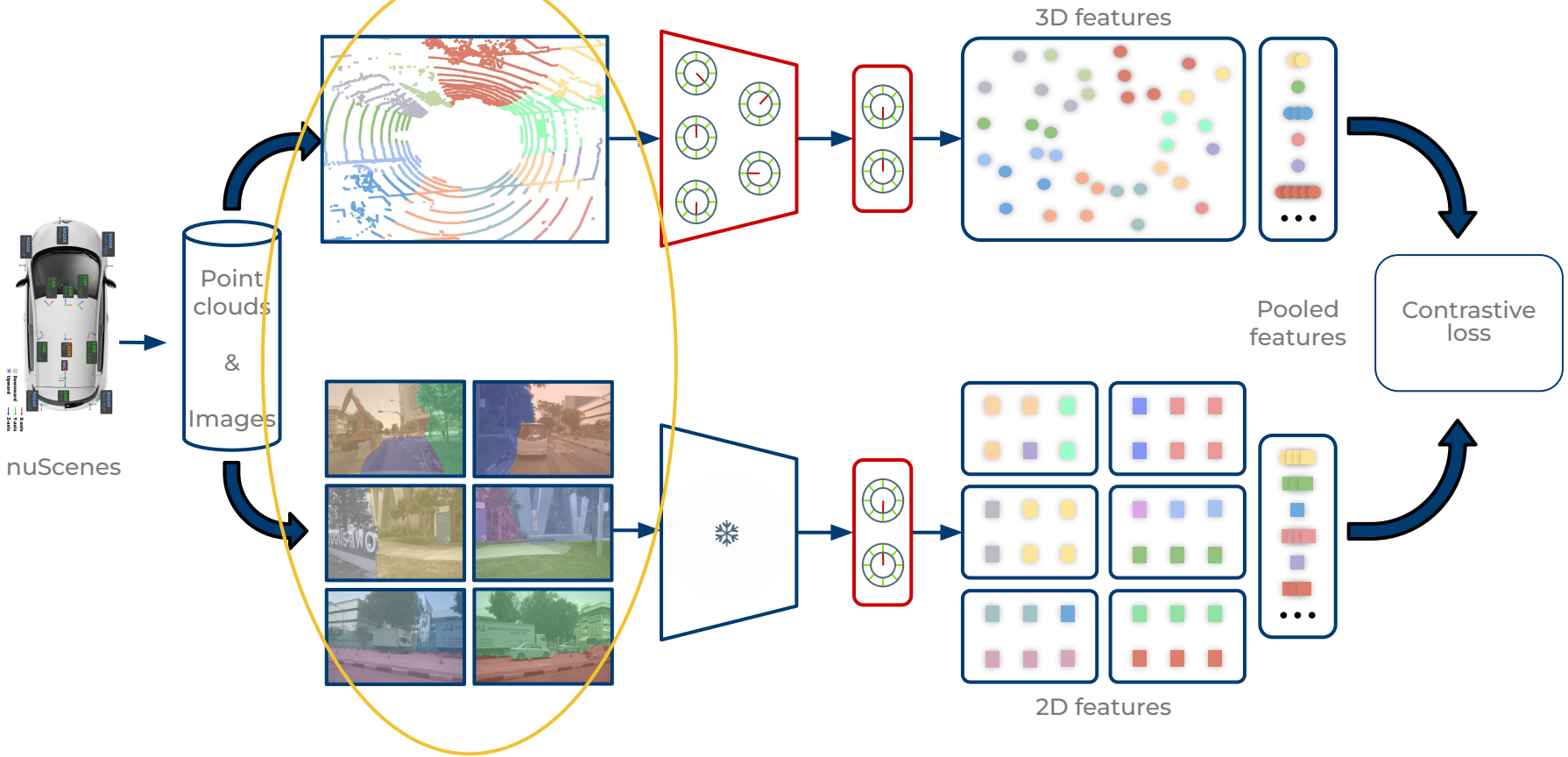


Motivation:
Distilled 3D features << supervised 3D features

Main result:
Large improvement of feature quality: +22.8 mIoU pts

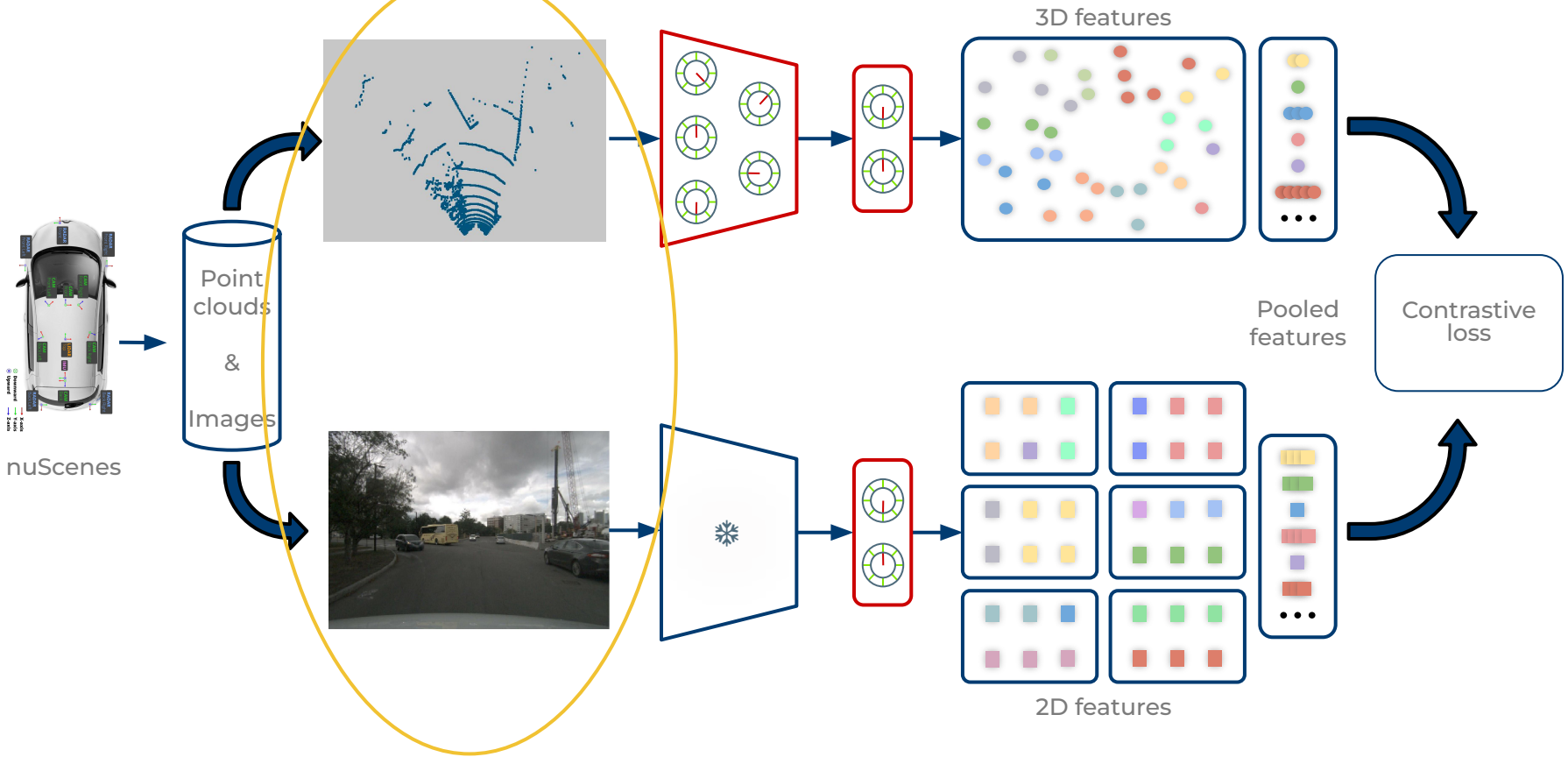
With ScaLR we revisit this pipeline and propose the following simplifications:

