MOS: Model Synergy for Test-Time Adaptation on LiDAR-Based 3D Object Detection

Zhuoxiao Chen[†], Junjie Meng[†], Mahsa Baktashmotlagh[†], Yonggang Zhang[‡], Zi (Helen) Huang[†], Yadan Luo[†]

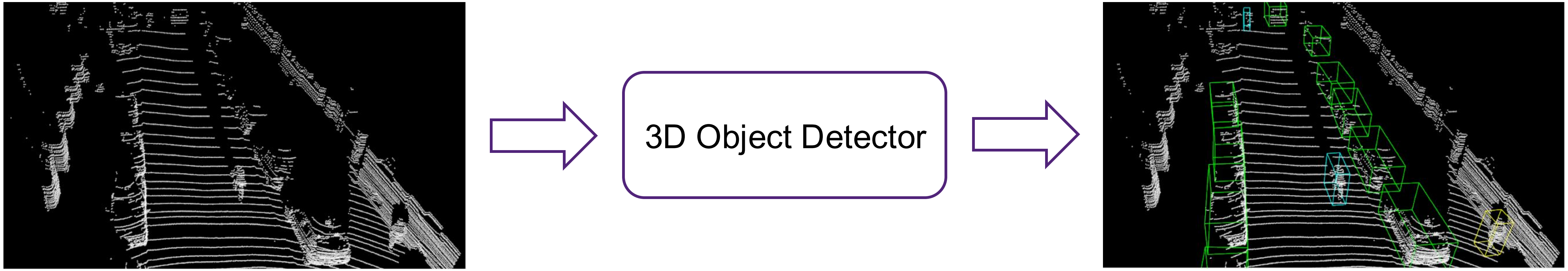
[†]The University of Queensland, [‡]Hong Kong Baptist University

International Conference on Learning Representations (ICLR 2025 Oral)

3D Object Detection & Challenges

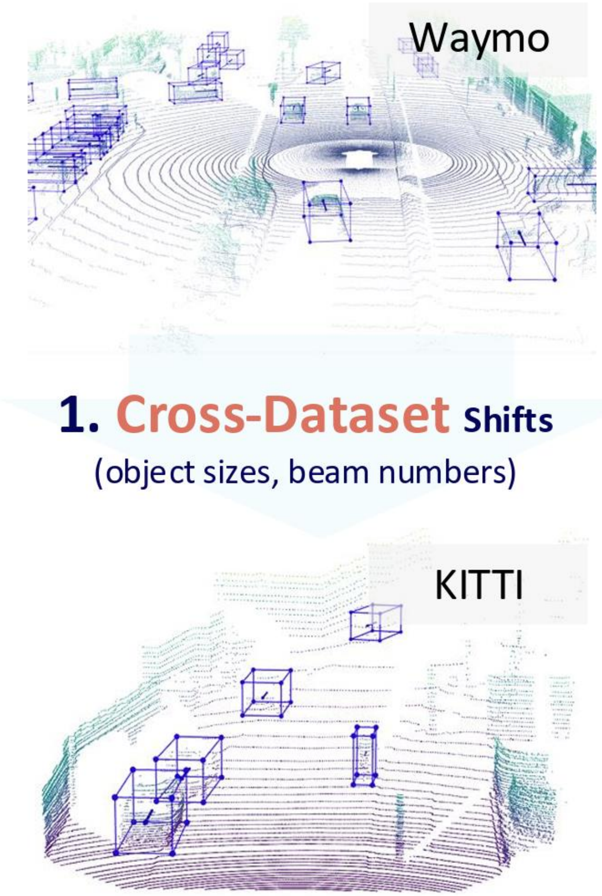
Goal: From LiDAR point clouds, output class-labelled 3D boxes.

Challenges: Domain shift – test data is different from training data, degrading the performance.



