NEURAL INFORMATION PROCESSING SYSTEMS

AD-PT: Autonomous Driving Pre-Training with Large-scale Point Cloud Dataset

Jiakang Yuan[1], Bo Zhang[2], Xiangchao Yan[2], Tao Chen[1], Botian Shi[2], Yikang Li[2], Yu Qiao[2]

[1] School of Information Science and Technology, Fudan University

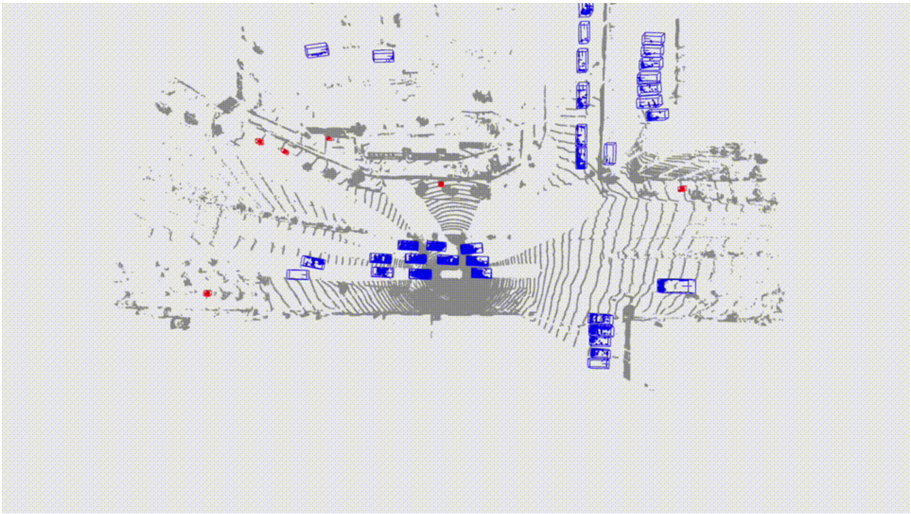
[2] Shanghai Artificial Intelligence Laboratory

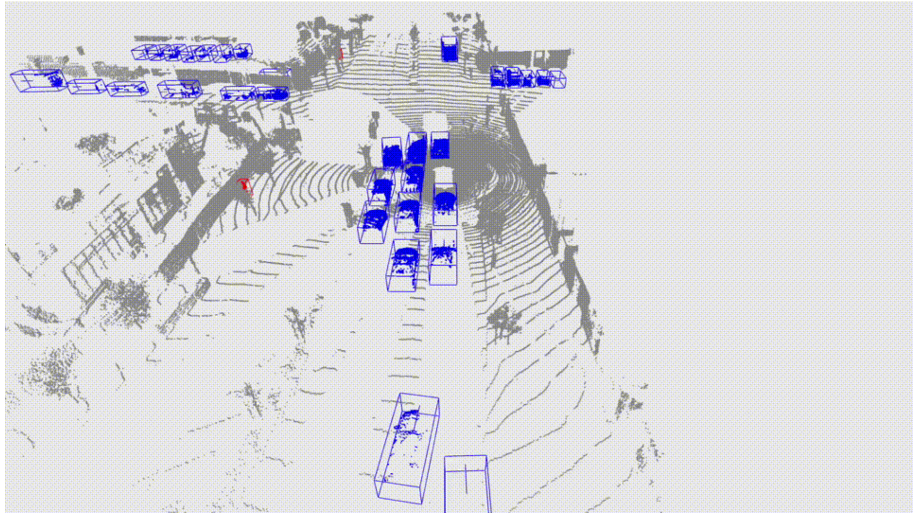
CONTENTS

CONTENTS

Review of Autonomous Driving-related Pre-training

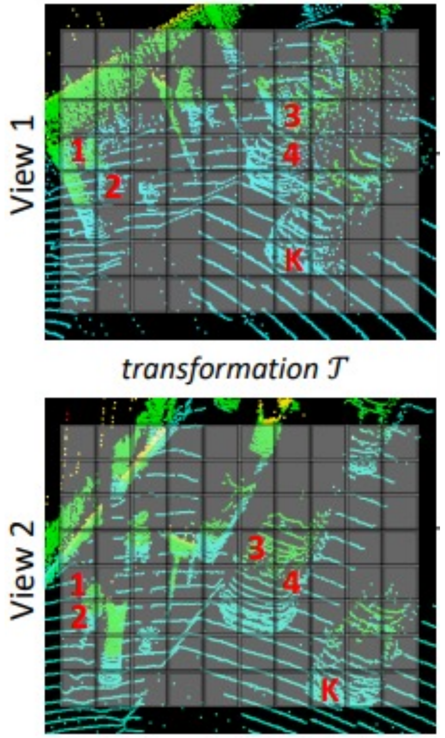
- The success of LiDAR-based 3D detectors depends on a large amount of point cloud data with accurate annotation
- Point cloud annotation is very difficult due to problems such as point cloud sparsity and occlusion.
- Unlabeled data is easy to obtain.
- Pre-training: make full use of the information in unlabeled data





Review of Autonomous Driving-related Pre-training

- Contrastive-learning-based methods
  - Using corresponding points of different views as positive pairs

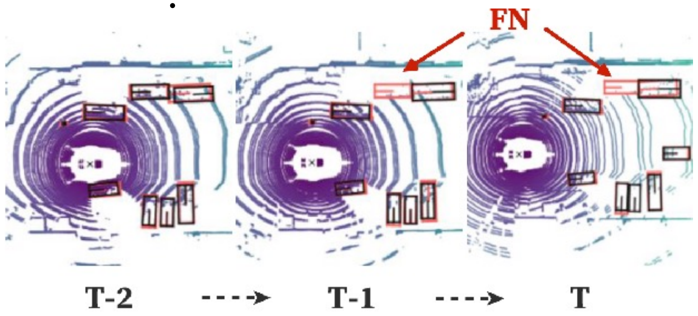


- *Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In: ECCV (2020)*
- *Exploring geometry-aware contrast and clustering harmonization for self-supervised 3d object detection. In: ICCV (2021)*
- *Proposalcontrast: Unsupervised pre-training for lidar-based 3d object detection. In: ECCV (2022)*

