[LabelDistill: Label-guided Cross-modal Knowledge Distillation for Camera-based 3D Object Detection]

ECCV 2024

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

- Problem/Objective
  - Camera-based 3D Object Detection

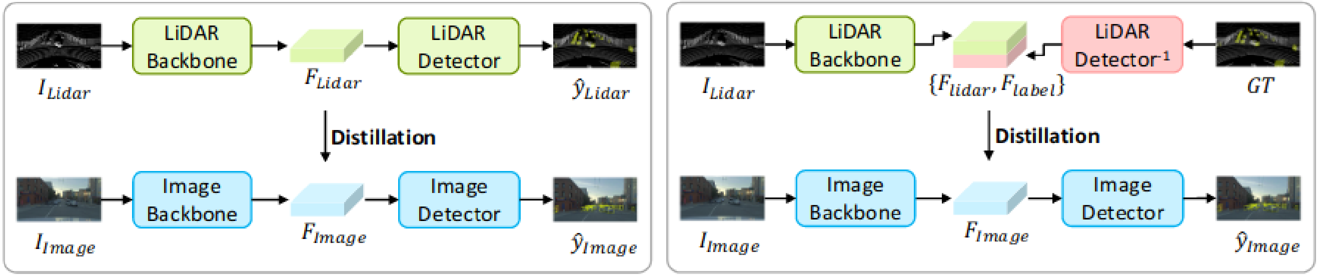
- Contribution/Key Idea
  - Novel label-guided cross-modal knowledge distillation
  - Introduce a feature partitioning
  - Improve performance

김범준

ECCV 2024

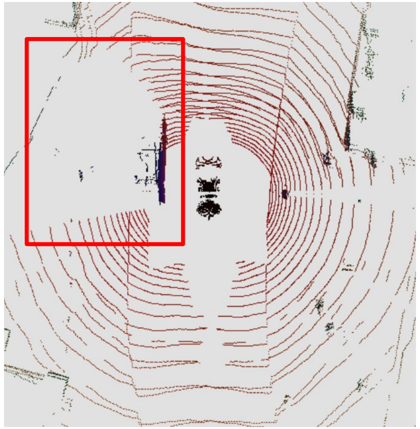
LabelDistill: Label-guided Cross-modal Knowledge Distillation for Camera-based 3D Object Detection

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



(a) Conventional Cross-modal Knowledge Distillation (b) LabelDistill

- 기존 cross-modality knowledge distillation의 문제점
  - Domain gap이 고려되지 않음
  - LiDAR의 거리에 따른 sparse
  - LiDAR occlusion이 고려되지 않음
  - 각 센서의 complementary하게 distill 못함
    → feature partitioning으로 보완



