Appendix

Anonymous Author(s)
Affiliation
Address
email

1 Method

1.1 Coordinate transformation among LiDAR, camera and EGB image in DAIR-V2X dataset

If the LiDAR coordinate system is regarded as the world coordinate system, the 3D coordinate of point W could be:

$$W_{world} = \begin{bmatrix} x_{world} \\ y_{world} \\ z_{world} \end{bmatrix} \tag{1}$$

We also have the camera coordinate and image coordinate of point W as

$$W_{cam} = \begin{bmatrix} x_{cam} \\ y_{cam} \\ z_{cam} \end{bmatrix}, W_{img} = \begin{bmatrix} x_{img} \\ y_{img} \end{bmatrix} \tag{2}$$

Its world homogeneous coordinate in camera coordinate system and its camera homogeneous coordinate in image coordinate system are

$$W_{world\_h} = \begin{bmatrix} x_{world} \\ y_{world} \\ z_{world} \\ 1 \end{bmatrix}, W_{img\_h} = \begin{bmatrix} x_{img} \\ y_{img} \\ 1 \end{bmatrix} \tag{3}$$

Suppose that E is the transformation matrix from LiDAR coordinate system to camera coordinate system and I is the transformation matrix from camera coordinate system to image coordinate system, The inverse matrix of E and matrix I are

$$E^{-1}_{4\times4} = \begin{bmatrix} r_{x1} & r_{y1} & r_{z1} & t_x \\ r_{x2} & r_{y2} & r_{z2} & t_y \\ r_{x3} & r_{y3} & r_{z3} & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix}, I_{3\times3} = \begin{bmatrix} f_x & 0 & u \\ 0 & f_y & v \\ 0 & 0 & 1 \end{bmatrix} \tag{4}$$

Generally, E and I are the extrinsic matrix and intrinsic matrix of camera, which are given by dataset [1].

Then we have

$$W_{cam\_h} = E_{4\times4} * W_{world_h}, W_{img} = \frac{1}{z_c am} * I_{3\times3} * W_{cam} \tag{5}$$

1.2 Generate preference map

We visualize four typical cases when generating one cell of preference map, which are shown in Fig 1.

Submitted to the 7th Conference on Robot Learning (CoRL 2023). Do not distribute.

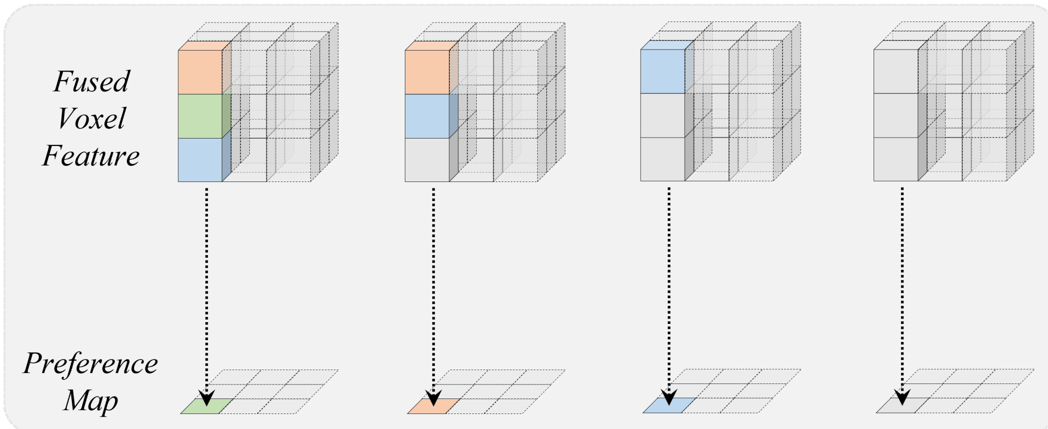


Figure 1: Four typical cases when generating one cell of preference map.

2 Experiments

2.1 Detailed settings of architecture

We follow the default settings in OpenCOOD [2] codebase, which is also shown in Tab 2.

Table 1: Details of unified network architecture.

