Image-Guided Semantic Pseudo-LiDAR Point Generation for 3D Object Detection

Minseung Lee[1], Seokha Moon[1], Seung Joon Lee[2], Reza Mahjourian[3], and Jinkyu Kim[1†]

[1]CSE, Korea University, [2]LG Innotek, [3]Waymo Research

Vision & AI Lab

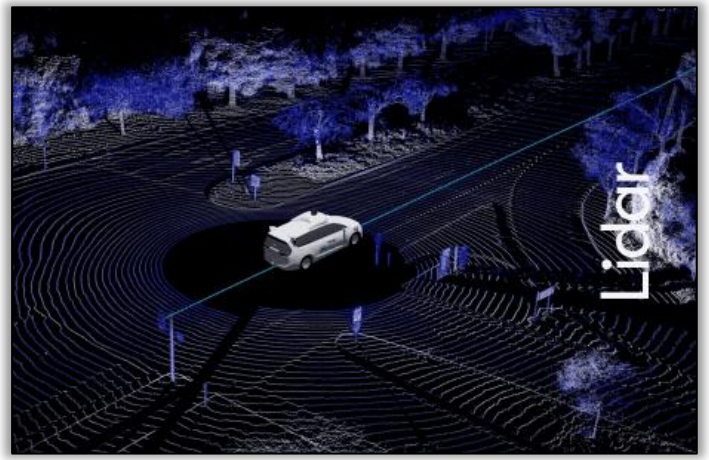
Korea University, Seoul, Korea

Introduction

❑ Motivation

- Types of Sensors
  - LiDAR: Provides accurate 3D environmental information in the form of point clouds.
  - Camera: Acquires rich semantic information from 2D images.

LiDAR



Camera

1) https://support.google.com/waymo/answer/9190838?hl=en

Introduction

❑ Motivation

- Limitations of LiDAR
  - Point clouds acquired by LiDAR become increasingly sparse as the distance from the sensor increases or when objects are small or occluded (Fig. 1), which leads to a degradation in 3D object detection performance (Fig. 2).

Points : 1942
Distance : 7.9m

Points : 355
Distance : 20.1m

Points : 74
Distance : 36.4m

Figure 1.

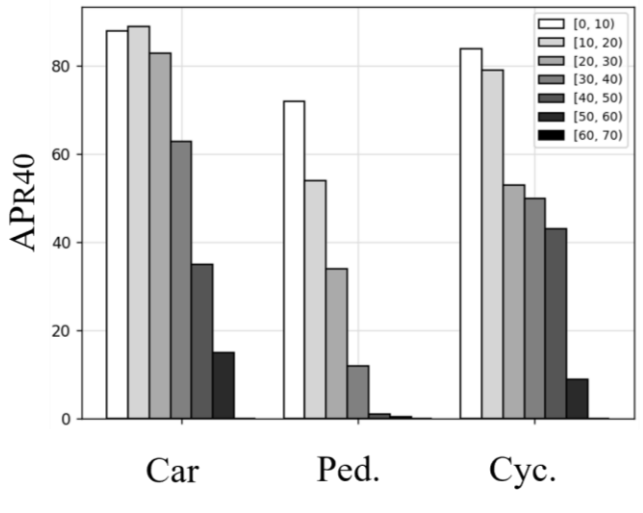


Figure 2.

1) Lang, A. H., et al., Pointpillars: Fast encoders for object detection from point clouds. CVPR 2019.

Introduction

❑ Motivation

- Limitations of LiDAR
  - To alleviate LiDAR sparsity, prior works generate additional points to densify the point cloud [1]. However, since these approaches rely solely on LiDAR, they may hallucinate points in background regions, resulting in false-positive detections (Fig. 1).
  - To address these limitations, we propose **Image-Guided Semantic Pseudo-LiDAR Point Generation** (ImagePG), which incorporates additional semantic cues extracted from camera images to guide point generation (Fig. 2).

