

3D perception in autonomous driving

From lidar self-supervision to multi-sensor distillation

Alexandre Boulch

Agenda

01	Introduction	X
02	Self-supervision for lidar	X
03	Distillation from images	X
04	Can self-supervision help you with domain gaps?	X
05	Conclusion	X

01 02 03 04 05

Introduction


Patrick


Renaud


Andrei


Nicolas


Vincent


Sophia


Corentin


Björn


Gilles


Tuan-Hung


Spyros

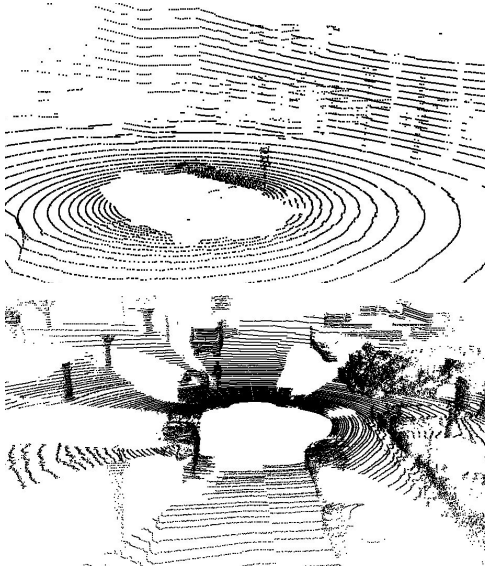

Oriane


Nermin


Antonin

Autonomous driving data

Automotive lidar



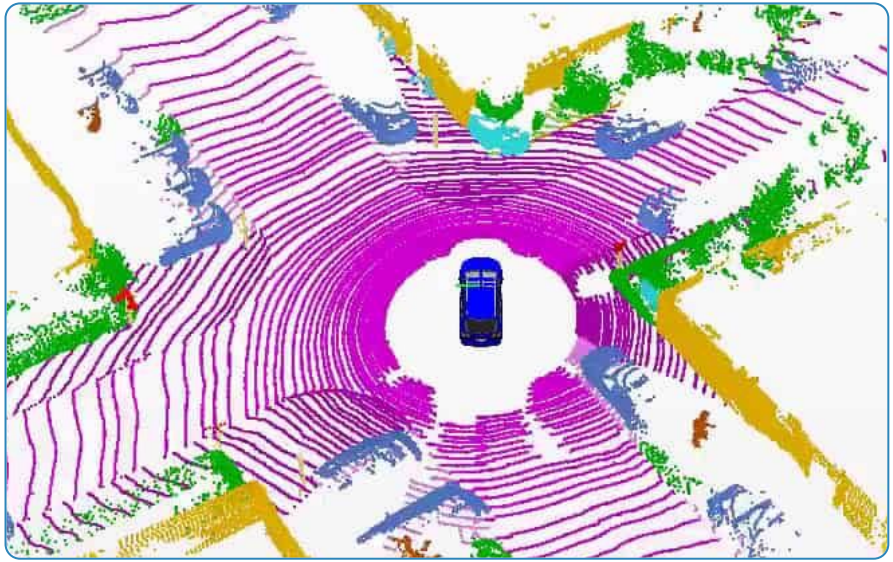
AD cameras

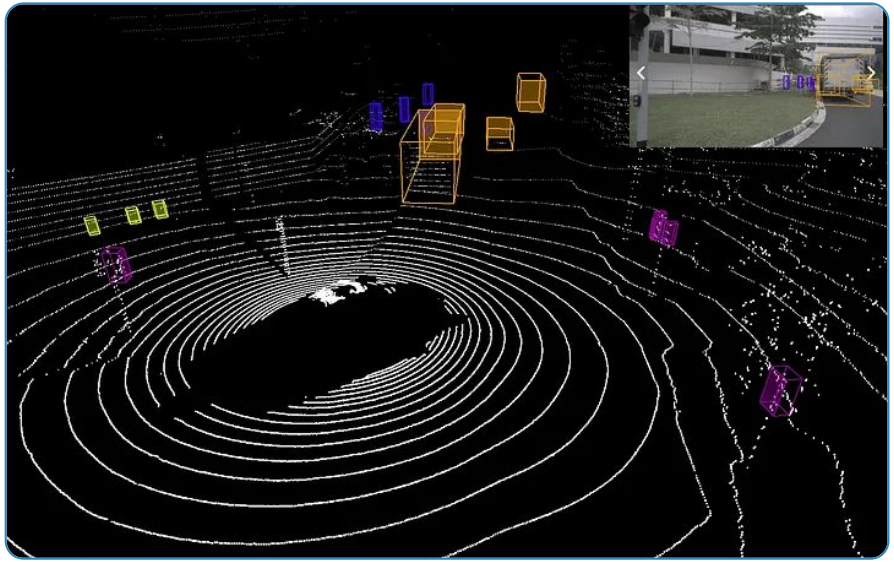


Datasets

Academic datasets
nuScenes
SemanticKITTI
SemanticPOSS
...

Evaluation tasks





Semantic segmentation

Assign a label to each input point

3D detection

Predict 3D oriented boxes containing objects of interest

Frugal learning

Reducing the annotation effort
(Reducing the data, not in this talk)

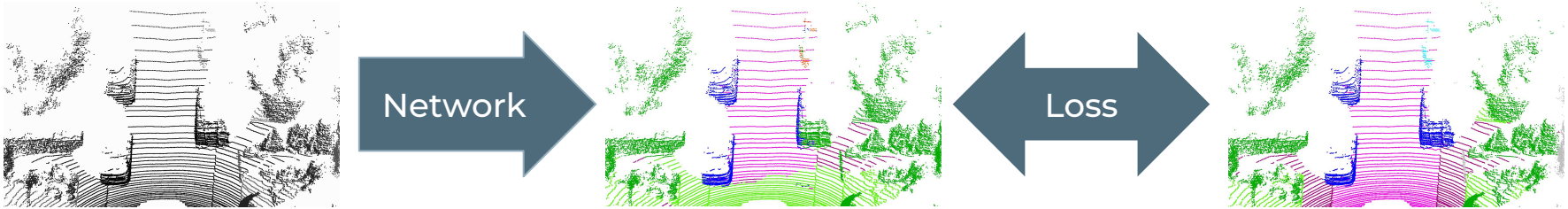
→ Lowering costs
→ improved performance
→ only way to scale?

Domain adaptation

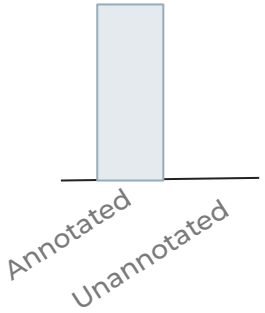
Transferring a model to another domain

→ practical use case in industry
→ train once apply anywhere with minimal change

Supervised learning

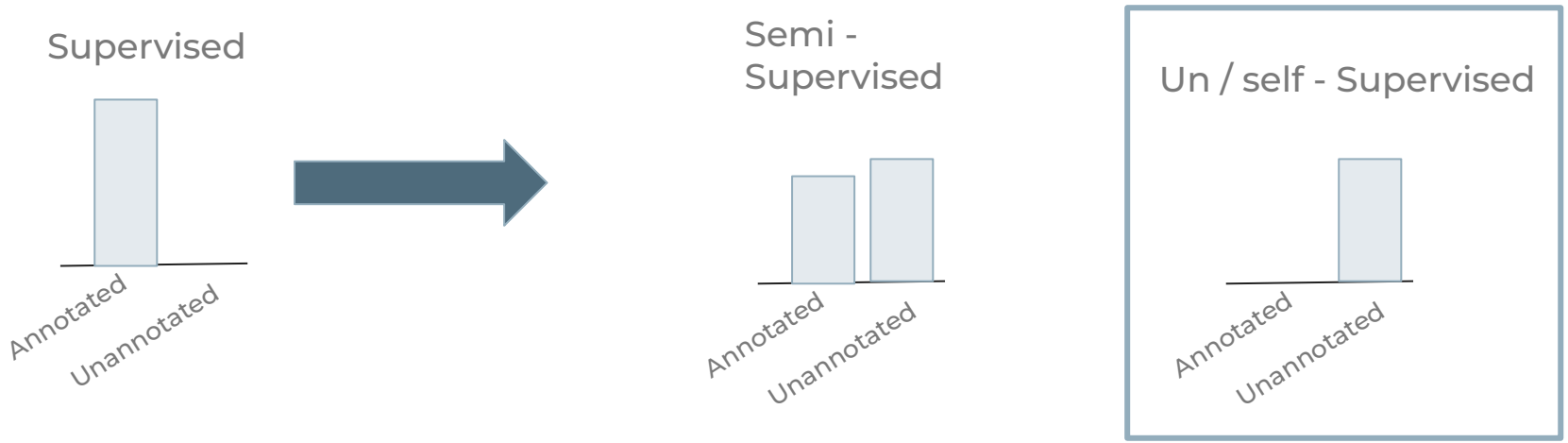


Train with annotated data
⇒ Annotations are costly
⇒ Limitation of the database size

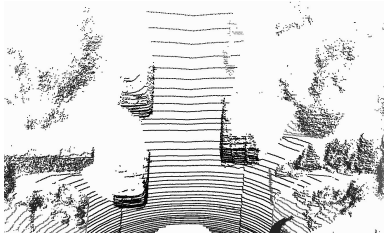


(Annotation-)Frugal learning

Objective: Learning with less annotations

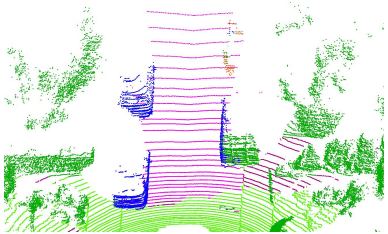


Models



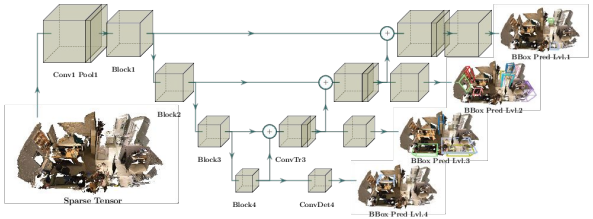
Input

Network



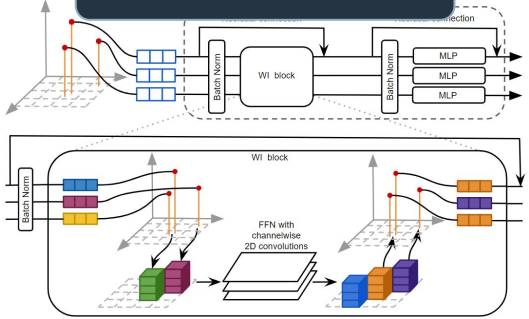
Predictions

Convolutional



Minkowski Net (MinkUNet) Choy et al.

Mixers



WaffleIron (Puy et al.)

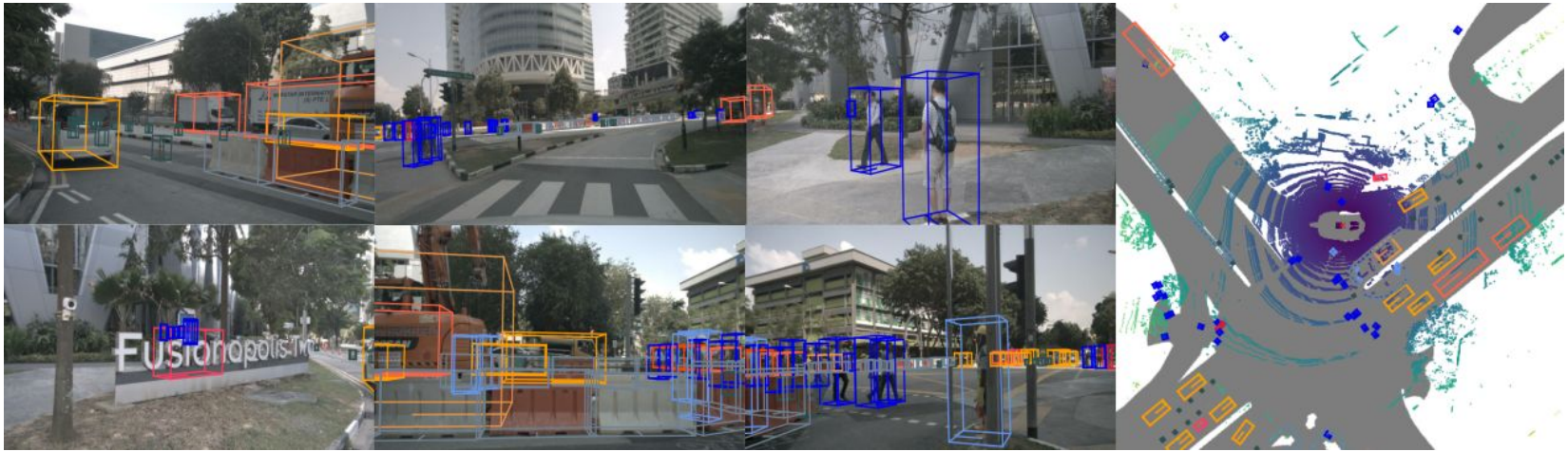
Transformers / Hybrids

PtV3
VoLT
VaVit
...

Bird’s Eye View (BEV) models

To drive and plan safely the AV needs an accurate representation of the surrounding environment

BEV representational space – the top-view is a suitable and natural space to fuse sensor information. (Multi camera and/or LiDAR etc.)

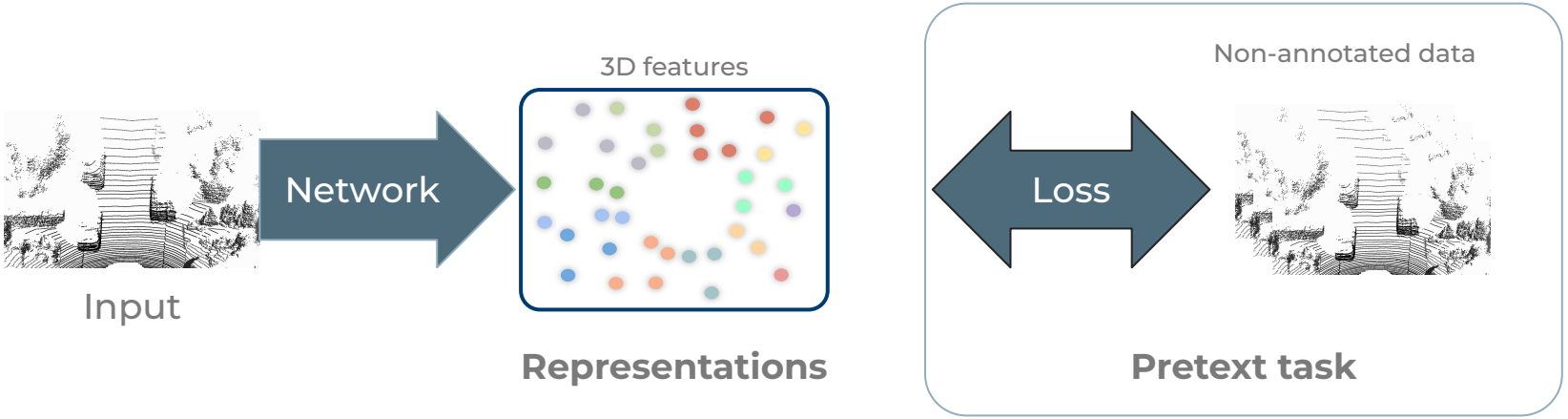


Lidar self supervision

1. Self-supervision
2. Geometric pretext task
3. Contrastive approaches

Self-supervised models

What? learn useful representations without annotations

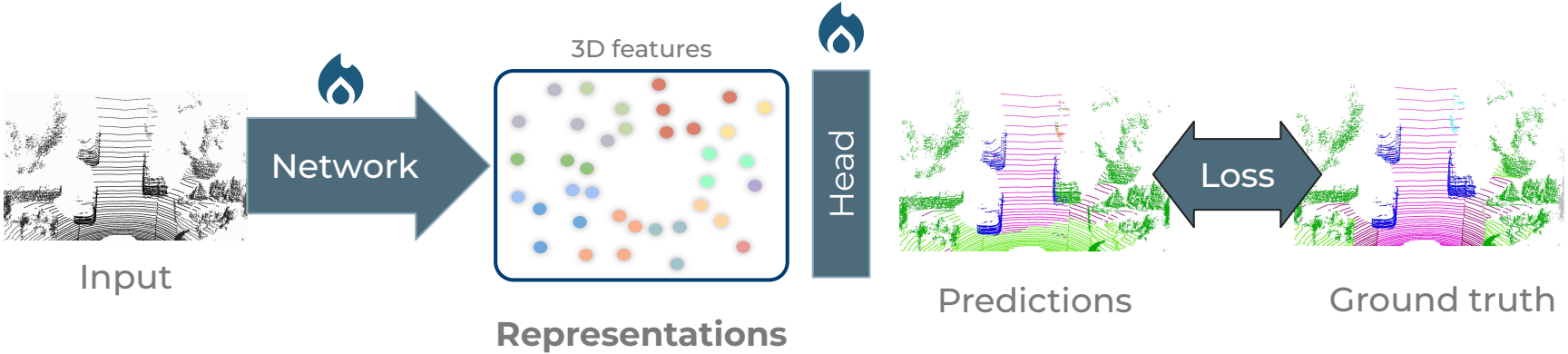


Objective:
after training, the network state will have **good properties**

- Good initialization
- Semantic space
- Hierarchical
- ...

Evaluation - finetune

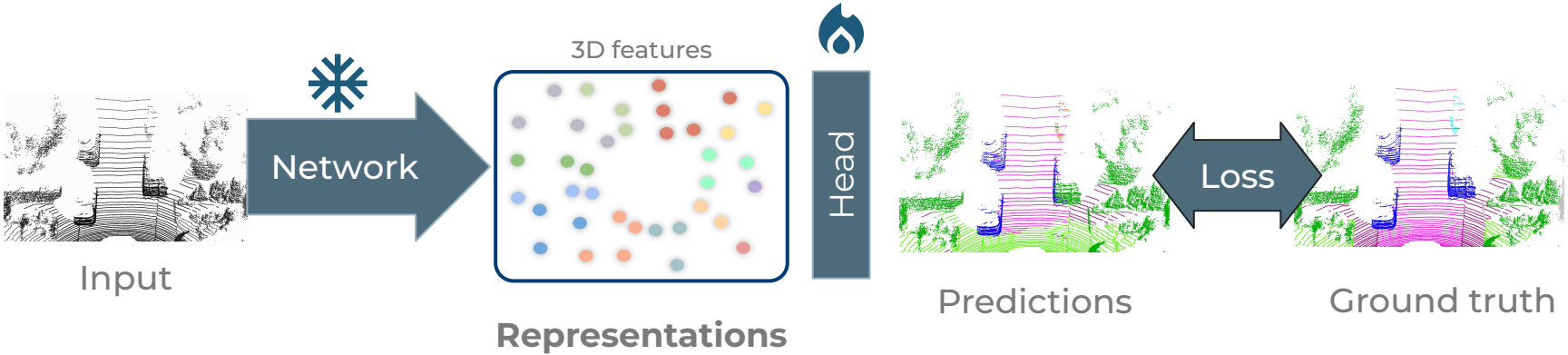
Evaluate if a pre-training is a good initialization for the task.



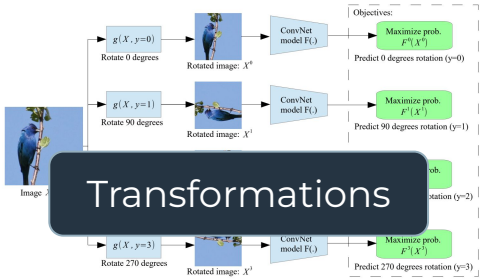
Evaluate the performance with 1%, 10%, ..., 100% annotated data

Evaluation - probing

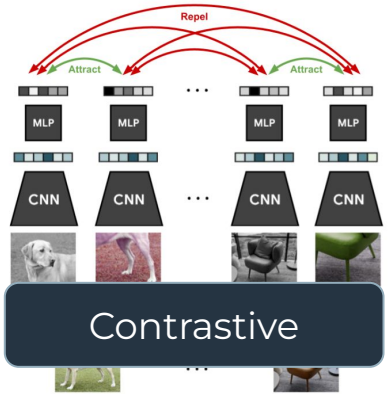
Evaluate if the feature space is well structured for the target task.



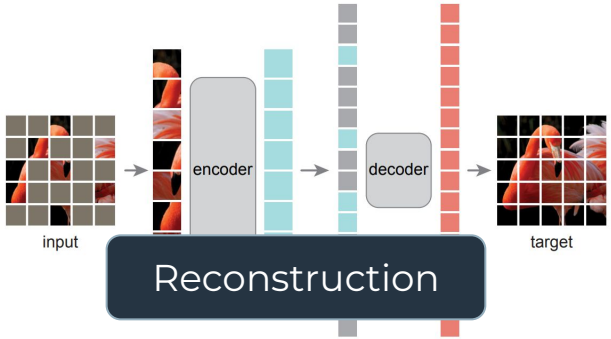
Pretext tasks - Image processing



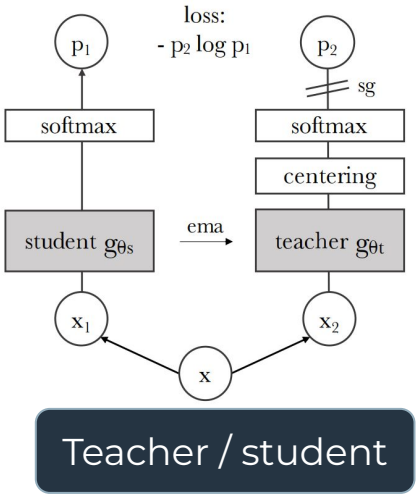
RotNet. Unsupervised representation learning by predicting image rotations



SimCLR: A Simple Framework for Contrastive Learning of Visual Representations



MAE: Masked Autoencoders Are Scalable Vision Learners

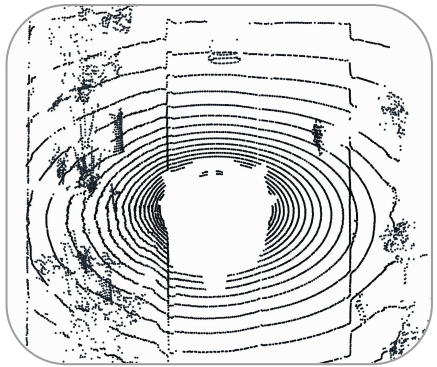


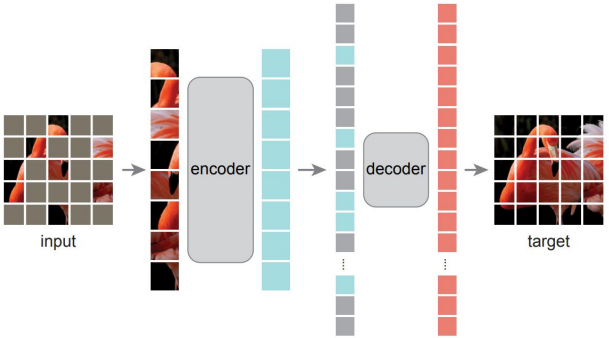
DINO: Emerging Properties in Self-Supervised Vision Transformers

Lidar self supervision

1. Self-supervision
2. Geometric pretext task
3. Contrastive approaches
4. Self-distillation

Geometric/reconstruction approaches to self-supervision





Lidar is a sparse signal

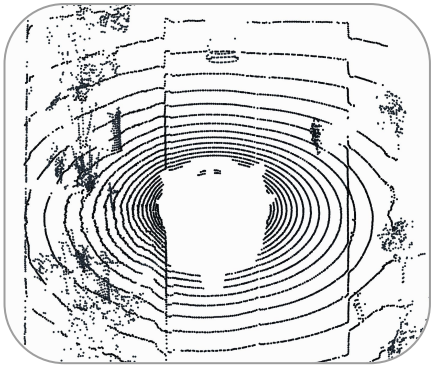
Lidar architectures are mostly convolutional
(e.g., MinkUnet, Ptv3, LightPT…)

Images are dense → reconstruction by dropping patches

MAE targets ViT (easy to drop tokens)

Surface-based self-supervision

Idea: is a network reconstruct a surface, it understands the scene.





