LabelDistill: Label-guided Cross-modal Knowledge Distillation for 3D Object Detection

Sanmin Kim, Youngseok Kim, Sihwan Hwang, Hyeonjun Jeong, and Dongsuk Kum

KAIST

Introduction

➢ Limitations in Image-based 3D Object Detection

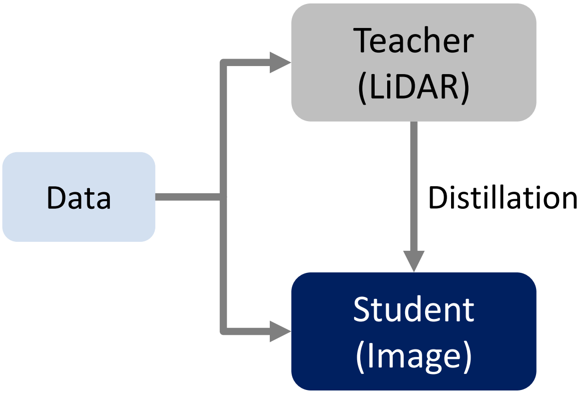
- Insufficient **spatial** information in images
  - Inherent 3D to 2D projection process in images leads to a loss of spatial information.
  - Depth estimation from images entails ambiguity.

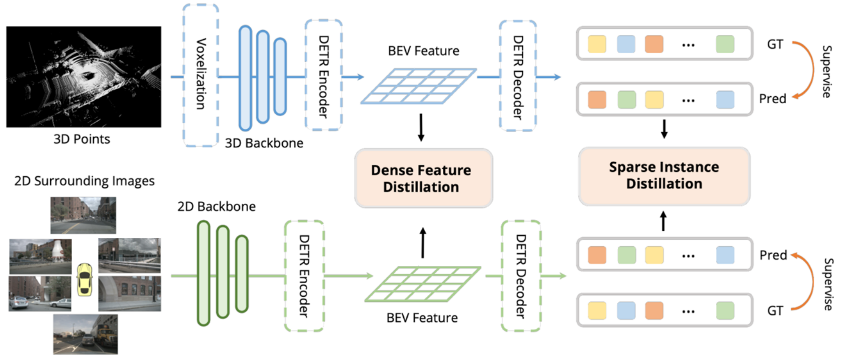
Image 3DOD

Ambiguity in depth estimation

➢ Cross-modal Knowledge Distillation for 3DOD

- Images are lack of spatial information.
- LiDAR point clouds have accurate spatial information.
- **Transferring accurate geometric knowledge** from LiDAR detector to image detector can improve the performance.

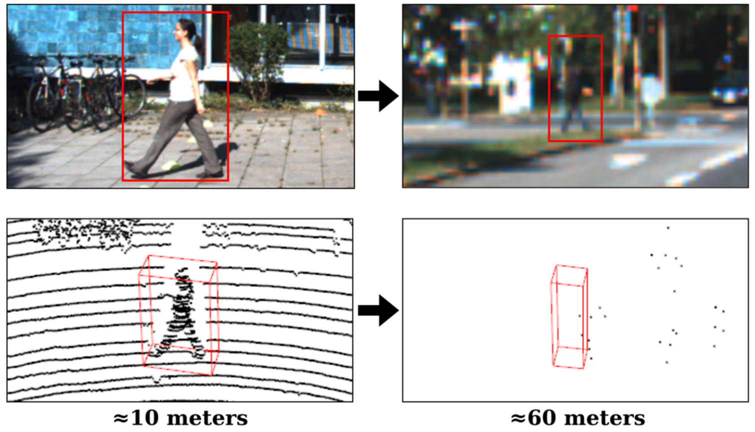




BEVDistill

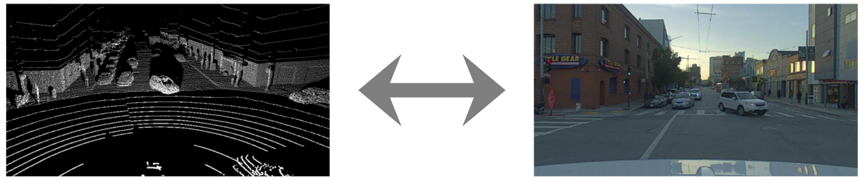
Motivation

- **Technical Challenges in Cross-modality Knowledge Distillation**
  - **Imperfection of LiDAR**
    - LiDAR point clouds contain aleatoric uncertainty
      - Limited range/sparsity
      - sensitivity to weather conditions
    - LiDAR feature can provide erroneous supervision