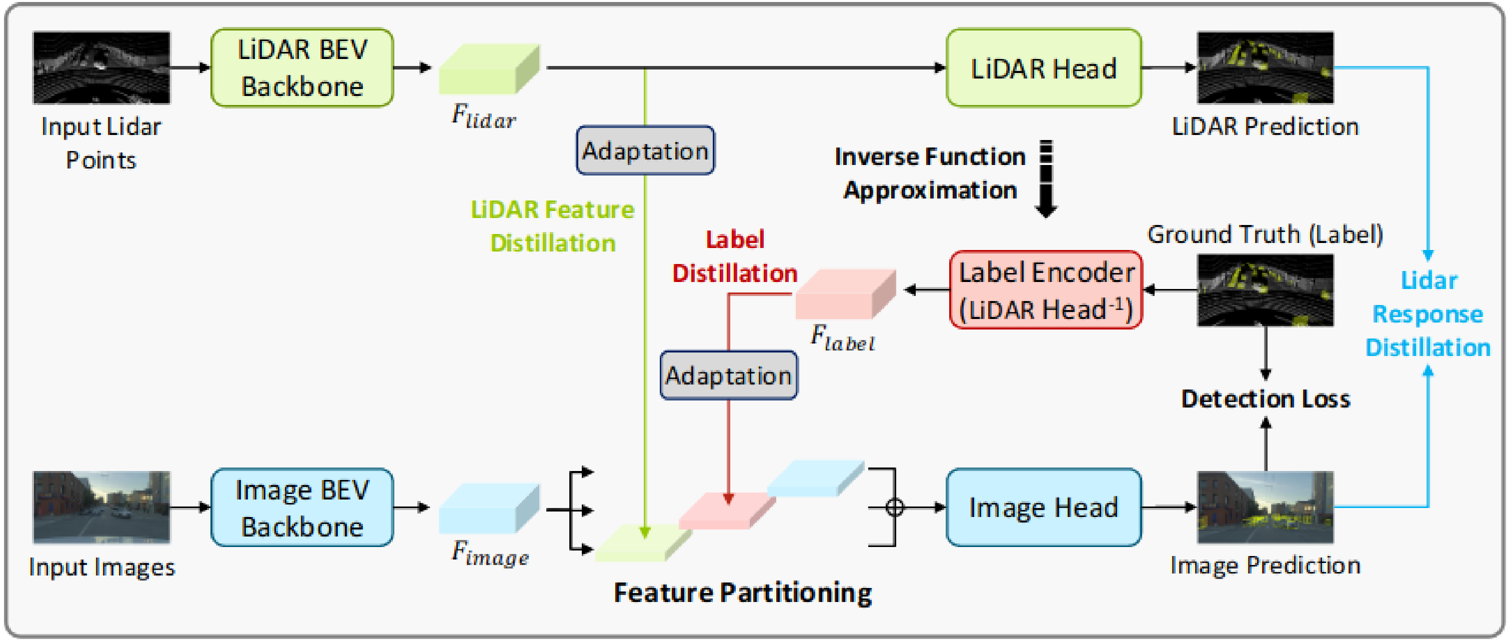
김범준

ECCV 2024

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

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

- **Method**



- Feature-level + Response-level + Label-level Distillation

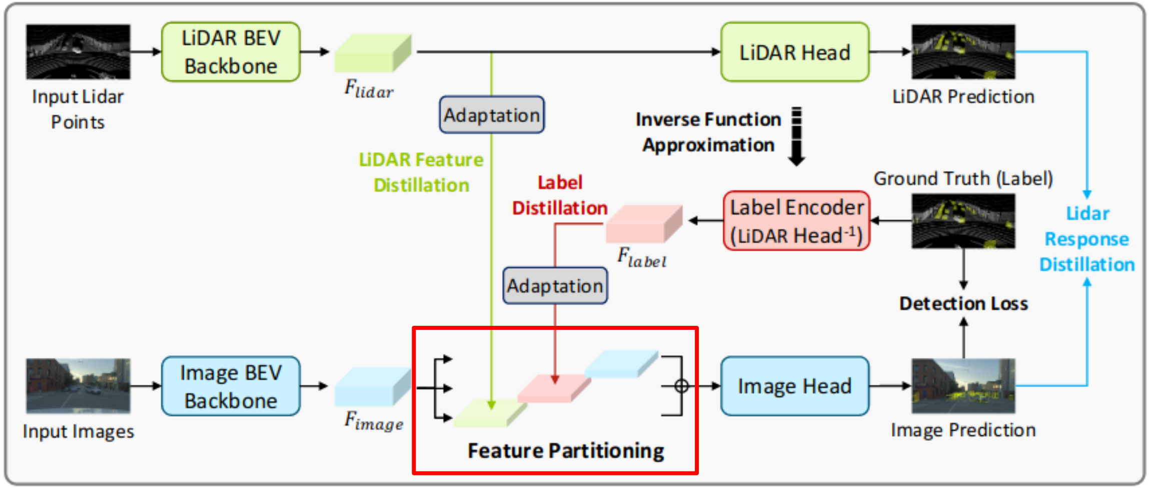
김범준

ECCV 2024

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

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

- **Method**



- Feature Partitioning

$$F_{\text{image}} \in \mathbb{R}^{H \times W \times C} = F_{image}^{image}, F_{image}^{lidar}, F_{image}^{label}$$