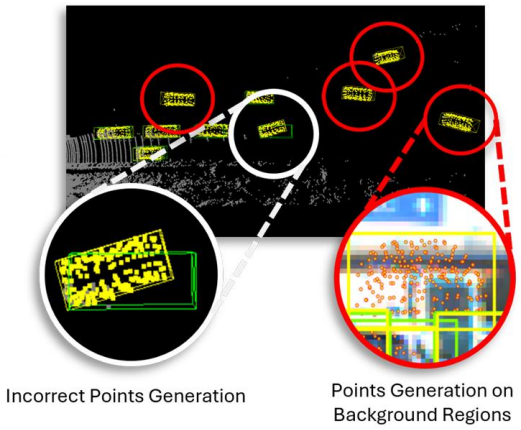


Figure 1.

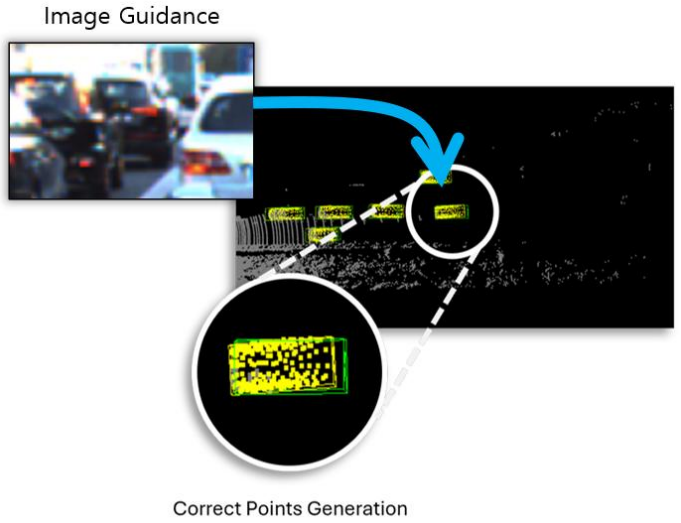


Figure 2.

1) I. Koo, et al, Pg-rcnn: Semantic surface point generation for 3d object detection. In Proceedings of the IEEE/CVF international conference on computer vision (pp.18142-18151).

Method

❑ Image-Guided Semantic Pseudo-LiDAR Point Generation (ImagePG)

- **ImagePG** consists of three components (Fig. 1).
  - **Input Feature Extraction**: Extract features from the LiDAR point cloud and the camera image.
  - **Multi-Stage Points Generation**: Iteratively generate points and predict 3D bounding boxes over multiple refinement stages.
  - **Box Voting**: Aggregate predicted boxes via voting to produce the final detections.
- Image-Guided RoI Points Generation (IG-RPG) and Image-aware Occupancy Prediction Network (I-OPN) explicitly leverage image features to enable accurate point generation and reliable 3D bounding box prediction.