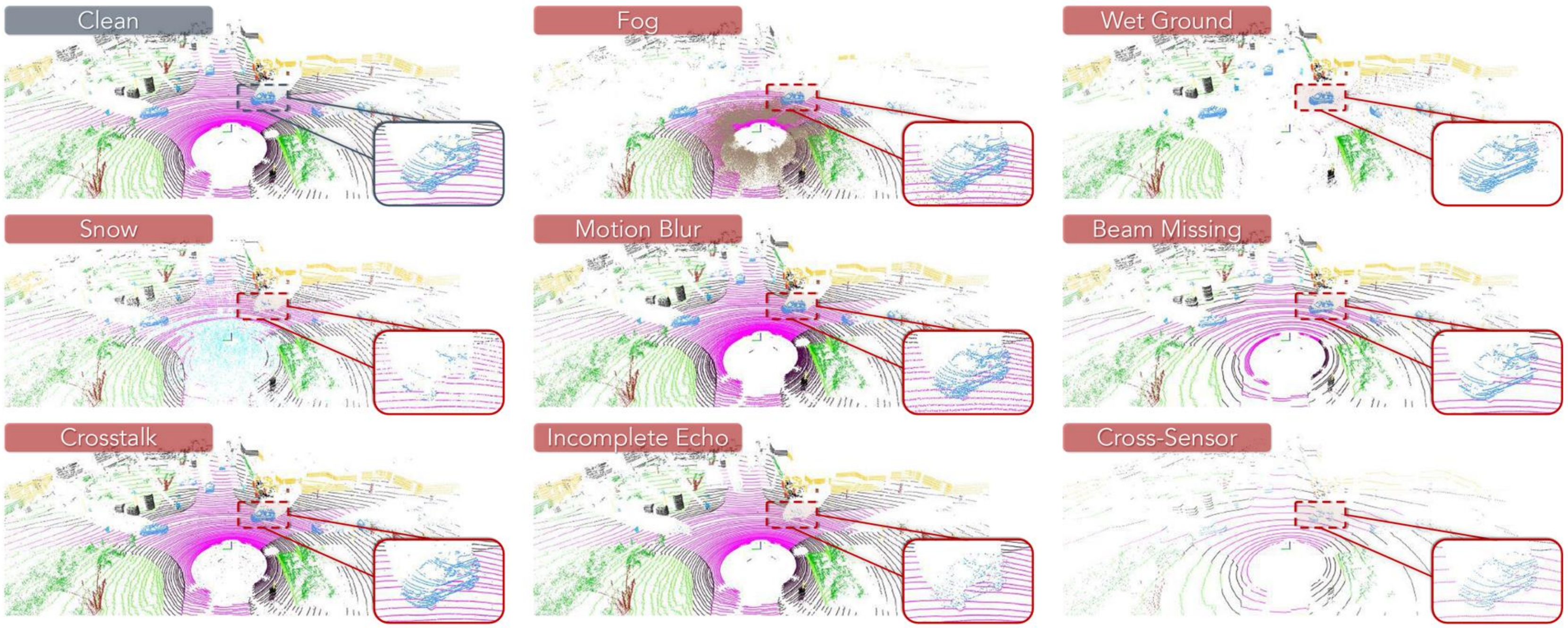
Cross-Dataset Shifts

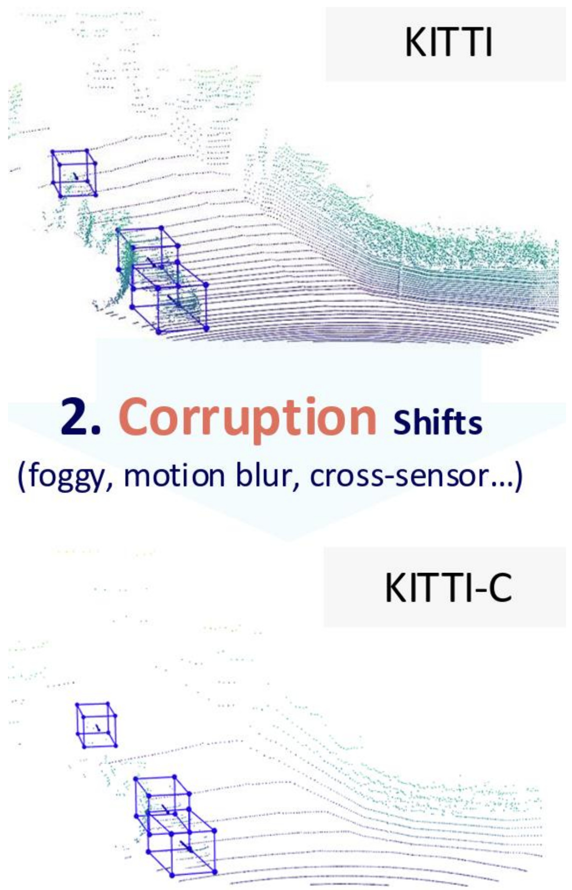
- Shifted **Object Size**:
  - Average car length: 4.7m (Waymo) vs. 3.9m (KITTI)
- Shifted **Point Cloud Density**:
  - 64-beam (KITTI, Waymo) vs. 32-beam (nuScnenes)
- Shifted **Environment**:
  - Germany (KITTI) vs. USA (nuScenes) vs. Singapore (Waymo)