Simplifications done in ScaLR



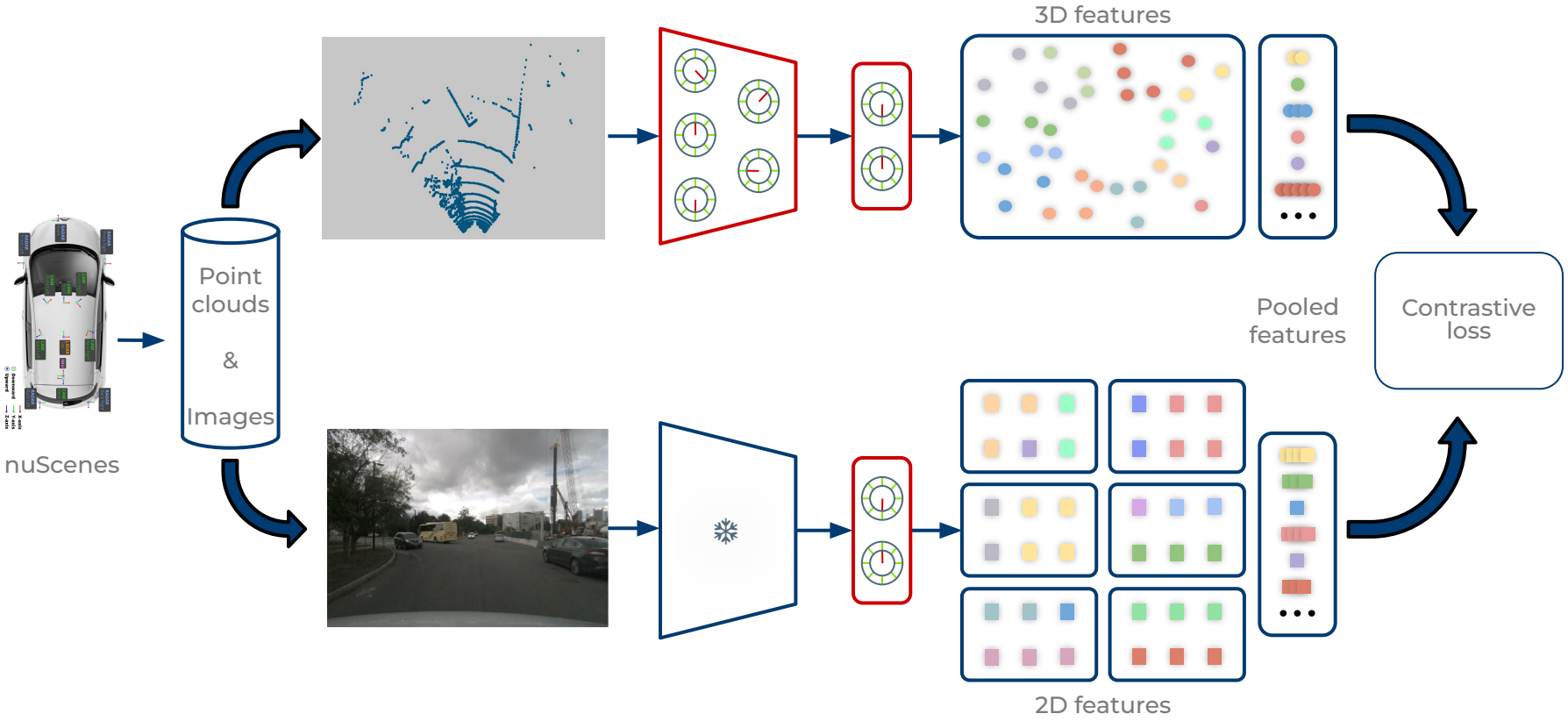
G. Puy et al., Three Pillars improving Vision Foundation Model Distillation for Lidar, CVPR 2024

Simplifications done in ScaLR



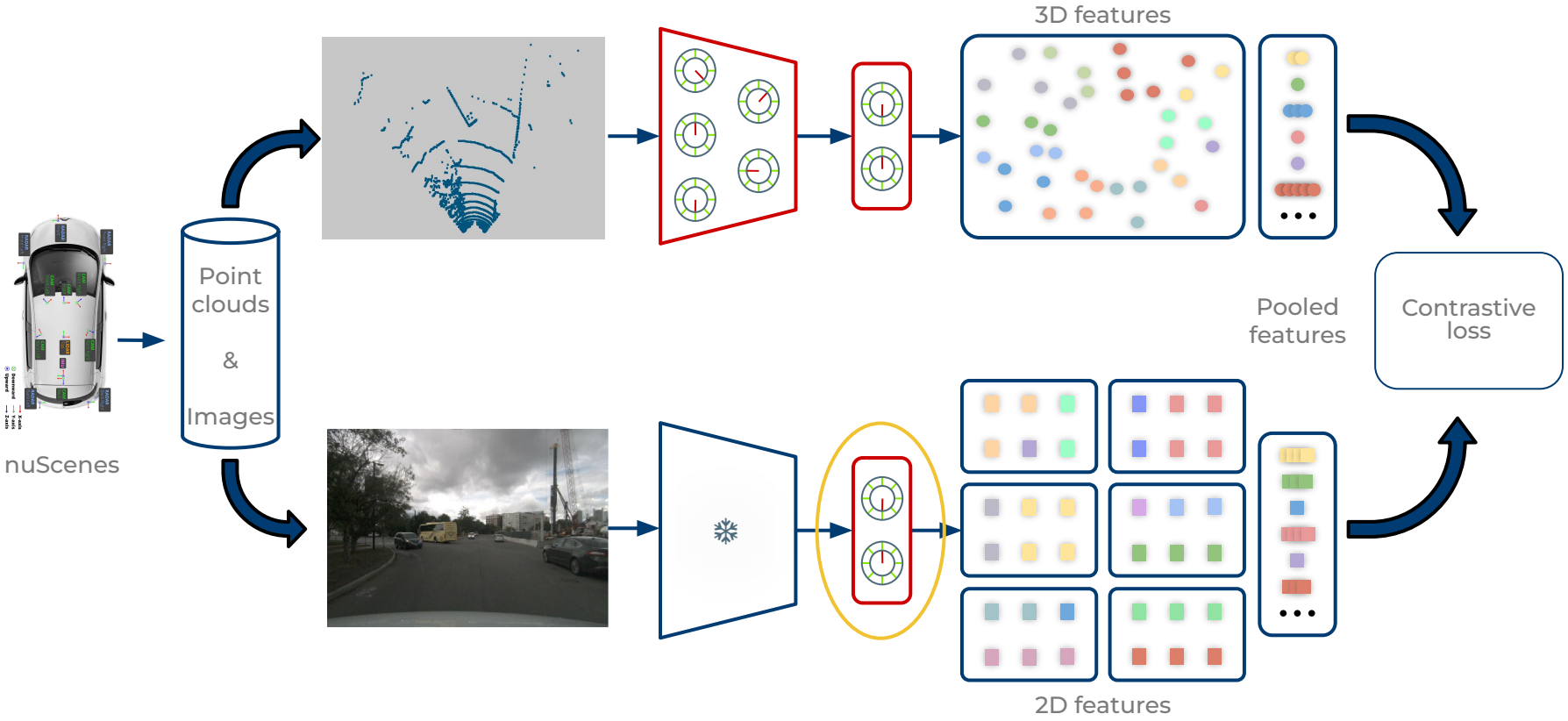
G. Puy et al., Three Pillars improving Vision Foundation Model Distillation for Lidar, CVPR 2024

