CMDA: Cross-Modal and Domain Adversarial Adaptation for LiDAR-Based 3D Object Detection

Gyusam Chang*, Wonseok Roh*, Sujin Jang, Dongwook Lee, Daehyun Ji,
Gyeongrok Oh, Jinsun Park, Jinkyu Kim†, Sangpil Kim†

KOREA UNIVERSITY | Samsung Advanced Institute of Technology

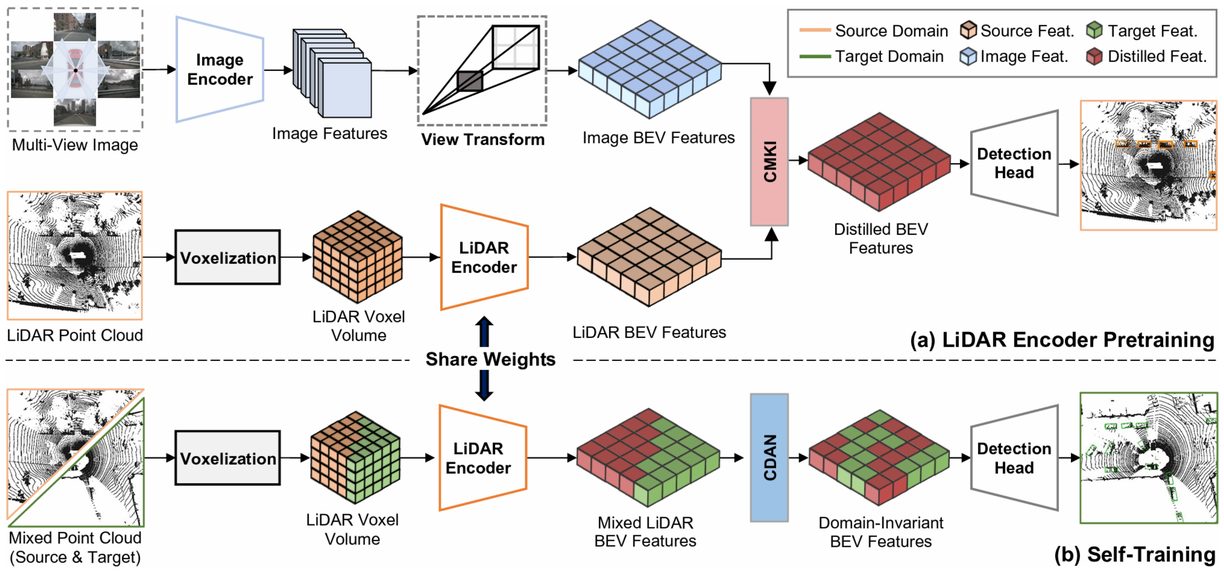
AAAI-24

Domain Shifts on LiDAR-based 3D Object Detection

Dataset	LiDAR Type	Beam Angles	Points per Beam	Annotation	Location
nuScenes	32-beam	[-30.0°, 10.0°]	1,084	car, bus, truck, construction vehicle, trailer	USA and Singapore
KITTI	64-beam	[-23.6°, 3.2°]	1,863	Car	Germany
Waymo	64-beam	[-18.0°, 2.0°]	2,258	Car	USA

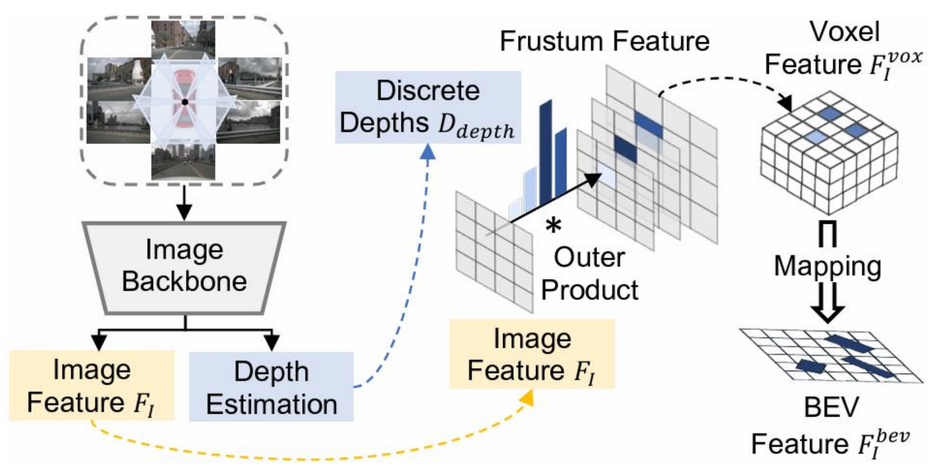
- ✓ LiDAR resolution (e.g., 64-bits → 32-bits)
- ✓ Annotation policy
- ✓ Location, weather, and day/night

Cross-Modal and Domain Adversarial Adaptation (CMDA)