Corruption Shifts

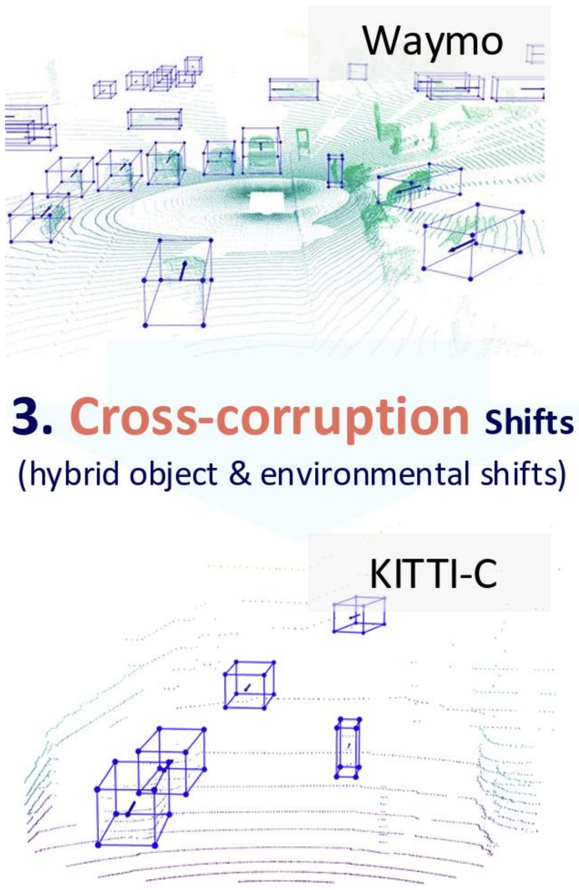
We follow Robo3D Benchmark (Kong, L., Et al.) to study **eight** types of corruptions, mainly caused by **severe weather** and **sensor failure**.