Review of Autonomous Driving-related Pre-training

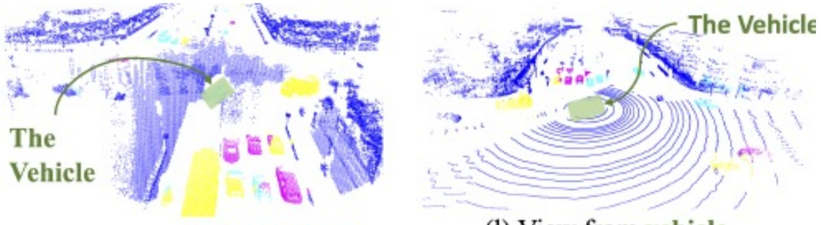
➢ Contrastive-learning-based methods

◆ Using corresponding points of different frames as positive



- *Spatio-temporal self-supervised representation learning for 3d point clouds. In ICCV (2021)*

◆ Using LiDAR point clouds from the vehicle- and infrastructure-side as positive pairs

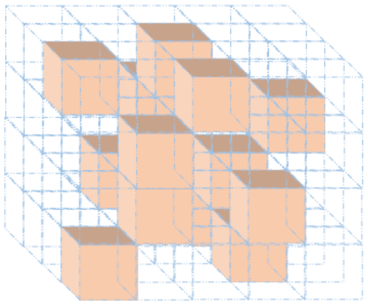


(k) View from infrastructure (l) View from vehicle

- *CO3: Cooperative unsupervised 3d representation learning for autonomous driving. In ICLR (2023)*

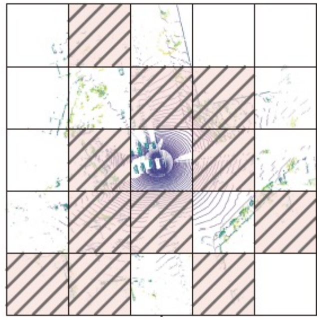
Review of Autonomous Driving-related Pre-training

➢ MAE-based methods



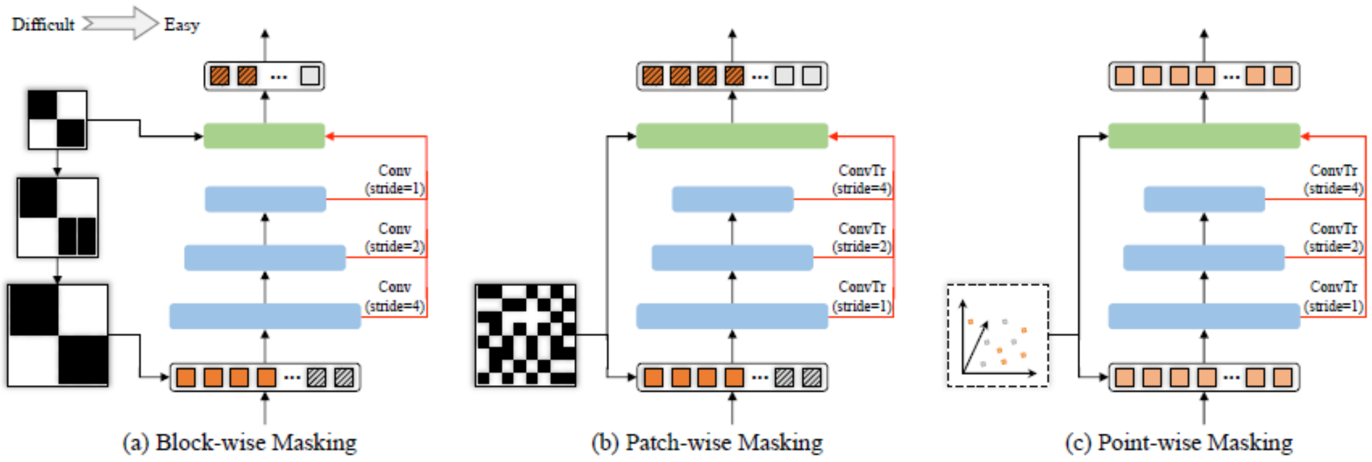
- Voxel space

- *Voxel-mae: Masked autoencoders for pre-training large-scale point clouds.*



- BEV space

- *BEV-MAE: Bird's Eye View Masked Auto-encoders for Outdoor Point Cloud Pre-training.*