Simplifications done in ScaLR



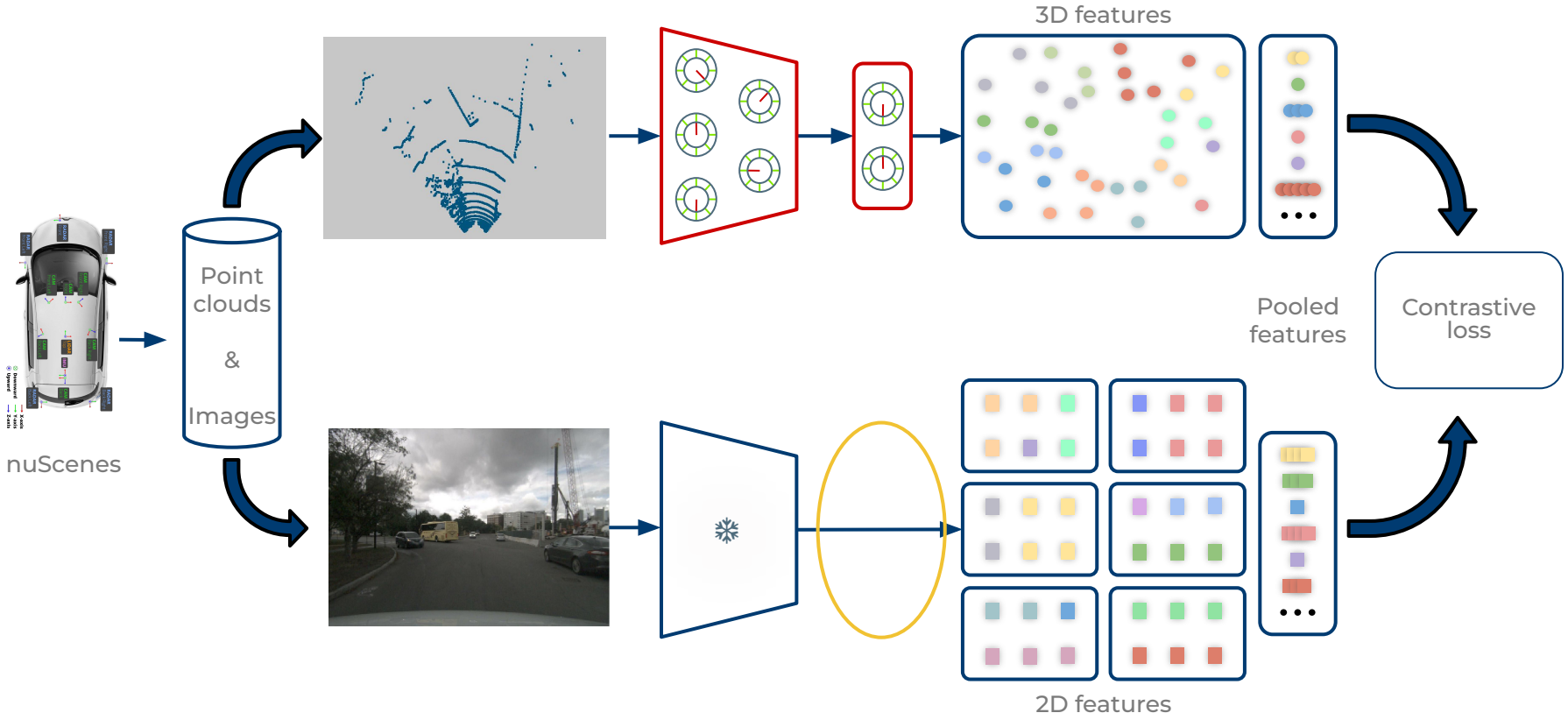
G. Puy et al., Three Pillars improving Vision Foundation Model Distillation for Lidar, CVPR 2024

Simplifications done in ScaLR



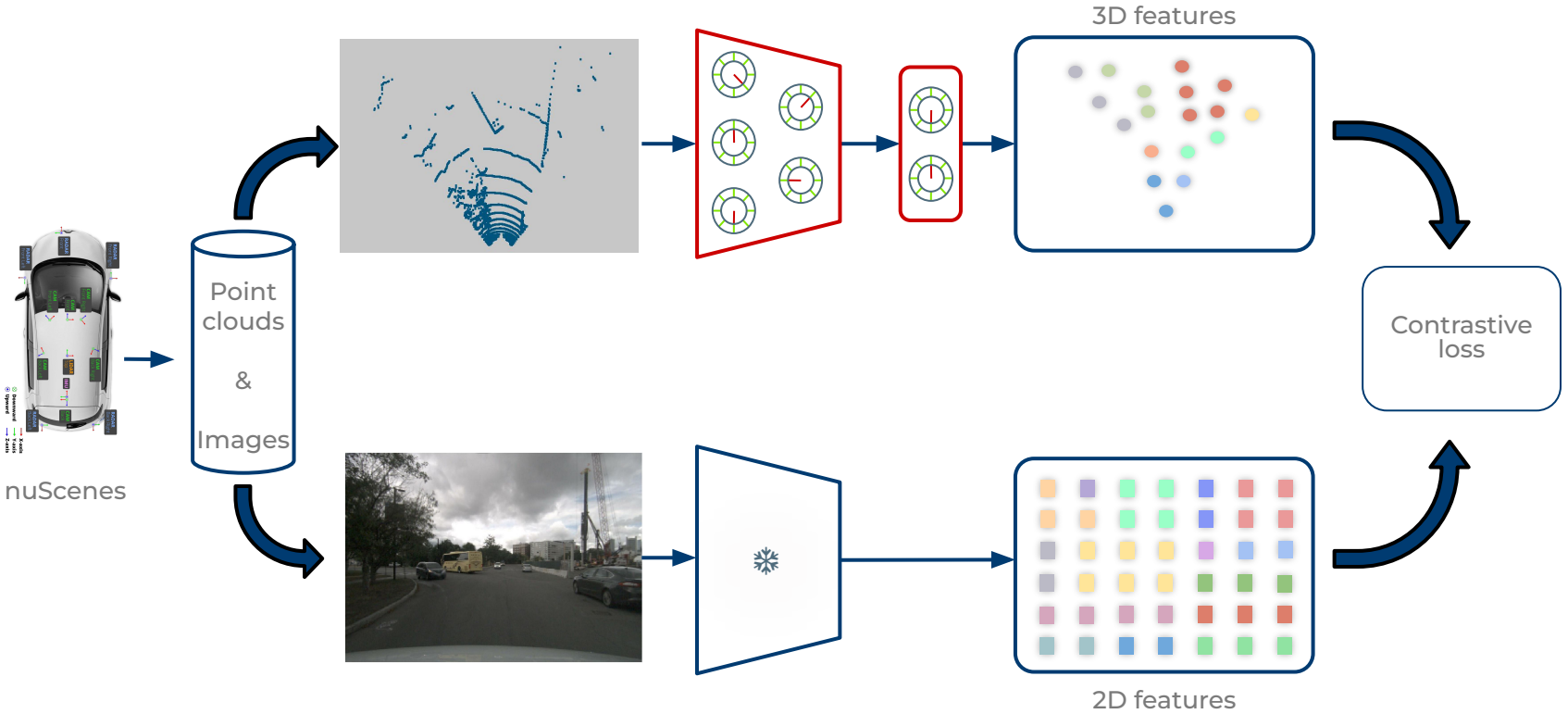
G. Puy et al., Three Pillars improving Vision Foundation Model Distillation for Lidar, CVPR 2024

