DuoSpaceNet: Leveraging Both Bird's-Eye-View and Perspective View Representations for 3D Object Detection

Zhe Huang*, Yizhe Zhao*, Hao Xiao, Chenyan Wu, Lingting Ge

Intro

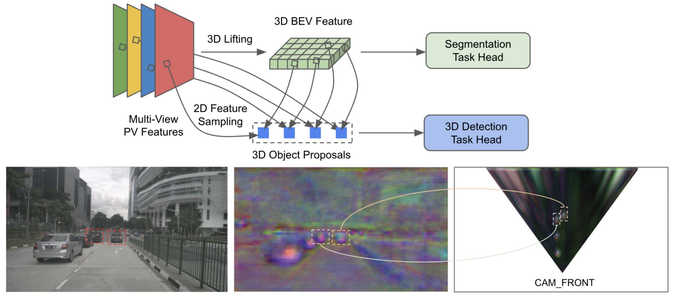

Figure 1. DuoSpaceNet (ours) where 3D detection benefits from both 3D BEV and 2D PV feature space. For instance, it is easier to interpret 3D positions from the above BEV heatmap, whereas PV heatmap preserves finer semantic details at higher resolution.

- In this paper, we propose DuoSpaceNet, a novel framework that fully unifies BEV and PV feature spaces within a single detection pipeline for comprehensive 3D perception. Our design includes a decoder to integrate BEV–PV features into unified detection queries, as well as a feature enhancement strategy that enriches different feature representations. In addition, DuoSpaceNet can be extended to handle multiframe inputs, enabling more robust temporal analysis.

Design

- Multi-view 2D perspective view (PV) features are extracted by the backbone and the feature pyramid network (FPN) [1]. Our 2D to 3D BEV lifting strategy consists of a parameter-free voxel projection following SimpleBEV [2] and a divergence enhancement process to make resulting BEV features more distinctive w.r.t. PV features. In our duo space framework, multi-view PV features and BEV features are identified as equally important and are fed into the decoder together. Each decoder layer has one self-attention layer and two deformable cross-attention layers [3]. The self-attention layer acts on both BEV and PV spaces, whereas each cross-attention layer only attends to either BEV features or PV features.

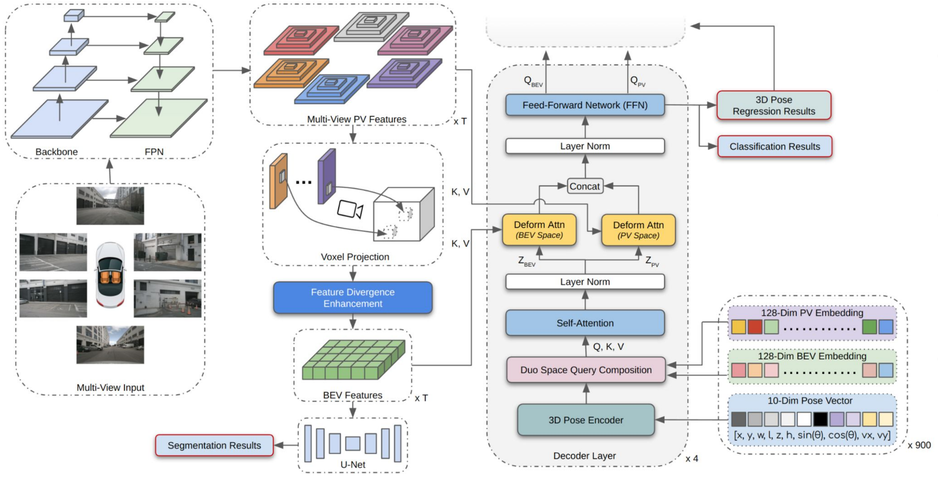

Figure 2. Overall architecture of the proposed DuoSpaceNet.