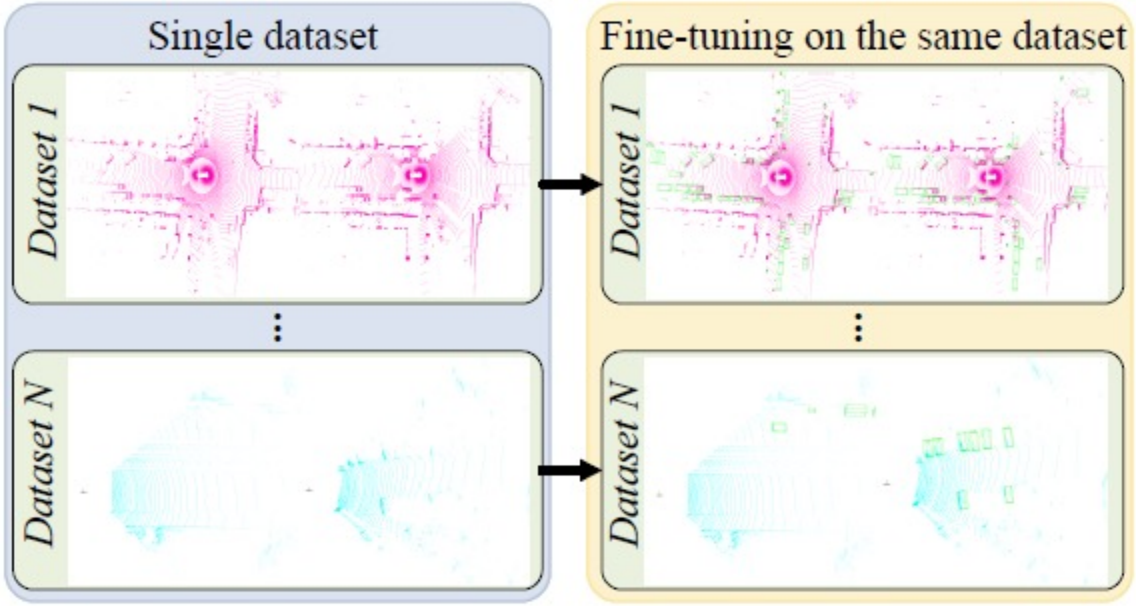


- Hierarchical space

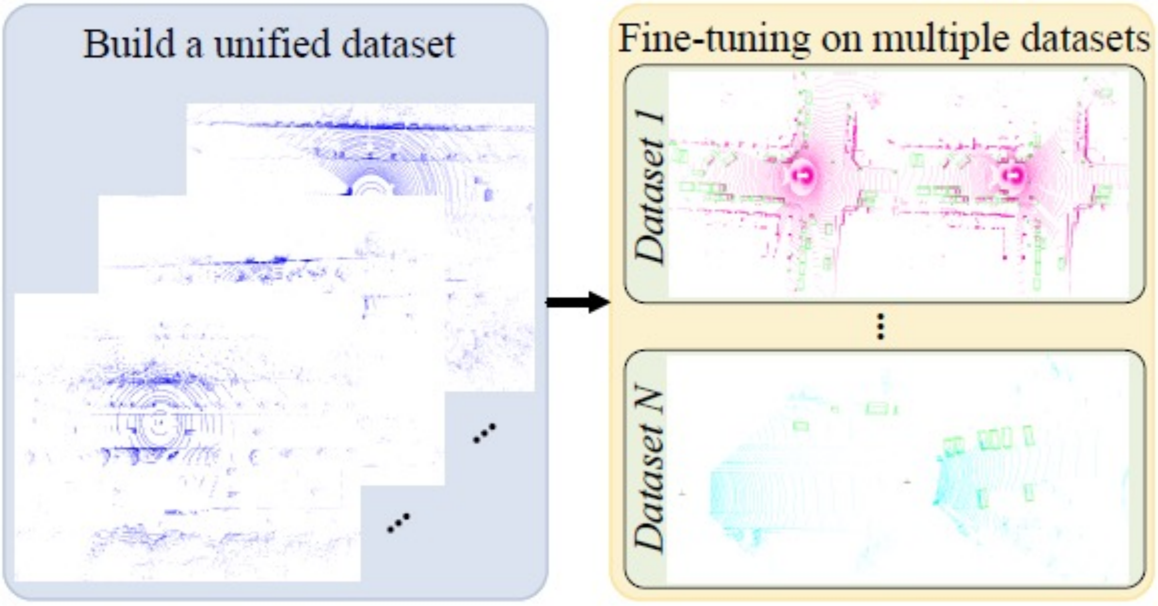
- *GD-MAE: generative decoder for MAE pre-training on lidar point clouds. In CVPR (2023).*

Review of Autonomous Driving-related Pre-training

➢ Previous methods



➢ AD-PT

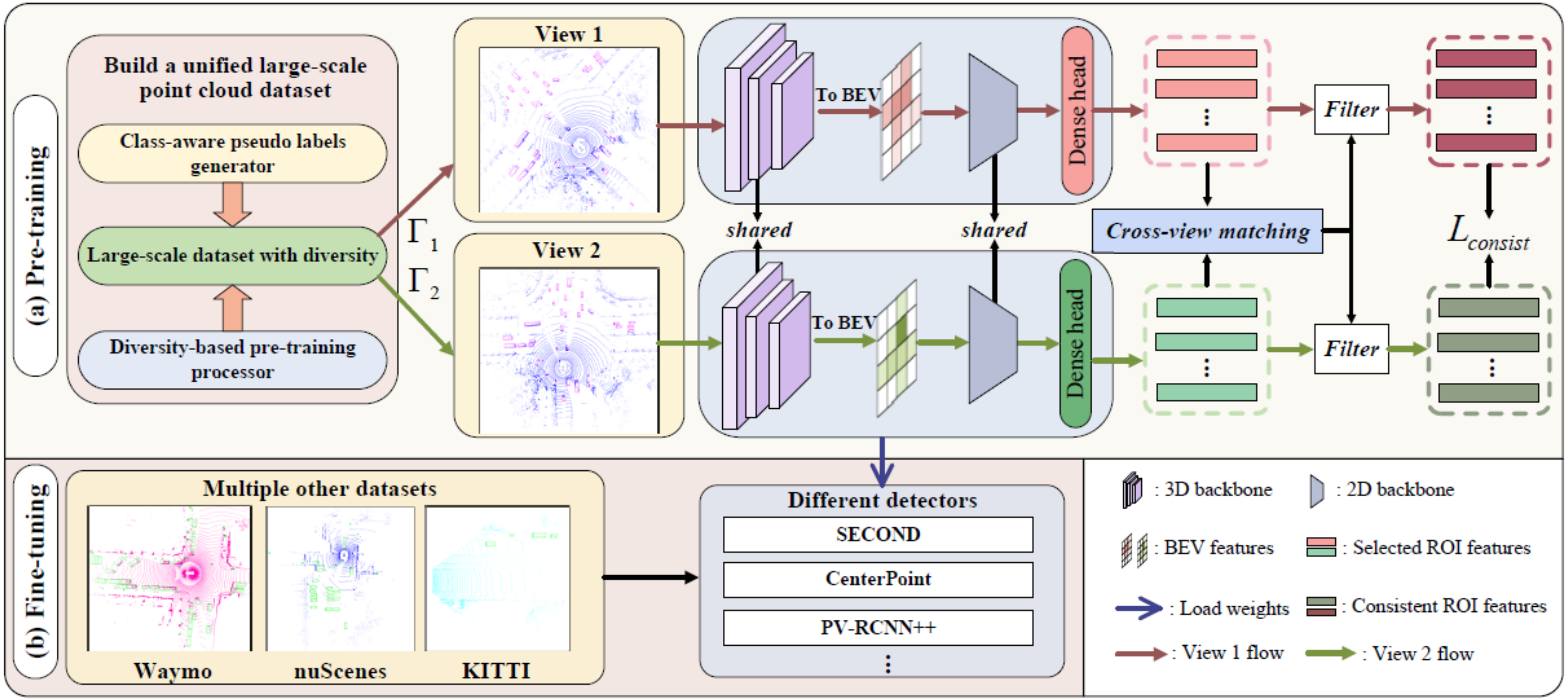


- Pre-training and fine-tuning data are sampled from the same single dataset
- Better generalized performance on different datasets

