LiDAR2Map: In Defense of LiDAR-Based Semantic Map Construction Using Online Camera Distillation

Song Wang, Wentong Li, Wenyu Liu, Xiaolu Liu,Jianke Zhu*

Zhejiang University

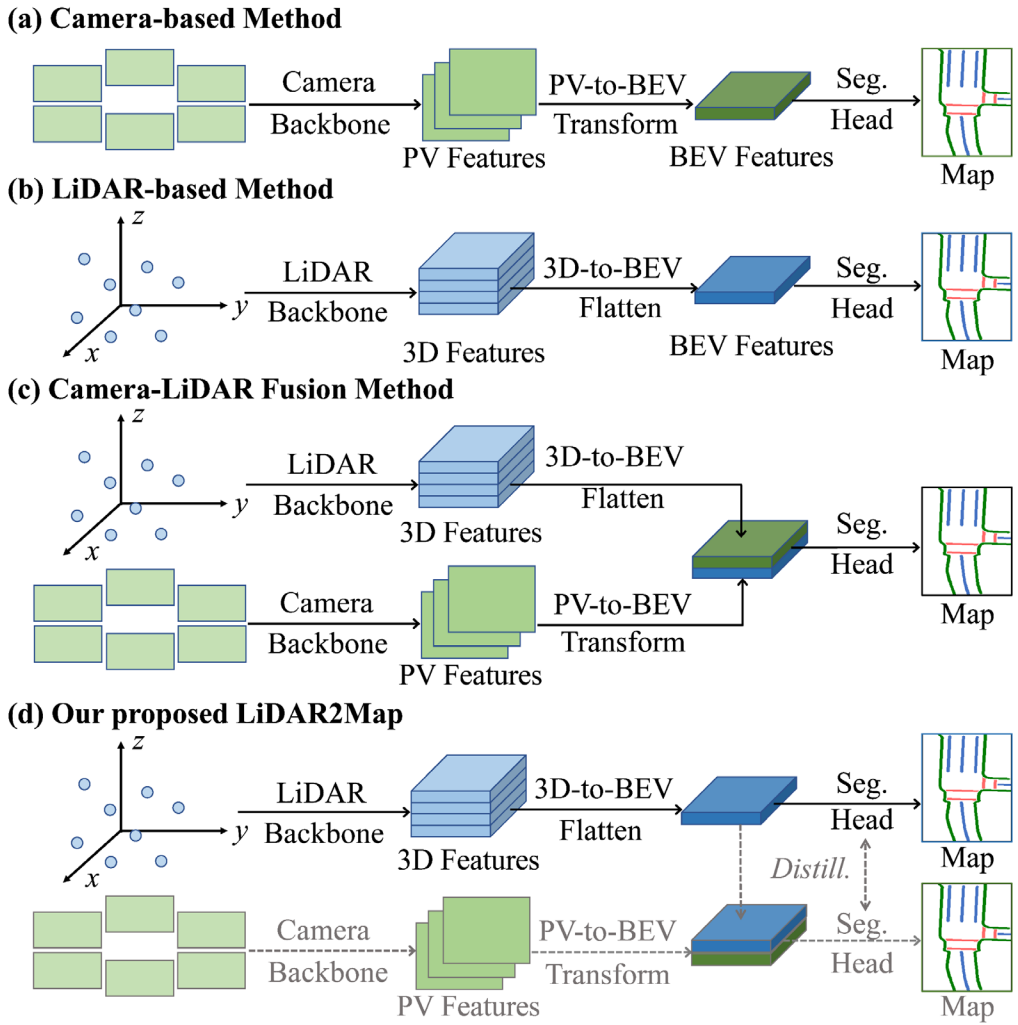
https://github.com/songw-zju/LiDAR2Map

Poster: TUE-PM-100

Contact: {songw, liwentong, liuwenyu.lwy, xiaoluliu, jkzhu}@zju.edu.cn

Overview

- An efficient framework LiDAR2Map with BEV feature pyramid decoder for semantic map construction.
- An effective online Camera-to-LiDAR distillation scheme that performs both feature-level and logit-level distillation.
- Our method achieves the state-of-the-art performance on semantic map construction including map and vehicle segmentation.