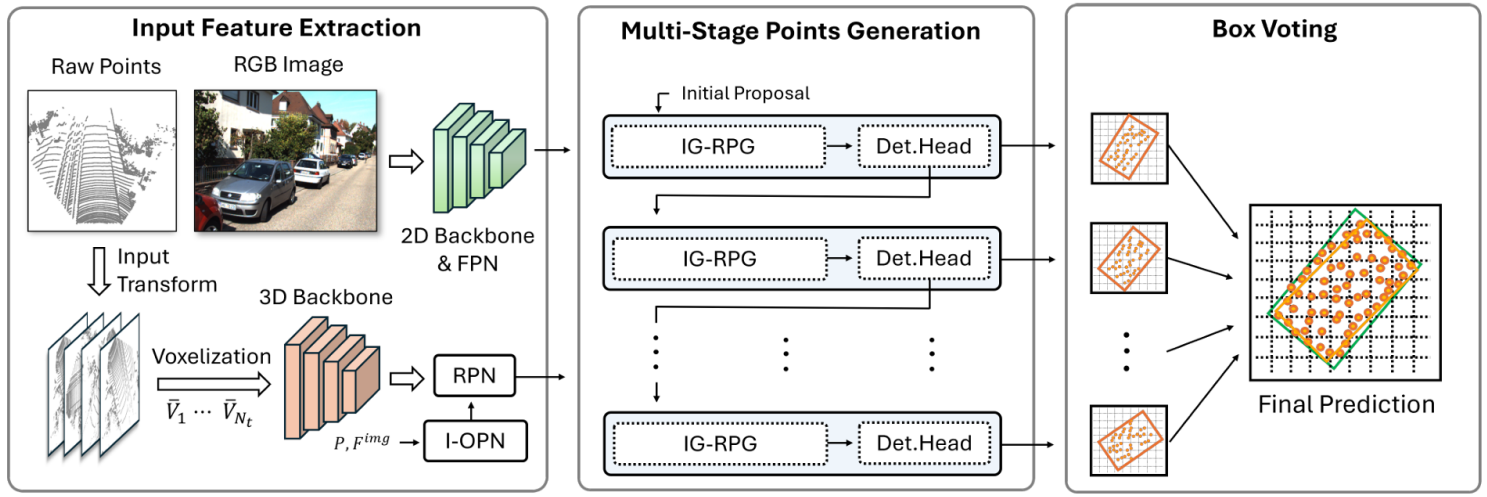


Figure 1.

Method

❑ Image-Guided RoI Points Generation (IG-RPG)

- **IG-RPG** fuses image features and voxel features at each grid location within the Region of Interest (RoI) via deformable attention. The fused representation is then used to generate points, and the generated points are subsequently leveraged to estimate the final 3D bounding boxes.

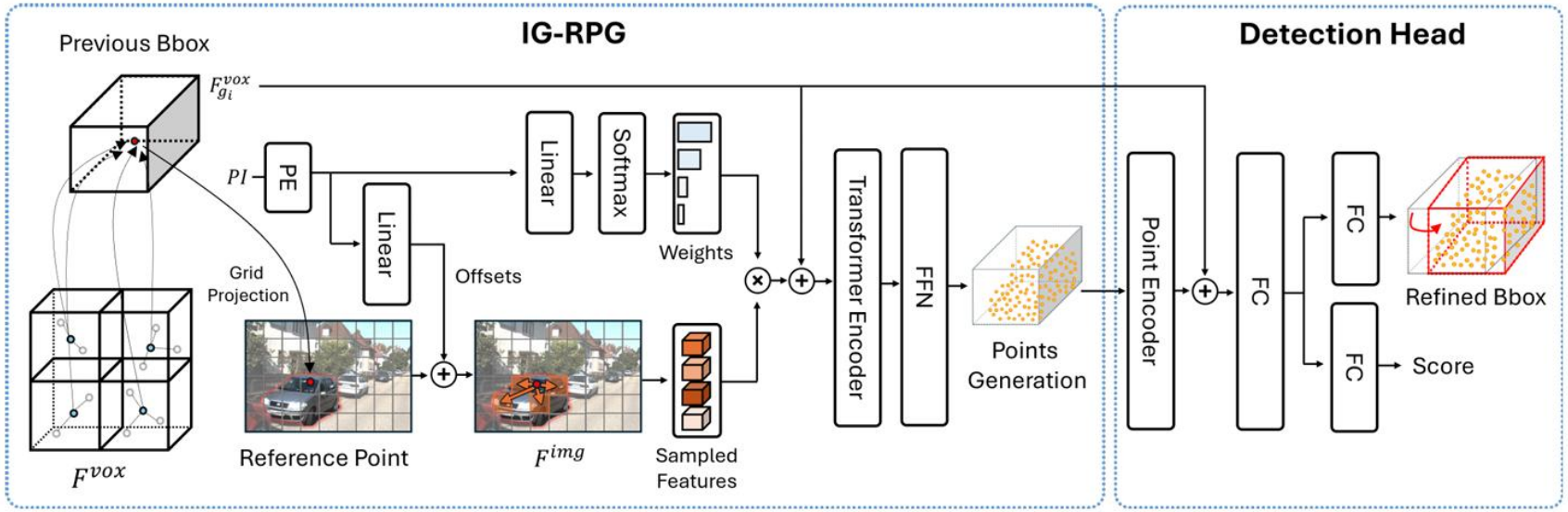


Figure 1.