CONTENTS

Method: AD-PT

➢ Overall Framework



Method: AD-PT

➢ Large-scale Point Cloud Dataset Preparation

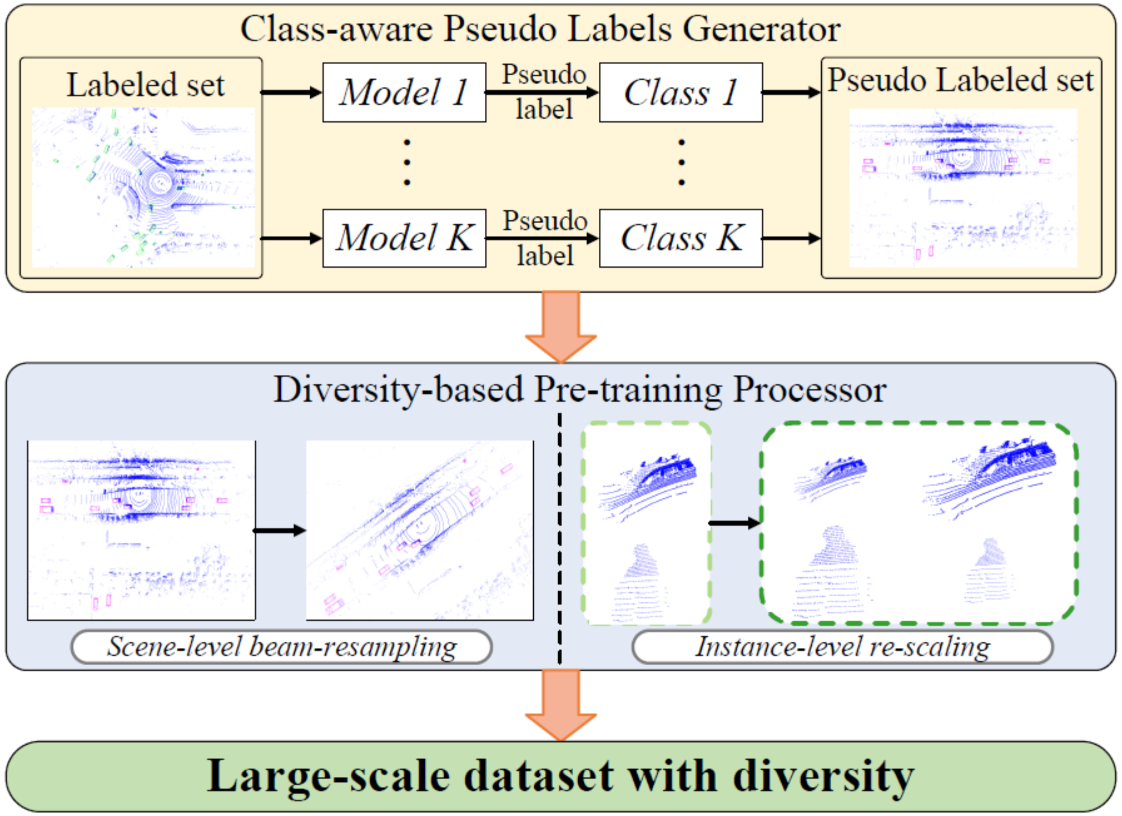
- Performs large-scale point cloud pre-training in a semi-supervised manner
- ONCE Dataset: ~5k *vs.* ~1M (labeled data *vs.* unlabeled data)
- Pseudo-labels with high accuracy on the pre-training dataset are beneficial to enhance the detection accuracy on downstream datasets

Pseudo-labeling Method	ONCE	Waymo L2 AP/APH				nuScenes	
	Overall	Overall	Vehicle	Pedestrian	Cyclist	mAP	NDS
SECOND (Low Performance)	57.10	65.96 / 63.29	65.95 / 65.46	66.87 / 60.36	65.07 / 64.06	41.49	50.82
CenterPoint (Middle Performance)	60.84	66.79 / 64.10	67.09 / 66.60	67.79 / 61.16	65.51 / 64.55	41.91	51.64
Ours (High Performance)	69.90	**67.77 / 65.09**	**68.01 / 67.61**	**68.32 / 61.69**	**66.99 / 65.98**	**43.11**	**52.41**

Mao J, Niu M, Jiang C, et al. One million scenes for autonomous driving: Once dataset[J]. arXiv preprint arXiv:2106.11037, 2021.

Method: AD-PT

- Large-scale Point Cloud Dataset Preparation



- Class-aware pseudo labels generator
  - Class-aware Pseudo Labeling

    Evaluate on ONCE validation set

Detector	Head Choice	Vehicle	Pedestrian	Cyclist
ONCE Benchmark (Best)	Center Head	66.79	49.90	63.45
CenterPoint (ours)	Center Head	-	**56.01**	-
PV-RCNN++ (ours)	Anchor Head	**82.50**	-	**71.19**

