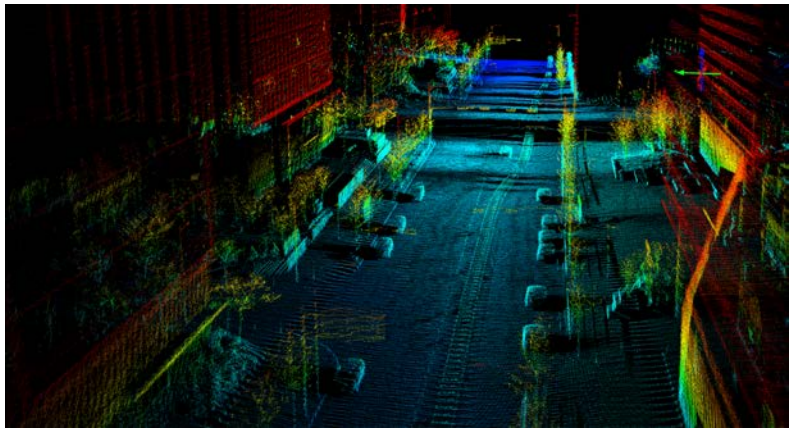
PointAugmenting: Cross-Modal Augmentation for 3D Object Detection

Recommend
火龙果软件

Background: 3D Object Detection

LiDAR



- Modality: Point cloud
- Input: (X, Y, Z, I, …)
- Advantages: accurate location
- Disadvantages: sparse, unordered

?

Fusion

Camera

- Modality: 2D Image
- Input: (R, G, B, …)
- Advantages: dense, rich semantics
- Disadvantages: lack of depth

1 **Lidar-only**

2 **Fusion-based**

Recommend
火龙果软件

Lidar-based 3D Object Detection

1 Grid-based

Methods: Divide point clouds into regular 3D voxels or BEV maps

- ✓ VoxelNet 2018 CVPR
- ✓ SECOND 2018 Sensors
- ✓ PointPillars 2019 CVPR
- ✓ SASSD 2020 CVPR

Advantages:

- computationally efficient
- high recall

Problems:

- **Voxelization** - degrade localization accuracy
- **Downscaled features** - lose spatial information
- **Uneven distribution in BEV** – get sparser with increasing depth

2 Point-based

Methods: Employ PointNet++ for feature extraction

- ✓ PointRCNN 2019 CVPR
- ✓ Fast Point RCNN 2019 ICCV
- ✓ STD 2019 ICCV
- ✓ 3DSSD 2020 CVPR
- ✓ PV-RCNN 2020 CVPR

Advantages:

- larger receptive field by the point set abstraction
- high localization accuracy

Disadvantages :