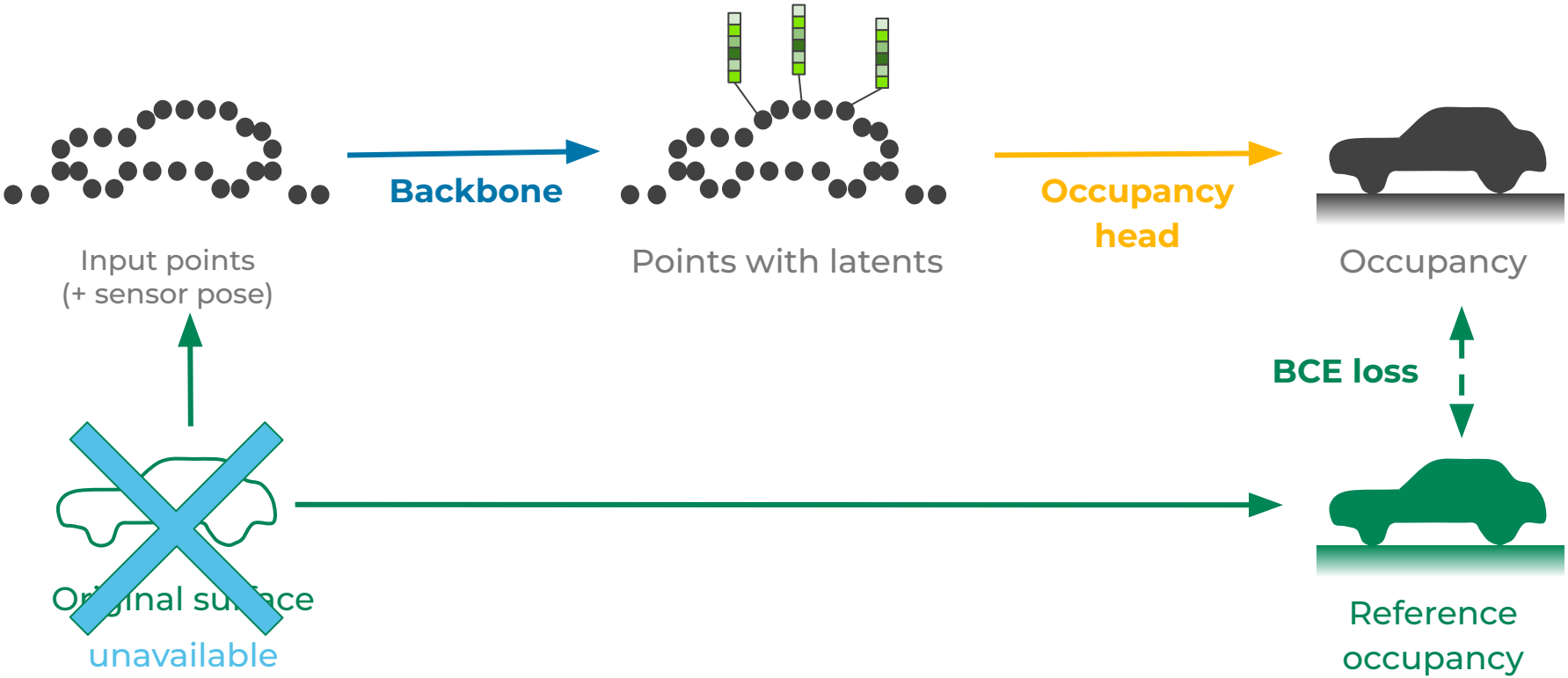
Lidar is a sparse signal

Lidar architectures are mostly convolutional
(e.g., MinkUnet, Ptv3, LightPT…)

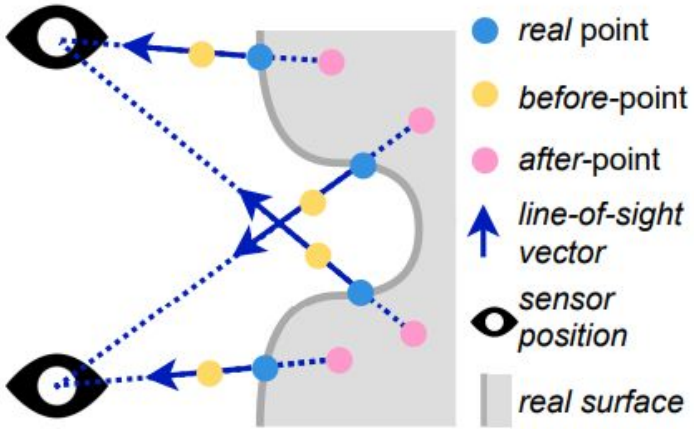
Reconstruct “between the points” (Complete the scene)

Estimate the occupancy (inside/outside the volume)

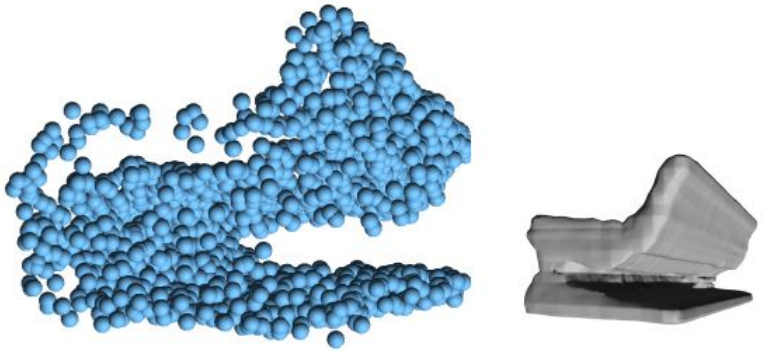
Supervision



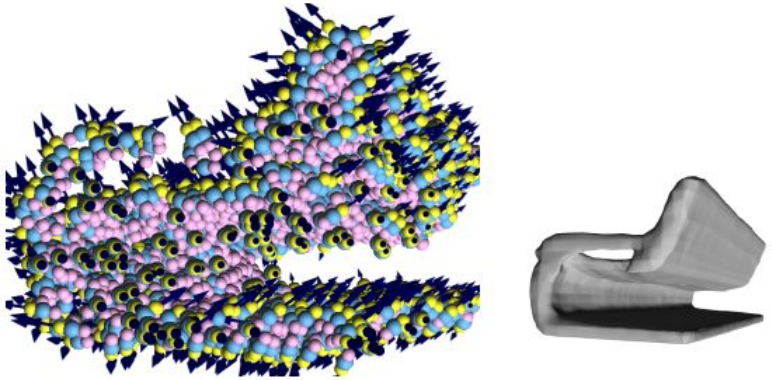
Deep Surface Reconstruction from Point Clouds with Visibility Information



Use the line of sights to generate additional points
→ improves surface reconstruction

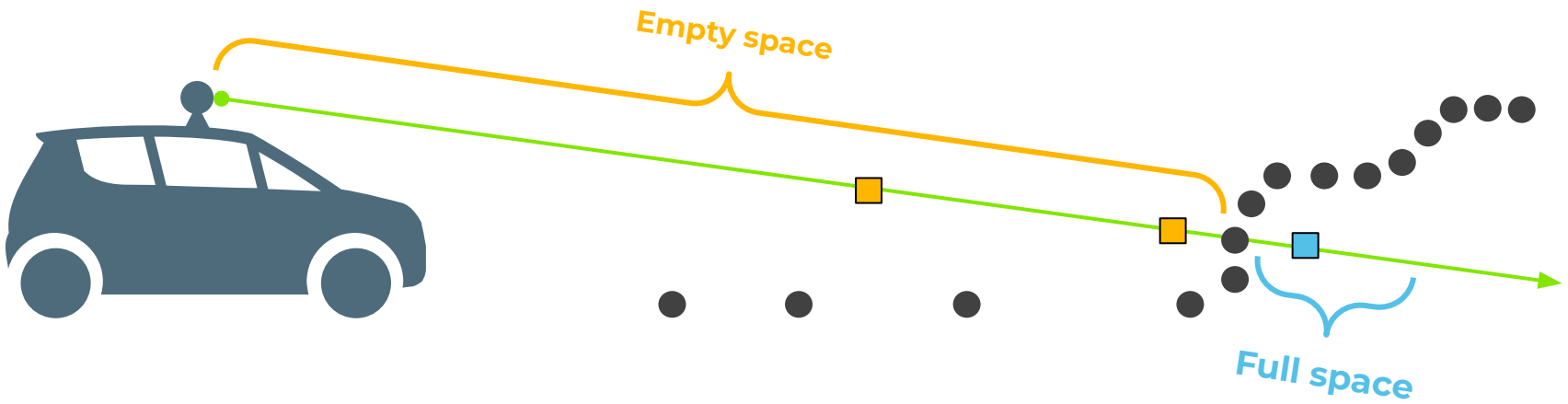


(a) Reconstruction using only the points position.



(b) Reconstruction with visibility augmented point cloud.

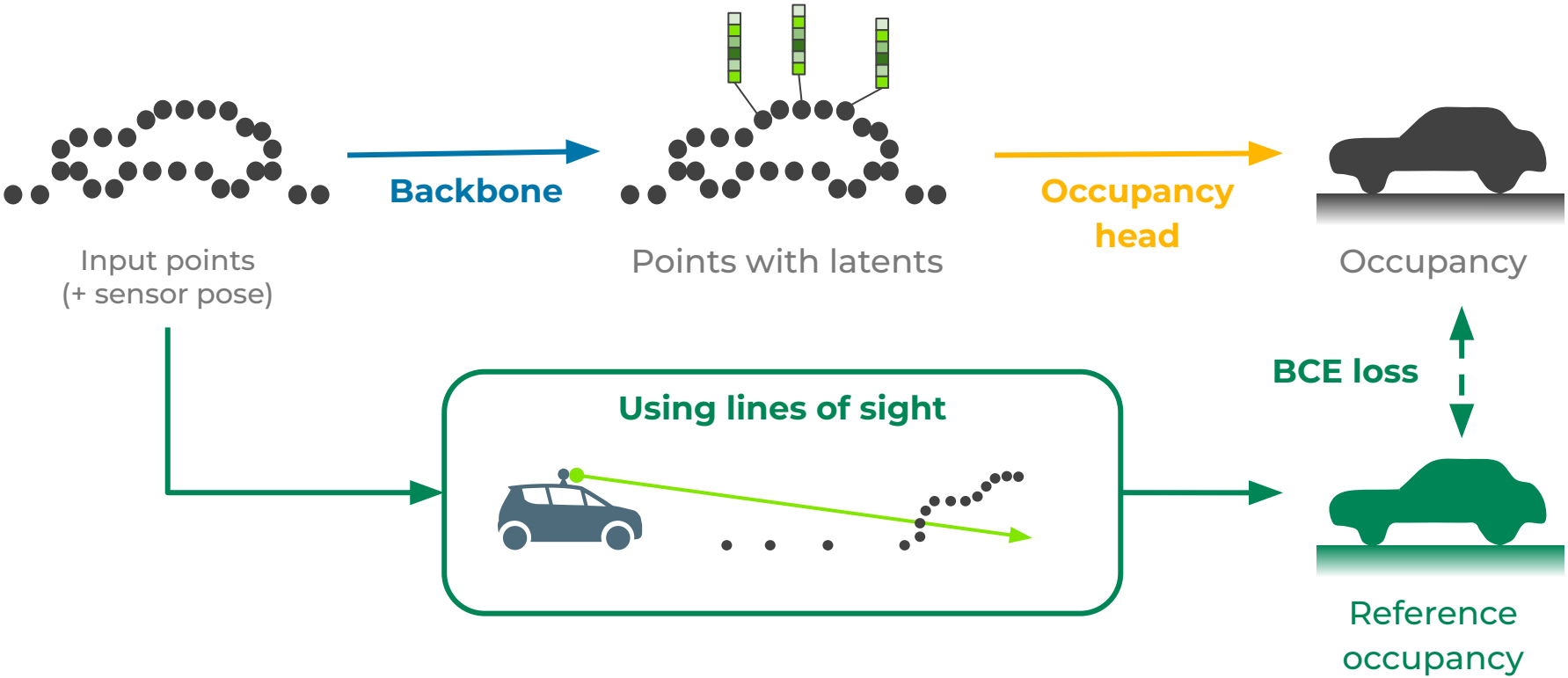
Pseudo labels for occupancy



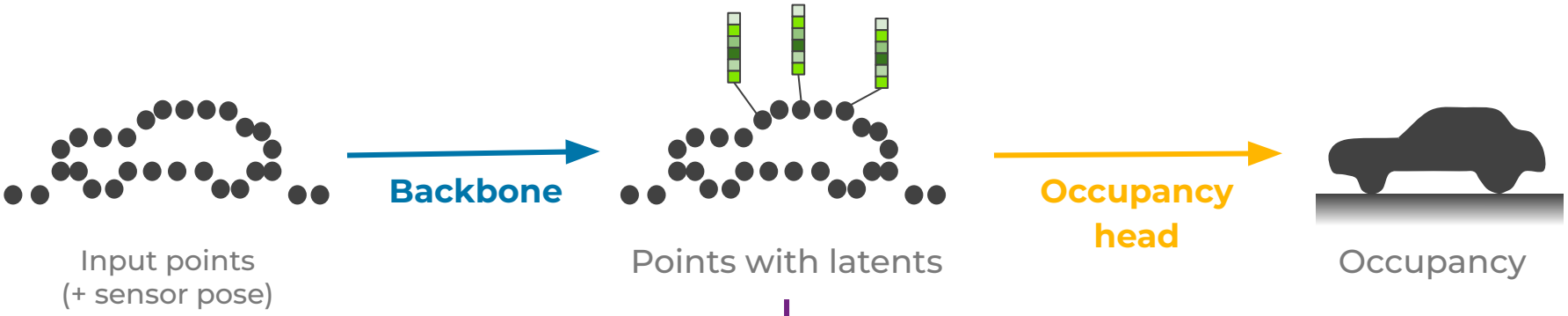
Along lidar lines of sight

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

Supervision



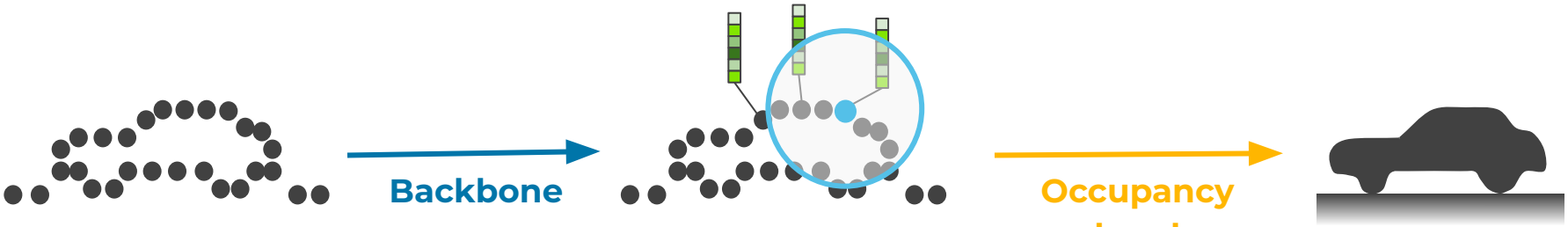
Downstream



Finetuning

- remove occupancy head
- add a single linear layer
- finetune the whole network

Downstream



Local reconstruction [POCO head]

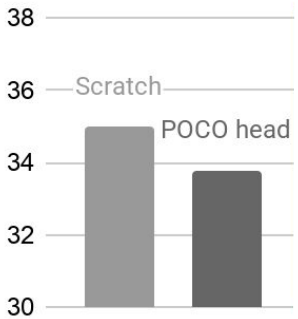
- everywhere, from features of neighboring points

⇒ (too) detailed geometry

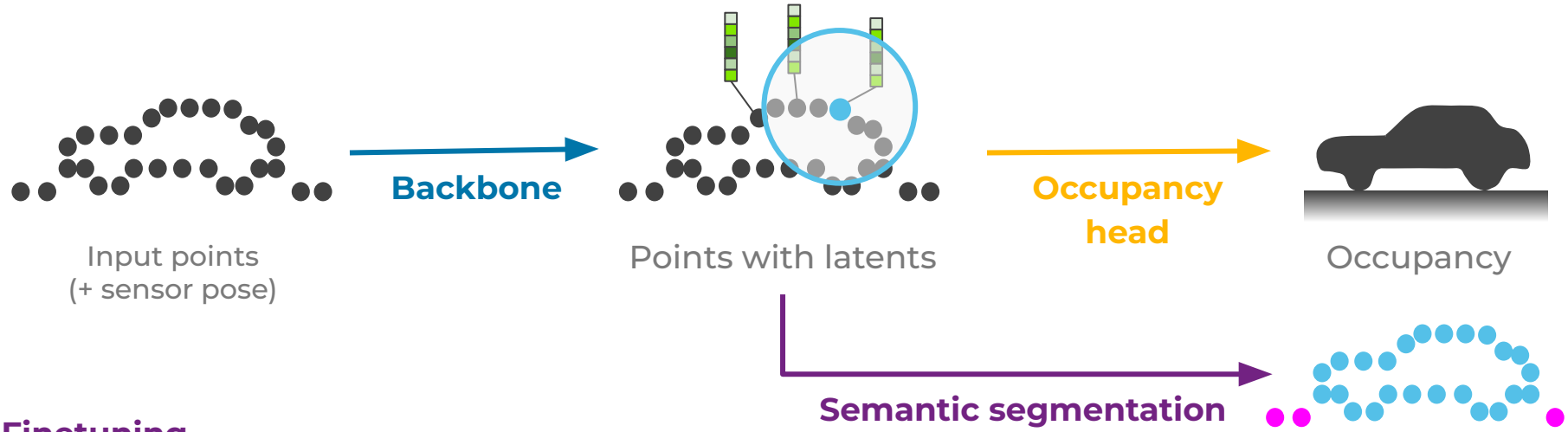
Context reconstruction [ALSO head]

- of a 1 meter ball, from each single feature point

⇒ rough geometry, more suited for object recognition



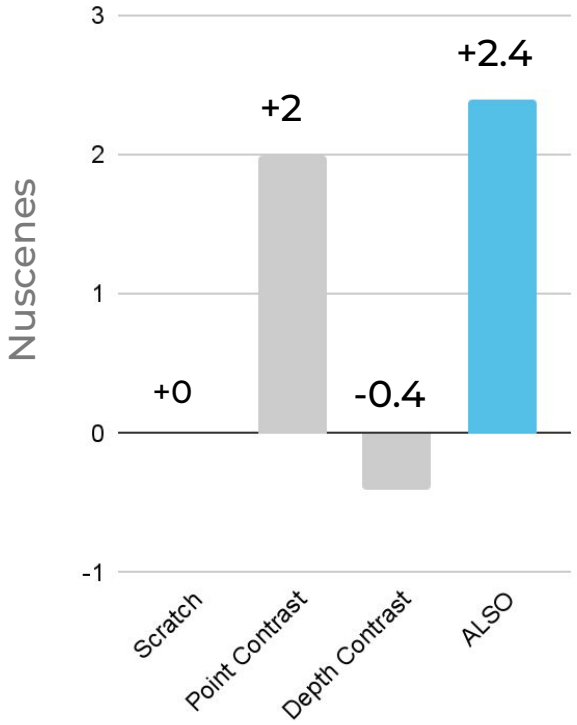
Downstream

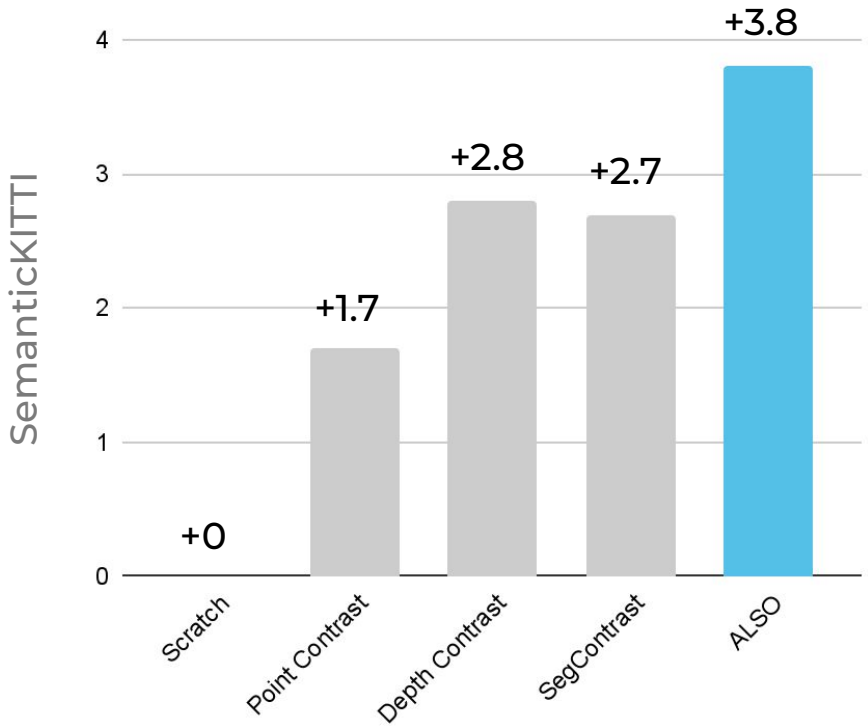


Finetuning

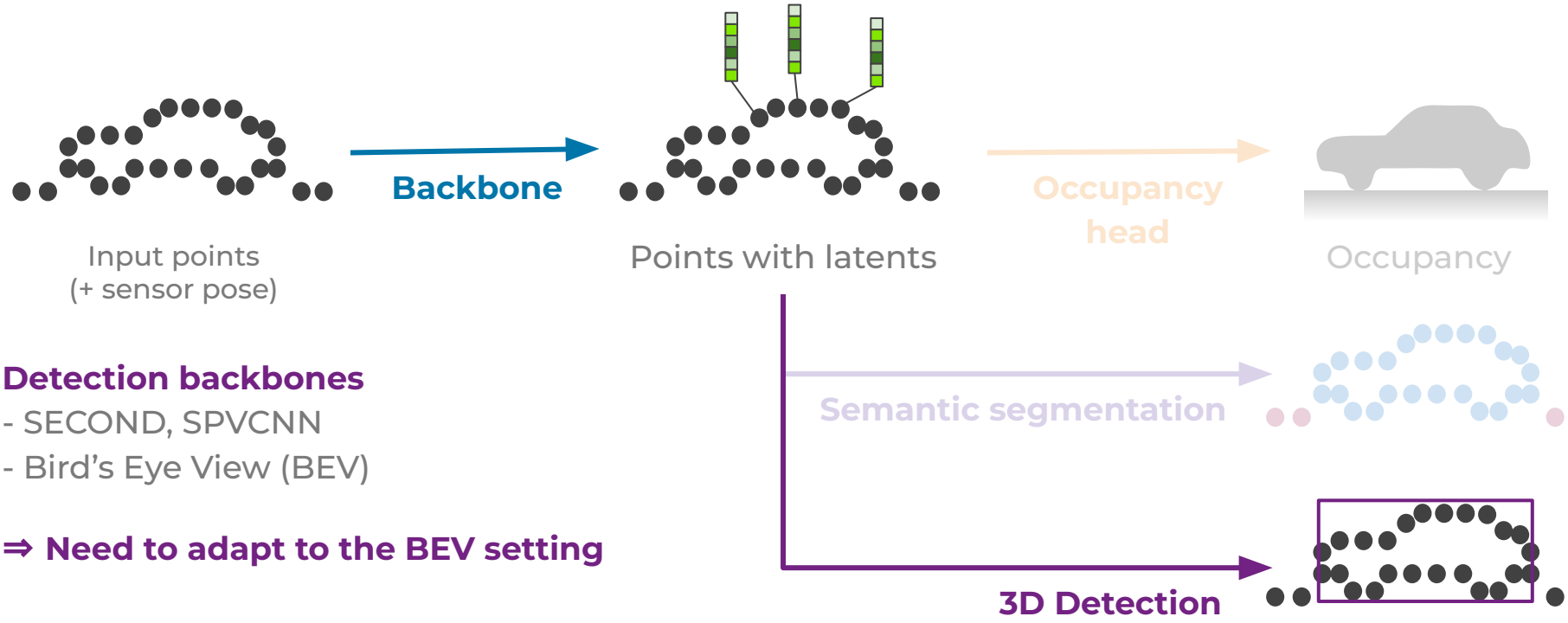
- remove occupancy head
- add a single linear layer
- finetune the whole network

Semantic segmentation - Finetuning 1% data





Downstream



Detection backbones

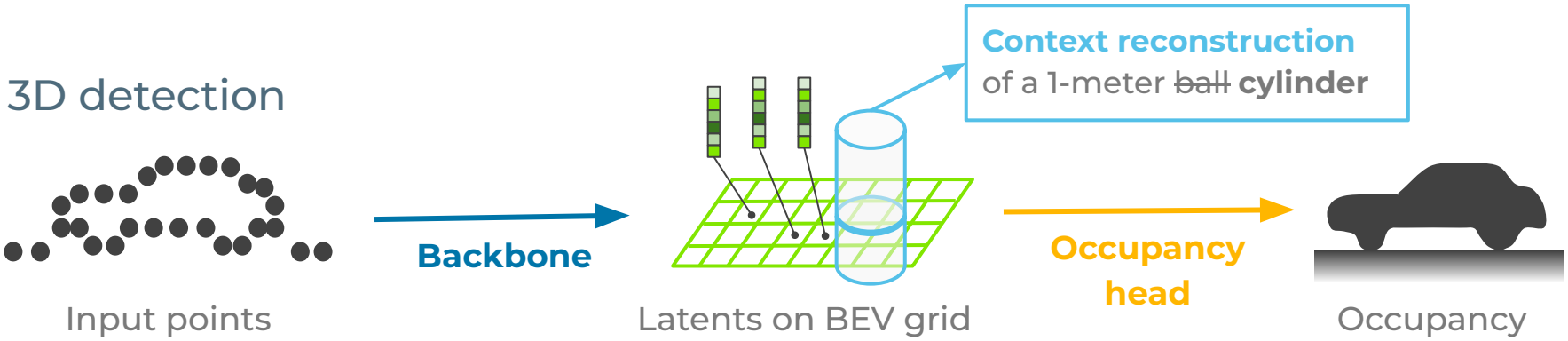
- SECOND, SPVCNN
- Bird's Eye View (BEV)

⇒ Need to adapt to the BEV setting

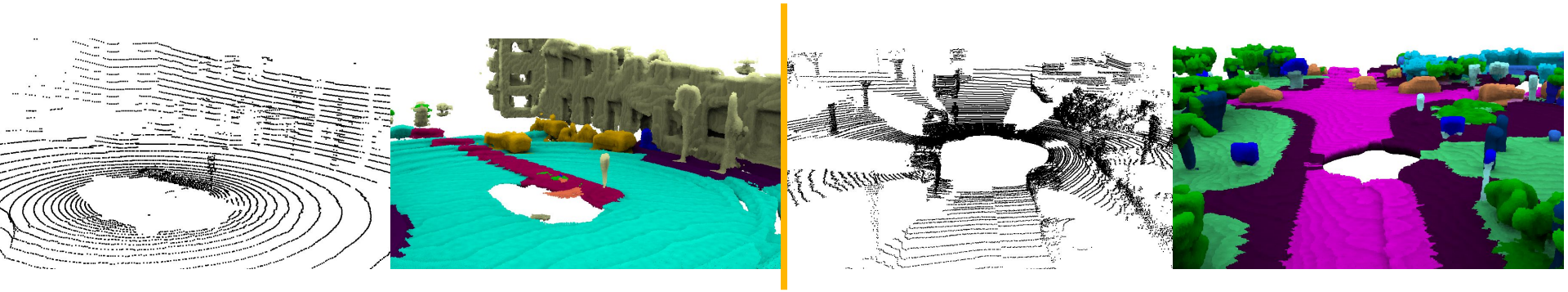
Pipeline for BEV networks

Semantic segmentation

3D detection



Surface-based self-supervision



Semantic labels on top of estimated occupancy (nuScenes and SemanticKITTI)

Scene reconstruction as a self-supervised pretext task

- ✔ Competitive results
- ✔ Single stream
  ⇒ memory efficient
- ✔ Sensor information can be leveraged for pretext tasks
- ✘ Latent space structure not suited for direct linear probing

  Similar as MAE

Lidar self supervision

1. Self-supervision
2. Geometric pretext task
3. Contrastive approaches
4. Self-distillation

Contrastive approaches

Car



Network

Z_1

Dissimilar

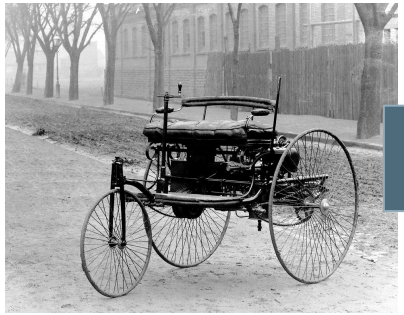
Z_3

Network



Bus

Similar

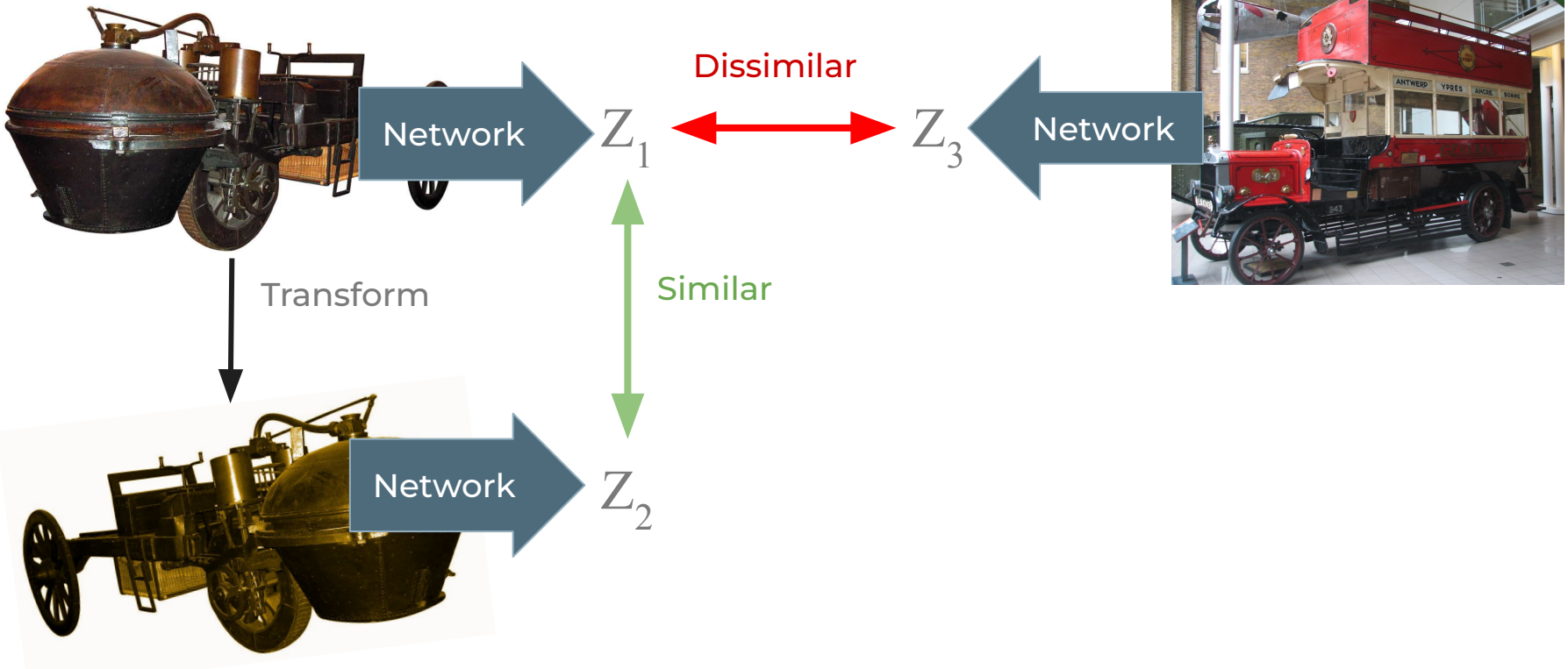


Network

Z_2

Car

Contrastive approaches



Contrastive approaches (InfoNCE loss)