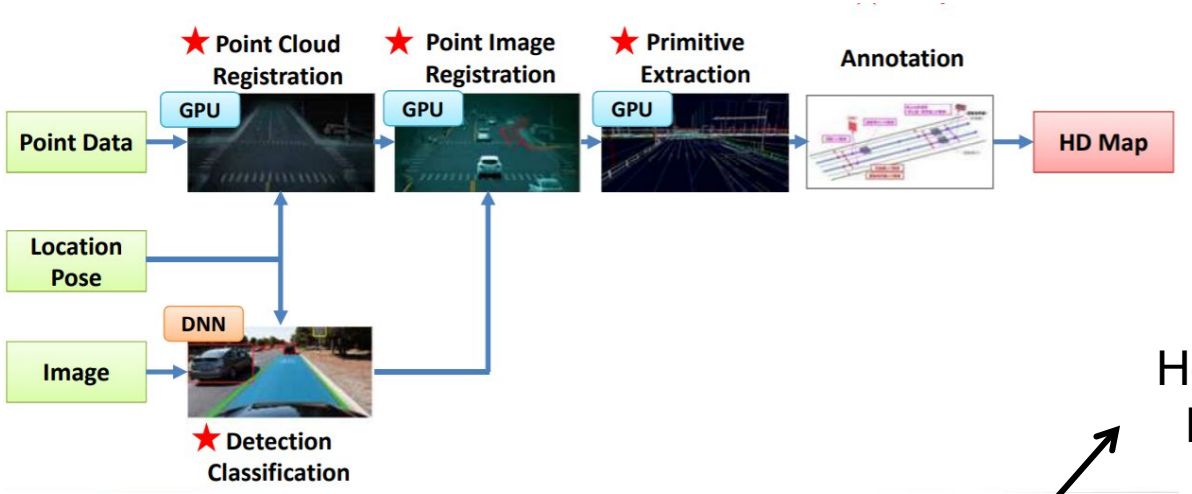


01

Motivation

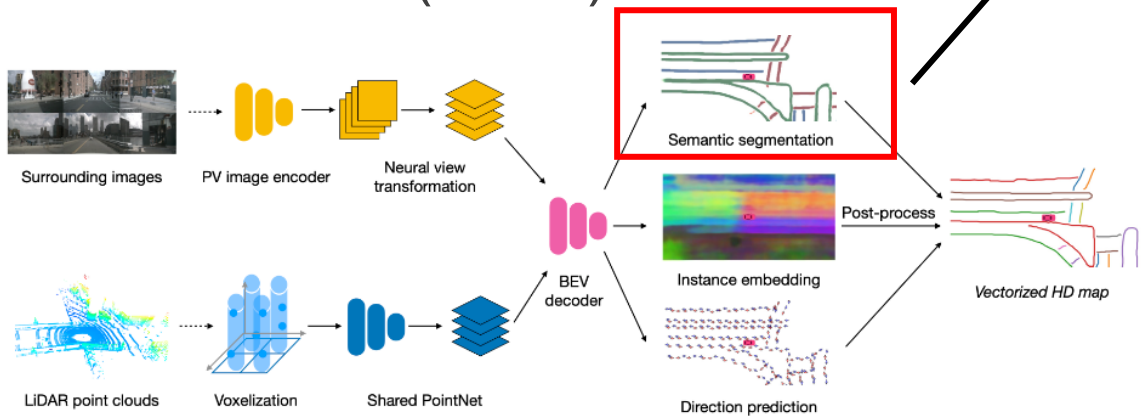
Motivation

➢ Traditional HD Map Construction (Global)



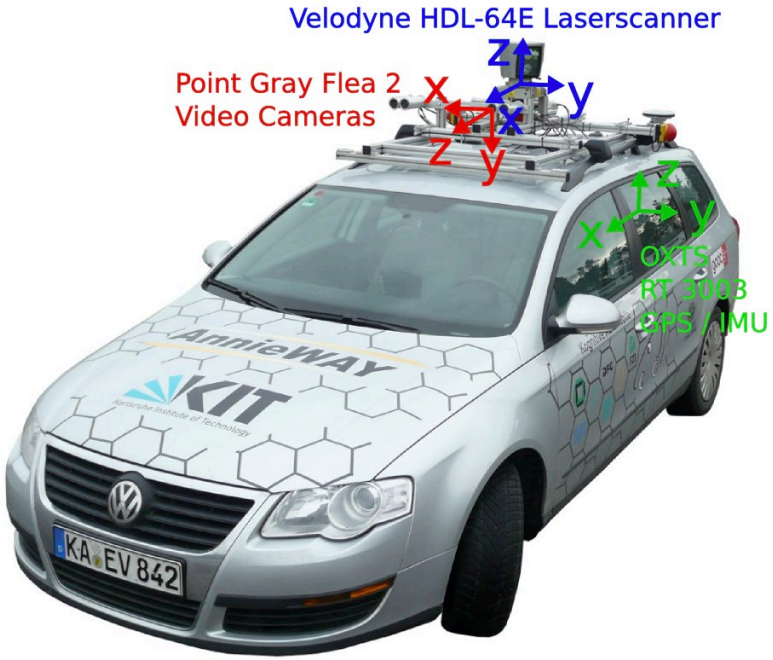
How to obtain accurate BEV semantic maps more effectively?

➢ Online HD Map Construction (Local)