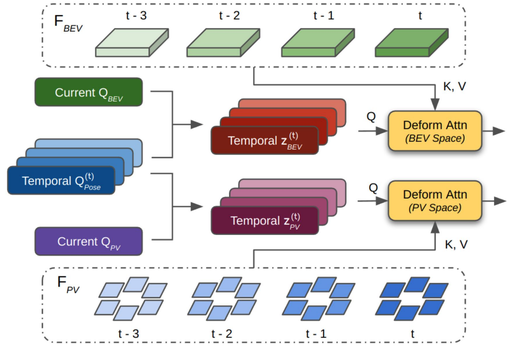

Figure 3. Diagram of the proposed duo space temporal modeling. Temporal pose embeddings. Pose are generated by warping pose vectors at current timestamp through motion compensation. Subsequently, temporal duo space queries are assembled by broadcasting current content embeddings over the time dimension and then combining them with the temporal pose embeddings. We then conduct space-specific cross-attention using recent BEV and PV feature maps, both of which are maintained by their respective memory queue.

Results

Method	Image Size	Frames	mAP↑	NDS↑	mATE↓	mASE↓	mAOE↓	mAVE↓	mAAE↓
DETR3D [48]	1600 × 900	1	0.349	0.434	0.716	0.268	0.379	0.842	0.200
PETR [26]	1600 × 900	1	0.370	0.442	0.711	0.267	0.383	0.865	0.201
BEVFormer-S [18]	1600 × 900	1	0.375	0.448	0.725	0.272	0.391	0.802	0.200
Sparse4D [21]	1600 × 640	1	0.382	0.451	0.710	0.279	0.411	0.806	0.196
SimMOD [64]	1600 × 900	1	0.366	0.455	0.698	0.264	0.340	0.784	0.197
DuoSpaceNet (Ours)	1600 × 640	1	**0.399**	**0.462**	0.683	0.279	0.376	0.829	0.205
UVTR [15]	1600 × 900	6	0.379	0.483	0.731	0.267	0.350	0.510	0.200
BEVFormer [18]	1600 × 900	4	0.416	0.517	0.673	0.274	0.372	0.394	0.198
PETRv2 [27]	1600 × 640	6	0.421	0.524	0.681	0.267	0.357	0.377	0.186
Sparse4D [21]	1600 × 640	4	0.436	0.541	0.633	0.279	0.363	0.317	0.177
OCBEV [38]	1600 × 900	4	0.417	0.532	0.629	0.273	0.339	0.342	0.187
DFA3D-MvACon [25]	1600 × 900	4	0.432	0.535	0.664	0.275	0.344	0.323	0.207
DuoSpaceNet (Ours)	1600 × 640	4	**0.443**	**0.547**	0.603	0.275	0.360	0.314	0.195

Table 1. Comparison on 3D detection results on nuScenes val set.

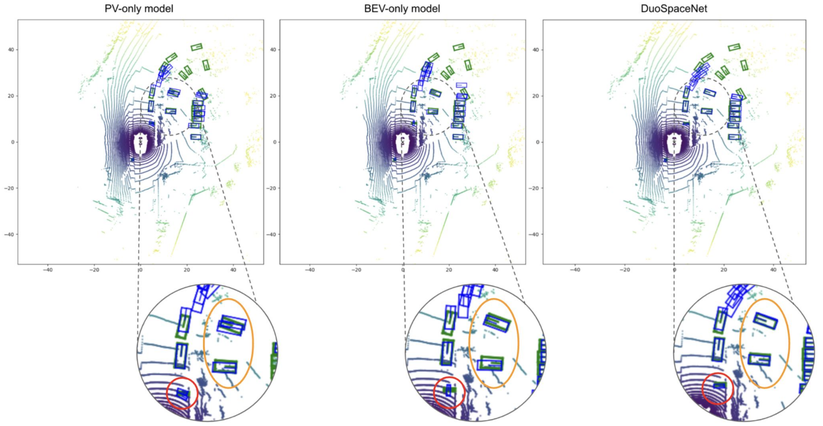

Figure 4. Qualitative comparison of top-down 3D detection results among our method, PV-only, and BEV-only models. Ground truth bounding boxes are in **green** and predictions are in **blue**. Our prediction aligns most accurately with the ground truth.

[1] Lin, Tsung-Yi, et al. "Feature pyramid networks for object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[2] Harley, Adam W., et al. "Simple-bev: What really matters for multi-sensor bev perception?." 2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023.

[3] Zhu, Xizhou, et al. "Deformable detr: Deformable transformers for end-to-end object detection." arXiv preprint arXiv:2010.04159 (2020).

For references in Tab. 1, please refer to the bibliography in our paper.