Blocks	Settings
Voxel Feature Encoder (VFE)	use normalization and absolute 3D coordinates, 64 filters
PointPillar Scatter	64-channel output
BEV backbone	ResNet backbone with ordered layers=[3, 4, 5], strides=[2, 2, 2], filters=[64, 128, 256], upsample_strides=[1, 2, 4], upsample_filters=[128, 128, 128]
Shrink Header	shrink from 384 channels to 256 channels with stride 3
Detect Head	256-channel output with 2 anchors

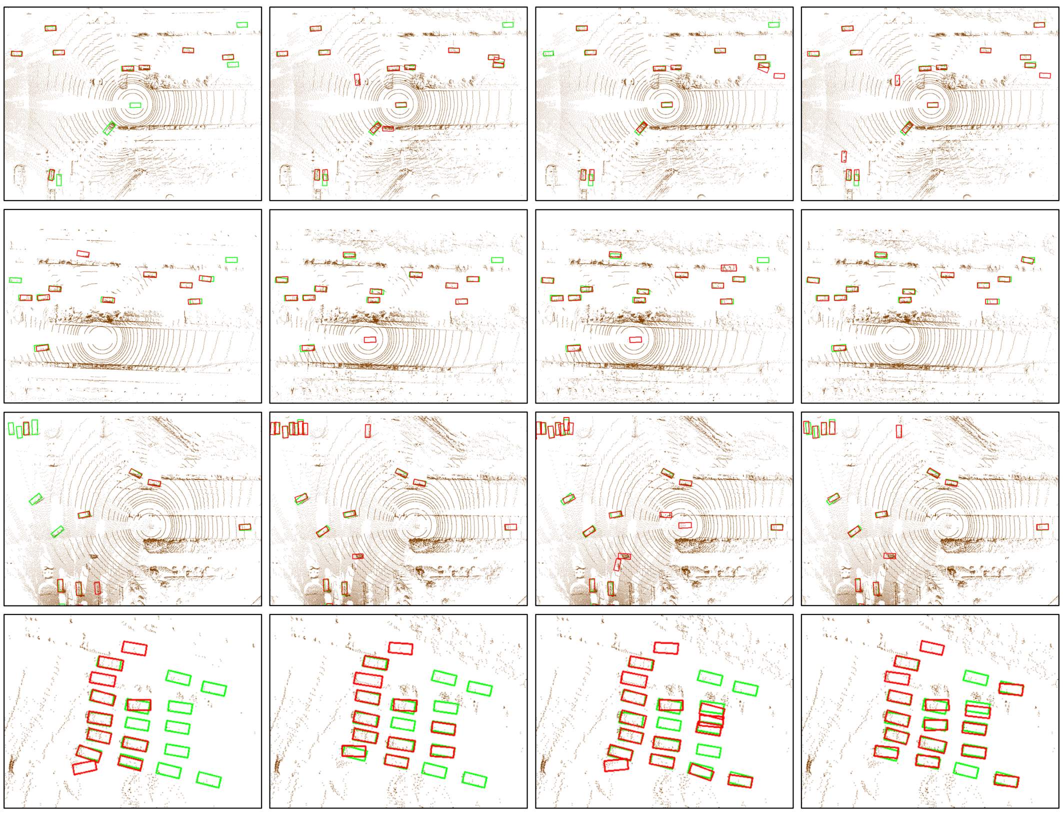
2.2 Detailed settings of experiments

Table 2: Details of unified network architecture.

Method	optimizer	lr schedule	initial lr
No Fusion	Adam	multistep	1e-3
Late Fusion	Adam	multistep	1e-3
When2com (CVPR'20)	Adam	multistep	1e-3
V2VNet (ECCV'20)	Adam	multistep	1e-3
DiscoNet (NeurIPS'21)	Adam	multistep	2e-3
CoBEVT (CoRL'22)	Adam	multistep	2e-3
V2X-ViT (ECCV'22)	Adam	multistep	2e-3
Where2comm (NeurIPS'22)	Adam	multistep	2e-3
BM2CP	Adam	multistep	1e-3

2.3 More Visualizations

Fig 2 shows more comparisons with *No Fusion*, *V2X-ViT* and *Where2comm*.



(a) No Fusion (b) V2X-ViT (c) Where2comm (d) BM2CP

Figure 2: More qualitative results in DAIR-V2X dataset. Ground truths are colored in green and predictions are colored in red.

References

[1] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan, et al. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21361–21370, 2022.

[2] R. Xu, H. Xiang, X. Xia, X. Han, J. Li, and J. Ma. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2583–2589. IEEE, 2022.