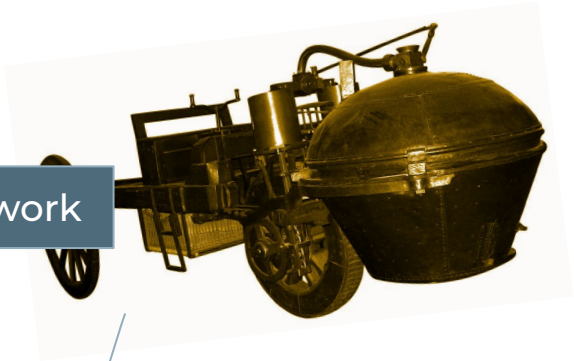


Network

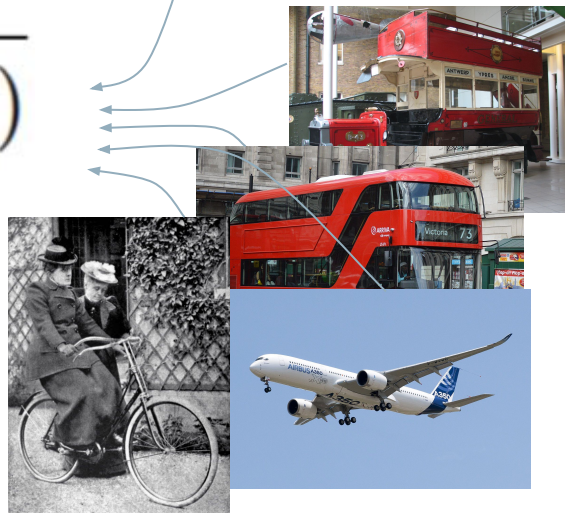
q

k_+

Network

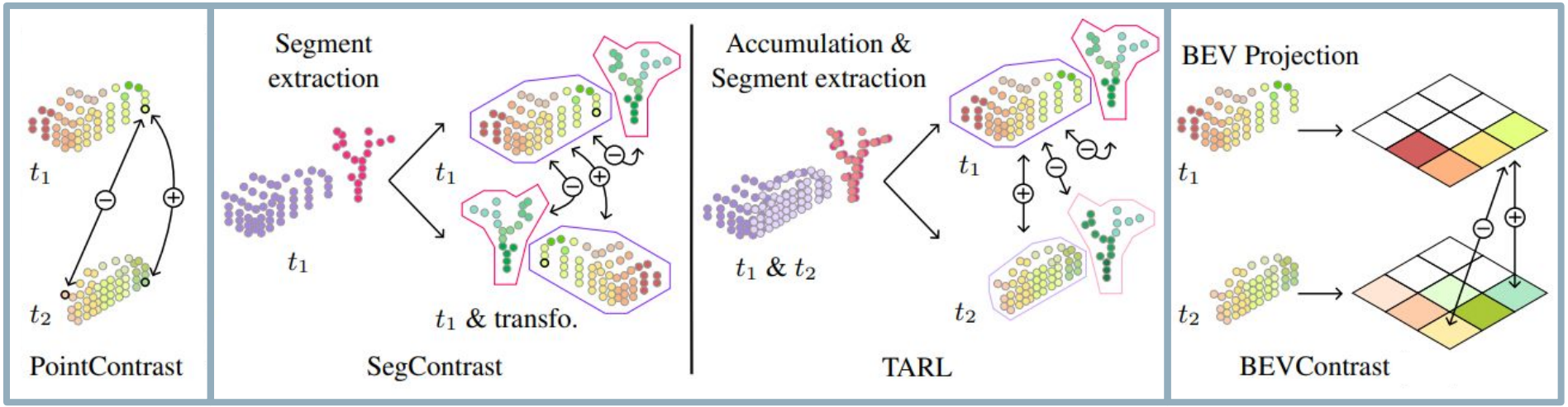


$$\mathcal{L}_q = -\log \frac{\exp(q \cdot k_+ / \tau)}{\sum_{i=0}^{K} \exp(q \cdot k_i / \tau)}$$



Contrastive self-supervised learning for point clouds

BEVContrast



PointContrast

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

SegContrast / TARL

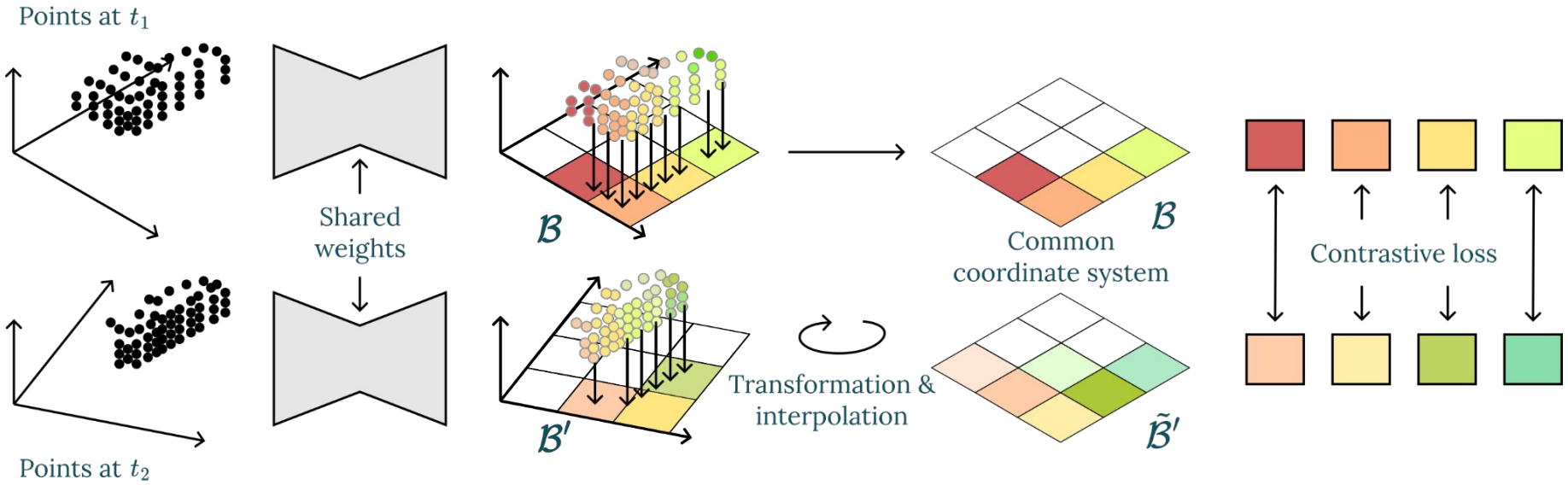
+++ Efficient thanks to object segmentation (temporal for TARL)
- - - Difficult to set up → rely on HDBScan (hyperparameters)

BEVContrast

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

BEVContrast: Main idea

BEVContrast's self-supervised learning principle



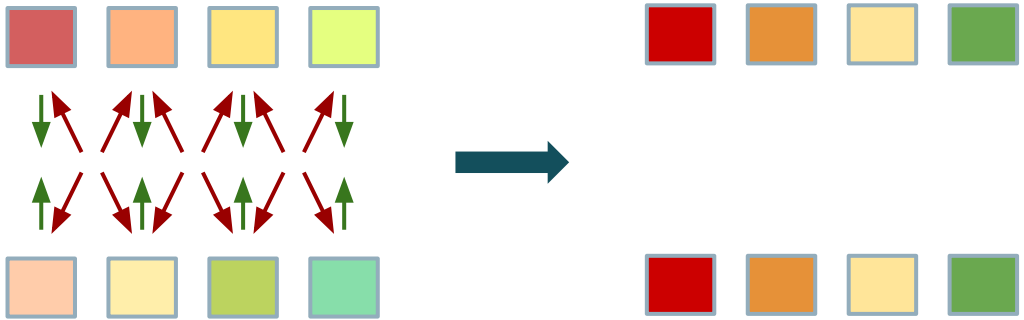
BEVContrast: InfoNCE Loss

$$\mathcal{L}_i = -\log \frac{\exp(\text{sim}(f_i, g_i)/\tau)}{\sum_{k=0}^{N-1} \exp(\text{sim}(f_i, g_j)/\tau)}$$

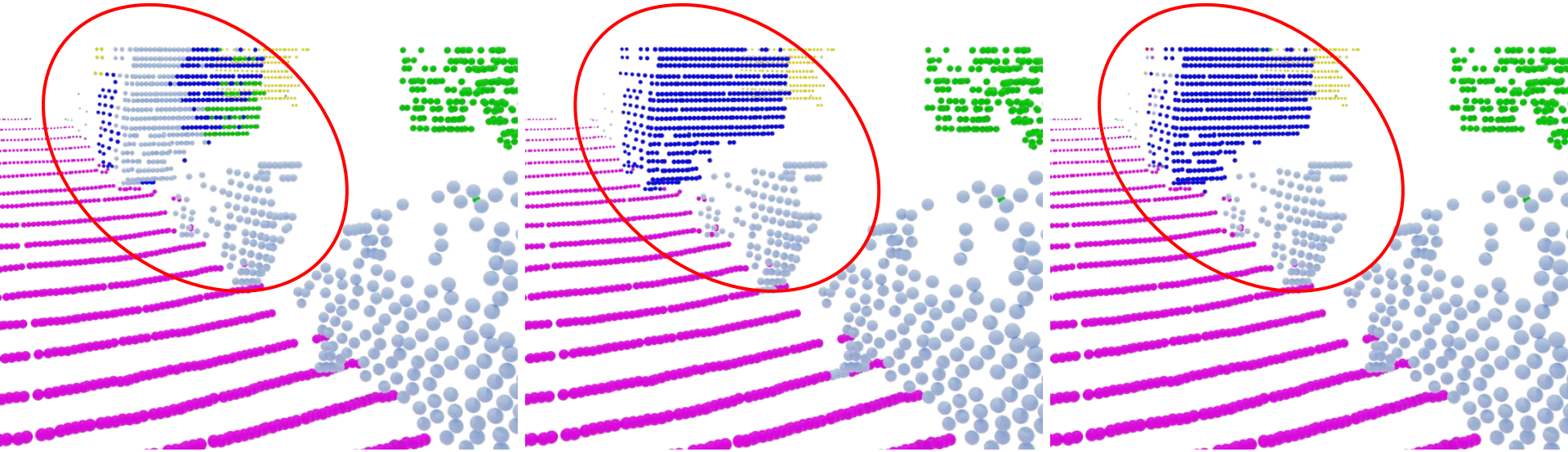
f, g are feature vectors

f_i, g_i is a positive pair

τ is a temperature parameter



BEVContrast: Data efficiency objective

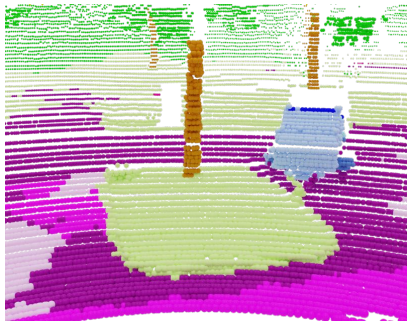


TARL | Ground Truth | BEVContrast (Ours)

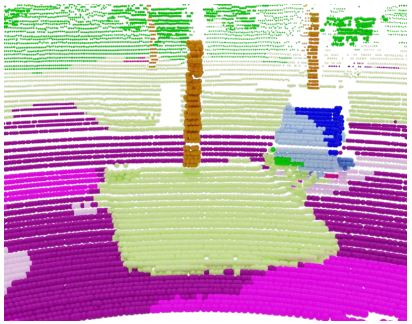
- Truck
- Car
- Road
- Vegetation

Training with limited annotations (1% SemanticKITTI)

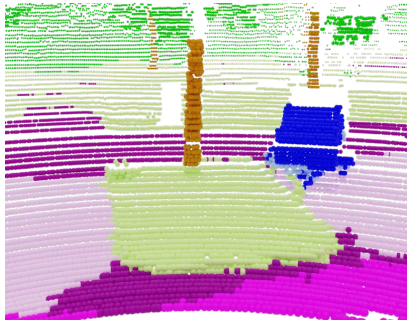
BEVContrast: Low data performance



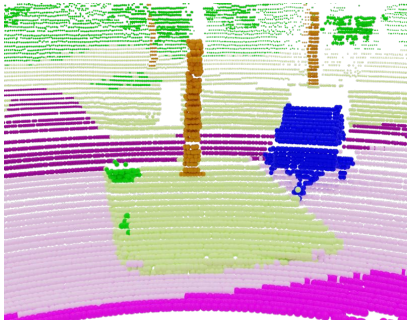
ALSO



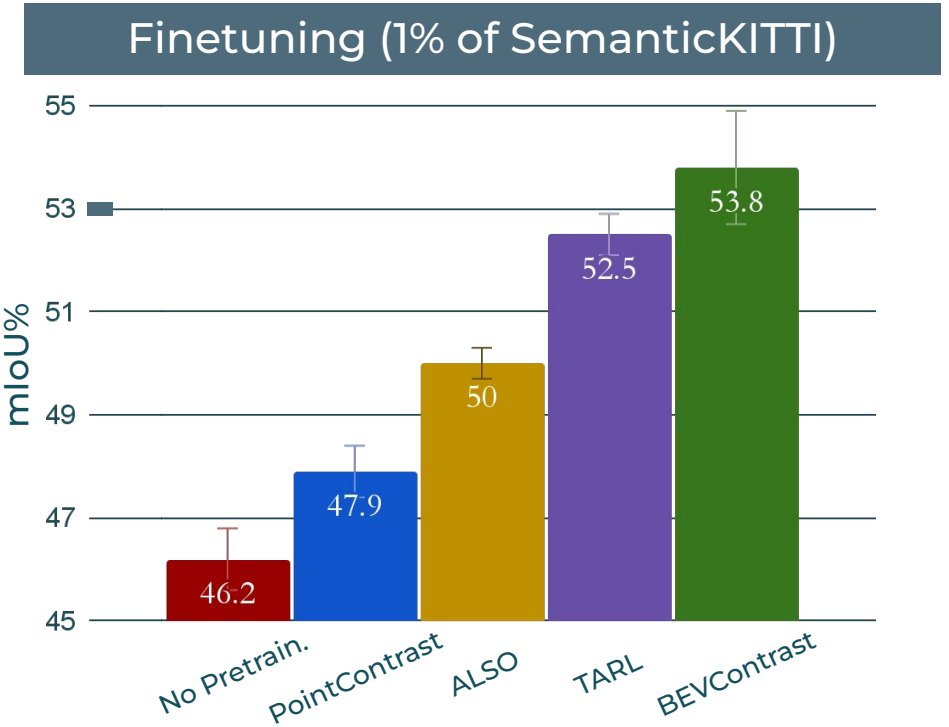
TARL



BEVContrast



Ground truth



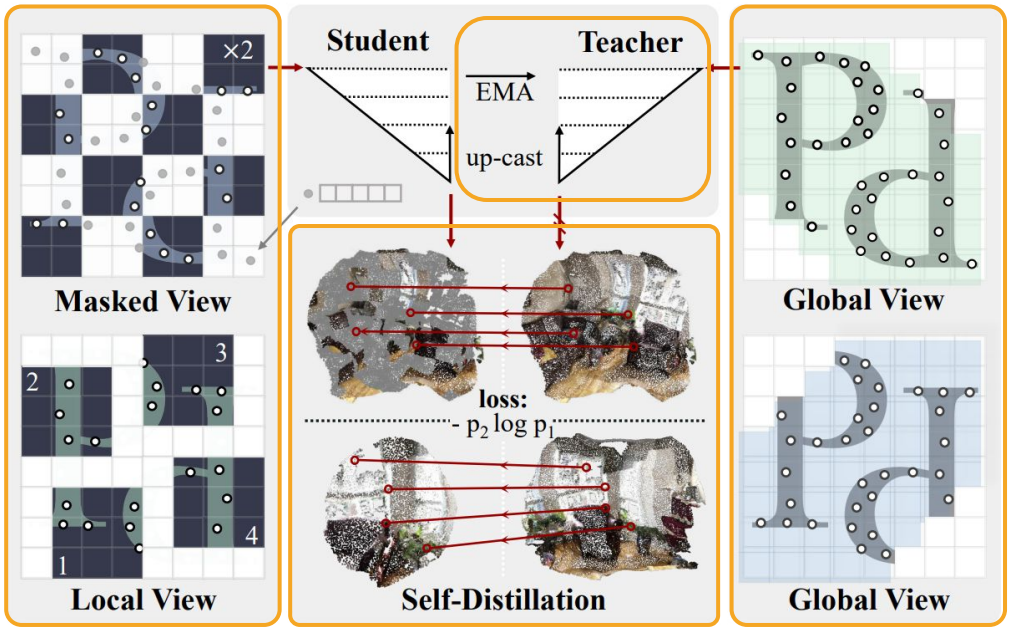
Boulch et al. ALSO: Automotive lidar self supervision by occupancy estimation. CVPR 2023
Nunes et al. Temporal consistent 3D lidar representation learning for semantic perception in autonomous driving. CVPR 2023

Lidar self supervision

1. Self-supervision
2. Geometric pretext task
3. Contrastive approaches
4. Self-distillation

Sonata: Self-Supervised Learning of Reliable Point Representations

DINO self-distillation applied to point clouds



- **Teacher**
  A slow evolving version of the student (EMA), no gradient update
- **Different views** for the student and the teacher
- They should **predict** the same features

Warning: retrain models (they were pretrain with test sets) ! (Same for Utonia and Concerto)

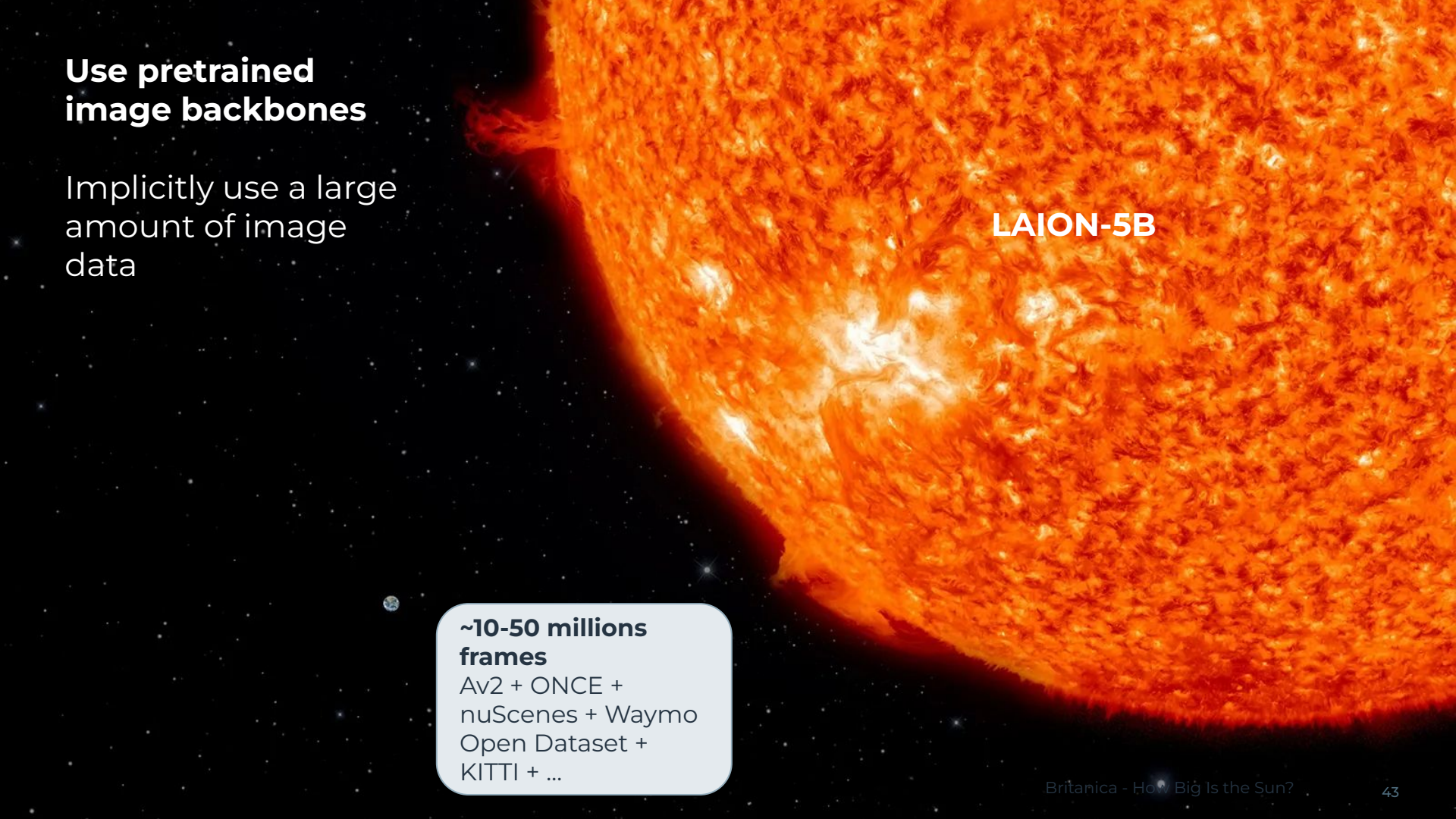
Sonata: Self-Supervised Learning of Reliable Point Representations, Wu et al, CVPR 2025

Distillation from images

1. A first distillation approach
2. Scaling up the distillation
3. Geometry and distillation

Use pretrained image backbones

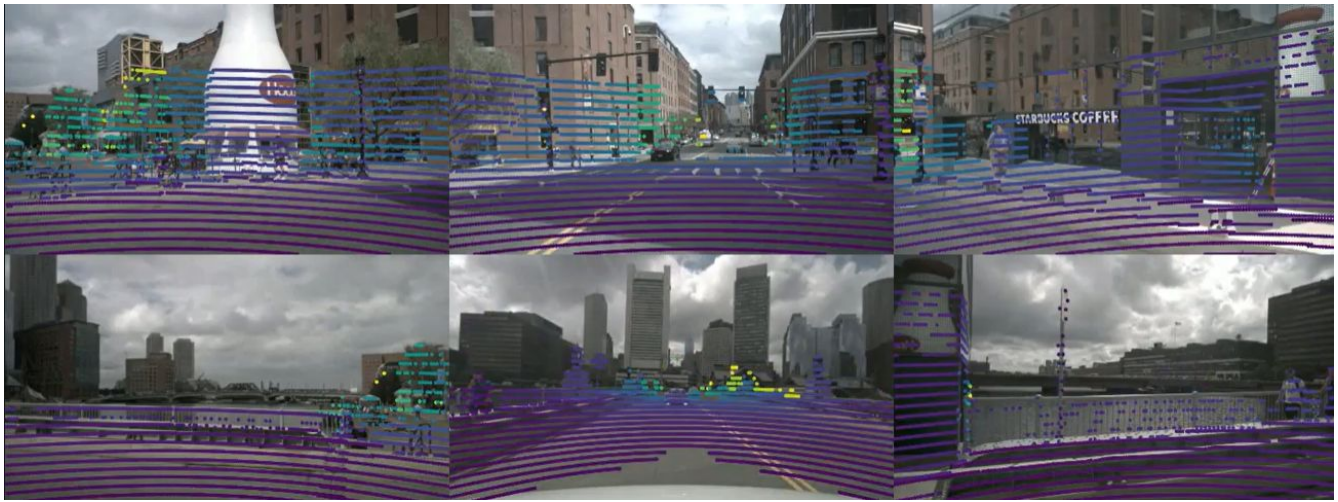
Implicitly use a large amount of image data



~10-50 millions frames
Av2 + ONCE + nuScenes + Waymo Open Dataset + KITTI + ...

Britanica - How Big Is the Sun?

Setup



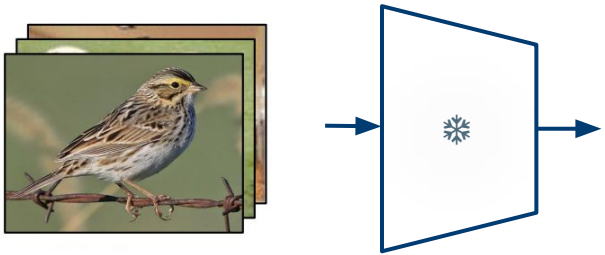
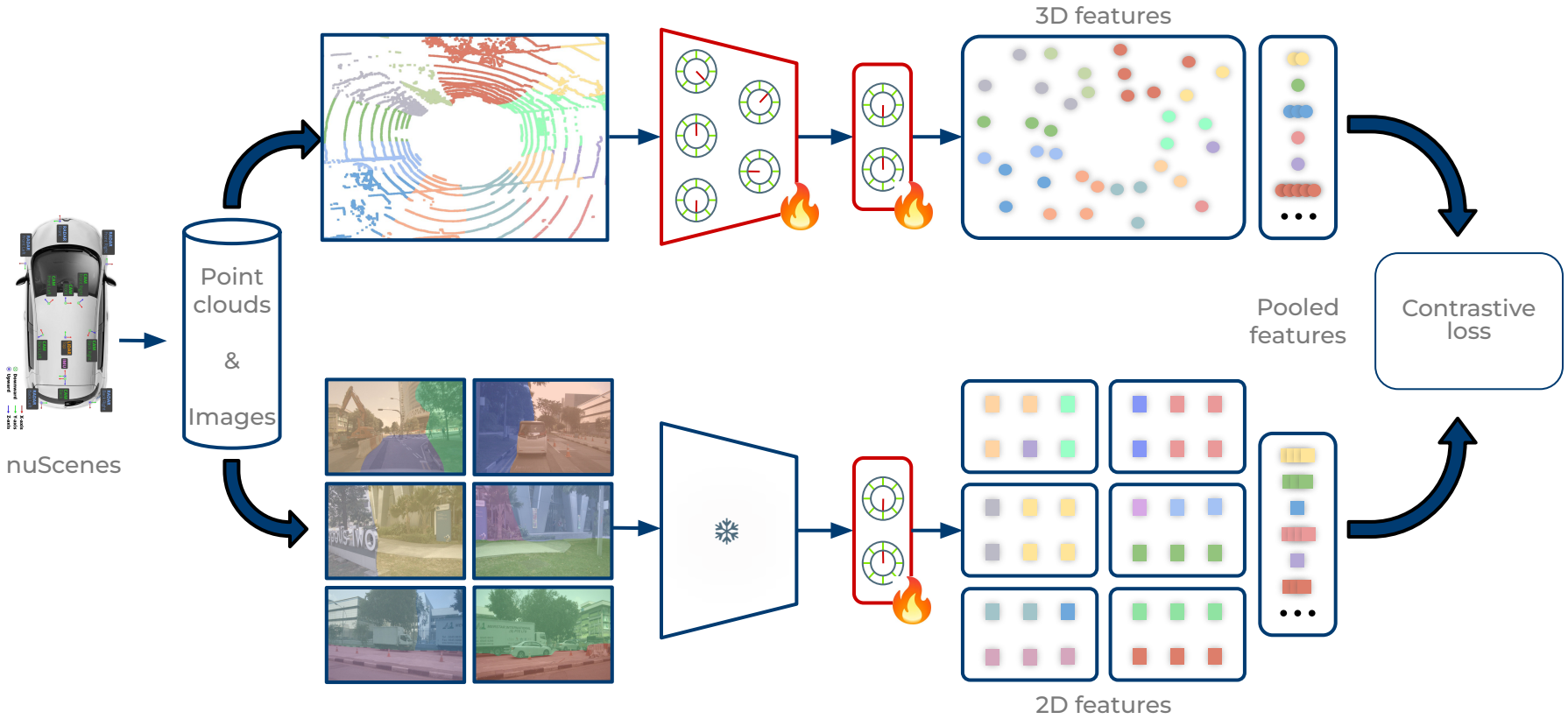


Image foundation model, e.g., DINOv2

Pretrained on natural images

H. Caesar et al., nuScenes: A multimodal dataset for autonomous driving. CVPR, 2020.
M. Oquab et al., DINOv2: Learning Robust Visual Features without Supervision. TMLR, 2024.

