IEEE IV 2023
IEEE INTELLIGENT VEHICLES SYMPOSIUM
Anchorage, Alaska, USA // June 4 – 7, 2023

3D-DLAD Workshop

Towards Robust Perception in Autonomous Driving

Speaker: Lingdong Kong

June 04, 2023

Short Bio

Ph.D. Student, School of Computing, National University of Singapore (https://ldkong.com)

Research Interests:

- 3D Perception
  - LiDAR Segmentation
  - 3D Object Detection
  - Monocular Depth Estimation
- Data-Efficient Learning
  - Semi-Supervised Learning
  - Unsupervised Domain Adaptation
  - Visual Representation Learning
- Robustness
  - Data Robustness
  - Model Robustness

IEEE IV 2023
IEEE INTELLIGENT VEHICLES SYMPOSIUM
Anchorage, Alaska, USA // June 4 – 7, 2023

3D-DLAD Workshop

Lingdong Kong

Ph.D. Student, NUS

Short Bio

Ph.D. Student, School of Computing, National University of Singapore (https://ldkong.com)

Industrial Experiences:

- Motional
  - Autonomous vehicle intern, semantic mapping team in Singapore
  - Unsupervised domain adaptation for LiDAR segmentation
- OpenMMLab
  - MMDetction3D codebase
  - Supporting LiDAR segmentation related model & dataset benchmark
  - https://github.com/open-mmlab/mmdetection3d

IEEE IV 2023
IEEE INTELLIGENT VEHICLES SYMPOSIUM
Anchorage, Alaska, USA // June 4 – 7, 2023

3D-DLAD Workshop

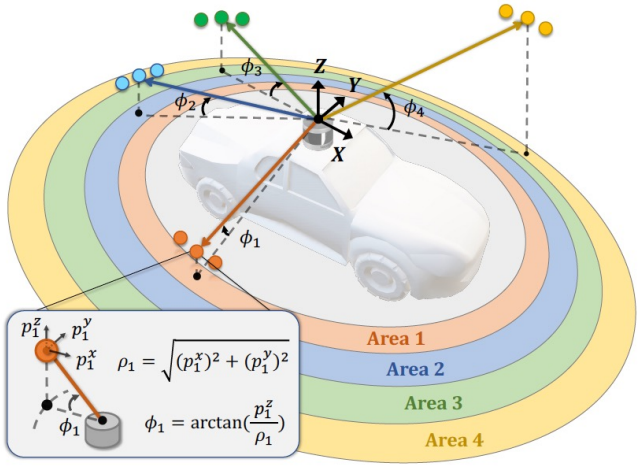
Lingdong Kong

Ph.D. Student, NUS

Robust Perception in Autonomous Driving

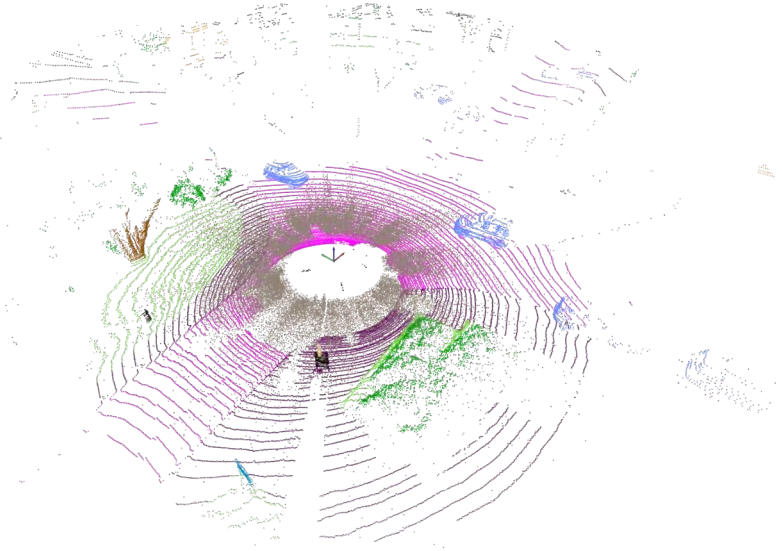
Topic #1: Robustness under Data-Efficient Settings

- Data collection is much easier than annotation
- Learning perception models with semi supervisions
- Goal: Achieve satisfactory perception performance with limited annotations



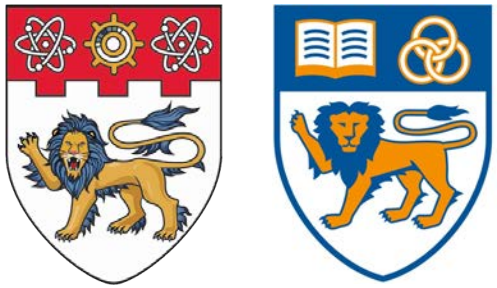
Topic #2: Robustness under Common Corruptions

- Data corruption and sensor failure are common issues
- Testing perception models with corrupted scenarios
- Goal: Achieve reliable perception performance under out-of-distribution corruptions



Topic #1

Robustness under Data-Efficient Settings



JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

LaserMix for Semi-Supervised LiDAR Semantic Segmentation

CVPR 2023 (Highlight)

Lingdong Kong

Jiawei Ren

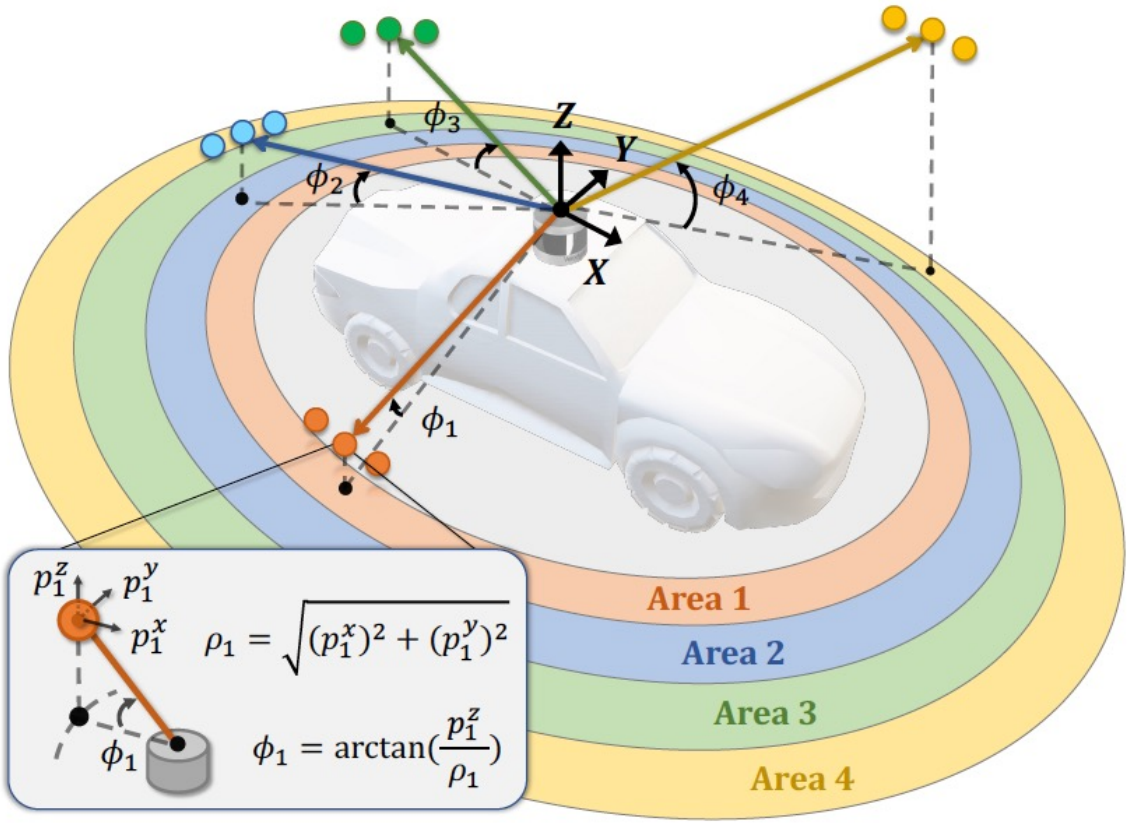
Liang Pan

Ziwei Liu

TL;DR

LaserMix is a data-efficient learning framework designed for LiDAR segmentation that:

- Leverages the spatial prior in driving scenes for data-efficient learning;
- Constructs low-variational areas via laser beam mixing;
- Encourages the model to make confident and consistent predictions before and after mixing;
- Achieves competitive results over full supervision counterparts with 2x to 5x fewer annotations

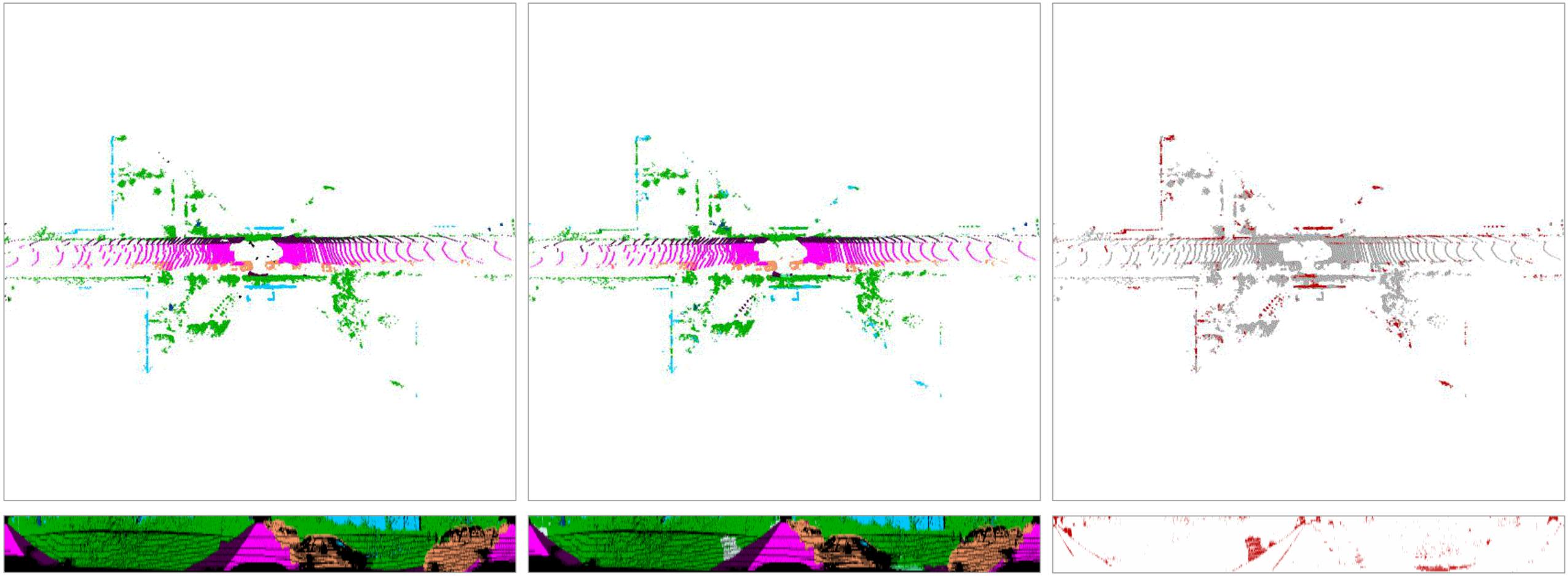


TL;DR

Groundtruth

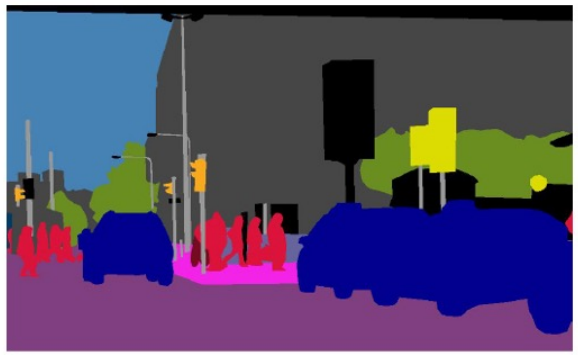
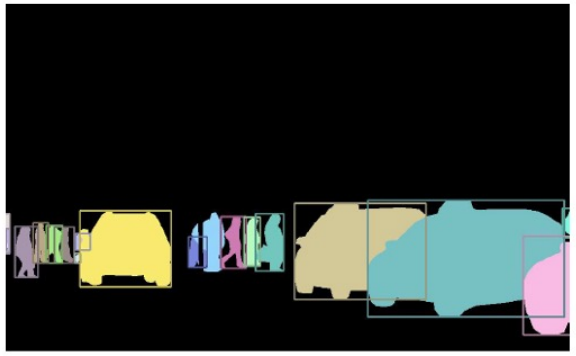
LaserMix (Prediction)

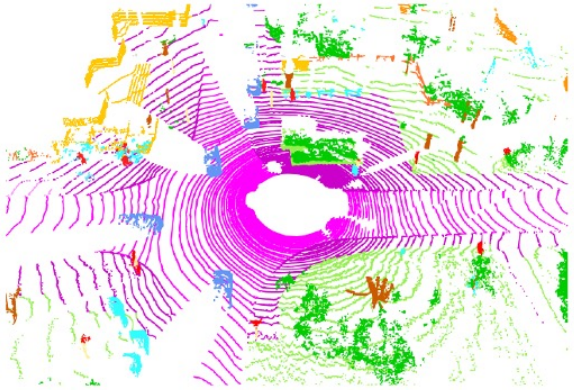
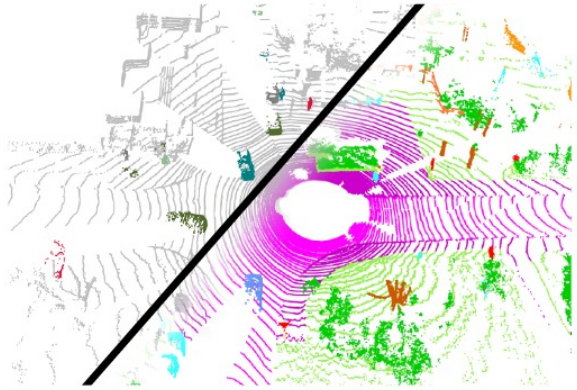
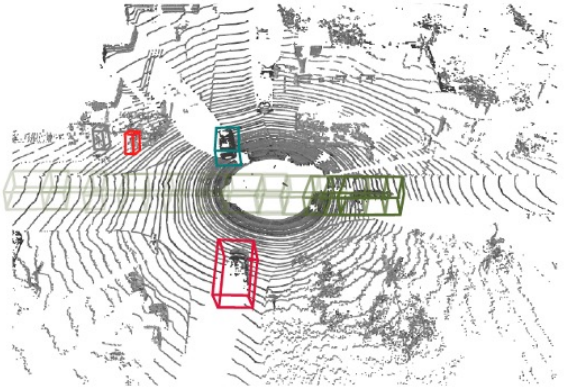
LaserMix (Error Map)



car bicy moto truc o.veh ped b.list m.list road park walk o.gro build fenc veg trun terr pole sign

Autonomous Driving Perception

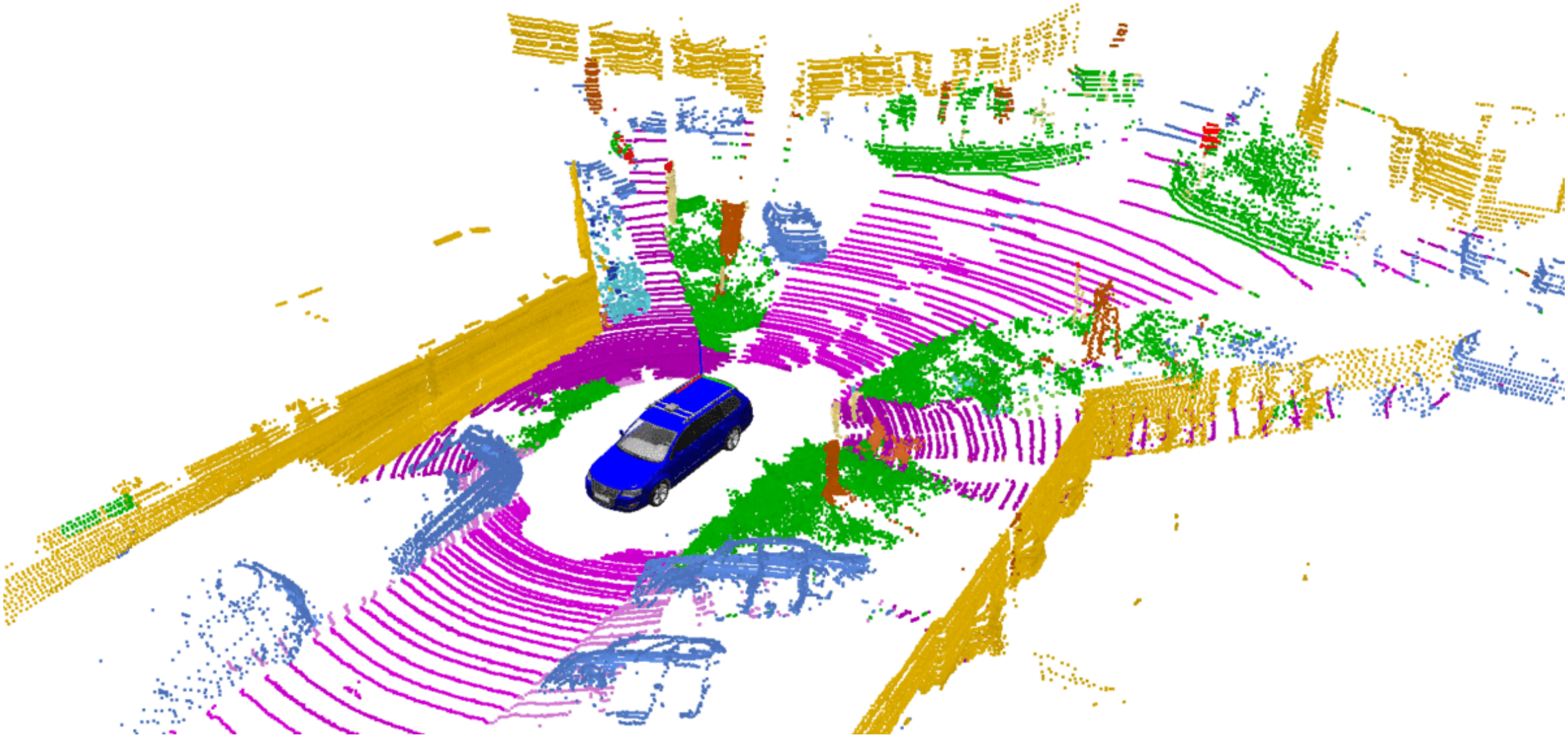

From left to right:

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

Why LiDAR sensors?

