# 3D Video Object Detection with Learnable Object-Centric Global Optimization Supplementary Material

Jiawei He[1,2] Yuntao Chen[3] Naiyan Wang[4] Zhaoxiang Zhang[1,2,3]
[1] CRIPAC, Institute of Automation, Chinese Academy of Sciences (CASIA)
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS)
[3] Centre for Artificial Intelligence and Robotics, HKISI_CAS [4] TuSimple
{hejiawei2019, zhaoxiang.zhang}@ia.ac.cn {chenyuntao08, winsty}@gmail.com

## A. Network Architecture

In this section, we explain the detailed network design of BA-Det. The backbone is a standard 'DLASeg' architecture, please refer to [4] for more details. Then two detection heads are added in the first-stage detector. The cls head contains a 3×3 Conv2d layer, and an FC to predict the classification. The reg head contains 8 independent '3×3 Conv2d layer + FC' modules, regressing '2d dim', '3d offset', 'corner offset', 'corner uncertainty','3d dim', 'ori cls', 'ori offset', 'depth', 'depth uncertainty' attributes. The 'ori cls' and 'ori offset' share the Conv2d layer. Note that the 'inplace-abn' module in MonoFlex [5] is changed back to a BatchNorm2d layer and a ReLU activation function. The architecture of OTCL module includes RoIAlign, conv layer, attention layer, and correlation layer. We adopt torchvision's RoIAlign implementation[1]. The output size of the RoI feature is 60×80. The conv layer includes two '3×3 Conv2d + BatchNorm2d + ReLU' modules. The attention layer contains 4 self-attention modules and 4 cross-attention modules, using standard MultiheadAttention implementation. The channel dimension is 64, and the number of attention heads is 4. The correlation layer contains the operation of the feature inner product and the softmax normalization.

## B. Training and Inference Details

### B.1. Training Settings

In Table A and Table B, we show the training details of BA-Det and the variant BA-Det$_{\text{FCOS3D}}$.

### B.2. Inference Details

In the main paper, we introduce the inference process in Sec. 4.4. Here, we make some additional explanations of inference details.

[1] https://github.com/pytorch/vision/blob/main/torchvision/ops/roi_align.py

| config | value |
|---|---|
| optimizer | Adam |
| optimizer momentum | $\beta_1 = 0.9, \beta_2 = 0.999$ |
| weight decay | $1 \times 10^{-5}$ |
| learning rate | $5 \times 10^{-4}$ |
| learning rate schedule | cosine decay |
| warmup iterations | 500 |
| epochs | 14 |
| augmentation | random flip |
| aug prob | 0.5 |
| batch size | 8 |
| gradient clip | 10 |
| image size | 1920 ×1280 |
| loss name | heatmap loss, bbox loss, depth loss<br>offset loss, orientation loss, dims loss<br>corner loss, keypoint loss, kp depth loss<br>trunc offset loss, featuremetric OBA loss |
| loss weight | 1.0 , 1.0 , 1.0<br>0.6 , 1.0 , 0.33<br>0.025 , 0.01 ,0.066<br>0.6 ,1.0 |
| sync BN | True |

Table A. Training config of BA-Det.

**Dense feature matching.** We match the dense RoI feature in two frames. The sliding window $\tau$ is 5 in the implementation. We adopt OPENCV's implementation of the RANSAC operation. To balance the number of valid features in each frame, we select the top $k$ feature correspondence for each frame based on the similarities. $k$ is set to 50 in the implementation. Note that if the number of correspondences is imbalanced between frames, the optimization tends to give high weight excessively to the frames with more tracklets and make other frames deviate from the correct pose.

**Object-centric bundle adjustment.** We use DeepLM to solve the non-linear least-square optimization problem. Note that in the inference stage, we adopt the original OBA formulation with reprojection error, according to the corre-

(a) segment-15948509588157321530_7187_290_7207_290

(b) segment-16229547658178627464_380_000_400_000

Figure A. Qualitative results in the WOD *val* set as videos. We use blue and red boxes to denote initial predictions and optimized predictions of the object we highlight. The green and **black** boxes denote the other box predictions and the ground truth boxes. The ego vehicle lies at the bottom of each figure. Please view with Adobe Acrobat Reader to see the videos.