SLidR pretraining

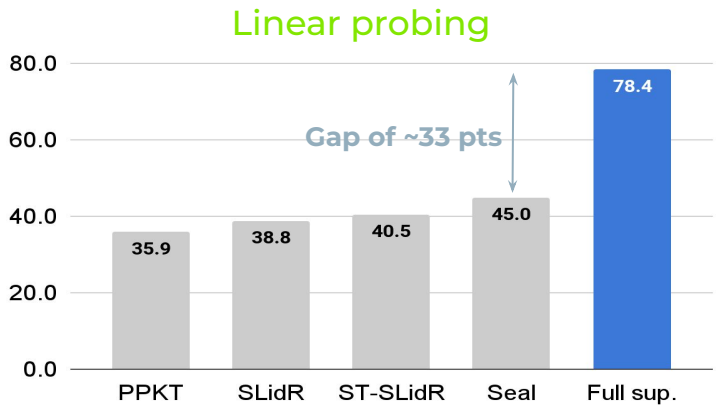


[SLidR] C. Sautier et al. Image-to-lidar self-supervised distillation for autonomous driving data. CVPR, 2022.

SLidR

Limitation

Distilled 3D features << supervised 3D features



[PPKT] Y.-C. Liu et al. Learning from 2D: Pixel-to-point knowledge transfer for 3D pretraining. arxiv:2104.04687, 2021
[SLidR] C. Sautier et al. Image-to-lidar self-supervised distillation for autonomous driving data. CVPR, 2022.
[ST-SLidR] A. Mahmoud et al., Self-supervised image-to-point distillation via semantically tolerant contrastive loss. CVPR, 2023.
[Seal] Y. Liu et al. Segment Any Point Cloud Sequences by Distilling Vision Foundation Models. NeurIPS, 2023.

Distillation from images

1. A first distillation approach
2. Scaling up the distillation
3. Geometry and distillation

SLidR

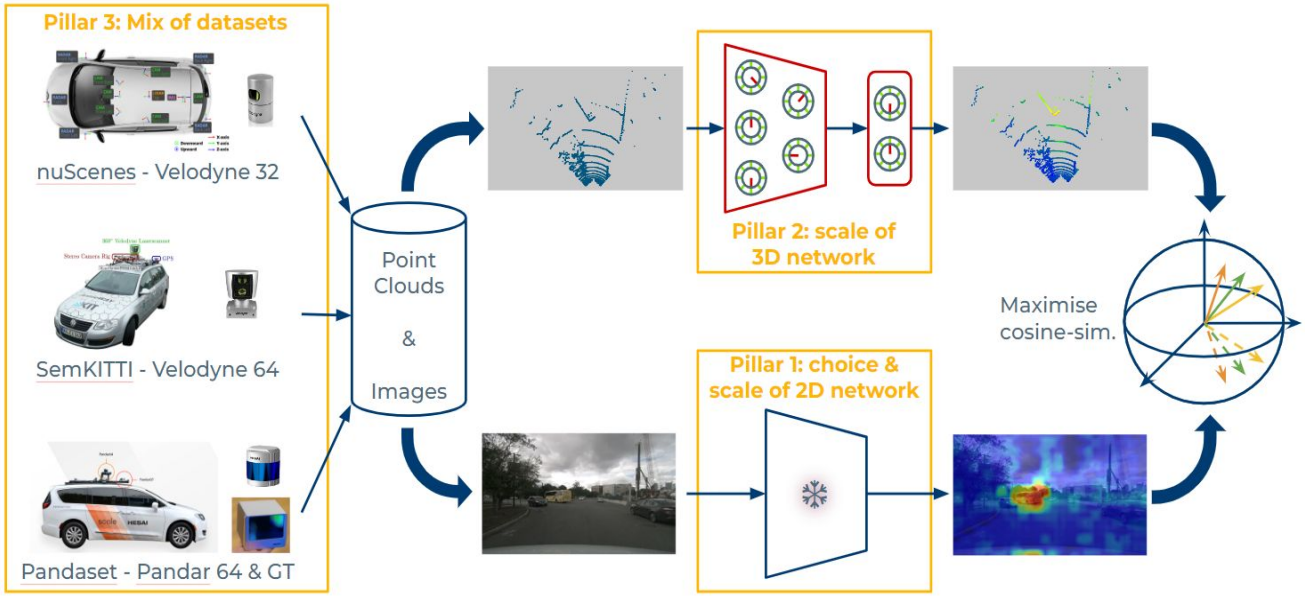
Limitation

Distilled 3D features << supervised 3D features

Axes of improvement

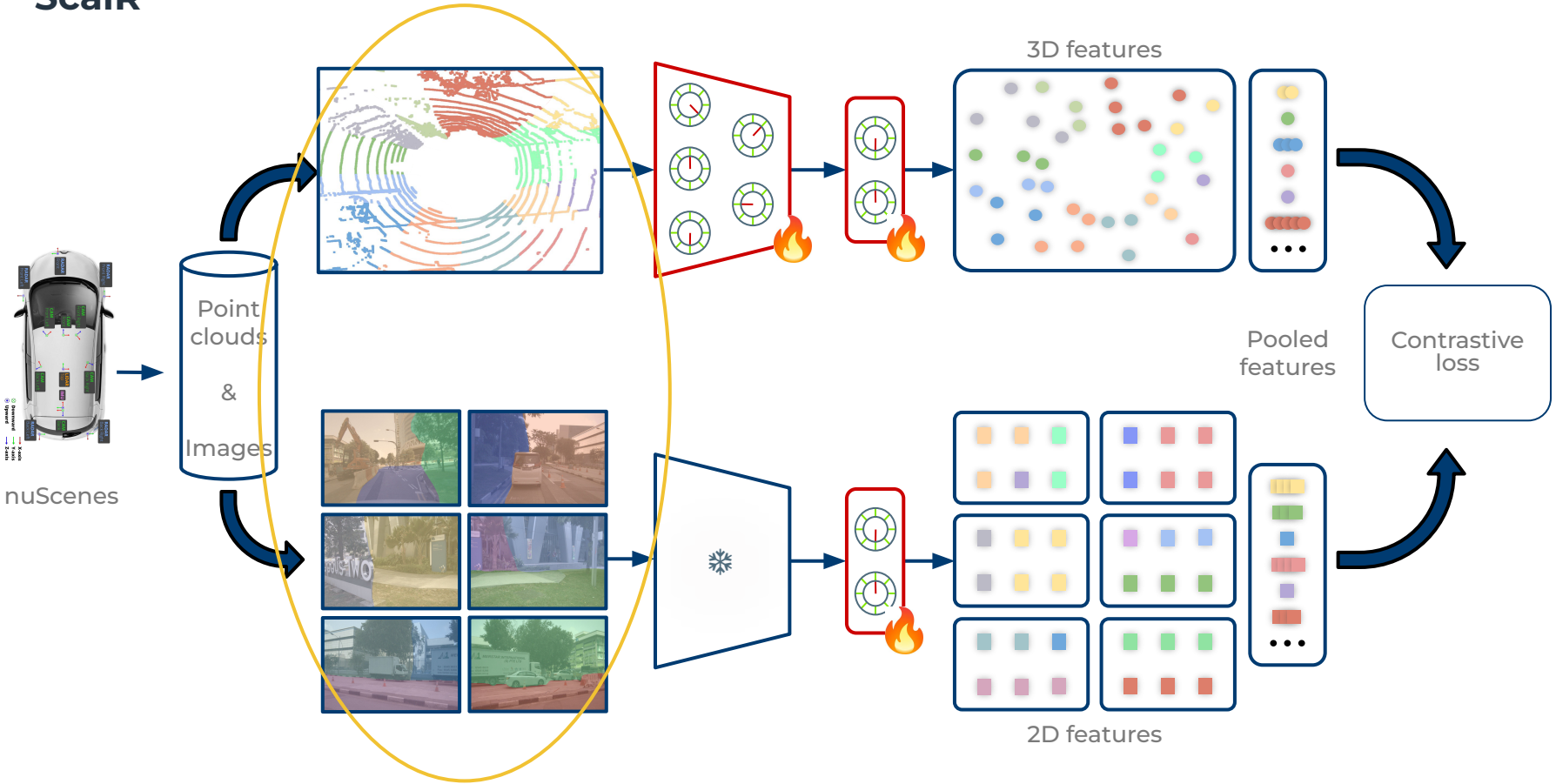
Study effect of 3 pillars:

- 2D backbone
- 3D backbone
- Datasets

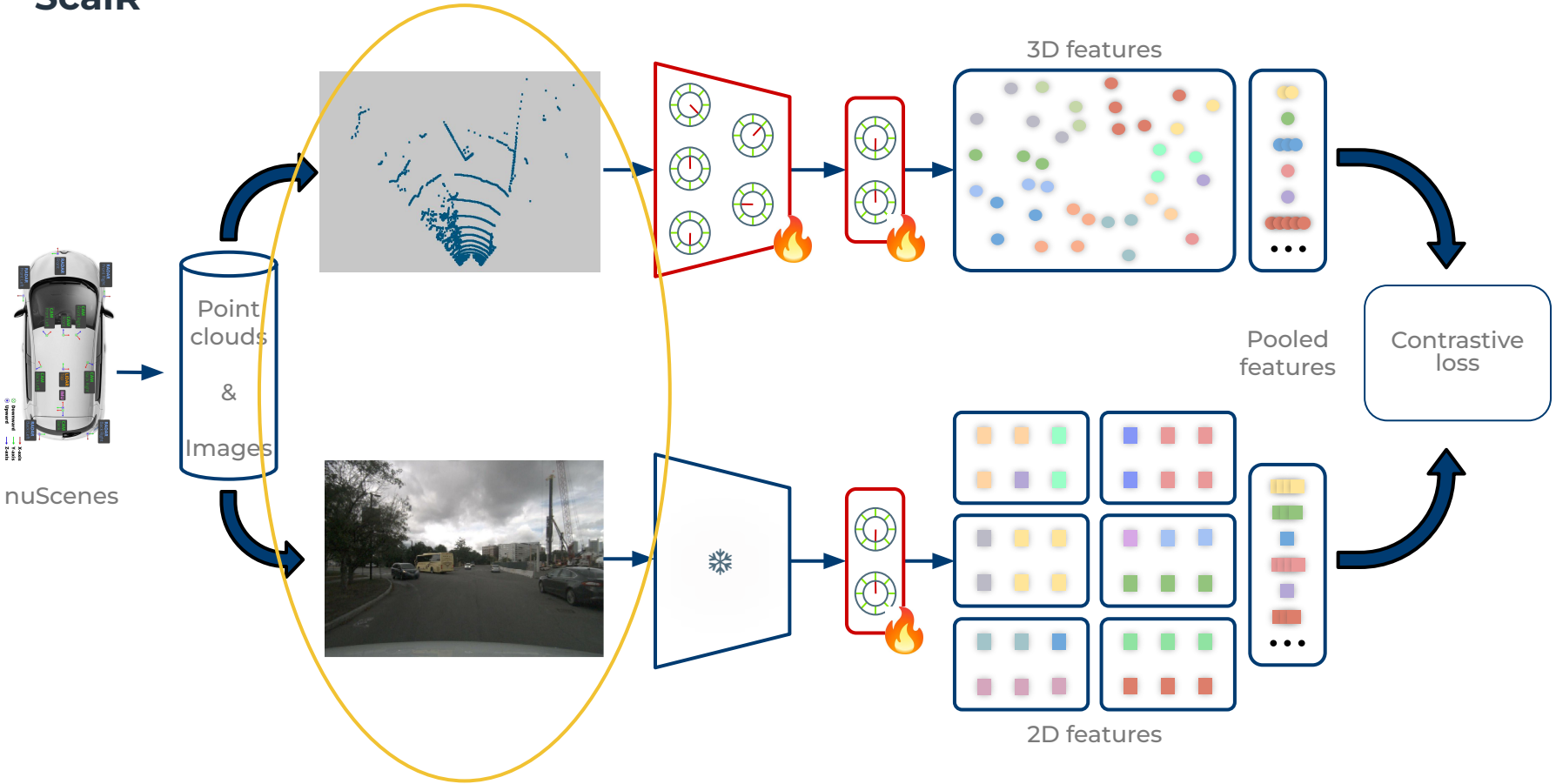


Three Pillars improving Vision Foundation Model Distillation for Lidar, Puy et al, CVPR 2024

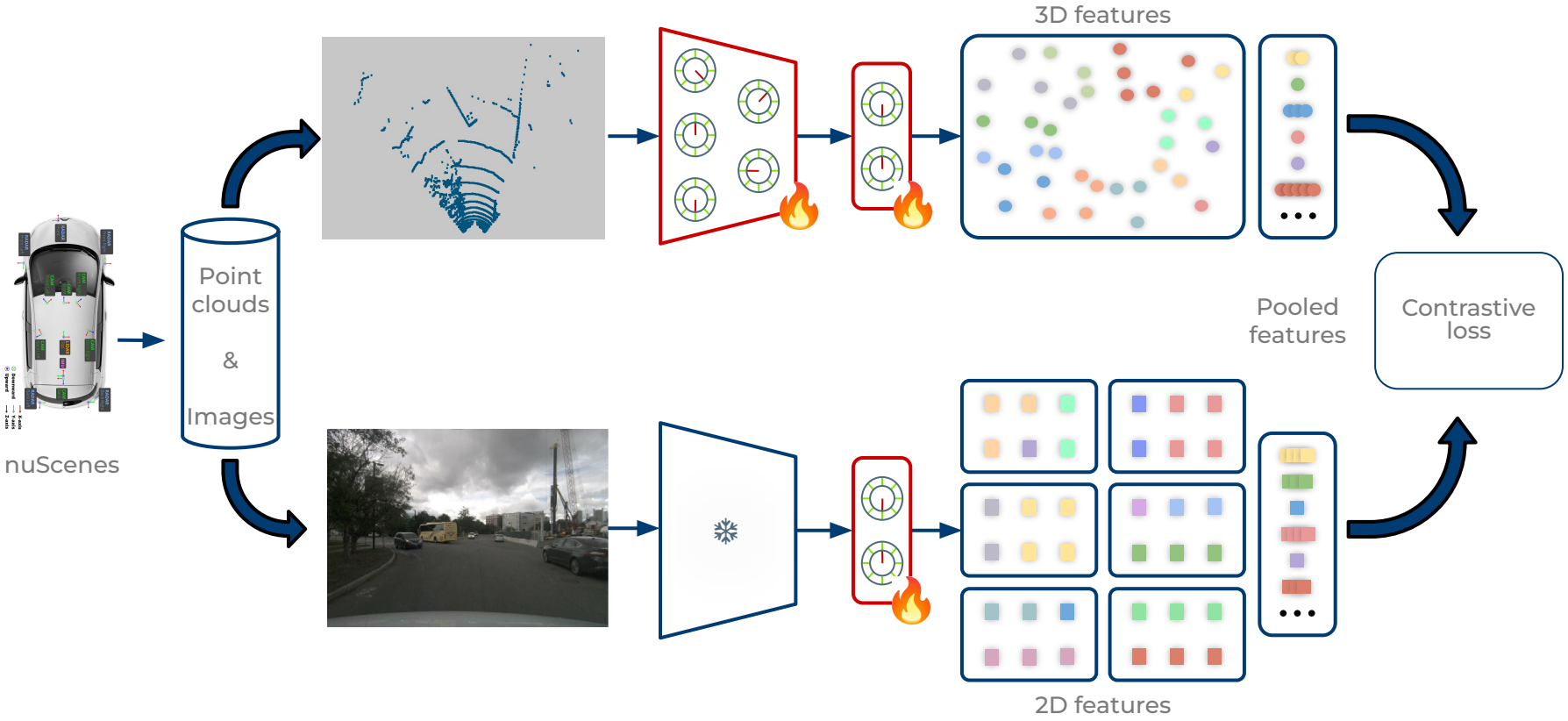
ScalR


3D features
Point clouds & Images
nuScenes
Pooled features
Contrastive loss
2D features

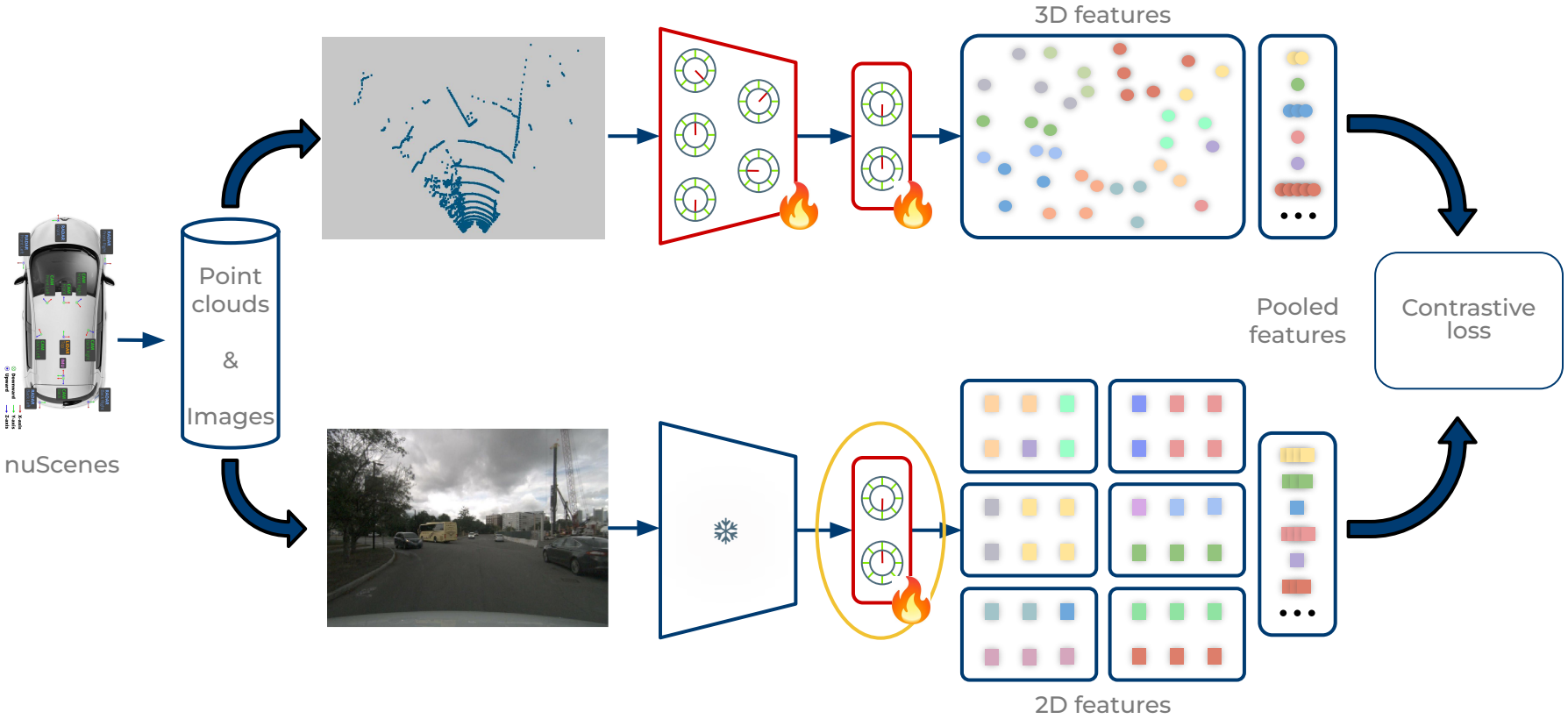
ScalR


3D features
Point clouds & Images
nuScenes
Pooled features
Contrastive loss
2D features

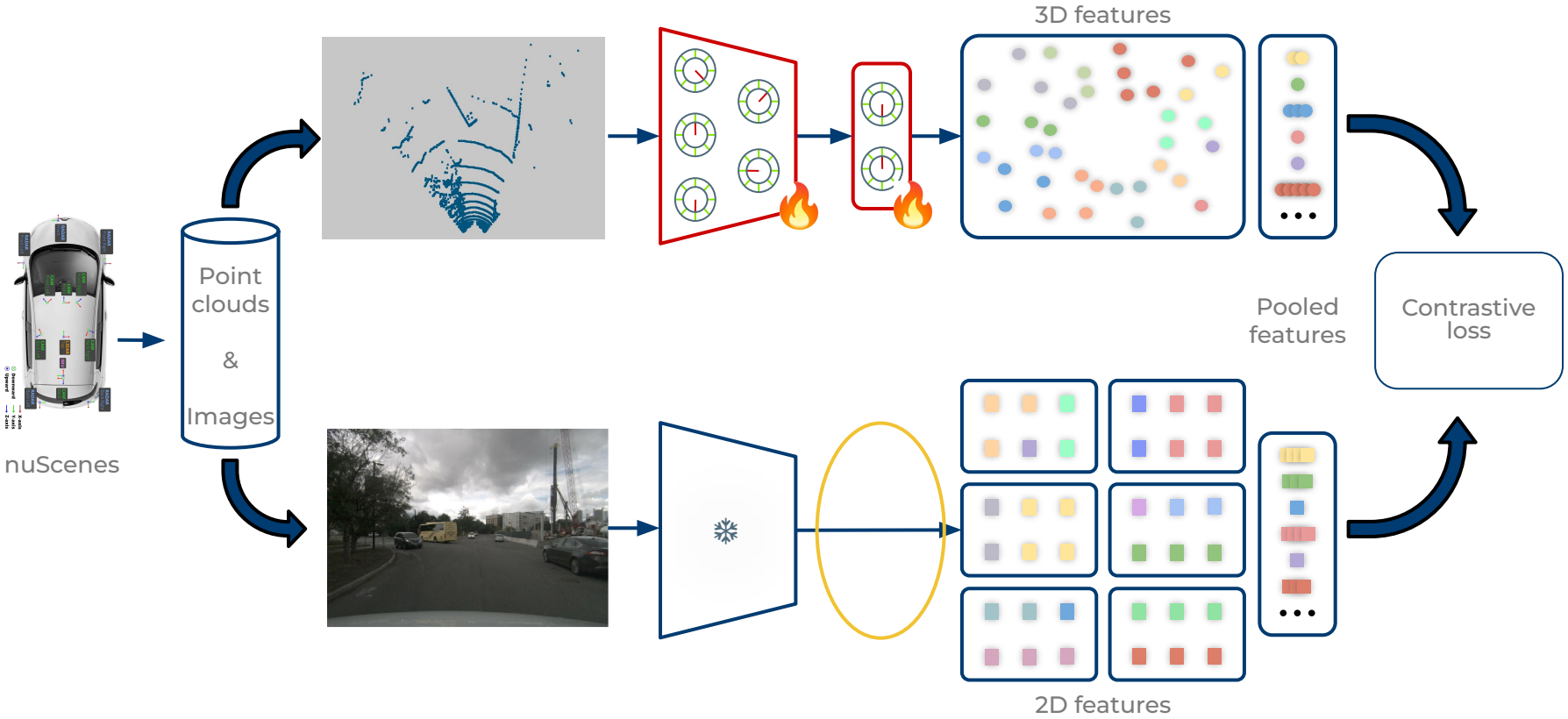
ScalR


3D features
Point clouds & Images
nuScenes
Pooled features
Contrastive loss
2D features

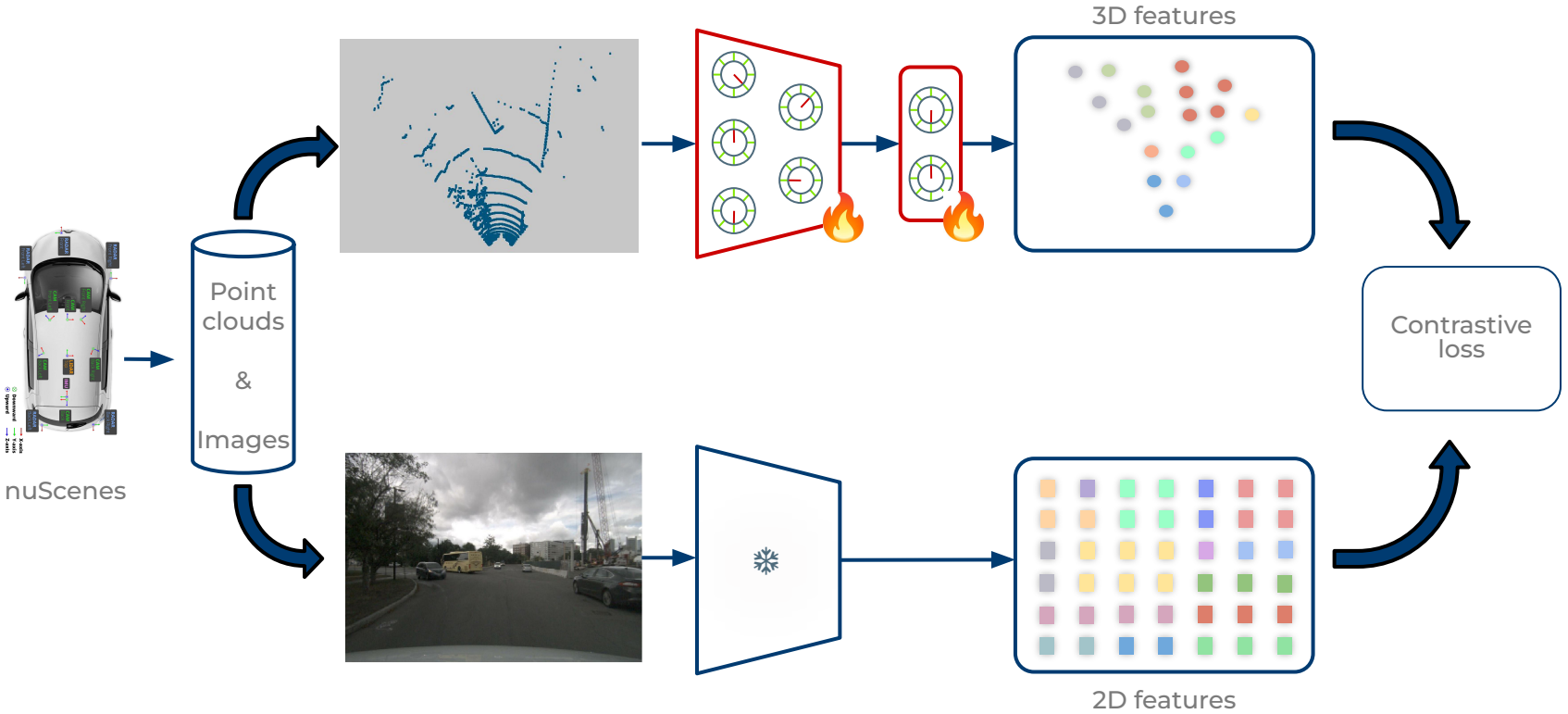
ScaIR


3D features
Point clouds & Images
nuScenes
Pooled features
Contrastive loss
2D features