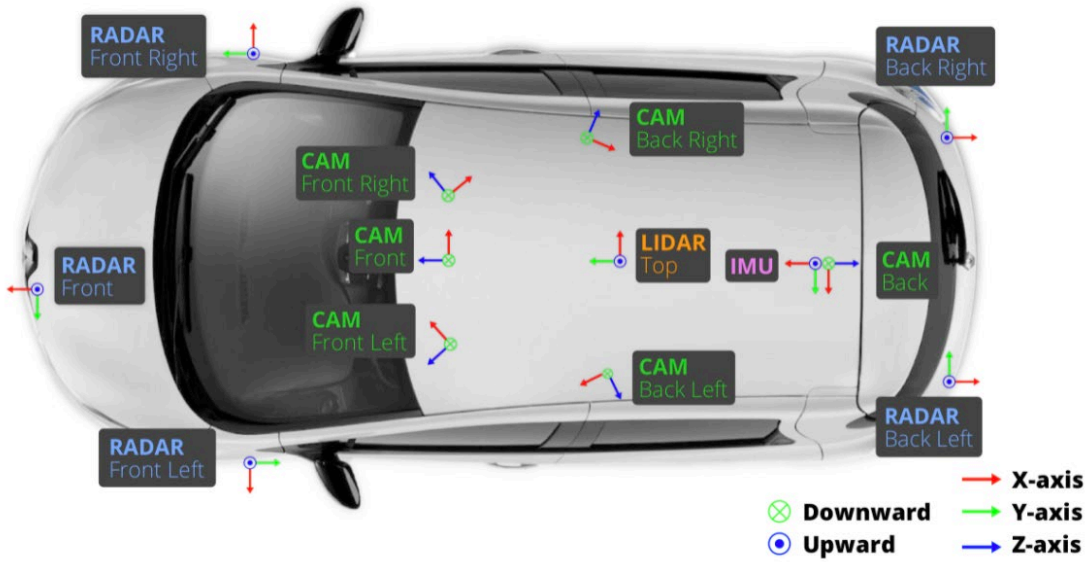


Li Q, Wang Y, Wang Y, et al. Hdmapnet: An online hd map construction and evaluation framework. ICRA 2022.

Motivation



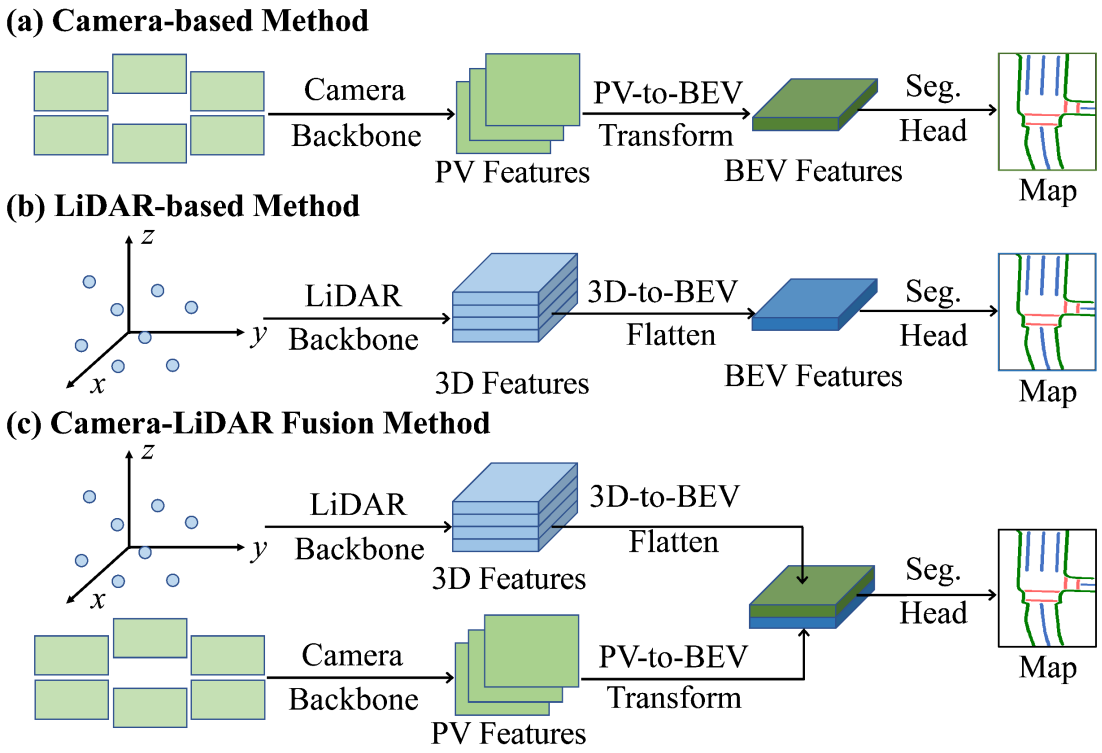
KITTI, CVPR2012



nuScenes, CVPR2020

- LiDAR-based methods are widely explored in 3D object detection while it is rarely investigated in semantic map construction.