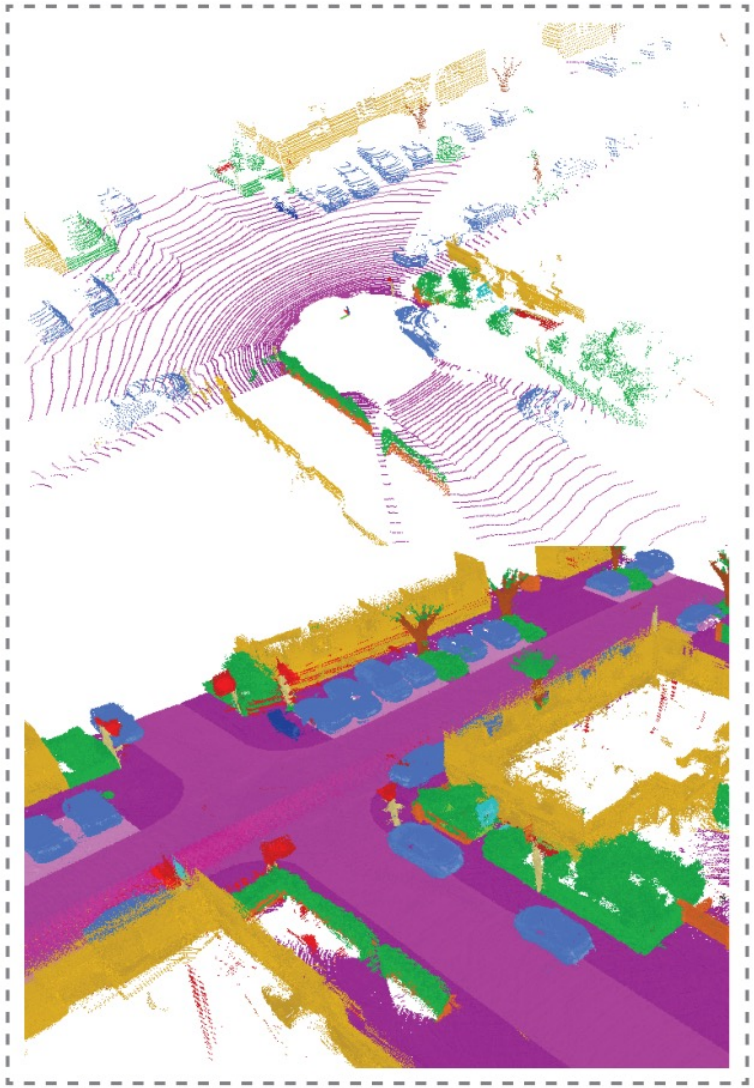
- Accurate depth sensing
- Robust at low-light conditions
- Dense perceptions
- …

LiDAR Semantic Segmentation



A. Milioto, et al. “RangeNet++: Fast and accurate LiDAR semantic segmentation,” IROS, 2019.

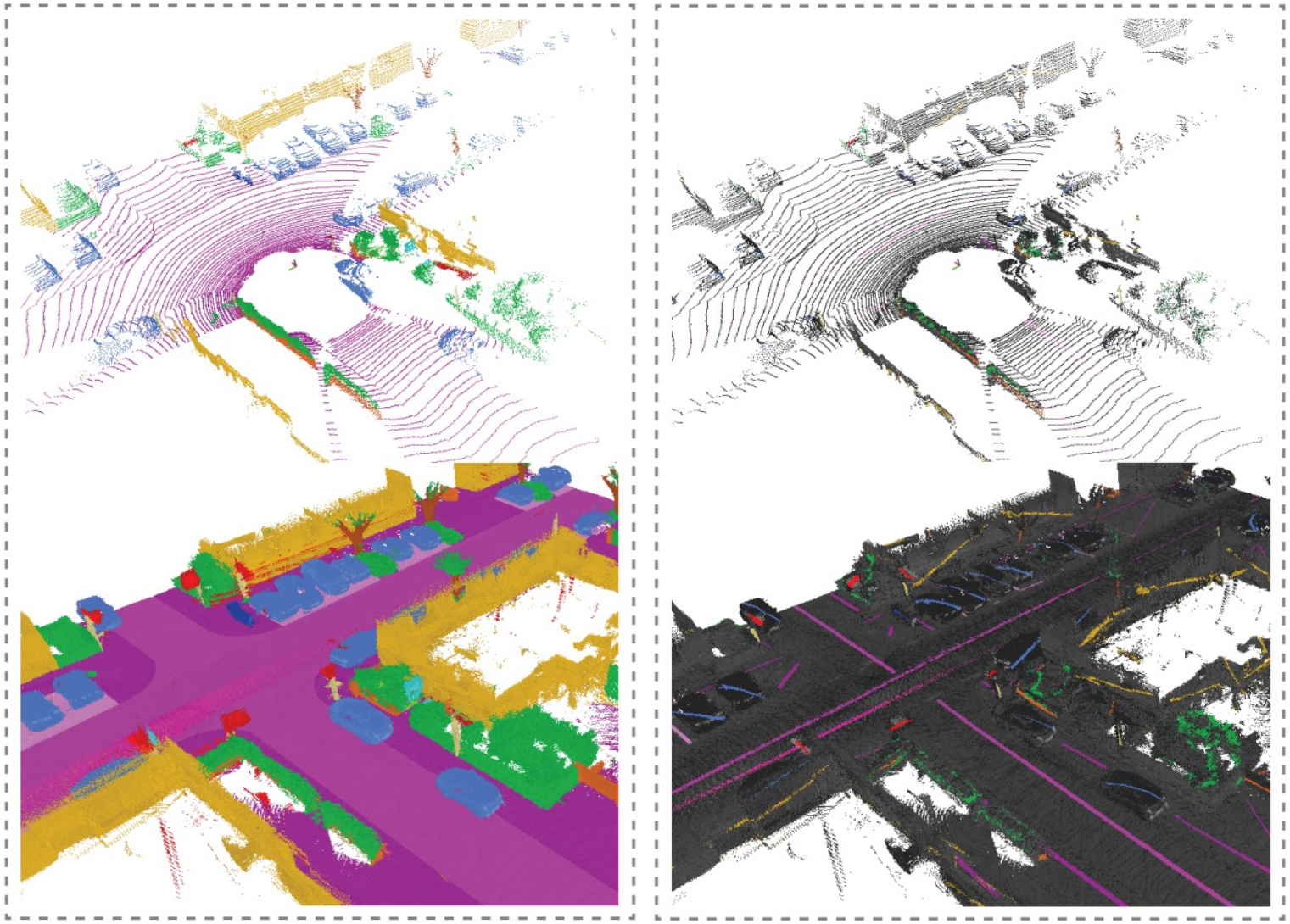
LiDAR Semantic Segmentation



- **SemanticKITTI**
  - Full labels (100%)
  - 19 semantic classes
  - 100 m x 100 m
  - Up to 4.5 hours

J. Behley, et al. “Semantickitti: A dataset for semantic scene understanding of LiDAR sequences,” ICCV, 2019.

LiDAR Semantic Segmentation



- **SemanticKITTI**
  - Full labels (100%)
  - 19 semantic classes
  - 100 m x 100 m
  - Up to 4.5 hours
- **ScribbleKITTI**
  - Weak (scribble) labels (8.06%)
  - 19 semantic classes
  - 100 m x 100 m
  - 10 - 25 min per scan
  - 90% time saving

O. Unal, et al. “Scribble-supervised LiDAR semantic segmentation,” CVPR, 2022.

Semi-Supervised LiDAR Segmentation

Objective

- We target on the less-explored semi-supervised LiDAR semantic segmentation.

Semi-Supervised LiDAR Segmentation

Objective

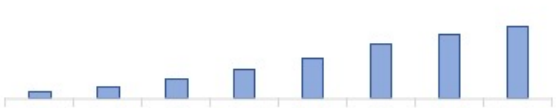
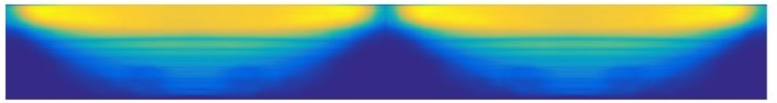
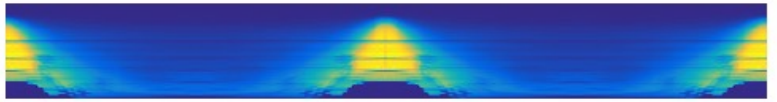
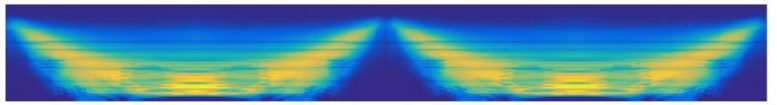
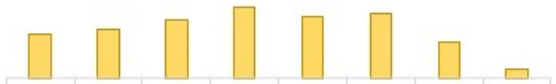
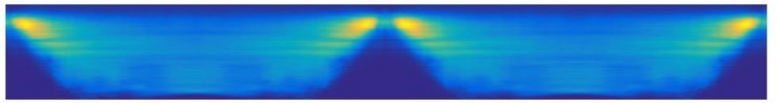
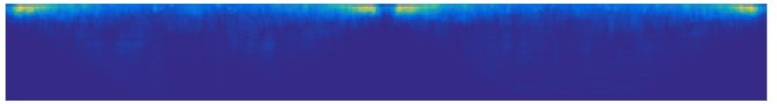
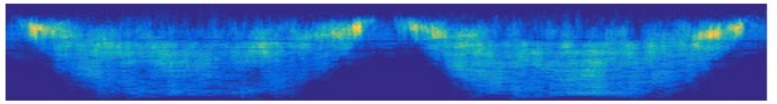
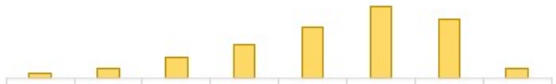
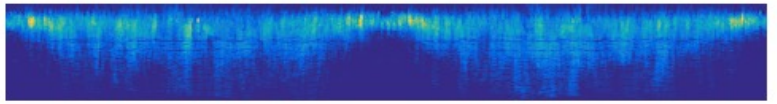
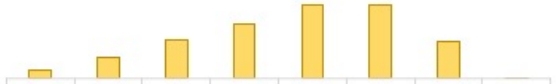
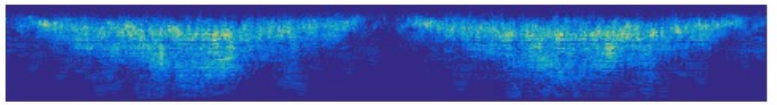
- We target on the less-explored semi-supervised LiDAR semantic segmentation.
- Our goal is to leverage the abundant raw LiDAR scans for training accurate segmentation models.

Semi-Supervised LiDAR Segmentation

Objective

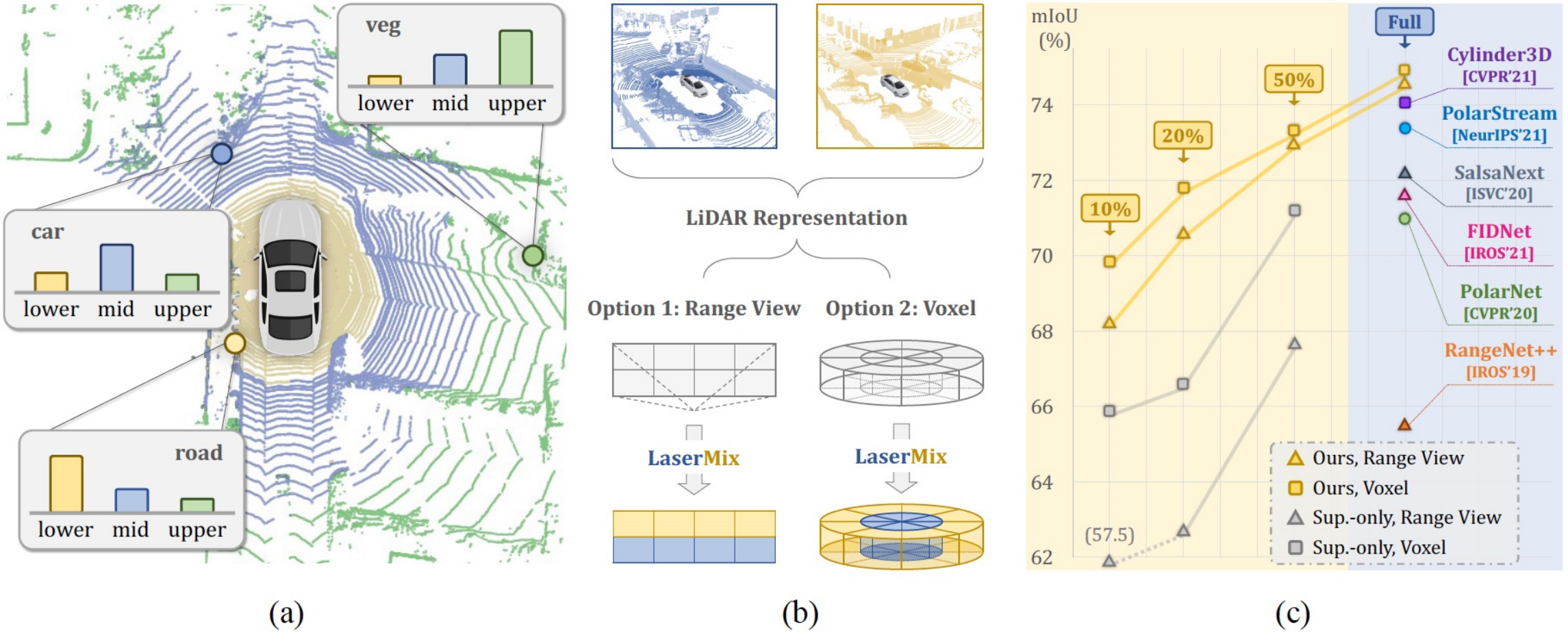
- We target on the less-explored semi-supervised LiDAR semantic segmentation.
- Our goal is to leverage the abundant raw LiDAR scans for training accurate segmentation models.
- We propose **LaserMix** to make advantages of the spatial prior in LiDAR scenes for effective learning with semi supervisions.

Spatial Prior

Class	Type	Proportion	Distribution	Heatmap
vegetation	static	24.825%		
road	static	22.545%		
sidewalk	static	16.353%		
car	dynamic	4.657%		
traffic-sign	static	0.061%		
motorcycle	dynamic	0.045%		
person	dynamic	0.036%		
bicycle	dynamic	0.018%		

Certain class tends to appear at certain areas around the ego-vehicle!

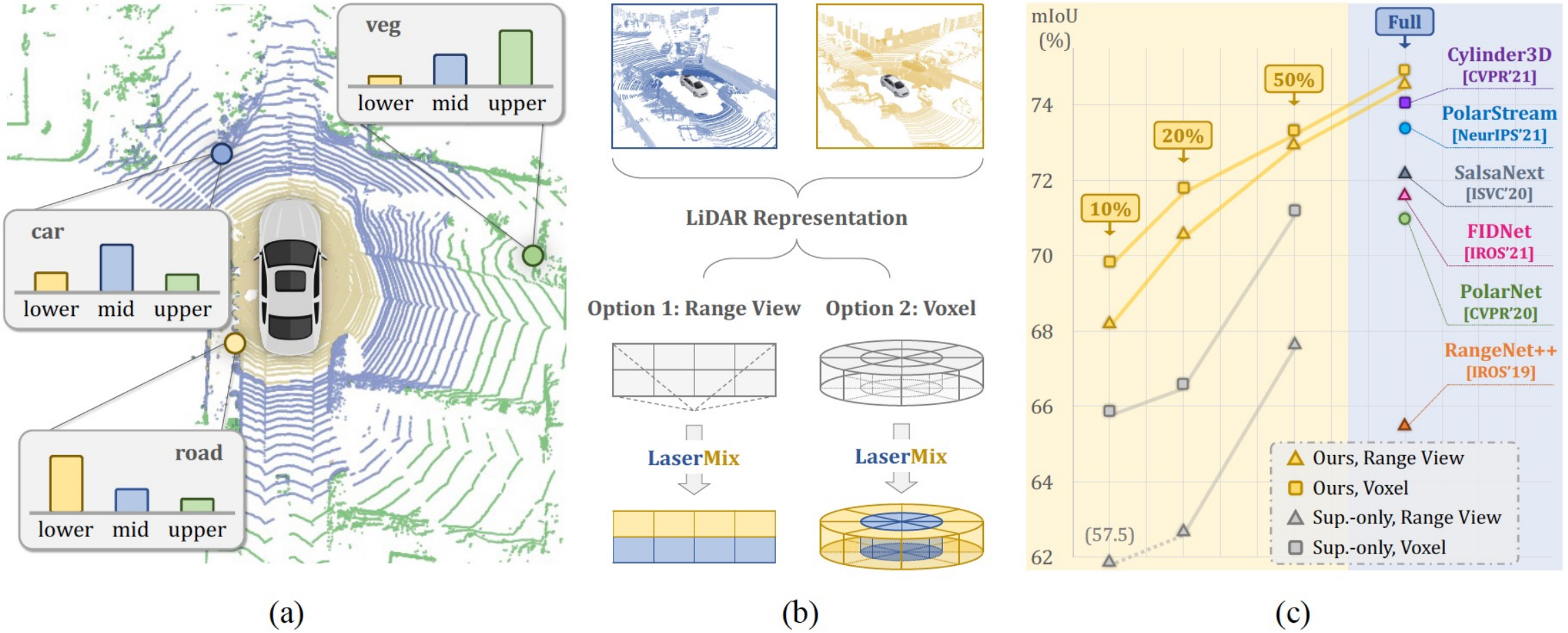
Overview



(a) (b) (c)

(a) Motivation. Semantic scene priors are overt for each category in LiDAR point clouds.

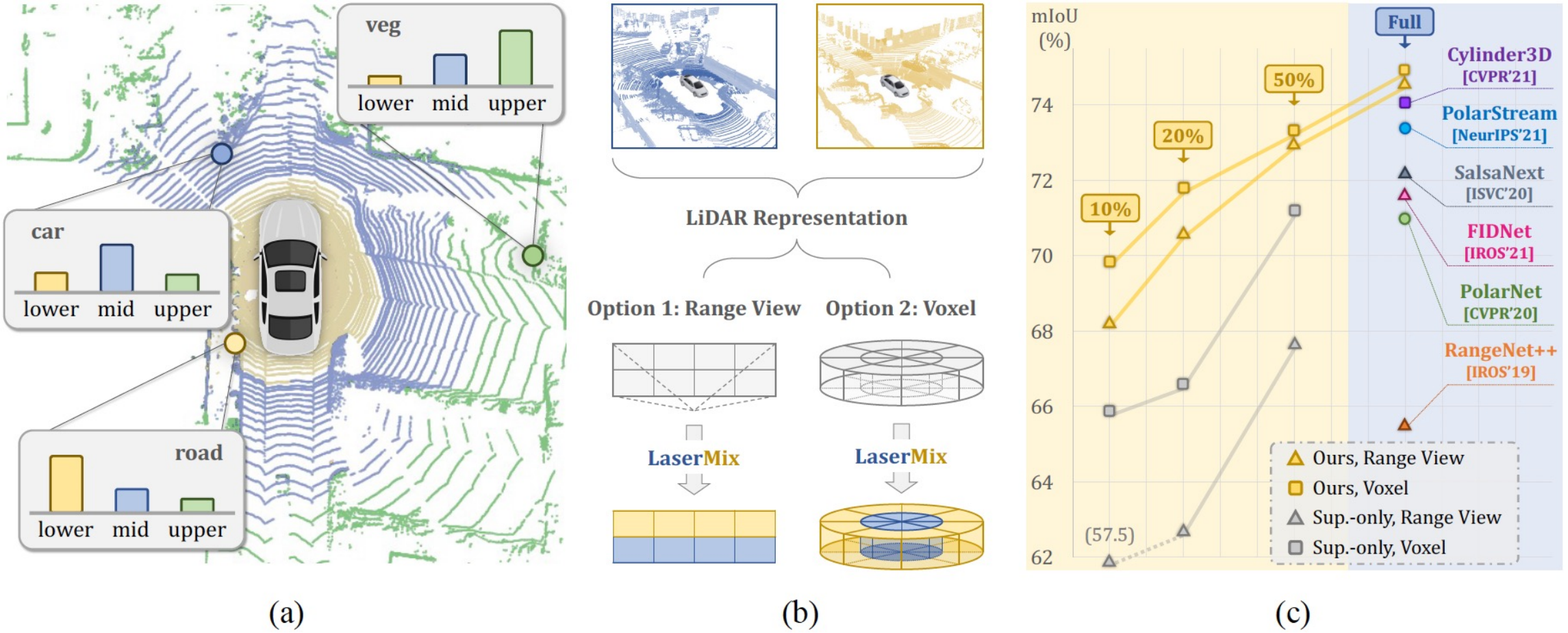
Overview



(a) (b) (c)

(a) Motivation. Semantic scene priors are overt for each category in LiDAR point clouds.
(b) Generalizability. LaserMix can be added into various popular LiDAR representations.

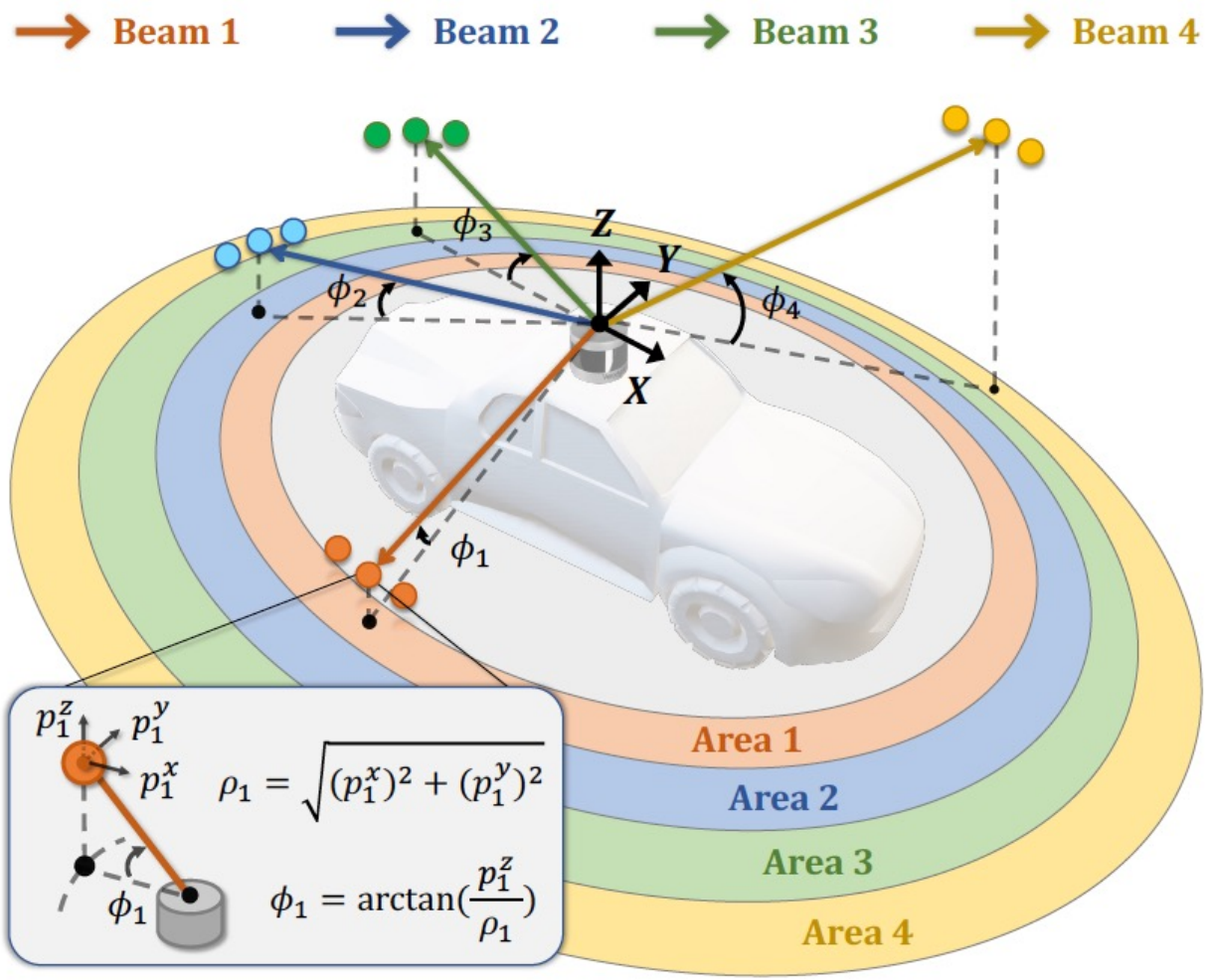
Overview



(a) (b) (c)

(a) Motivation. Semantic scene priors are overt for each category in LiDAR point clouds.
(b) Generalizability. LaserMix can be added into various popular LiDAR representations.
(c) Effectiveness. LaserMix helps to improve both semi- and fully-supervised settings.