- higher computation cost

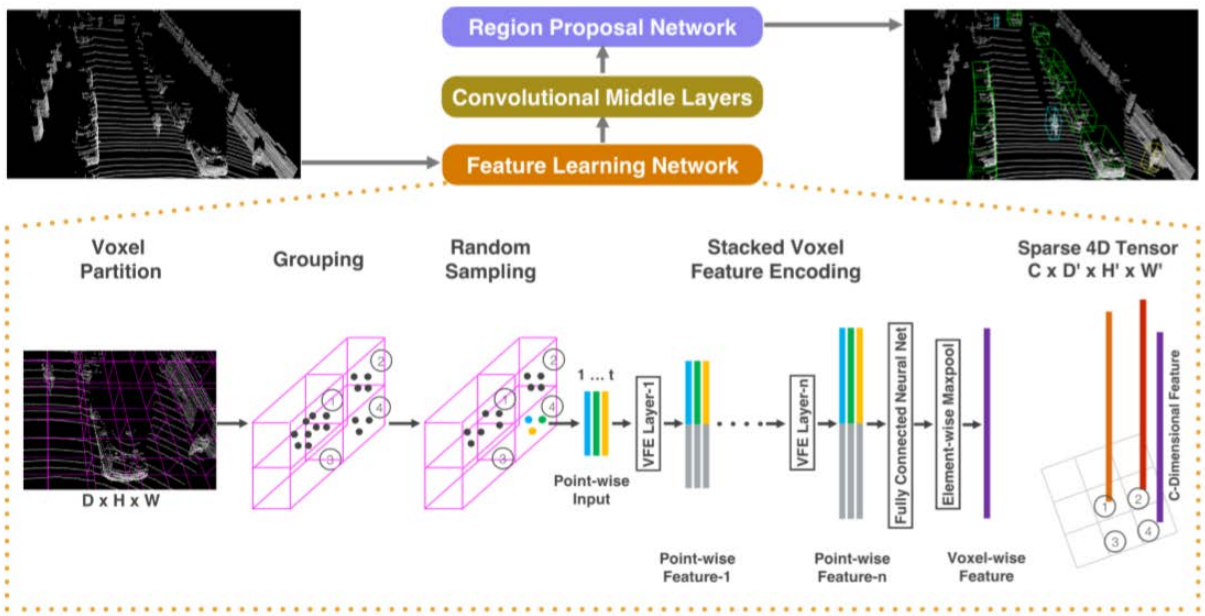
Recommend
火龙果软件

Lidar-based 3D Object Detection

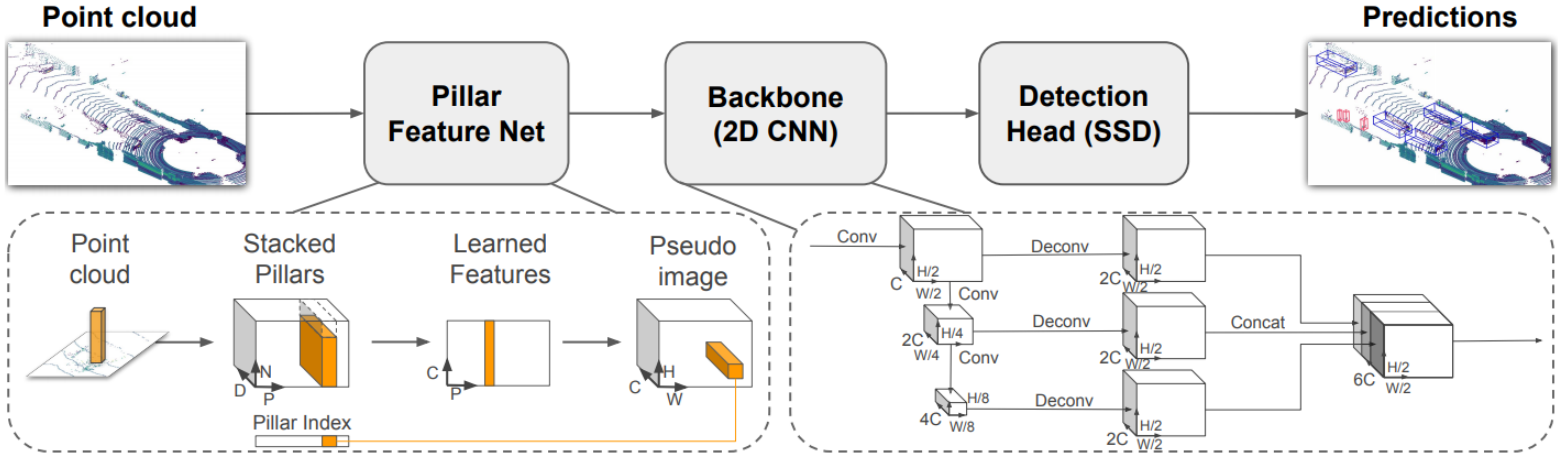
Grid-based: VoxelNet 2018 CVPR

- Voxelization
- Conv middle layer → Conv3D → reshape to BEV
- RPN

SECOND 2018 Sensors:
3D sparse convolution
- computation efficient



Grid-based: PointPillars 2019 CVPR



Recommend
火龙果软件

Lidar-based 3D Object Detection

1 Grid-based

Methods: Divide point clouds into regular 3D voxels or BEV maps

- ✓ VoxelNet 2018 CVPR
- ✓ SECOND 2018 Sensors
- ✓ PointPillars 2019 CVPR
- ✓ SASSD 2020 CVPR

Advantages:

- computationally efficient
- high recall

Problems:

- **Voxelization** - degrade localization accuracy
- **Downscaled features** - lose spatial information
- **Uneven distribution in BEV** – get sparser with increasing depth

2 Point-based

Methods: Employ PointNet++ for feature extraction

- ✓ PointRCNN 2019 CVPR
- ✓ Fast Point RCNN 2019 ICCV
- ✓ STD 2019 ICCV
- ✓ 3DSSD 2020 CVPR
- ✓ PV-RCNN 2020 CVPR

Advantages:

- larger receptive field by the point set abstraction
- high localization accuracy

Disadvantages :

- higher computation cost

Recommend
火龙果软件

Fusion-based 3D Object Detection

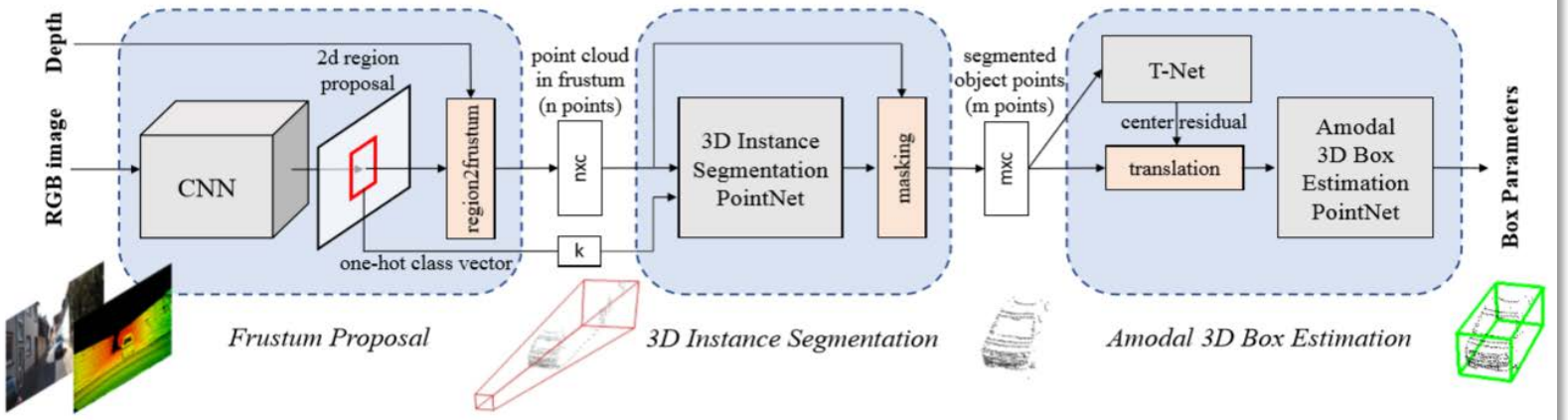
1 Result Level

Methods: adopt off-the-shelf 2D object detectors.
Disadvantages: The performance of 2D detectors set an upper bound on 3D detection.

