EUROPEAN CONFERENCE ON COMPUTER VISION

SimPB: A Single Model for 2D and 3D Object Detection from Multiple Cameras

Yingqi Tang*, Zhaotie Meng*, Guoliang Chen, and Erkang Cheng✉

Nullmax

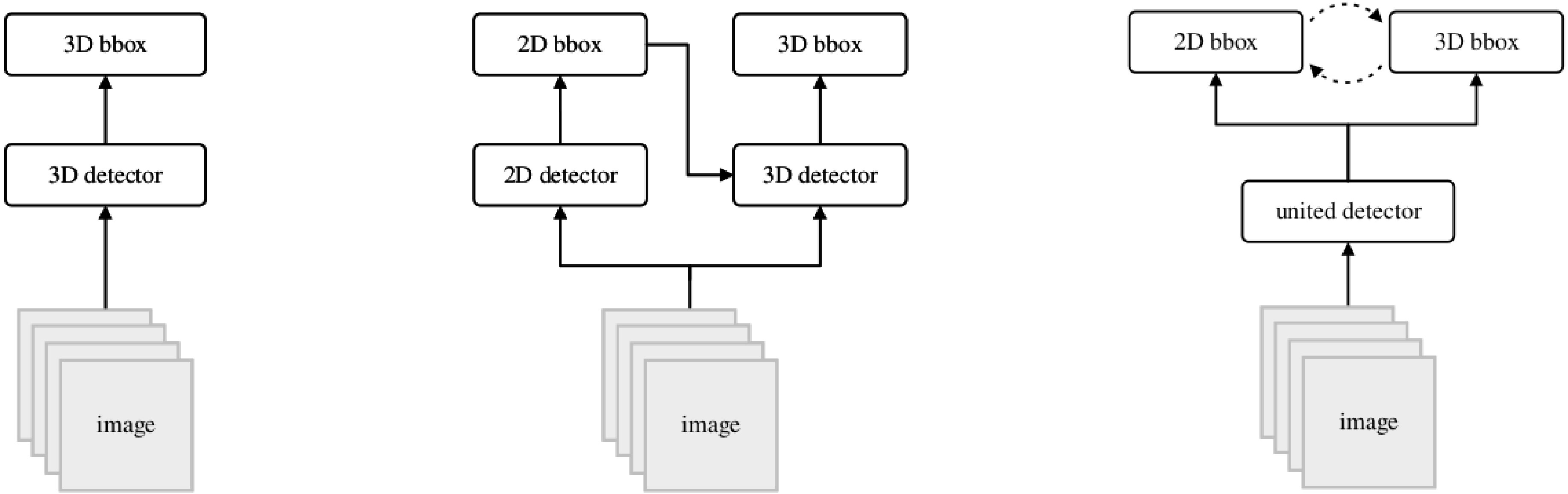
* Equal contribution

✉ Corresponding author

Motivation

Limitation of utilizing 2D boxes as priors with independent detectors in 3D object detection.

- Focus on local parts rather than capturing the global information
- 2D information is only used once during initialization
- May introduces challenges in model optimization and efficiency