Motivation



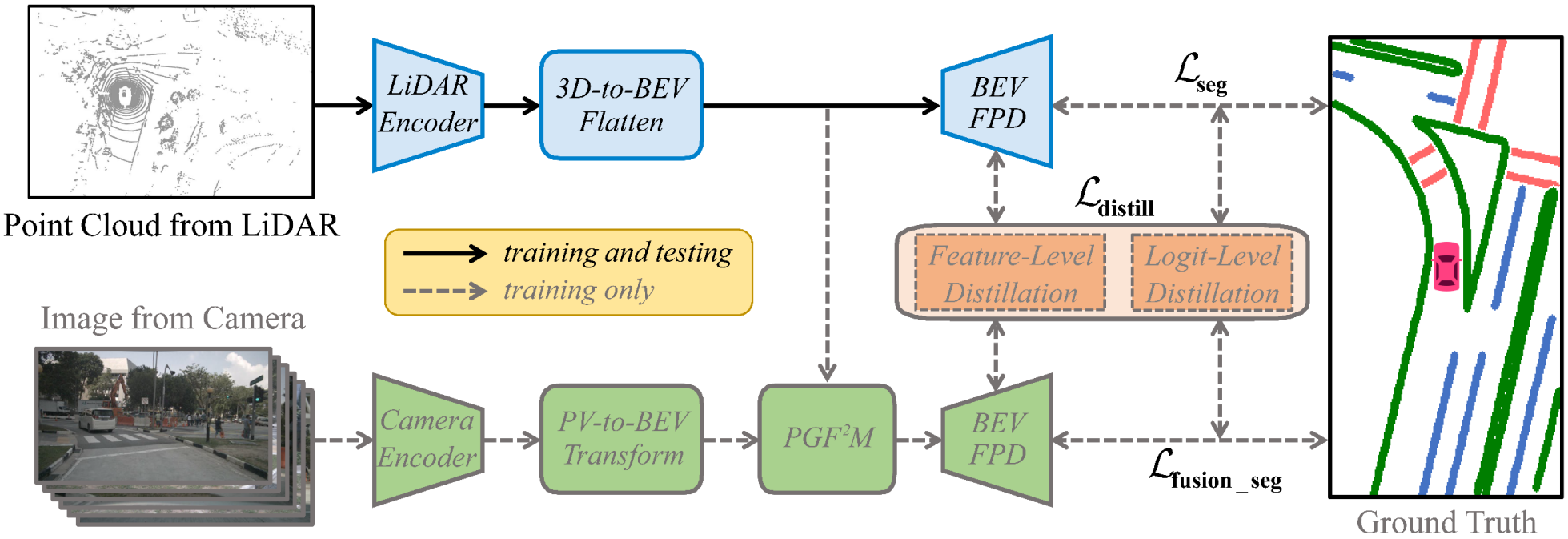
- Camera-based methods make full use of multi-view images with the enriched semantic information.
- LiDAR outputs the accurate 3D spatial information that can be used to project the captured features onto the BEV space.
- We intend to construct the semantic map from LiDAR point cloud effectively.

02

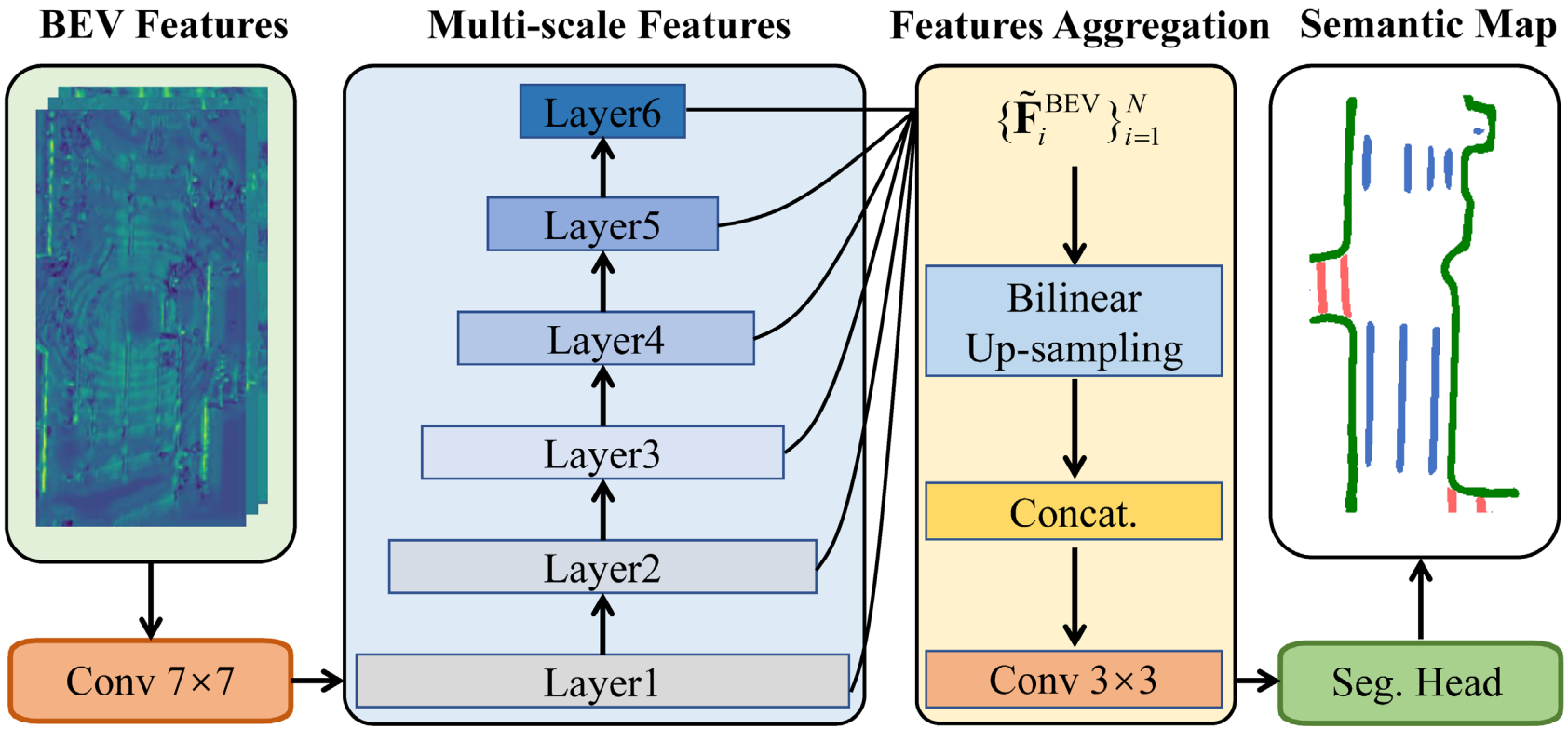
Method

Our Framework

- Design a **BEV feature pyramid decoder** can learn the robust BEV feature representations to boost the LiDAR-based model
- Propose an effective online **Camera-to-LiDAR distillation** scheme that performs both feature-level and logit-level distillation



BEV Feature Pyramid Decoder



- As the layer number increases, the corresponding BEV features can better capture the robust spatial features with accurate responses.