- ✓ F-PointNets 2018 CVPR
- ✓ F-ConvNet 2019 IROS

Result-Level: F-PointNets 2018 CVPR

Stage 1: 2D proposal → frustum
Stage 2: segmentation
Stage 3: Amodal bounding box estimation



Recommend
火龙果软件

Fusion-based 3D Object Detection

1 Result Level

Methods: adopt off-the-shelf 2D object detectors.
Disadvantages: The performance of 2D detectors set an upper bound on 3D detection.

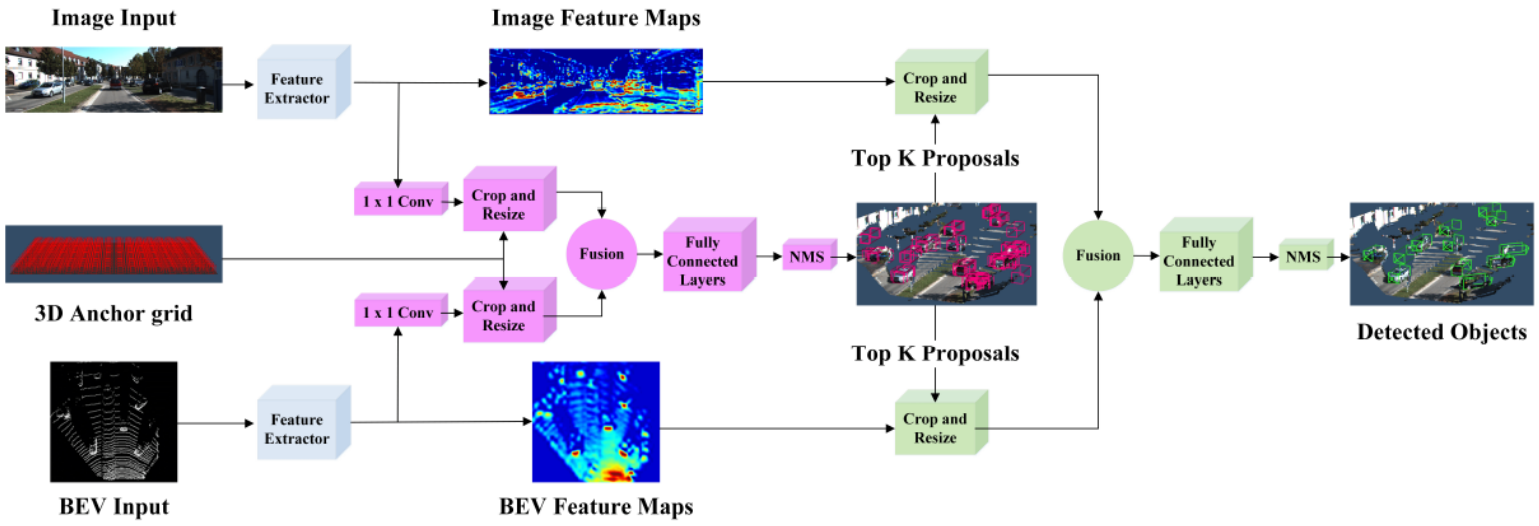
- ✓ F-PointNets 2018 CVPR
- ✓ F-ConvNet 2019 IROS

2 Proposal Level

Methods: perform fusion at the region proposal level
Disadvantages: slow and cumbersome

- ✓ MV3D 2017 CVPR
- ✓ AVOD 2018 IROS

Proposal Level: AVOD 2018 IROS



Recommend
火龙果软件

Fusion-based 3D Object Detection

1 Result Level

Methods: adopt off-the-shelf 2D object detectors.
Disadvantages: The performance of 2D detectors set an upper bound on 3D detection.

- ✓ F-PointNets 2018 CVPR
- ✓ F-ConvNet 2019 IROS

2 Proposal Level

Methods: perform fusion at the region proposal level
Disadvantages: slow and cumbersome

- ✓ MV3D 2017 CVPR
- ✓ AVOD 2018 IROS

3 Point Level

Methods: fetch point-wise image features by projecting point clouds onto image plane.

a

Methods: construct BEV camera features before fusing with LiDAR BEV features.
Disadvantages: Feature blurring