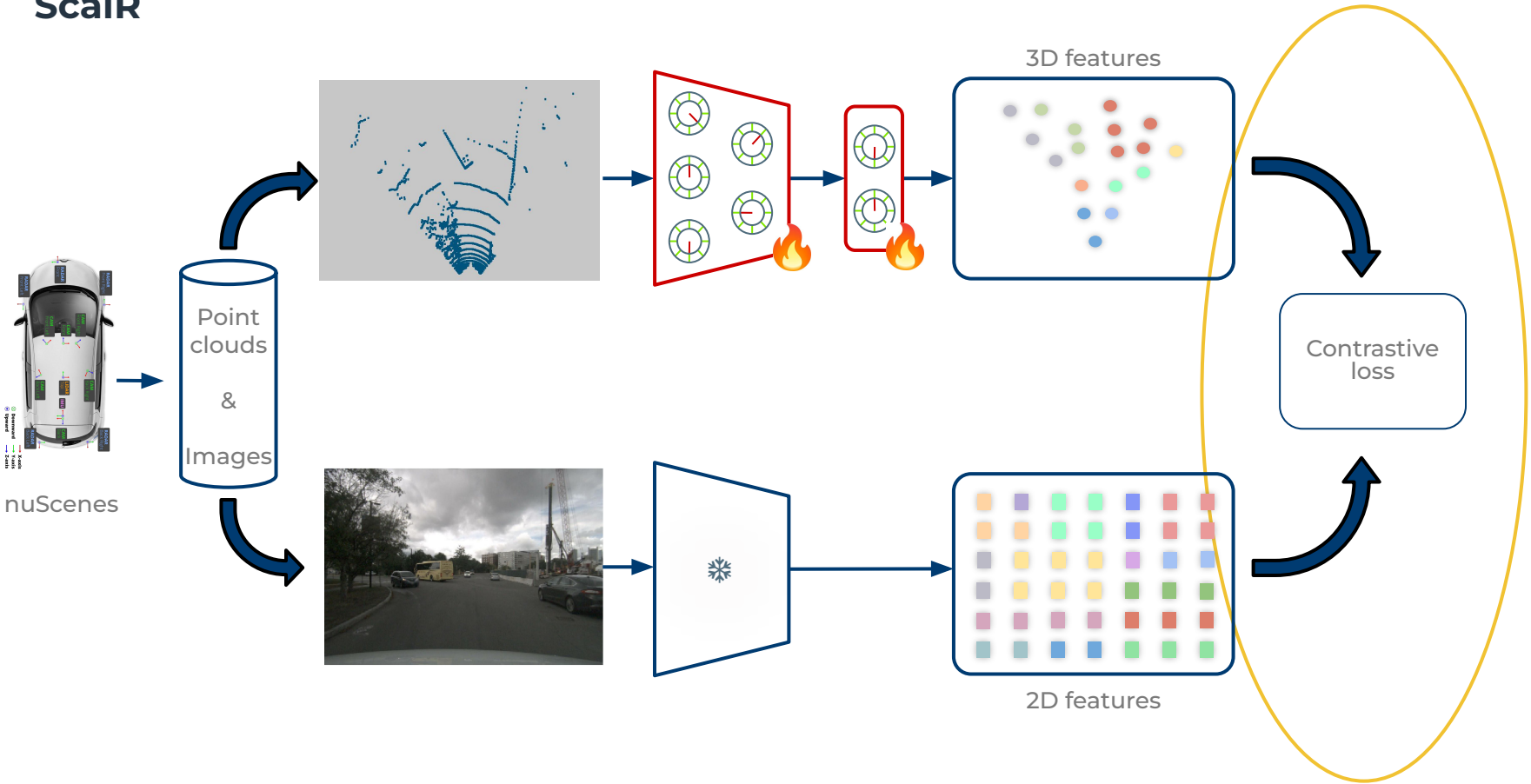
ScalR


3D features
Point clouds & Images
nuScenes
Pooled features
Contrastive loss
2D features

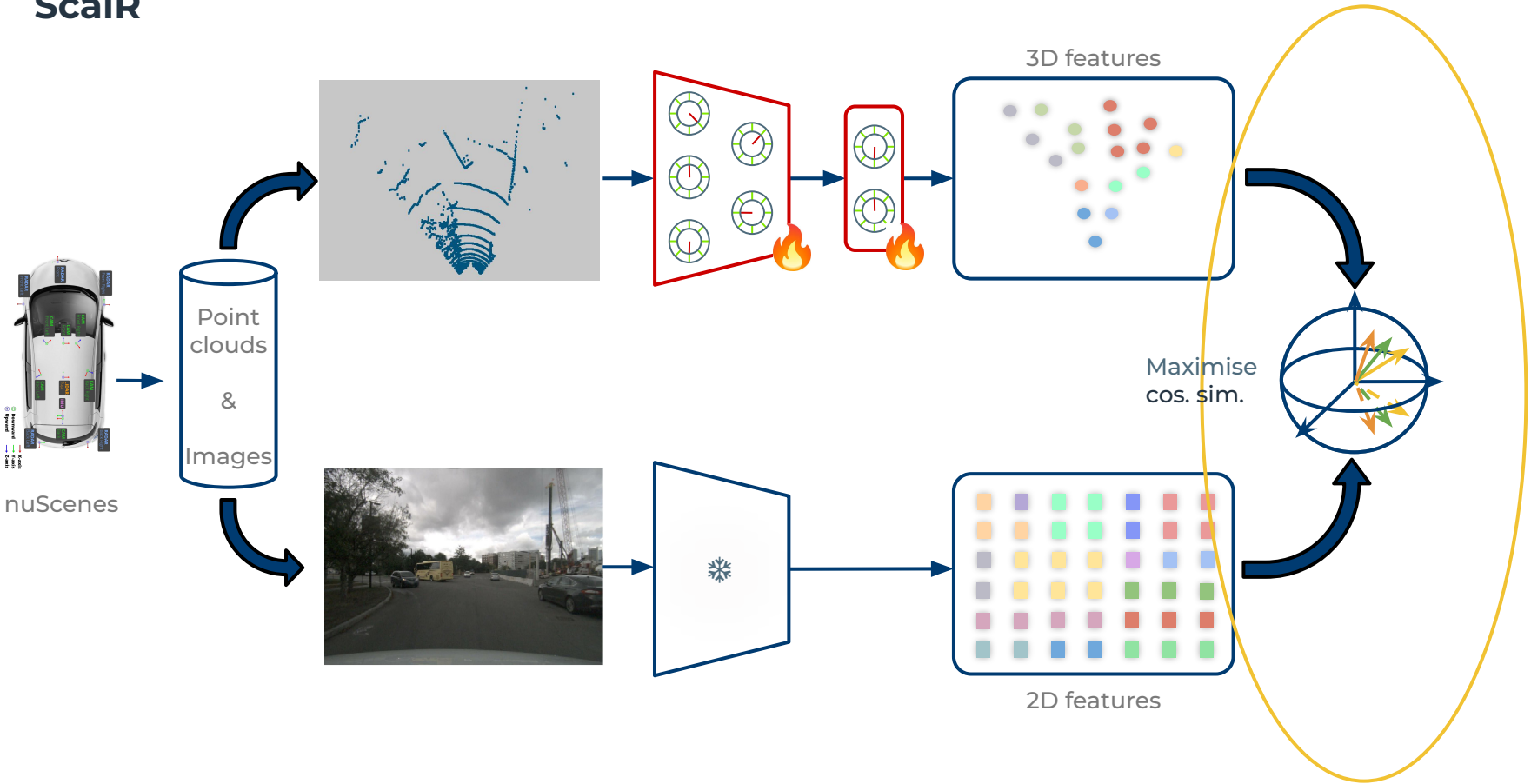
ScaIR


3D features
Point clouds & Images
nuScenes
Contrastive loss
2D features

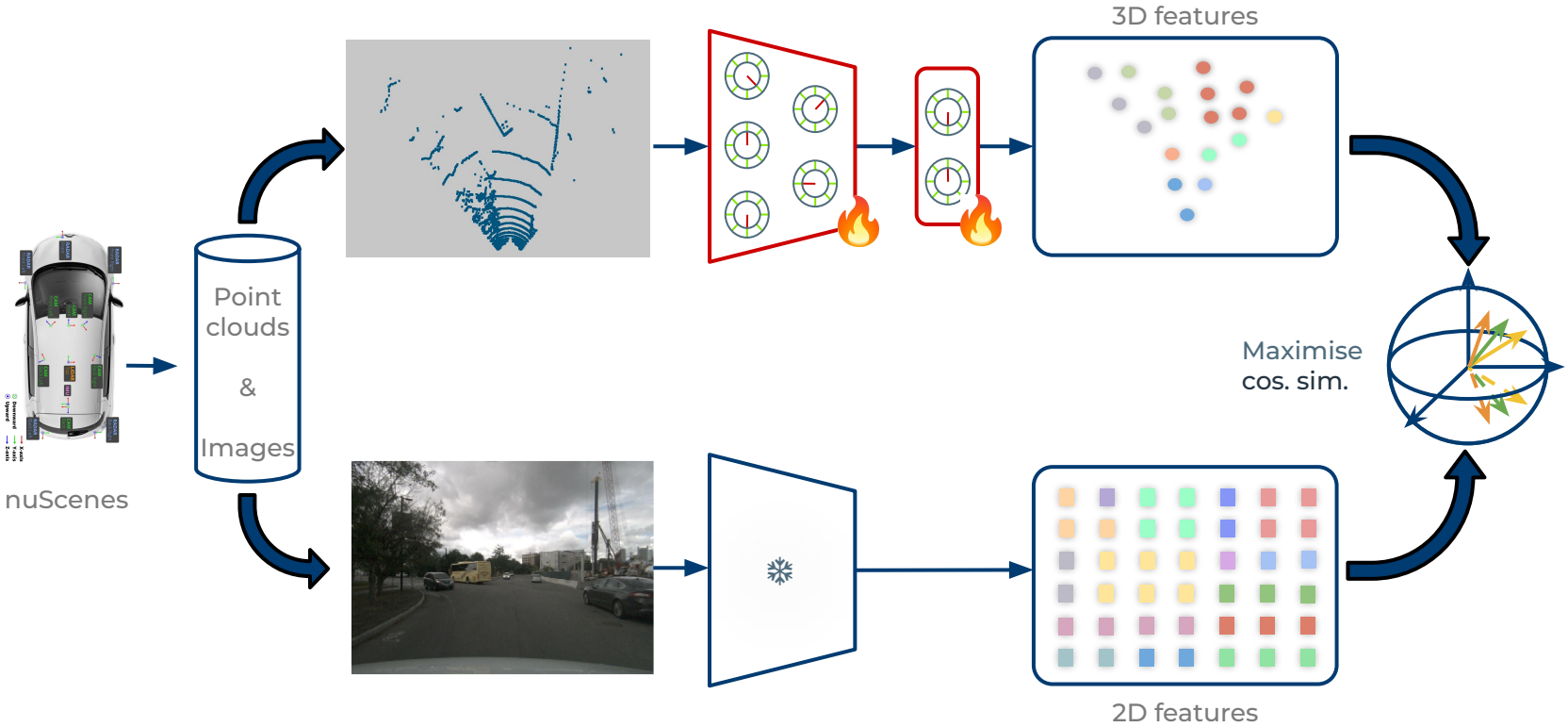
ScalR


3D features
Point clouds & Images
nuScenes
Contrastive loss
2D features

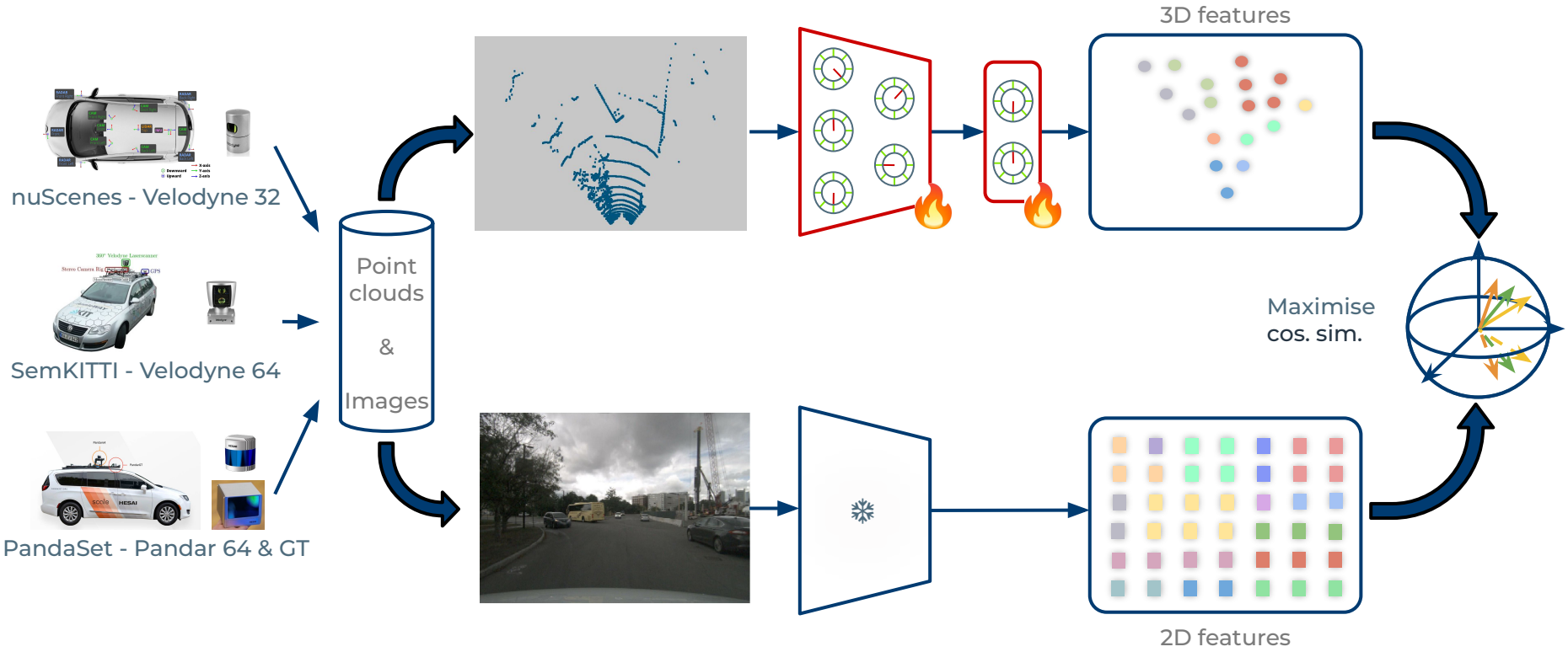
ScaIR


3D features
Point clouds & Images
nuScenes
Maximise cos. sim.
2D features

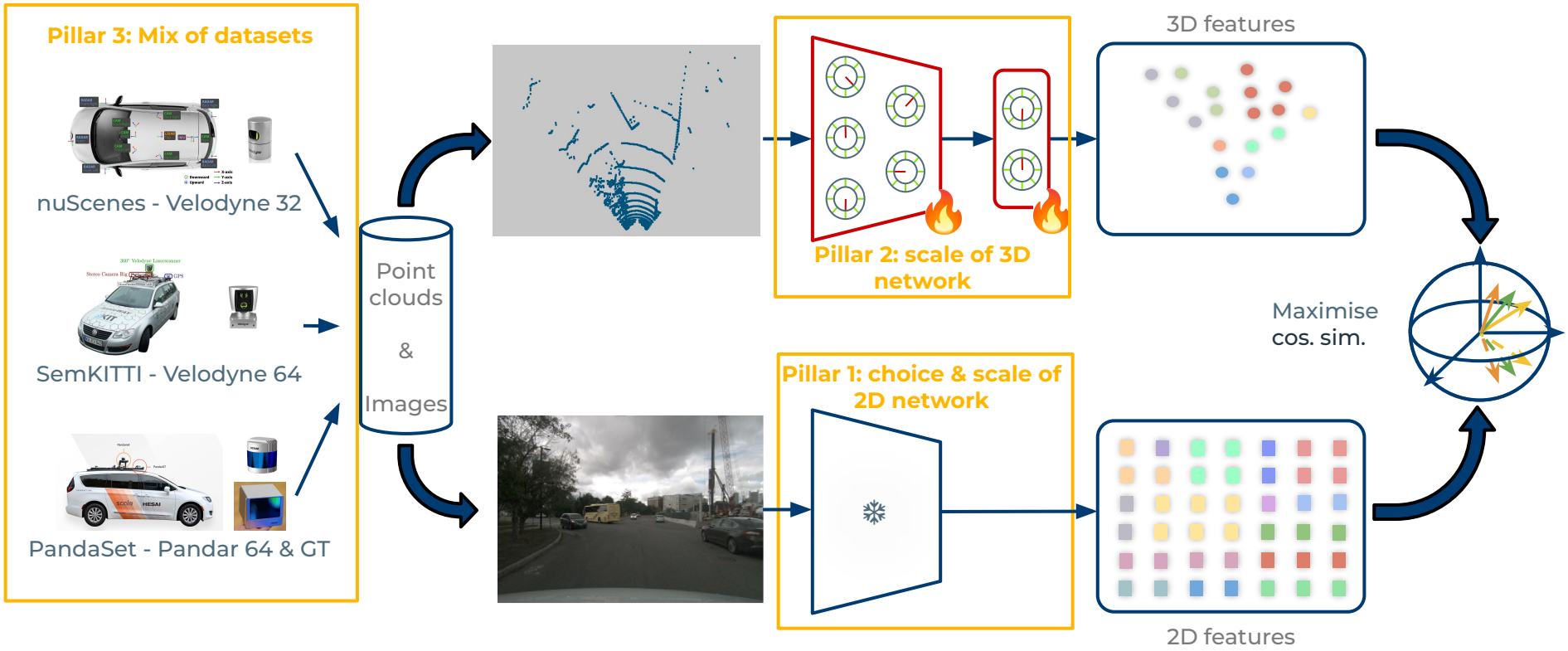
ScaIR


3D features
Point clouds & Images
nuScenes
Maximise cos. sim.
2D features

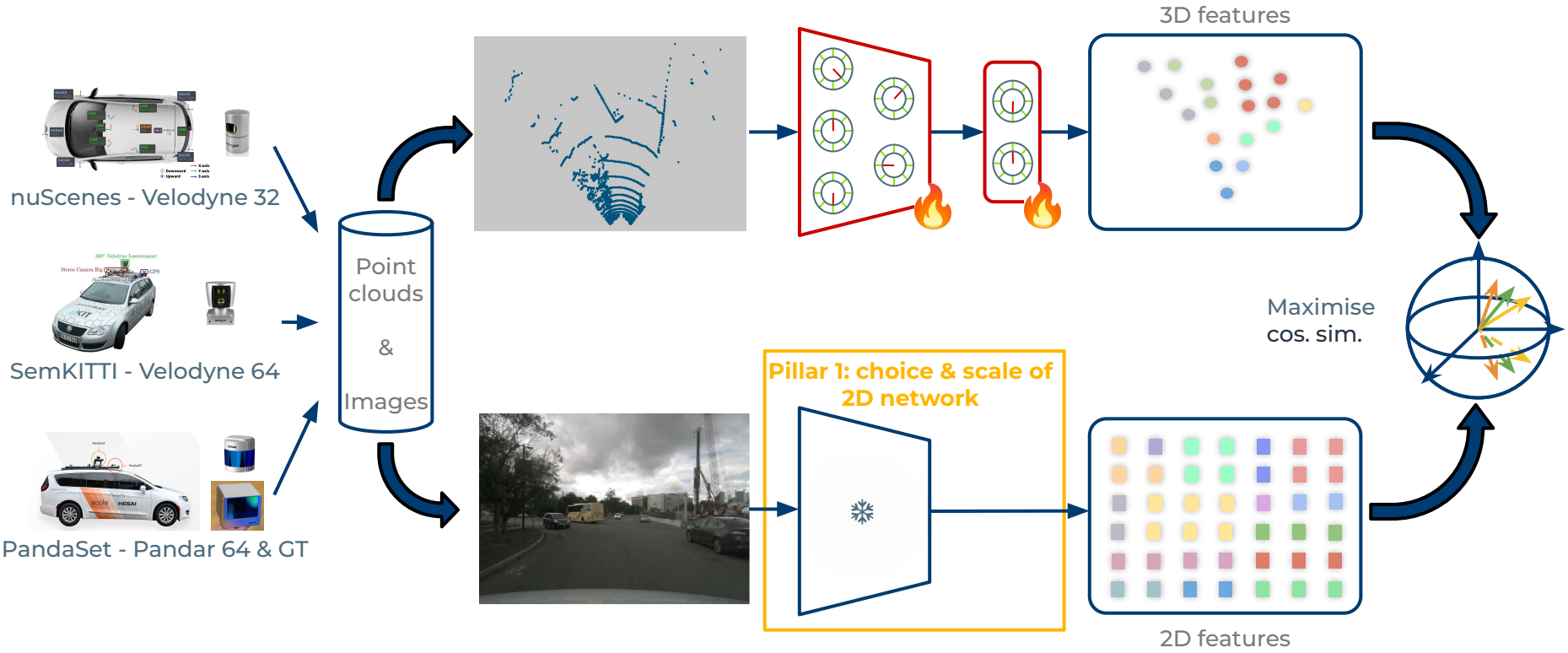
ScalR


nuScenes - Velodyne 32
SemKITTI - Velodyne 64
PandaSet - Pandar 64 & GT
Point clouds & Images
3D features
2D features
Maximise cos. sim.

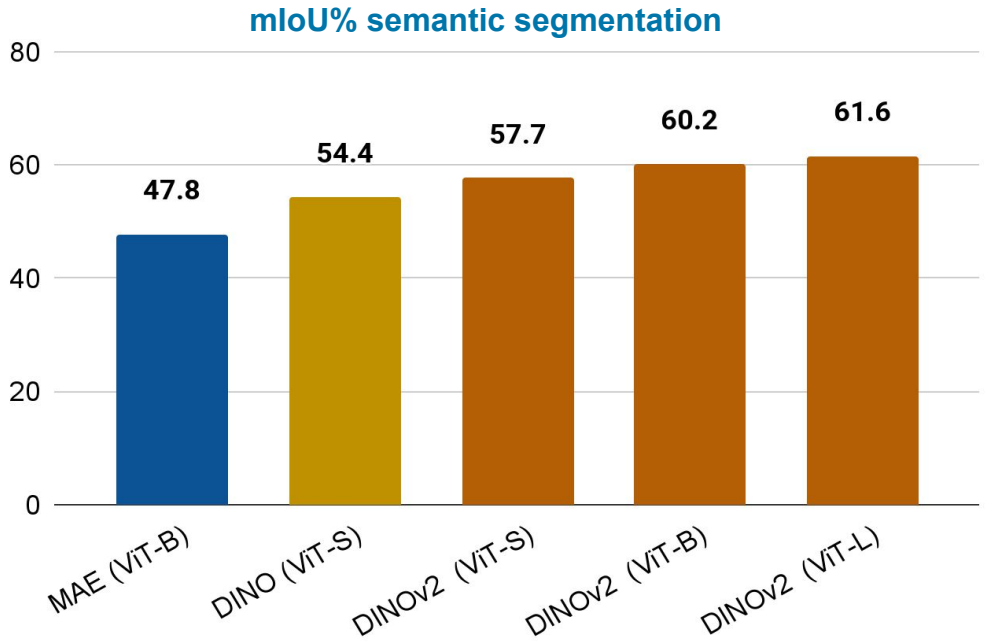
ScalR


Pillar 3: Mix of datasets
nuScenes - Velodyne 32
SemKITTI - Velodyne 64
PandaSet - Pandar 64 & GT
Point clouds & Images
Pillar 2: scale of 3D network
3D features
Pillar 1: choice & scale of 2D network
2D features
Maximise cos. sim.

ScalR

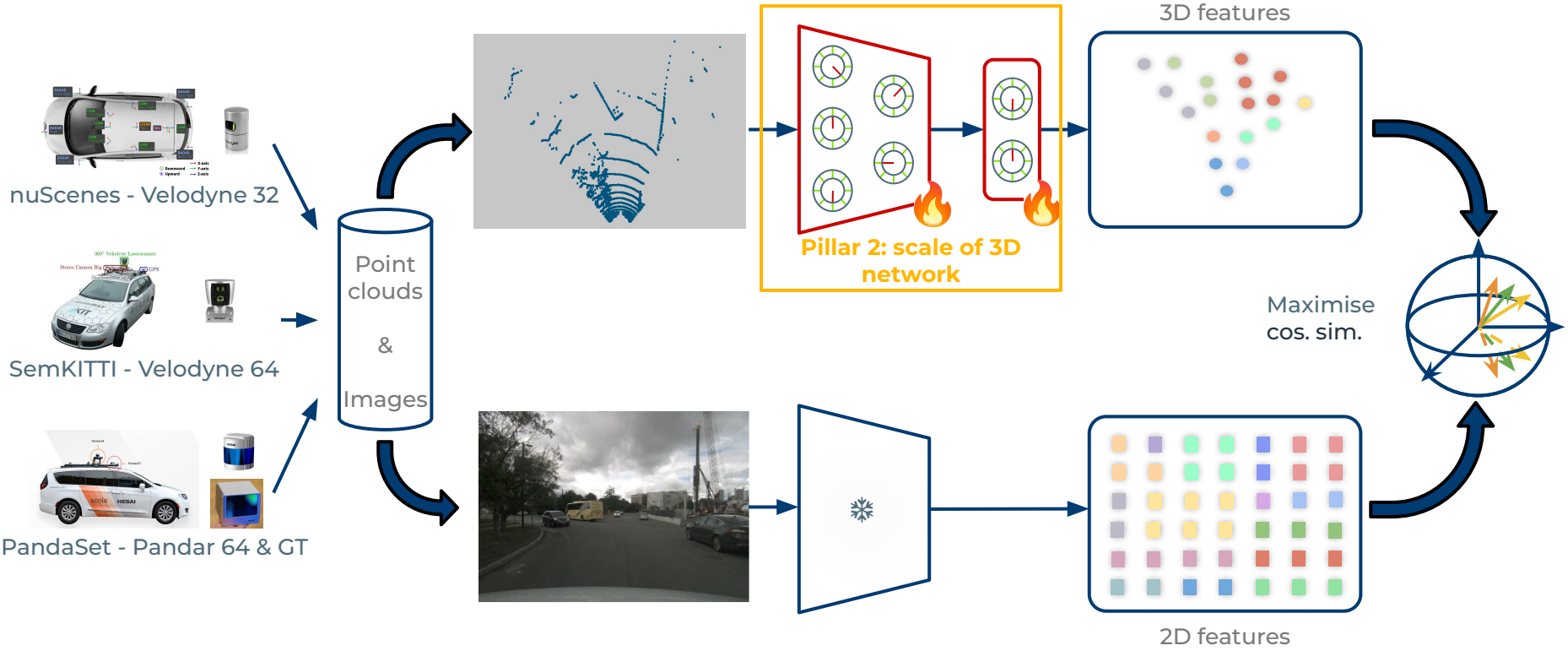

nuScenes - Velodyne 32
SemKITTI - Velodyne 64
PandaSet - Pandar 64 & GT
Point clouds & Images
3D features
Pillar 1: choice & scale of 2D network
2D features
Maximise cos. sim.