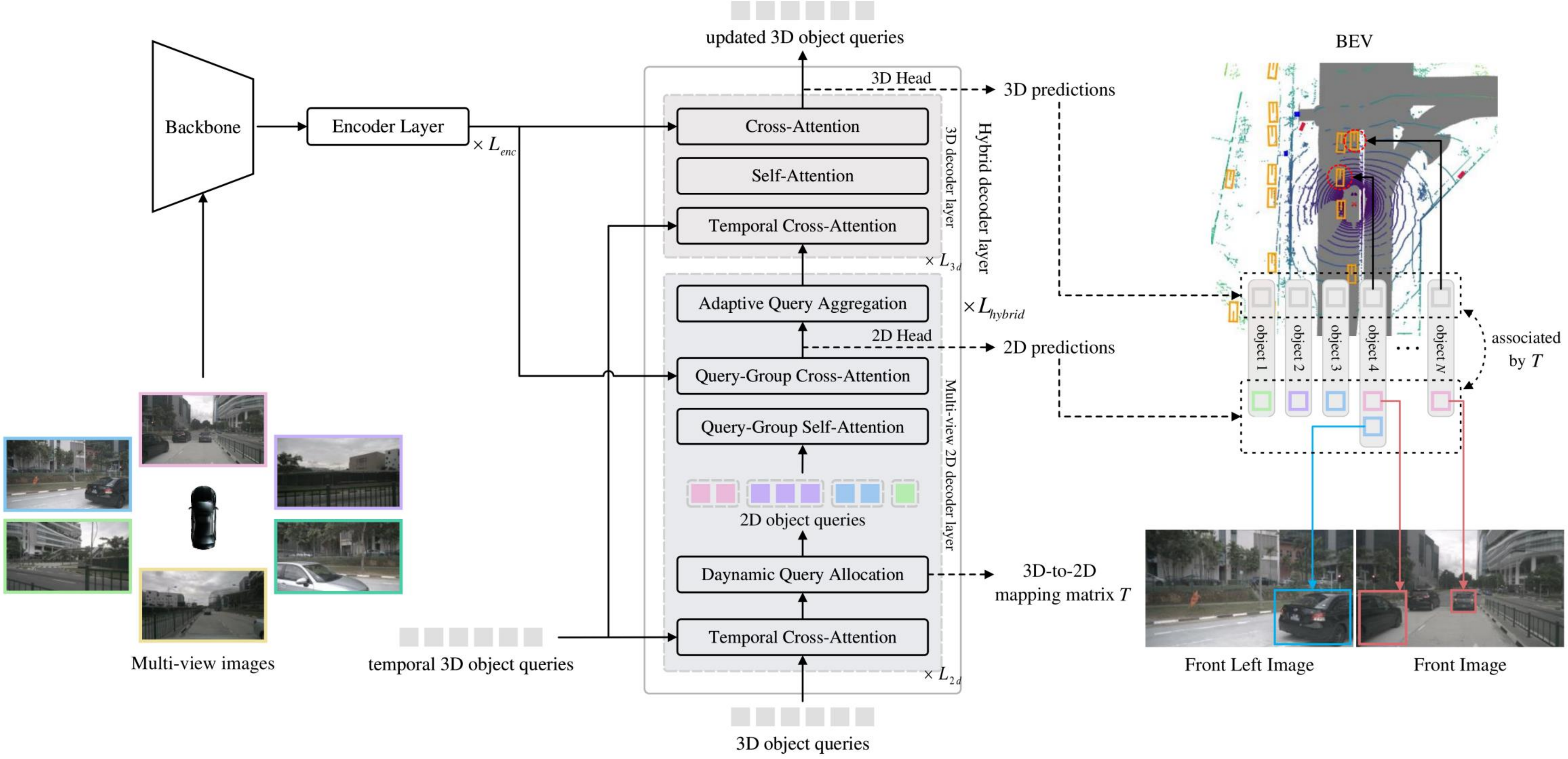
(a) Multi-view 3D Object Detection

(b) Multi-view 3D Object Detection with 2D Detection Priors

(c) United Paradigm for Multi-view 2D&3D Object Detection

Method

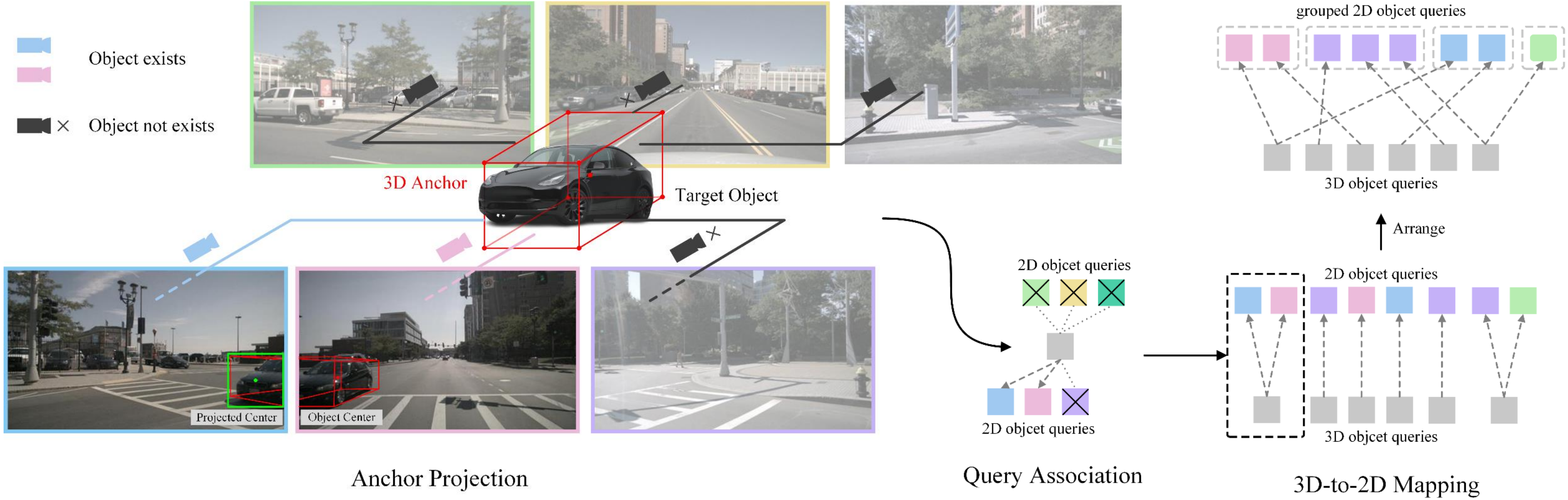
Architecture



Method

Dynamic Query Allocation

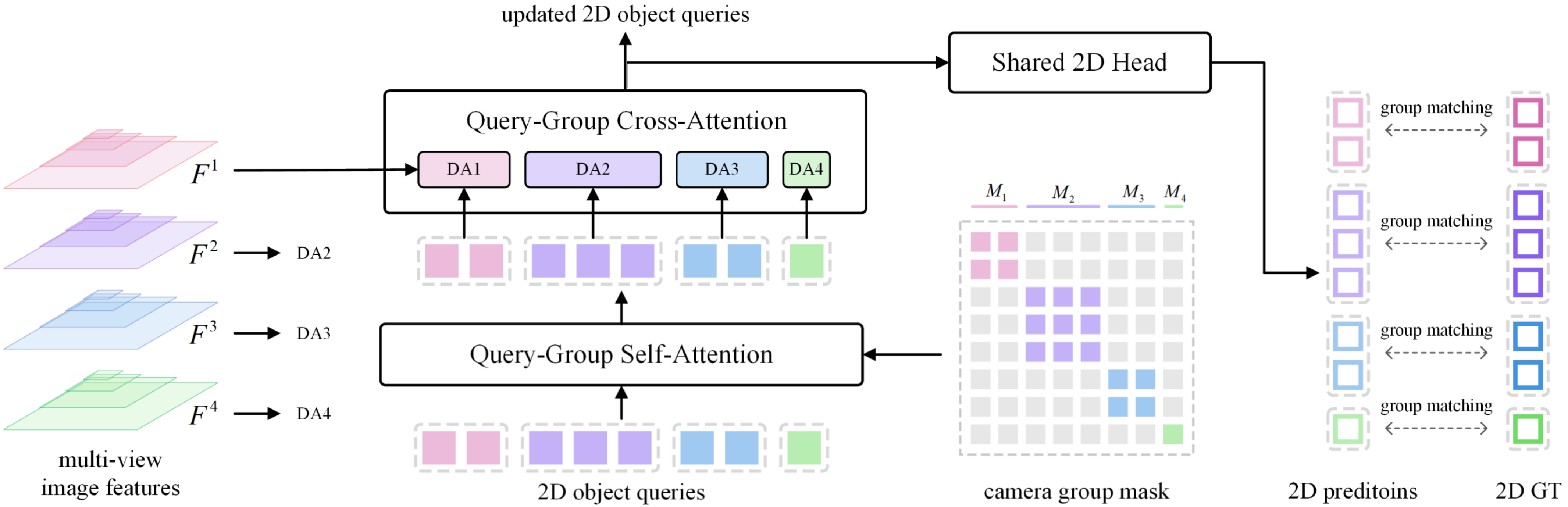