- ✓ ContFuse 2018 ECCV
- ✓ MMF 2019 CVPR
- ✓ 3D-CVF 2020 ECCV

b

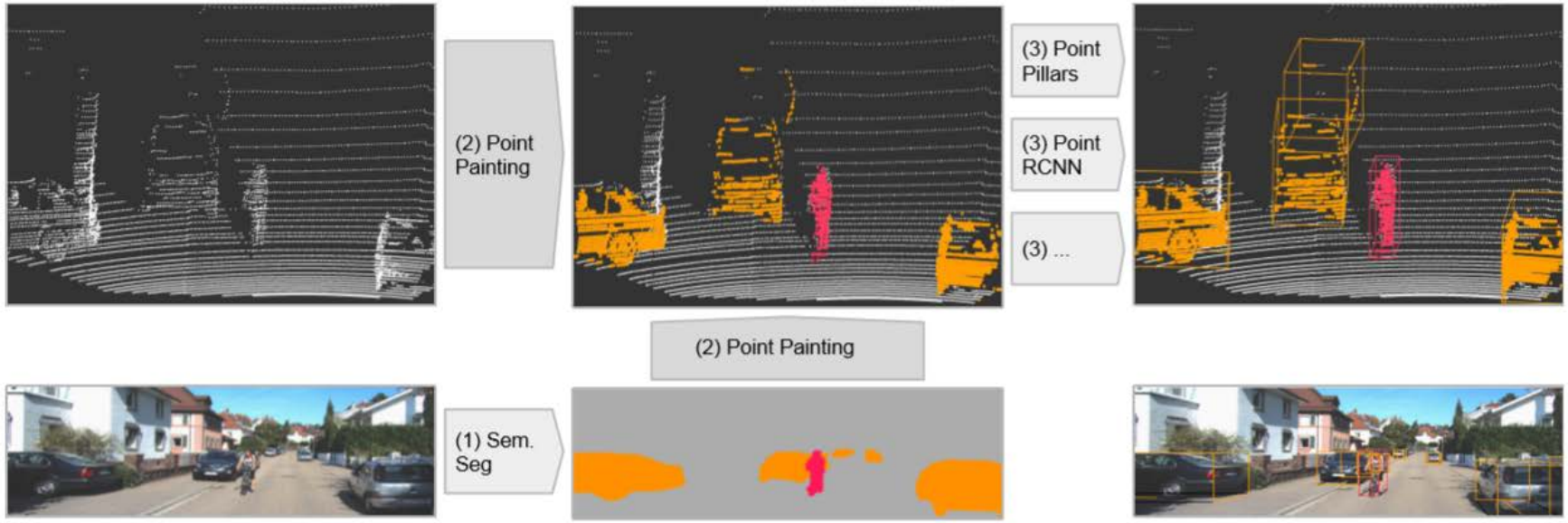
Methods: augment each LiDAR point with image features or segmentation scores.

- ✓ MVX-Net 2019 ICRA
- ✓ **PointPainting 2020 CVPR**

Recommend
火龙果软件

Fusion-based 3D Object Detection

Point-Level: PointPainting 2020 CVPR



Method	mAP	Car			Pedestrian			Cyclist		
	Mod.	Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
PointPillars [10]	73.84	90.21	87.75	84.92	71.97	67.84	62.41	85.74	65.92	62.40
Painted PointPillars	76.46	90.01	87.65	85.26	79.93	72.96	67.51	81.72	68.76	63.99
Delta	+2.62	-0.2	-0.1	+0.34	**+7.96**	**+5.12**	+5.10	-4.02	+2.84	+1.59
VoxelNet [33, 28]	72.11	89.82	87.36	79.49	70.76	62.73	55.05	86.93	66.25	59.52
Painted VoxelNet	74.22	90.05	87.51	86.66	71.09	67.06	59.79	87.46	68.08	65.59
Delta	+2.10	+0.23	+0.15	**+7.17**	+0.33	+4.33	**+7.74**	+0.53	+1.83	+6.07
PointRCNN [20]	72.65	89.96	86.64	86.07	68.89	63.54	57.63	84.99	67.78	63.71
Painted PointRCNN	75.80	90.19	87.64	86.71	72.65	66.06	61.24	86.33	73.69	70.17
Delta	**+3.15**	**+0.23**	**+1.0**	+0.64	+3.76	+2.52	+3.61	**+1.34**	**+5.91**	**+6.46**

Table 1. PointPainting applied to state of the art lidar based object detectors. All lidar methods show an improvement in bird's-eye view (BEV) mean average precision (mAP) of car, pedestrian, and cyclist on KITTI *val* set, moderate split.