Method

❑ Image-aware Occupancy Prediction Network (I-OPN)

- **I-OPN** aggregates image features using deformable attention and predicts BEV occupancy, thereby implicitly learning where points should be filled in the scene.

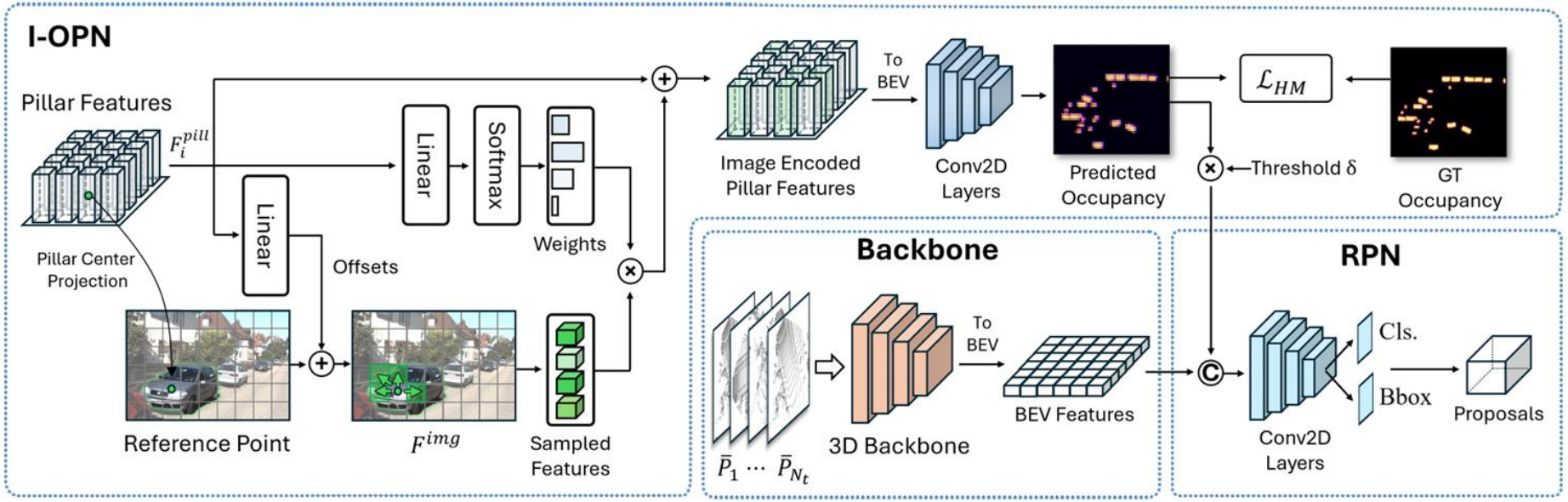


Figure 1.

Experiments

❑ Quantitative Results

- On KITTI, our method improves performance over the baseline by +1.38 pp (Car), +7.91 pp (Pedestrian), and +5.21 pp (Cyclist) (Tab. 1), achieving State-of-the-Art performance on KITTI leaderboard for Cyclist. It also outperforms the baseline on Waymo (Tab. 2).

