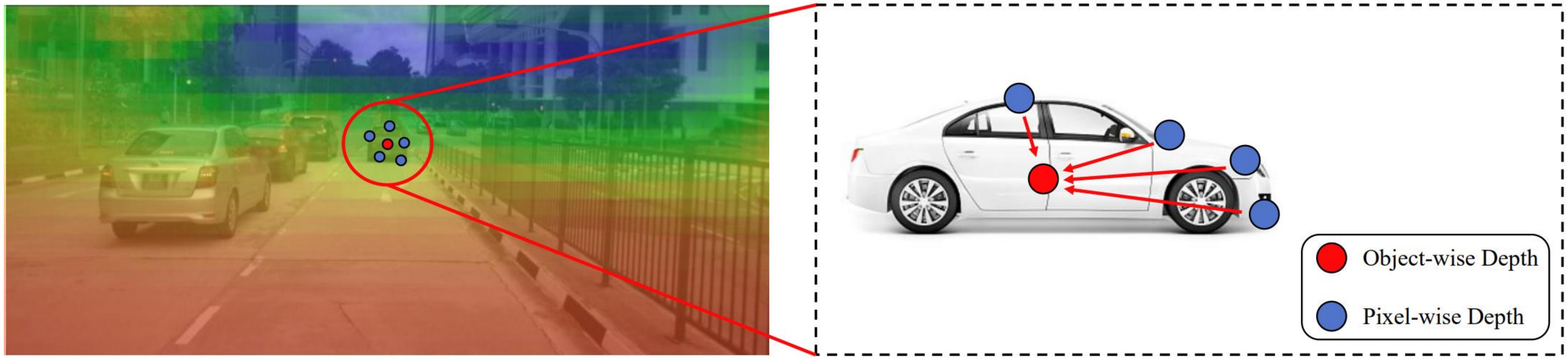
OPEN: Object-wise Position Embedding for Multi-view 3D Object Detection

Jinghua Hou, Tong Wang, Xiaoqing Ye, Zhe Liu, Shi Gong, Xiao Tan, Errui Ding, Jingdong Wang, Xiang Bai

Huazhong University of Science and Technology, Baidu Inc.

Motivation

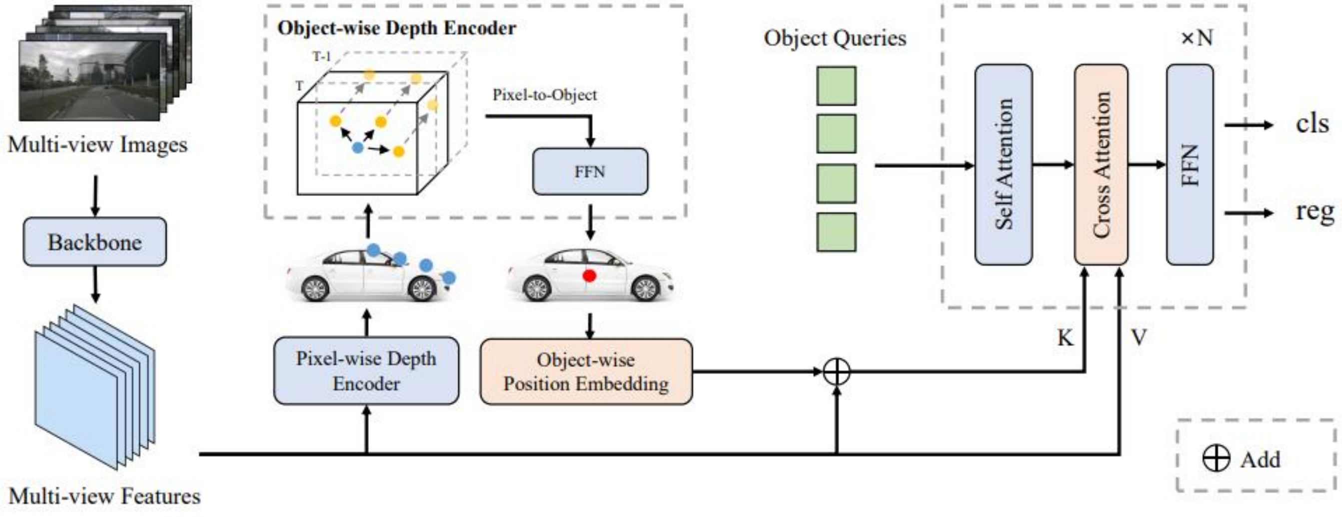
- The depth supervision obtained from LiDAR points is usually distributed on the surface of the object, which is not so friendly to existing DETR-based 3D detectors due to the lack of the depth of 3D object center
- For distant objects, fine-grained depth estimation of the whole object is more challenging
- We argue that the object-wise depth (or 3D center of the object) is essential for accurate detection