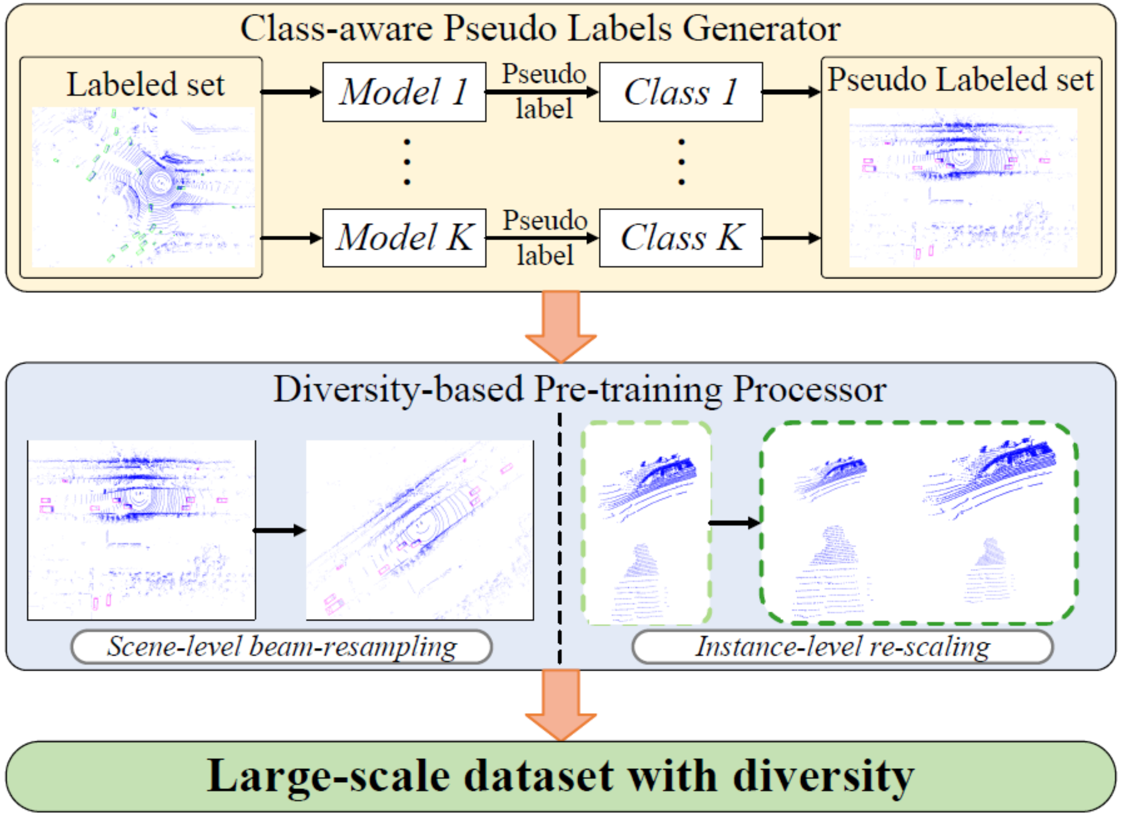
  - Semi-supervised Data Labeling

    Further improve the accuracy

https://once-for-auto-driving.github.io/benchmark.html

Method: AD-PT

➢ Large-scale Point Cloud Dataset Preparation



- Diversity-based Pre-training Processor

  Highly diverse data can greatly improve the generalization ability of the model

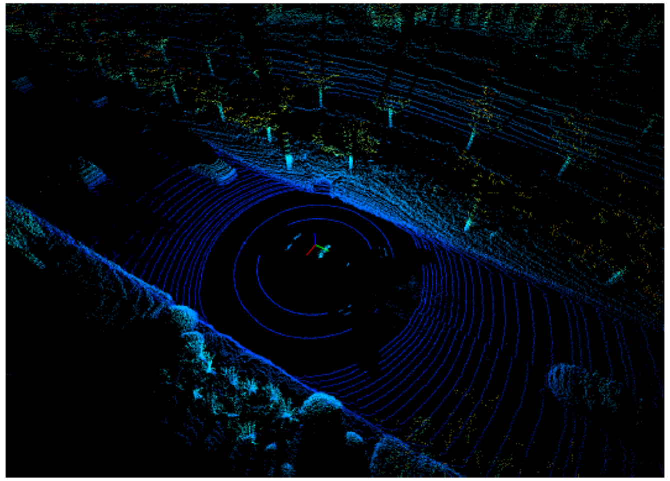
  - ◆ Data with More Beam-Diversity
  - Range image as an intermediate variable for point data up-sampling and downsampling

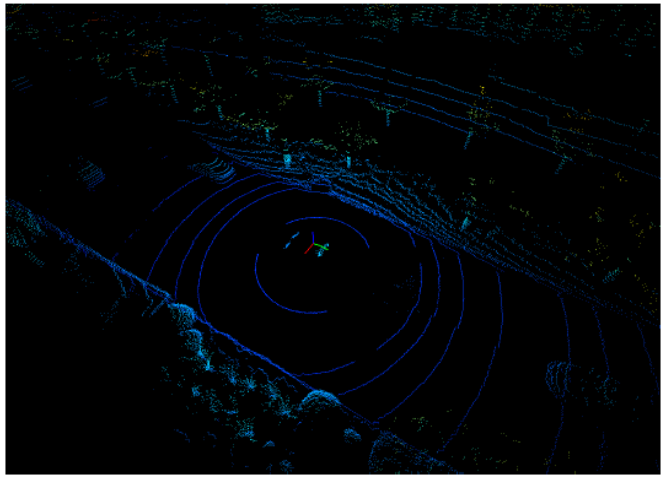
  - ◆ Data with More RoI-Diversity
  - Randomly re-scale the length, width and height of each object