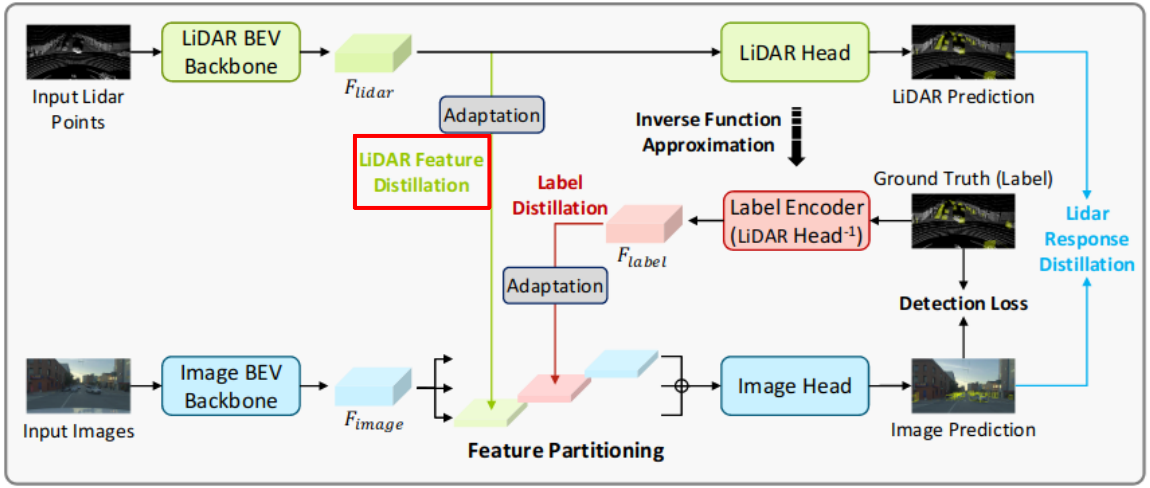
김범준

[LabelDistill: Label-guided Cross-modal Knowledge Distillation for Camera-based 3D Object Detection]()

ECCV 2024

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

- **Method**



- Feature-level distillation

$$\mathcal{L}^{feat}_{lidar} = \frac{1}{N_p}\sum_{i}^{H}\sum_{j}^{W}\mathcal{M}_{ij}\{F^{lidar}_{ij} - \alpha(F^{image}_{ij})\}^2$$

location (i, j)