Contribution

- We propose a new multi-view 3D object detector named OPEN, which utilizes the 3D object-wise depth representation to achieve better detection performance
- We introduce the object-wise position embedding to effectively inject object-wise depth information into the transformer decoder, leading to 3D object-aware features
- The proposed OPEN outperforms previous state-of-the-art methods on the nuScenes dataset

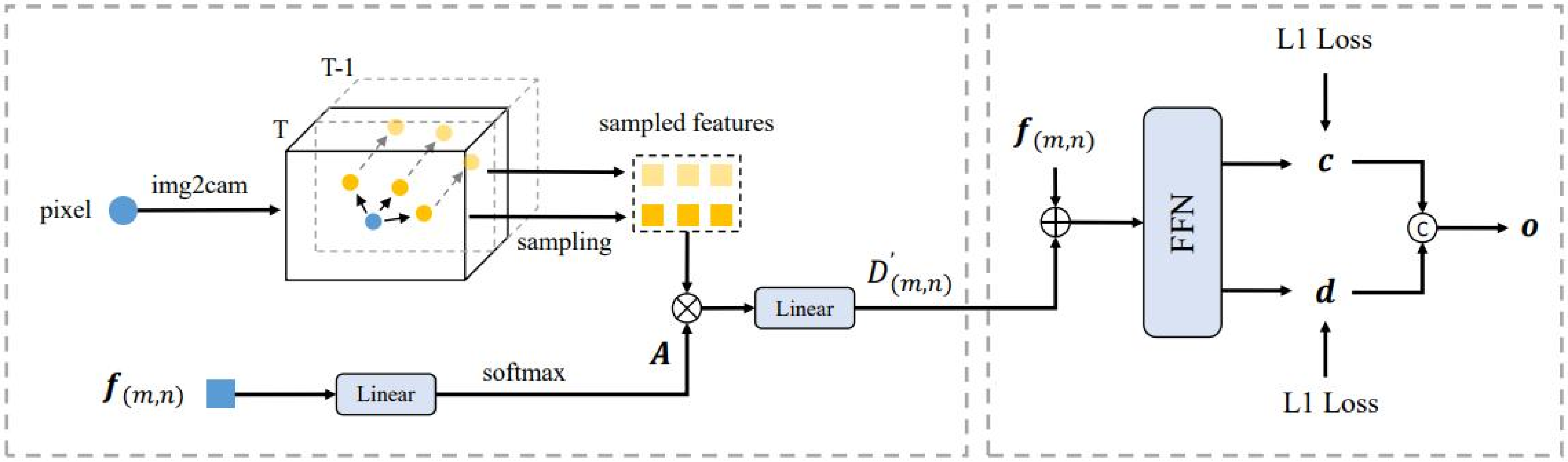
Method



OPEN consists of the pixel-wise depth encoder (PDE), the object-wise depth encoder (ODE), and object-wise position embedding (OPE)