Simplifications done in ScaLR



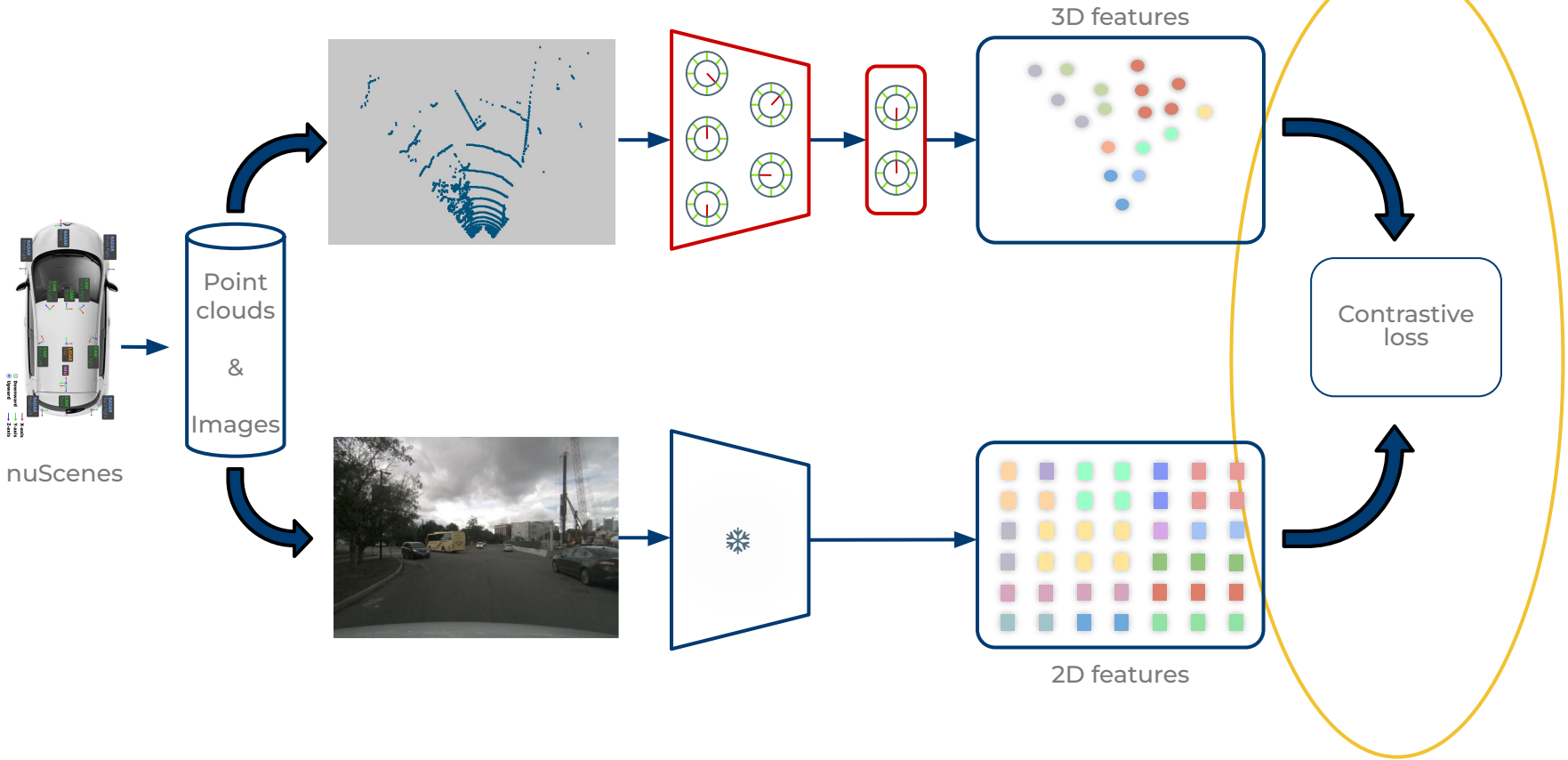
G. Puy et al., Three Pillars improving Vision Foundation Model Distillation for Lidar, CVPR 2024

Simplifications done in ScaLR



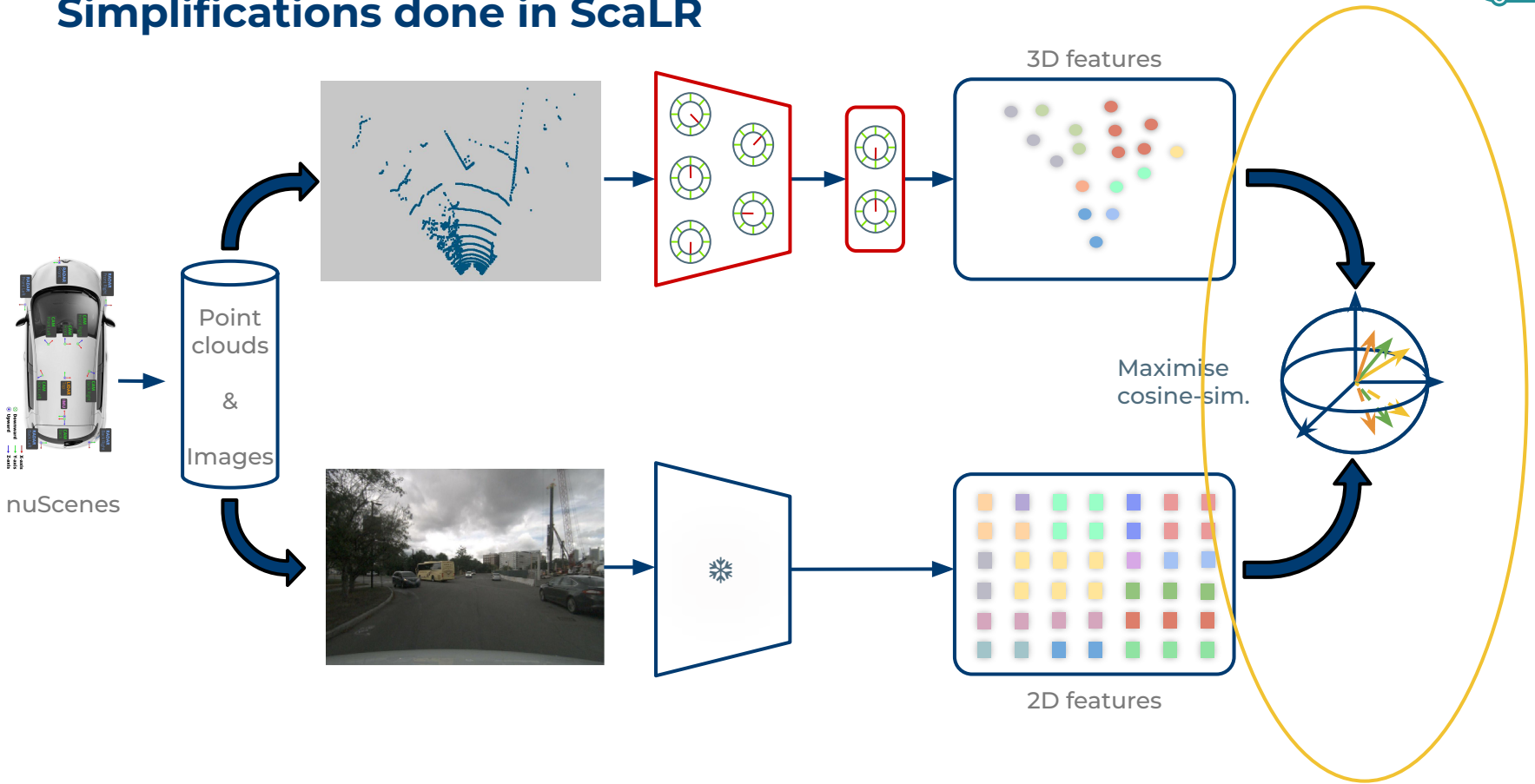
G. Puy et al., Three Pillars improving Vision Foundation Model Distillation for Lidar, CVPR 2024