object-specific mask $\mathcal{M}$

module α, aligns the dimensionality

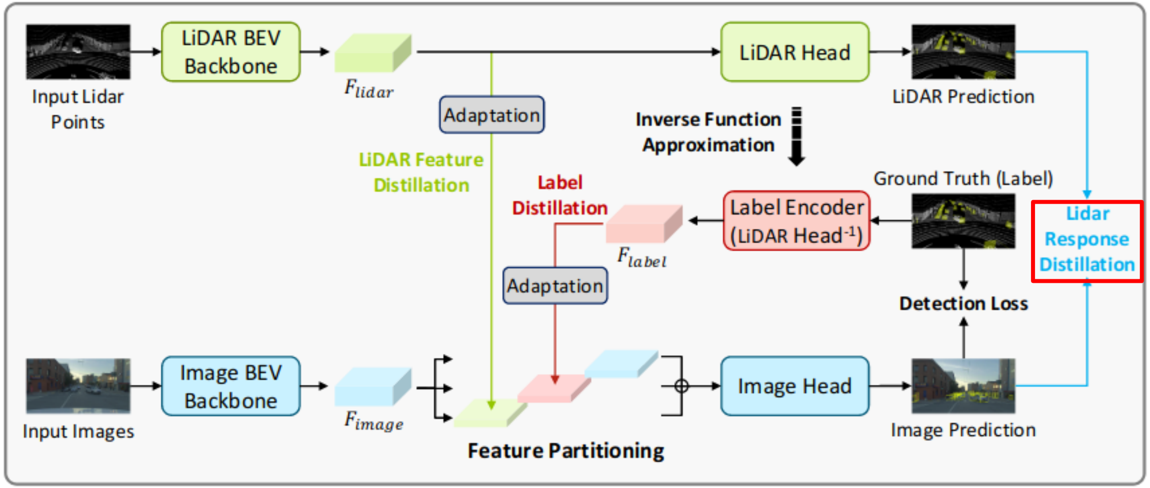
김범준

ECCV 2024

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

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

- **Method**



- Response-level distillation

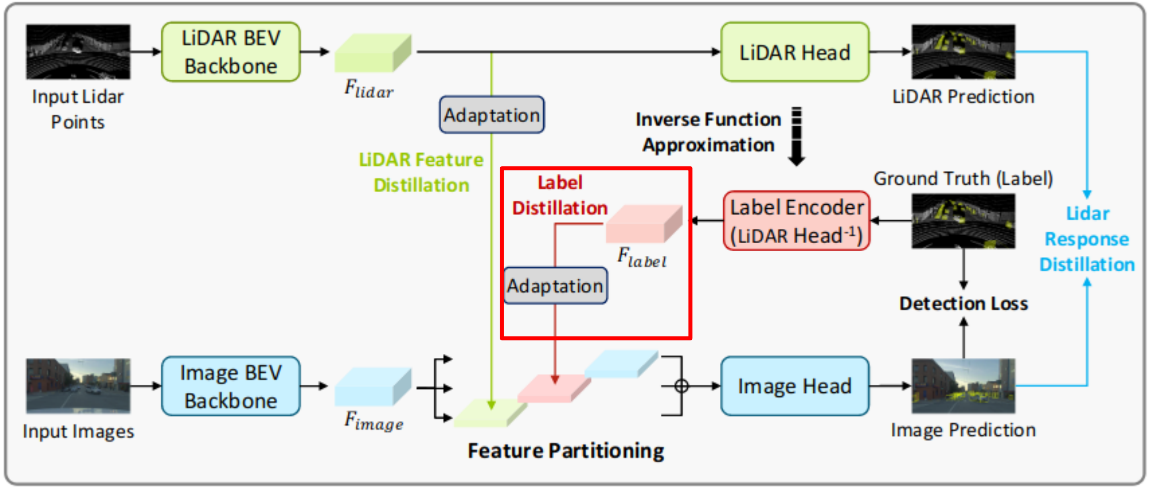
$$\mathcal{L}_{lidar}^{resp} = \mathcal{L}_{cls}(c_{lidar}, c_{image}) + \mathcal{L}_{bbox}(b_{lidar}, b_{image}),$$

c and b denote the class heatmap and bounding box

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

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

- **Method**



- Label-guided distillation

  ⟶기존에는 LiDAR의 높은 detect 성능에 불완전성이 간과됨

  ⟶이를 해결하기 위해 Novel distillation 방법 제시

$$\hat{y} = h(F_{lidar}; \theta_h) \Rightarrow F_{label} = h^{-1}(y; \theta_{h^{-1}})$$

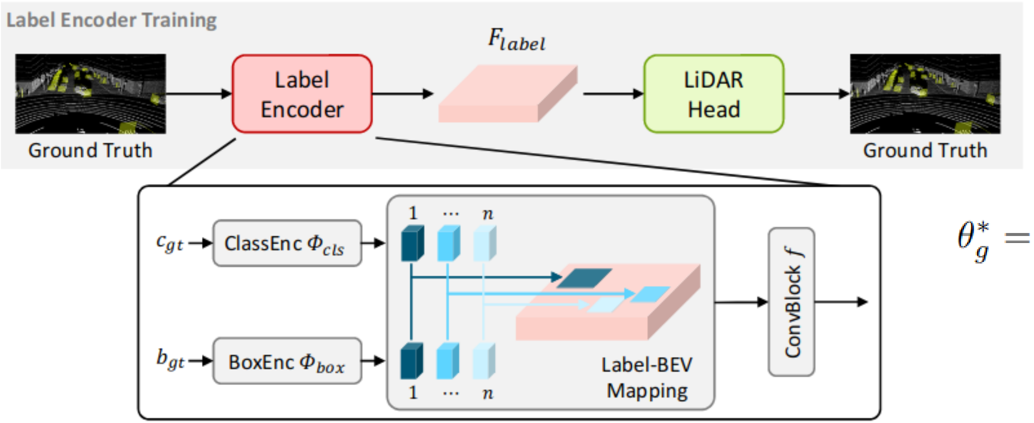
Inverse Function

[LabelDistill: Label-guided Cross-modal Knowledge Distillation for Camera-based 3D Object Detection]