- ✓ We advocate leveraging **multi-modal inputs** during the training phase to enhance the generalizability across diverse domains
- ✓ First, we encourage the LiDAR BEV features to **learn rich-semantic knowledge from camera BEV features**
- ✓ Second, we **explicitly guide such cross-modal learning via cross-domain adversarial pipeline**, achieving generalized perception against unseen target conditions

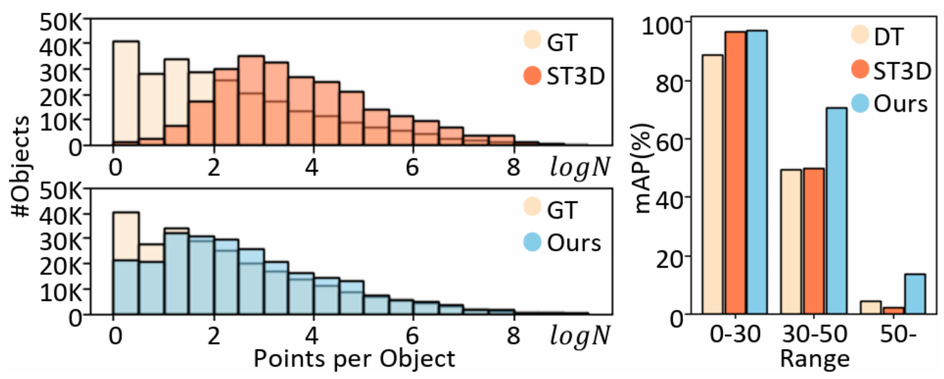
1. Cross-Modal Knowledge Interaction (CMKI)



- ✓ **Precise geometric alignment** is essential to ensure the quality of both image and point cloud features
- ✓ We project multi-modal inputs into **BEV (Bird's Eye View) joint representation**, facilitating effective cross-modal knowledge interaction
- ✓ We optimize 3D LiDAR-based features to contain **highly informative semantic clues** from 2D image-based features

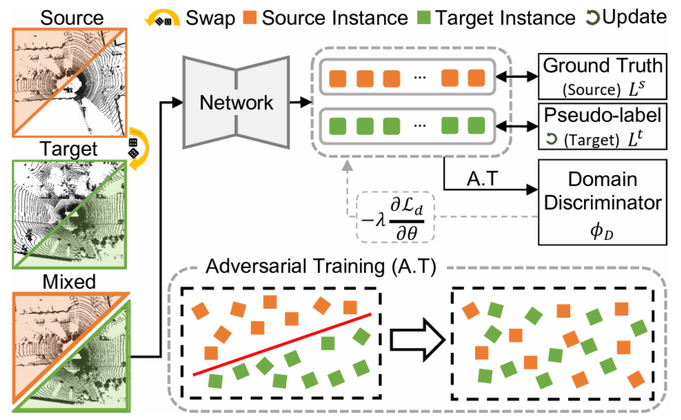
$$\mathcal{L}_{cmki} = \frac{1}{XY}\sum_{i=1}^{X}\sum_{j=1}^{Y}\left\|F_P^{bev}(i,j) - F_I^{bev}(i,j)\right\|_2$$

- **Impact of Utilizing Visual Semantic Priors**



- ✓ CMKI **effectively capture hard samples (i.e., low resolution and far distant objects)** that hinder self-training paradigm
- ✓ Specifically, CMKI **suppresses type 1 and 2 errors, inducing high quality pseudo ground truths**

2. Cross-Domain Adversarial Network (CDAN)



- ✓ To ensure an explicit connection across domains, we first introduce the **point cloud mix-up technique**, which swaps points sector with random azimuth angles.
- ✓ Then, we further apply **adversarial regularization to reduce the representational gap across domains**, guiding the model to learn domain-invariant information.
- ✓ Besides, we design a function that **minimizes independent BEV grid-wise entropy to suppress ambiguous and uncertain features** derived from mixed inputs.

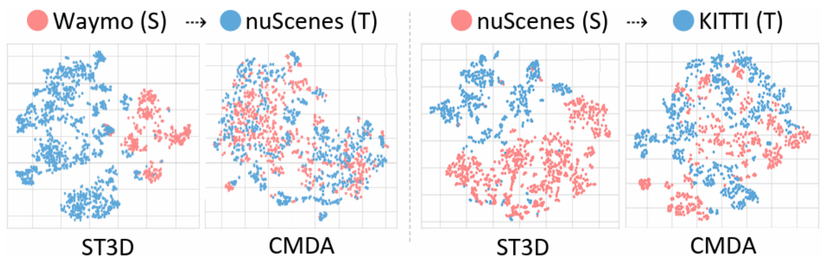
$$\mathcal{L}_d = -\mathbb{E}_{f,y_r\sim\mathbb{D}}\left[\sum_{r\in\mathcal{R}} y_r \log\phi_D(f)_r\right] \qquad \mathcal{L}_{ent} = \frac{-1}{\log ZC}\sum_{i=1}^{X}\sum_{j=1}^{Y}\sum_{c=1}^{ZC} F_P^{bev}(i,j,c)\log F_P^{bev}(i,j,c)$$

$$\mathcal{L}_{cdan} = \lambda_d\mathcal{L}_d + \lambda_{ent}\mathcal{L}_{ent}$$

- **Total Loss Function**

$$\mathcal{L}_{total} = \lambda_{det}\mathcal{L}_{det} + \mathcal{T}_{cmki}\lambda_{cmki}\mathcal{L}_{cmki} + \mathcal{T}_{cdan}\lambda_{cdan}\mathcal{L}_{cdan}$$