Source: Faraway-Frustum

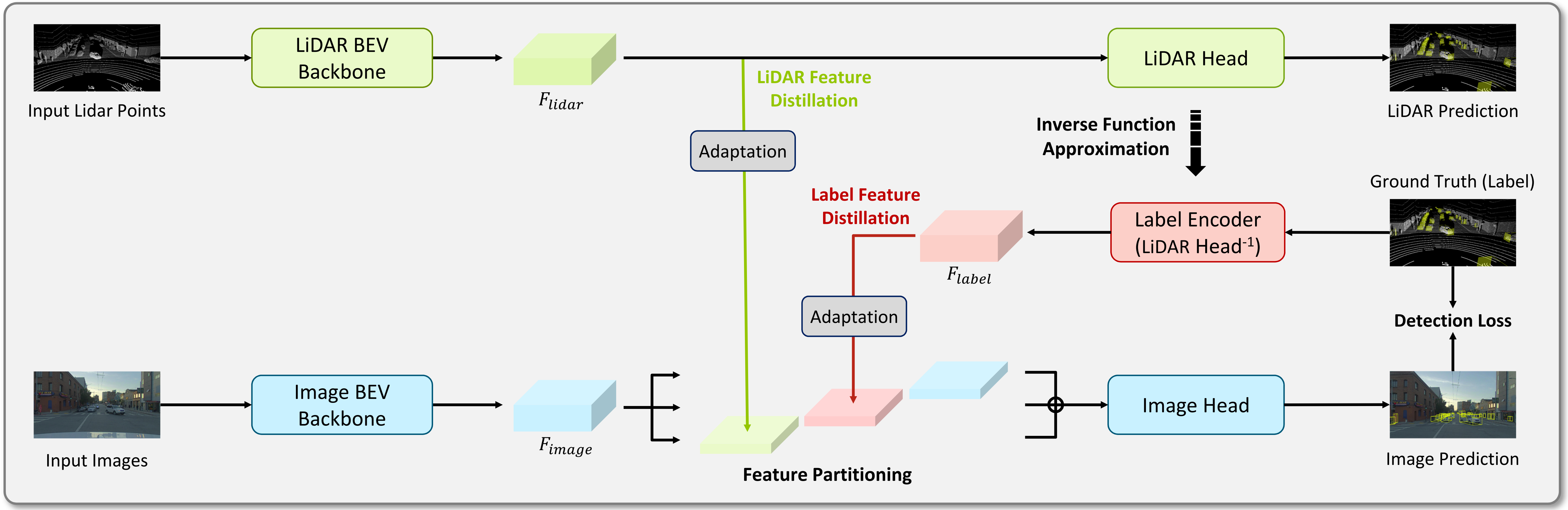
  - **Complementary characteristics in different modalities**
    - Camera and LiDAR have complementary properties
      - **LiDAR – 3D information**
      - **Camera – dense and semantic information**
    - Features can be diverse under different modality
    - Directly pushing the camera model to mimic the LiDAR model can degrade detection performance