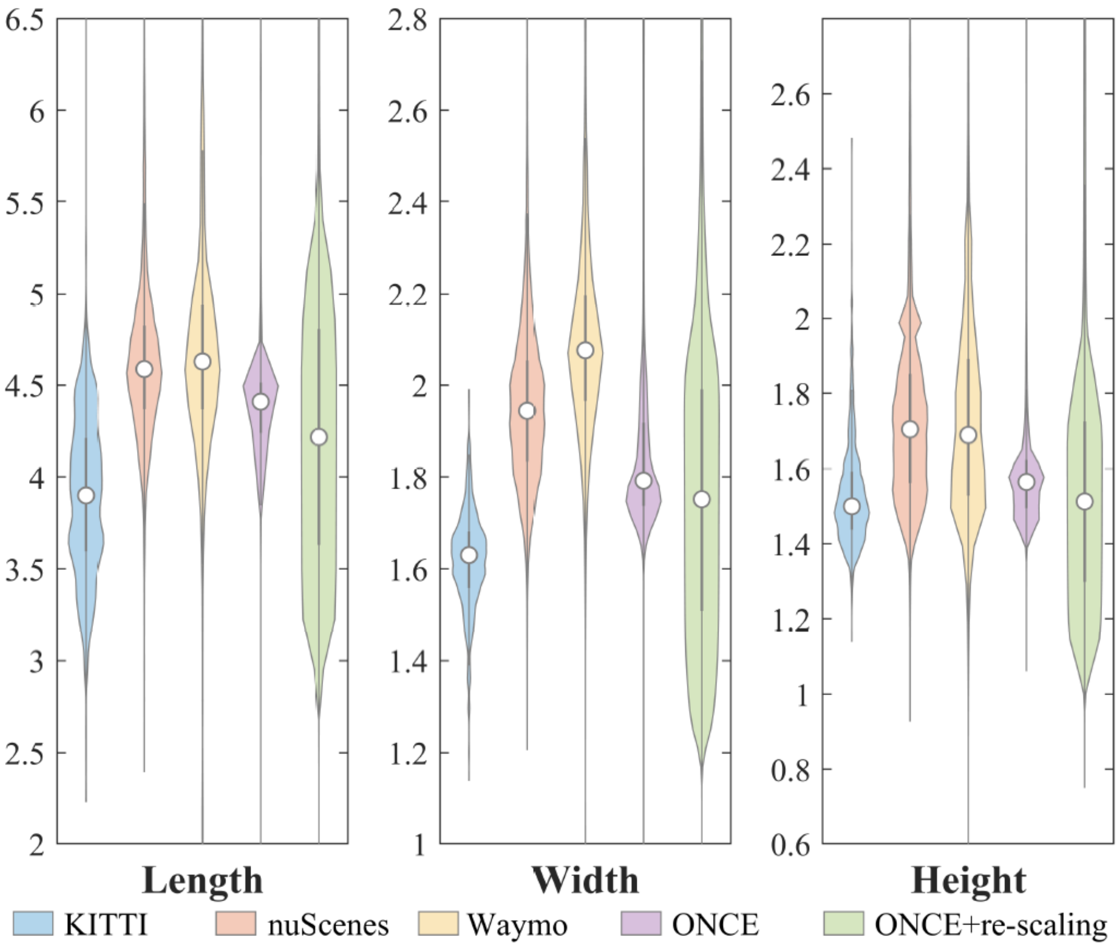
https://once-for-auto-driving.github.io/benchmark.html

Method: AD-PT

➢ Large-scale Point Cloud Dataset Preparation







CONTENTS

Method: AD-PT

- Learning Unified Representations under Large-scale Point Cloud Dataset
  - Taxonomy difference

Dataset	classes
ONCE (Pre-train)	Car, Truck, Bus, Pedestrian, Cyclist
Waymo (Fine-tune)	Vehicle, Pedestrian, Cyclist
nuScenes (Fine-tune)	Car, Truck, Construction vehicle, Bus, Trailer, Barrier, Motorcycle, Bicycle, Pedestrian, Traffic cone
KITTI (Fine-tune)	Car, Pedestrian, Cyclist

  - Undetected hard instances

ONCE labeled set			Pseudo label set		
Vehicle	Ped.	Cyclist	Vehicle	Ped.	Cyclist
19.01	4.52	5.63	15.67	1.63	1.90

Be suppressed during the pre-training process

Method: AD-PT

- Learning Unified Representations under Large-scale Point Cloud Dataset
  - Consider as an open-set learning problem

- Consider background region proposals with relatively high objectness scores to be unknown instances
- Two-branch head as a committee
- Discover corresponding features using positional relationship

$$(\hat{\mathbf{F}}^{\Gamma_1}, \hat{\mathbf{F}}^{\Gamma_2}) = \{(\tilde{f}_i^{\Gamma_1}, \tilde{f}_j^{\Gamma_2}) | \sqrt{(c_{i,x}^{\Gamma_1} - c_{j,x}^{\Gamma_2})^2 + (c_{i,y}^{\Gamma_1} - c_{j,y}^{\Gamma_2})^2 + (c_{i,z}^{\Gamma_1} - c_{j,z}^{\Gamma_2})^2} < \tau\}$$

- Consistency loss

$$\mathcal{L}_{consist} = \frac{1}{BK} \sum_{i=1}^{B} \sum_{j=1}^{K} (\hat{f}_j^{\Gamma_1} - \hat{f}_j^{\Gamma_2})^2$$

