DualDistill: A Unified Cross-Modal Knowledge Distillation Framework for Camera-Based BEV Representation

[1]Gaeun Kim, [1]Daeil Han, [2]Yeong Jun Koh, [1]Hanul Kim

[1] Seoul National University of Science and Technology [2] Chungnam National University

Project Page: github.com/Gaeun-Kimm/DualDistill

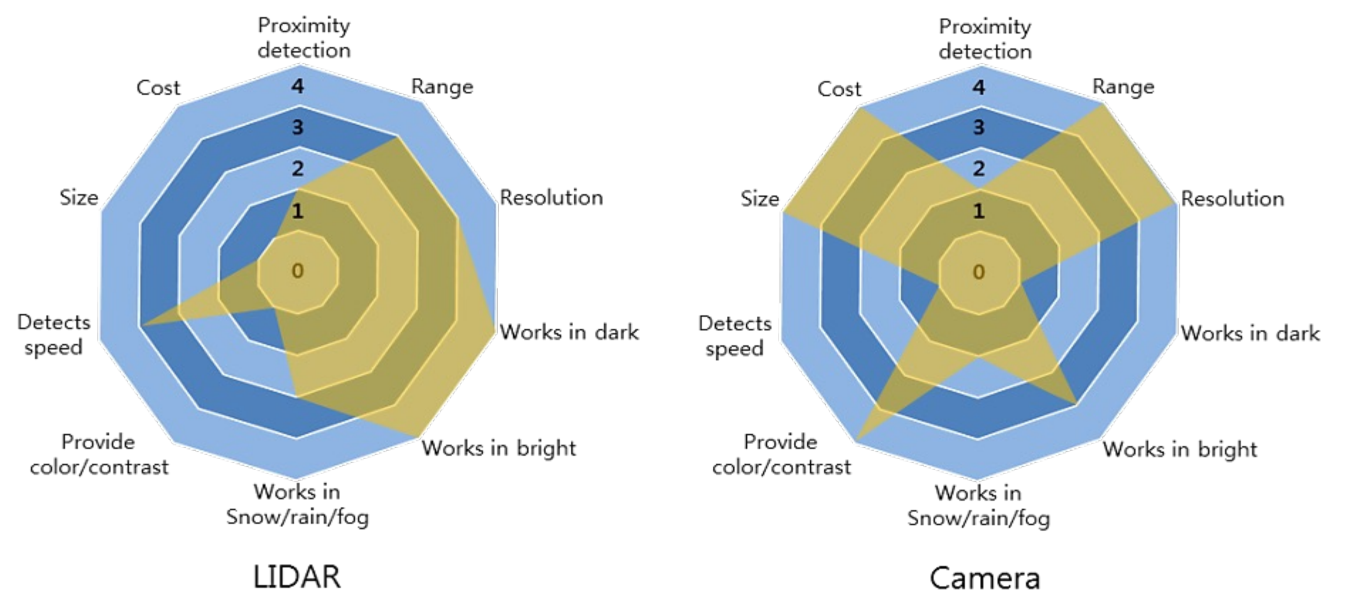
BMVC 2025

The 36th British Machine Vision Conference 24th - 27th November 2025, Sheffield, UK

Motivation

Camera-Based Autonomous Driving System

- Cost-effectiveness and rich semantic information
- Loss of explicit depth information leads to a performance gap compared to lidar-based systems