Points cloud Image

Method: Overview

➢ **LabelDistill**



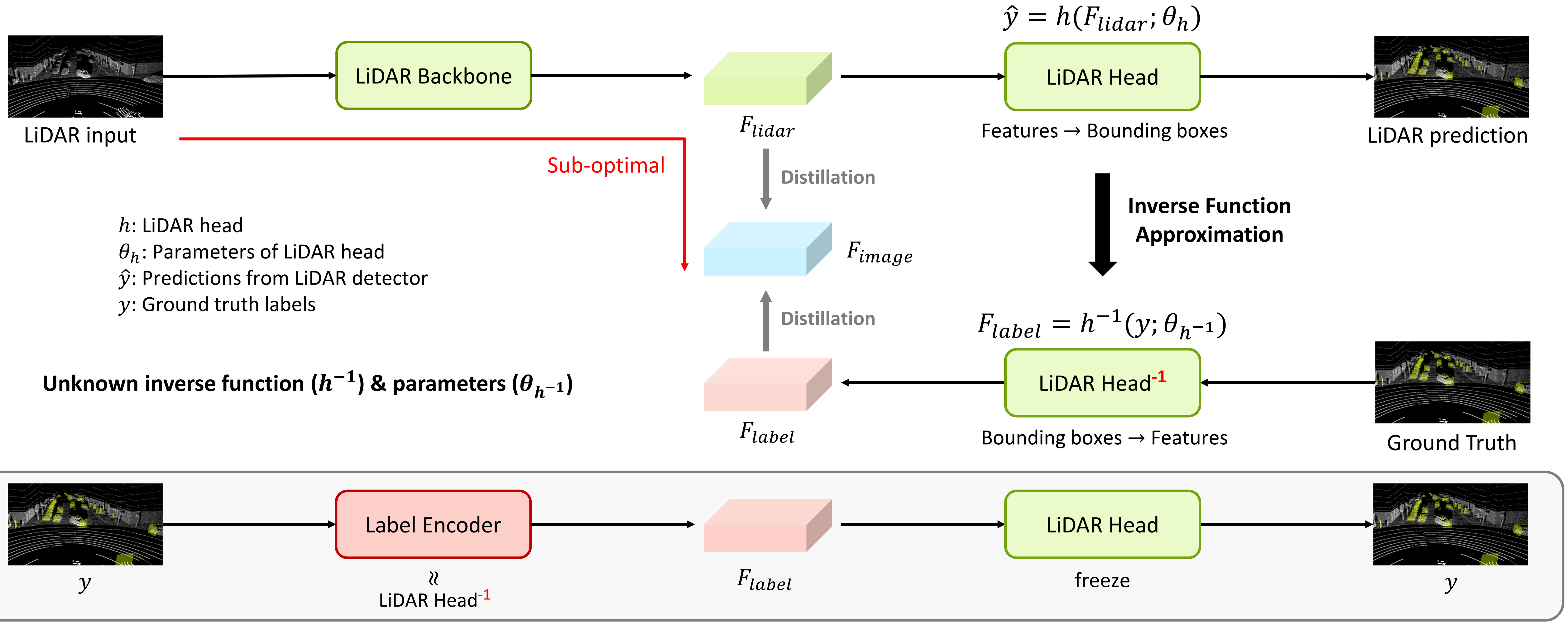
➢ **Main Contribution**

- Overcoming Limitation of LiDAR: **Label Encoder with the Inverse Function of LiDAR Head**
- Preserving Complementary characteristics: **Feature Partitioning**