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

ECCV 2024

- **Method - Label-guided distillation**



Pre-trained Head (Decoder)

GT Label

$$\theta_g^* = \underset{\theta_g}{\arg\min} \mathbb{E}_{(I,y)\sim\mathcal{D}} \ \mathcal{L}_{det}\Big(h\big(\underline{g(y;\theta_g)};\theta_h^*\big), y\Big)$$

Label Encoder
(Encoder)

$$g(y;\theta_g) = f\Big(q\big(\Phi_{cls}(c_{gt}) + \Phi_{box}(b_{gt})\big)\Big)$$

- NN의 non-linearity로 역함수 계산 어려워 **근사**하여 **별도 학습**
  - 기존 Autoencoder 방식을 차용
    → 차이점은 scratch로부터 학습x, decoder를 pre-trained 사용

[LabelDistill: Label-guided Cross-modal Knowledge Distillation for Camera-based 3D Object Detection]

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

- ## Method - Label-guided distillation

Pre-trained Head (Decoder)

GT Label

$$\theta_g^* = \arg\min_{\theta_g} \mathbb{E}_{(I,y)\sim\mathcal{D}}\ \mathcal{L}_{det}\Big(h\big(\underline{g(y;\theta_g)};\theta_h^*\big), y\Big)$$

Label Encoder (Encoder)

$$g(y;\theta_g) = f\Big(q\big(\Phi_{cls}(c_{gt}) + \Phi_{box}(b_{gt})\big)\Big)$$

Table 1: Evaluation of the autoencoder consists of the label encoder and LiDAR detection head on nuScenes validation set.

	mAP ↑	NDS ↑	mATE ↓	mAOE ↓	mAVE ↓
Label Encoder + LiDAR Head	94.14	90.25	0.192	0.048	0.128

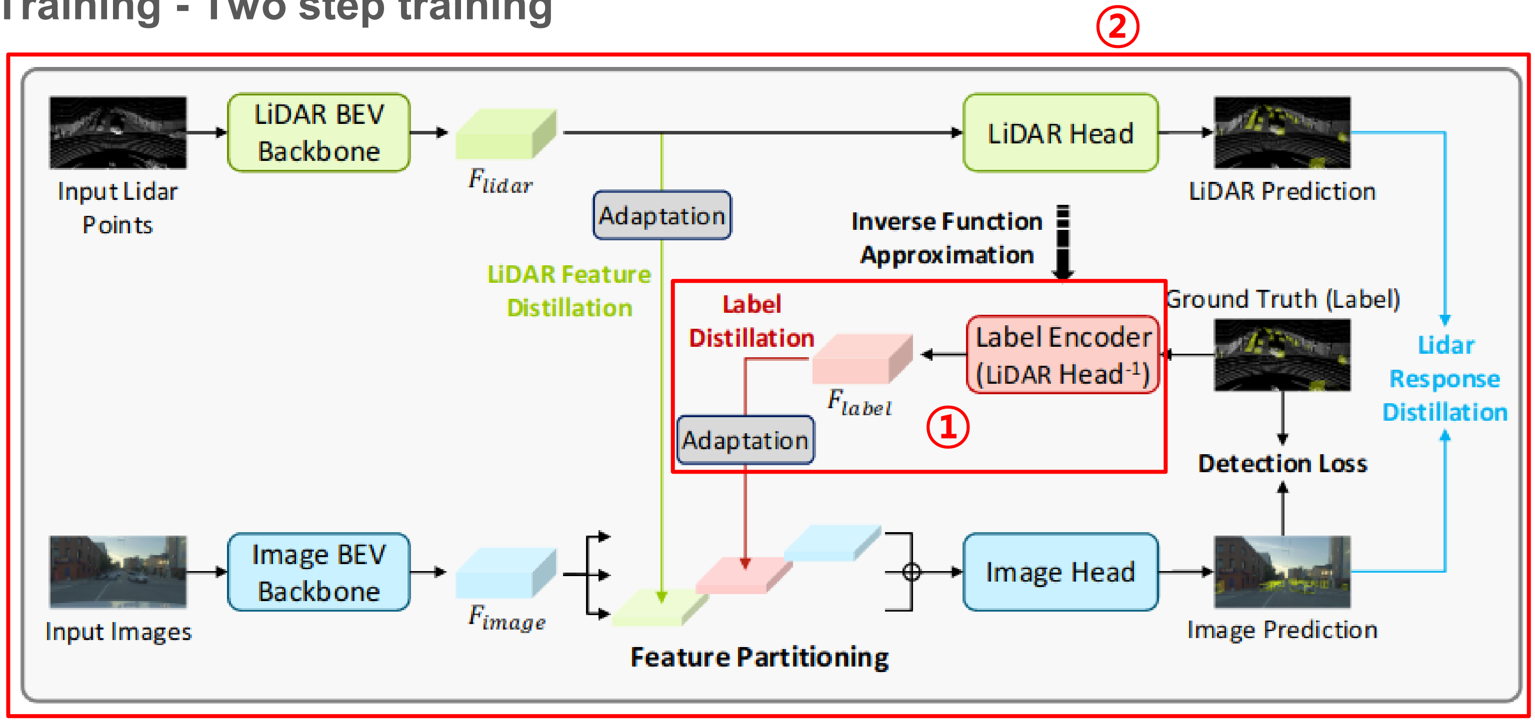
Table 6: Evaluation on the effectiveness of the inverse function approximation. AutoEncoder trains the label encoder and the detection head from the scratch.