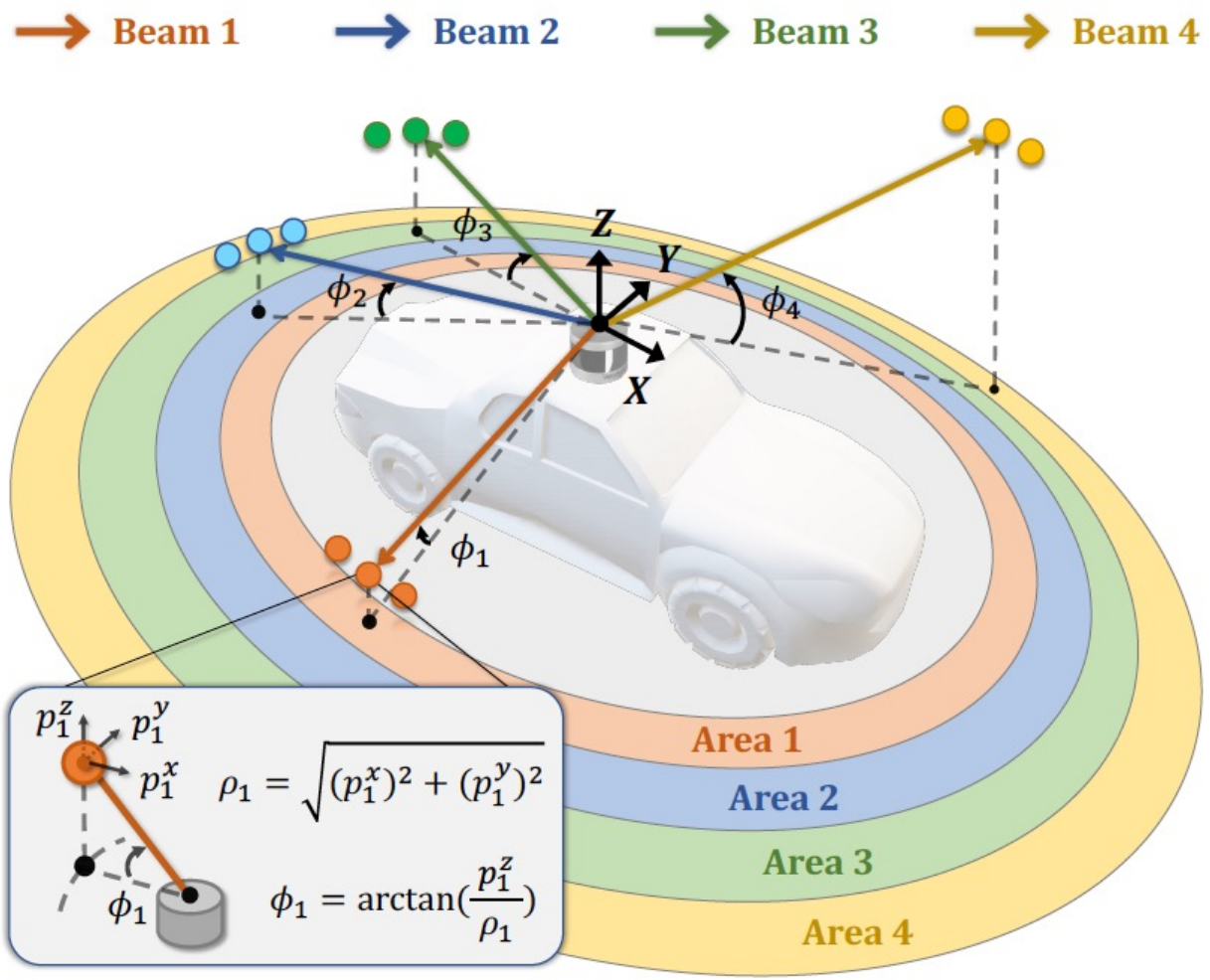
Laser Partition & Mixing



Three-Step Procedure

1. Partitioning the captured LiDAR scan into low-variational areas.

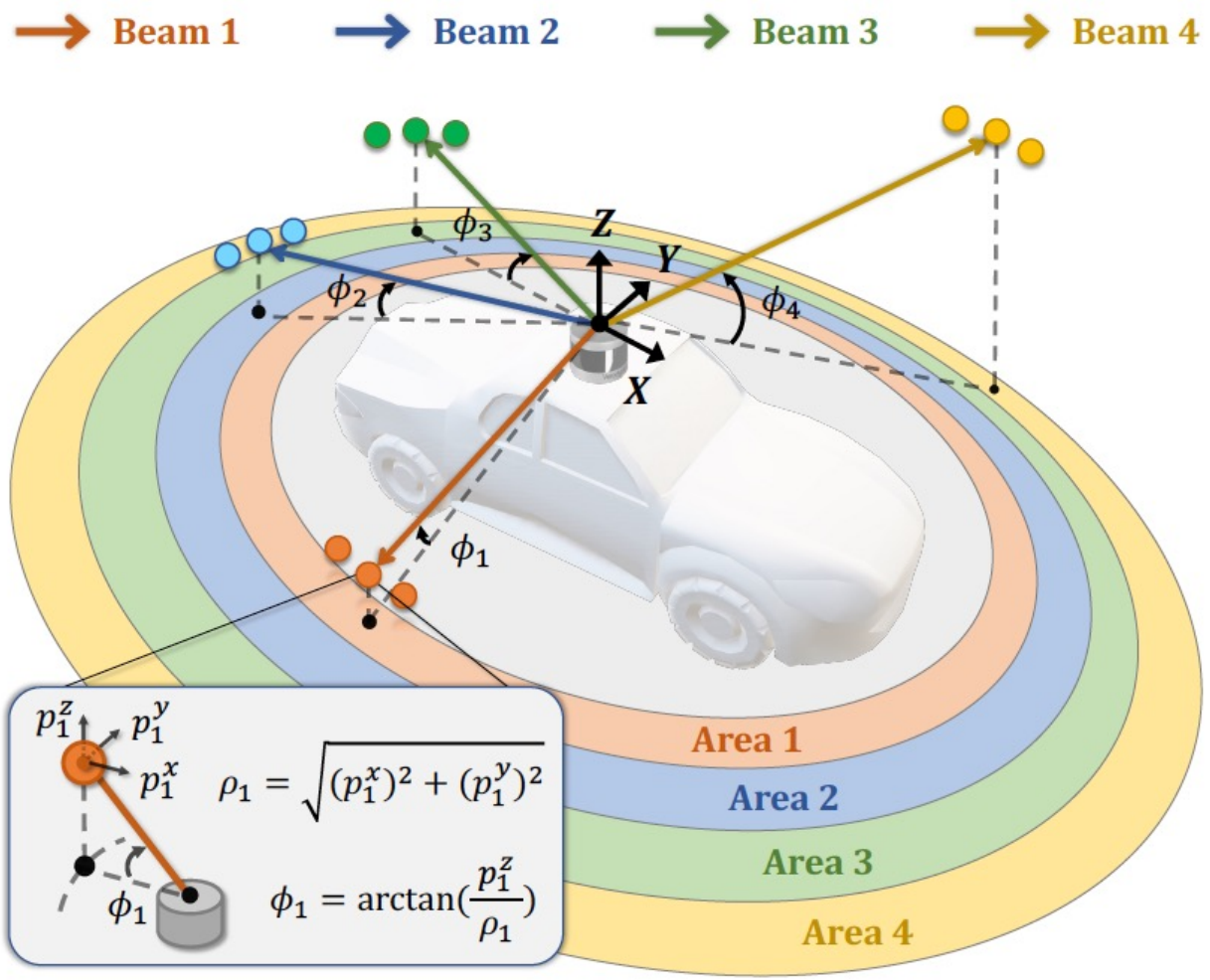
Laser Partition & Mixing



Three-Step Procedure

1. Partitioning the captured LiDAR scan into low-variational areas.
2. Efficiently mixing every area in the LiDAR scan with foreign data.

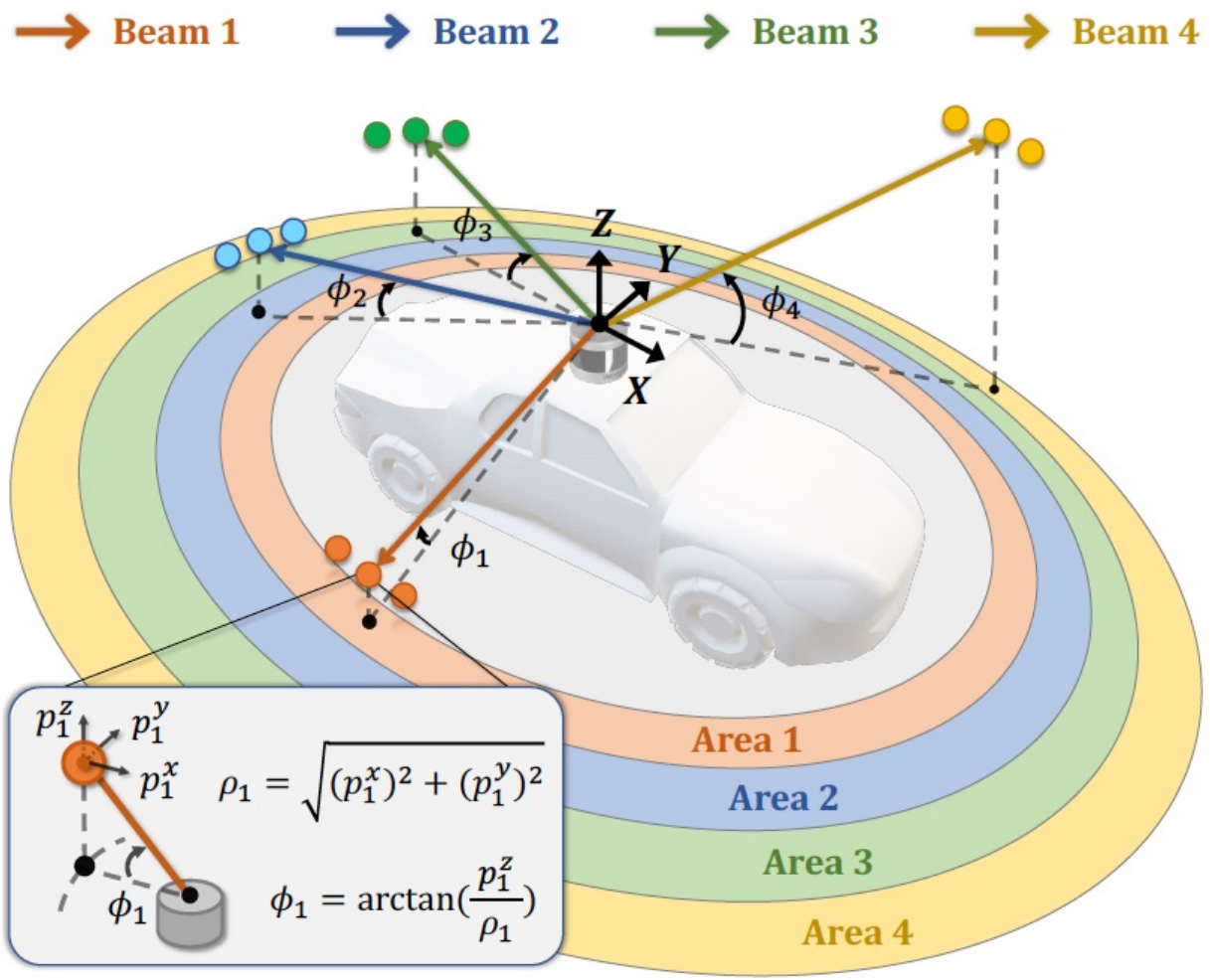
Laser Partition & Mixing



Three-Step Procedure

1. Partitioning the captured LiDAR scan into low-variational areas.
2. Efficiently mixing every area in the LiDAR scan with foreign data.
3. Encouraging the LiDAR segmentation models to make confident and consistent predictions on the same area in different mixing.

Laser Partition & Mixing



- Inclination:

$$\phi_i = \arctan(\frac{p_i^z}{\sqrt{(p_i^x)^2+(p_i^y)^2}})$$

- Depth: $\rho_i = \sqrt{(p_i^x)^2+(p_i^y)^2}$

- Azimuth: $\alpha_i = \arctan(\frac{p_i^y}{p_i^x})$

Laser Partition & Mixing

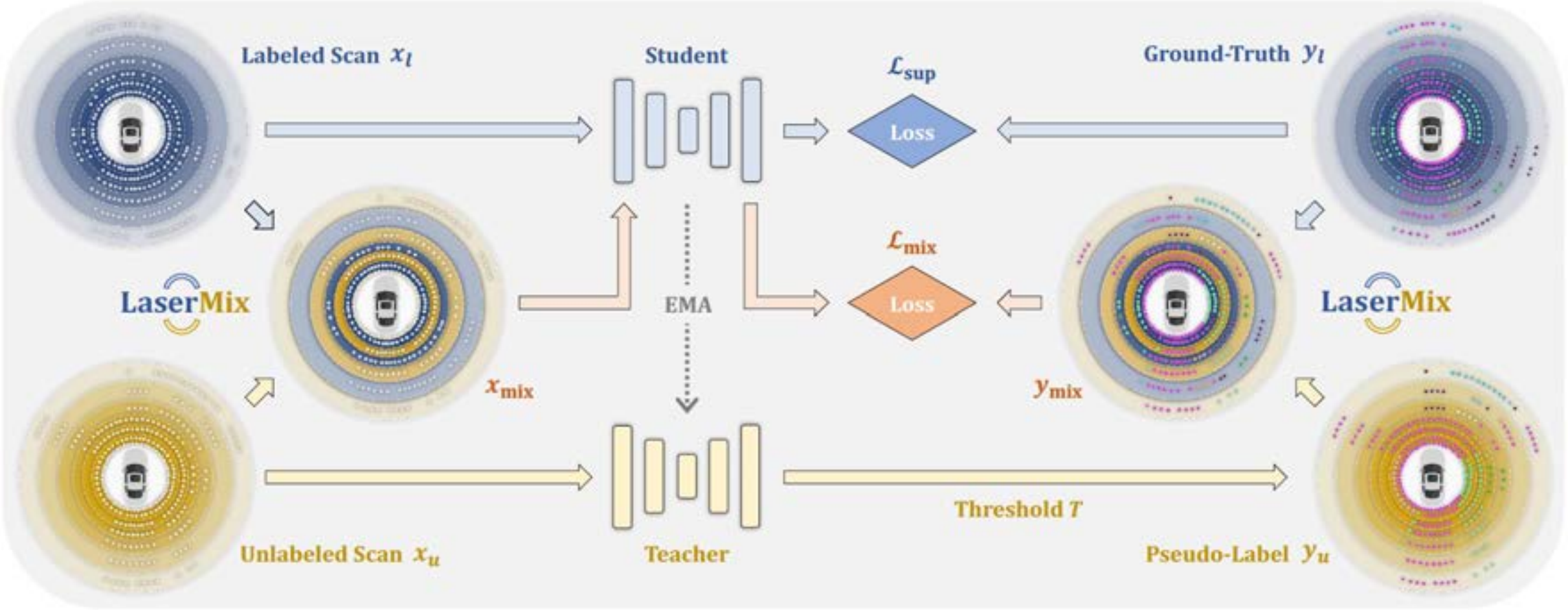
- Inclination:

$$\phi_i = \arctan\left(\frac{p_i^z}{\sqrt{(p_i^x)^2+(p_i^y)^2}}\right)$$

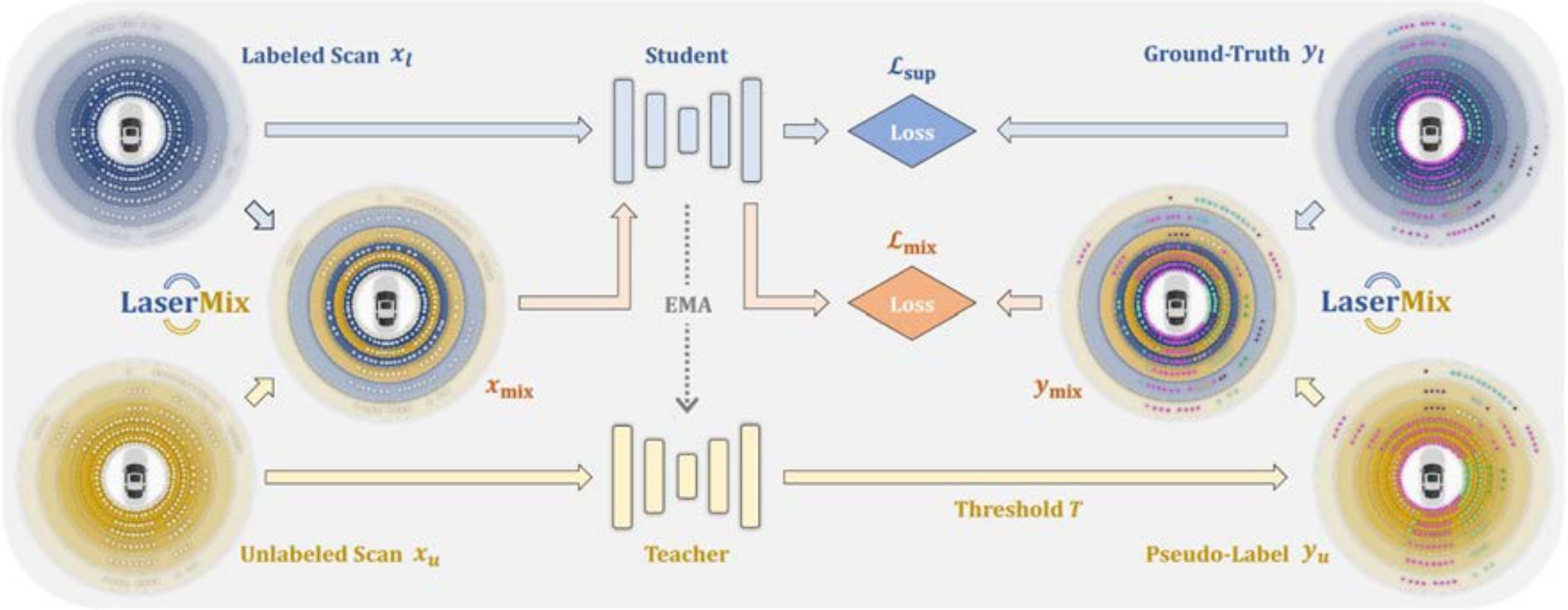
- Depth: $\rho_i = \sqrt{(p_i^x)^2+(p_i^y)^2}$

- Azimuth: $\alpha_i = \arctan\left(\frac{p_i^y}{p_i^x}\right)$

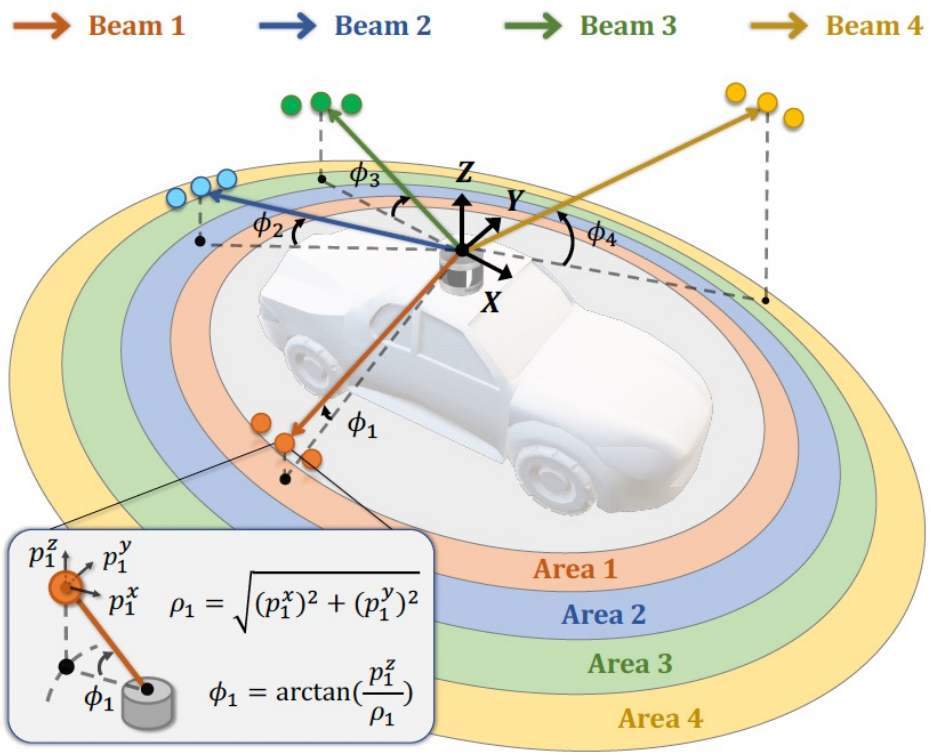
Consistency Regularization



Consistency Regularization



Derivation



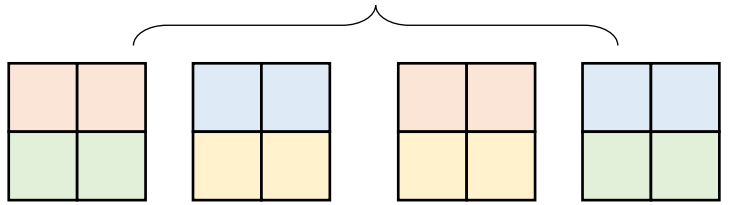
$H(X_{\text{in}}, Y_{\text{in}} \mid A = \text{Outer Rings})$