Recommend
火龙果软件

Image Representation for Lidar Points

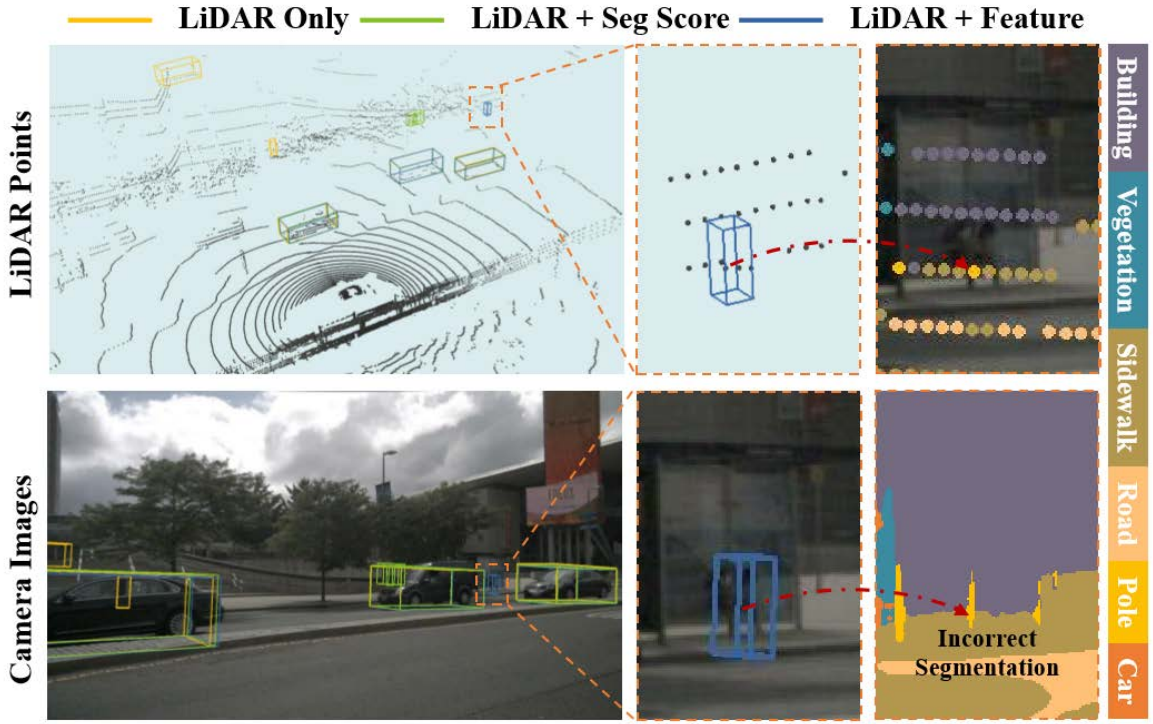
Segmentation Scores

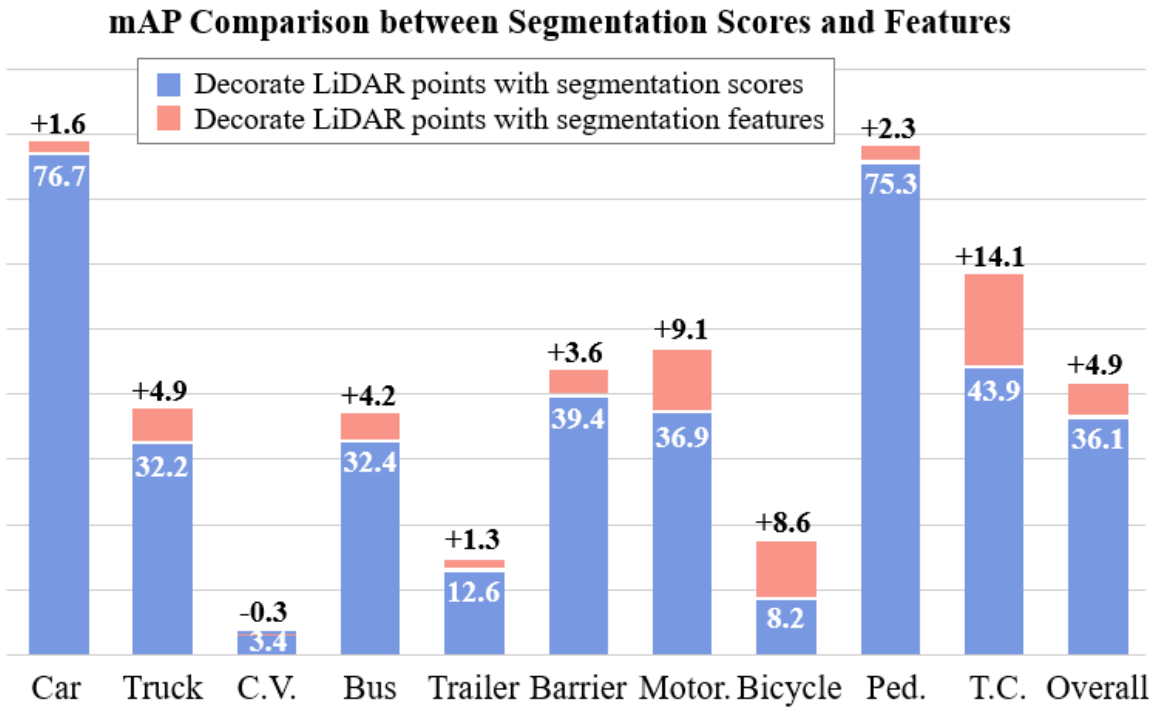
- Provide semantic labels
- Straightforward and compact semantic cues

VS

CNN Features

- Provide richer semantic cues rather than the object class only
- Larger receptive field