Online Camera-to-LiDAR Distillation

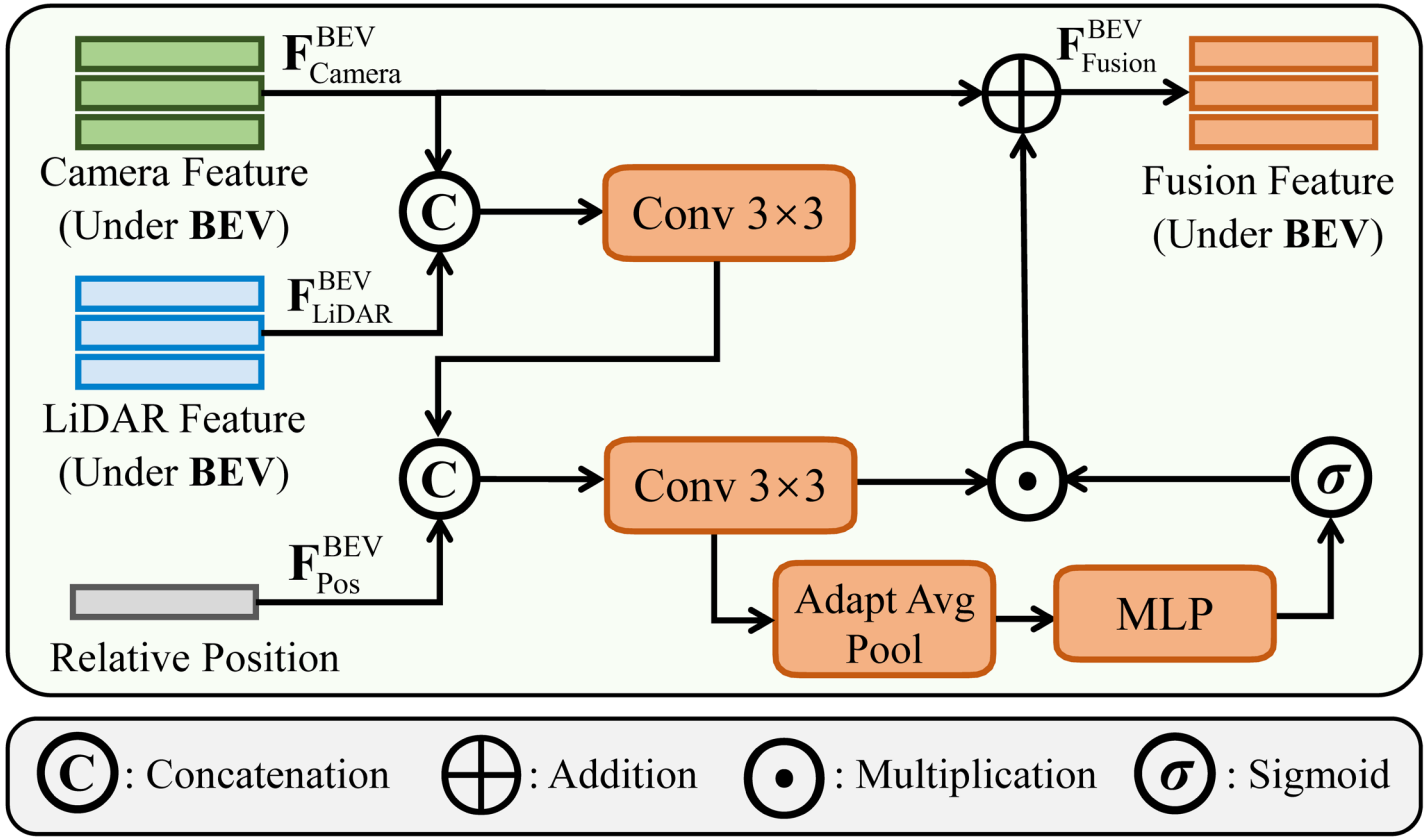
- The module consists of three components, including Position-Guided Feature Fusion Module (*PGF2M*), Feature-level Distillation (*FD*) and Logit-level Distillation (*LD*).

- We suggest the online distillation to make the LiDAR-based "Student" model learn from the Camera-LiDAR fusion model as a "Teacher".

