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

Supplementary Material

6. Additional details on Duo Space Temporal Modeling.

Specifically, we generate temporal duo space queries $\mathbf{z}_{BEV}^{(t)}$ and $\mathbf{z}_{PV}^{(t)}$ by infusing past information into shared 3D poses. Assuming the current timestamp is T and the temporal length is l frames, we compute temporal poses $\mathcal{P}_i^{(t)}, i \in \{1, 2, ..., k\}, t \in \{T-l+1, T-l+2, ..., T\}$. Ego-motion compensation can be done via a warp transformation matrix from timestamp $t-1$ to t, denoted as $[\boldsymbol{R} \,|\, \boldsymbol{t}]_{(t)}^{(t-1)}, t \in \{T-l+1, T-l+2, ..., T\}$, where $\boldsymbol{R}$ and $\boldsymbol{t}$ refer to the rotational and translational components in the matrix. Object-motion compensation, on the other hand, can be tackled using the predicted velocity of each query, assuming a constant velocity motion model over the l-length sequence. Adding up both compensations, we update the object location x, y, z at $t-1$, dubbed $\mathcal{P}_i^{(t-1)}|_{x,y,z}$, the object orientation $\sin\theta, \cos\theta$ at $t-1$, dubbed $\mathcal{P}_i^{(t-1)}|_{\theta}$ and the object velocity vx, vy at $t-1$, dubbed $\mathcal{P}_i^{(t-1)}|_{vx,vy}$, through

$$\mathcal{P}_i^{(t)}|_{x',y'} = \mathcal{P}_i^{(t)}|_{x,y} - \Delta t \cdot \mathcal{P}_i^{(t)}|_{vx,vy}, \tag{10}$$

$$\mathcal{P}_i^{(t-1)}|_{x,y,z} = [\boldsymbol{R} \,|\, \boldsymbol{t}]_{(t)}^{(t-1)} \mathcal{P}_i^{(t)}|_{x',y',z}, \tag{11}$$

$$\mathcal{P}_i^{(t-1)}|_{\theta} = \boldsymbol{R}_{(t)}^{(t-1)} \mathcal{P}_i^{(t)}|_{\theta}, \tag{12}$$

$$\mathcal{P}_i^{(t-1)}|_{vx,vy} = \boldsymbol{R}_{(t)}^{(t-1)} \mathcal{P}_i^{(t)}|_{vx,vy}, \tag{13}$$

$$i \in \{1, 2, ..., k\}, t \in \{T-l+1, T-l+2, ..., T\},$$

where Δt represents the wall-clock time difference between adjacent frames. We compute the temporal pose embedding for each timestamp t as follows:

$$\left(Q_{Pose}^{i}\right)^{(t)} = \xi\left(\texttt{Enc}(\mathcal{P}_i^{(t)})\right). \tag{14}$$

We then generate $\mathbf{z}_{BEV}^{(t)}$, $\mathbf{z}_{PV}^{(t)}$, $\hat{\mathbf{p}}_{BEV}^{(t)}$ and $\hat{\mathbf{p}}_{PV}^{(t)}$, $t \in \{T-l+1, T-l+2, ..., T\}$ according to Eq. 2, 3, 6 & 8 as temporal inputs. Finally, after cross-attention layers (Eq. 7 and 9), we aggregate temporal outputs via 3-layer MLP before the FFN of each decoder layer. An illustration of temporal cross-attention in PV space is shown in Fig. 6. The temporal cross-attention in BEV space is identical expect for the use of BEV space queries and BEV features as input.

7. Additional details on experiment settings.

Following [21], we initialize the x, y, z coordinates of pose vectors using K-Means centroids on nuScenes `training` set [1]. For all experiments, we use AdamW optimizer [30] and a cosine learning rate scheduler [29]. The initial learning rate for backbone and other modules are 2e-5 and 2e-4, respectively. No data augmentation is used other than the grid mask used in DETR3D [48]. The perception ranges for both the X and Y axes are $[-51.2m, 51.2m]$, which are consistent for both the 3D object detection and map segmentation tasks.

When it comes to the loss functions we use to train the 3D detection, we utilize Focal Loss [20] for bounding box classification and L1 Loss for attribute regression. Duo space queries are assigned to their ground truth via Hungarian Matching introduced in DETR [2]. For segmentation, we use a combination of L1 Loss, Cross Entropy Loss and Dice Loss [40] for each predicted mask.

8. Additional Visualizations

In Fig. 7, the 3D detection results are displayed in the perspective camera view for the same example as shown in Fig. 5, comparing three different detection methods along with the ground truth.

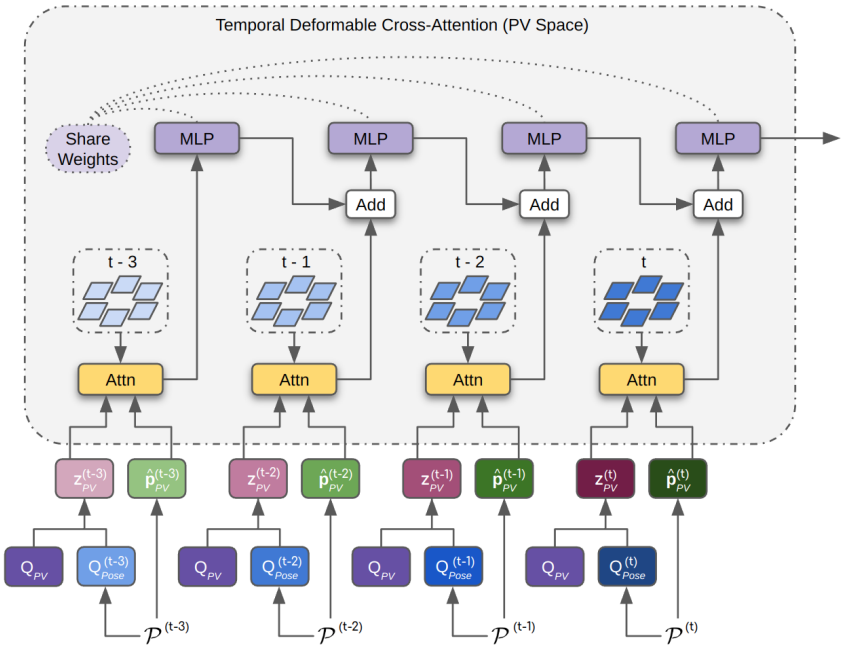

Figure 6. An illustration of space-specific temporal deformable cross-attention in perspective view (PV) space, with 4 temporal frames. Temporal pose vectors are generated by transforming current the pose vector at current frame to previous frames with motion compensation. After the duo space query composition, duo space temporal queries are formulated for both spaces. Subsequently, in PV space, attention queries, $\mathbf{z}_{PV}^{(\cdot)}$, and their reference points, $\hat{\mathbf{p}}_{PV}^{(\cdot)}$, from timestamp $t-3$ to t are used as input queries for multi-scale deformable attention [68]. Each set of queries at a given timestamp only attends to corresponding PV features at the same timestamp. After the attention mechanism, results are aggregated by MLPs in a recurrent manner. Note that weights are shared across all MLPs.

Figure 7. Visualization of 3D detection results in perspective camera view. Different colors represent different categories. Our method achieves the best prediction result for the motorcycle instance w.r.t. its 3D position as well as its orientation, showcasing the effectiveness of incorporating both PV and BEV information in detection queries.