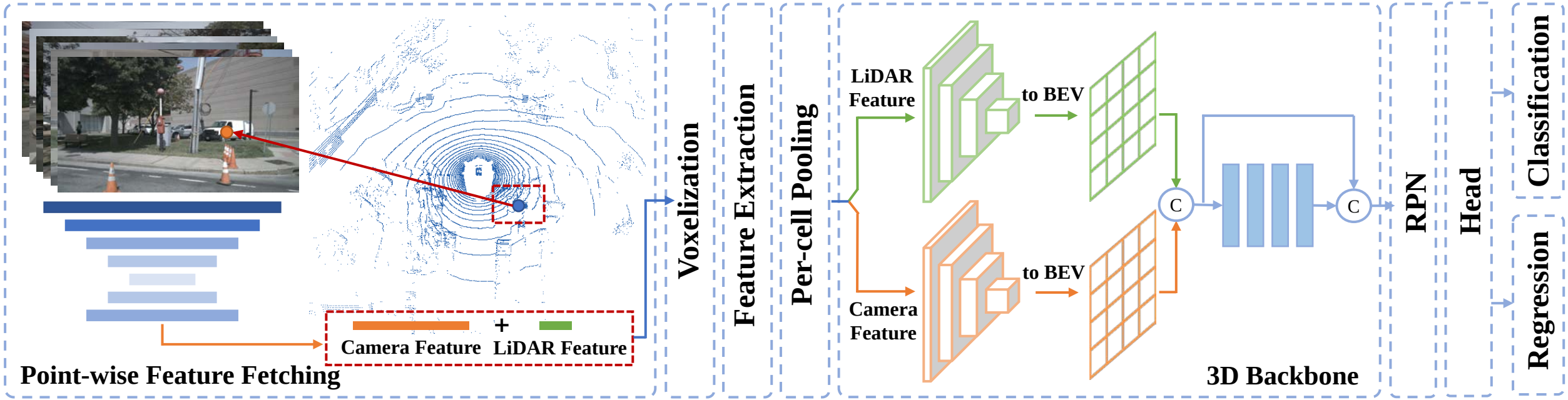




- ***PointPainting fails due to segmentation failures on small objects***

- ***CNN Feature is better than Segmentation scores***

PointAugmenting Network Architecture



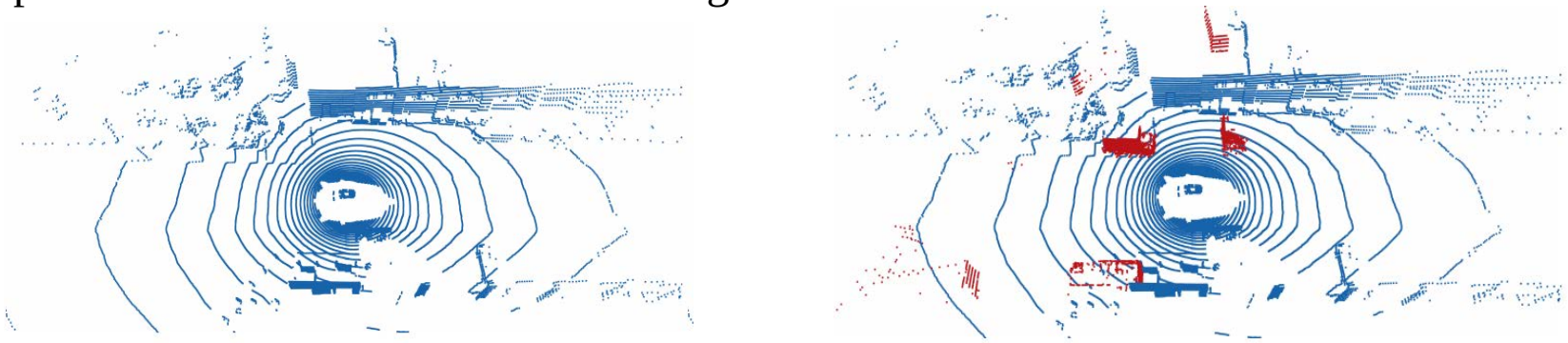
- **Lidar only Baseline:** CenterPoint
- **Point-wise Feature Fetching**: . LiDAR points are projected onto image plane and then appended by the fetched point-wise CNN features
- **3D Detection**: a late fusion mechanism across modalities

Recommend
火龙果软件

Data Augmentation for Cross-modality

- **Data Augmentation for Lidar Points**

  GT-Paste: pastes virtual objects in the forms of ground-truth boxes and LiDAR points from other scenes to the training scenes.



Method	Car	Truck	C.V.	Bus	Trailer	Barrier	Motor.	Bicycle	Ped.	T.C.	mAP	NDS
CenterPoint w/o GT-Paste	74.2	30.9	3.7	27.0	12.5	37.2	30.3	1.7	68.2	42.4	32.8	42.3
CenterPoint w/ GT-Paste	78.6	39.2	2.0	33.5	13.5	46.8	32.2	8.6	74.2	47.5	37.6	49.5
Gains of GT-Paste	+4.4	+8.3	-1.7	+6.5	+1.0	+9.6	+1.9	+6.9	+6.0	+5.1	+4.8	+7.2

Table 1. Effectiveness of the GT-Paste data augmentation scheme. Applying GT-Paste data augmentation for LiDAR points achieves an improvement of $+4.8\%$ 3D mAP. We use CenterPoint as baseline with 1/8 training data on the nuScenes dataset.