Label Encoder Training	mAP ↑	NDS ↑	mATE ↓	mASE ↓	mAOE ↓
AutoEncoder	34.9	46.7	0.656	0.270	0.476
LabelEnc [14]	34.8	46.8	0.658	0.267	0.479
Inverse Function Approximation	**36.8**	**48.1**	**0.646**	**0.263**	**0.474**

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

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

- **Training - Two step training**



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

ECCV 2024

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

- ## Experiments

Set	Method	Backbone	Size	mAP	NDS	mATE	mASE	mAOE	mAVE	mAAE
Validation	BEVDet4D [17]	ResNet50	256×704	32.3	45.3	0.674	0.272	0.503	0.429	0.208
	BEVDepth [29]	ResNet50	256×704	33.3	44.1	0.683	0.276	0.545	0.526	0.226
	BEVStereo [28]	ResNet50	256×704	34.4	44.9	0.659	0.276	0.579	0.503	0.216
	VEDet† [4]	ResNet50	384×1056	34.7	44.3	0.726	0.282	0.542	0.555	0.198
	PETR v2 [36]	ResNet50	256×704	34.9	45.6	0.700	0.275	0.580	0.437	**0.187**
	FB-BEV† [31]	ResNet50	256×704	35.0	47.9	0.642	0.275	0.459	0.391	0.193
	AeDet† [10]	ResNet50	256×704	35.8	47.3	0.655	0.273	0.493	0.427	0.216
	P2D [24]	ResNet50	256×704	37.4	48.6	0.631	0.272	0.508	0.384	0.212
	BEVFormer v2† [61]	ResNet50	640×1600	38.8	49.8	0.679	0.276	0.417	0.403	0.189
	SOLOFusion [45]	ResNet50	256×704	40.6	49.7	0.609	0.284	0.650	**0.315**	0.204
	LabelDistill	ResNet50	256×704	**41.9**	**52.8**	**0.582**	**0.258**	**0.413**	0.346	0.220
Validation	DETR3D† [56]	ResNet101	900×1600	34.9	43.4	0.716	0.268	0.379	0.842	0.200
	BEVDepth [29]	ResNet101	512×1408	40.6	49.0	0.626	0.278	0.513	0.489	0.226
	BEVFormer [30]	ResNet101	900×1600	41.6	51.7	0.673	0.274	0.372	0.394	0.198
	VEDet† [4]	ResNet101	512×1408	43.2	52.0	0.638	0.275	0.362	0.498	0.191
	PolarFormer [23]	ResNet101	900×1600	43.2	52.8	0.648	0.270	0.348	0.409	0.201
	P2D [24]	ResNet101	512×1408	43.3	52.8	0.619	0.265	0.432	0.364	0.211
	Sparse4D [33]	ResNet101	900×1600	43.6	54.1	0.633	0.279	0.363	**0.317**	**0.177**
	LabelDistill	ResNet101	512×1408	**45.1**	**55.3**	**0.579**	**0.252**	**0.331**	0.357	0.207
Test	BEVDepth* [29]	ConvNeXt-B	900×1600	47.5	56.1	0.474	0.259	0.463	0.432	**0.134**
	LabelDistill	ConvNeXt-B	900×1600	**52.6**	**61.0**	**0.443**	**0.241**	**0.339**	**0.370**	0.136