Simplifications done in ScaLR



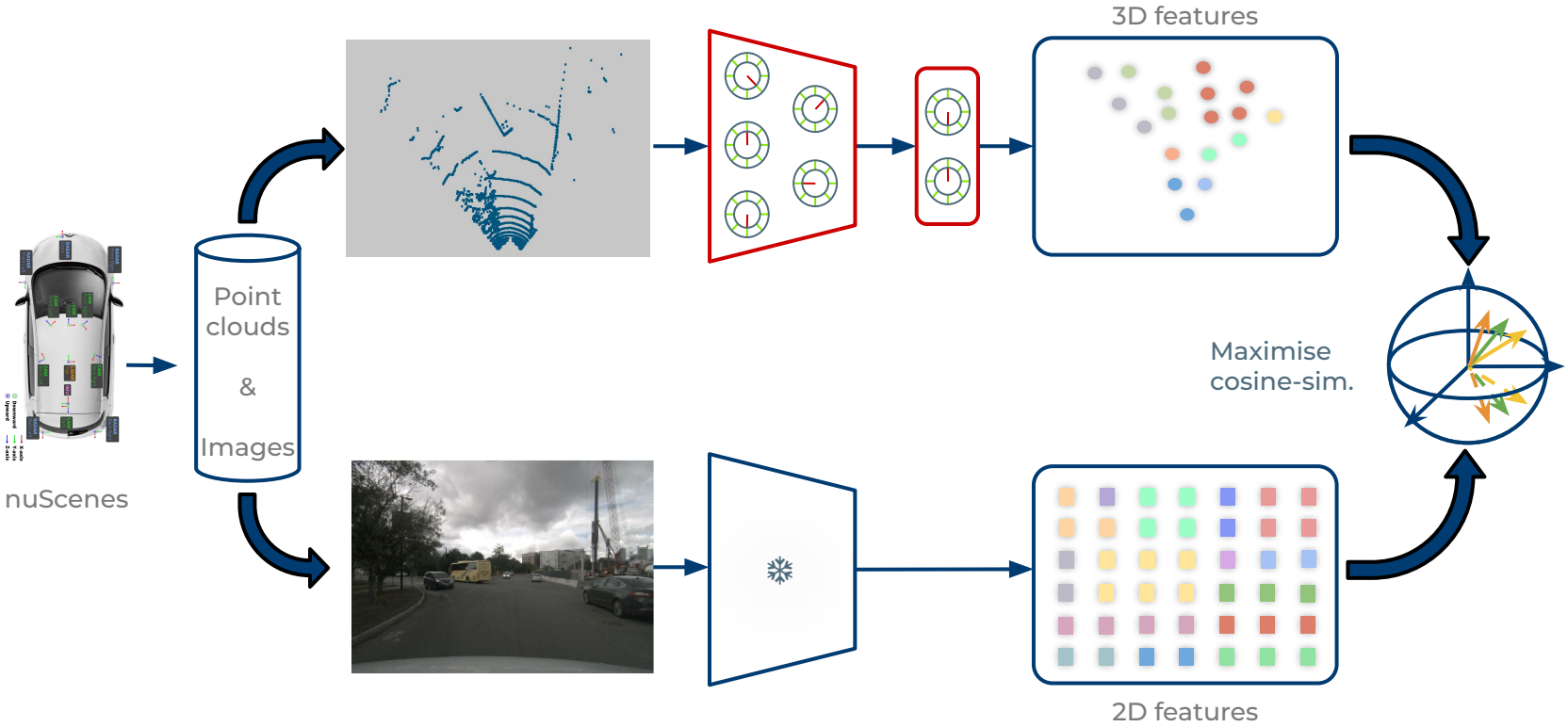
G. Puy et al., Three Pillars improving Vision Foundation Model Distillation for Lidar, CVPR 2024

Simplifications done in ScaLR



G. Puy et al., Three Pillars improving Vision Foundation Model Distillation for Lidar, CVPR 2024

Simplifications done in ScaLR



G. Puy et al., Three Pillars improving Vision Foundation Model Distillation for Lidar, CVPR 2024

Impact of these simplifications

Impact of simplifications

	2D head	Contrast. loss	Cosine loss	No. cam.	Mem. / Time ↓	mIoU% ↑
MinkUNet	✓	✓	-	6	1.0 / 1.0	39.3
	✗	✓	-	6	1.6 / 1.0	43.4
	✗	-	✓	6	1.6 / 1.1	43.6
	✗	-	✓	1	0.8 / 0.3	42.1
WI-48-256	✓	✓	-	1	-	50.0
	✗	✓	-	1	-	54.1
	✗	-	✓	1	-	56.7

Impact of simplifications

	2D head	Contrast. loss	Cosine loss	No. cam.	Mem. / Time ↓	mIoU% ↑
MinkUNet	✓	✓	-	6	1.0 / 1.0	39.3
	✗	✓	-	6	1.6 / 1.0	43.4
	✗	-	✓	6	1.6 / 1.1	43.6
	✗	-	✓	1	0.8 / 0.3	42.1
WI-48-256	✓	✓	-	1	-	50.0
	✗	✓	-	1	-	54.1
	✗	-	✓	1	-	56.7

Impact of simplifications

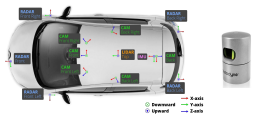
	2D head	Contrast. loss	Cosine loss	No. cam.	Mem. / Time ↓	mIoU% ↑
MinkUNet	✓	✓	-	6	1.0 / 1.0	39.3
	✗	✓	-	6	1.6 / 1.0	43.4
	✗	-	✓	6	1.6 / 1.1	43.6
	✗	-	✓	1	0.8 / 0.3	42.1
WI-48-256	✓	✓	-	1	-	50.0
	✗	✓	-	1	-	54.1
	✗	-	✓	1	-	56.7

Impact of simplifications

	2D head	Contrast. loss	Cosine loss	No. cam.	Mem. / Time ↓	mIoU% ↑
MinkUNet	✓	✓	-	6	1.0 / 1.0	39.3
	✗	✓	-	6	1.6 / 1.0	43.4
	✗	-	✓	6	1.6 / 1.1	43.6
	✗	-	✓	1	0.8 / 0.3	42.1
WI-48-256	✓	✓	-	1	-	50.0
	✗	✓	-	1	-	54.1
	✗	-	✓	1	-	56.7

Study influence of three pillars

ScaLR

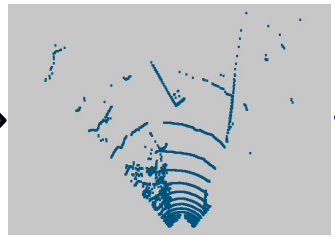


nuScenes - Velodyne 32

SemKITTI - Velodyne 64

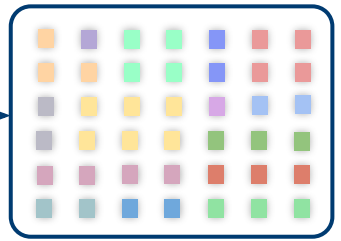
Pandaset - Pandar 64 & GT

Point clouds & Images



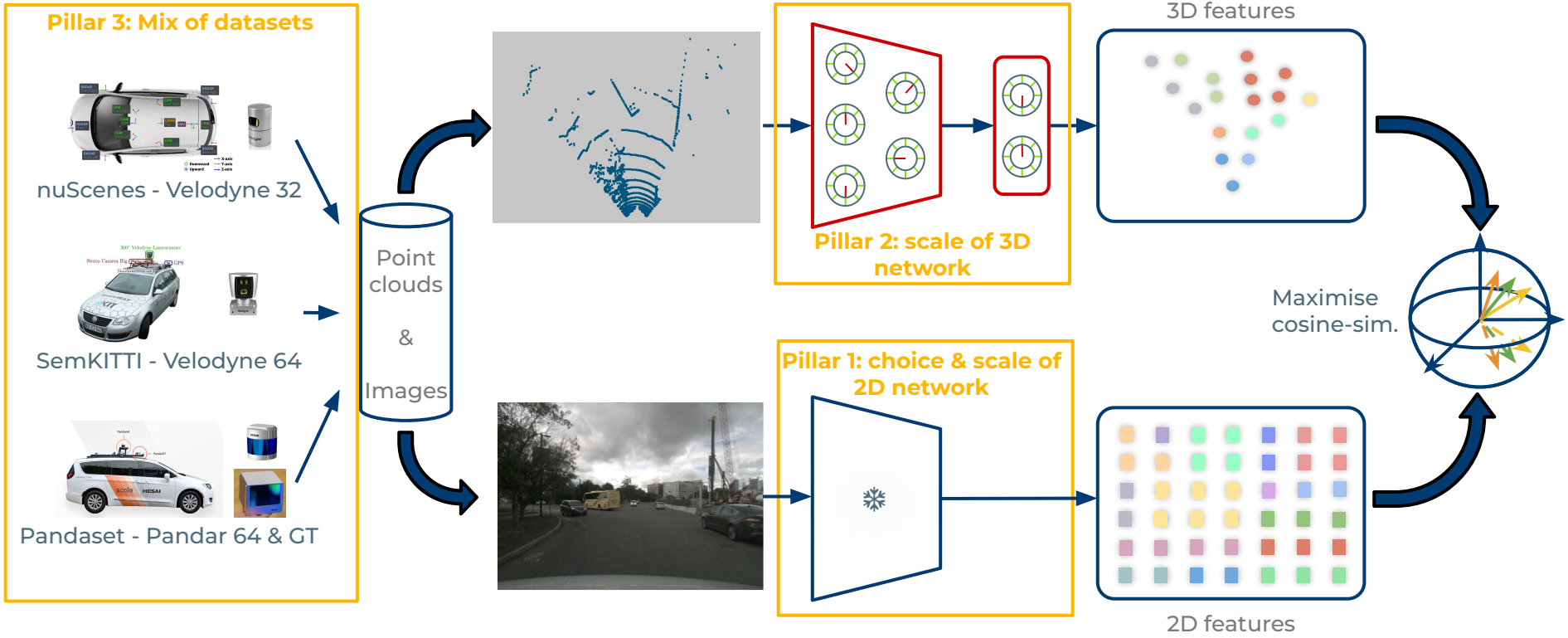
3D features

2D features

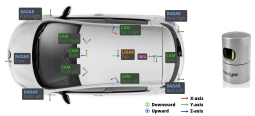


Maximise cosine-sim.