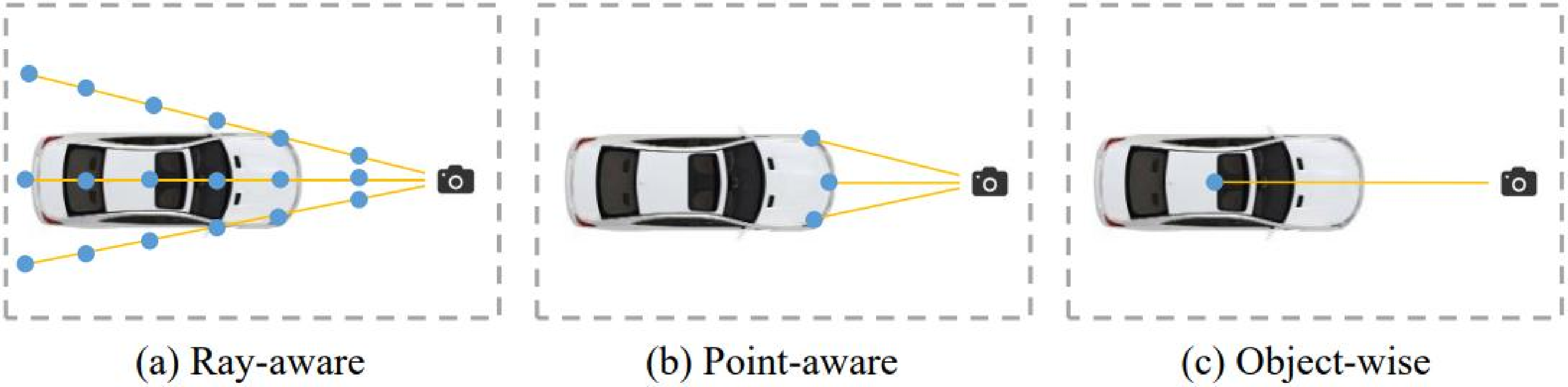
- PDE utilizes a DepthNet to predict the pixel-wise depth map supervised by projected LiDAR points
- ODE predicts the object-wise depth, supervised by the center of projected 3D bounding boxes
- OPE is used to convert the multi-view image features to object-wise 3D features

ODE



ODE predicts the object-wise depth based on pixel-wise depth information and streaming temporal fusion strategy in 3D camera space

OPE



• For ray-aware position embedding, the uncertainty depth estimation without depth supervision makes it difficult to generate accurate 3D-aware features

• For point-aware position embedding, although it adopts projected LiDAR points to supervise the pixel-wise depth prediction and encodes 3D points for position embedding to improve performance, it ignores the importance of object-wise depth for DETR-based 3D object detectors, leading to sub-optimal performance

OPE

Given the object-wise depth and corresponding object center predicted by ODE on the image, OPE converts in the pixel coordinate to the 3D object center in the LiDAR coordinate. Finally, OPE adopts a multi-layer perceptron to generate the object-wise position embedding.

$$\mathbf{o}_j^{'} = (x \times d_j, y \times d_j, d_j, 1)^{\mathrm{T}},$$

$$\mathbf{O}_j = \mathbf{R}^{-1}\mathbf{K}^{-1}\mathbf{o}_j^{'},$$

$$\mathbf{OPE}_j = \mathrm{MLP}((\mathrm{PE}_{3D}(\mathrm{Norm}(\mathbf{O}_j)))),$$

DFL

- DFL aims to further encourage OPEN to pay more attention to the object center

predicted 3D object center: $\hat{\mathbf{C}}$

predicted classification probability: $\hat{\mathbf{p}}$

ground truth 3D object center: $\mathbf{C}$

binary target class label: $\mathbf{t}$

$$\mathcal{L}_{DFL} = -\boldsymbol{\alpha}^{'} \cdot |\mathbf{t} \cdot \mathbf{s} - \hat{\mathbf{p}}|^{\gamma} \cdot \log(|1 - \mathbf{t} - \hat{\mathbf{p}}|),$$

$$\text{where } \mathbf{s} = e^{-\mathrm{L2}(\hat{\mathbf{C}} - \mathbf{C})},\ \boldsymbol{\alpha}^{'} = \alpha \cdot \mathbf{t} \cdot \mathbf{s} + (1 - \alpha) \cdot (1 - \mathbf{t} \cdot \mathbf{s}).$$

$$\mathcal{L} = \lambda_1 \mathcal{L}_{PDE} + \lambda_2 \mathcal{L}_{ODE} + \lambda_3 \mathcal{L}_{DFL} + \lambda_4 \mathcal{L}_{reg},$$

Total loss consists of depth-aware focal loss, 3D bounding box regression loss, pixel-wise depth prediction loss, and object-wise depth prediction loss.

Results

Comparison of other methods on the nuScenes validation set. OPEN achieves SOTA performance with ResNet 50 and ResNet 101 backbone.