2D queries can be dynamically allocated and grouped by $Q_{2d} = T^T \cdot Q_{3d}$, where T is the 3D-to-2D mapping matrix



Method

Query Group Attention

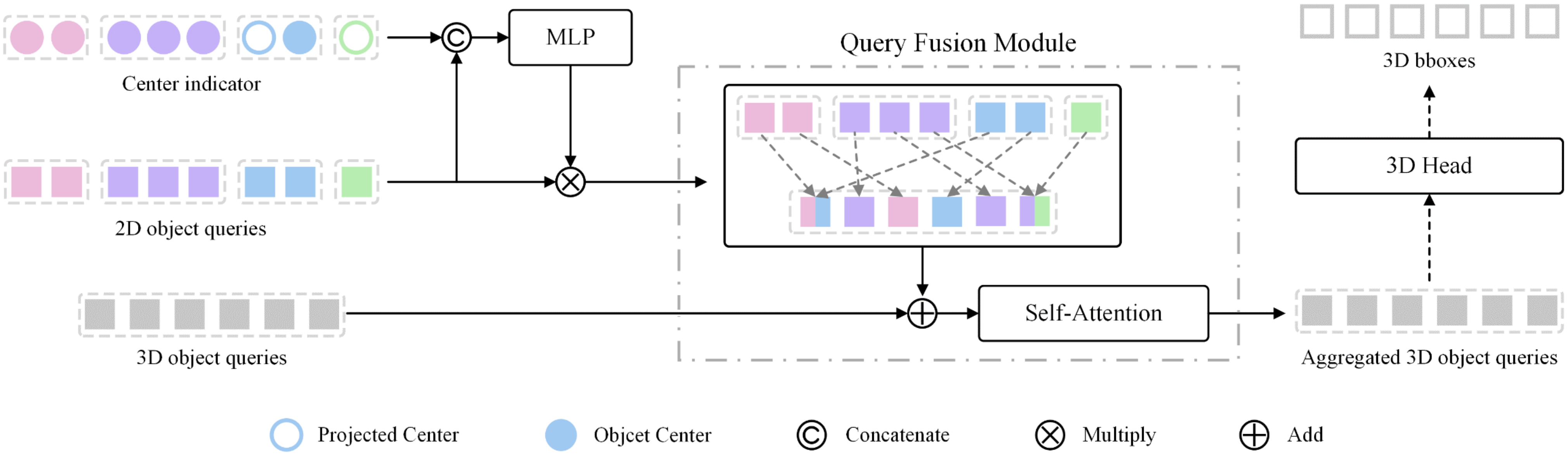
We introduce query-group self-attention and query-group cross-attention to detect targets in each view