The 3 pillars



Pillar 1: choice & scale of 2D network

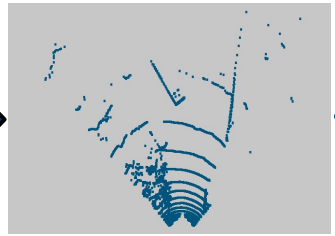


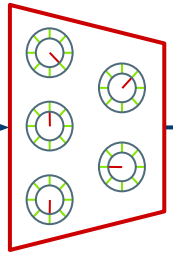
nuScenes - Velodyne 32

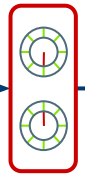
SemKITTI - Velodyne 64

Pandaset - Pandar 64 & GT

Point clouds & Images

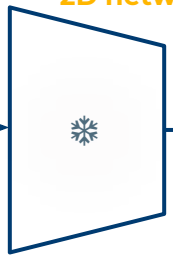


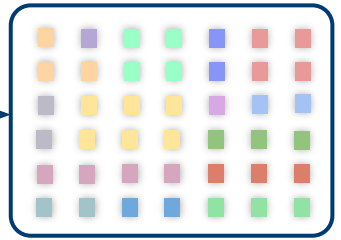




3D features

Pillar 1: choice & scale of 2D network

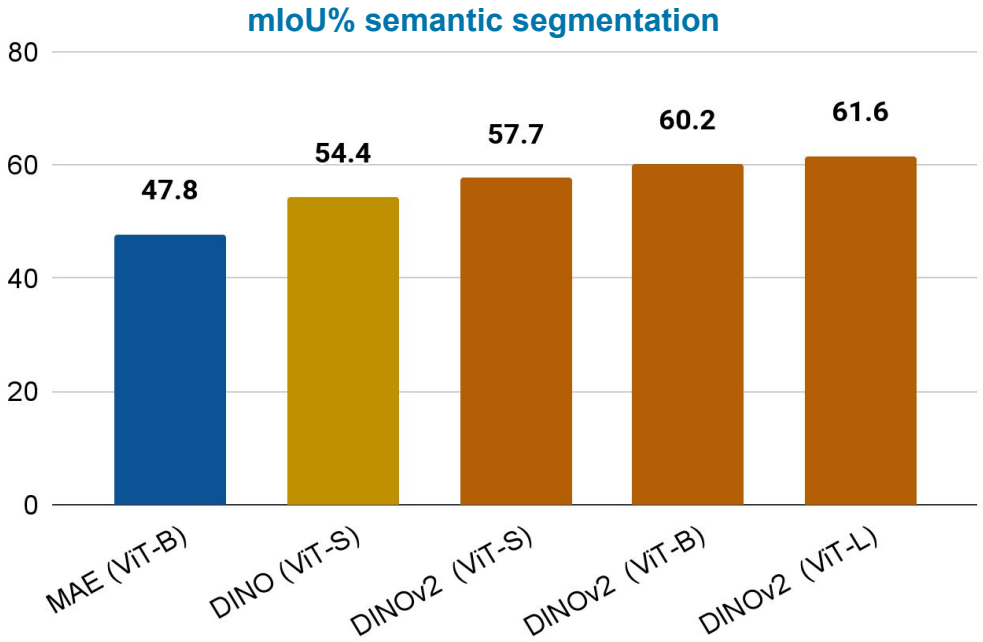




2D features

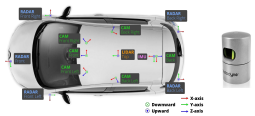
Maximise cosine-sim.

Pillar 1: choice & scale of 2D network



Pretraining & linear probing on nuScenes - Distilled into WaffleIron-256 (features of size 256)