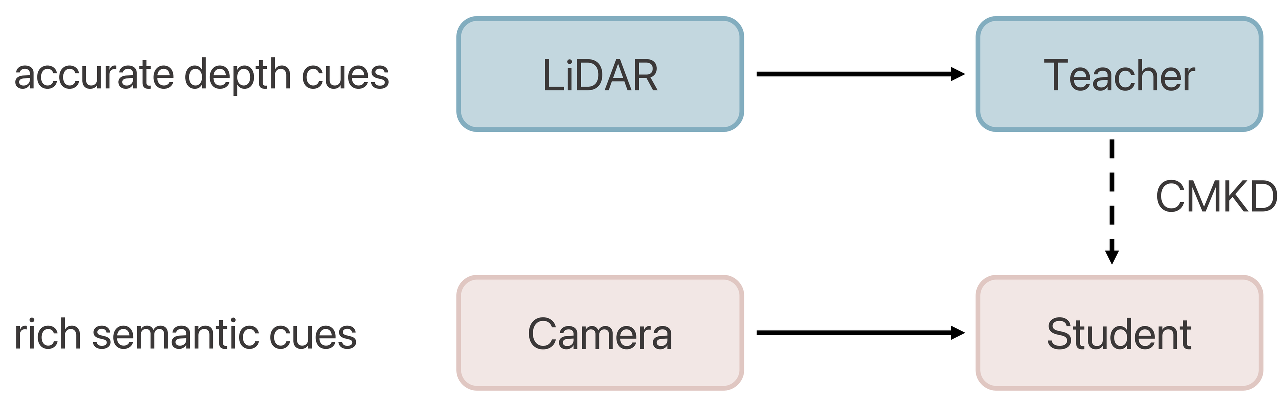


Cross-Modality Knowledge Distillation (CMKD)

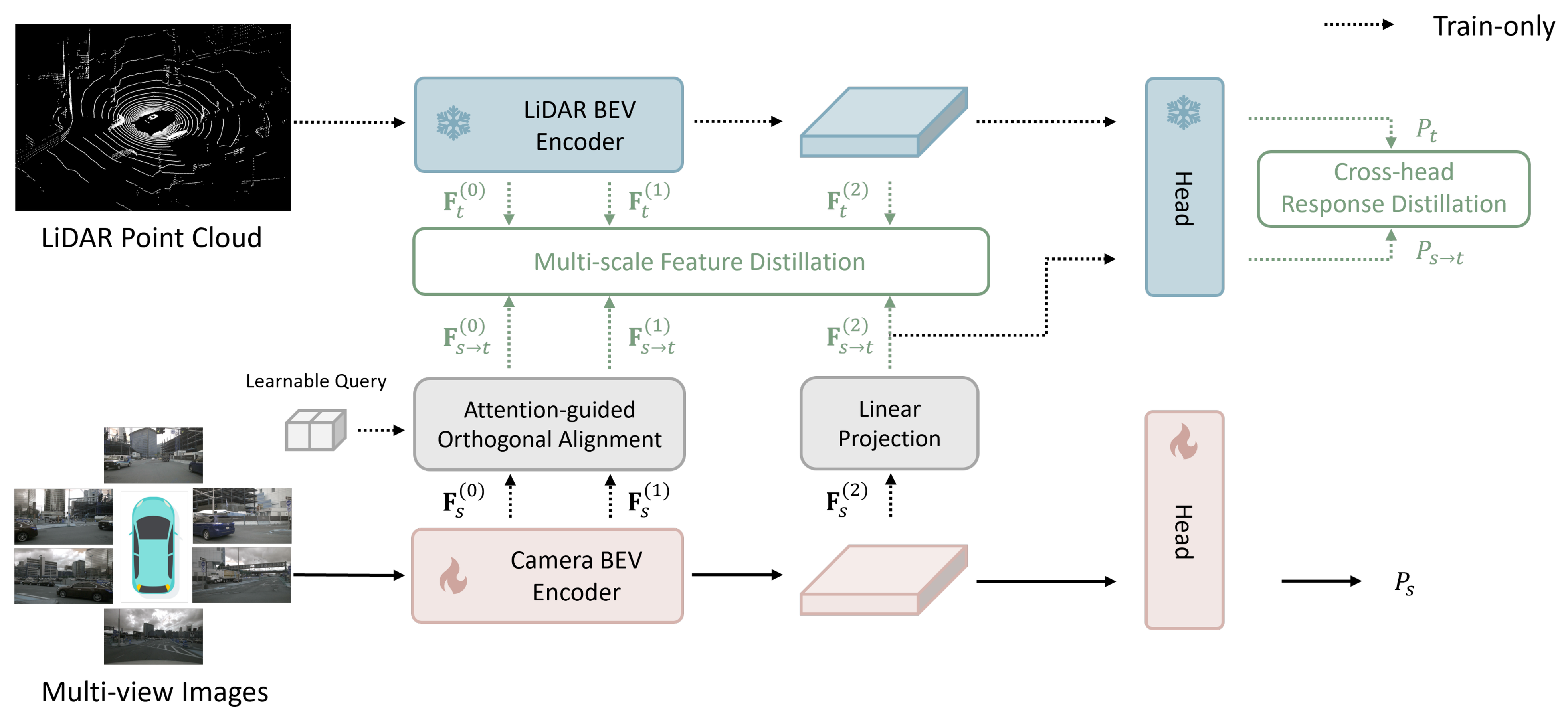
- Method to transfer knowledge from a teacher in one modality to a student in another, injecting modality cues while keeping single-modality inference