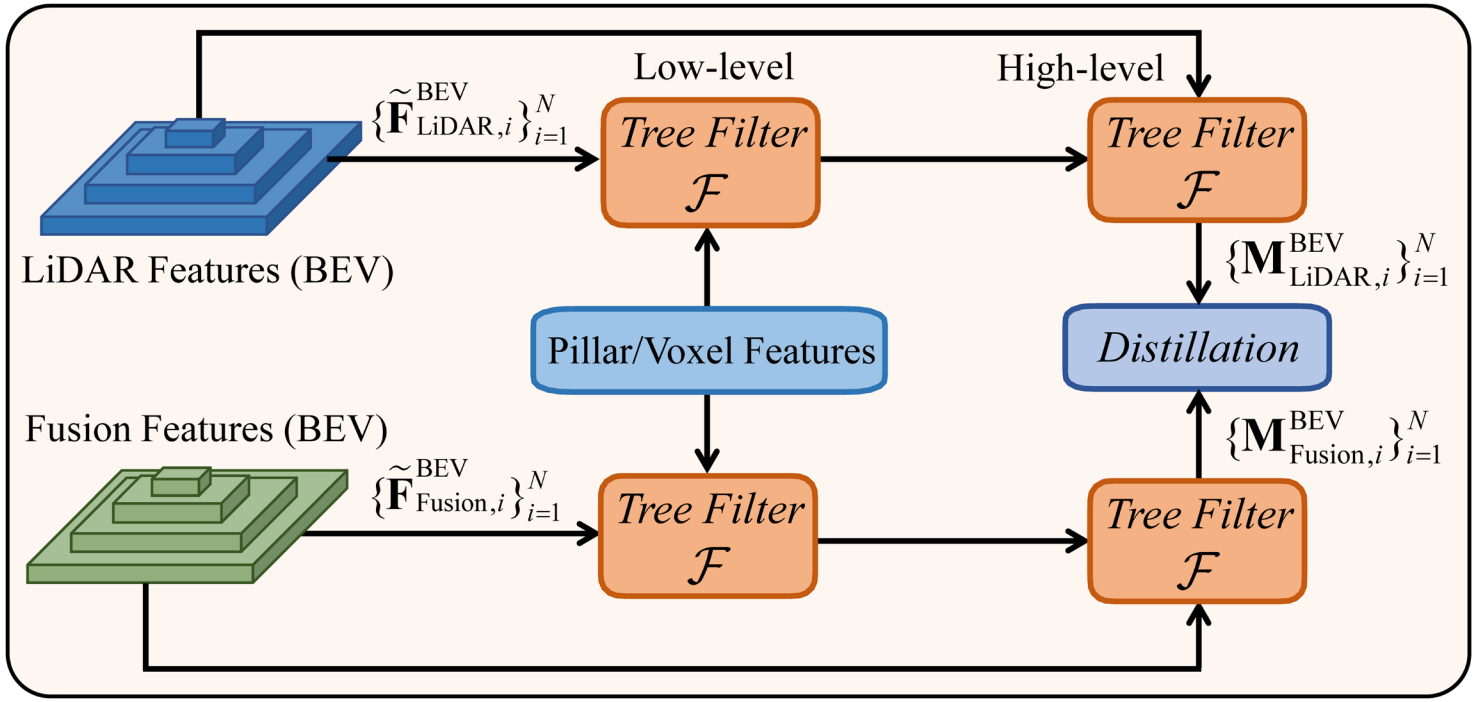
$$\mathcal{L}_{\mathbf{distill}} = \mathcal{L}_{\mathbf{feature}} + \mathcal{L}_{\mathbf{logit}}$$

$$\mathcal{L}_{\mathbf{logit}} = \mathbf{D}_{\mathrm{KL}}\left(\mathbf{P}^{\mathrm{BEV}}_{\mathrm{Fusion}} \| \mathbf{P}^{\mathrm{BEV}}_{\mathrm{LiDAR}}\right)$$

Position-Guided Feature Fusion Module



Feature-level Distillation



- We employ the tree filter as the transform function to model the long-range dependencies of dense BEV features in each modality by minimal spanning tree.

03

Experiments

Performance Comparison

—— Map Segmentation

Method	Image Size	Modality	Backbone	Divider	Ped Crossing	Boundary	mIoU
VPN* [30]	352×128	Camera	EfficientNet-B0 [41]	36.5	15.8	35.6	29.3
Lift-Splat* [32]	352×128	Camera	EfficientNet-B0	38.3	14.9	39.3	30.8
HDMapNet-Camera [20]	352×128	Camera	EfficientNet-B0	40.6	18.7	39.5	32.9
BEVSegFormer [31]	800×448	Camera	ResNet-101	51.1	32.6	50.0	44.6
BEVFormer† [21]	1600×900	Camera	ResNet-50	53.0	36.6	54.1	47.9
BEVerse [53]	1408×512	Camera	Swin-Tiny	56.1	44.9	58.7	53.2
UniFusion [33]	1600×900	Camera	Swin-Tiny	58.6	43.3	59.0	53.6
HDMapNet-Fusion [20]	352×128	Camera & LiDAR	EfficientNet-B0 & PointPillars	46.1	31.4	56.0	44.5
HDMapNet-LiDAR [20]	-	LiDAR	PointPillars	26.7	17.3	44.6	29.5
LiDAR2Map	-	LiDAR	PointPillars	60.4	45.5	66.4	57.4
LiDAR2Map	-	LiDAR	VoxelNet	**61.5**	**46.3**	**68.1**	**58.6**

On the validation set of nuScenes with the 60m × 30m setting for map segmentation