ScalR



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

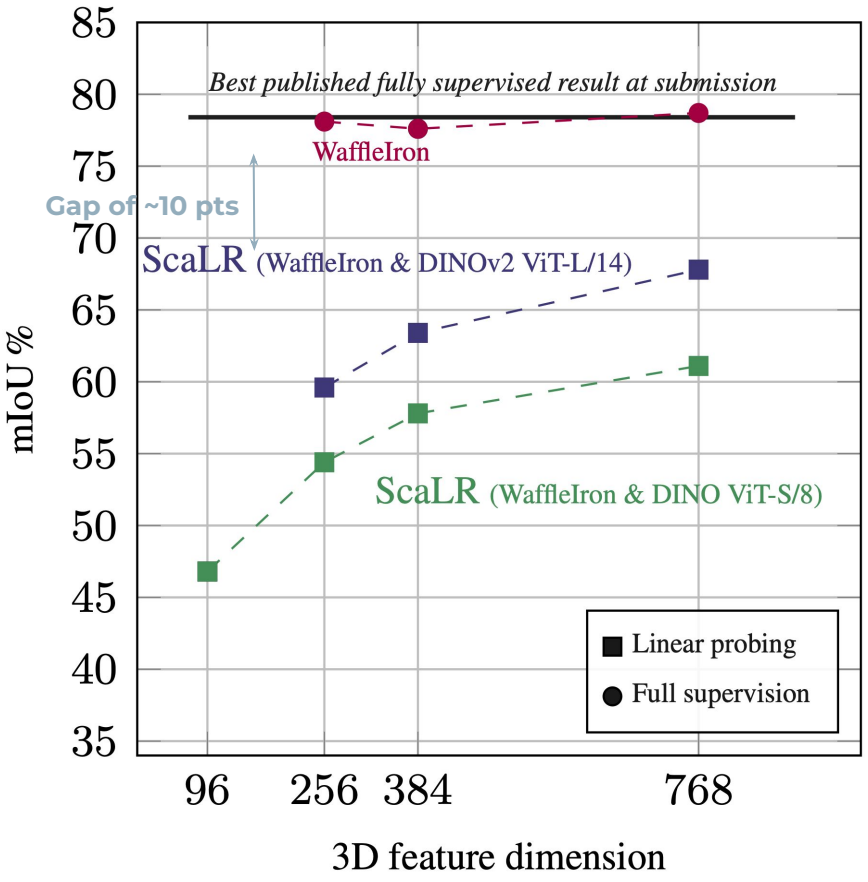
ScalR


nuScenes - Velodyne 32
SemKITTI - Velodyne 64
PandaSet - Pandar 64 & GT
Point clouds & Images
Pillar 2: scale of 3D network
3D features
2D features
Maximise cos. sim.

ScaIR

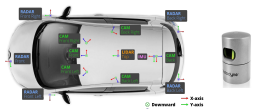
Pretraining & linear probing on nuScenes

Distilled into WaffleIron with varying feature sizes



ScalR

Pillar 3: Mix of datasets



nuScenes - Velodyne 32

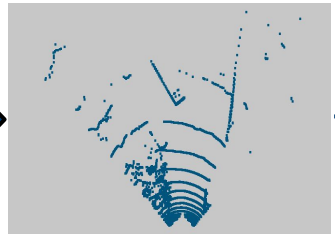


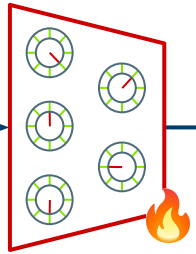
SemKITTI - Velodyne 64

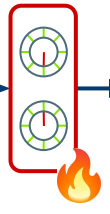


PandaSet - Pandar 64 & GT

Point clouds & Images



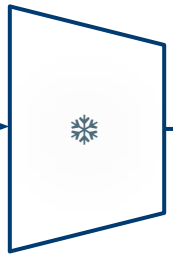


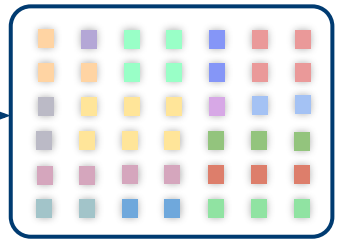


3D features









2D features

Maximise cos. sim.



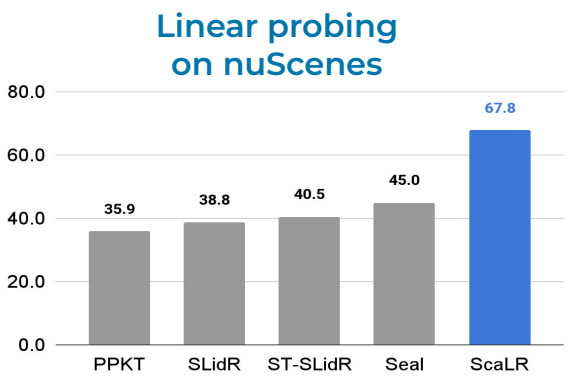
ScalR

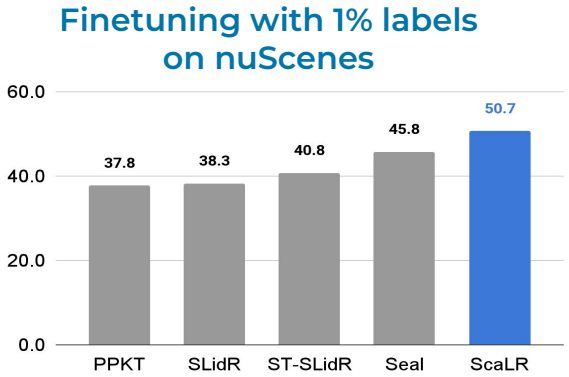
	Pretrain. Dataset	Downstream & Test Dataset			
		nuScenes	SemKITTI	Pand. 64	Pand. GT
WI-256	*Pretraining with DINO-ViT-S/8 and linear probing*				
	nuScenes	54.4	28.8	26.9	25.2
	SemKITTI	39.5	46.6	25.3	25.7
	Pand. 64	39.6	25.6	30.0	24.7
	Pand. GT	29.9	26.9	23.5	28.5
	All datasets	**54.6**	**50.6**	**33.1**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

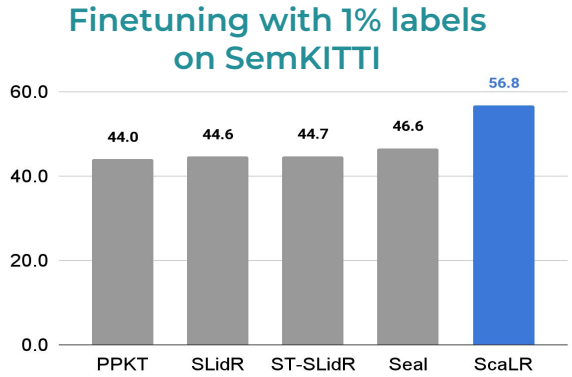
Pretraining on individual datasets or mix of all - Linear probing on each individual dataset

Progress after optimizing the three pillars

ScaIR





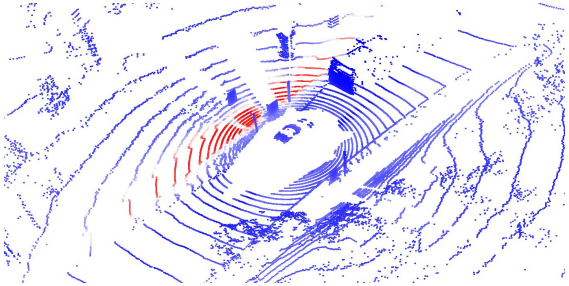


Idea to improve ScaLR further:
Exploit segments as in SLidR, ST-SLidR, Seal and/or time consistency used in Seal.

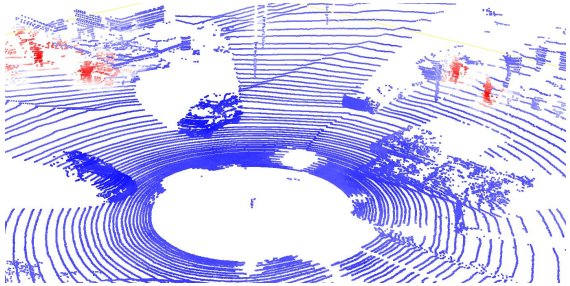
[PPKT] Y.-C. Liu et al. Learning from 2D: Pixel-to-point knowledge transfer for 3D pretraining. arxiv:2104.04687, 2021
[SLidR] C. Sautier et al. Image-to-lidar self-supervised distillation for autonomous driving data. CVPR, 2022.
[ST-SLidR] A. Mahmoud et al., Self-supervised image-to-point distillation via semantically tolerant contrastive loss. CVPR, 2023.
[Seal] Y. Liu et al. Segment Any Point Cloud Sequences by Distilling Vision Foundation Models. NeurIPS, 2023.

ScalR

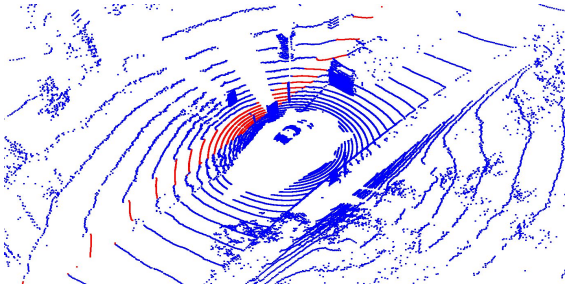
nuScenes - Sidewalk

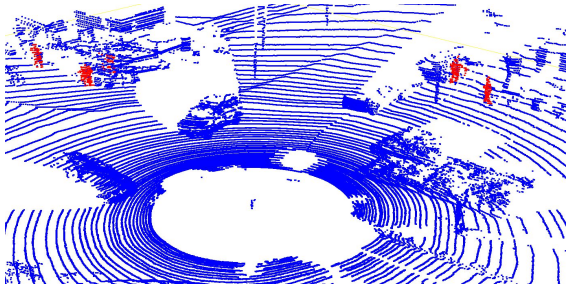


SemKITTI - Pedestrian



Correlation maps with class prototypes



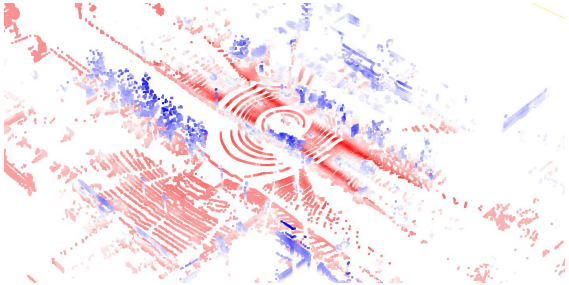


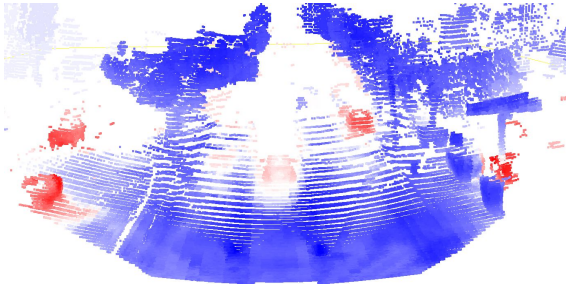
Ground truth

ScalR

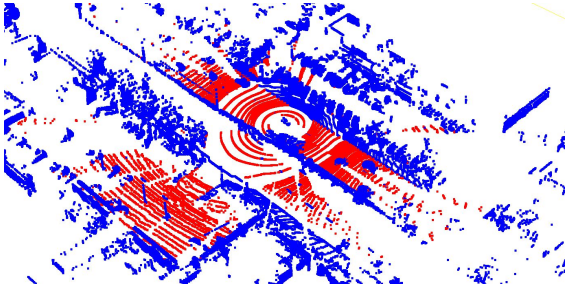
PandaSet 64 - Road

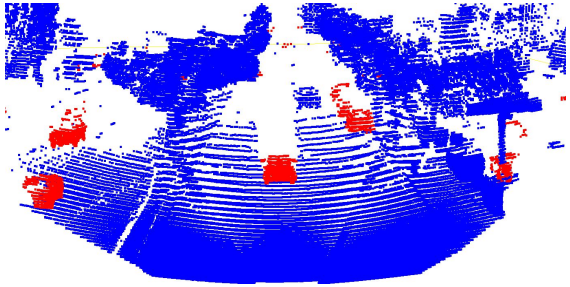
PandaSet GT - Car





Correlation maps with class prototypes





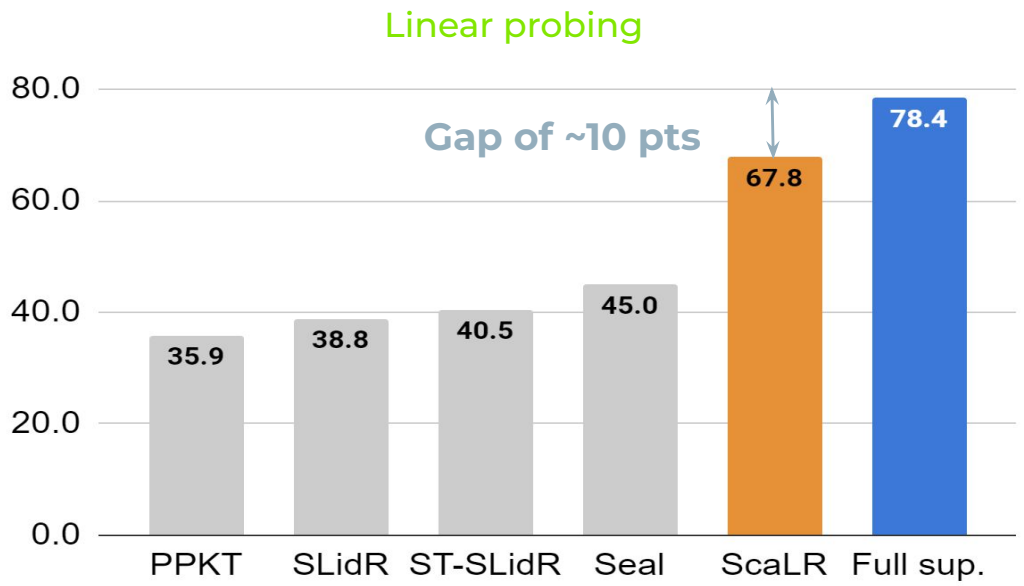
Ground truth

ScaLR

Axes of improvement

Study effect of 3 pillars:

- 2D backbone
- 3D backbone
- Datasets

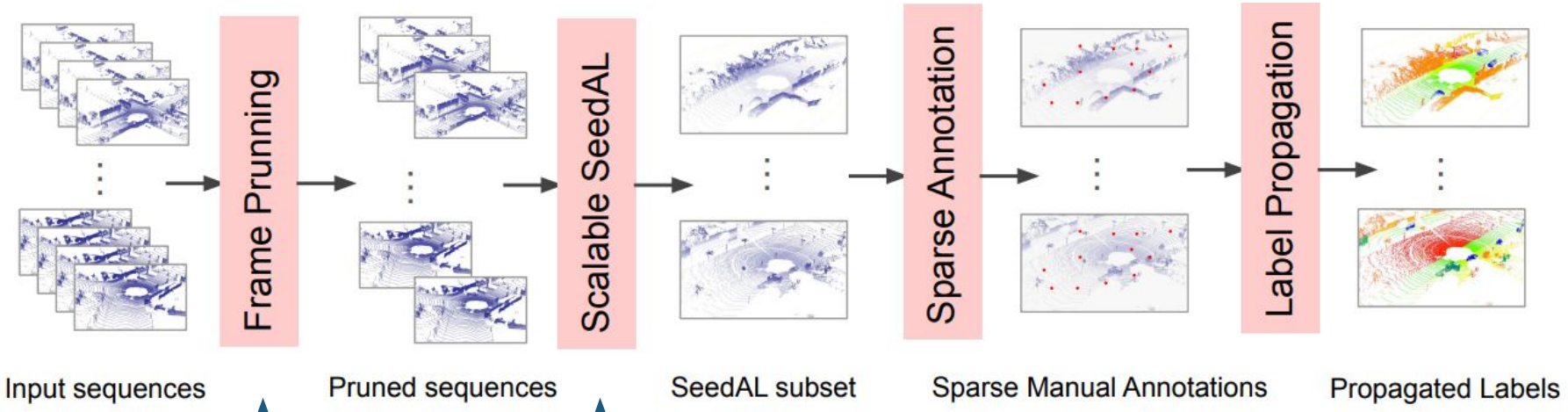


ScaLR

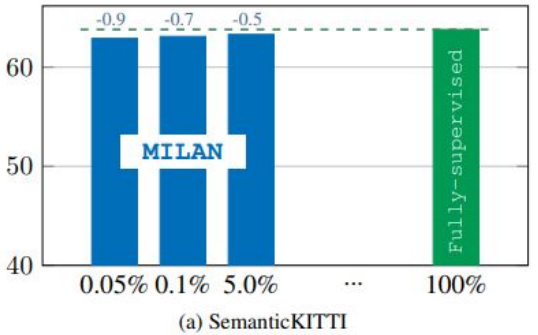
Large improvement of feature quality: +22.8 mIoU pts

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

Helping annotation effort



ScaLR features



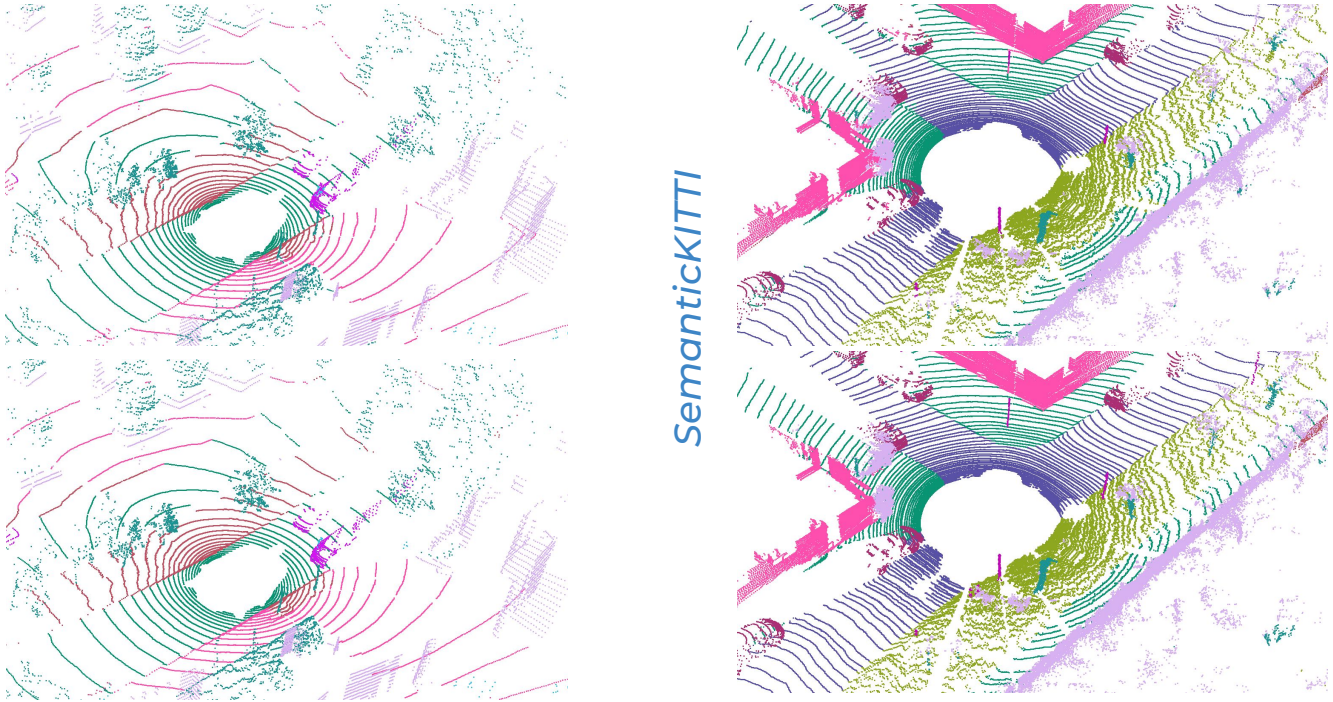
MILAN: Milli-Annotations for Lidar Semantic Segmentation, Samet et al.

Helping annotation effort

Propagated labels after clustering ScaLR features

nuScenes

SemanticKITTI



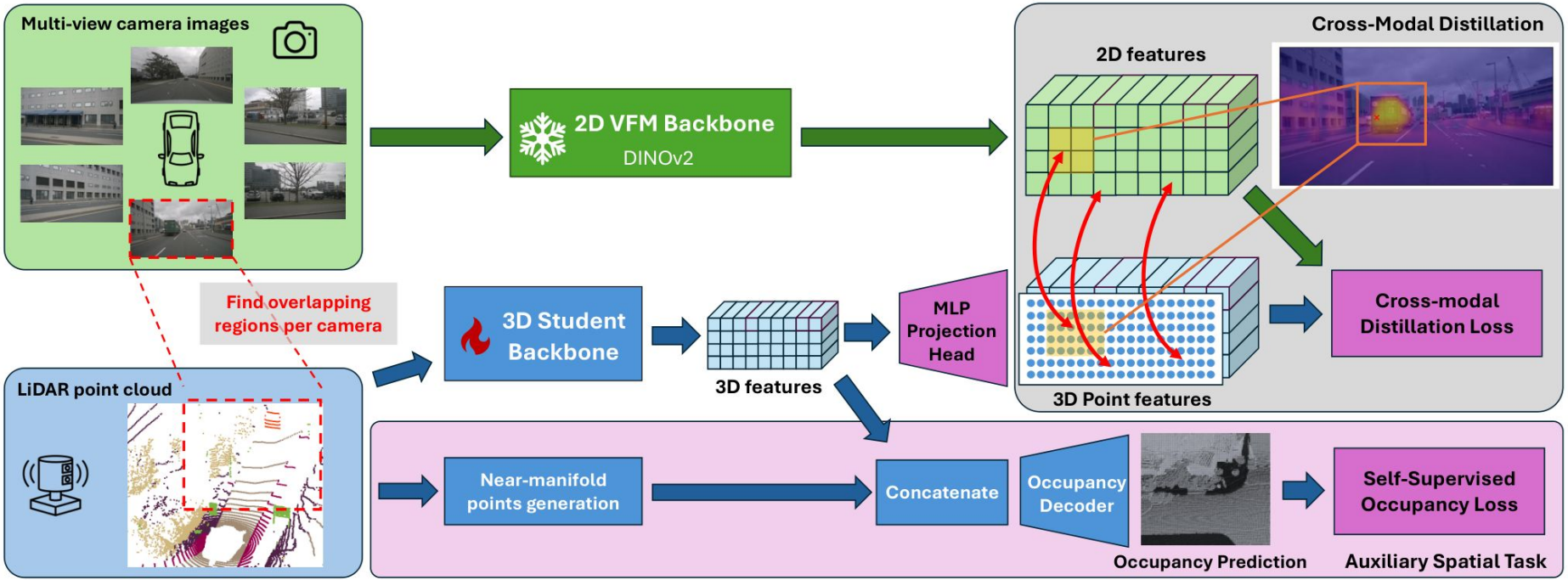
Ground truth

MILAN: Milli-Annotations for Lidar Semantic Segmentation, Samet et al.

Distillation from images

1. A first distillation approach
2. Scaling up the distillation
3. Geometry and distillation

Clever distiller



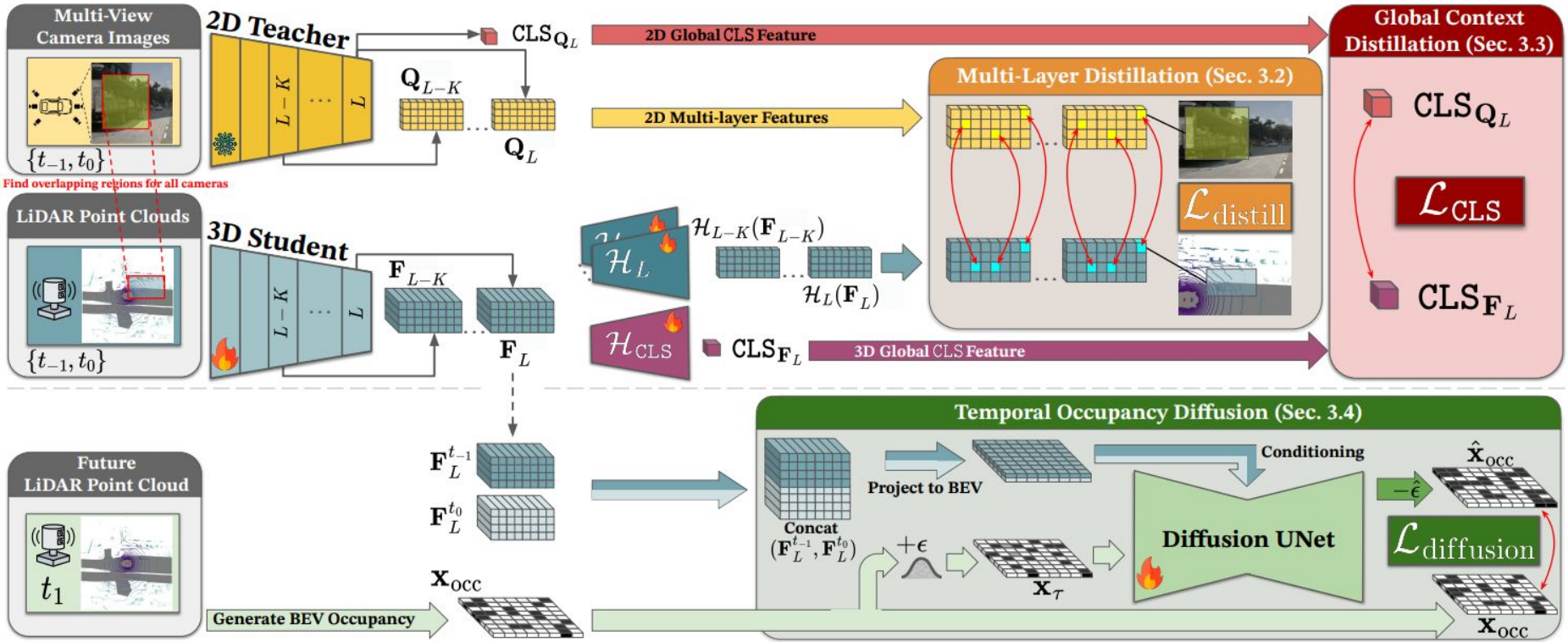
CleverDistiller: Simple and Spatially Consistent Cross-modal Distillation, Govindarajan et al, BMVC 2025

Clever distiller

Method	nuScenes					SKITTI	Waymo
	LP	1%	5%	10%	25%	1%	1%
ViT-S							
ScaLR	44.99	62.07	69.10	72.91	75.69	56.92	60.09
Ours	**58.49**	**62.54**	**71.09**	**73.89**	**76.64**	**61.95**	**62.19**
ViT-B							
ScaLR	41.62	63.54	71.74	73.84	76.56	60.74	57.17
Ours	**60.54**	**63.93**	**72.27**	**74.11**	**76.93**	**63.31**	**62.20**
ViT-L							
ScaLR	45.54	66.17	72.34	74.96	77.56	58.48	61.47
Ours	**61.92**	**66.18**	**73.13**	**75.18**	**77.57**	**62.23**	**65.94**

Table 5: Comparisons of SOTA methods pretrained on nuScenes and fine-tuned on datasets subset, using PTV3 [63] for 3D semantic segmentation. All scores in mIoU.