Main Idea & Challenge

- Distill LiDAR geometric priors into a camera-based student during training
- But LiDAR is an imperfect teacher. We should retain camera-specific advantages



Proposed Method



Attention-Guided Orthogonal Alignment (AOA)

- Cross-attention projector samples student features for distillation

$$\tilde{\mathbf{F}}_{s\to t} = \mathrm{Attn}(\mathbf{Q}, \mathbf{F}_s, \mathbf{F}_s)$$

- Orthogonal projector preserves the student's inherent feature structure while transforming it into the teacher's space

$$\mathbf{F}_{s\to t} = \tilde{\mathbf{P}}\tilde{\mathbf{F}}_{s\to t},\ \text{where}\ \tilde{\mathbf{P}} = \underset{\mathbf{P}}{\operatorname{argmin}} \|\mathbf{P}^{\intercal}\mathbf{P} - \mathbf{I}\|_F$$

Multi-Scale Feature Distillation (MFD)

- MFD aligns BEV representations between camera and LiDAR models with adaptive region weighting, distilling knowledge from an imperfect teacher

$$C(i,j) = \begin{cases} \dfrac{\alpha}{N_{\mathrm{TP}} + N_{\mathrm{FN}}} & \text{if } (i,j) \in \text{TP or FN} \\ \dfrac{\beta}{N_{\mathrm{FP}}} & \text{if } (i,j) \in \text{FP} \\ \dfrac{\gamma}{N_{\mathrm{TN}}} & \text{if } (i,j) \in \text{TN} \end{cases}$$

Category-aware region weights

$$S_t = N_{\mathrm{F}_t} \cdot \mathrm{softmax}(\bar{\mathbf{F}}_t/\omega)$$
$$S_s = N_{\mathrm{F}_s} \cdot \mathrm{softmax}(\bar{\mathbf{F}}_s/\omega)$$
$$S = \frac{1}{2}(S_t + S_s)$$