Method	Modality	mAP	Car 3D (IoU=0.7)				Pedestrian 3D (IoU=0.5)				Cyclist 3D (IoU=0.5)			
			Easy	Mod.	Hard	mAP	Easy	Mod.	Hard	mAP	Easy	Mod.	Hard	mAP
PointRCNN [26]	L	60.33	86.96	75.64	70.70	77.77	47.98	39.37	36.01	41.12	74.96	58.82	52.53	62.10
PV-RCNN [27]	L	64.91	90.25	81.43	76.82	82.83	52.17	43.29	40.29	45.25	78.60	63.71	57.65	66.65
PDV [7]	L	65.31	90.43	81.86	77.36	83.21	47.80	40.56	38.46	42.27	83.04	67.81	60.46	70.44
PG-RCNN [10]	L	65.38	89.38	82.13	77.33	82.95	47.99	41.04	38.71	42.58	82.77	67.82	61.25	70.61
CasA-V [38]	L	69.77	91.58	83.06	80.08	84.91	54.04	47.09	44.56	48.56	87.91	73.47	66.17	75.85
CAT-Det [56]	L+C	67.05	89.87	81.32	76.68	82.62	54.26	45.44	41.94	47.21	83.68	68.81	61.45	71.31
LoGoNet [12]	L+C	69.35	**91.80**	85.06	**80.74**	**85.87**	53.07	47.43	45.22	48.57	84.47	71.70	64.67	73.61
GraphAlign [31]	L+C	63.07	90.90	82.23	79.67	84.27	41.38	36.89	34.95	37.74	78.42	64.43	58.71	67.19
UPIDet [57]	L+C	70.13	89.13	82.97	80.05	84.05	55.59	48.77	46.12	50.16	86.74	74.32	**67.45**	76.17
TED-M [39]	L+C	**70.99**	91.61	**85.28**	80.68	85.86	55.85	**49.21**	**46.52**	**50.53**	**88.82**	74.12	66.84	**76.59**
Baseline [10]	L	65.38	89.38	82.13	77.33	82.95	47.99	41.04	38.71	42.58	82.77	67.82	61.25	70.61
+ ImagePG (Ours)	L+C	70.21 (4.83%↑)	91.00 (1.62%↑)	82.46 (0.33%↑)	79.54 (2.21%↑)	84.33 (1.38%↑)	**56.66** (8.67%↑)	48.81 (7.77%↑)	45.99 (7.28%↑)	50.49 (7.91%↑)	86.53 (3.76%↑)	**74.68** (6.86%↑)	66.24 (4.99%↑)	75.82 (5.21%↑)

Table 1.

Method	Vehicle 3D (IoU=0.7)				Pedestrian 3D (IoU=0.5)				Cyclist 3D (IoU=0.5)			
	L1		L2		L1		L2		L1		L2	
	AP	APH	AP	APH	AP	APH	AP	APH	AP	APH	AP	APH
Baseline† [13]	75.00	74.52	66.49	66.06	75.14	68.79	66.19	60.44	69.82	68.74	67.23	66.19
+ ImagePG (Ours)	77.31 (2.31%↑)	76.85 (2.33%↑)	68.84 (2.35%↑)	68.41 (2.35%↑)	78.11 (2.97%↑)	72.53 (3.74%↑)	69.28 (3.09%↑)	64.14 (3.70%↑)	71.31 (1.49%↑)	70.29 (1.55%↑)	68.71 (1.48%↑)	67.73 (1.54%↑)

Table 2.

Experiments

❑ Quantitative Results

- Consistent gains are observed across distance ranges (Tab. 1) and across point-density within objects (Tab. 2). It also achieves consistent improvements for occluded objects (Tab. 3).
- Our approach reduces the false-positive rate by 50% compared to the baseline (Fig. 1) on KITTI.

Method	Car 3D (IoU=0.7)			Pedestrian 3D (IoU=0.5)			Cyclist 3D (IoU=0.5)		
	[0, 20)	[20, 40)	[40, inf)	[0, 20)	[20, 40)	[40, inf)	[0, 20)	[20, 40)	[40, inf)
Baseline† [10]	95.12	84.40	38.23	70.44	35.03	0.87	91.42	69.51	33.56
+ ImagePG (Ours)	94.97 (0.15%↓)	85.99 (1.59%↑)	42.56 (4.33%↑)	80.79 (10.35%↑)	43.65 (8.62%↑)	3.20 (2.33%↑)	92.76 (1.34%↑)	74.89 (5.38%↑)	40.60 (7.04%↑)

Table 1.