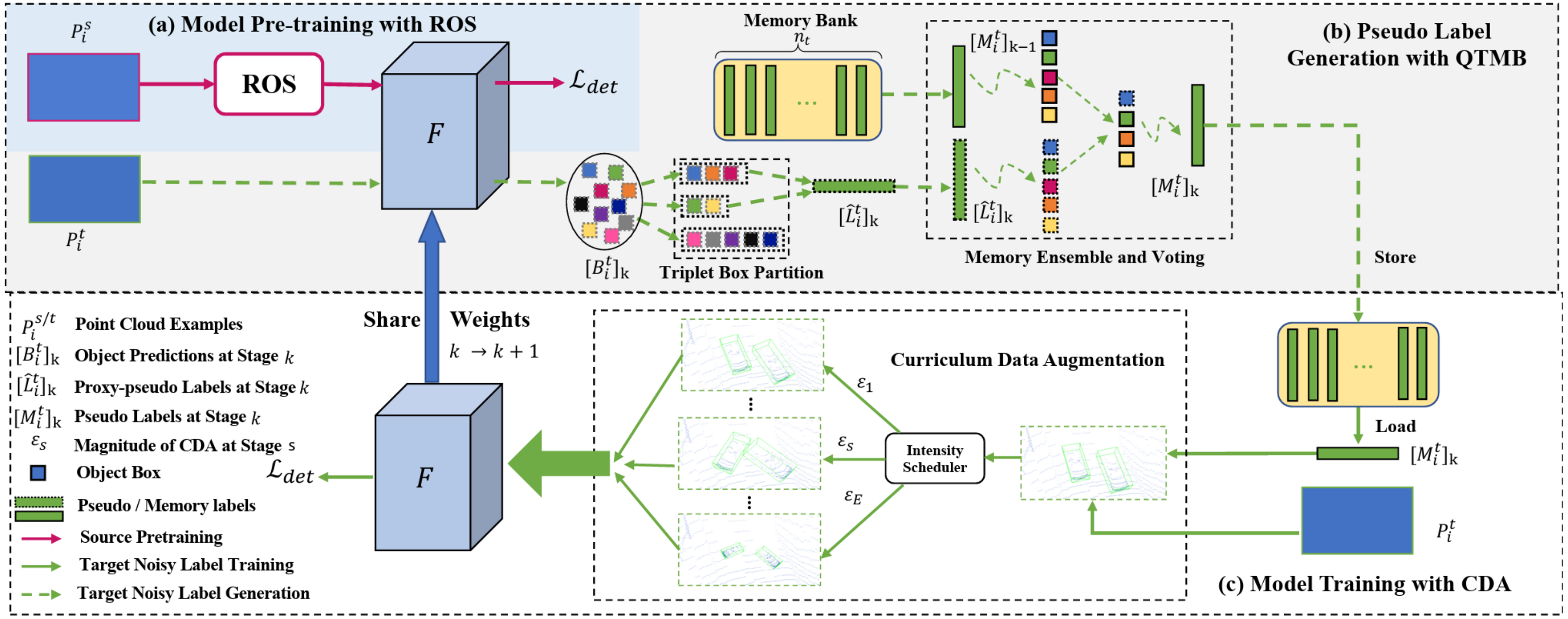
Cross-Corruption Shifts in 3D Object Detection

Question: what if a 3D detection system is deployed at a **new location** at the same time there is a **heavy snow**?



Existing Works – Unsupervised Domain Adaptation (UDA)

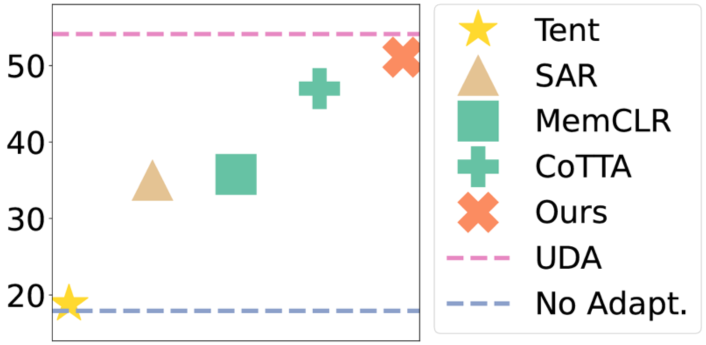
- UDA methods self-train the 3D detector **offline** for **many epochs** on the **full** target dataset.
- **Question: what if the test domain changes dynamically (e.g. travelling to a new location, weather variations, sensor issues)?**



ST3D (Yang, J., Et al.)