Method: Label Encoder

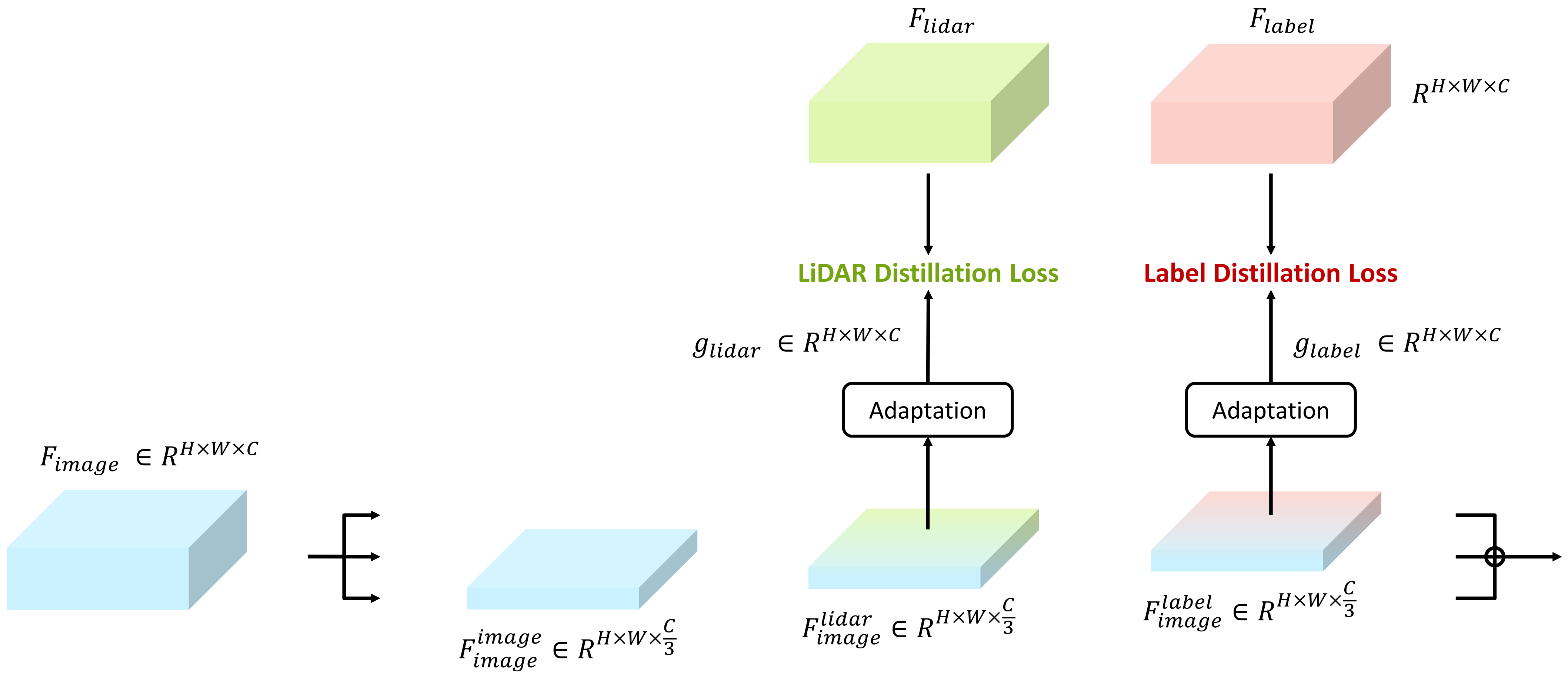
➢ Inverse Function Approximation - Overcoming Limitation of LiDAR data

- Features encoded from LiDAR point cloud are sub-optimal since **aleatoric uncertainty in LiDAR data**
- To handle this problem, we leverage ground truth labels as input to extract **aleatoric uncertainty-free features**



➢ Feature Partitioning - Overcoming Domain Discrepancy

- Train a student network by
  - **Partially following** the knowledge from the teacher network
  - **Partially exploring** for new knowledge that are complementary to the teacher network