Method	Car 3D (IoU=0.7)				Pedestrian 3D (IoU=0.5)				Cyclist 3D (IoU=0.5)			
	[Min, Q1)	[Q1, Q2)	[Q2, Q3)	[Q3, Max]	[Min, Q1)	[Q1, Q2)	[Q2, Q3)	[Q3, Max]	[Min, Q1)	[Q1, Q2)	[Q2, Q3)	[Q3, Max]
Baseline [6]	15.59	66.69	87.14	95.74	6.34	28.05	37.82	59.96	7.34	35.36	77.35	86.13
+ ImagePG (Ours)	19.34 (3.75%↑)	69.38 (2.69%↑)	88.56 (1.42%↑)	96.07 (0.34%↑)	8.99 (2.65%↑)	43.73 (15.68%↑)	57.62 (19.80%↑)	78.40 (18.44%↑)	9.26 (1.92%↑)	40.75 (5.39%↑)	80.19 (2.84%↑)	90.70 (4.57%↑)

Table 2.

Method	Car 3D (IoU=0.7)			Pedestrian 3D (IoU=0.5)			Cyclist 3D (IoU=0.5)		
	LVL_1	LVL_2	LVL_3	LVL_1	LVL_2	LVL_3	LVL_1	LVL_2	LVL_3
Baseline [13]	84.62	77.01	54.41	60.48	19.56	4.49	83.50	25.60	0.93
+ ImagePG (Ours)	84.77 (0.15%↑)	79.58 (2.57%↑)	60.47 (6.06%↑)	73.96 (13.48%↑)	33.93 (14.37%↑)	9.02 (4.53%↑)	85.53 (2.03%↑)	28.30 (2.70%↑)	2.79 (1.86%↑)

Table 3.

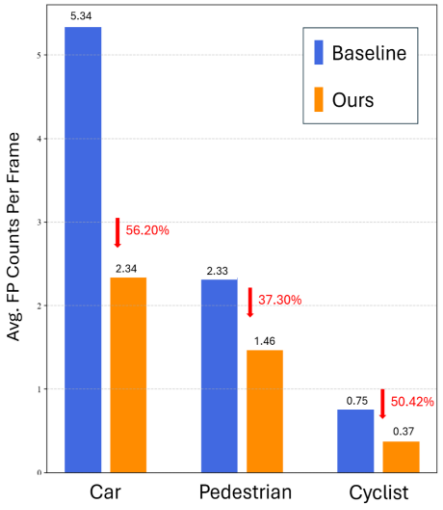


Figure 1.

Experiments

❑ Qualitative Results

- The baseline generates hallucinating points and producing incorrect detections for distant or occluded objects. In contrast, our method demonstrates improved robustness under these conditions (Fig. 1-2).