Method

Adaptive Query Aggregation

We adaptively aggregate these grouped 2D queries using the pre-computed 3D-to-2D mapping matrix to reconstruct 3D object queries



Method

Loss Functions

$$\mathcal{L} = \mathcal{L}_{2d} + \mathcal{L}_{3d}$$

$$\mathcal{L}_{2d} = \mathcal{L}_{detr2d} + \lambda_{alpha}\mathcal{L}_{alpha}$$

$$\mathcal{L}_{alpha} = \frac{1}{M}\sum_{i=1}^{n} |\sin(\theta) - \hat{\sin}(\theta)| + |\cos(\theta) - \hat{\cos}(\theta)|$$

Experiment

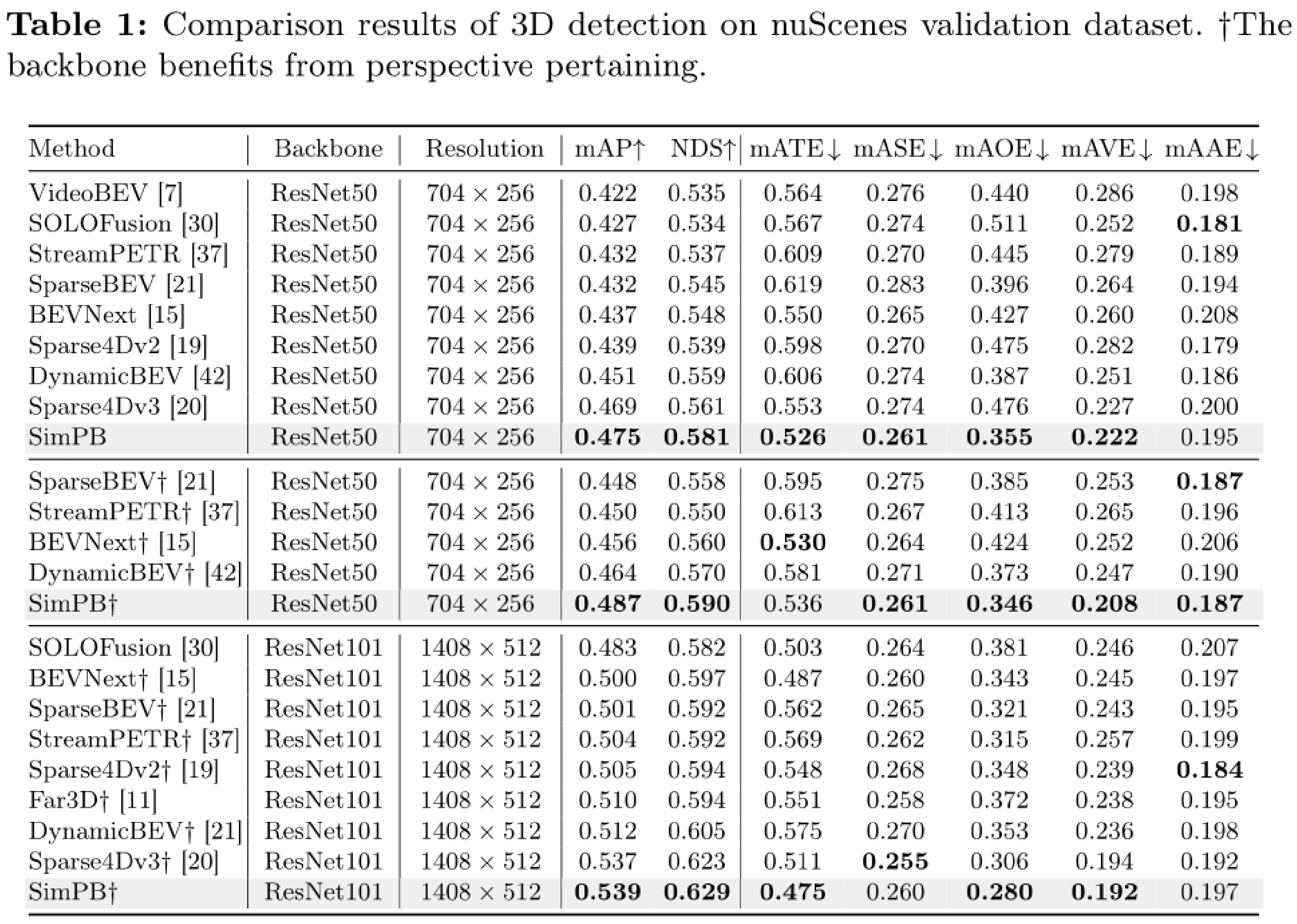


Table 1: Comparison results of 3D detection on nuScenes validation dataset. †The backbone benefits from perspective pertaining.

Method	Backbone	Resolution	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
VideoBEV [7]	ResNet50	704 × 256	0.422	0.535	0.564	0.276	0.440	0.286	0.198
SOLOFusion [30]	ResNet50	704 × 256	0.427	0.534	0.567	0.274	0.511	0.252	**0.181**
StreamPETR [37]	ResNet50	704 × 256	0.432	0.537	0.609	0.270	0.445	0.279	0.189
SparseBEV [21]	ResNet50	704 × 256	0.432	0.545	0.619	0.283	0.396	0.264	0.194
BEVNext [15]	ResNet50	704 × 256	0.437	0.548	0.550	0.265	0.427	0.260	0.208
Sparse4Dv2 [19]	ResNet50	704 × 256	0.439	0.539	0.598	0.270	0.475	0.282	0.179
DynamicBEV [42]	ResNet50	704 × 256	0.451	0.559	0.606	0.274	0.387	0.251	0.186
Sparse4Dv3 [20]	ResNet50	704 × 256	0.469	0.561	0.553	0.274	0.476	0.227	0.200