LiDAR data and labels strongly correlate with the area A

$H(X_{\text{in}}, Y_{\text{in}}|A)$ is low

A Simplified Case:

Color correlates with the row; each row has two colors



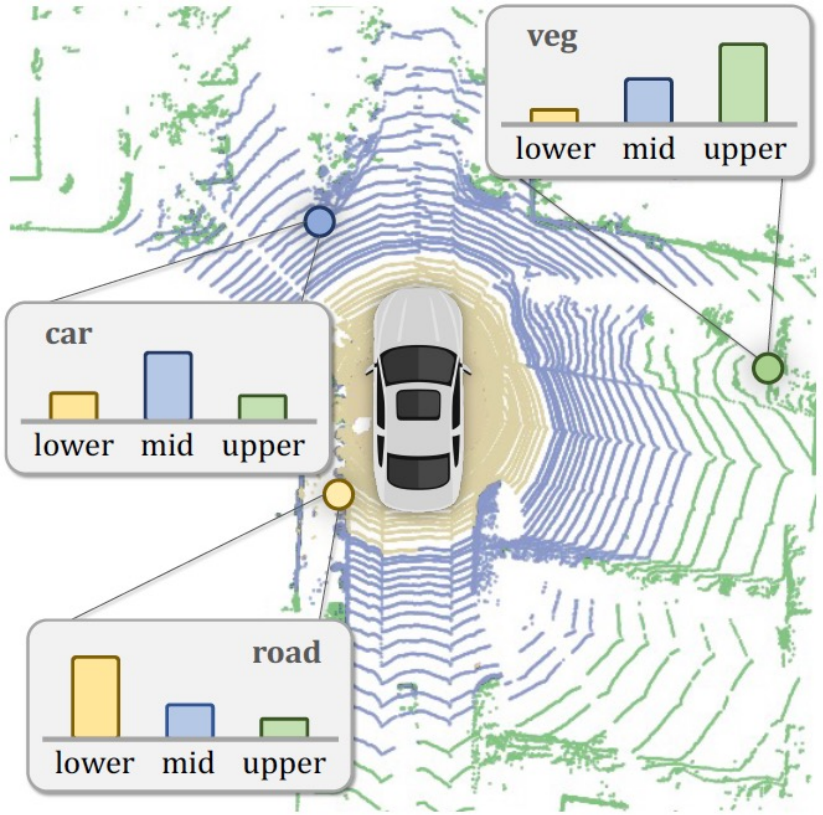
$H(X, Y|A \in \{2\times2\}) = \log 4$

$H(X, Y|A \in \{1\times2, 1\times2\}) = \log 4$

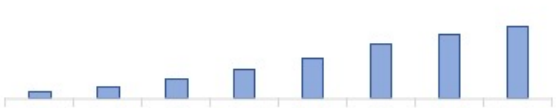
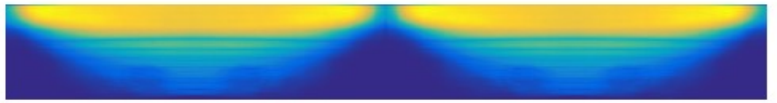
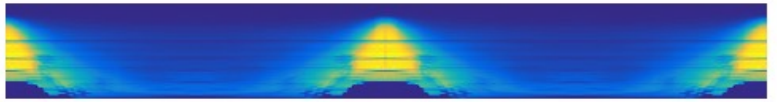
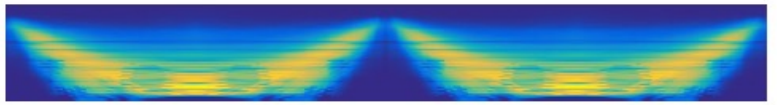
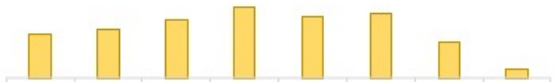
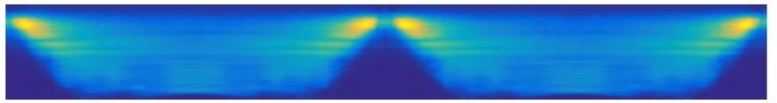
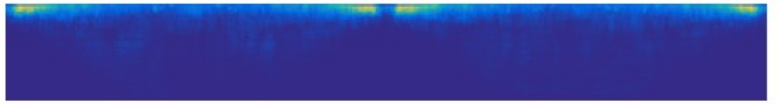
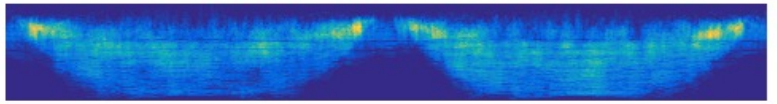
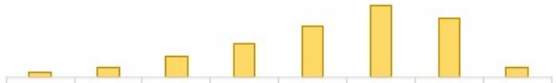
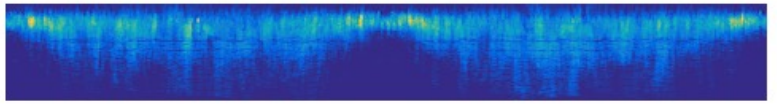
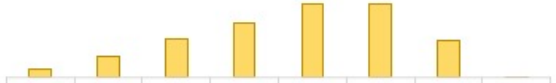
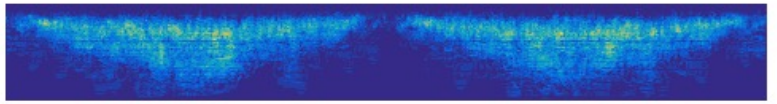
$H(X, Y|A \in \{2\times1, 2\times1\}) = \log 2$

Derivation

- $H(X_{\text{in}}, Y_{\text{in}}|A)$ is low => $H(Y_{\text{in}}|X_{\text{in}}, A)$ is low (conditional entropy).
- Let θ be the parameter of the LiDAR segmentation network.
- We would like to solve the following:
  - $E_\theta[H_\theta(Y_{\text{in}}|X_{\text{in}}, A)] = c$, where c is a constant.
  - $\sum_\theta P(\theta) = 1$ (sum to one).
- Principle of Maximum Entropy:
  - $P(\theta) \propto \underbrace{\exp(-\lambda\, H_\theta(Y_{\text{in}}|X_{\text{in}}, A))}_{\text{spatial prior}}$, where λ is the Lagrange multiplier.



Derivation

Class	Type	Proportion	Distribution	Heatmap
vegetation	static	24.825%		
road	static	22.545%		
sidewalk	static	16.353%		
car	dynamic	4.657%		
traffic-sign	static	0.061%		
motorcycle	dynamic	0.045%		
person	dynamic	0.036%		
bicycle	dynamic	0.018%		

Certain class tends to appear at certain areas around the ego-vehicle!

Derivation

- $P(\theta) \propto \exp(-\lambda\, H_\theta(Y_{\text{in}}|X_{\text{in}}, A))$ => spatial prior.
- Compute the empirical entropy:
- $\hat{H}_\theta(Y_{\text{in}}|X_{\text{in}}, A) = E_{X_{\text{in}},Y_{\text{in}},A}[P_\theta(Y_{\text{in}}|X_{\text{in}}, A) \log P_\theta(Y_{\text{in}}|X_{\text{in}}, A)]$.
- $P_\theta(Y_{\text{in}}|X_{\text{in}}, A)$ means predicting the labels by the data inside an area A.
- The segmentation network predicts from full data. Therefore, we need X_{out} to complement the remaining area outside A and marginalize X_{out}.
- $P_\theta(y_{\text{in}}|x_{\text{in}}, a) = \frac{1}{|X_{\text{out}}|} \sum_{x_{\text{out}} \in X_{\text{out}}} P_\theta(y_{\text{in}}|x_{\text{in}}, a, x_{\text{out}})$.

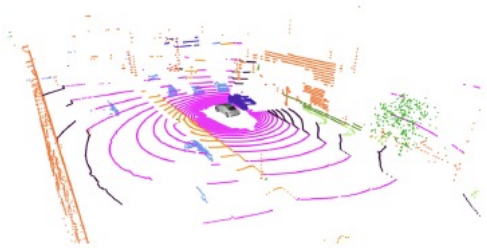
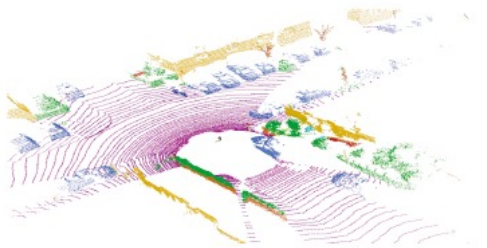
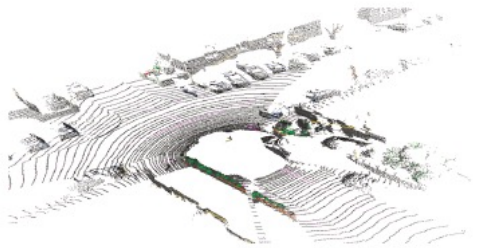
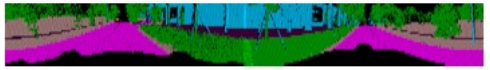
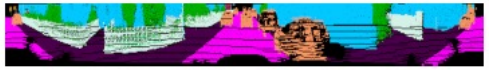
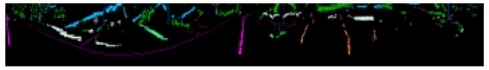
Derivation

- $P(\theta) \propto \exp(-\lambda\, H_\theta(Y_{\text{in}}|X_{\text{in}}, A))$ => spatial prior.
- $P_\theta(y_{\text{in}}|x_{\text{in}}, a) = E_{X_{\text{out}}}[P_\theta(y_{\text{in}}|x_{\text{in}}, a, x_{\text{out}})]$ => marginalization.
- We maximize the following posterior:
  - $C(\theta) = -\lambda \hat{E}_{x_{\text{in}} \in X_{\text{in}}, y_{\text{in}} \in Y_{\text{in}},\, a \in A}[H]$,
  - $H = P_\theta(y_{\text{in}}|x_{\text{in}}, a) \cdot \log P_\theta(y_{\text{in}}|x_{\text{in}}, a)$,
  - H is minimized only when $P_\theta(y_{\text{in}}|x_{\text{in}}, a, x_{\text{out}})$ is certain and consistent to x_{out}.

Derivation

- $H = \frac{1}{|X_{\text{out}}|} \sum_{x_{\text{out}} \in X_{\text{out}}} P_\theta(y_{\text{in}} | x_{\text{in}}, a, x_{\text{out}}) \log P_\theta(y_{\text{in}} | x_{\text{in}}, a, x_{\text{out}})$.
- $H = 0$ only when $P_\theta(y_{\text{in}} | x_{\text{in}}, a, x_{\text{out}})$ is certain and consistent to x_{out}.
- For every selected area and the data *inside* that area, a LiDAR segmentation network should make certain and consistent predictions regardless of the data *outside* the area.
- Directly compute $E_{y_{\text{in}} \in Y_{\text{in}}}[H]$ is infeasible / intractable, since $|y_{\text{in}}| = C^{H_{\text{in}} \times W_{\text{in}}}$ is too large.
- Instead, we use the pseudo-label to make sure that $P_\theta(y_{\text{in}} | x_{\text{in}}, a, x_{\text{out}})$ is certain and consistent.

Experimental Settings

	nuScenes [15]	SemanticKITTI [16]	ScribbleKITTI [4]
Vis.			
#Class	16	19	19
#Train	29130	19130	19130
#Val	6019	4071	4071
Res. (RV)	32×1920	64×2048	64×2048
Res. (voxel)	$[240, 180, 20]$	$[240, 180, 20]$	$[240, 180, 20]$
#Beam	32	64	64
$[\phi_{\text{up}}, \phi_{\text{low}}]$	$[10°, -30°]$	$[3°, -25°]$	$[3°, -25°]$
$[p^x_{\text{max}}, p^x_{\text{min}}]$	$[50m, -50m]$	$[50m, -50m]$	$[50m, -50m]$
$[p^y_{\text{max}}, p^y_{\text{min}}]$	$[50m, -50m]$	$[50m, -50m]$	$[50m, -50m]$
$[p^z_{\text{max}}, p^z_{\text{min}}]$	$[3m, -5m]$	$[2m, -4m]$	$[2m, -4m]$
#Label	100%	100%	8.06%
Intensity			
Range			
Semantics			

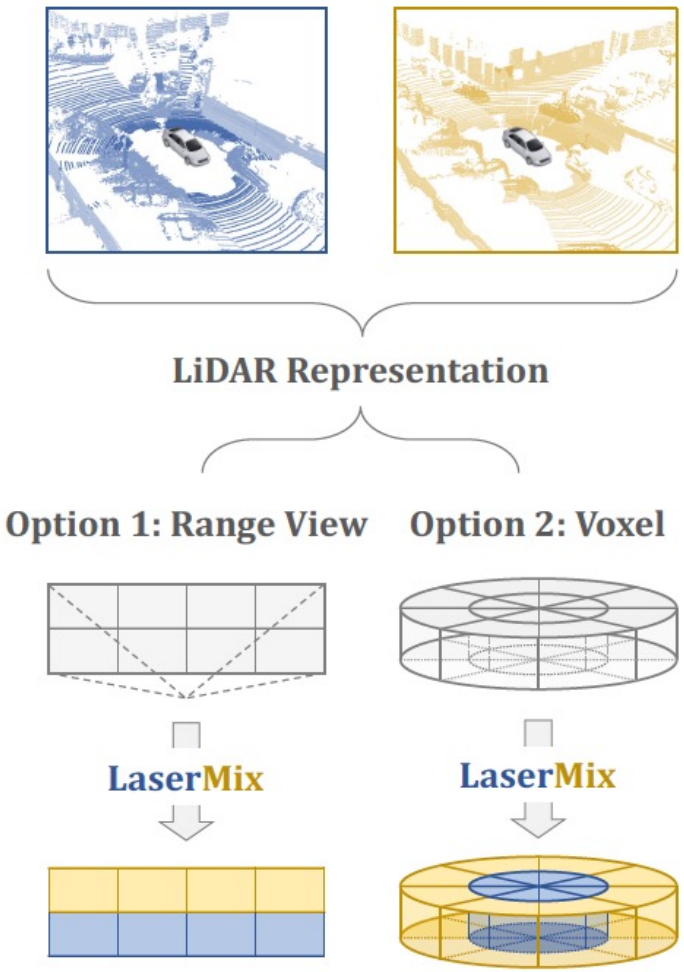
High-res LiDAR:
- SemanticKITTI
- Denser scenes

Low-res LiDAR:
- nuScenes
- Sparser scenes

Weak supervision:
- ScribbleKITTI
- Sparse labels

Experimental Settings

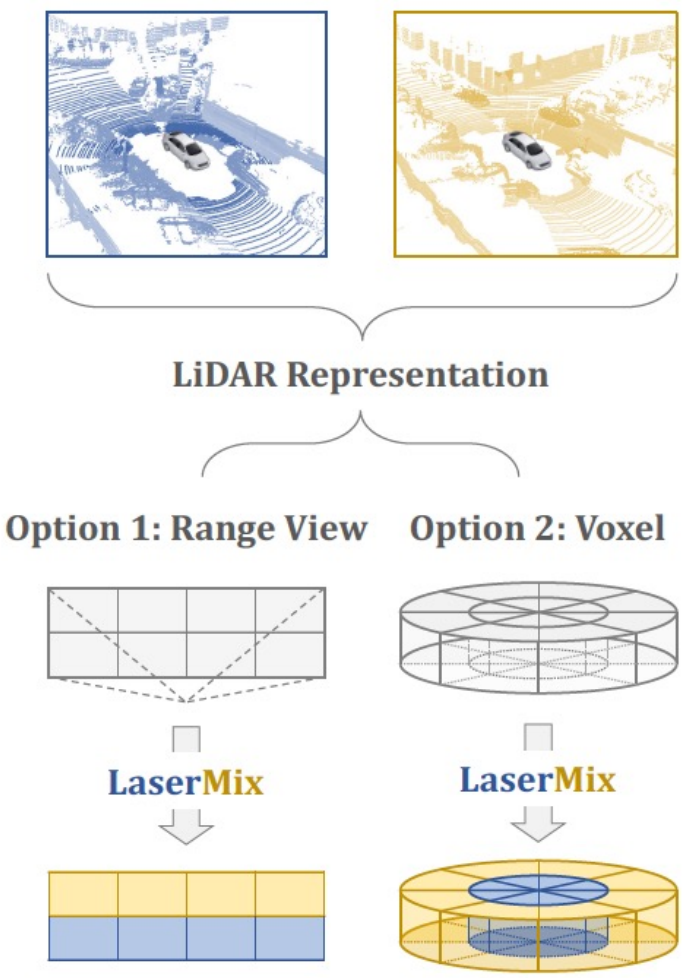
- **Range View**
  - Backbone: FIDNet [IROS'21]
  - # Param: 6.05M
  - 6 x 32 x 1920 (nuScenes)
  - 6 x 64 x 2048 (SemanticKITTI/ScribbleKITTI)
- **Voxel**
  - Backbone: Cylinder3D [CVPR'21]
  - # Param: 28.13M
  - [240, 180, 20]



Y. Zhao, et al. "FIDNet: LiDAR point cloud semantic segmentation with fully interpolation decoding," IROS, 2021.
X. Zhu, et al. "Cylindrical and asymmetrical 3D convolution networks for LiDAR segmentation," CVPR, 2021.

Experimental Settings

- **Range View**
  - Backbone: FIDNet [IROS'21]
  - # Param: 6.05M
  - 6 x 32 x 1920 (nuScenes)
  - 6 x 64 x 2048 (SemanticKITTI/ScribbleKITTI)
- **Voxel**
  - Backbone: Cylinder3D [CVPR'21]
  - # Param: 28.13M
  - [240, 180, 20]
- **Data Split**
  - 1%, 10%, 20%, 50% (labeled)
  - Random sampling
  - Assume the remaining ones are unlabeled



Y. Zhao, et al. "FIDNet: LiDAR point cloud semantic segmentation with fully interpolation decoding," IROS, 2021.
X. Zhu, et al. "Cylindrical and asymmetrical 3D convolution networks for LiDAR segmentation," CVPR, 2021.

Experimental Results

Repr.	Method	nuScenes [15] 1%	nuScenes [15] 10%	nuScenes [15] 20%	nuScenes [15] 50%	SemanticKITTI [16] 1%	SemanticKITTI [16] 10%	SemanticKITTI [16] 20%	SemanticKITTI [16] 50%	ScribbleKITTI [4] 1%	ScribbleKITTI [4] 10%	ScribbleKITTI [4] 20%	ScribbleKITTI [4] 50%
Range View	*Sup.-only*	38.3	57.5	62.7	67.6	36.2	52.2	55.9	57.2	33.1	47.7	49.9	52.5
	MeanTeacher [26]	42.1	60.4	65.4	69.4	37.5	53.1	56.1	57.4	34.2	49.8	51.6	53.3
	CBST [30]	40.9	60.5	64.3	69.3	39.9	53.4	56.1	56.9	35.7	50.7	52.7	54.6
	CutMix-Seg [29]	43.8	63.9	64.8	69.8	37.4	54.3	56.6	57.6	36.7	50.7	52.9	54.3
	CPS [13]	40.7	60.8	64.9	68.0	36.5	52.3	56.3	57.4	33.7	50.0	52.8	54.6
	LaserMix (Ours)	**49.5**	**68.2**	**70.6**	**73.0**	**43.4**	**58.8**	**59.4**	**61.4**	**38.3**	**54.4**	**55.6**	**58.7**
	$\Delta\uparrow$	+11.2	+10.7	+7.9	+5.4	+7.2	+6.6	+3.5	+4.2	+5.2	+6.7	+5.7	+6.2
Voxel	*Sup.-only*	50.9	65.9	66.6	71.2	45.4	56.1	57.8	58.7	39.2	48.0	52.1	53.8
	MeanTeacher [26]	51.6	66.0	67.1	71.7	45.4	57.1	59.2	60.0	41.0	50.1	52.8	53.9
	CBST [30]	53.0	66.5	69.6	71.6	48.8	58.3	59.4	59.7	41.5	50.6	53.3	54.5
	CPS [13]	52.9	66.3	70.0	72.5	46.7	58.7	59.6	60.5	41.4	51.8	53.9	54.8
	LaserMix (Ours)	**55.3**	**69.9**	**71.8**	**73.2**	**50.6**	**60.0**	**61.9**	**62.3**	**44.2**	**53.7**	**55.1**	**56.8**
	$\Delta\uparrow$	+4.4	+4.0	+5.2	+2.0	+5.2	+3.9	+4.1	+3.6	+5.0	+5.7	+3.0	+3.0

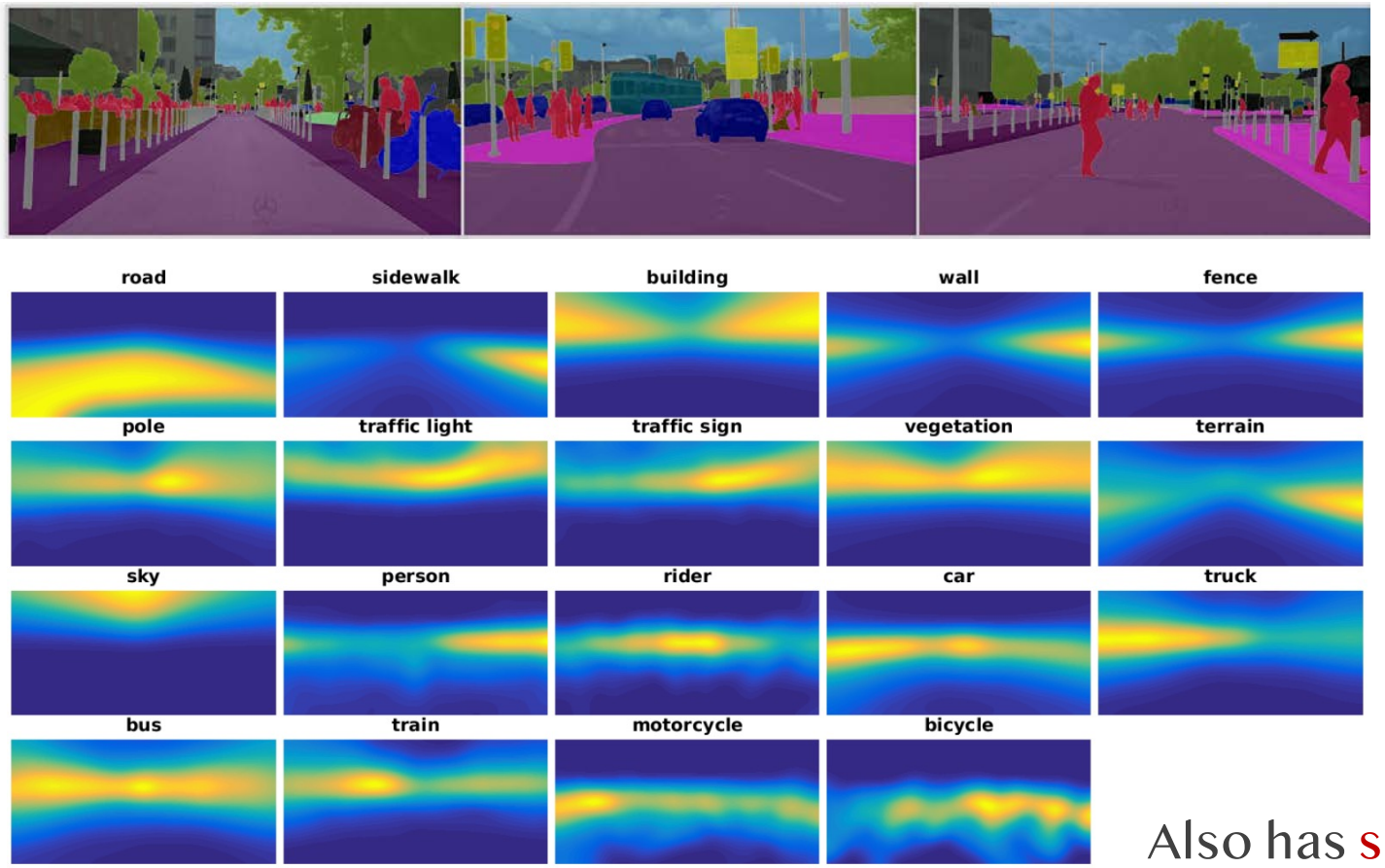
A. Tarvainen and H. Valpola. "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," NeurIPS, 2017.