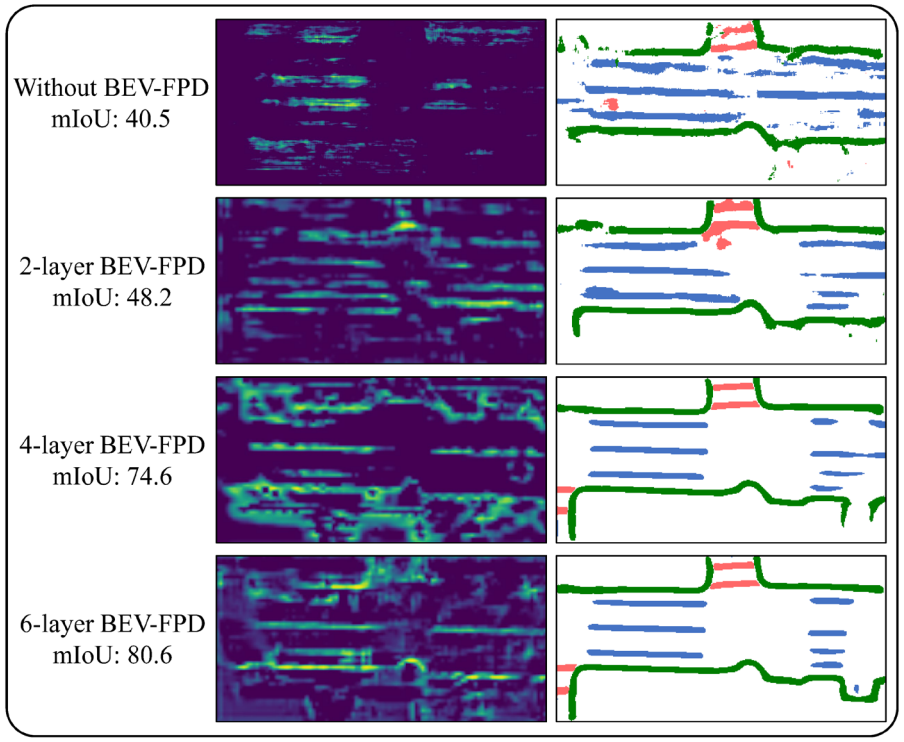
Performance Comparison

—— Vehicle Segmentation

Method	Image Size	Modality	Backbone	Setting 1	Setting 2	#Params(M)	FPS
VED [29]	800×600	Camera	ResNet-50	8.8	-	-	-
PON [34]	800×600	Camera	ResNet-50	24.7	-	38	30
VPN [30]	800×600	Camera	ResNet-50	25.5	-	18	-
STA [37]	1280×720	Camera	ResNet-50	36.0	-	-	-
Lift-Splat [32]	352×128	Camera	EfficientNet-B0	-	32.1	14	25
FIERY Static [15]	448×224	Camera	EfficientNet-B4	37.7	35.8	7.4	8
PolarBEV [26]	960×448	Camera	EfficientNet-B4	45.4	41.2	7.4	10
SimpleBEV [9]	800×448	Camera	ResNet-101	-	47.4	37	7.3
TransFuseGrid [38]	352×128	Camera & LiDAR	EfficientNet-B0 & PointPillars	-	35.9	-	18.4
Pillar feature Net [38]	-	LiDAR	PointPillars	-	23.4	-	-
LiDAR2Map	-	LiDAR	PointPillars	**58.9**	**52.1**	8.8	**35**

On the validation set of nuScenes with two commonly used settings
for vehicle segmentation without masking invisible vehicles

Ablation Study



Layer Num.	Div.	P. C.	Bound.	mIoU	FPS
2	49.3	34.1	58.4	47.3	8.2
	45.4	30.5	55.6	43.8	23.3
4	56.9	45.1	64.0	55.3	7.2
	55.7	43.9	63.2	54.3	16.3
6	60.8	47.2	66.3	58.1	6.3
	60.4	45.5	66.4	57.4	12.6

Accuracy and speed performance with different layer number of BEV-FPD

Ablation Study

Baseline	PGF^{2}M	FD	LD	Map	Vehicle
✓				52.2	49.1
✓	✓			52.6	50.6
✓	✓	✓		53.5	51.8
✓	✓		✓	53.7	51.3
✓	✓	✓	✓	54.3	52.1

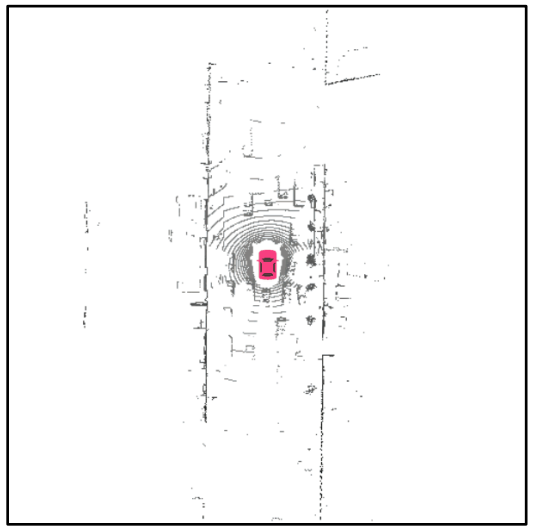
The effectiveness of our online Camera-to-LiDAR distillation scheme

Method	Div.	P. C.	Bound.	mIoU
Baseline	53.9	41.2	61.6	52.2
MonoDistill [6]	47.2	31.4	55.1	44.6
MGD [47]	52.0	38.7	59.6	50.1
xMUDA [16]	54.7	42.6	62.5	53.3
2DPASS [45]	55.3	43.0	62.4	53.6
LiDAR2Map (Ours)	55.7	43.9	63.2	54.3