CONTENTS

Experimental Results

➢ Results on Waymo

Method	Paradigm	Data amount	L2 AP / APH			
			Overall	Vehicle	Pedestrian	Cyclist
From scratch (SECOND)	-	3%	52.00 / 37.70	58.11 / 57.44	51.34 / 27.38	46.57 / 28.28
From scratch (SECOND)	-	20%	60.62 / 56.86	64.26 / 63.73	59.72 / 50.38	57.87 / 56.48
ProposalContrast (SECOND) [30]	SS-PT	20%	60.91 / 57.16	64.50 / 63.90	**60.33** / 51.00	57.90 / 56.60
BEV-MAE (SECOND) [12]	SS-PT	20%	61.03 / 57.30	64.42 / 63.87	59.97 / 50.65	58.69 / 57.39
MeanTeacher (SECOND) [20]	Semi	20%	60.93 / 57.31	64.22 / 63.73	59.54 / 50.80	58.66 / 57.41
Ours (SECOND)	AD-PT	3%	55.41 / 51.78	60.53 / 59.93	54.91 / 45.78	50.79 / 49.65
Ours (SECOND)	AD-PT	20%	**61.26 / 57.69**	**64.54 / 64.00**	60.25 / **51.21**	**59.00 / 57.86**
From scratch (CenterPoint)	-	3%	59.00 / 56.29	57.12 / 56.57	58.66 / 52.44	61.24 / 59.89
From scratch (CenterPoint)	-	20%	66.47 / 64.01	64.91 / 64.42	66.03 / 60.34	68.49 / 67.28
GCC-3D (CenterPoint) [11]	SS-PT	20%	65.29 / 62.79	63.97 / 63.47	64.23 / 58.47	67.68 / 66.44
ProposalContrast (CenterPoint) [30]	SS-PT	20%	66.67 / 64.20	65.22 / 64.80	66.40 / 60.49	68.48 / 67.38
BEV-MAE (CenterPoint) [12]	SS-PT	20%	66.92 / 64.45	64.78 / 64.29	66.25 / 60.53	**69.73 / 68.52**
MeanTeacher (CenterPoint) [20]	Semi	20%	66.66 / 64.23	64.94 / 64.43	66.35 / 60.61	68.69 / 67.65
Ours (CenterPoint)	AD-PT	3%	61.21 / 58.46	60.35 / 59.79	60.57 / 54.02	62.73 / 61.57
Ours (CenterPoint)	AD-PT	20%	**67.17 / 64.65**	**65.33 / 64.83**	**67.16 / 61.20**	69.39 / 68.25
From scratch (PV-RCNN++)	-	3%	63.81 / 61.10	64.42 / 63.93	64.33 / 57.79	62.69 / 61.59
From scratch (PV-RCNN++)	-	20%	69.97 / 67.58	69.18 / 68.75	70.88 / 65.21	69.84 / 68.77
ProposalContrast (PV-RCNN++) [30]	SS-PT	20%	70.30 / 67.78	69.45 / 69.00	71.42 / 65.68	70.04 / 69.05
BEV-MAE (PV-RCNN++) [12]	SS-PT	20%	70.54 / 68.11	69.53 / 69.07	71.50 / 65.69	70.60 / 69.56
MeanTeacher (PV-RCNN++) [20]	Semi	20%	70.62 / 68.14	69.21 / 68.81	71.96 / 66.42	70.17 / 69.21
Ours (PV-RCNN++)	AD-PT	3%	68.33 / 65.69	68.17 / 67.70	68.82 / 62.39	68.00 / 67.00
Ours (PV-RCNN++)	AD-PT	20%	**71.55 / 69.23**	**70.62 / 70.19**	**72.36 / 66.82**	**71.69 / 70.70**

Experimental Results

- Results on nuScenes

Method	Setting	Data amount	mAP	Car			Pedestrian			Cyclist		
			(Mod.)	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
From scratch (SECOND)	-	20%	61.70	89.78	78.83	76.21	52.08	47.23	43.37	76.35	59.06	55.24
From scratch (SECOND)	-	100%	66.70	89.63	80.78	78.21	58.05	52.61	48.24	84.25	66.71	62.50
Ours (SECOND)	AD-PT	20%	65.95	90.23	80.70	78.29	55.63	49.67	45.12	83.78	67.50	63.40
Ours (SECOND)	AD-PT	100%	**67.58**	**90.36**	**81.39**	**78.41**	**58.30**	**53.58**	**48.72**	**86.04**	**67.78**	**63.95**
From scratch (PV-RCNN)	-	20%	66.71	91.81	82.52	80.11	58.78	53.33	47.61	86.74	64.28	59.53
ProposalContrast (PV-RCNN) [30]	SS-PT	20%	68.13	91.96	82.65	80.15	62.58	55.05	50.06	88.58	66.68	62.32
From scratch (PV-RCNN)	-	100%	70.57	-	84.50	-	-	57.06	-	-	70.14	-
GCC-3D (PV-RCNN) [11]	SS-PT	100%	71.26	-	-	-	-	-	-	-	-	-
STRL (PV-RCNN) [6]	SS-PT	100%	71.46	-	84.70	-	-	57.80	-	-	71.88	-
PointContrast (PV-RCNN) [24]	SS-PT	100%	71.55	91.40	84.18	82.25	65.73	57.74	52.46	91.47	72.72	67.95
ProposalContrast (PV-RCNN) [30]	SS-PT	100%	72.92	**92.45**	84.72	82.47	68.43	60.36	55.01	**92.77**	**73.69**	**69.51**
Ours (PV-RCNN)	AD-PT	20%	69.43	92.18	82.75	82.12	65.50	57.59	51.84	84.15	67.96	64.73
Ours (PV-RCNN)	AD-PT	100%	**73.01**	91.96	**84.75**	**82.53**	**68.87**	**60.79**	**55.42**	91.81	73.49	69.21

Experimental Results

- Results on KITTI

Method	Setting	Data amount	mAP (Mod.)	Car Easy	Car Mod.	Car Hard	Pedestrian Easy	Pedestrian Mod.	Pedestrian Hard	Cyclist Easy	Cyclist Mod.	Cyclist Hard
From scratch (SECOND)	-	20%	61.70	89.78	78.83	76.21	52.08	47.23	43.37	76.35	59.06	55.24
From scratch (SECOND)	-	100%	66.70	89.63	80.78	78.21	58.05	52.61	48.24	84.25	66.71	62.50
Ours (SECOND)	AD-PT	20%	65.95	90.23	80.70	78.29	55.63	49.67	45.12	83.78	67.50	63.40
Ours (SECOND)	AD-PT	100%	**67.58**	**90.36**	**81.39**	**78.41**	**58.30**	**53.58**	**48.72**	**86.04**	**67.78**	**63.95**
From scratch (PV-RCNN)	-	20%	66.71	91.81	82.52	80.11	58.78	53.33	47.61	86.74	64.28	59.53
ProposalContrast (PV-RCNN) [30]	SS-PT	20%	68.13	91.96	82.65	80.15	62.58	55.05	50.06	88.58	66.68	62.32
From scratch (PV-RCNN)	-	100%	70.57	-	84.50	-	-	57.06	-	-	70.14	-
GCC-3D (PV-RCNN) [11]	SS-PT	100%	71.26	-	-	-	-	-	-	-	-	-
STRL (PV-RCNN) [6]	SS-PT	100%	71.46	-	84.70	-	-	57.80	-	-	71.88	-
PointContrast (PV-RCNN) [24]	SS-PT	100%	71.55	91.40	84.18	82.25	65.73	57.74	52.46	91.47	72.72	67.95
ProposalContrast (PV-RCNN) [30]	SS-PT	100%	72.92	**92.45**	84.72	82.47	68.43	60.36	55.01	**92.77**	**73.69**	**69.51**
Ours (PV-RCNN)	AD-PT	20%	69.43	92.18	82.75	82.12	65.50	57.59	51.84	84.15	67.96	64.73
Ours (PV-RCNN)	AD-PT	100%	**73.01**	91.96	**84.75**	**82.53**	**68.87**	**60.79**	**55.42**	91.81	73.49	69.21