CleverDistiller: Simple and Spatially Consistent Cross-modal Distillation, Govindarajan et al, BMVC 2025



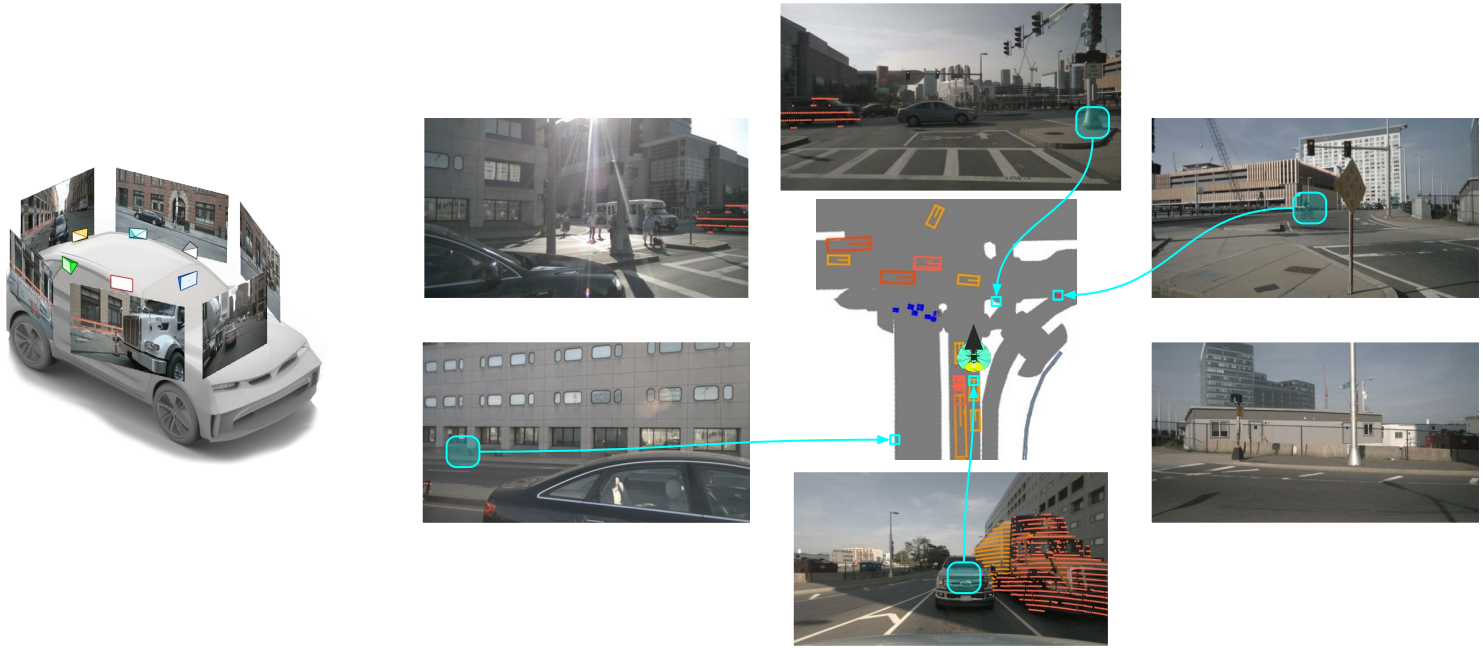
HilDA: Hierarchical Distillation with Diffusion for Advancing Self-Supervised LiDAR Pre-training, Wozniak et al, 2026

Method	S.P.	nuScenes						SKITTI	Waymo
		LP	1%	5%	10%	25%	Full	1%	1%
SLidR [68]	✗	45.35	41.64	55.83	62.68	67.61	74.98	45.50	48.32
Seal [49]	✓	46.59	45.98	57.15	62.79	68.18	75.41	47.24	48.91
SuperFlow [88]	✓	47.66	48.09	59.66	64.52	69.79	76.57	48.40	50.20
ScaLR [63]	✗	41.80	55.83	63.46	65.24	68.70	74.76	45.59	49.60
LiMoE [87]	✓	49.07	50.23	61.51	66.17	71.56	77.81	50.30	51.77
CleverDistiller [24]	✗	51.89	59.80	66.44	67.65	69.53	78.49	51.48	53.56
SuperFlow++ [89]	✓	48.86	49.56	60.75	65.46	70.19	77.29	49.90	51.65
LiMA [90]	✗	56.65	51.29	61.11	65.62	70.43	76.91	50.44	51.35
HilDA	✗	**58.95**	**62.71**	**70.19**	**71.00**	**73.68**	**79.12**	**53.44**	**53.89**

SP: Semantic priors

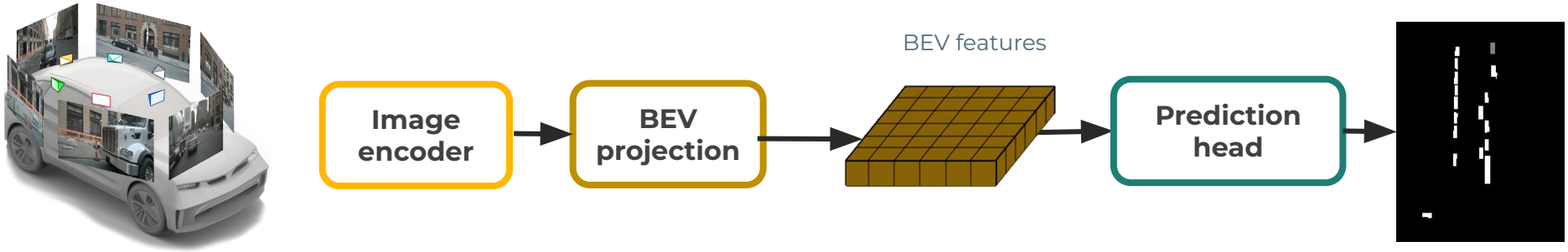
HilDA: Hierarchical Distillation with Diffusion for Advancing Self-Supervised LiDAR Pre-training, Wozniak et al, 2026

OccFeat – BEV Pretraining



- AV need an accurate representation of the surrounding environment
- The top-view is a natural space to fuse sensor information (multi-camera and/or LiDAR, Radar etc.)

OccFeat – BEV Pretraining

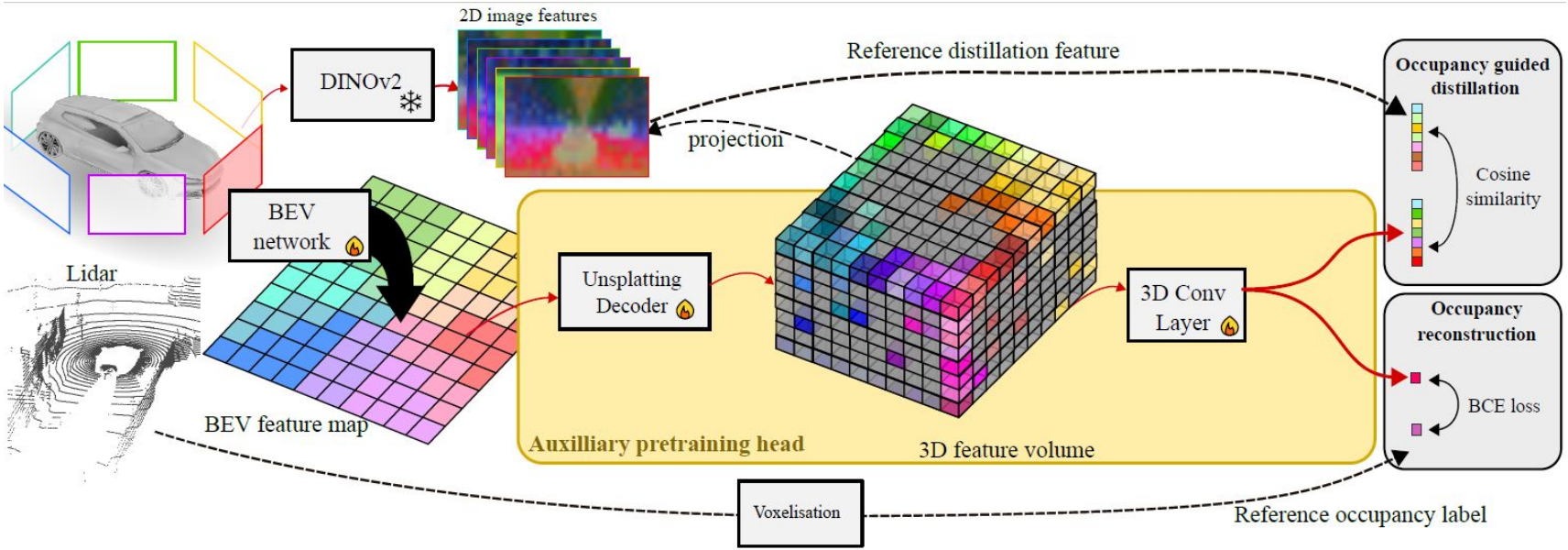


- AV need an **accurate representation** of the surrounding environment
- The **top-view** is a natural space to fuse sensor information (multi-camera and/or LiDAR, Radar etc.)

OccFeat

Architecture-agnostic pretraining that teaches the model semantic-aware 3D geometry information
Leveraging :

- aligned **Image & LiDAR data**
- SSL **pretrained image encoder**

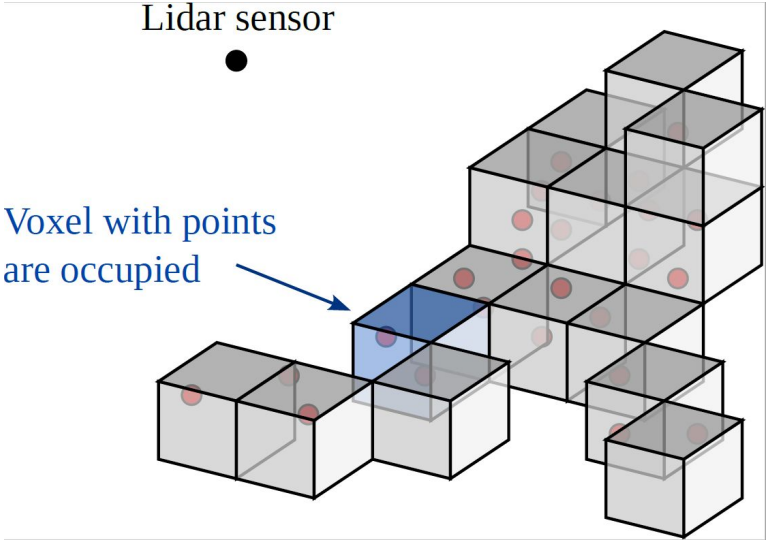


OccFeat: Self-supervised Occupancy Feature Prediction for Pretraining BEV Segmentation Networks, Sirko-Galouchenko et al, CVPRW 2024

Occupancy estimation task

Predict occupancy volume.

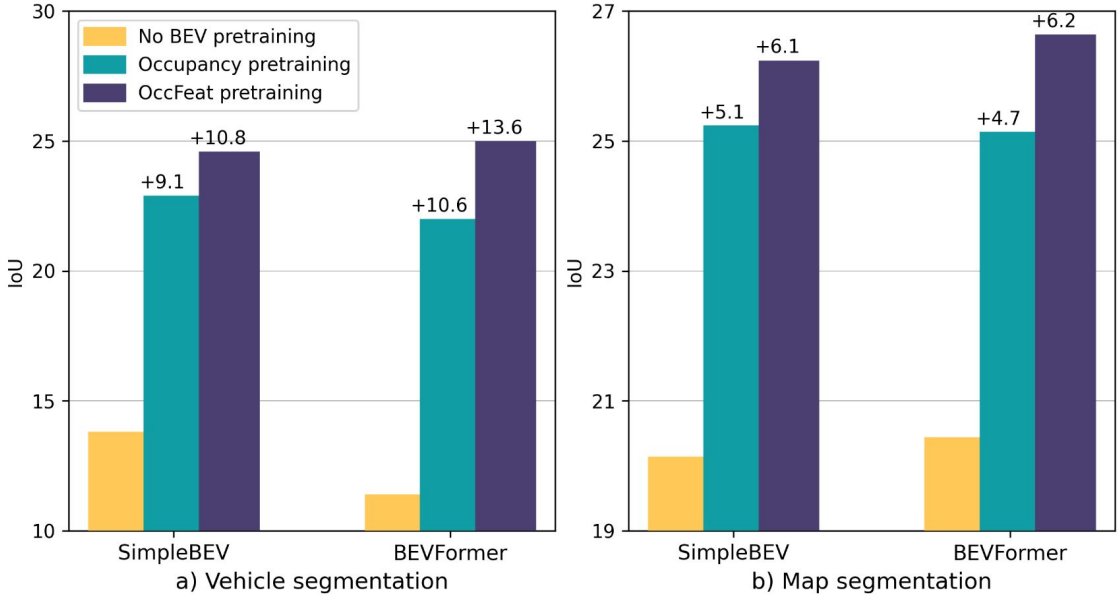
Supervision from LiDAR sensor:
A voxel is occupied if at least one Lidar point falls inside



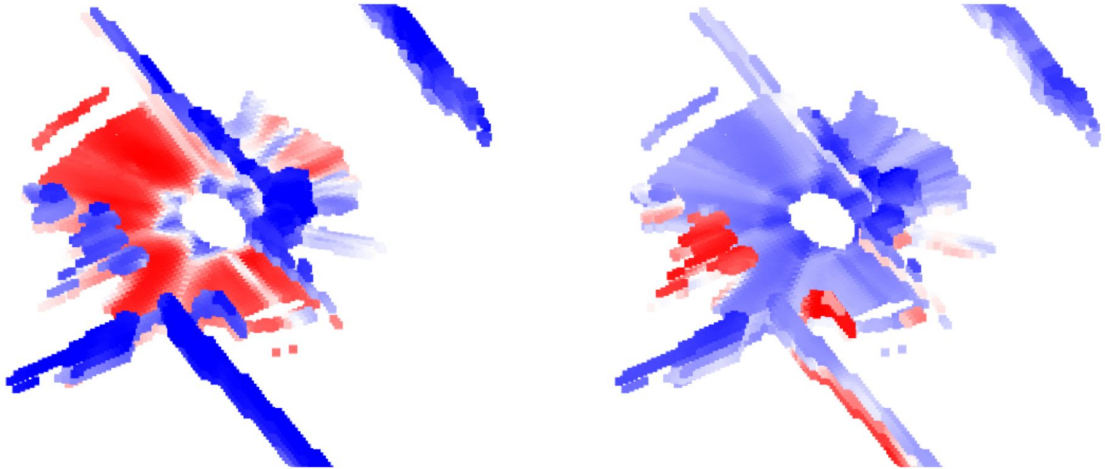
OccFeat quantitative results

Performance comparison in **low data regime** (1% annotated data of nuScenes)
BEV semantic segmentation of vehicles and map layout

No BEV pretraining - Imagenet pretraining of the image encoder



OccFeat - Qualitative results



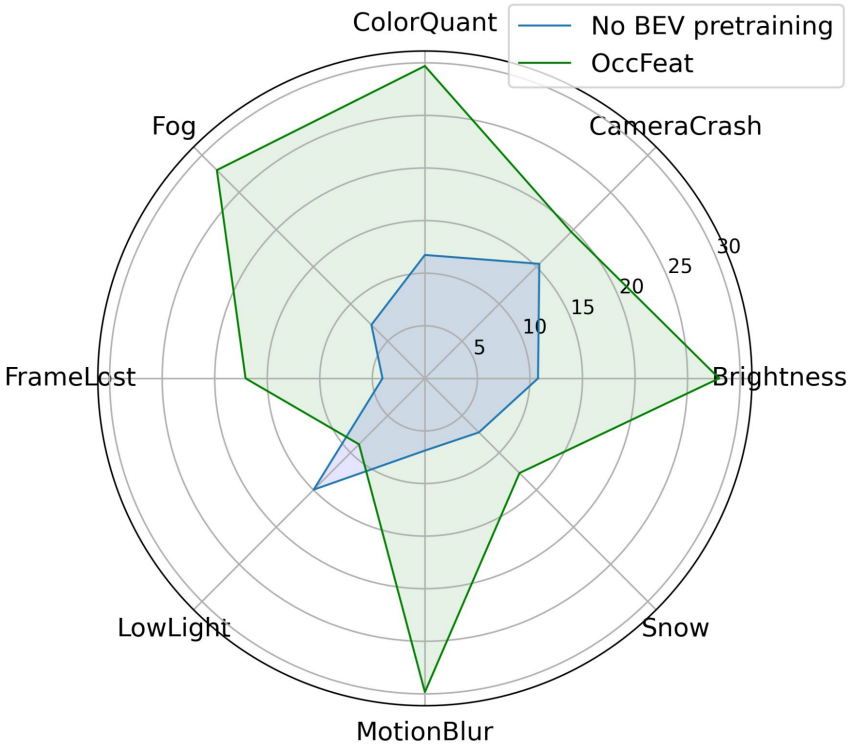
(a) Road (b) Car

Correlation maps

Robustness study 💪

Segmentation results on nuScenes-C dataset (8 corruptions on nuScenes validation set)

OccFeat vs no BEV pretraining

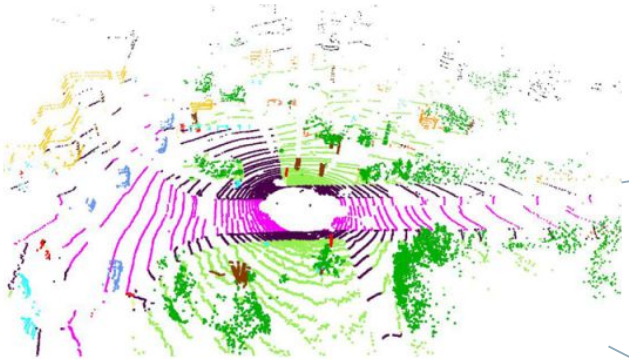


OccFeat: Self-supervised Occupancy Feature Prediction for Pretraining BEV Segmentation Networks, Sirko-Galouchenko et al, CVPRW 2024

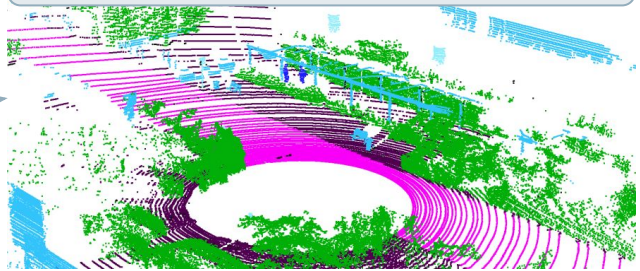
Can self-supervision help with domain gaps?

1. Domain adaptation
2. Surface-based regularization
3. Distillation

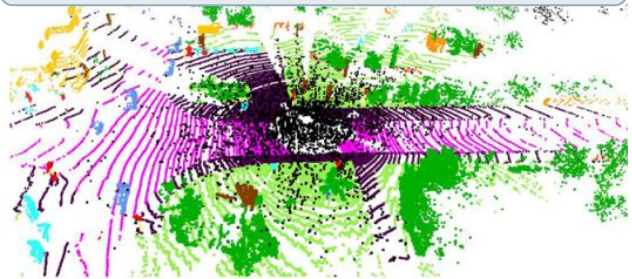
Domain adaptation



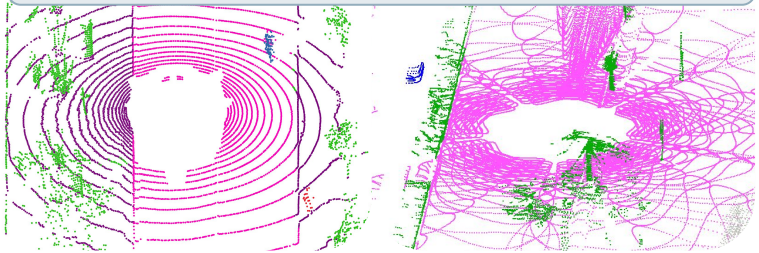
Location



Snowfall



Sensor



Benchmarking the Robustness of LiDAR Semantic Segmentation Models
Online Segmentation of LiDAR Sequences: Dataset and Algorithm

Domain adaptation

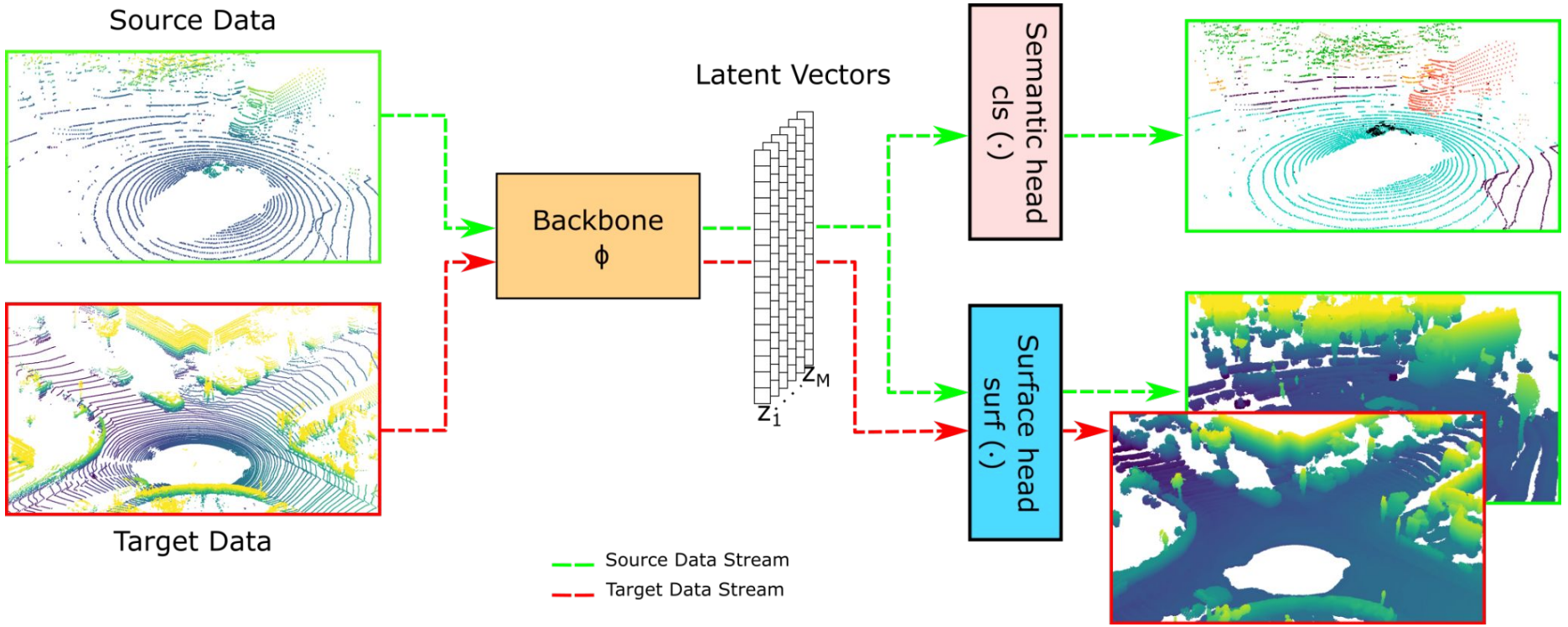
Setting	Source	Target	Target Adaptation
Domain Adaptation	X, Y	X, Y	Training
Unsupervised Domain Adaptation (UDA)	X, Y	X	Training
Multimodal UDA	X, Y, I	X, I	Training
Source-Free UDA (SFUDA)	$f[\theta^s]$	X	Training
Test-Time Adaptation	$f[\theta^s]$	X	Inference
Domain Generalization	$X_{1,\ldots,i}, Y_{1,\ldots,i}$		No

Notation: point clouds X, labels Y, images I and source-trained model $f[\theta^s]$.

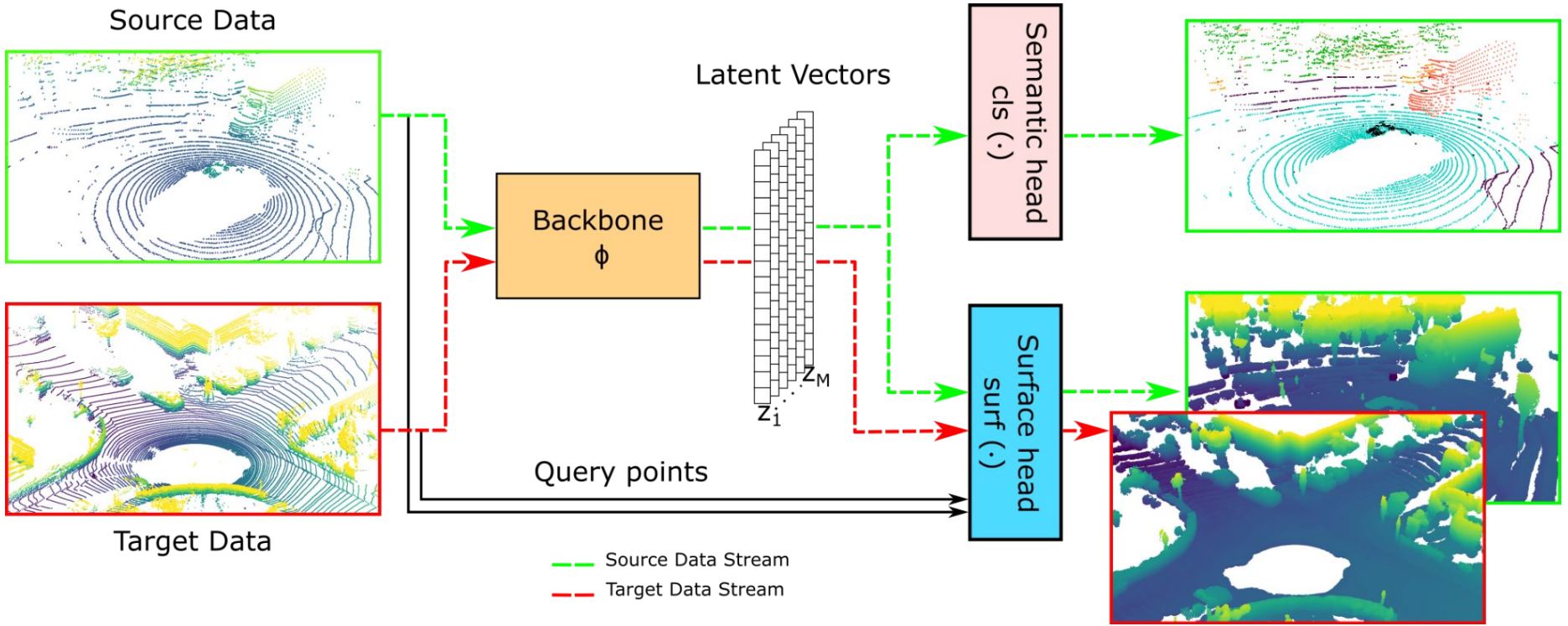
Can self-supervision help with domain gaps?