SimPB	ResNet50	704 × 256	**0.475**	**0.581**	**0.526**	**0.261**	**0.355**	**0.222**	0.195
SparseBEV† [21]	ResNet50	704 × 256	0.448	0.558	0.595	0.275	0.385	0.253	**0.187**
StreamPETR† [37]	ResNet50	704 × 256	0.450	0.550	0.613	0.267	0.413	0.265	0.196
BEVNext† [15]	ResNet50	704 × 256	0.456	0.560	**0.530**	0.264	0.424	0.252	0.206
DynamicBEV† [42]	ResNet50	704 × 256	0.464	0.570	0.581	0.271	0.373	0.247	0.190
SimPB†	ResNet50	704 × 256	**0.487**	**0.590**	0.536	**0.261**	**0.346**	**0.208**	**0.187**
SOLOFusion [30]	ResNet101	1408 × 512	0.483	0.582	0.503	0.264	0.381	0.246	0.207
BEVNext† [15]	ResNet101	1408 × 512	0.500	0.597	0.487	0.260	0.343	0.245	0.197
SparseBEV† [21]	ResNet101	1408 × 512	0.501	0.592	0.562	0.265	0.321	0.243	0.195
StreamPETR† [37]	ResNet101	1408 × 512	0.504	0.592	0.569	0.262	0.315	0.257	0.199
Sparse4Dv2† [19]	ResNet101	1408 × 512	0.505	0.594	0.548	0.268	0.348	0.239	**0.184**
Far3D† [11]	ResNet101	1408 × 512	0.510	0.594	0.551	0.258	0.372	0.238	0.195
DynamicBEV† [21]	ResNet101	1408 × 512	0.512	0.605	0.575	0.270	0.353	0.236	0.198
Sparse4Dv3† [20]	ResNet101	1408 × 512	0.537	0.623	0.511	**0.255**	0.306	0.194	0.192
SimPB†	ResNet101	1408 × 512	**0.539**	**0.629**	**0.475**	0.260	**0.280**	**0.192**	0.197

Experiment

SimPB consistently delivers the best results across all 2D evaluation metrics.