Pillar 2: scale of 3D network

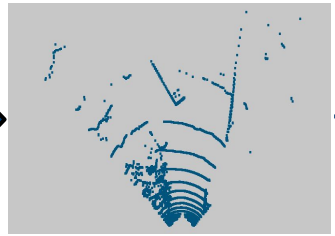


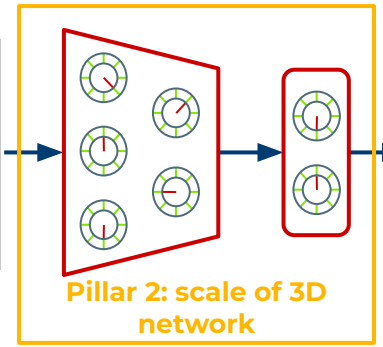
nuScenes - Velodyne 32

SemKITTI - Velodyne 64

Pandaset - Pandar 64 & GT

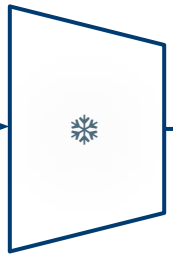
Point clouds & Images

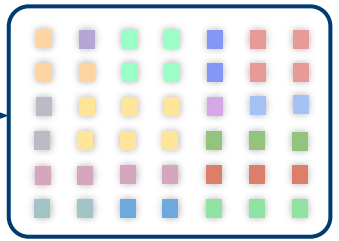




Pillar 2: scale of 3D network

3D features

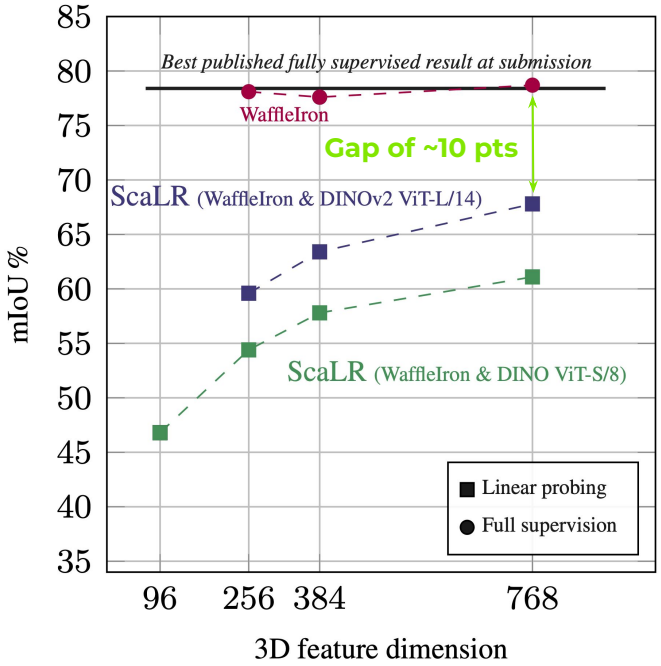




2D features

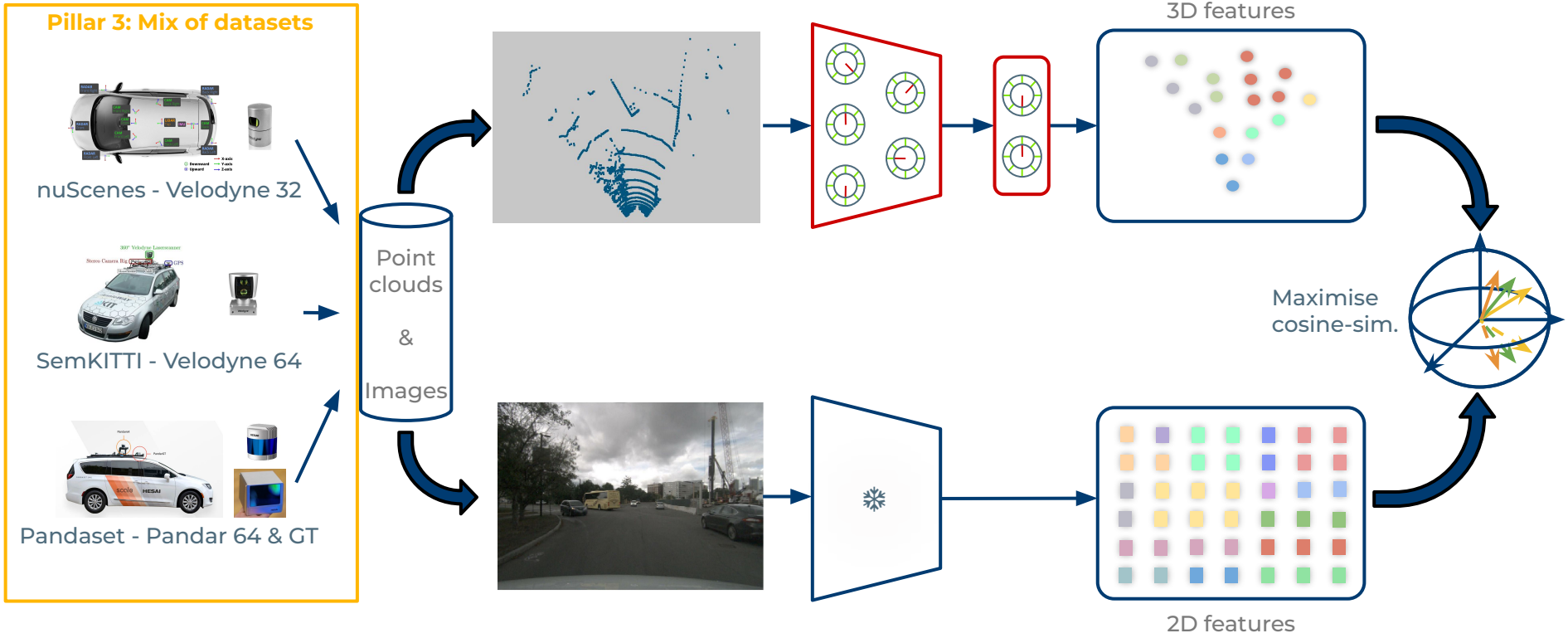
Maximise cosine-sim.

Pillar 2: scale of 3D network



Pretraining & linear probing on nuScenes - Distilled into WaffleIron with varying feature sizes

Pillar 3: Mix of datasets



Pillar 3: Mix of datasets

	Pretrain. Dataset	Downstream & Test Dataset			
		nuScenes	SemKITTI	Pand. 64	Pand. GT
WI-256	*Pretraining with DINO-ViT-S/8 and linear probing*				
	nuScenes	54.4	28.8	26.9	25.2
	SemKITTI	39.5	46.6	25.3	25.7
	Pand. 64	39.6	25.6	30.0	24.7
	Pand. GT	29.9	26.9	23.5	28.5
	All datasets	**54.6**	**50.6**	**33.1**	**32.3**
WI-768	*Pretraining with DINOv2-ViT-L/14 and linear probing*				
	nuScenes	**67.8**	43.1	33.9	29.9
	All datasets	**67.8**	**55.8**	**37.9**	**34.5**

Pretraining on individual datasets or mix of all - Linear probing on each datasets

Pillar 3: Mix of datasets

	Pretrain. Dataset	Downstream & Test Dataset			
		nuScenes	SemKITTI	Pand. 64	Pand. GT
WI-256	*Pretraining with DINO-ViT-S/8 and linear probing*				
	nuScenes	54.4	28.8	26.9	25.2
	SemKITTI	39.5	46.6	25.3	25.7
	Pand. 64	39.6	25.6	30.0	24.7
	Pand. GT	29.9	26.9	23.5	28.5
	All datasets	**54.6**	**50.6**	**33.1**	**32.3**
WI-768	*Pretraining with DINOv2-ViT-L/14 and linear probing*				
	nuScenes	**67.8**	43.1	33.9	29.9
	All datasets	**67.8**	**55.8**	**37.9**	**34.5**

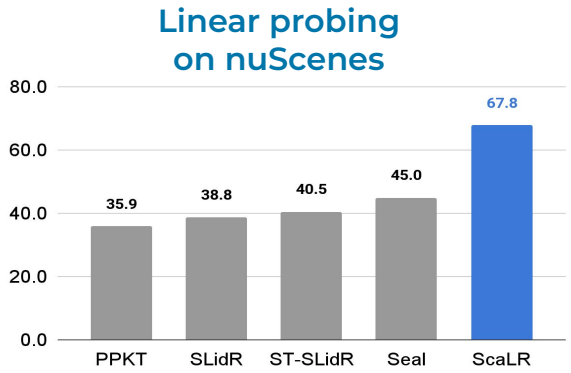
Pretraining on individual datasets or mix of all - Linear probing on each datasets

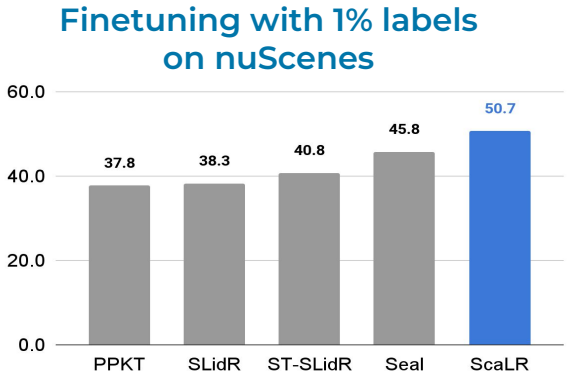
Pillar 3: Mix of datasets

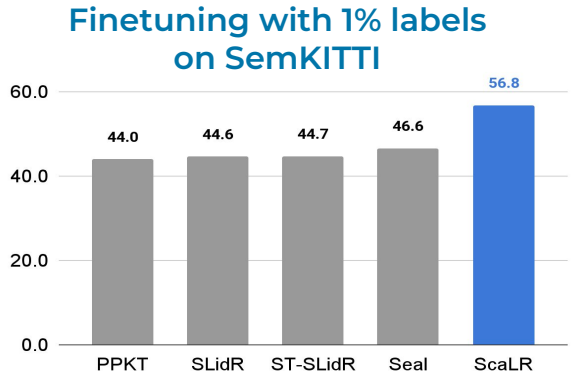
	Pretrain. Dataset	Downstream & Test Dataset			
		nuScenes	SemKITTI	Pand. 64	Pand. GT
WI-256	*Pretraining with DINO-ViT-S/8 and linear probing*				
	nuScenes	54.4	28.8	26.9	25.2
	SemKITTI	39.5	46.6	25.3	25.7
	Pand. 64	39.6	25.6	30.0	24.7
	Pand. GT	29.9	26.9	23.5	28.5
	All datasets	**54.6**	**50.6**	**33.1**	**32.3**
WI-768	*Pretraining with DINOv2-ViT-L/14 and linear probing*				
	nuScenes	**67.8**	43.1	33.9	29.9
	All datasets	**67.8**	**55.8**	**37.9**	**34.5**

Pretraining on individual datasets or mix of all - Linear probing on each datasets