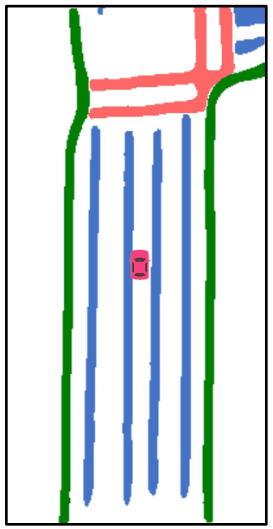
Comparison with different knowledge distillation strategies

Cam. Num.	Div.	P. C.	Bound.	mIoU
0	53.9	41.2	61.6	52.2
1	55.4	43.1	63.3	53.9
2	56.3	43.7	63.4	54.5
4	56.0	43.1	63.0	54.0
6	55.7	43.9	63.2	54.3

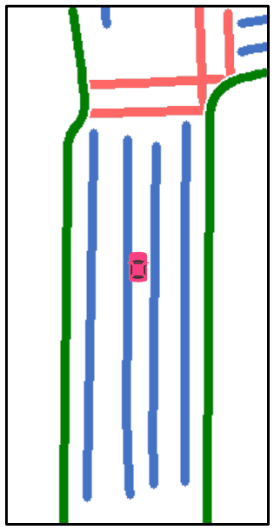
Comparison with different camera number during training



LiDAR Data (Input)

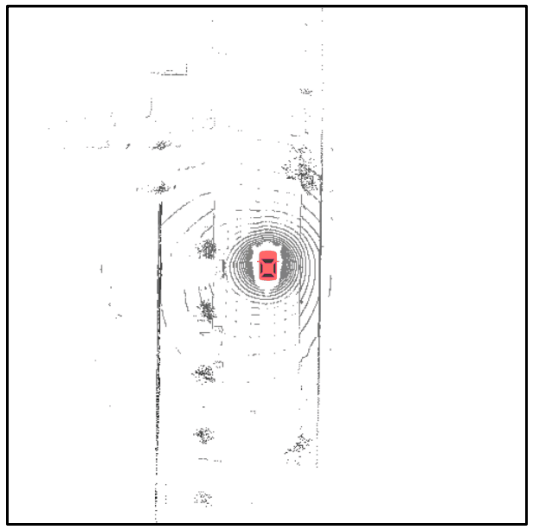


Pred. (Output)

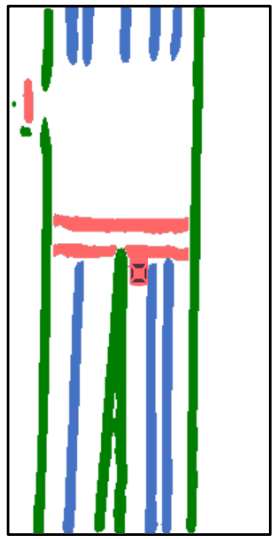


Ground Truth

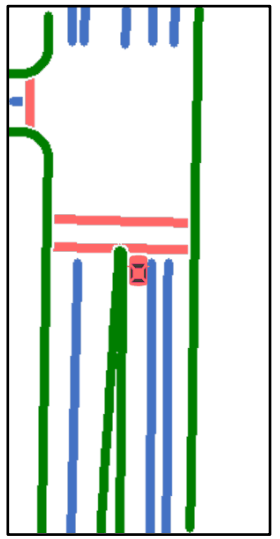
Camera Data (Visualization Only)



LiDAR Data (Input)

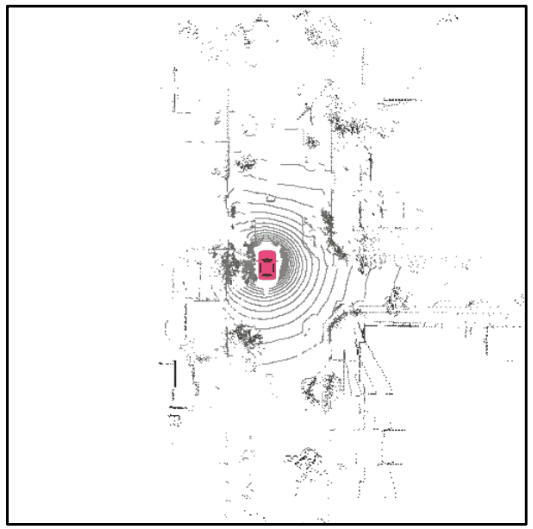


Pred. (Output)

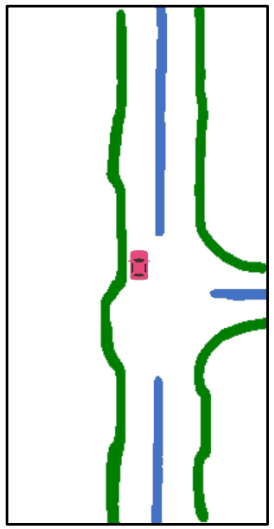


Ground Truth

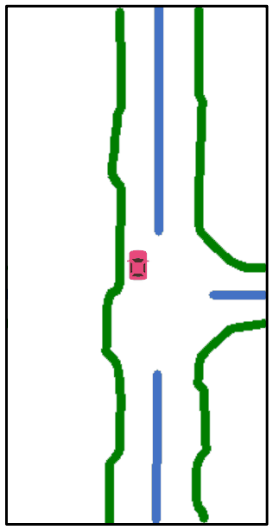
Camera Data (Visualization Only)



LiDAR Data (Input)



Pred. (Output)



Ground Truth

Camera Data (Visualization Only)

Thanks!

IEEE / CVF Computer Vision and Pattern Recognition Conference