김범준

LabelDistill: Label-guided Cross-modal Knowledge Distillation for Camera-based 3D Object Detection
Sanmin Kim, Youngseok Kim∗, Sihwan Hwang, Hyeonjun Jeong, and Dongsuk Kum

- **Experiments**

Table 3: Comparison to other LiDAR-guided cross-modal knowledge distillation strategies. †: methods with CBGS.

Model	Baseline	Image Size	Backbone	mAP (Δ)	NDS (Δ)
UniDistill [71]	BEVDet	704×256	ResNet50	29.6 (3.2)	39.3 (3.2)
BEVDistill [6]	BEVDepth	704×256	ResNet50	33.0 (1.3)	45.2 (1.2)
TiG-BEV [20]	BEVDepth	704×256	ResNet50	36.6 (3.7)	46.1 (3.0)
BEVSimDet [70]	BEVFusion-C	704×256	ResNet50	37.3 (1.7)	43.8 (2.6)
$X^3KD^\dagger$ [25]	BEVDepth	704×256	ResNet50	39.0 (3.1)	50.5 (3.3)
DistillBEV† [59]	BEVDepth	704×256	ResNet50	40.3 (3.9)	51.0 (2.6)
LabelDistill	BEVDepth	704×256	ResNet50	**41.9 (5.1)**	**52.8 (4.5)**
UVTR [27]	-	1600×900	ResNet101	39.2 (1.3)	48.8 (0.5)
BEVDistill† [6]	BEVFormer	1600×900	ResNet101	41.7 (1.2)	52.4 (1.8)
TiG-BEV [20]	BEVDepth	1408×512	ResNet101	43.0 **(2.4)**	51.4 (2.3)
DistillBEV† [59]	BEVDepth	1408×512	ResNet101	45.0 (2.3)	54.7 (3.1)
LabelDistill	BEVDepth	1408×512	ResNet101	**45.1 (2.4)**	**55.3 (3.7)**

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

- ## Experiments

Table 4: Ablation study on the proposed method. LiDAR, Label, and Partition represent LiDAR distillation, label distillation, and feature partitioning, respectively.

	LiDAR	Label	Partition	mAP ↑	NDS ↑	mATE ↓	mASE ↓
(a)				33.6	44.8	0.694	0.273
(b)	✓			35.4	48.6	0.648	0.262
(c)	✓	✓		37.0	49.5	0.663	0.258
(d)	✓	✓	✓	**37.9**	**50.1**	**0.641**	**0.256**