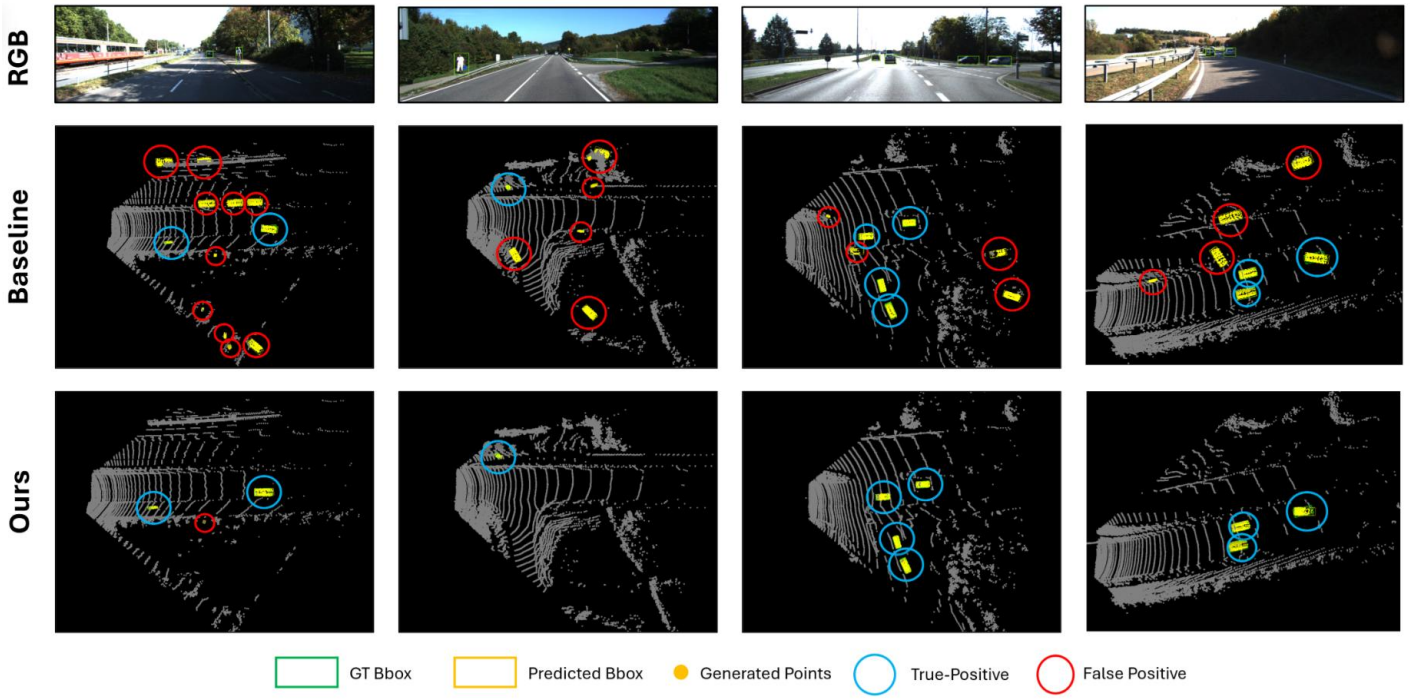


Figure 1.

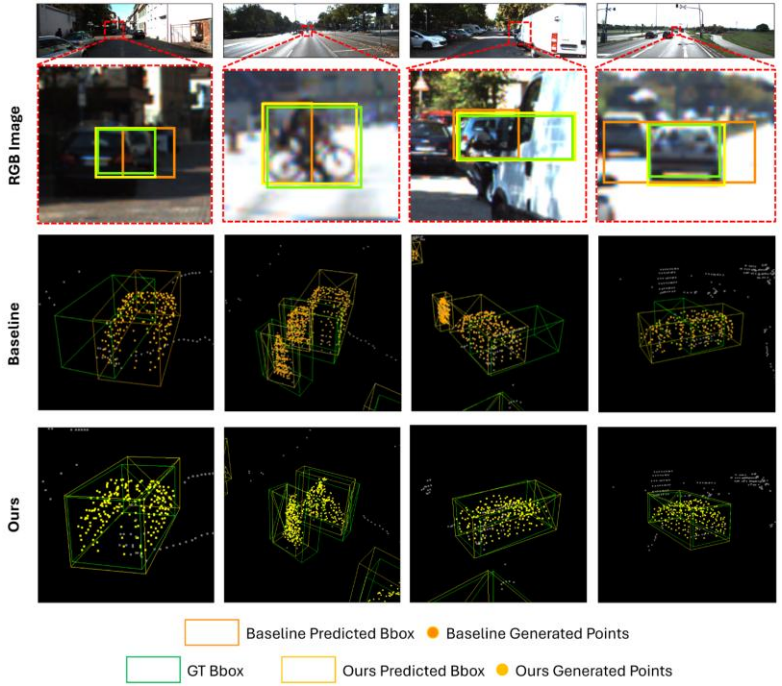


Figure 2.

Experiments

❑ Qualitative Results

- Fig. 1 provides qualitative visualizations of the points generated by the proposed approach on KITTI.