G. French, et al. "Semi-supervised semantic segmentation needs strong, high-dimensional perturbations," BMVC, 2020.

Y. Zou, et al. "Domain adaptation for semantic segmentation via class-balanced self-training," ECCV, 2018.

X. Chen, et al. "Semi-supervised semantic segmentation with cross pseudo supervision," CVPR, 2021.

Experimental Results



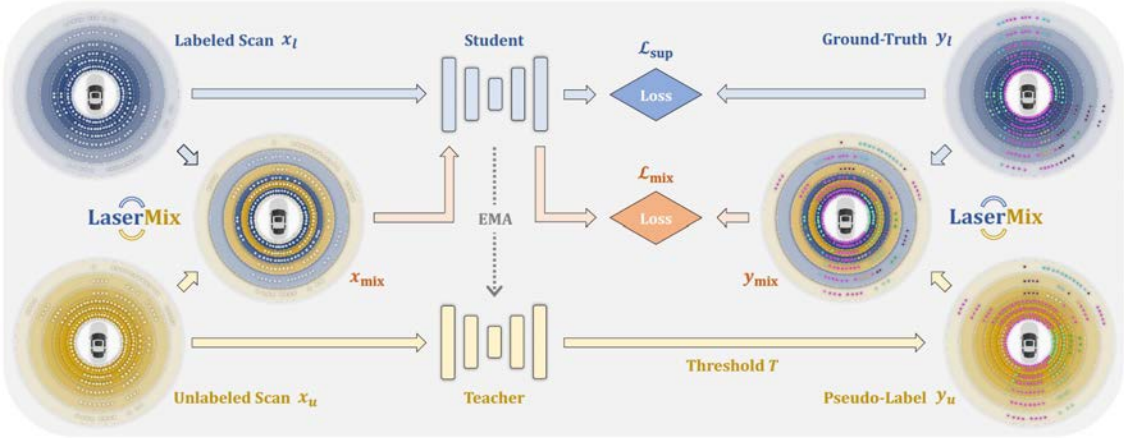
Method	1/16	1/8	1/4	1/2
MeanTeacher [26]	66.1	71.2	74.4	76.3
***w/* Ours**	68.7	72.3	75.7	76.8
$\Delta\uparrow$	+2.6	+1.1	+1.3	+0.5
CCT [11]	66.4	72.5	75.7	76.8
GCT [12]	65.8	71.3	75.3	77.1
CPS [13]	69.8	74.4	76.9	78.6
CPS-CutMix [13]	74.5	76.6	77.8	78.8
***w/* Ours**	75.5	77.1	78.3	79.1
$\Delta\uparrow$	+1.0	+0.5	+0.5	+0.3

Also has spatial priors in scenes!

Y. Ouali, et al. “Semi-supervised semantic segmentation with cross-consistency training,” CVPR, 2020.
Z. Ke, et al. “Guided collaborative training for pixel-wise semi-supervised learning,” ECCV, 2020.

Ablation Study

#	$\mathcal{L}_{\text{mt}}$	$\mathcal{L}_{\text{mix}}$	SS	TS	1%	10%	20%	50%
(1)	✓				42.1	60.4	65.4	69.4
(2)		✓	✓		45.6	64.3	67.8	71.6
	✓	✓	✓		47.0	65.5	69.5	72.0
(3)		✓		✓	46.0	64.1	69.5	72.3
	✓	✓		✓	49.5	68.2	70.6	73.0

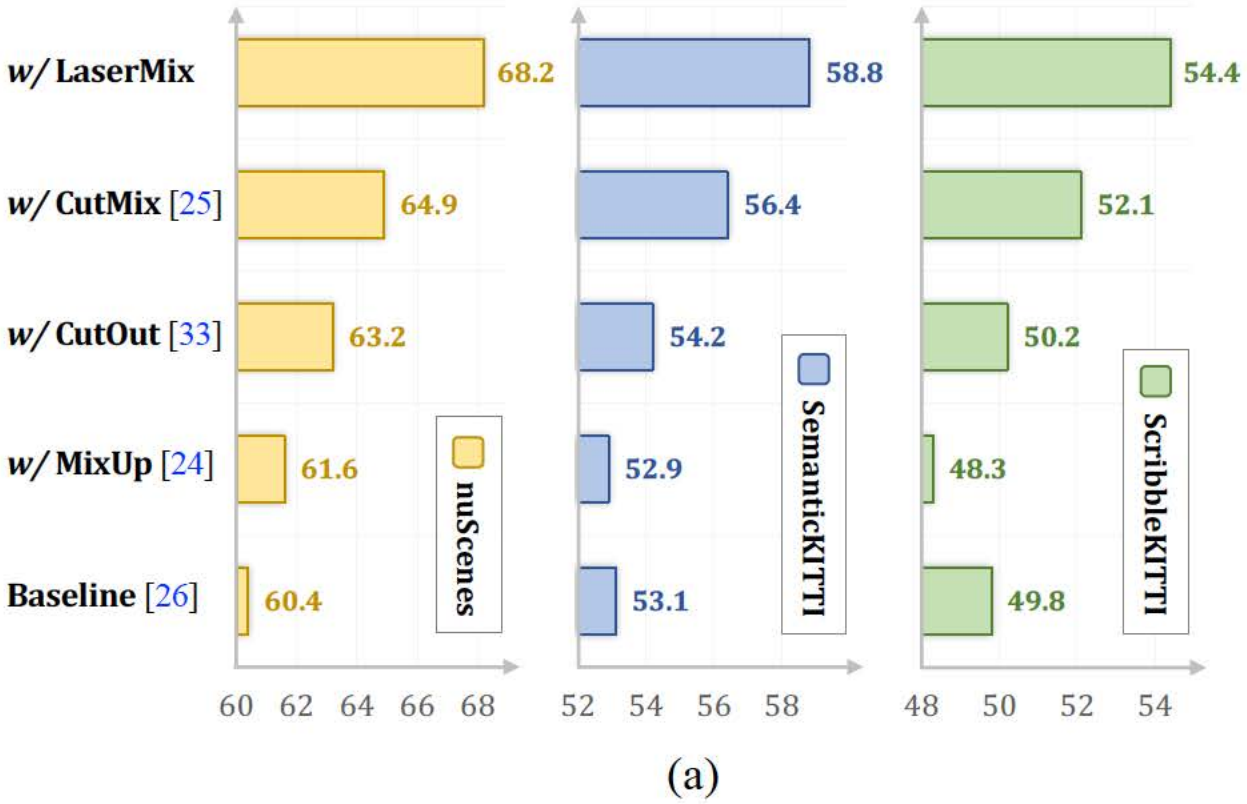


(1) Results of MeanTeacher.

(2) Results of LaserMix w/ student supervisions; much better than the counterpart.

(3) Results of LaserMix w/ teacher supervisions; much better than the counterpart.

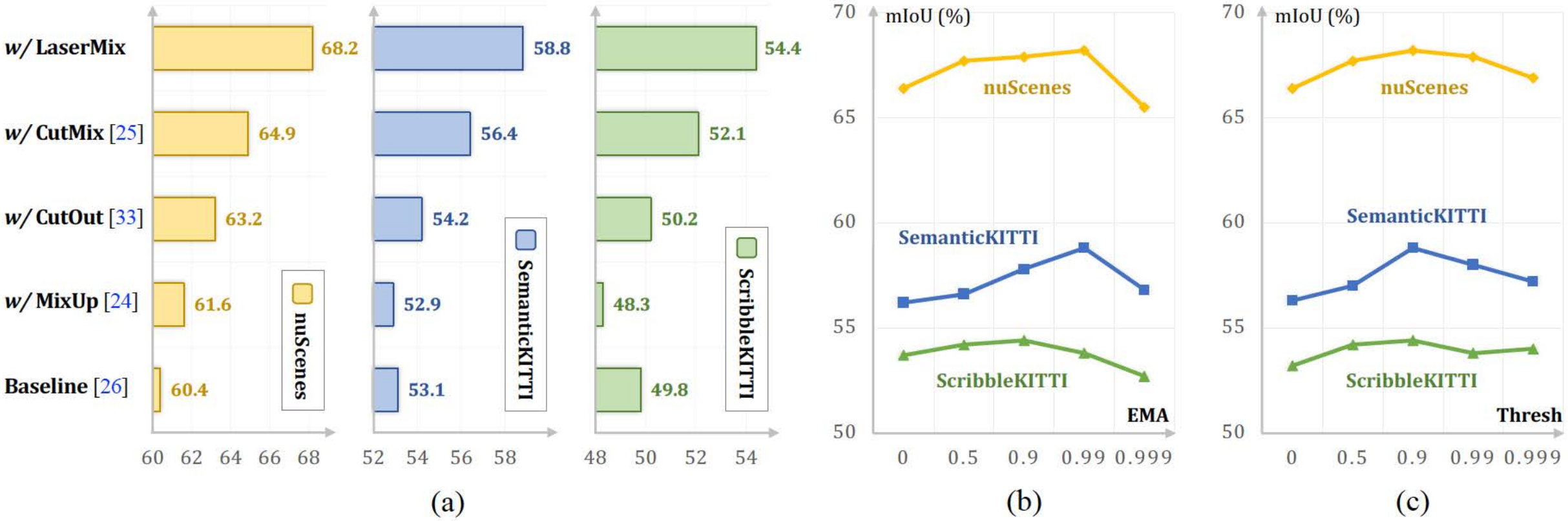
Ablation Study



(a) Comparisons among different mixing techniques.

A. Nekrasov, et al. "Mix3D: Out-of-context data augmentation for 3D scenes," 3DV, 2021.
S. Yun, et al. "Cutmix: Regularization strategy to train strong classifiers with localizable features," ICCV, 2019
T. DeVries and G. W. Taylor. "Improved regularization of convolutional neural networks with cutout," arXiv, 2017
H. Zhang, et al. "Mixup: Beyond empirical risk minimization," ICLR, 2018.

Ablation Study



(a) Comparisons among different mixing techniques. (b) EMA. (c) Confidence threshold.

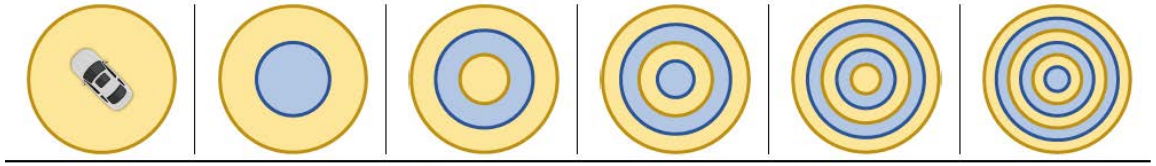
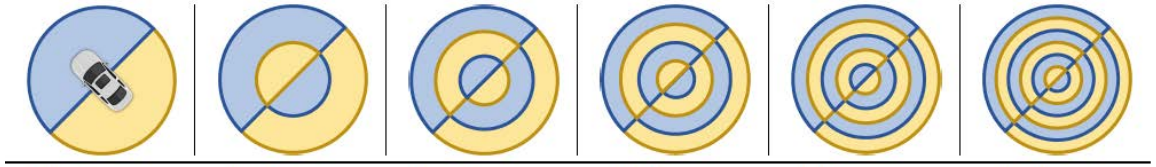
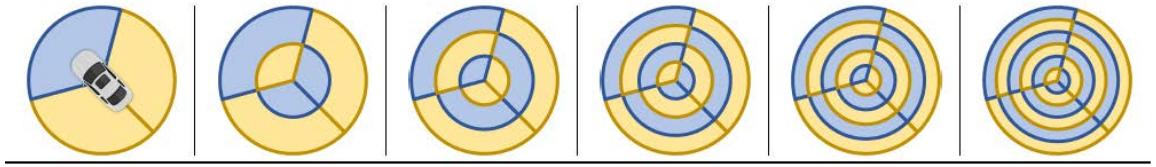
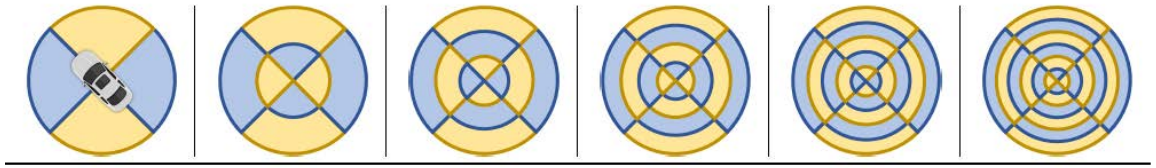
A. Nekrasov, et al. "Mix3D: Out-of-context data augmentation for 3D scenes," 3DV, 2021.

S. Yun, et al. "Cutmix: Regularization strategy to train strong classifiers with localizable features," ICCV, 2019

T. DeVries and G. W. Taylor. "Improved regularization of convolutional neural networks with cutout," arXiv, 2017

H. Zhang, et al. "Mixup: Beyond empirical risk minimization," ICLR, 2018.

Ablation Study

Baseline	$(1\alpha, 2\phi)$	$(1\alpha, 3\phi)$	$(1\alpha, 4\phi)$	$(1\alpha, 5\phi)$	$(1\alpha, 6\phi)$
					
60.4	$63.5_{(+3.1)}$	$65.2_{(+4.8)}$	$66.5_{(+6.1)}$	$66.2_{(+5.8)}$	$65.4_{(+5.0)}$
$(2\alpha, 1\phi)$	$(2\alpha, 2\phi)$	$(2\alpha, 3\phi)$	$(2\alpha, 4\phi)$	$(2\alpha, 5\phi)$	$(2\alpha, 6\phi)$
					
$61.5_{(+1.1)}$	$63.3_{(+2.9)}$	$65.9_{(+5.5)}$	$66.1_{(+5.7)}$	$66.7_{(+6.3)}$	$65.3_{(+4.9)}$
$(3\alpha, 1\phi)$	$(3\alpha, 2\phi)$	$(3\alpha, 3\phi)$	$(3\alpha, 4\phi)$	$(3\alpha, 5\phi)$	$(3\alpha, 6\phi)$
					
$60.9_{(+0.6)}$	$64.2_{(+3.8)}$	$65.9_{(+5.5)}$	$66.3_{(+5.9)}$	$66.0_{(+5.6)}$	$65.2_{(+4.8)}$
$(4\alpha, 1\phi)$	$(4\alpha, 2\phi)$	$(4\alpha, 3\phi)$	$(4\alpha, 4\phi)$	$(4\alpha, 5\phi)$	$(4\alpha, 6\phi)$
					
$60.9_{(+0.6)}$	$64.7_{(+4.3)}$	$65.3_{(+4.9)}$	$65.6_{(+5.2)}$	$65.7_{(+5.3)}$	$65.2_{(+4.8)}$

- Inclination:

$$\phi_i = \arctan\left(\frac{p_i^z}{\sqrt{(p_i^x)^2+(p_i^y)^2}}\right)$$

- Depth: $\rho_i = \sqrt{(p_i^x)^2+(p_i^y)^2}$

- Azimuth: $\alpha_i = \arctan\left(\frac{p_i^y}{p_i^x}\right)$

Public Resources

- **Paper:** https://arxiv.org/abs/2207.00026
- **Code:** https://github.com/ldkong1205/LaserMix
- **Tutorial:** https://zhuanlan.zhihu.com/p/528689803
- **Project Page:** https://ldkong.com/LaserMix

Topic #2

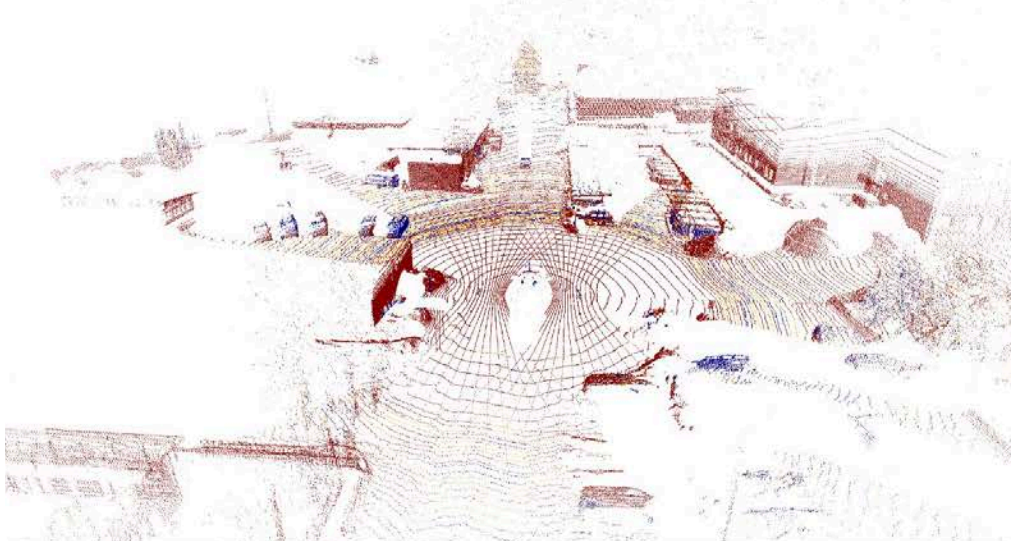
Robustness under Common Corruptions

Benchmarking 3D Perception Robustness to Common Corruptions and Sensor Failure

Lingdong Kong[1,2,*], Youquan Liu[1,3,*], Xin Li[1,4,*], Runnan Chen[1,5], Wenwei Zhang[1,6]
Jiawei Ren[6], Liang Pan[6], Kai Chen[1], Ziwei Liu[6]

[1]Shanghai AI Laboratory [2]NUS [3]Hochschule Bremerhaven
[4]ECNU [5]HKU [6]S-Lab, NTU