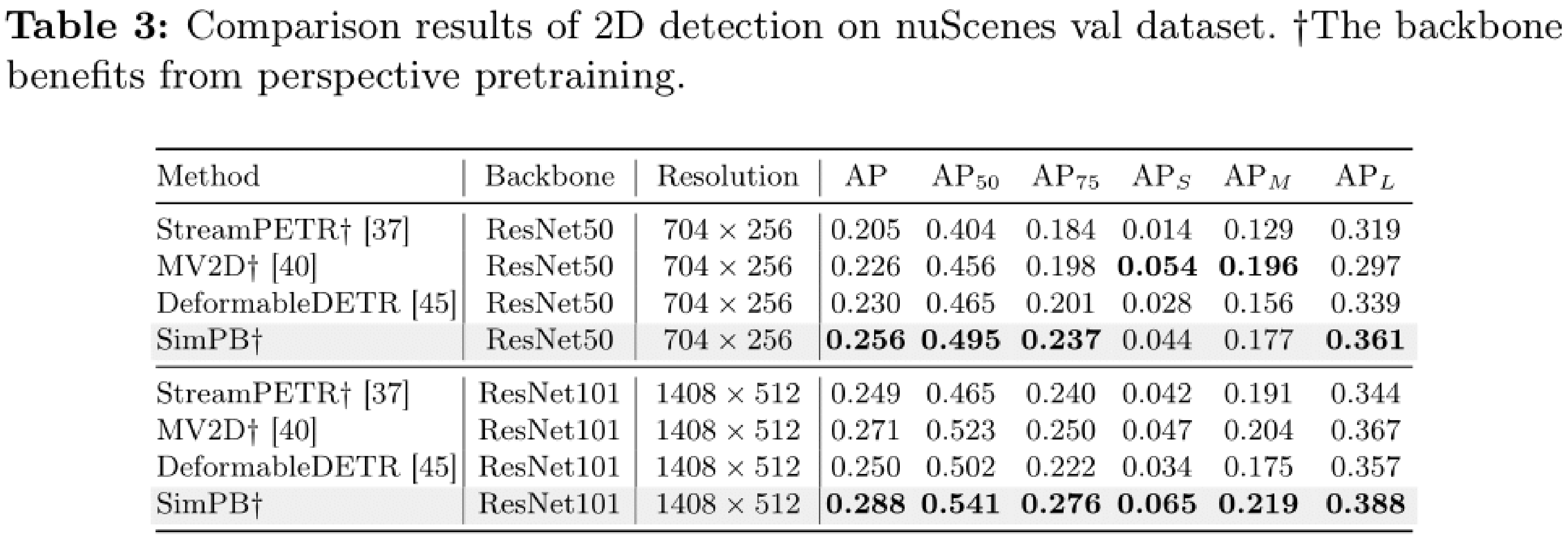


Table 3: Comparison results of 2D detection on nuScenes val dataset. †The backbone benefits from perspective pretraining.

Method	Backbone	Resolution	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
StreamPETR† [37]	ResNet50	704 × 256	0.205	0.404	0.184	0.014	0.129	0.319
MV2D† [40]	ResNet50	704 × 256	0.226	0.456	0.198	**0.054**	**0.196**	0.297
DeformableDETR [45]	ResNet50	704 × 256	0.230	0.465	0.201	0.028	0.156	0.339
SimPB†	ResNet50	704 × 256	**0.256**	**0.495**	**0.237**	0.044	0.177	**0.361**
StreamPETR† [37]	ResNet101	1408 × 512	0.249	0.465	0.240	0.042	0.191	0.344
MV2D† [40]	ResNet101	1408 × 512	0.271	0.523	0.250	0.047	0.204	0.367
DeformableDETR [45]	ResNet101	1408 × 512	0.250	0.502	0.222	0.034	0.175	0.357
SimPB†	ResNet101	1408 × 512	**0.288**	**0.541**	**0.276**	**0.065**	**0.219**	**0.388**

Experiment

Cyclic interaction between multi-view 2D and 3D layers and provides the best performance