Experimental Results

- Ablation studies on data preparation

Method	Enhancement	Waymo L2 AP/APH				nuScenes	
		Overall	Vehicle	Pedestrian	Cyclist	mAP	NDS
Baseline	None	67.12 / 64.55	67.45 / 66.97	67.74 / 61.15	66.19 / 65.24	36.26	45.04
Baseline+re-scaling	Object-size	67.39 / 64.68	67.52 / 67.03	67.82 / 61.24	66.83 / 65.79	39.72	49.93
Baseline+re-sampling	LiDAR-beam	67.37 / 64.70	67.70 / 67.21	68.21 / 61.71	66.15 / 65.18	41.35	51.03
Baseline+re-scaling+re-sampling	Both	**67.77 / 65.09**	**68.01 / 67.61**	**68.32 / 61.69**	**66.99 / 65.98**	**43.11**	**52.41**

- Ablation studies on training algorithm

Method	Waymo L2 AP/APH				nuScenes	
	Overall	Vehicle	Pedestrian	Cyclist	mAP	NDS
Baseline	67.77 / 65.09	68.01 / 67.61	68.32 / 61.69	66.99 / 65.98	43.11	52.41
Baseline+UIL	67.97 / 65.35	67.99 / 67.58	68.62 / 62.12	67.32 / 66.35	43.92	52.65
Baseline+UIL+CL	**68.33 / 65.69**	**68.17 / 67.70**	**68.82 / 62.39**	**68.00 / 67.00**	**44.99**	**52.99**

Experimental Results

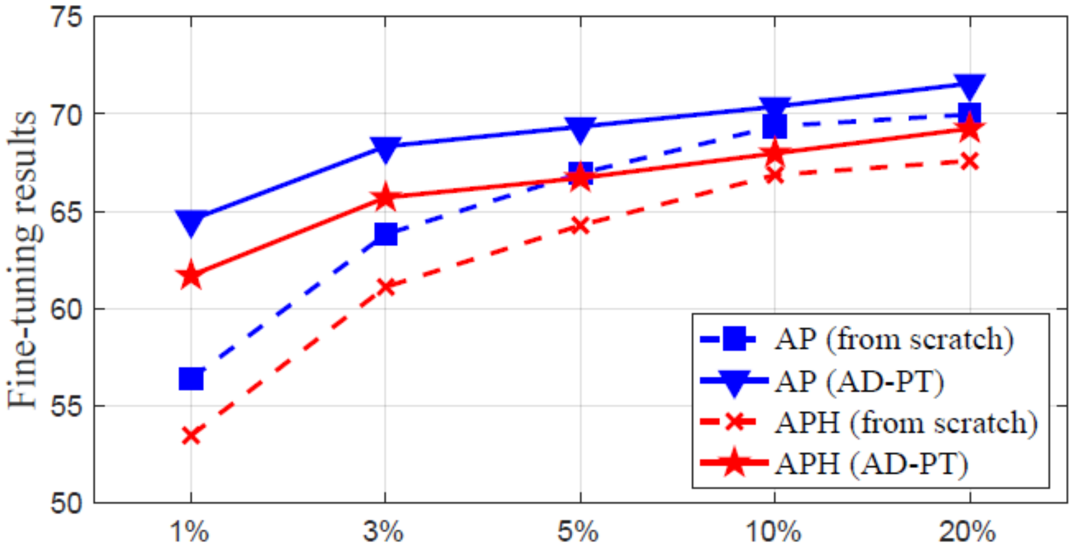
- Increasing pre-training data

Pre-training dataset	Waymo L2 AP/APH			
	Overall	Vehicle	Pedestrian	Cyclist
KITTI (~4k)	64.28 / 63.16	64.73 / 64.19	64.43 / 57.30	63.69 / 62.60
ONCE (~4k)	64.28 / 61.36	66.11 / 65.64	66.26 / 59.51	65.39 / 64.35
ONCE (~10k)	66.94 / 64.24	67.41 / 66.91	67.97 / 61.39	65.45 / 64.43
ONCE (~100k)	68.33 / 65.69	68.17 / 67.70	68.82 / 62.39	68.00 / 67.00
ONCE (~500k)	**69.04 / 66.52**	**68.69 / 68.23**	**69.81 / 63.74**	**68.61 / 67.60**

Pre-training dataset	Waymo L2 AP/APH				KITTI Moderate mAP			
	Overall	Vehicle	Pedestrian	Cyclist	Overall	Car	Pedestrian	Cyclist
ONCE (~100k)	68.33 / 65.69	68.17 / 67.70	68.82 / 62.39	68.00 / 67.00	69.43	82.75	57.59	67.96
ONCE (~500k)	69.04 / 66.52	68.69 / 68.23	69.81 / 63.74	68.61 / 67.60	71.36	83.17	58.14	72.78
ONCE (~1M)	**69.63 / 67.08**	**69.03 / 68.57**	**70.54 / 64.34**	**69.33 / 68.33**	**72.37**	**83.47**	**59.84**	**73.81**

Experimental Results

➢ Increasing fine-tuning data



➢ Fine-tuning on the same dataset

Init.	SECOND				CenterPoint			
	Overall	0-30m	30-50m	>50m	Overall	0-30m	30-50m	>50m
Random Initialization	56.47	65.94	51.05	36.44	64.94	74.52	59.47	44.28
AD-PT Initialization	**64.10**	**74.34**	**57.69**	**41.23**	**67.73**	**76.48**	**61.85**	**46.29**

3DTrans Team

Fudan EDL Lab