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

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

- ## Experiments

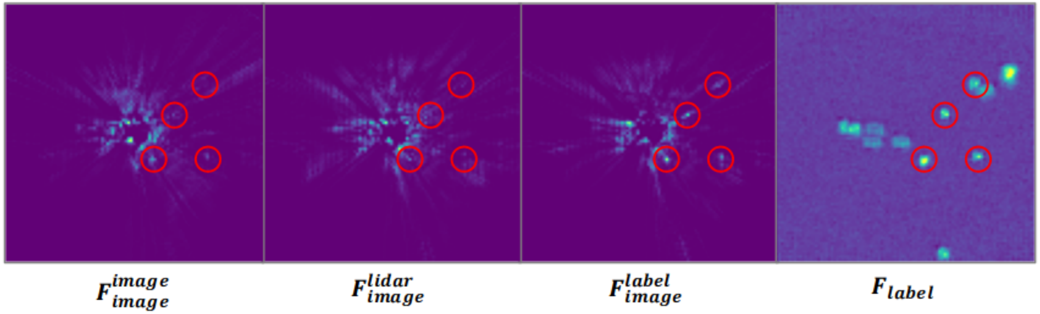


Fig. 4: Illustration of BEV feature maps in the inference stage. F_{image}^{image} is undistilled image feature, F_{image}^{lidar} is lidar-distilled image feature, and F_{image}^{label}, label-distilled image feature, and F_{label} denotes label feature from the label encoder.

Table 5: Experiments on different channel ratio for the feature partitioning.

Channel Ratio			mAP ↑	NDS ↑	mATE ↓	mASE ↓
F_{lidar}^{image}	F_{label}^{image}	F_{image}^{image}				
1	3	2	36.6	48.8	0.655	0.260
3	1	2	37.1	49.4	0.646	0.258
2	2	2	**37.6**	**49.6**	**0.643**	**0.256**

[LabelDistill: Label-guided Cross-modal Knowledge Distillation for Camera-based 3D Object Detection]
Sanmin Kim, Youngseok Kim, Sihwan Hwang, Hyeonjun Jeong, and Dongsuk Kum*

Experiments

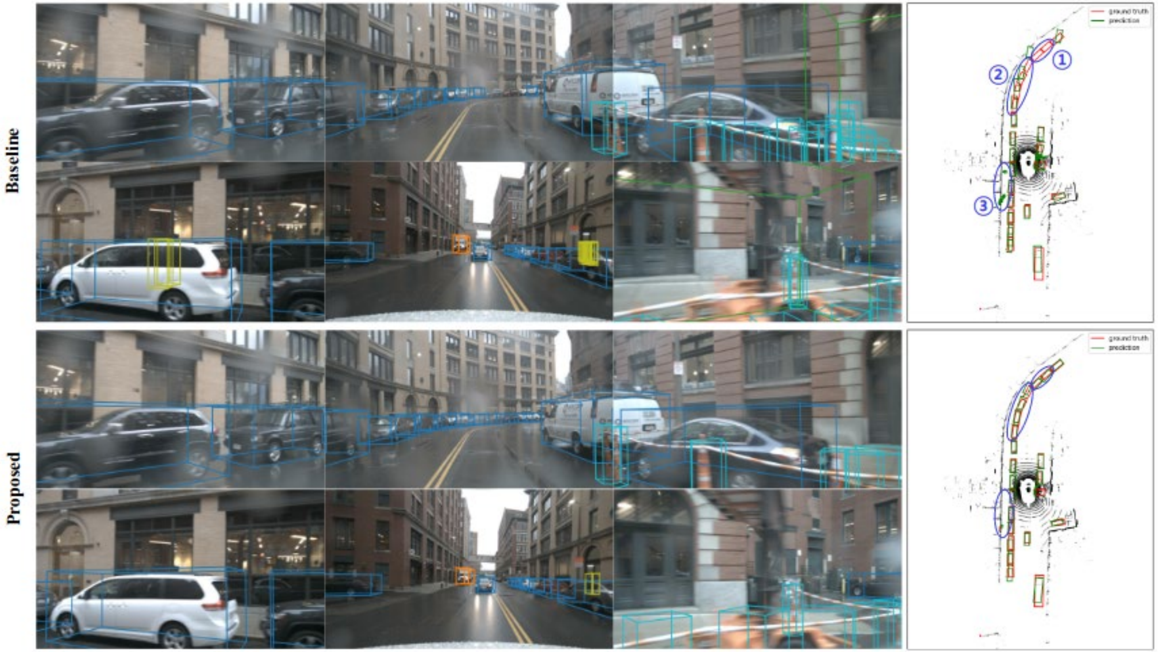


Fig. 5: Comparison of the baseline (BEVDepth) and our approach. The blue circles in the BEV view highlight cases that demonstrate the advantages of our approach, including: 1) higher recall, 2) more accurate localization, and 3) fewer false positives.