Perception Environment



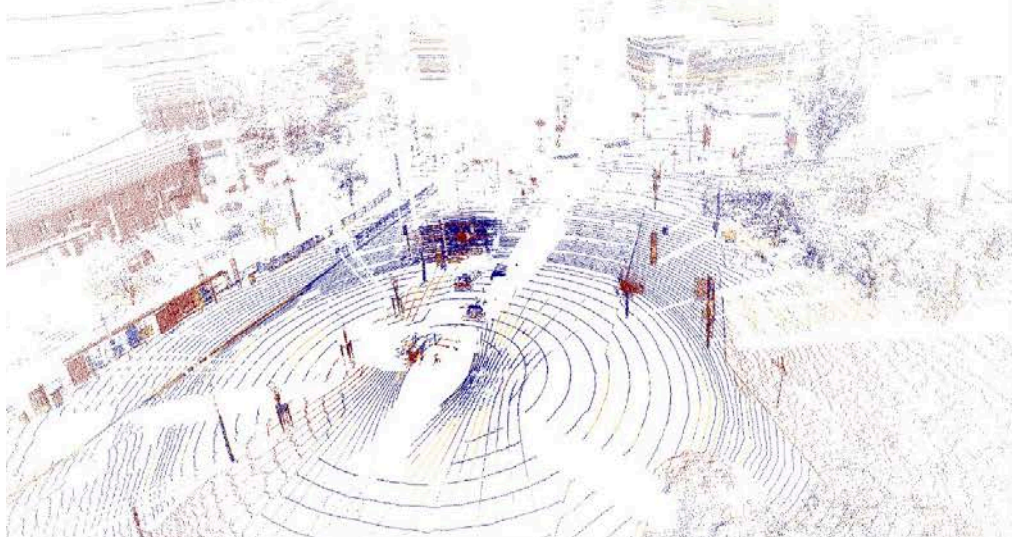
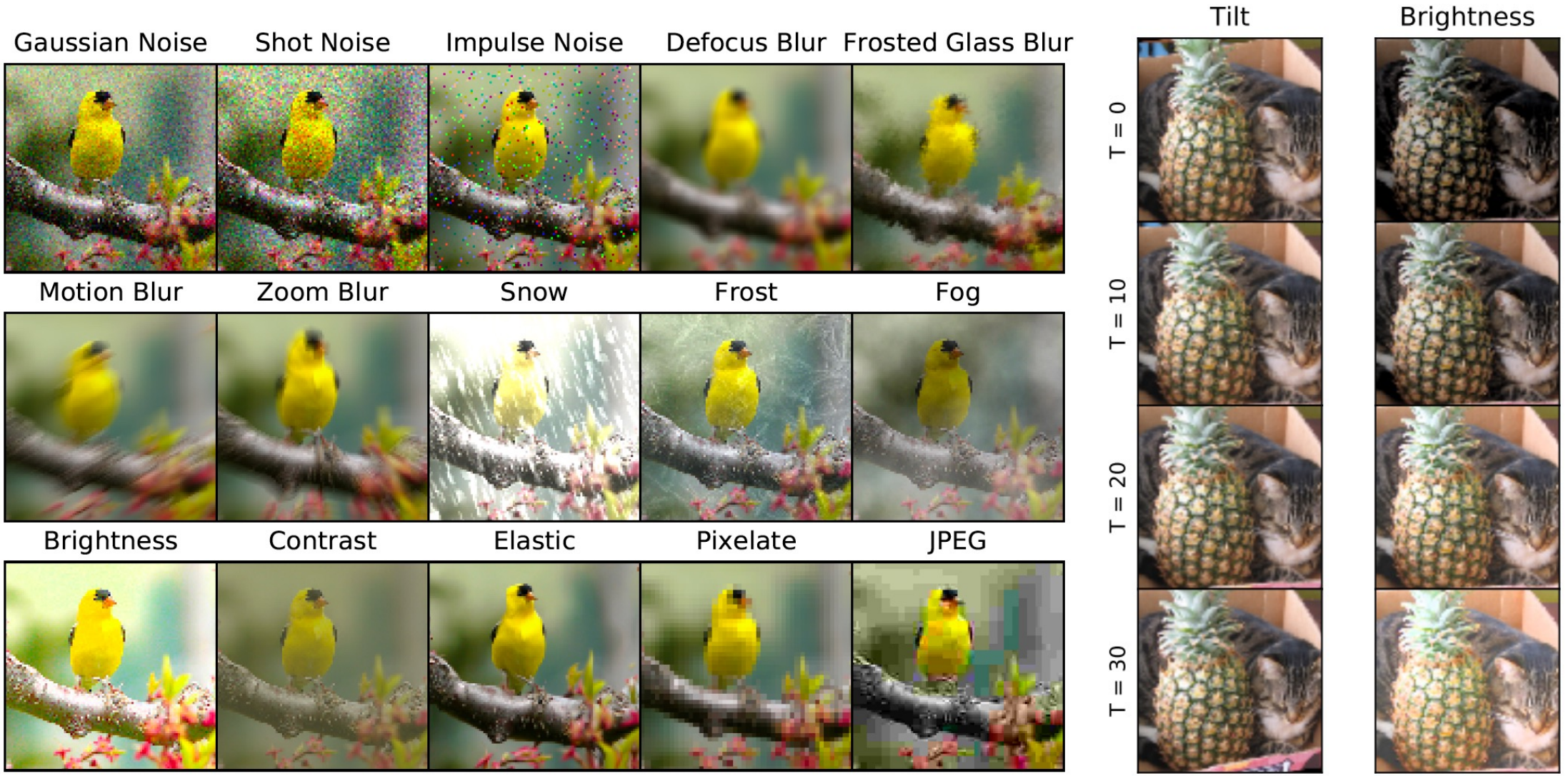


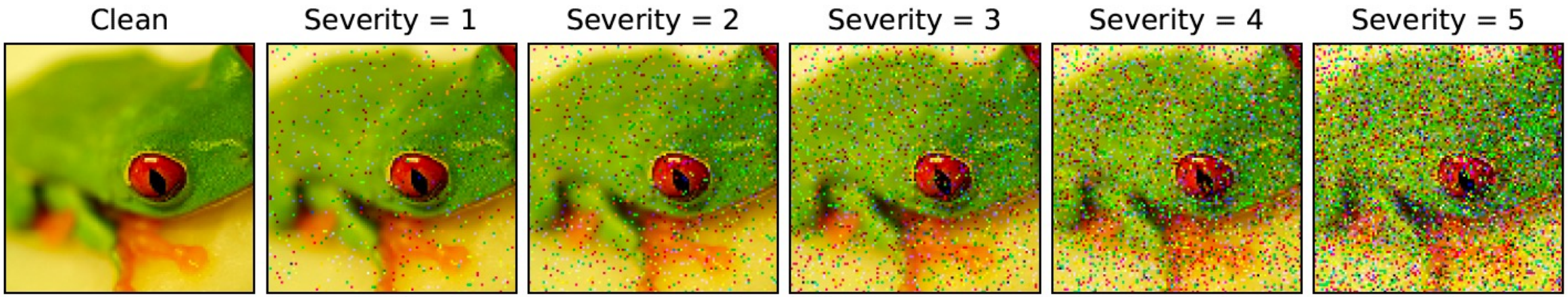
*Image credit: https://zod.zenseact.com

Robustness in RGB Images



D. Hendrycks, et al. "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations," ICLR, 2019

Robustness in RGB Images



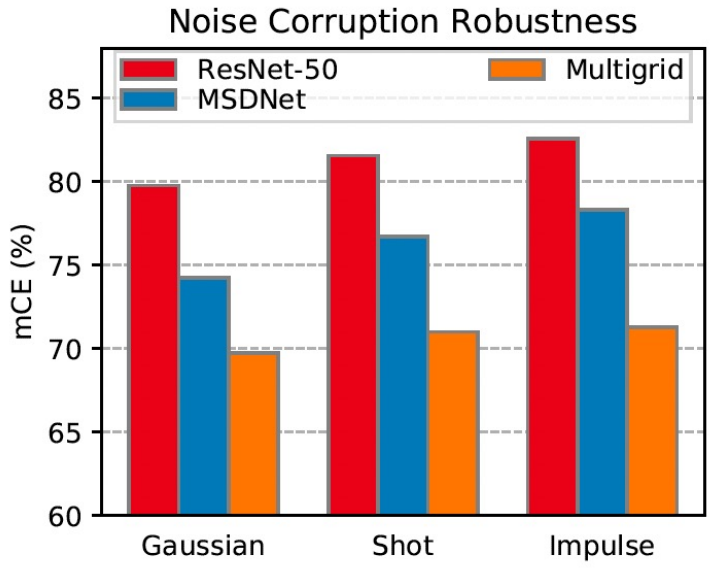
Corruption Error:

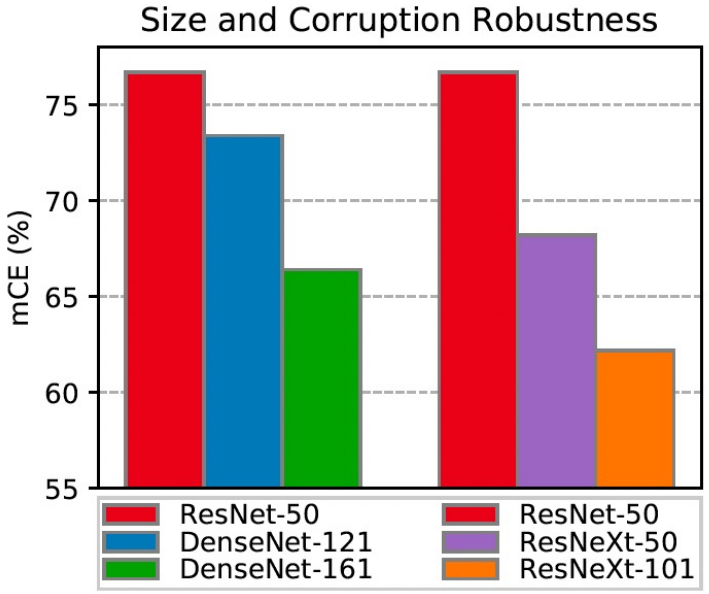
$$\mathrm{CE}_c^f = \left(\sum_{s=1}^{5} E_{s,c}^f\right) \Big/ \left(\sum_{s=1}^{5} E_{s,c}^{\mathrm{AlexNet}}\right)$$

D. Hendrycks, et al. "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations," ICLR, 2019

Robustness in RGB Images

			Noise			Blur				Weather				Digital			
Network	Error	**mCE**	Gauss.	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG
AlexNet	43.5	100.0	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
SqueezeNet	41.8	104.4	107	106	105	100	103	101	100	101	103	97	97	98	106	109	134
VGG-11	31.0	93.5	97	97	100	92	99	93	91	92	91	84	75	86	97	107	100
VGG-19	27.6	88.9	89	91	95	89	98	90	90	89	86	75	68	80	97	102	94
VGG-19+BN	25.8	81.6	82	83	88	82	94	84	86	80	78	69	61	74	94	85	83
ResNet-18	30.2	84.7	87	88	91	84	91	87	89	86	84	78	69	78	90	80	85
ResNet-50	23.9	76.7	80	82	83	75	89	78	80	78	75	66	57	71	85	77	77





D. Hendrycks, et al. "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations," ICLR, 2019

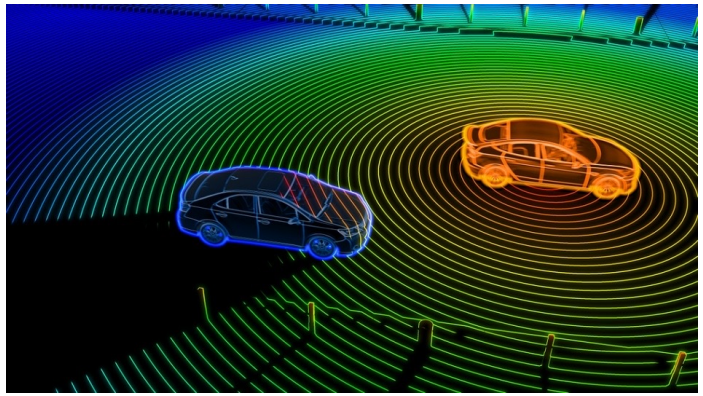
Robustness in Point Clouds

- Data in the point cloud format are used in safety-critical applications, such as autonomous driving and robot navigation.
- However, this data format often suffers from severe Out-of-Distribution (OoD) corruptions in real-world deployment.

Clean

Real-World

Corruptions are severe and OoD
e.g., occlusion, sensory noise

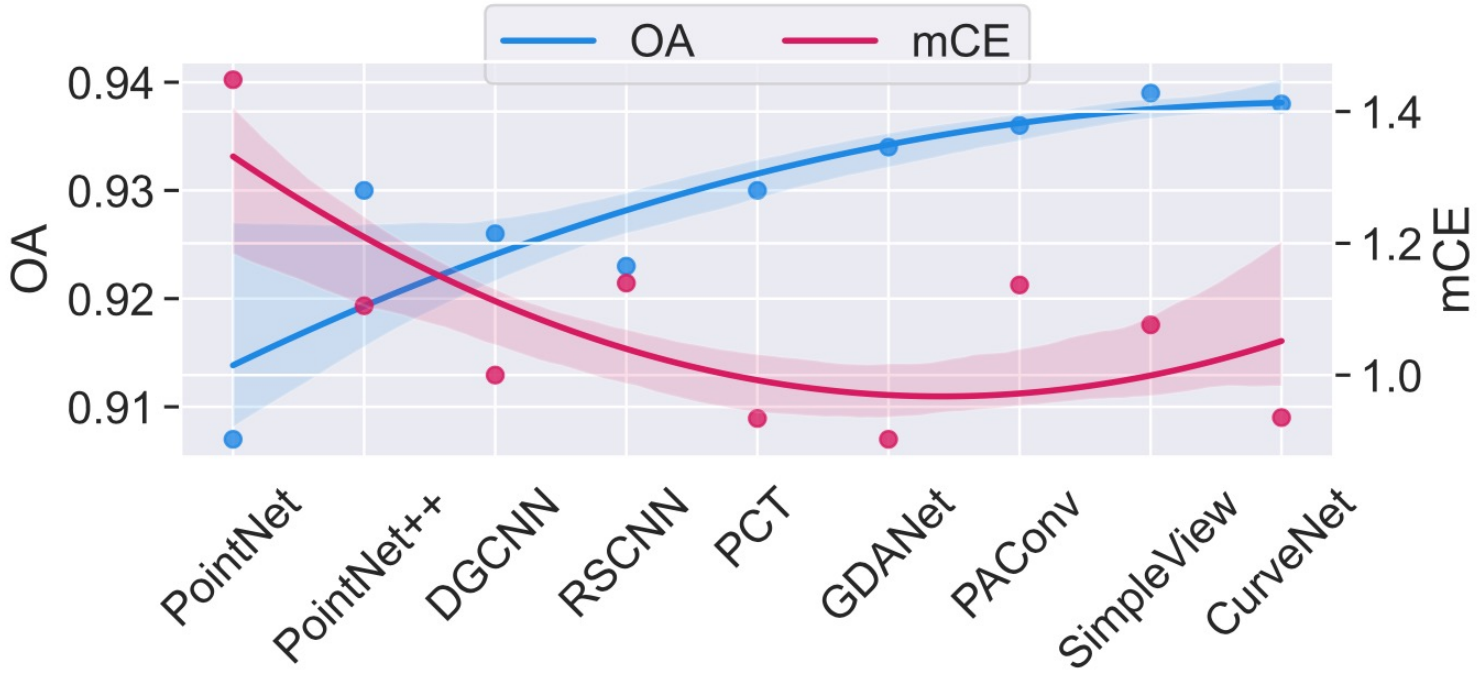


Applications are safety-critical
e.g., autonomous driving

J. Ren, et al. "PointCloud-C: Benchmarking and Analyzing Point Cloud Perception Robustness under Corruptions," *preprint*, 2022

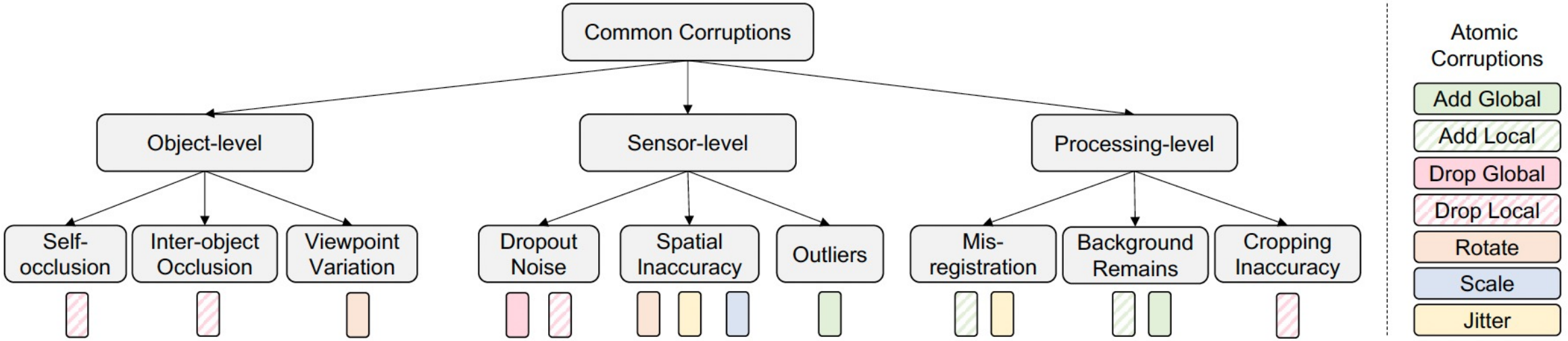
Robustness in Point Clouds

- **Question:** Are point cloud classifiers getting more robust?
- **Answer:** No. Although the accuracy indicator on ModelNet40 gradually saturates, the robustness is at the risk of getting worse, due to the lack of a standard test suite.



J. Ren, et al. "PointCloud-C: Benchmarking and Analyzing Point Cloud Perception Robustness under Corruptions," *preprint*, 2022

Robustness in Point Clouds

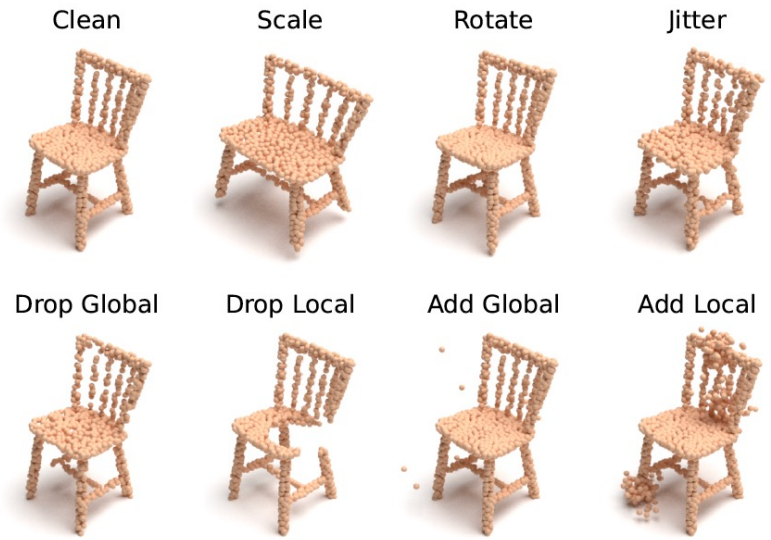


- Three levels of corruption sources: Object, sensor, and processing.
- Nine potential corruption types.
- Simplified into a combination of seven atomic corruptions for a more controllable empirical analysis.

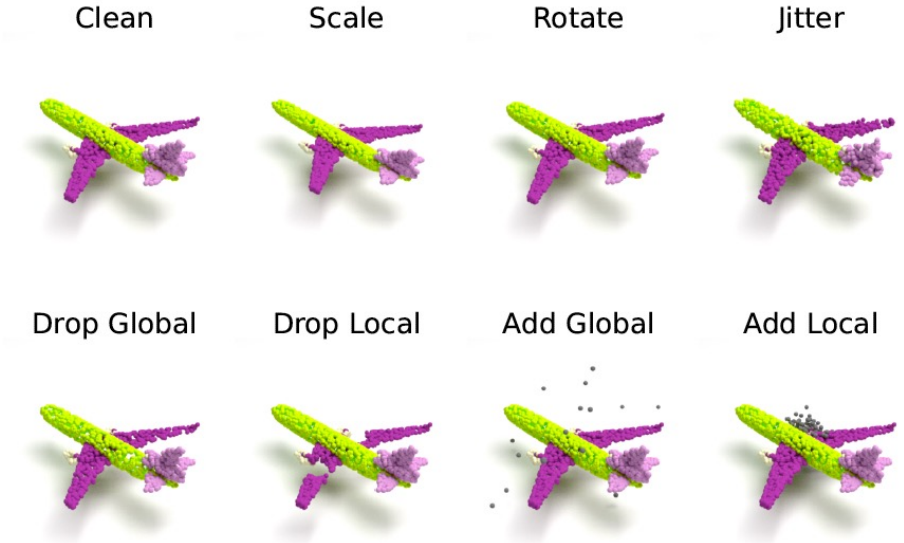
J. Ren, et al. "PointCloud-C: Benchmarking and Analyzing Point Cloud Perception Robustness under Corruptions," *preprint*, 2022

Robustness in Point Clouds

- PointCloud-C is the first competition that targets the robustness of point cloud understanding under corruptions.
- The benchmark result suggests that point cloud classifiers are at the risk of getting less robust, thus highlighting the importance of proposing new designs to improve the robustness.