- **Effectiveness of Adversarial representational learning**



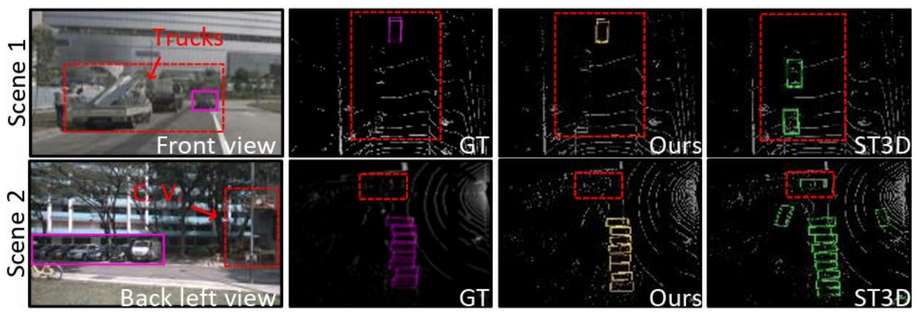
- ✓ CMDA results in a **harmoniously dispersed feature** space encompassing both target and source domains

Experimental Results

- **Overall Performance**

Task	Model	SECOND-IoU (Yan, Mao, and Li 2018)		PV-RCNN (Shi et al. 2020)	
		BEV AP ↑ / 3D AP ↑	Closed Gap ↑	BEV AP ↑ / 3D AP ↑	Closed Gap ↑
nuScenes → Waymo	Direct Transfer	39.18 / 20.78		41.30 / 25.89	
	ST3D (Yang et al. 2021)	45.35 / 27.12	+21.62% / +19.08%	52.50 / 36.21	+38.63% / +31.07%
	ST3D++ (Yang et al. 2022)	44.87 / 25.79	+19.94% / +15.08%	–.– / –.–	–.–% / –.–%
	CMDA (Ours)	**46.79 / 29.42**	**+26.66% / +26.00%**	**58.57 / 45.58**	**+59.57% / +59.29%**
	Oracle	67.72 / 54.01		70.29 / 59.10	
nuScenes → KITTI	Direct Transfer	51.84 / 17.92		68.15 / 37.17	
	SN (Wang et al. 2020)	40.03 / 21.23	-37.55% / +05.96%	60.48 / 49.47	-36.82% / +27.13%
	ST3D (Yang et al. 2021)	75.94 / 54.13	+76.63% / +59.50%	78.36 / 70.85	+49.02% / +74.30%
	ST3D++ (Yang et al. 2022)	80.52 / 62.37	+91.19% / +80.05%	–.– / –.–	–.–% / –.–%
	DTS (Hu, Liu, and Hu 2023)	81.40 / 66.60	+93.99% / +87.66%	83.90 / 71.80	+75.61% / +76.40%
	CMDA (Ours)	**82.13 / 68.95**	**+96.31% / +91.90%**	**84.85 / 75.02**	**+80.17% / +83.50%**
	Oracle	83.29 / 73.45		88.98 / 82.50	
Waymo → nuScenes	Direct Transfer	32.91 / 17.24		34.50 / 21.47	
	SN (Wang et al. 2020)	33.23 / 18.57	+01.69% / +07.54%	34.22 / 22.29	-01.50% / +04.80%
	ST3D (Yang et al. 2021)	35.92 / 20.19	+15.87% / +16.73%	36.42 / 22.99	+10.32% / +08.89%
	ST3D++ (Yang et al. 2022)	35.73 / 20.90	+14.87% / +20.76%	–.– / –.–	–.–% / –.–%
	LD (Wei et al. 2022)	40.66 / 22.86	+40.85% / +31.88%	43.31 / 25.63	+47.34% / +24.34%
	DTS (Hu, Liu, and Hu 2023)	41.20 / 23.00	+43.70% / +32.67%	44.00 / 26.20	+51.04% / +27.68%
	CMDA (Ours w/ LD)	**42.81 / 24.64**	**+52.19% / +41.97%**	**44.44 / 26.41**	**+53.41% / +28.91%**
	Oracle	51.88 / 34.87		53.11 / 38.56	

- **Qualitative Visualization of Waymo → nuScenes**



Summary

- ✓ To reduce the gap between source and target (where its labels are not accessible during training) domains, we propose CMKD composed two main steps: (i)Cross-modal LiDAR Encoder Pre-training and (ii) Cross-Domain LiDAR-Only Self Training.
- ✓ In (i), a pair of image-based and LiDAR-based BEV features is aligned to learn modality-agnostic (and thus more domain-invariant) features.
- ✓ Further, in (ii), we apply an adversarial regularization to reduce the representation gap between source and target domains.