Progress after optimizing the 3 pillars

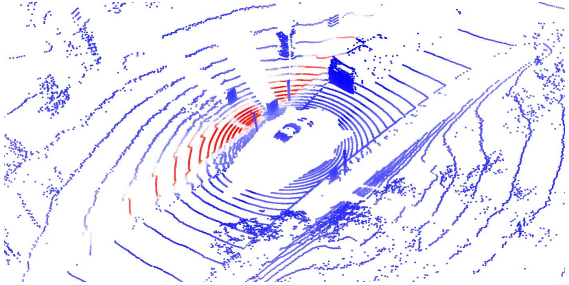




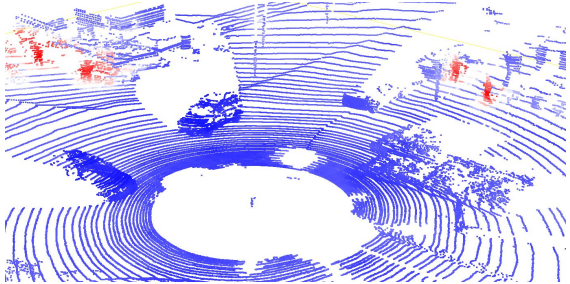


Qualitative results

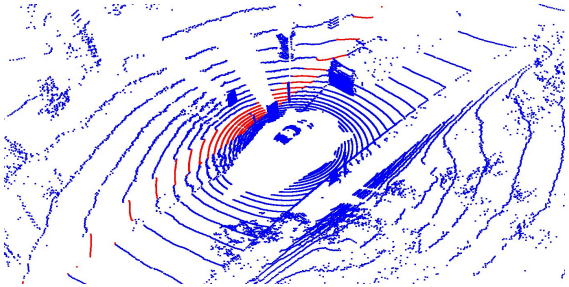
Correlation maps with class prototypes

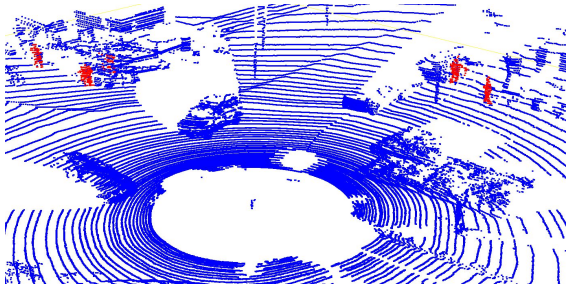


nuScenes - Sidewalk



Sem.KITTI - Pedestrian

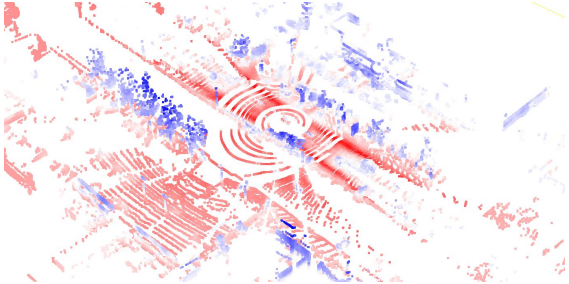




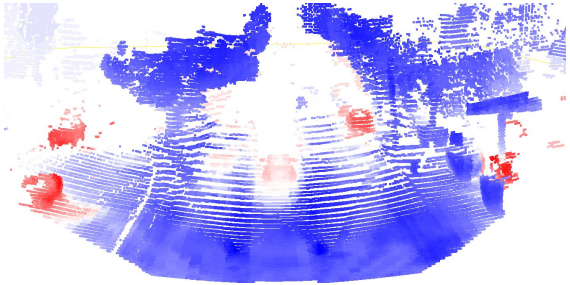
Ground truth

Qualitative results

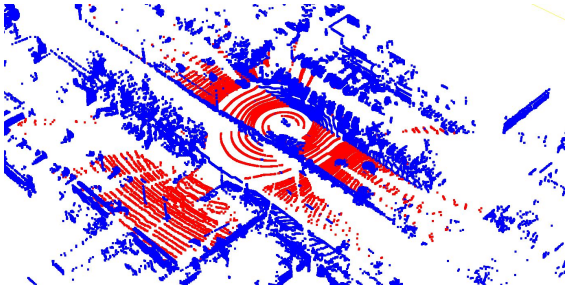
Correlation maps with class prototypes

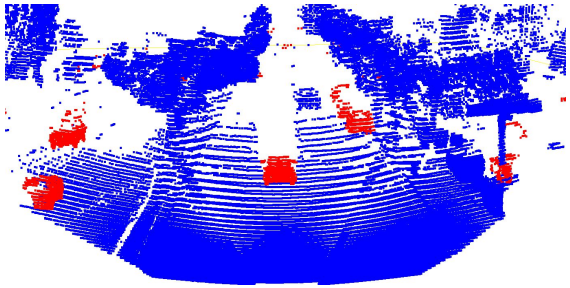


PandaSet 64 - Road



PandaSet GT - Car



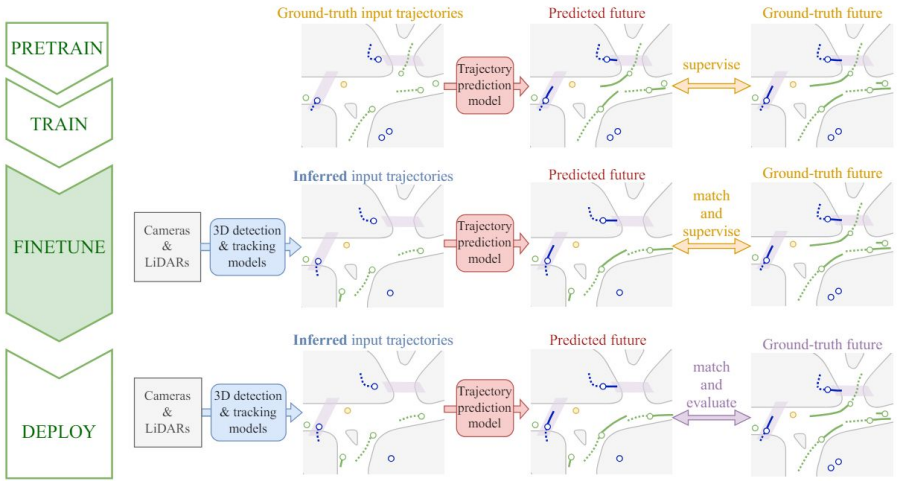


Ground truth

Conclusion and perspectives

Conclusion

- Scalable and simple pretraining strategies become first class citizens in AD
- Data outside the AD domain (internet, robotics) is key for camera perception
- Training over multiple datasets and sensor configurations shows promising results (ScaLR, UniTraj)
- Increased reliance on previously trained modules



Valeo4Cast approach at the Argoverse 2 end-to-end forecasting challenge 2024

		Forecasting			Detection			Tracking
		mAP_f (↑)	ADE (↓)	FDE (↓)	Inputs	Training range	mCDS (↑)	HOTA (↑)
2023	CenterPoint [28]	-	-	-	L	150	14	-
	BEVFusion [13]	-	-	-	L+C	150	37	-
	Anony_3D (v0) [11]	-	-	-	L	150	31	44.36
	Host_4626_Team [17]	14.51	5.10	7.32	-	-	-	39.98
	Dgist-cvlab [25]	42.91	4.11	4.59	L+C	150	34	41.49
	Le3DE2E [22]	46.70	3.22	3.76	L+C	150	39	56.19
2024	Dgist-cvlab	45.83	4.09	4.53	L+C	150	34	41.49
	Le3DE2E	50.53	4.07	4.60	L+C	150	**43**	**64.60**
	Valeo4Cast	**63.82**	**2.14**	**2.43**	L	75	31	61.28

Valeo4Cast results across sub-challenges

Y. Xu et al., Valeo4Cast: A Modular Approach to End-to-End Forecasting, arXiv 2024
L. Feng et al., UniTraj: A Unified Framework for Scalable Vehicle Trajectory Prediction, arXiv 2024

Perspectives

- Different forms of world models taking over paving the path to embodied AI
- More downstream use-cases around foundation models
- Arrival of new larger scale datasets in the community
- Need for more open research in the field (data, code, models, collabs)

nuPlan
(calibrated multi-cam + LiDAR)
~ 100h of driving data @10Hz

OpenDV-Youtube
(only non-calib. front-cam)
~ 1700h of driving data @10Hz

N. Karnchanachari et al., Towards learning-based planning:The nuPlan benchmark for real-world autonomous driving, ICRA 2024
J. Yang et al., Generalized Predictive Model for Autonomous Driving, CVPR 2024

https://valeoai.github.io