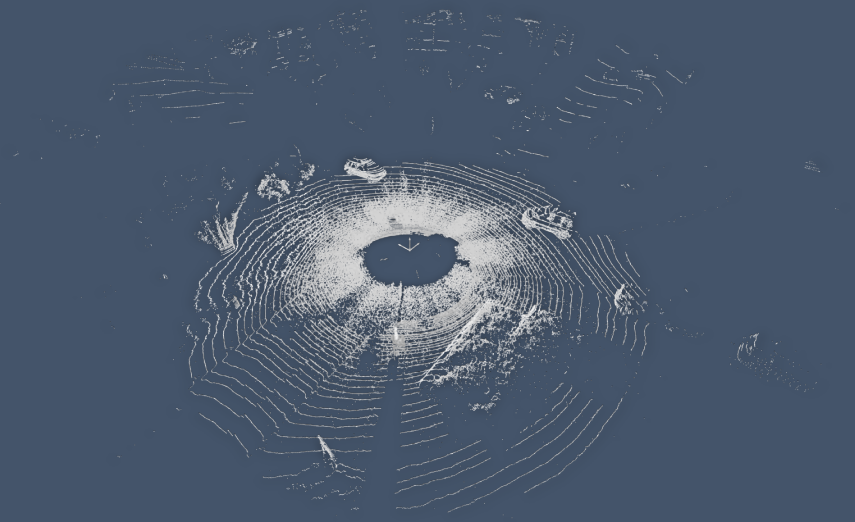
ModelNet-C



ShapeNet-C

J. Ren, et al. "PointCloud-C: Benchmarking and Analyzing Point Cloud Perception Robustness under Corruptions," *preprint*, 2022

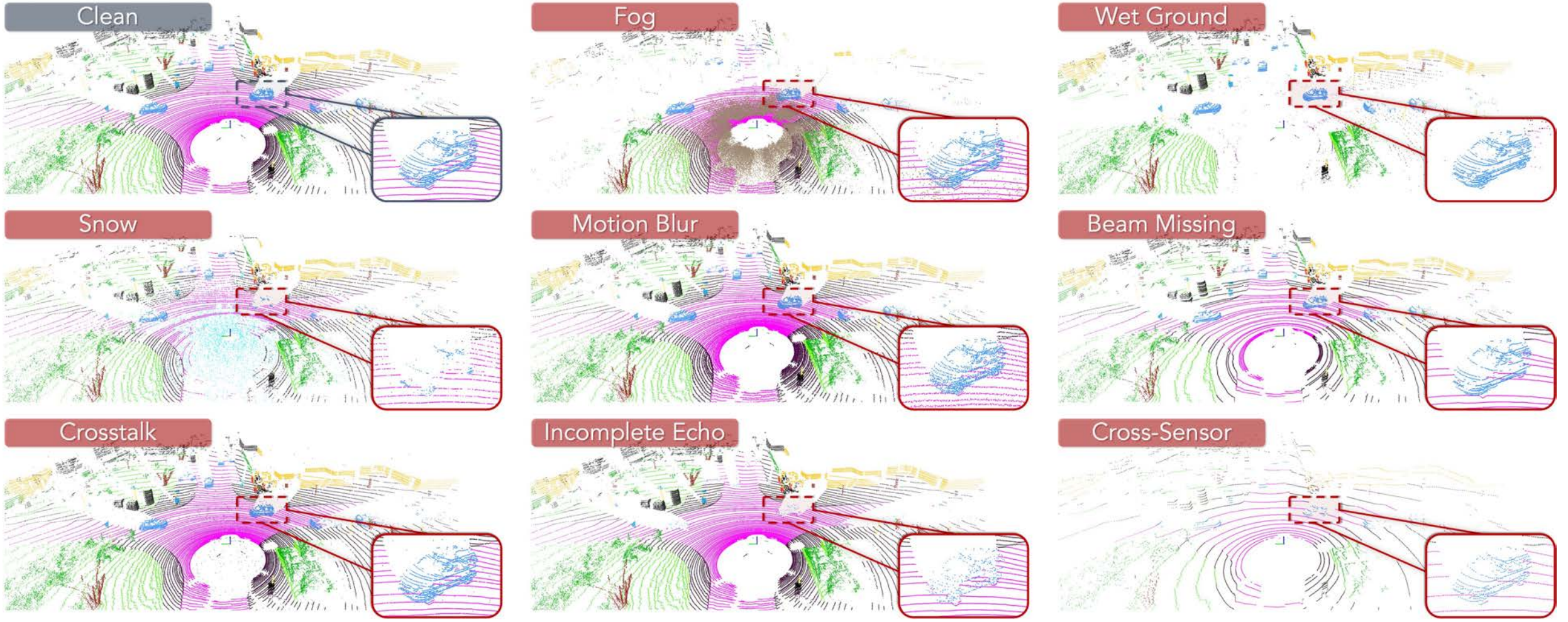
Robo3D



TL;DR

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

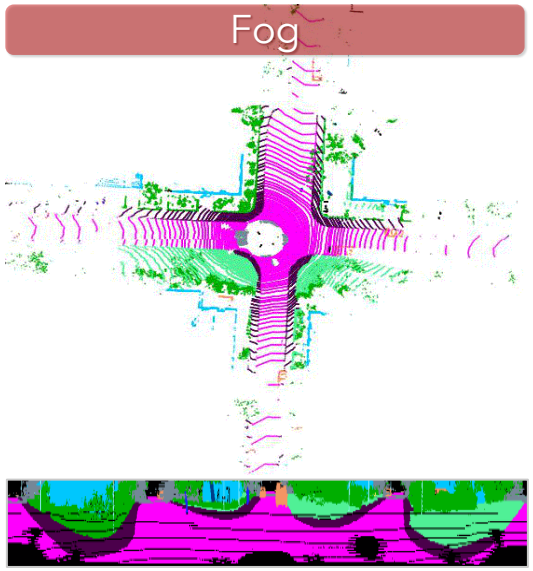
Robo3D: Taxonomy



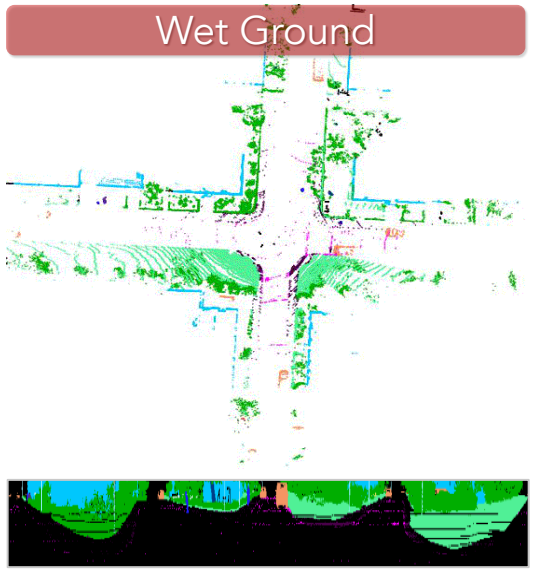
*More examples at: https://ldkong.com/Robo3D

Robo3D: Example

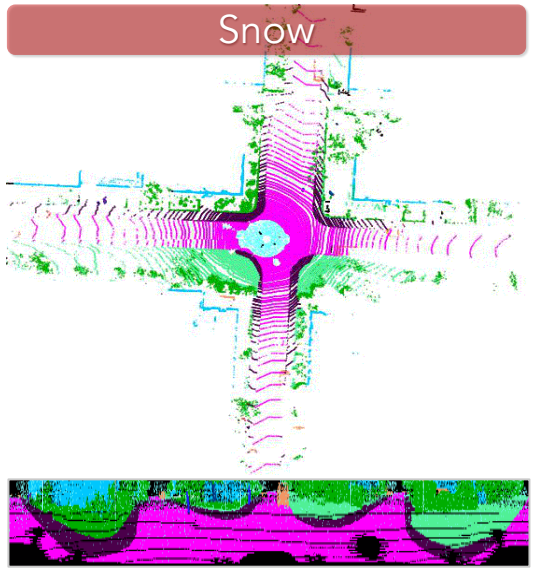
Fog



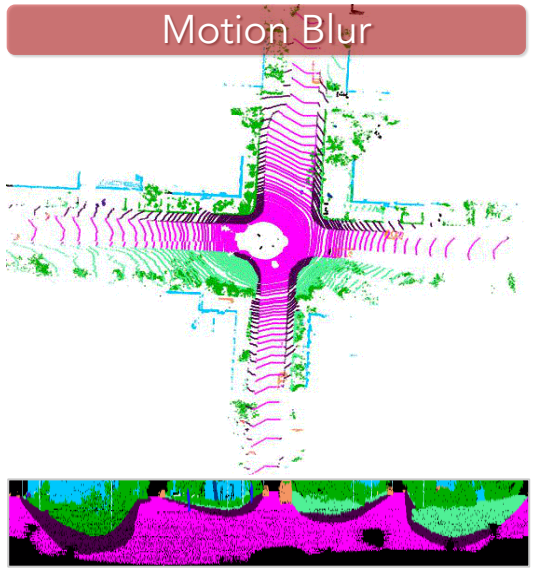
Wet Ground



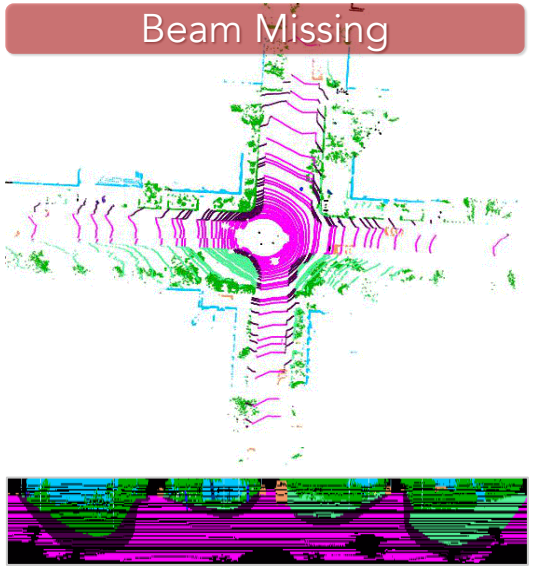
Snow



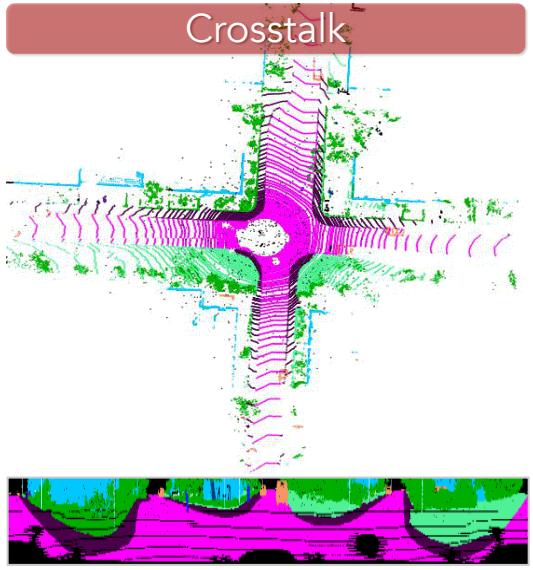
Motion Blur



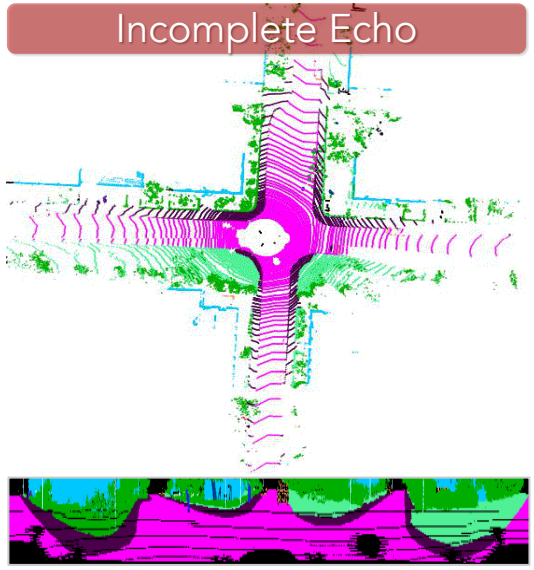
Beam Missing



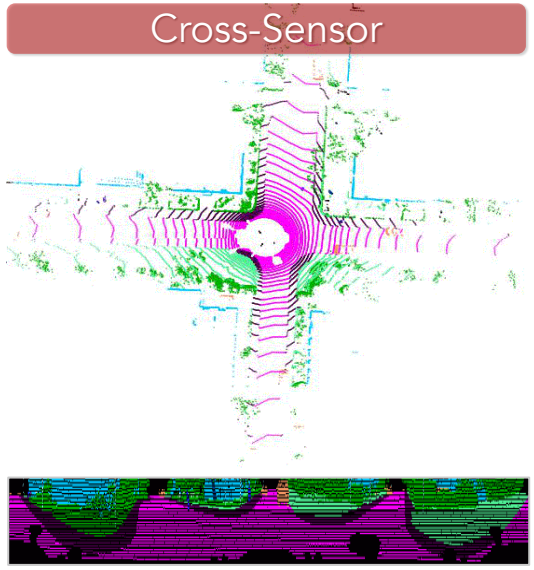
Crosstalk



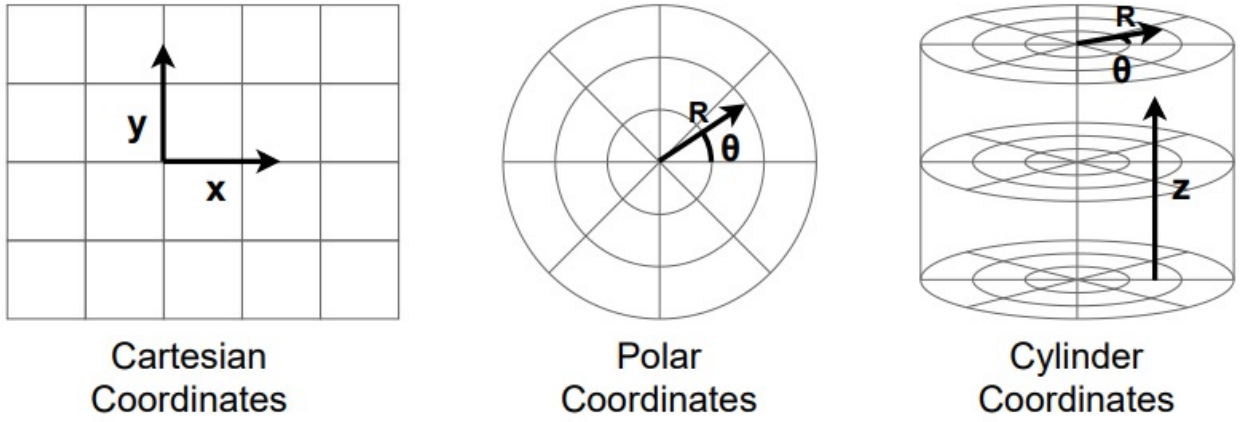
Incomplete Echo



Cross-Sensor

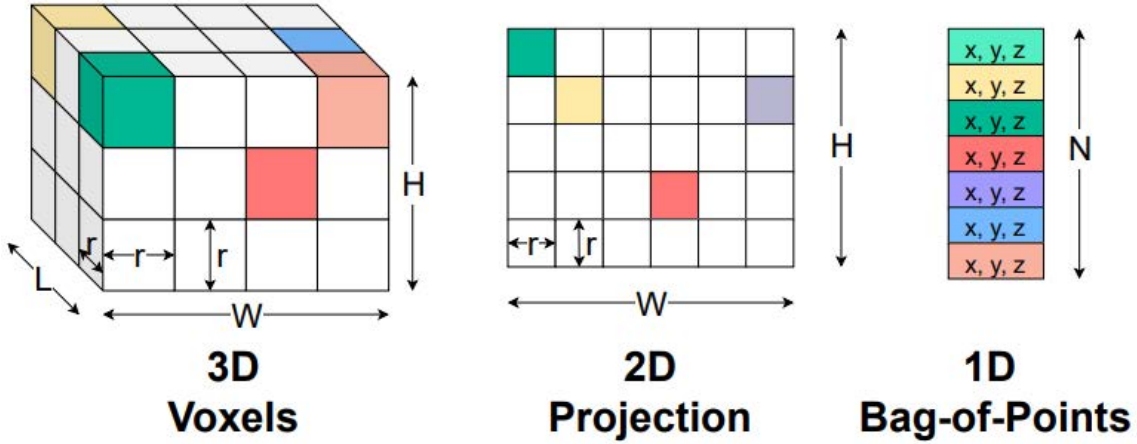


Robo3D: Representation



Representation:

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



Operator:

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

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

Robo3D: Statistics

Corruption Type:

- Include 8 types, each with 3 severity levels

Dataset (6 different sets):

- **LiDAR Semantic Segmentation:** [1]SemanticKITTI-C, [2]nuScenes-C (Seg3D), [3]WOD-C (Seg3D)
- **3D Object Detection:** [4]KITTI-C, [5]nuScenes-C (Det3D), [6]WOD-C (Det3D)

Model & Algorithm (34 perception models):

- **LiDAR Semantic Segmentation:** 22 segmentors
- **3D Object Detection:** 12 detectors
- **Data Augmentation:** 3 augmentation techniques

Robo3D: Metrics

Task-Specific Accuracy (Acc):

- **LiDAR Semantic Segmentation:** mean IoU (**mIoU**)
- **3D Object Detection:** mean AP (**mAP**), nuScenes Detection Score (**NDS**)

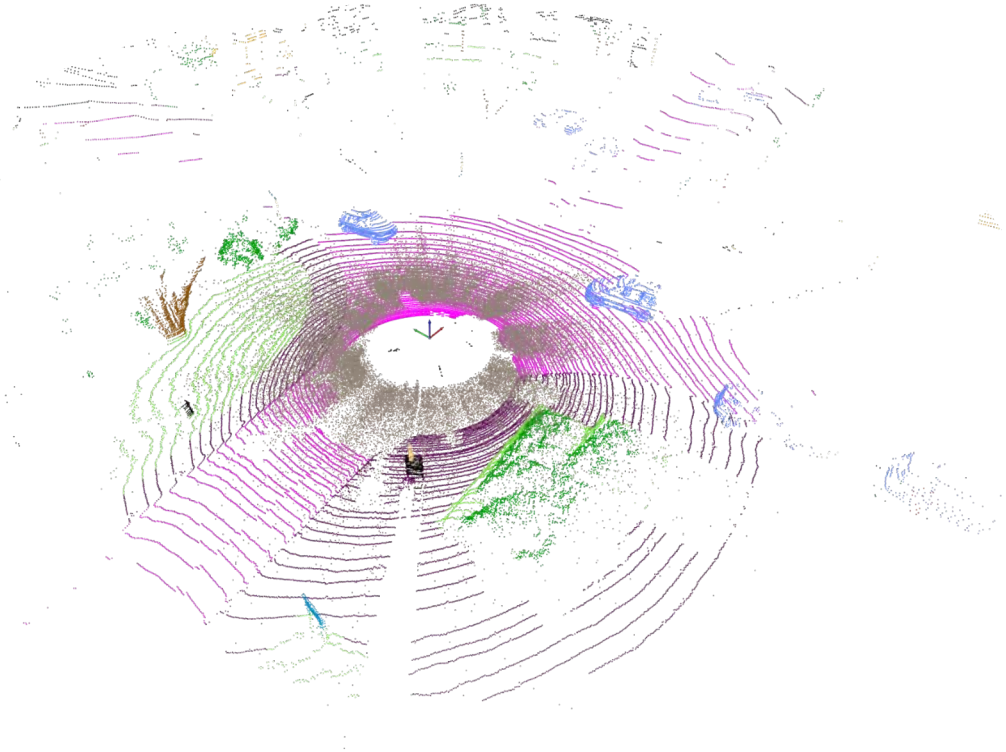
Robustness Metrics:

- **Mean Corruption Error (mCE):**

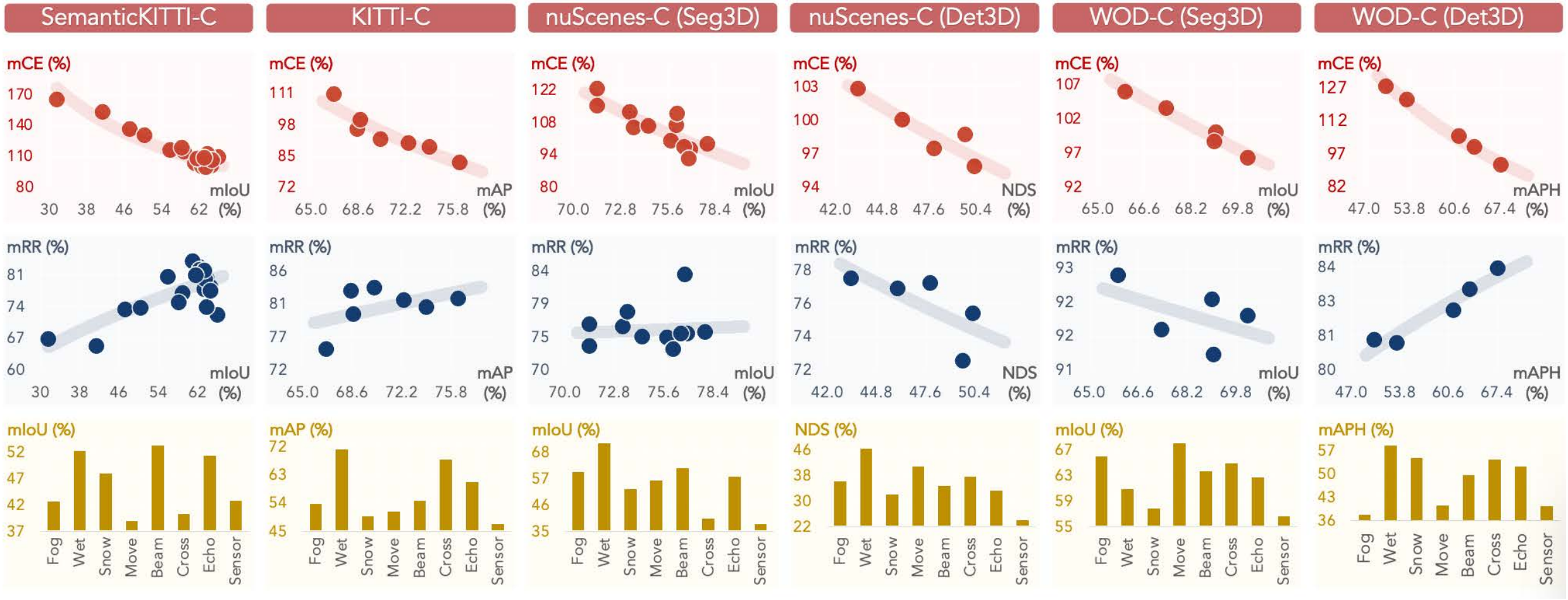
$$\mathrm{CE}_i = \frac{\sum_{l=1}^{3}(1-\mathrm{Acc}_{i,l})}{\sum_{l=1}^{3}(1-\mathrm{Acc}_{i,l}^{\mathrm{baseline}})}, \quad \mathrm{mCE} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{CE}_i$$