Figure B. Video with detection and tracking results. Open with Adobe Acrobat Reader to play the video. Best viewed by zooming in.

spondence prediction from the OTCL module. The initial 3D position of the keypoint is set to (0,0,0) in the object reference frame. In DeepLM, no robust loss (e.g., Huber loss) is used.

| config | value |
|---|---|
| optimizer | AdamW |
| optimizer momentum | $\beta_1 = 0.9, \beta_2 = 0.999$ |
| weight decay | $1 \times 10^{-4}$ |
| learning rate | $2.5 \times 10^{-4}$ |
| learning rate schedule | step decay |
| lr decay epoch | [16, 22] |
| warmup iterations | 500 |
| epochs | 24 |
| data interval | 5 |
| augmentation | random flip |
| aug prob | 0.5 |
| batch size | 8 |
| gradient clip | 35 |
| image size | $1248 \times 832$ |
| loss name | cls loss, bbox loss,dir loss, OBA loss |
| loss weight | 1.0,2.0,0.2,0.8 |
| gt assigner | MaxIoUAssigner |
| sync BN | True |
| pre-train | FCOS3D++_r101_DCN_2x_D3 |

Table B. Training config of BA-Det$_{\text{FCOS3D}}$.

| | MOTA | MOTP | Miss | Mismatch | FP |
|---|---|---|---|---|---|
| 1st + ImmortalTracker | 0.190 | 0.282 | 0.724 | 9.84e-04 | 0.085 |
| BA-Det (Ours) | 0.238 | 0.282 | 0.652 | 5.44e-04 | 0.110 |

Table C. Tracking results evaluated on WOD val set. We report the metrics on LEVEL 2 @IoU0.5.

**Post-processing.** For the tracklet rescoring process, we adopt the maximum predicted score of the tracklet to replace the first-stage predicted score in each frame. The bounding box interpolation is for the missing detection in the tracklet. This process is to interpolate the 3D location of the object center, 3D box size, and orientation in the global reference frame. Only the nearest two observations are considered in interpolation.

## C. Qualitative Results

We show some additional qualitative results in the appendix. In Fig. A, the qualitative results of two sequences (segment-1594850958815732153 0_7187_290_7207_290 and segment-16229547658178627464_380_000_400_000) in the WOD [1] *val* set are shown as videos. For better understanding, we show a more detailed video (Fig. B) with detection and tracking results in BEV. Please view with Adobe Acrobat to see the videos.

## D. Additional Experiments

We evaluate the tracking results compared with the baseline of first-stage predictions + ImmortalTracker in Table C. Besides, to show how tracking quality affects detection, we use a weaker tracker, AB3DMOT [3]. The results are shown in Table D. The results reveal that our BA-Det improves tracking performance and is robust to tracking quality.

| Tracking methods | 3D AP$_{70}$ | 3D AP$_{50}$ |
|---|---|---|
| AB3DMOT [3]* | 16.3 | 38.9 |
| ImmortalTracker [2] | 16.6 | 40.9 |

Table D. Influence of tracking methods. *: implemented by us.

## E. Reproducibility Statement

We will release the training and inference codes to help reproducing our work. Limited by the license of WOD, the checkpoint of the model cannot be publicly available. However, we will provide the checkpoint by email. The implementation details, including the training and inference settings, network architecture, and the BA-Det$_{\text{FCOS3D}}$ variant, are mentioned in Sec. 4 and 5.2 in the main paper and Sec. A and B in the appendix. WOD is a publicly available dataset.

## References


[1] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 3

[2] Qitai Wang, Yuntao Chen, Ziqi Pang, Naiyan Wang, and Zhaoxiang Zhang. Immortal tracker: Tracklet never dies. *arXiv preprint arXiv:2111.13672*, 2021. 3

[3] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. Ab3dmot: A baseline for 3d multi-object tracking and new evaluation metrics. In *ECCV Workshops*, 2020. 3

[4] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *CVPR*, 2018. 1

[5] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3d object detection. In *CVPR*, 2021. 1