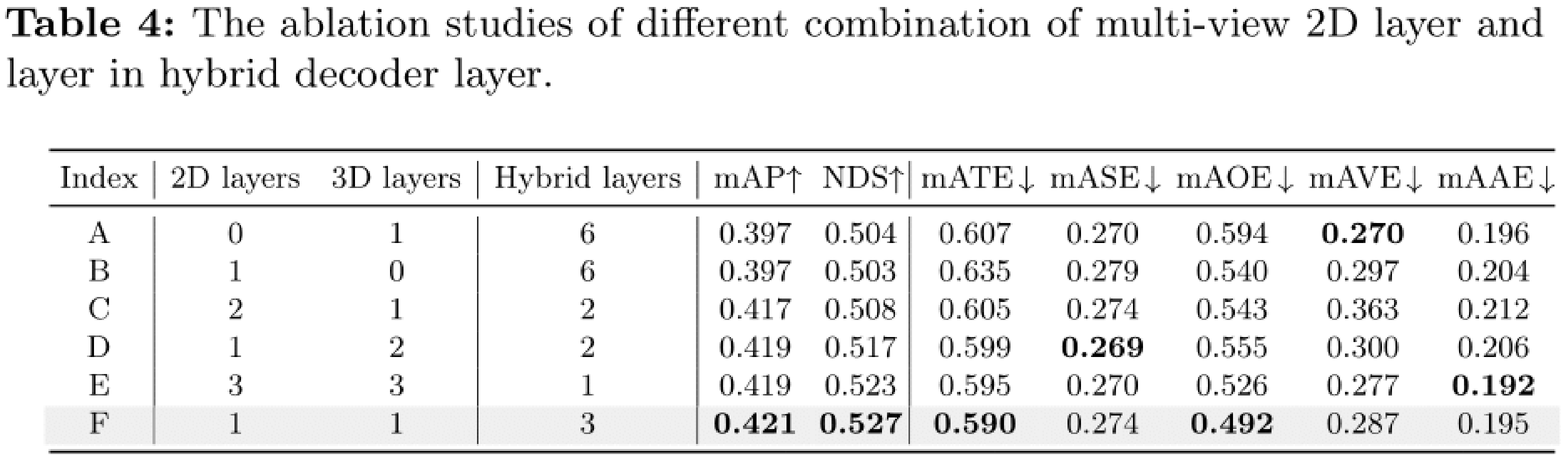
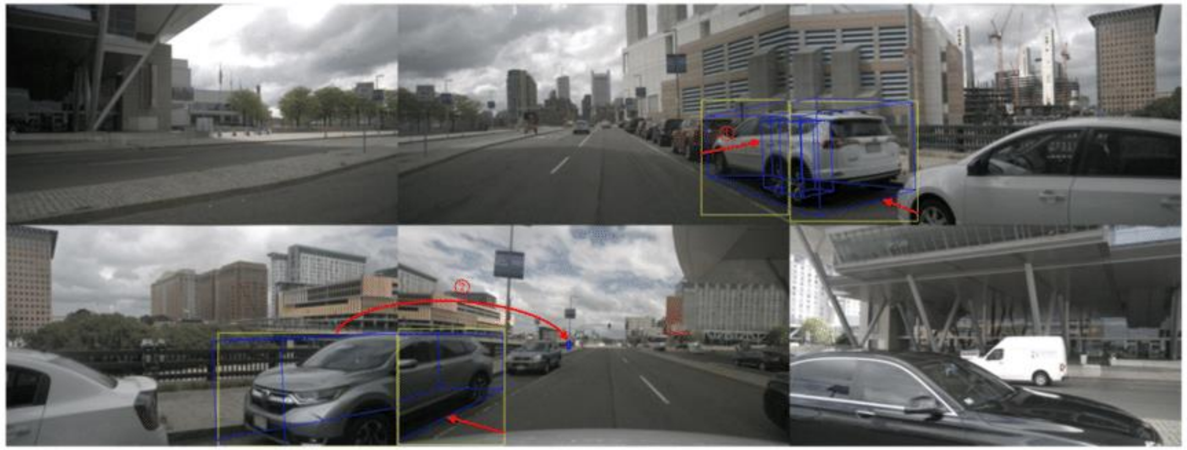


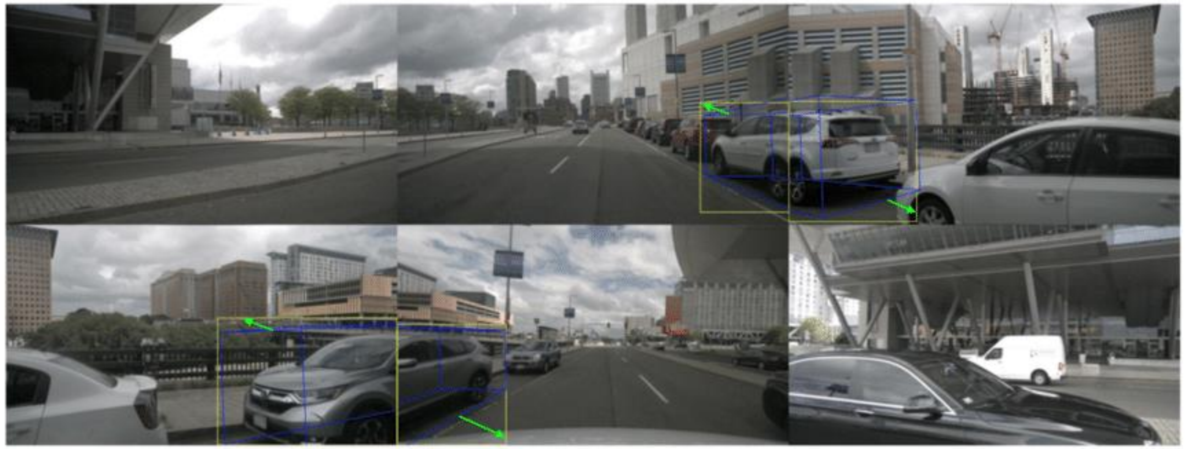
Table 4: The ablation studies of different combination of multi-view 2D layer and 3D layer in hybrid decoder layer.