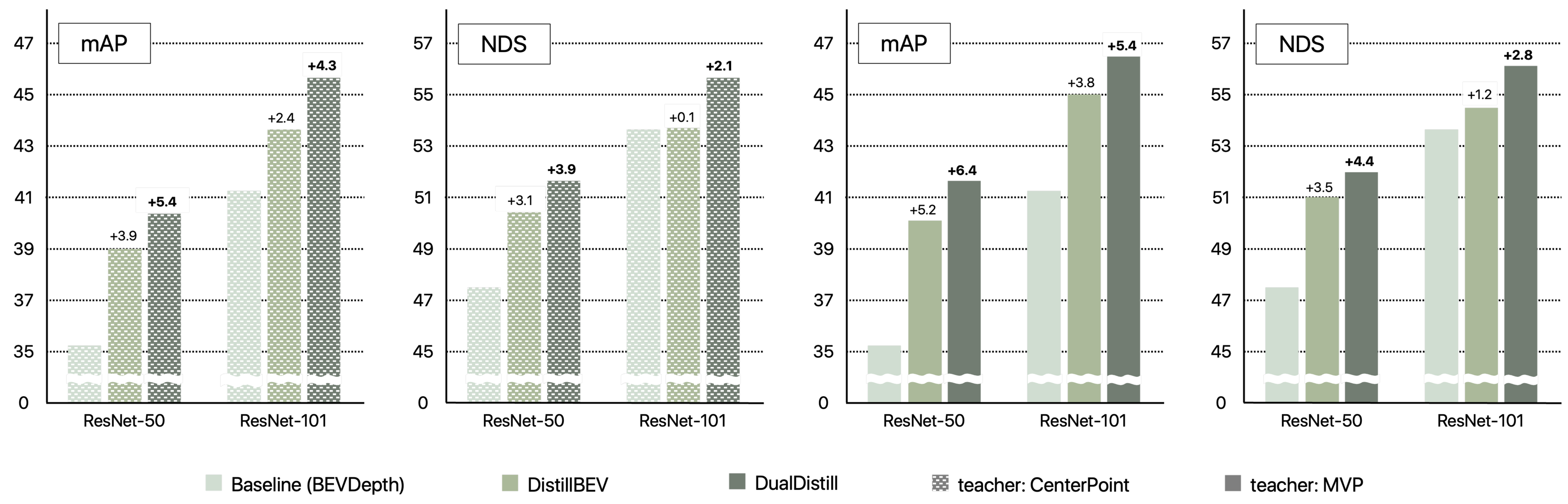
Saliency-aware region weights

Cross-Head Response Distillation (CRD)

- CRD enhances cross-modal consistency by comparing predictions of the shared detection head from teacher's and aligned student's features

Experiments on nuScenes 3D Object Detection

Model	Base	Teacher	mAP↑ (Δ)	NDS↑ (Δ)
CenterPoint	-	-	56.4	64.8
MVP	-	-	67.1	78.0
Backbone: ResNet-50				
BEVDepth	-	-	35.1	47.5
TiG-BEV	BEVDepth	CenterPoint	36.6 (+1.5)	46.1 (-1.4)
X[3]KD	BEVDepth	CenterPoint	39.0 (+3.9)	50.5 (+3.0)
DistillBEV	BEVDepth	CenterPoint	39.0 (+3.9)	50.6 (+3.1)
LabelDistill	BEVDepth	CenterPoint	**41.9 (+6.8)**	**52.8 (+5.3)**
Ours	BEVDepth	CenterPoint	<u>40.5 (+5.4)</u>	<u>51.4 (+3.9)</u>
DistillBEV	BEVDepth	MVP	<u>40.3 (+5.2)</u>	<u>51.0 (+3.5)</u>
Ours	BEVDepth	MVP	**41.5 (+6.4)**	**51.9 (+4.4)**
Backbone: ResNet-101				
BEVDepth	-	-	41.2	53.5
TiG-BEV	BEVDepth	CenterPoint	44.0 (+2.8)	54.4 (+0.9)
X[3]KD	BEVDepth	CenterPoint	44.8 (+3.6)	55.3 (+1.8)
DistillBEV	BEVDepth	CenterPoint	43.6 (+2.4)	53.6 (+0.1)
LabelDistill	BEVDepth	CenterPoint	<u>45.1 (+3.9)</u>	<u>55.3 (+1.8)</u>
Ours	BEVDepth	CenterPoint	**45.5 (+4.3)**	**55.6 (+2.1)**
DistillBEV	BEVDepth	MVP	<u>45.0 (+3.8)</u>	<u>54.7 (+1.2)</u>
Ours	BEVDepth	MVP	**46.6 (+5.4)**	**56.3 (+2.8)**



MFD	CRD	mAP↑	NDS↑
		35.1	47.5
✓		40.4	51.3
✓	✓	**41.5**	**51.9**

Alignment	mAP↑	NDS↑
Upsample-Conv×3	41.0	51.3
AOA	**41.5**	**51.9**

Labels	CRD temperature	mAP↑	NDS↑
GT	-	40.3	51.1
Teacher	1.0	40.9	51.5
	0.5	41.2	51.9
	0.1	**41.5**	**51.9**