Method	Backbone	Input Size	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
BevDet4D [8]	ResNet50	256 × 704	45.7	32.2	0.703	0.278	0.495	0.354	0.206
PETRv2 [22]	ResNet50	256 × 704	45.6	34.9	0.700	0.275	0.580	0.437	0.187
BEVDepth [15]	ResNet50	256 × 704	47.5	35.1	0.639	0.267	0.479	0.428	0.198
BEVStereo [14]	ResNet50	256 × 704	50.0	37.2	0.598	0.270	0.438	0.367	0.190
BEVFormerv2† [40]	ResNet50	-	52.9	42.3	0.618	0.273	0.413	0.333	0.188
SOLOFusion [26]	ResNet50	256 × 704	53.4	42.7	0.567	0.274	0.511	0.252	0.181
Sparse4Dv2 [18]	ResNet50	256 × 704	53.8	43.9	0.598	0.270	0.475	0.282	0.179
StreamPETR† [35]	ResNet50	256 × 704	55.0	45.0	0.613	0.267	0.413	0.265	0.196
SparseBEV† [19]	ResNet50	256 × 704	55.8	44.8	0.581	0.271	0.373	0.247	0.190
OPEN†	ResNet50	256 × 704	**56.4**	**46.5**	0.573	0.275	0.413	0.235	0.193
3DPPE [33]	ResNet101	512 × 1408	45.8	39.1	0.674	0.282	0.395	0.830	0.191
BEVDepth [15]	ResNet101	512 × 1408	53.5	41.2	0.565	0.266	0.358	0.331	0.190
SOLOFusion [26]	ResNet101	512 × 1408	58.2	48.3	0.503	0.264	0.381	0.246	0.207
SparseBEV† [19]	ResNet101	512 × 1408	59.2	50.1	0.562	0.265	0.321	0.243	0.195
StreamPETR† [35]	ResNet101	512 × 1408	59.2	50.4	0.569	0.262	0.315	0.257	0.199
Sparse4Dv2† [18]	ResNet101	512 × 1408	59.4	50.5	0.548	0.268	0.348	0.239	0.184
Far3D† [10]	ResNet101	512 × 1408	59.4	51.0	0.551	0.258	0.372	0.238	0.195
OPEN†	ResNet101	512 × 1408	**60.6**	**51.6**	0.528	0.266	0.312	0.222	0.190

Results

Comparison of other methods on the nuScenes test set. OPEN achieves SOTA performance with V2-99 backbone

Method	Backbone	Input Size	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
BEVDepth [15]	V2-99	640 × 1600	60.0	50.3	0.445	0.245	0.378	0.320	0.126
BEVStereo [14]	V2-99	640 × 1600	61.0	52.5	0.431	0.246	0.358	0.357	0.138
CAPE-T [38]	V2-99	640 × 1600	61.0	52.5	0.503	0.242	0.361	0.306	0.114
FB-BEV [19]	V2-99	640 × 1600	62.4	53.7	0.439	0.250	0.358	0.270	0.128
HoP [48]	V2-99	640 × 1600	61.2	52.8	0.491	0.242	0.332	0.343	0.109
StreamPETR [35]	V2-99	640 × 1600	63.6	55.0	0.479	0.239	0.317	0.241	0.119
SparseBEV [19]	V2-99	640 × 1600	63.6	55.6	0.485	0.244	0.332	0.246	0.117
Sparse4Dv2 [18]	V2-99	640 × 1600	63.8	55.6	0.462	0.238	0.328	0.264	0.115
OPEN	V2-99	640 × 1600	**64.4**	**56.7**	0.456	0.244	0.325	0.240	0.129

Ablation Studies

#	PDE	ODE	OPE	DFL	NDS↑	mAP↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
I					59.4	50.3	0.575	0.258	0.300	0.243	0.196
II	✓				59.4	50.5	0.564	0.257	0.320	0.252	0.190
III	✓	✓			59.7	50.6	0.568	0.257	0.305	0.245	**0.187**
IV	✓	✓	✓		60.8	**52.4**	0.553	0.258	0.291	0.242	0.197
V	✓	✓	✓	✓	**61.3**	52.1	**0.525**	**0.256**	**0.281**	**0.216**	0.199

- pixel-wise depth supervision can not significantly boost detection performance
- encoding object-wise depth information into the network is effective
- paying more attention to the 3D object center information can significantly reduce the mean translation error

Visualization

Compared with the ray-aware position embedding (a) and the point-aware position embedding (b), our OPEN can generate better attention weight maps for some hard-detected objects, which are highlighted by red circles.

Thanks!