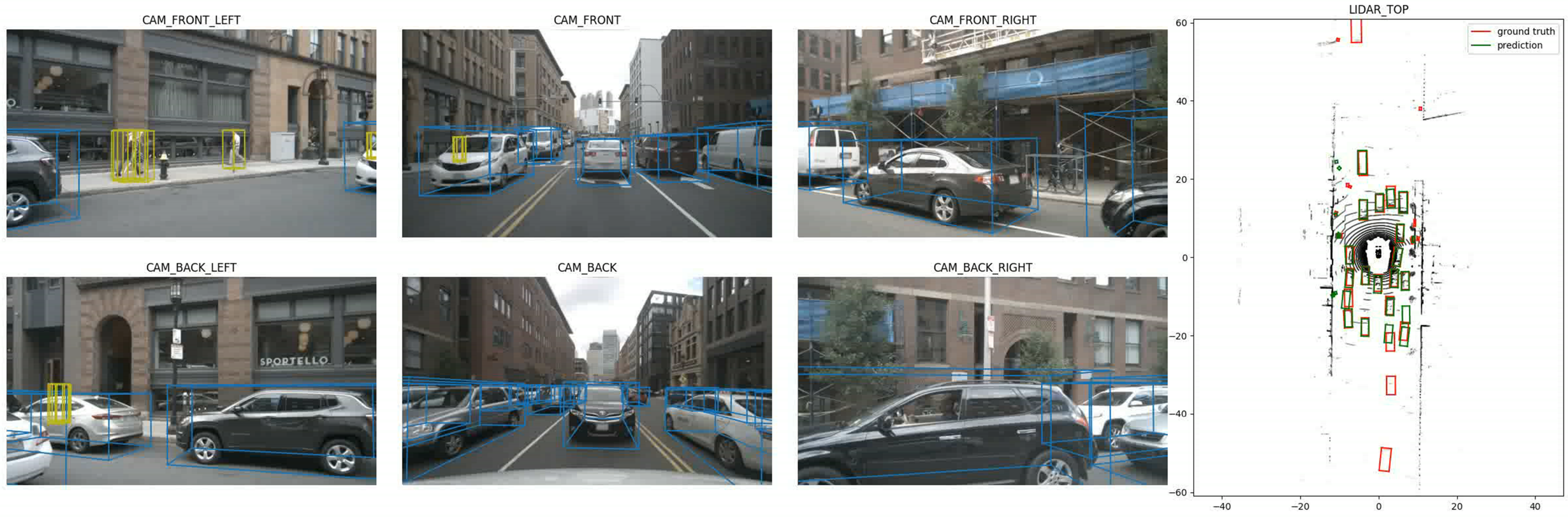


Experiments: Quantitative Results

- **NuScenes *Validation* Set**

Method	Baseline	Backbone	Image Size	mAP (Δ)	NDS (Δ)
UniDistill	BEVDet	ResNet50	256 × 704	29.6 (+3.2)	39.3 (+3.2)
BEVDistill	BEVDepth	ResNet50	256 × 704	33.0 (+1.3)	45.2 (+1.2)
TiG-BEV	BEVDepth	ResNet50	256 × 704	36.6 (+3.7)	46.1 (+3.0)
SimDistill	BEVFusion-C	ResNet50	256 × 704	37.3 (+1.7)	43.8 (+2.6)
X^3KD*	BEVDepth	ResNet50	256 × 704	39.0 (+3.1)	50.5 (+3.3)
DistillBEV*	BEVDepth	ResNet50	256 × 704	40.3 (+3.9)	51.0 (+2.6)
LabelDistill	**BEVDepth**	**ResNet50**	**256 × 704**	**41.9 (+5.1)**	**52.8 (+4.5)**
UVTR	-	ResNet101	900 × 1600	39.2 (+1.3)	48.8 (+0.5)
BEVDistill*	BEVFormer	ResNet101	900 × 1600	41.7 (+1.2)	52.4 (+1.8)
TiG-BEV	BEVDepth	ResNet101	512 × 1408	43.0 **(+2.4)**	51.4 (+2.3)
DistillBEV*	BEVDepth	ResNet101	512 × 1408	45.0 (+2.3)	54.7 (+3.1)
LabelDistill	**BEVDepth**	**ResNet101**	**512 × 1408**	**45.1 (+2.4)**	**55.3 (+3.7)**

*: models trained with CBGS
Δ: improvement from the baseline



Thank you for watching.

Sanmin Kim

VDCLab@KAIST

Contact: sanmin.kim@kaist.ac.kr

LinkedIn **Google Scholar** **Lab Homepage**

KAIST VDCLAB