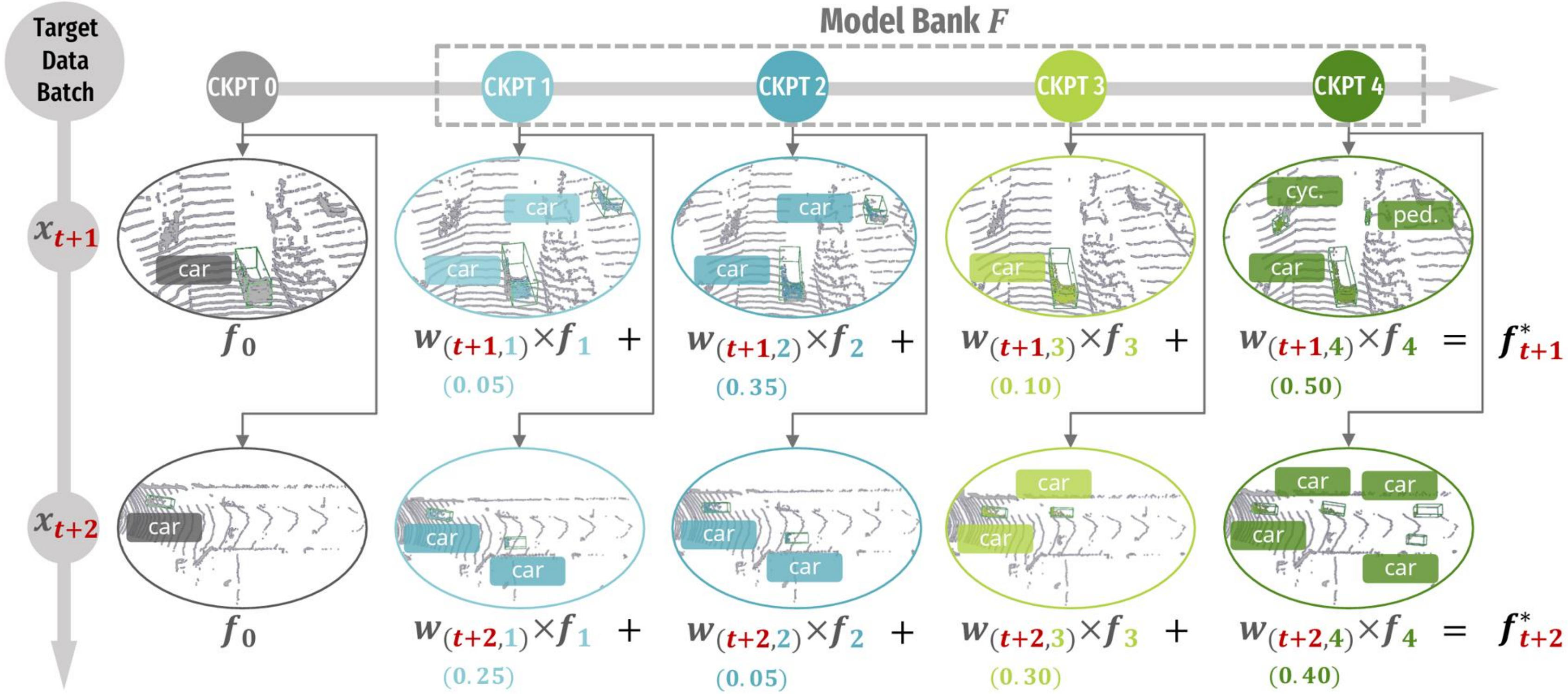
Test-Time Adaptation on 3D Object Detection

- On-the-fly model update per test sample.
- Mean-teacher variants (CoTTA (Wang, Q., Et al.) / MemCLR (VS, V., Et al.)) → less forgetting.
- Naively averaging all saved checkpoints → sub-optimal.
- **Question**: how to identify and reuse only the most informative checkpoints?



Model Synergy – Core Idea

- Identify checkpoints that **best fit the current test batch, then** assemble them.
- Minimize redundancy & focus on the unique knowledge.



Model Synergy – Compute Optimal Ensemble Weights

- **Goal**: build an on-the-fly “super-model” by weighting K saved checkpoints.

$$\mathbf{F}\mathbf{w} = f_t^*, \quad \mathbf{w} = (\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T f_t^* = (\mathbf{F}^T\mathbf{F})^{-1}\mathbb{1}^K.$$

- **Inverse of Generalized Gram Matrix**: prioritize diverse checkpoints and minimize redundancy:

$$\mathbf{G} = \mathbf{F}^T\mathbf{F} = \Big(\langle f_i, f_j \rangle\Big)_{i,j=1}^{K} \in \mathbb{R}^{K\times K}, \;\; f_i, f_j \in \mathbf{F}$$

- **Similarity Measure**: combines feature overlap and bounding-box agreement to gauge redundancy:

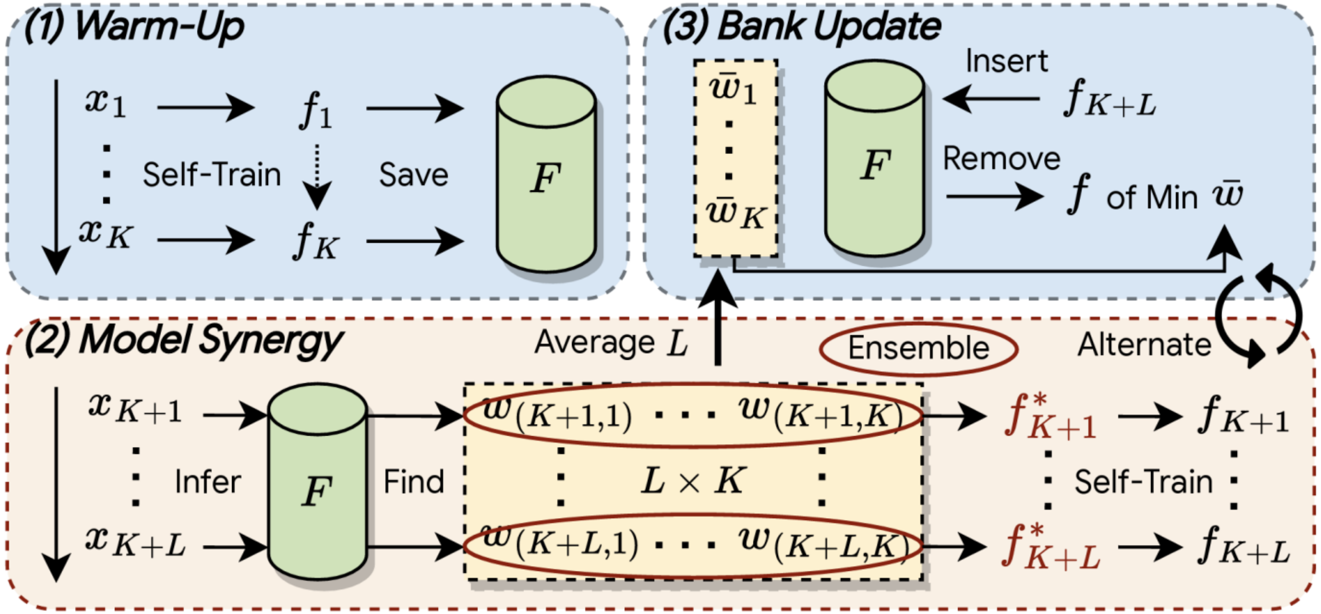
- The assembled tation:

$$\tilde{\mathbf{w}} = (\mathbf{F}^T\mathbf{F})^{-1}\mathbb{1}^K = \tilde{\mathbf{G}}^{-1}\mathbb{1}^K, \;\; \tilde{\mathbf{G}} = \Big(\mathsf{S}_{\text{box}}\langle f_i, f_j \rangle \times \mathsf{S}_{\text{feat}}\langle f_i, f_j \rangle\Big)_{i,j=1}^{K}$$

$$f_t^* = \sum_{i=1}^{K} w_i f_i, \;\; w_i \in \tilde{\mathbf{w}}, f_i \in \mathbf{F}, \quad \hat{\mathbf{B}}^t \leftarrow f_t^*(x_t), \;\; f_t \xleftarrow{\text{train}} \mathbf{aug}(x_t, \hat{\mathbf{B}}^t)$$

Model Synergy – Overall TTA Framework

- **Phase 1 (Warm-up)**: self-train on initial batches → build a checkpoint bank.
- **Phase 2 (Model Synergy)**: assemble the best checkpoints into one super model for self-training.
- **Phase 3 (Bank Update)**: drop redundant models, add new ones, then repeat.



Experiments – Address Cross-Dataset Shift

- **Transfer Tasks**: Waymo → KITTI and nuScenes → KITTI.
- Outperform the best baseline by 81.6% in AP 3D.
- Achieve **comparable or better performance** than UDA methods.