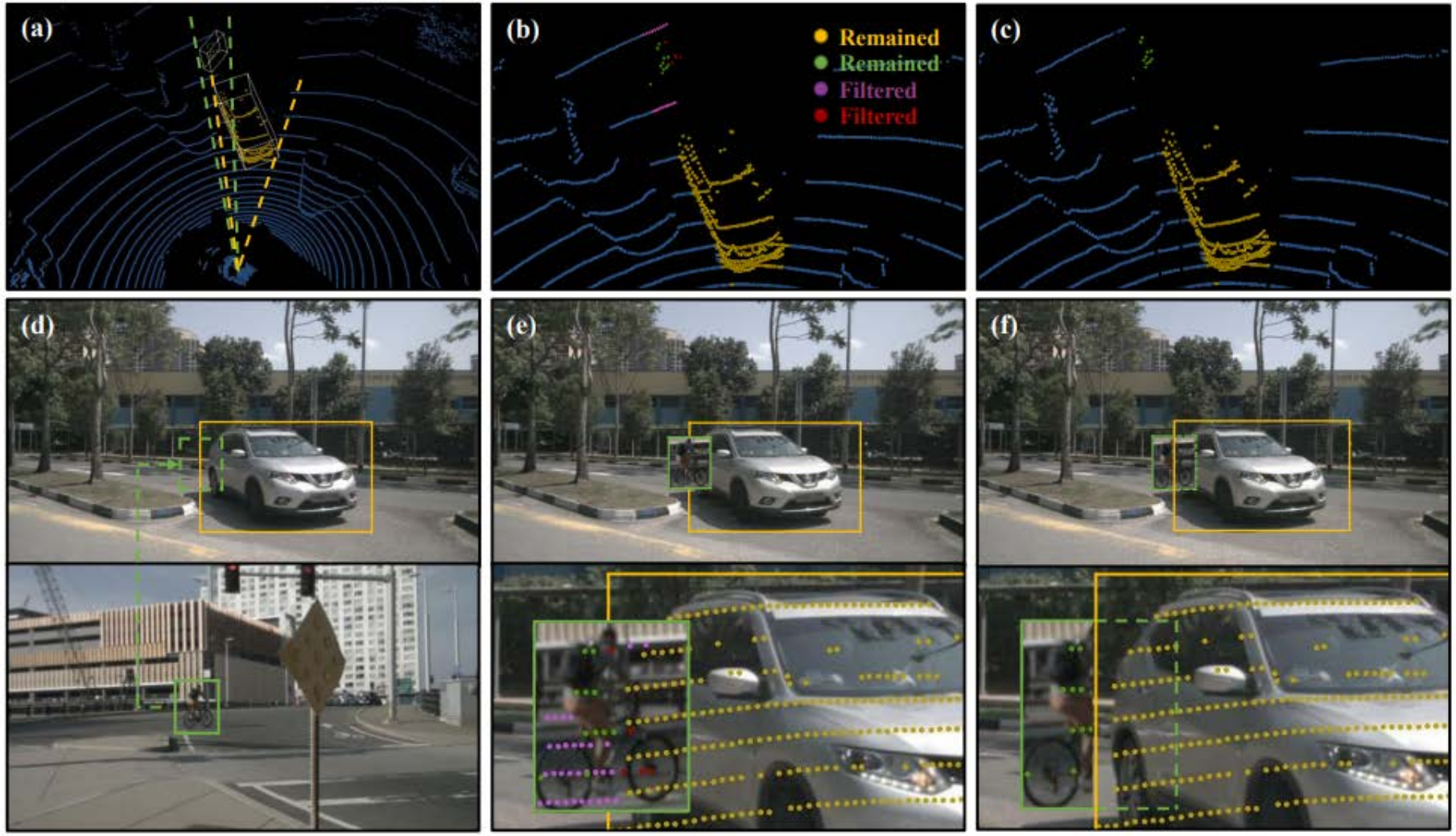
Extend to Cross-modality – Consistency Destruction

propose a simple yet effective cross-modal augmentation method to make GT-Paste applicable to both point clouds and images.

Recommend
火龙果软件

Data Augmentation for Cross-modality

- **Methods**: simultaneously attach a virtual object onto Lidar scene and images.
- **Challenge**: consistency preservation between camera and LiDAR data.



Recommend
火龙果软件

Experiments Results

nuScenes datatset

- ***Rank 2 on nuScenes Leaderboard (rank 1 with single model)***

Method	mAP	NDS	Car	Truck	C.V.	Bus	Trailer	Barrier	Motor.	Bicycle	Ped.	T.C.
PointPillars [9]	30.5	45.3	68.4	23.0	4.1	28.2	23.4	38.9	27.4	1.1	59.7	30.8
3DSSD [25]	42.6	56.4	81.2	47.2	12.6	61.4	30.5	47.9	36.0	8.6	70.2	31.1
PointPainting [19]	46.4	58.1	77.9	35.8	15.8	36.2	37.3	60.2	41.5	24.1	73.3	62.4
CBGS [35]	52.8	63.3	81.1	48.5	10.5	54.9	42.9	65.7	51.5	22.3	80.1	70.9
CenterPoint [27]	60.3	67.3	85.2	53.5	20.0	63.6	56.0	71.1	59.5	30.7	84.6	78.4
Ours	**66.8**	**71.0**	**87.5**	**57.3**	**28.0**	**65.2**	**60.7**	**72.6**	**74.3**	**50.9**	**87.9**	**83.6**
					+8.0					**+20.2**		**+5.2**

Table 2. Performance comparisons of 3D object detection on the nuScenes test set. We report the NDS, mAP, and mAP for each class.

Waymo datatset

Method	Vehicle		Pedestrian		Cyclist		All	
	L1 mAP	L2 mAP	L1 mAP	L2 mAP	L1 mAP	L2 mAP	L1 mAP/mAPH	L2 mAP/mAPH
CenterPoint [27]	66.70	62.00	73.55	68.64	72.51	70.00	70.92 / 68.26	66.88 / 64.36
Ours	67.41	62.70	75.42	70.55	76.29	74.41	73.04 / 70.39	69.22 / 66.70
Gains of fusion	+0.71	+0.70	+1.87	+1.91	+3.78	+4.41	+2.12 / +2.13	+2.34 / +2.34

Table 3. Performance comparisons of 3D object detection on the Waymo validation set. We show the mAP and mAPH in the L1 and L2 difficulty levels. The results of CenterPoint are reproduced by ourselves.

Ablation Study

1 Cross-Modal Network Design

	Seg Score	DetFeat.	CC	LF	mAP	NDS
(a)					37.4	49.9
(b)	✓		✓		42.3	51.4
(c)		✓	✓		46.0	53.9
(d)		✓		✓	47.5	55.6

Table 4. Comparison of fusion policies. Seg Score: decorating LiDAR points with segmentation scores as suggested by PointPainting [19]. DetFeat: decorating LiDAR points with image features from the detection task. CC: fusing LiDAR and image features by point-wise concatenation. LF: our late fusion mechanism.

