# – Supplementary material –
# AeDet: Azimuth-invariant Multi-view 3D Object Detection

## 1. Discussion

**Difference between AeConv and previous works.** The main difference between AeConv and previous works (*e.g.* DCN [1], STN [3] and MVDeTr [2]) is: the offset of AeConv is azimuth-equivalent and designed according to the inherent property of the BEV features, while the offset/transformation of the previous works is learned from the features. In practice, the learning of the offset is difficult in BEV space of multi-view 3D detection for the following reasons: (1) The offset/transformation convolution is translation-invariant, which is unfriendly for learning the azimuth-equivalent offset. (2) The projected BEV features are inaccurate and noisy because the depth predicted from the monocular camera is inaccurate. Experimentally, we replaced the convolution with DCN in BEV network, and the model crashed during training. Instead, the offset of AeConv is designed according to the radial symmetry of the BEV features, providing a strong prior to the network. It thus largely reduces the variability of data and the learning difficulty.

**Cooperation between AeConv and azimuth-equivariant anchor.** AeConv and azimuth-equivariant anchor should work collaboratively to preserve the consistency between the representation learning and predictions. As shown in Figure 3 in the paper, if we use AeConv only, we'll get the same network output for the three cars. However, the typical anchor yields different targets for the three cars, which conflicts with AeConv. In contrast, the azimuth-equivariant anchor yields the same targets for the three cars, ensuring the consistency between the network outputs and targets. As shown in Table S1, if we apply them separately, worse results (*i.e.* 44.2%→31.8% NDS and 44.2%→41.8% NDS) are obtained. However, if we apply them together, it yields a better result (*i.e.* 44.2%→47.3% NDS).

| AeConv | AeAnchor | mAP↑ | NDS↑ |
|---|---|---|---|
| | | 0.330 | 0.442 |
| ✓ | | 0.323 | 0.318 |
| | ✓ | 0.335 | 0.418 |
| ✓ | ✓ | 0.358 | 0.473 |

Table S1. Ablation study of AeConv and azimuth-equivariant anchor. AeAnchor denotes the azimuth-equivariant anchor.

## References


[1] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, pages 764–773, 2017. 1

[2] Yunzhong Hou and Liang Zheng. Multiview detection with shadow transformer (and view-coherent data augmentation). In *ACMMM*, pages 1673–1682, 2021. 1

[3] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *NIPS*, 28, 2015. 1