1. Domain adaptation
2. Surface-based regularization
3. Distillation

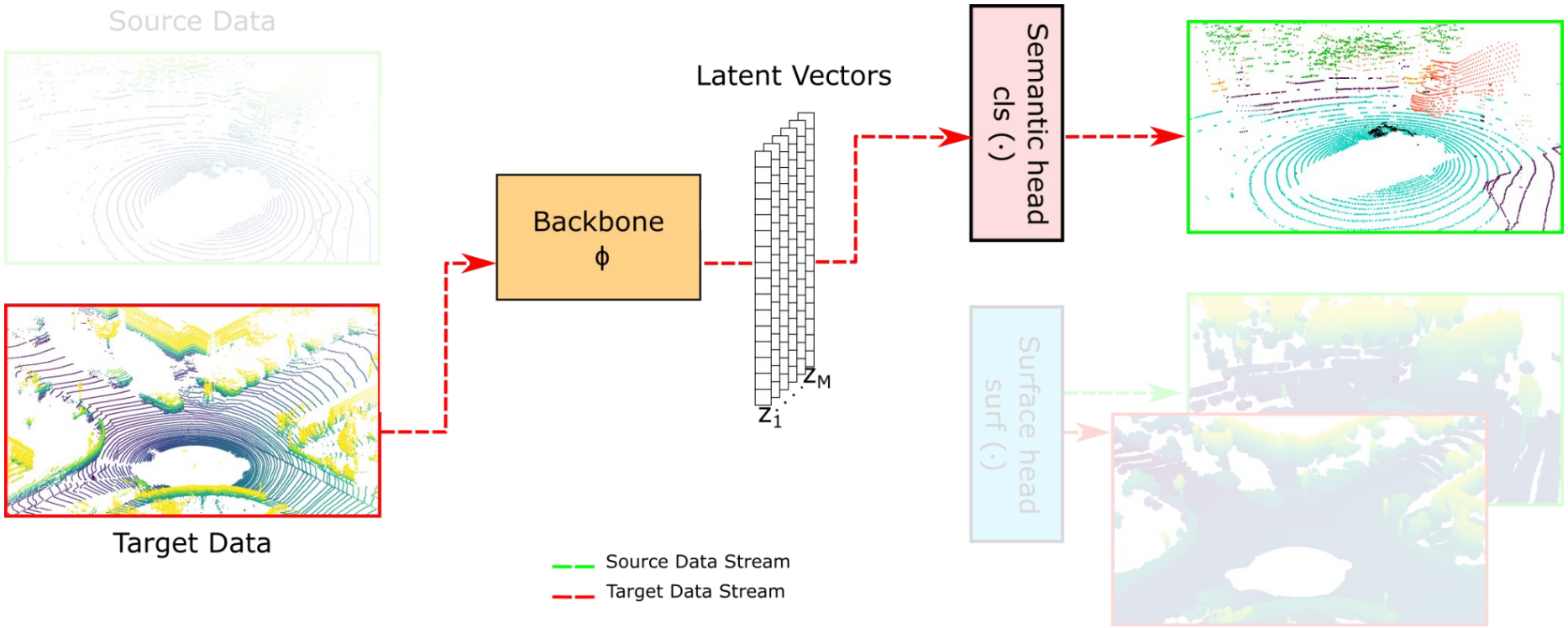
SALUDA


Source Data
Target Data
Backbone
ϕ
Latent Vectors
z_1
z_M
Semantic head
cls (·)
Surface head
surf (·)
Source Data Stream
Target Data Stream

SALUDA

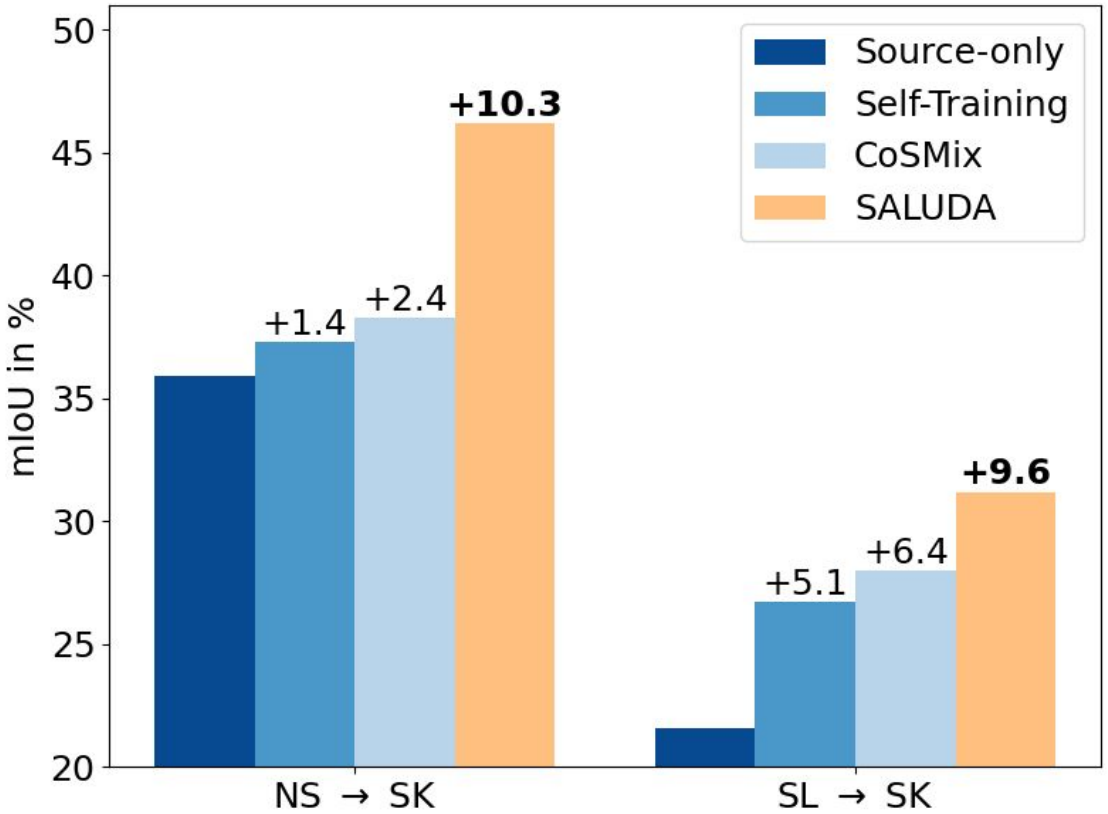

Source Data
Target Data
Backbone
ϕ
Latent Vectors
z_1
z_M
Semantic head
cls (·)
Surface head
surf (·)
Query points
Source Data Stream
Target Data Stream

SALUDA


Source Data
Target Data
Backbone
ϕ
Latent Vectors
z_1
z_M
Semantic head
cls (·)
Surface head
surf (·)
Source Data Stream
Target Data Stream

SALUDA

Real-to-Real, different sensor: nuScenes to SemanticKITTI (32 Beams to 64 Beams): NS→SK



Synthetic-to-Real, same sensor: SynLiDAR to SemanticKITTI (64 Beams to 64 Beams): SL→SK

Validation for UDA

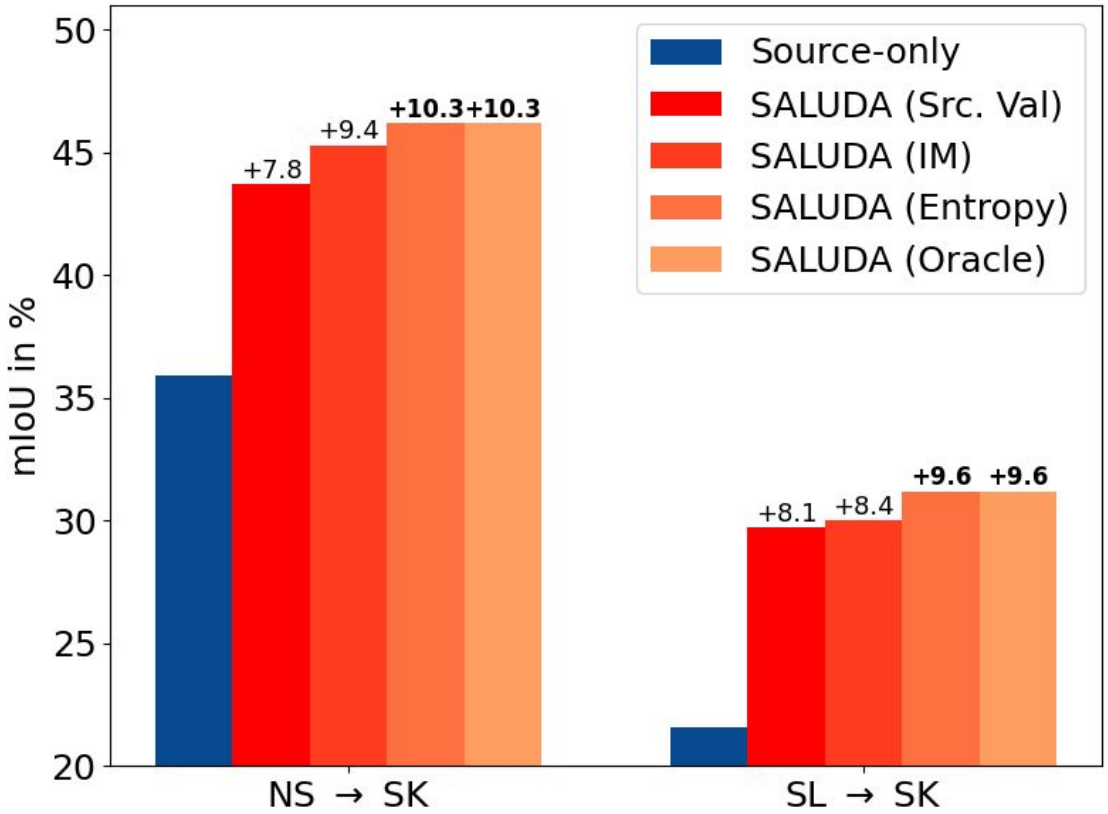
❌ No target labels available

How to find hyperparameters ?

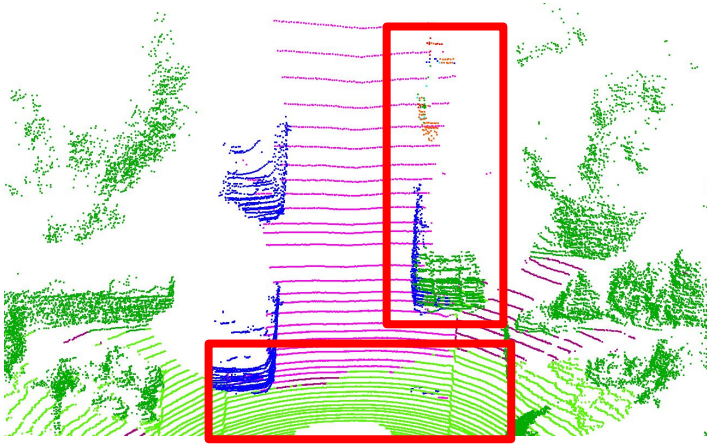
- Use a target validation set
- Keep hyperparameters the same across different settings
- Look on proxies and heuristics (=validators)[1)]
  - Entropy
  - Performance on source data

Musgrave et al., *Three New Validators and a Large-Scale Benchmark Ranking for Unsupervised Domain Adaptation*, Arxiv 2022*

Hyperparameter selection with validators

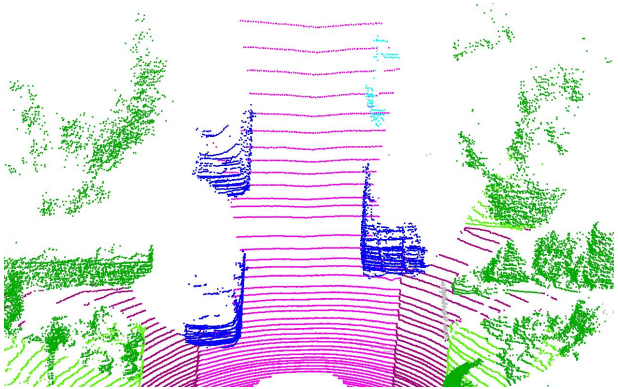


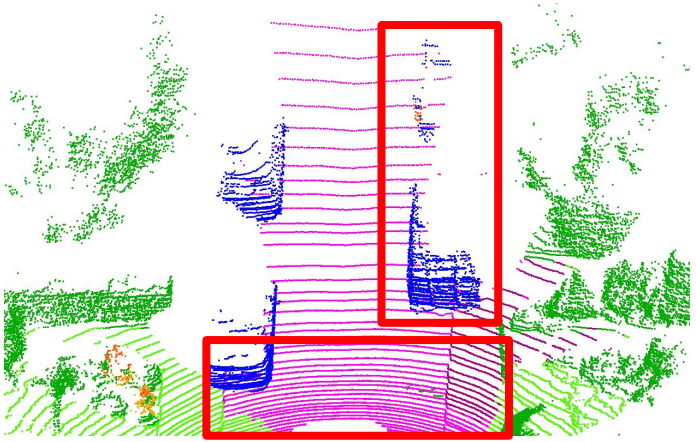
Qualitative results



Source Only

GT





SALUDA

SALUDA

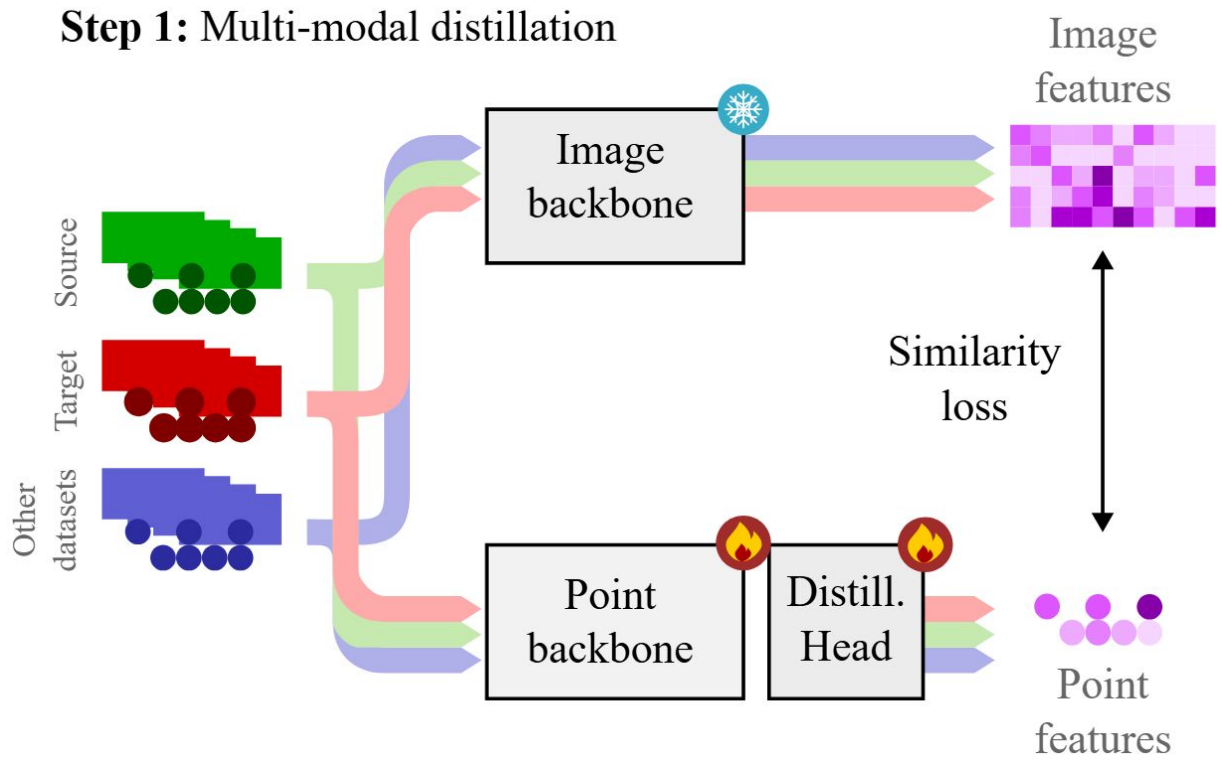
- **Competitive results** with geometric regularization.
- **Robust** in unsupervised hyperparameter selection.
- Can be **combined with other SOTA** methods.
- **Published code.**

SALUDA: Surface-based Automotive Lidar Unsupervised Domain Adaptation, Michele et al, 3DV 2024

Can self-supervision help with domain gaps?

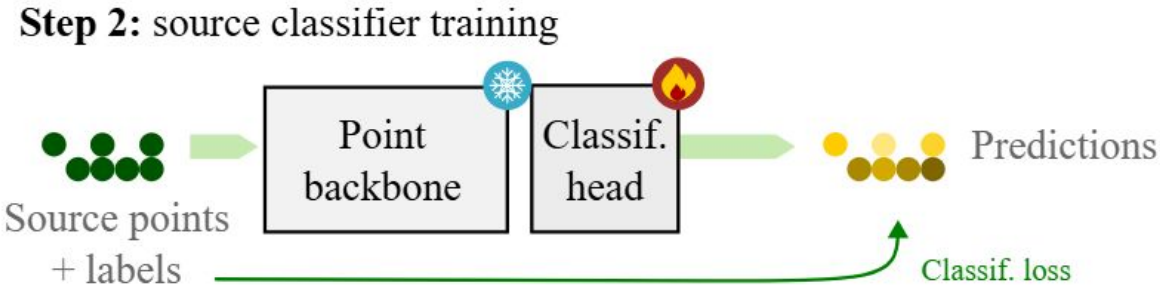
1. Domain adaptation
2. Surface-based regularization
3. Distillation

Joint VFM distillation



Improving Multimodal Distillation for 3D Semantic Segmentation under Domain Shift, Michele et al. BMVC 2025

Joint VFM distillation



How to get the best out of VFMs for 3D UDA

We study:

✔️ Best practices of unimodal UDA

✔️ Choice of a visual foundation model for pretraining

✔️ Choice of pretraining datasets

✔️ Downstream training

On 4 settings:

Real-to-Real, different sensor:
nuScenes to SemanticKITTI (32 to 64 Beams): NS→SK
nuScenes to Waymo (32 to 64 Beams): NS→WO

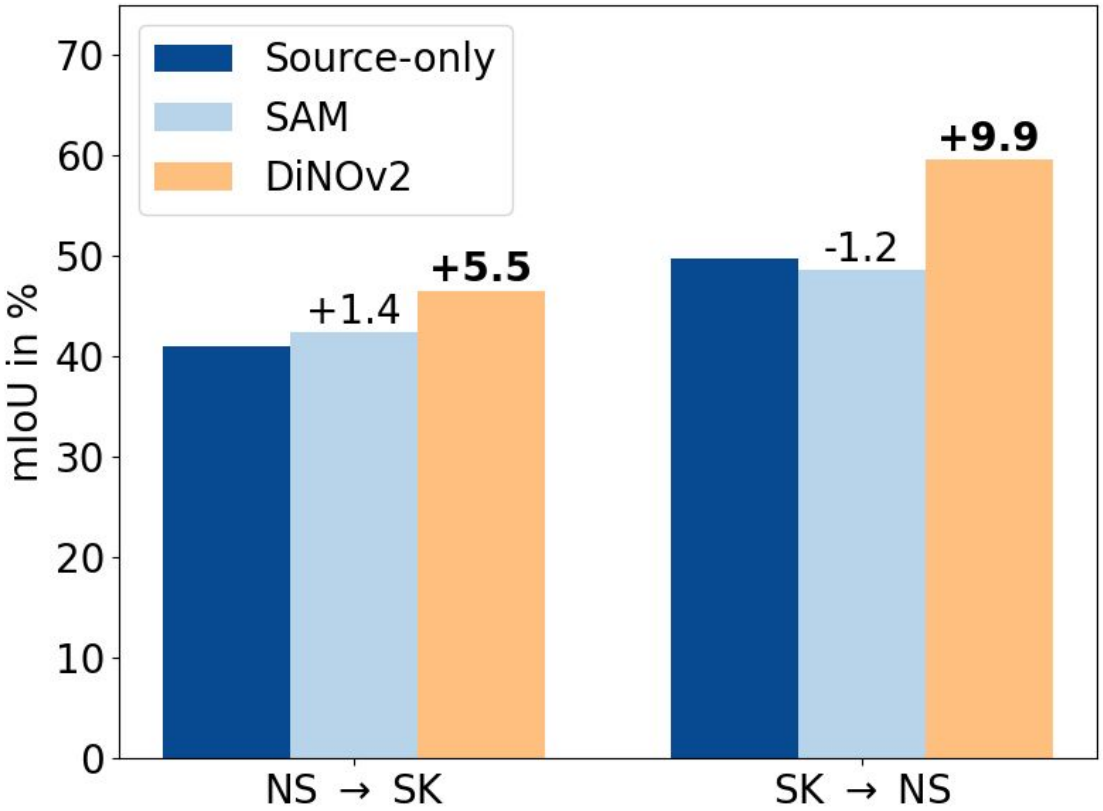
SemanticKITTI to nuScenes (64 to 32 Beams): SK→NS
Waymo to nuScenes (64 to 32 Beams): WO→NS

Best practices of Unimodal UDA

✔️ No intensity
✔️ LayerNorm (LN) instead of BatchNorm (BN)
✔️ WaffleIron(WI) instead of Minkowski-U-Net (MUNet)
✔️ Bigger than smaller

Method	Backbone	Int.	Norm	NS→SK	SK→NS	NS→WO	WO→NS
Source-only	WI-768	✗	LN	44.6	**55.1**	37.1	**64.6**
State of the art in lidar multi-modal UDA.							
Adapt-SAM	MUNet	✓	BN	**48.5**	42.9	**44.9**	48.2

Choice of visual foundation model



Choice of the pretraining dataset

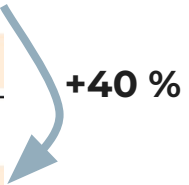
Back-bone	Pretrain S	Pretrain T	Pretrain E	Frozen back-bone	Avg. mIoU in %
WI-256	✗	✗	✗	✗	47.0
	✓	✗	✗	✓	33.3
	✓	✓	✗	✓	56.8
WI-768	✗	✗	✗	✗	50.4
	✓	✓	✗	✓	61.6
	✓	✓	✓	✓	63.6

Downstream training

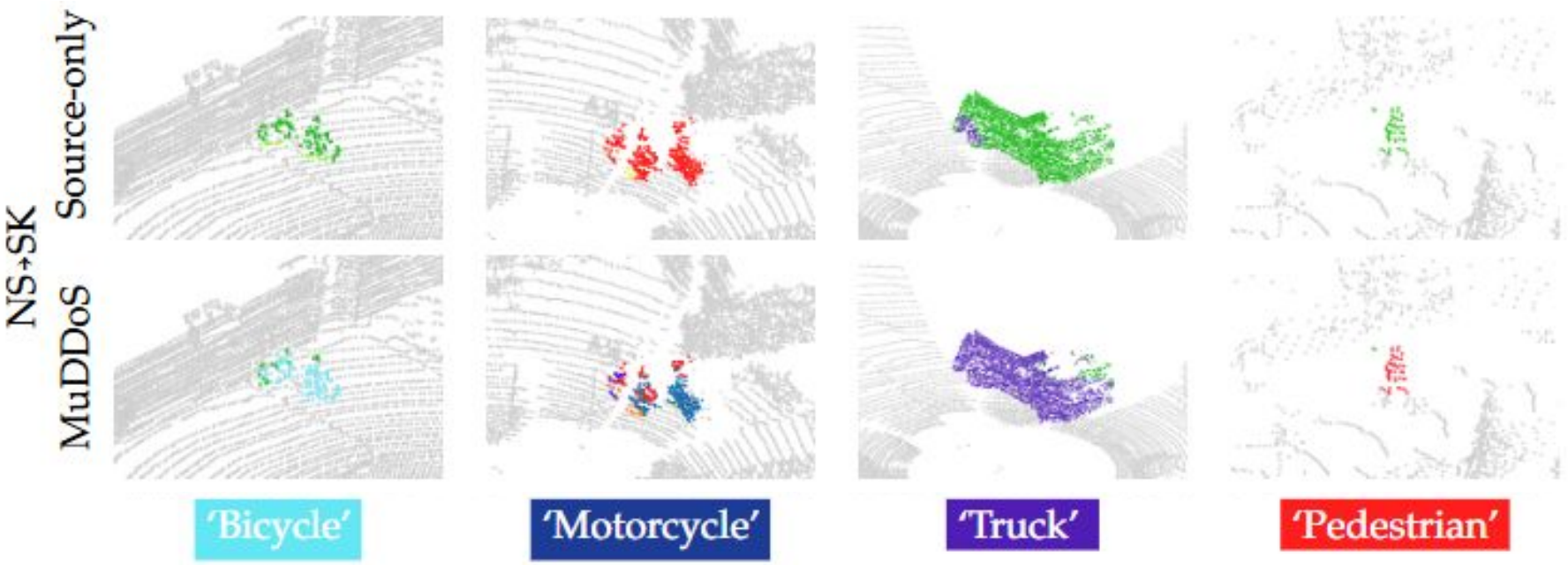
Downstream training	NS→SK	SK→NS
Frozen Back. + Linear	43.4	<u>60.1</u>
Frozen Back. + MLP	<u>47.5</u>	**62.1**
Full finetuning	**50.5**	58.9
Joint distill. & classif.	44.9	56.0

Comparison to the SOTA

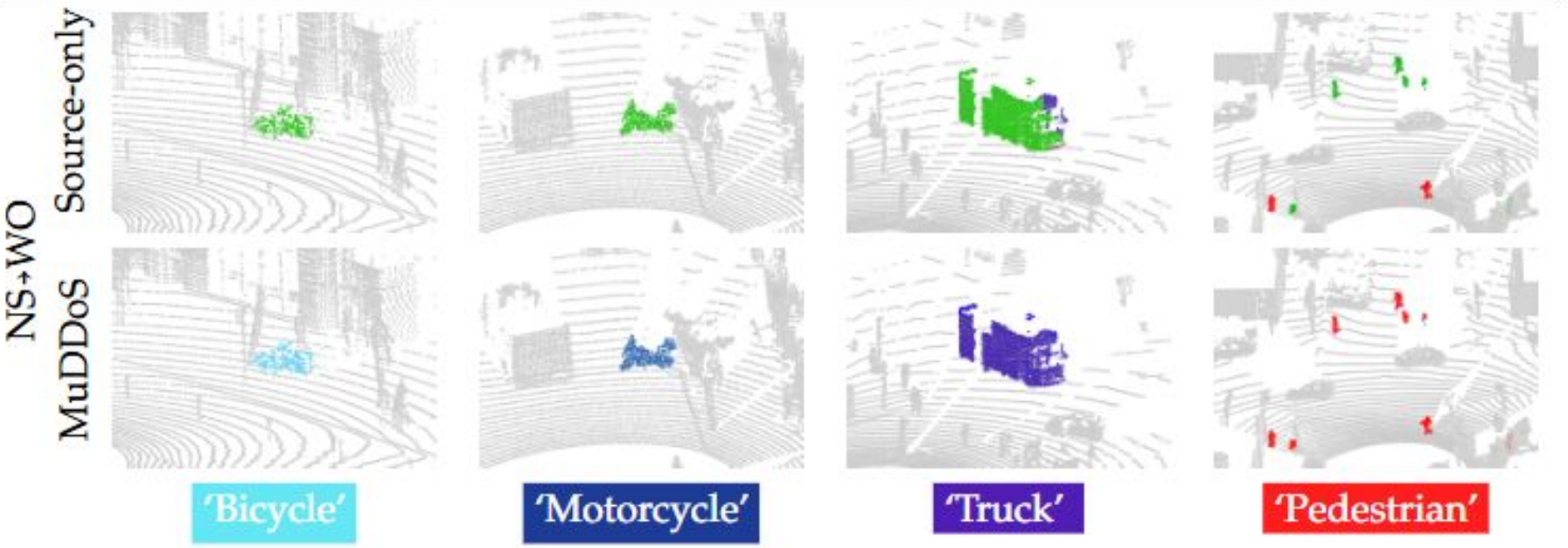
Method	Back.	Mod.	Avg. mIoU in %
Source-only	MUNet	-	26.8
Source-only	WI-768	-	50.4
PL	MUNet	U	28.3
CosMix	MUNet	U	30.5
MM2D3D	MUNet	M	34.6
Adapt-SAM	MUNet	M	46.1
MuDDoS (ours)	MUNet	M	50.6
Self-Training	WI-768	U	44.2
MuDDoS (ours)	WI-768	M	**64.5**



Where do we improve



Where do we improve



Summary MuDDos and Domain adaptation

☑ **Best unimodal UDA practices** are a key part.

☑ **Joint distillation** with one model for many domain shifts.

☑ **Careful downstream training** to avoid degrading the quality of the pretrained features.

Improving Multimodal Distillation for 3D Semantic Segmentation under Domain Shift, Michele et al. BMVC 2025

Conclusion

Takeaways: self-supervision and domain adaptation



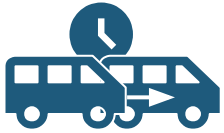
Use sensor information



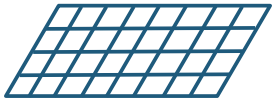
When image is available: distillation



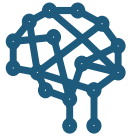
Scale data (multi dataset)



Time helps self-supervision



BEV is an important space



Architecture matters for scaling

Valeo

Driving Change Together



valeo.com

valeo.ai