2 Cross-Modal Data Augmentation

	Naive	CM	Fade	Fusion	mAP	NDS
(e)					32.8	42.3
(f)	✓				37.6	49.5
(g)		✓			37.4	49.9
(h)				✓	42.6	50.0
(i)		✓		✓	47.5	55.6
(j)		✓	✓	✓	48.8	56.8

Table 5. Effectiveness of cross-modal data augmentation. Naive: the original GT-Paste applied to CenterPoint. CM: Our cross-modal GT-Paste data augmentation. Fade: the training strategy that discontinues our data augmentation in the last 5 epochs. Fusion: adding camera stream by our late fusion mechanism.

Recommend
火龙果软件

Ablation Study

Figure 6. Qualitative results of 2D detection after our data augmentation. Top row: 2D detection results of original scenes. Bottom row: the results after our GT-Paste data augmentation. Yellow and green boxes respectively denote the detection results of original and pasted objects. Red boxes are the false negative predictions.

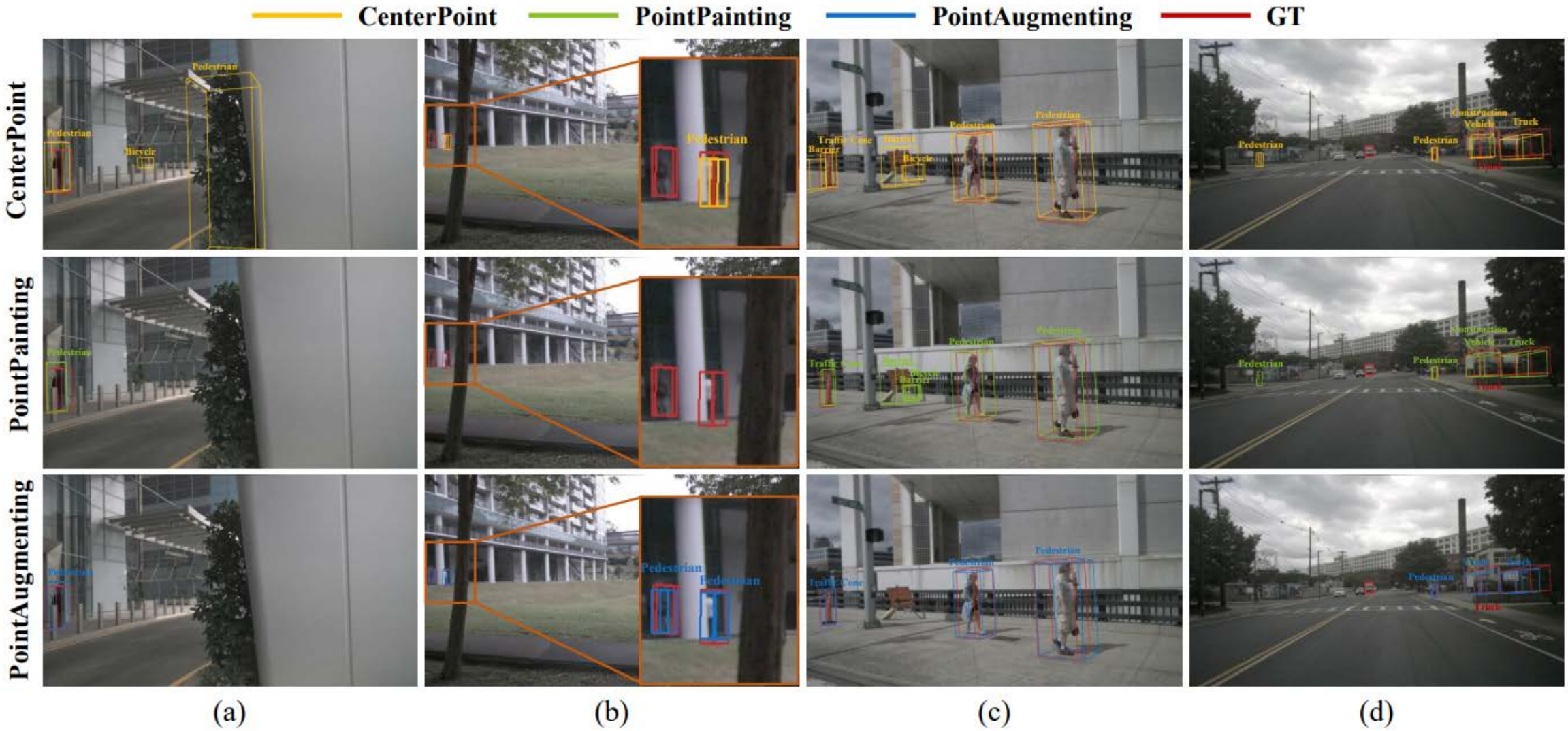
Runtime

Methods	Image size	Fusion	mAP	2D time	Total time
CenterPoint	-	-	37.6	-	85ms
Ours	896 × 1600	LF	47.5	383ms	542ms
Ours lite 1	448 × 800	LF	47.3	95ms	238ms
Ours lite 2	448 × 800	CC	46.4	95ms	178ms

Table 6. Runtime per frame on the nuScenes dataset. CC: point-wise concatenation. LF: our late fusion mechanism. The runtime is on a NVIDIA 1080Ti GPU.

Recommend
火龙果软件

Result Comparison



Recommend
火龙果软件