- **Mean Resilience Rate (mRR):**

$$\mathrm{RR}_i = \frac{\sum_{l=1}^{3}\mathrm{Acc}_{i,l}}{3 \times \mathrm{Acc}_{\mathrm{clean}}}, \quad \mathrm{mRR} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{RR}_i$$



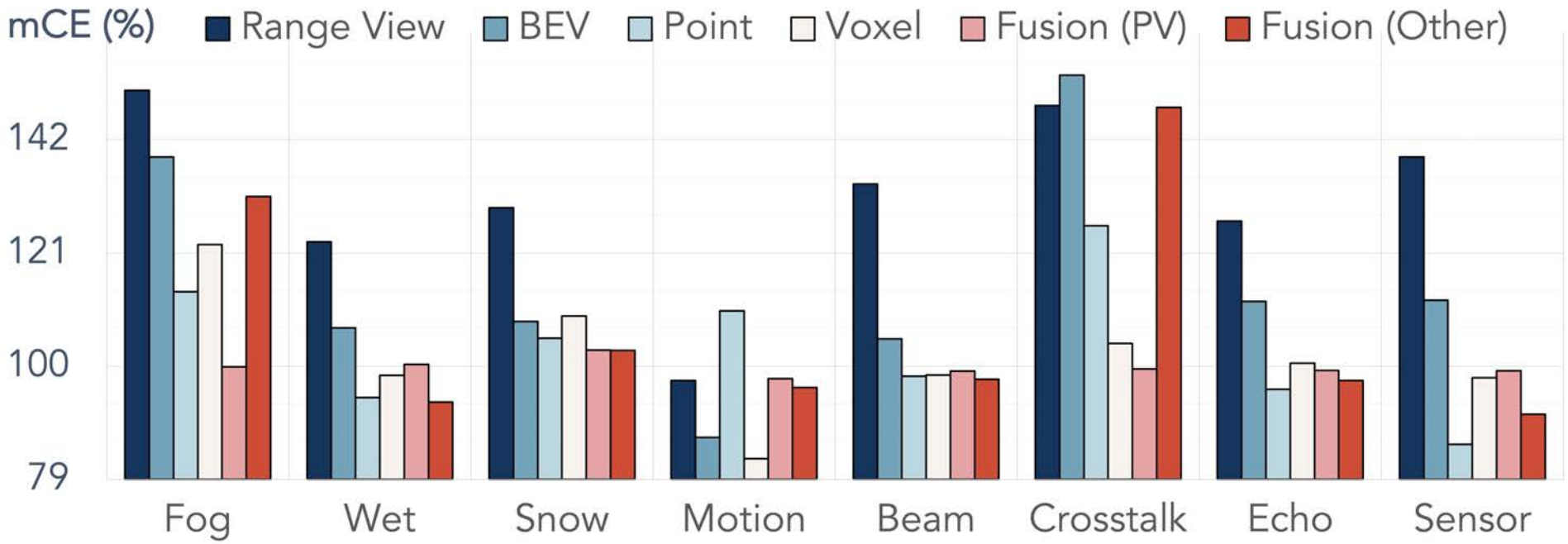
Robo3D: Benchmarking Result



*More results and analysis at: https://github.com/ldkong1205/Robo3D

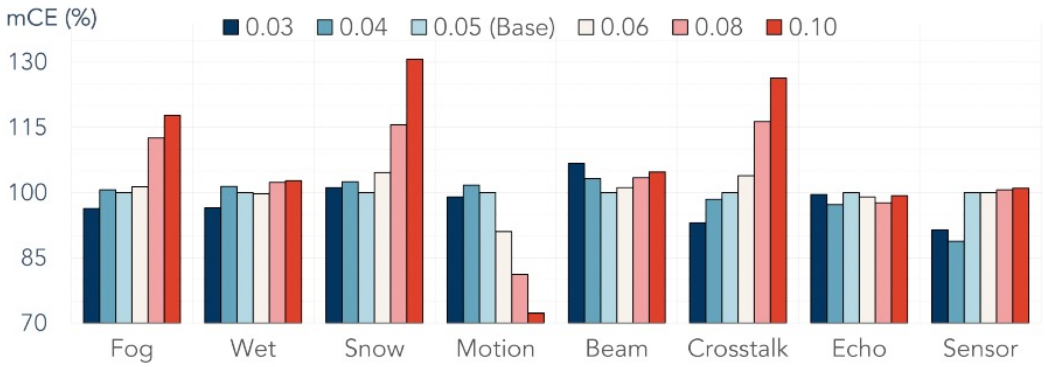
Robo3D: Key Observation

1. Existing 3D detectors and segmentors are **vulnerable** to real-world corruptions.
2. Models trained with LiDAR data from different sources (sensor setups) exhibit **inconsistent sensitivities** to each corruption type.
3. Representing the LiDAR data as raw **points**, sparse **voxel**, or the **fusion** of them tend to yield better robustness.

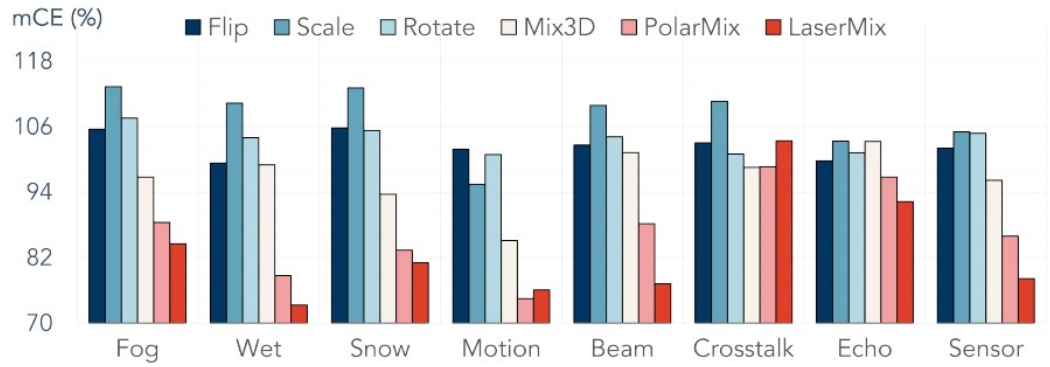


Robo3D: Key Observation

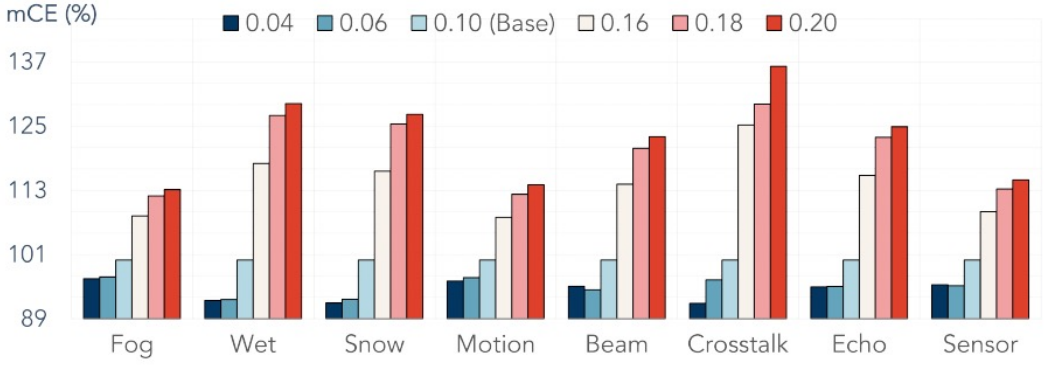
4. The 3D detectors and segmentors show different **sensitivities** to corruption scenarios.
5. The recent **out-of-context augmentation techniques** improve 3D robustness by large margins; the flexible rasterization strategies help learn more robust features.



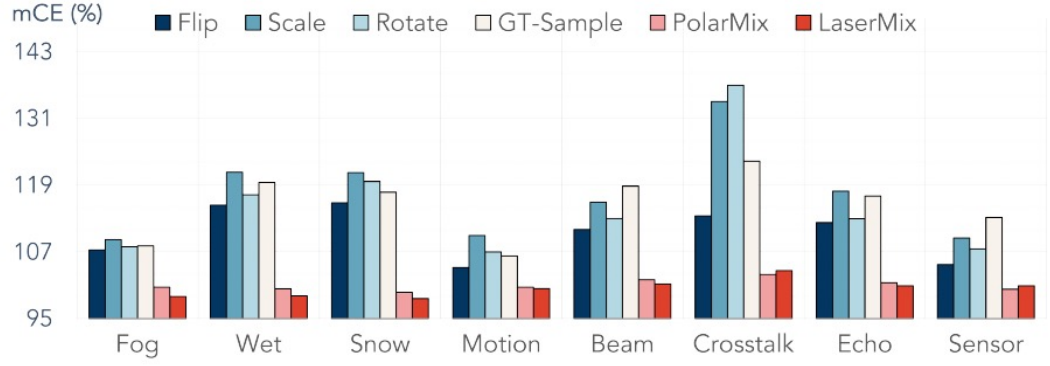
(a) Voxel Size on *SemanticKITTI-C* (Seg3D)



(b) Augmentation on *SemanticKITTI-C* (Seg3D)

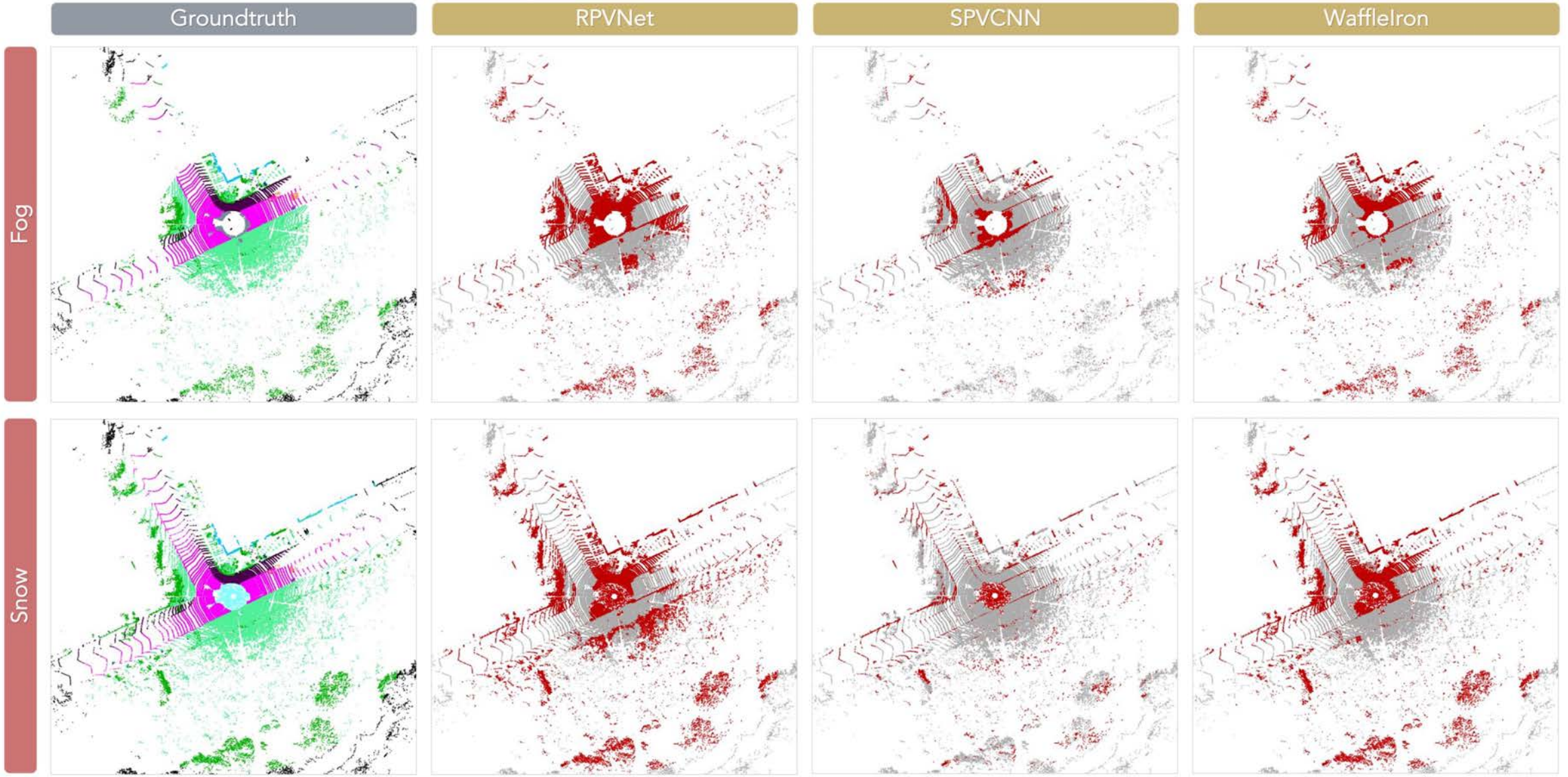


(c) Voxel Size on *WOD-C* (Det3D)



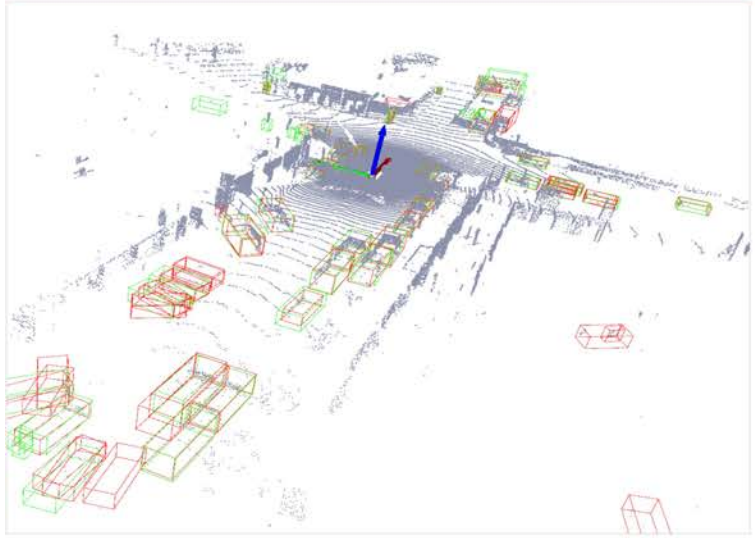
(d) Augmentation on *WOD-C* (Det3D)

Robo3D: Qualitative Assessment

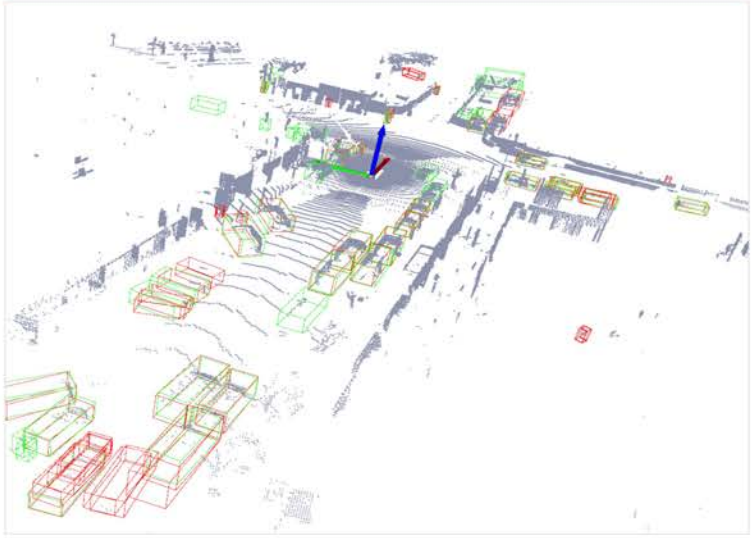


Robo3D: Qualitative Assessment

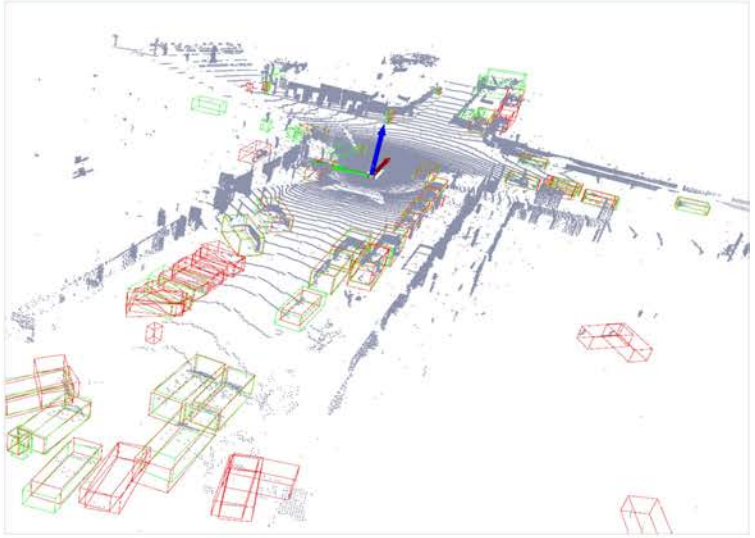
Fog



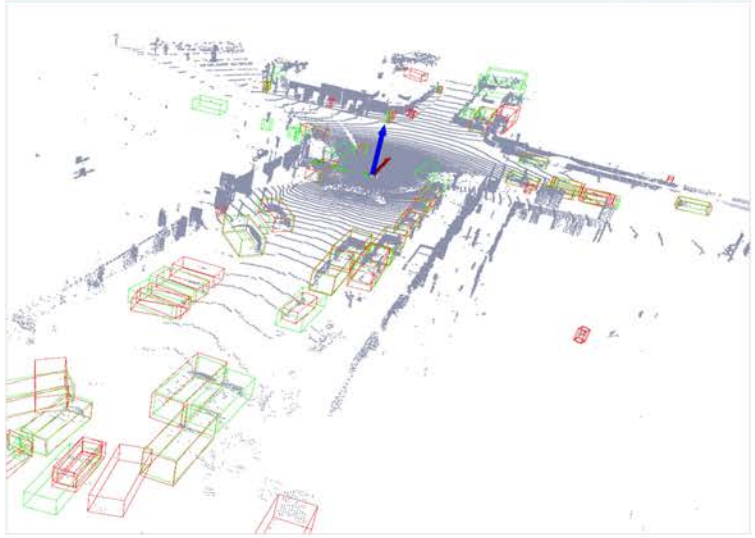
Wet Ground



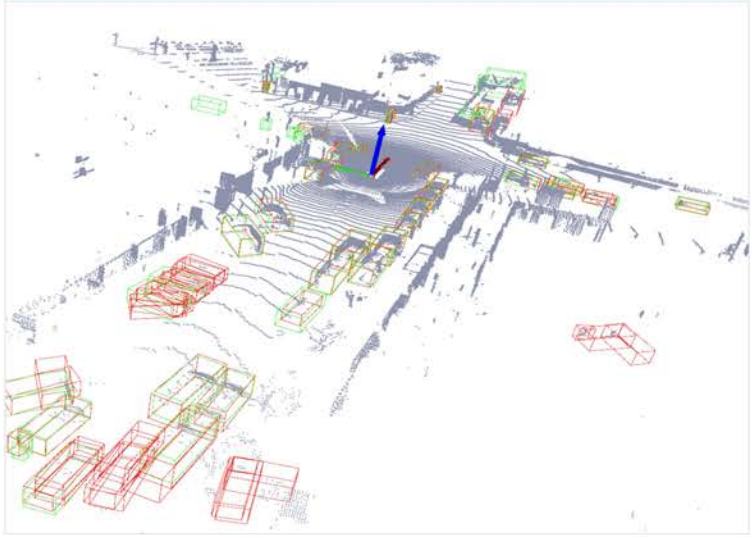
Crosstalk



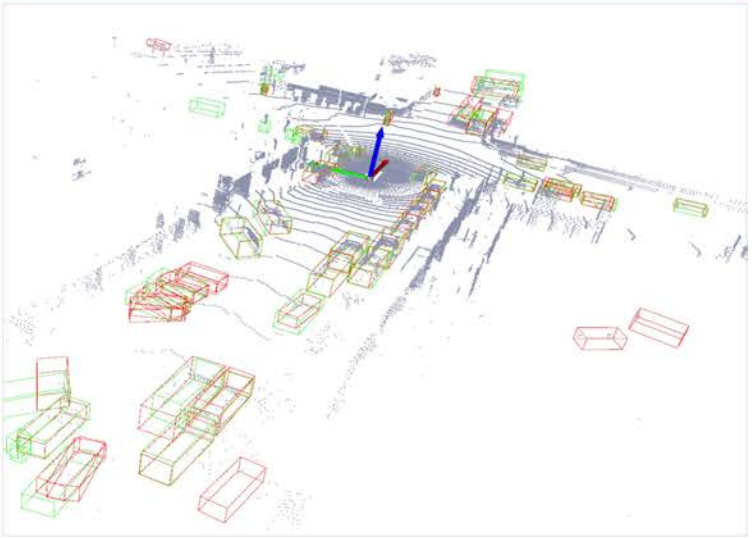
Snow



Motion Blur



Cross-Sensor



Thank you for your attention!

Robo3D

Paper

Code

MMDet3D