KI DELTA LEARNING
Scalable AI for Automated Driving

SpatialDETR: 3D Object Detection from Multi-View Camera Images with Global Cross-Sensor Attention

Simon Doll, Richard Schulz,
Lukas Schneider, Viviane Benzin,
Markus Enzweiler, Hendrik P. A. Lensch

Overview

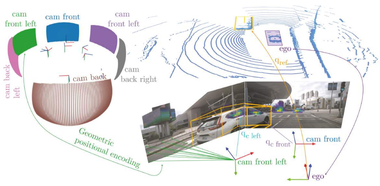

Figure 1: Overview of the Spatial Attention block and Geometric Positional Encoding. (© Mercedes-Benz, Motional AD Inc.)

Objective and Contribution

We propose a novel method for multi-view 3D object detection.

Threefold Contribution

- A 3D geometric positional encoding
- Cross-sensor global attention for arbitrary sensor setups
- Explicit modelling of sensor extrinsics

Architecture

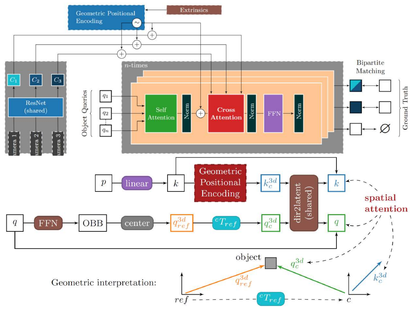

Figure 2: SpatialDETR Architecture. (© Mercedes-Benz)

- Each transformer layer uses dot product attention

$$Attn(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q} \cdot \boldsymbol{K}^T}{\sqrt{d_k}}\right) \cdot \boldsymbol{V}$$

- Geometric similarity of queries and keys is incorporated with a shared MLP (*dir2latent*).

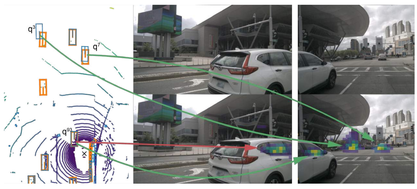

Figure 3: Visualization of attention weights for different objects queries in the last decoder layer (© Mercedes-Benz, Motional AD Inc.)

Sensor-Relative Attention

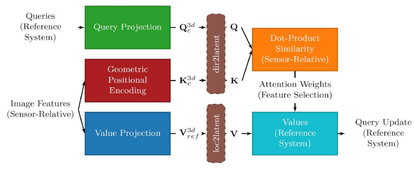

Figure 4: Sensor-Relative Similarity Computation. (© Mercedes-Benz)

Query-key similarity is computed in the sensor frame whilst the query updates are performed in reference coordinates.

Ablations & Insights

Performance for different object classes in terms of mAP.

Name	Truck	Bus	Trailer	Pedestrian	Traffic Cone	FPS
DETR3D	0.286	0.347	0.167	**0.424**	**0.529**	**2.5**
SpatialDETR	**0.302**	**0.378**	**0.176**	0.418	0.514	2.4

Table 1: SpatialDETR evaluation on the nuScenes val-set. Runtime on a single NVIDIA V100 GPU.

- Global cross-sensor context improves the detection of large objects
- Incorporating multi-scale information could boost the performance for small objects even further

Project Page & Code

For more details check out our paper and code!

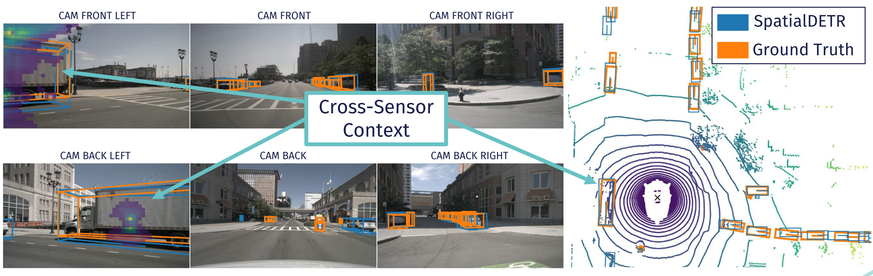

Figure 5: Multi-camera input images (left), top-view model output (right). Predictions are marked in blue, ground truth annotations in orange. LiDAR point cloud is shown for visualization only. (© Mercedes-Benz, Motional AD Inc.)

Knowledge Transfer

Partners

External partners

For more information contact:
simon.doll@mercedes-benz.com
richard.schulz@mercedes-benz.com

KI Delta Learning is a project of the KI Familie. It was initiated and developed by the VDA Leitinitiative autonomous and connected driving and is funded by the Federal Ministry for Economic Affairs and Climate Action.

Supported by:
Federal Ministry for Economic Affairs and Climate Action
on the basis of a decision by the German Bundestag

www.ki-deltalearning.de @KI_Familie KI Familie