Method	Venue	TTA	Waymo →KITTI		nuScenes →KITTI	
			AP_{BEV} / AP_{3D}	Closed Gap	AP_{BEV} / AP_{3D}	Closed Gap
No Adapt.	-	-	67.64 / 27.48	-	51.84 / 17.92	-
SN	CVPR'20	×	78.96 / 59.20	+72.33% / +69.00%	40.03 / 21.23	+37.55% / +5.96%
ST3D	CVPR'21	×	82.19 / 61.83	+92.97% / +74.72%	75.94 / 54.13	+76.63% / +65.21%
Oracle	-	-	83.29 / 73.45	-	83.29 / 73.45	-
Tent	ICLR'21	✓	65.09 / 30.12	−16.29% / +5.74%	46.90 / 18.83	−15.71% / +1.64%
CoTTA	CVPR'22	✓	67.46 / 35.34	−1.15% / +17.10%	68.81 / 47.61	+53.96%/ +53.47%
SAR	ICLR'23	✓	65.81 / 30.39	−11.69% / +6.33%	61.34 / 35.74	+30.21% / +32.09%
MemCLR	WACV'23	✓	65.61 / 29.83	−12.97% / +5.11%	61.47 / 35.76	+30.62% / +32.13%
MOS	-	✓	**81.90 / 64.16**	**+91.12% / +79.79%**	**71.13 / 51.11**	**+61.33% / +59.78%**

Experiments – Address Cross-Corruption Shift

MOS lifts AP 3D on the **toughest** cases (Incomplete Echo, Cross Sensor) by **97.9%** and **76.4%** under the **hard** setting, significantly surpassing all baselines.

Easy

	No Adapt.	Tent	CoTTA	SAR	MemCLR	MOS
Fog	20.25	23.49	32.54	23.17	22.80	51.23
Wet.	30.76	34.03	44.86	33.36	32.70	69.97
Snow	12.44	14.22	26.75	14.72	14.28	28.77
Moti.	8.16	9.12	13.74	8.85	8.54	26.21
Beam.	18.25	21.13	32.01	20.90	20.90	46.79
CrossT.	23.13	25.07	28.74	24.54	24.74	42.44
Inc.	4.91	7.38	10.97	7.22	6.66	22.43
CrossS.	11.01	16.94	15.24	16.08	17.40	32.12

Moderate

	No Adapt.	Tent	CoTTA	SAR	MemCLR	MOS
Fog	17.94	21.25	29.11	20.95	20.61	45.70
Wet.	27.74	30.94	39.74	30.14	30.06	60.14
Snow	13.03	14.82	26.32	14.61	14.37	31.14
Moti.	8.38	9.82	13.83	9.56	9.36	24.65
Beam.	13.57	15.82	24.13	15.68	15.78	34.96
CrossT.	21.71	23.77	28.65	23.18	23.39	40.98
Inc.	3.84	6.14	8.36	5.59	5.25	16.14
CrossS.	8.09	11.67	10.67	10.55	11.83	21.63

Hard

	No Adapt.	Tent	CoTTA	SAR	MemCLR	MOS
Fog	16.85	19.52	27.27	19.34	19.08	43.26
Wet.	26.31	29.92	38.23	28.23	28.24	58.32
Snow	11.40	13.71	24.50	13.81	13.50	28.87
Moti.	8.16	9.87	13.75	9.50	9.34	24.50
Beam.	12.53	14.99	22.86	14.83	14.98	32.20
CrossT.	20.03	22.12	27.63	21.64	21.77	39.63
Inc.	3.51	5.63	7.56	5.18	4.87	14.96
CrossS.	7.40	10.60	9.36	8.99	10.88	19.19