Index	2D layers	3D layers	Hybrid layers	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
A	0	1	6	0.397	0.504	0.607	0.270	0.594	**0.270**	0.196
B	1	0	6	0.397	0.503	0.635	0.279	0.540	0.297	0.204
C	2	1	2	0.417	0.508	0.605	0.274	0.543	0.363	0.212
D	1	2	2	0.419	0.517	0.599	**0.269**	0.555	0.300	0.206
E	3	3	1	0.419	0.523	0.595	0.270	0.526	0.277	**0.192**
F	1	1	3	**0.421**	**0.527**	**0.590**	0.274	**0.492**	0.287	0.195

Qualitative Results



(a) The detection results of MV2D. 2D-to-3D association (red arrow) may produce duplicate 3D results or unrelated results from 2D priors for a cross-camera target.

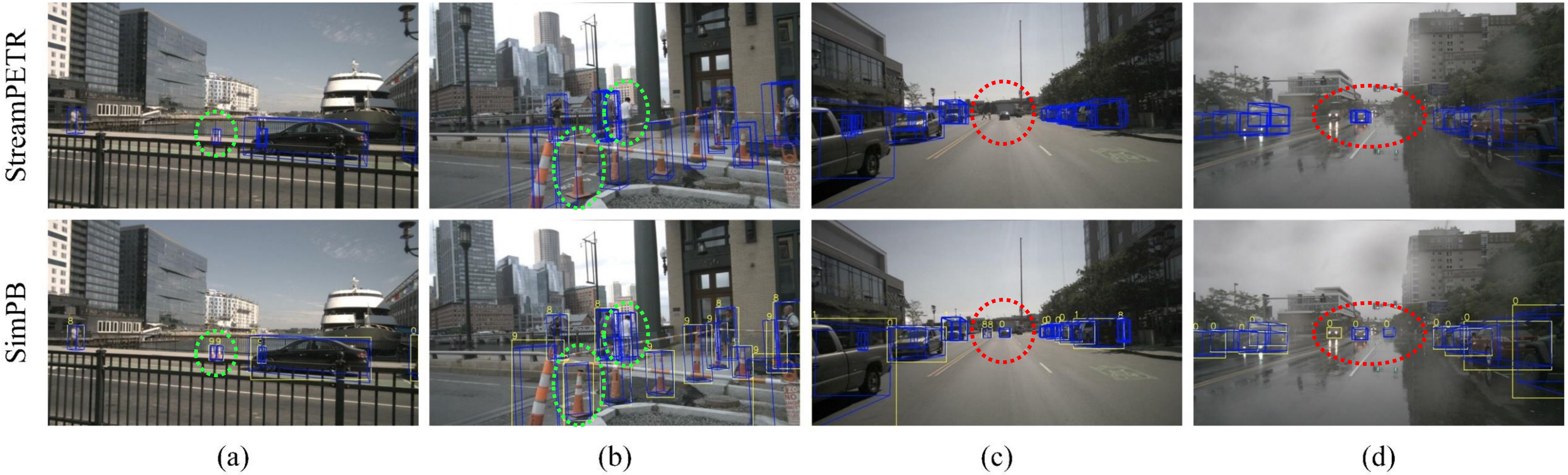


(b) The detection results of SimPB. The process of 3D-to-2D association (green arrow) effectively yields accurate 3D results along with their corresponding 2D boxes for cross-camera targets.

Wang, Z. et al. Equipping any 2d object detector with 3d detection ability. ICCV 2023

Qualitative Results

SimPB provides more precise results and successfully distinguishes crowded and small objects



Wang, S. et al. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. ICCV 2023

Conclusion

- We introduce a **single-stage** query-based method called **SimPB** for multi-view 2D and 3D object detection.

- To interact 2D and 3D objects in a cyclic manner, we propose a **dynamic query allocation** and **adaptive query aggregation** module within the hybrid decoder.

- We also apply **query-group attention** to strengthen the interaction among 2D queries within a specific camera.

- We extensively evaluate SimPB in the Nuscenes dataset with **comprehensive experiments** for both 2D and 3D tasks.