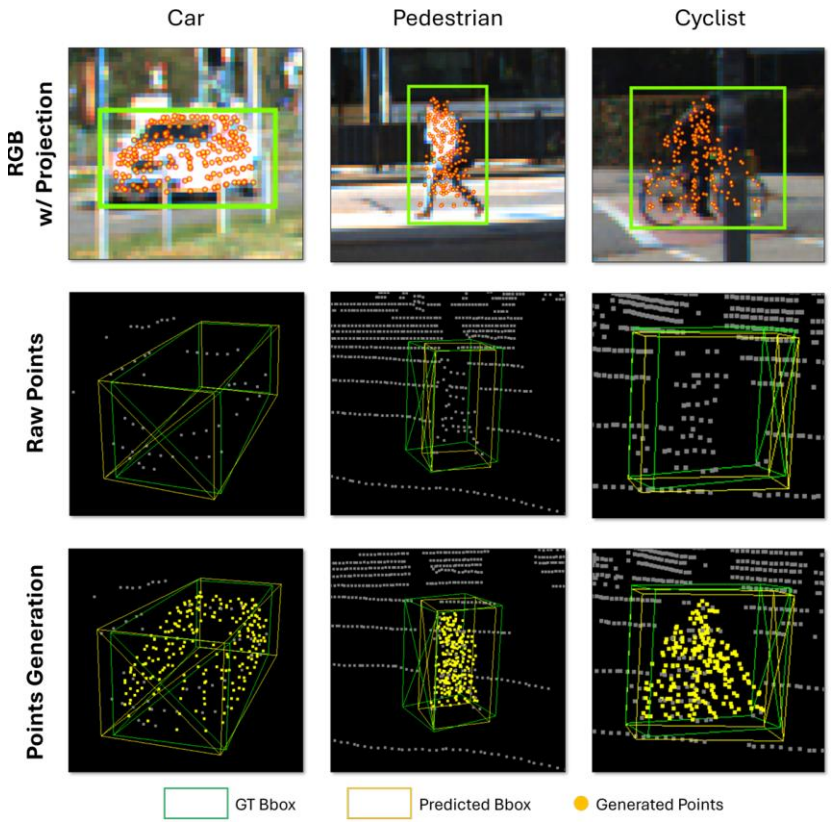


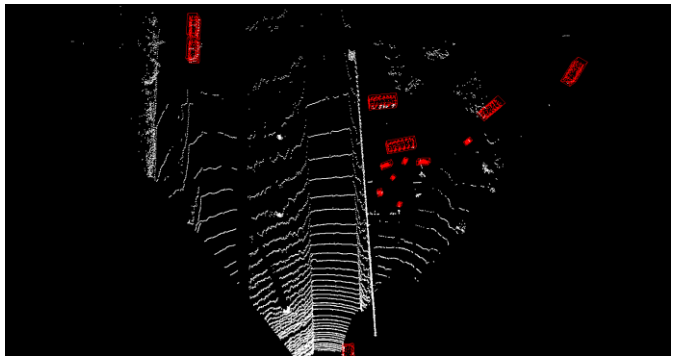
Figure 1.

Experiments

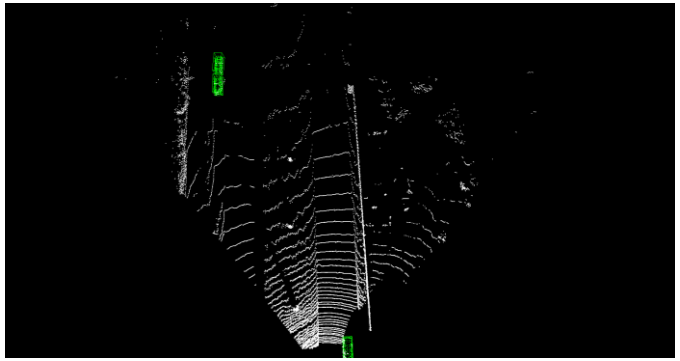
❑ Qualitative Results

- Inference videos comparing the baseline and ImagePG show that ours produces significantly fewer false positives.

Baseline



Ours



Confidence score ≥ 0.3

Conclusion

- ImagePG is the first framework to directly leverage semantic features from camera images for generating Pseudo-LiDAR points.
- ImagePG achieves significant improvements in detecting small and distant objects (e.g., pedestrians and cyclists) on KITTI and Waymo, reducing false positives by approximately 50%.
- ImagePG introduces a new perspective toward more reliable multimodal 3D object perception.

WACV 2026
TUCSON, AZ 3/6 - 3/10

KOREA UNIVERSITY

LG Innotek

WAYMO

Thank You