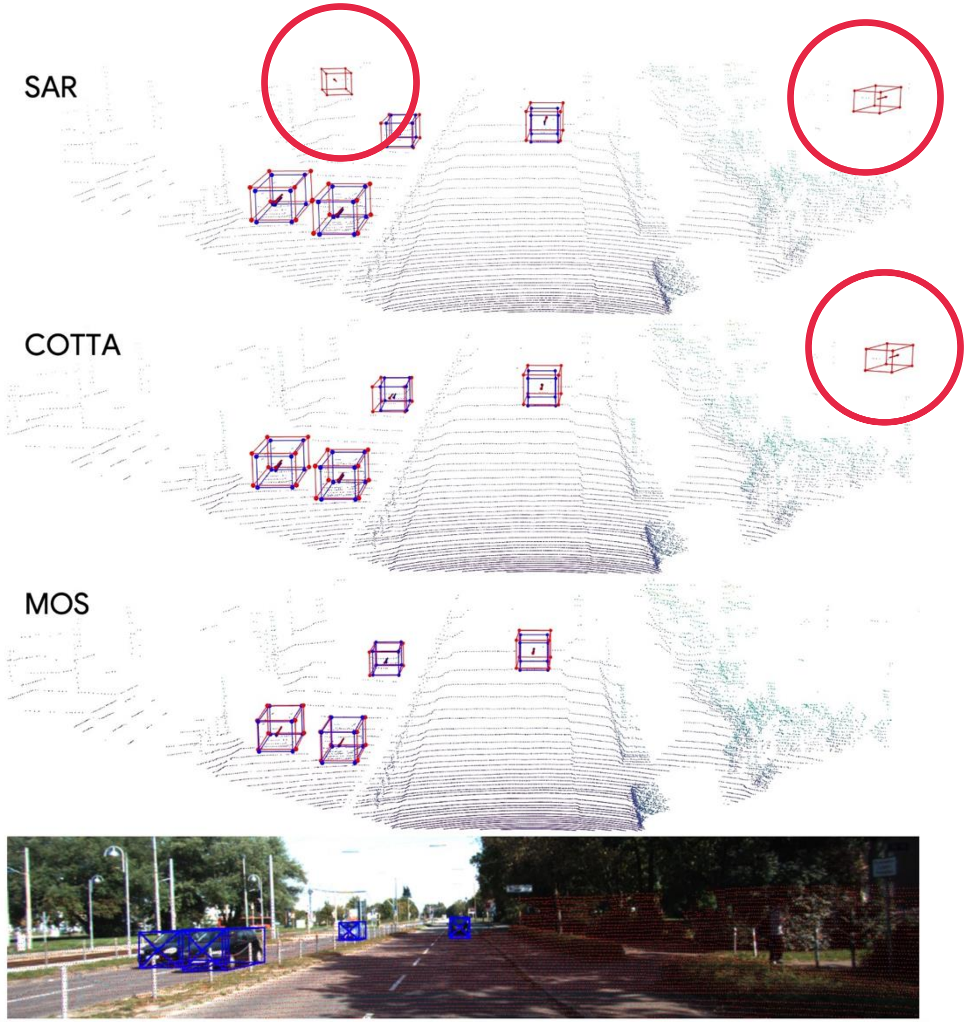
Experiments – Qualitative Results

Fewer False Positives

- Reduced misclassification of background as objects.

More Accurate Bounding Boxes

- Tighter alignment with ground truth.



Conclusion, Limitations & Future Directions

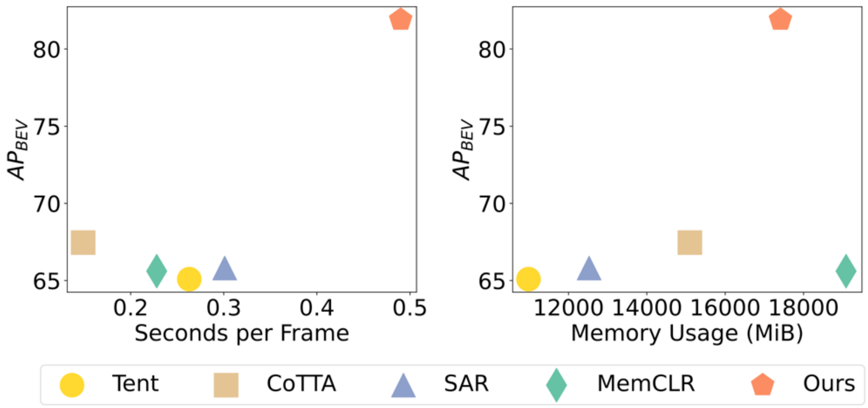
Contributions:

- Pioneered test-time adaptation for 3D detection.
- Mitigated three real-world shifts directly at deployment.

Limitations:

- Inferring **multiple checkpoints** can increase **test-time latency** (up to **0.4 s/frame**).
- Loading/unloading models strains CPU/GPU memory.

Future Directions: explore **partial weight storage** or **lightweight ensembles.**



The University of Queensland
AUSTRALIA
CREATE CHANGE

Thank You! ANY QUESTIONS?

Our poster session will be held on this afternoon (Friday, April 25 at 15:00).

We welcome further discussion and questions there!

Mr Zhuoxiao Chen | PhD Candidate
School of Electrical Engineering and Computer Science (EECS)
zhuoxiao.chen@uq.edu.au