Supplementary Material for Synchronize Feature Extracting and Matching: A Single Branch Framework for 3D Object Tracking

Teli Ma[1,2]*, Mengmeng Wang[2]*, Jimin Xiao[3], Huifeng Wu[4], Yong Liu[2]†
[1]The Hong Kong University of Science and Technology, Guangzhou [2]Zhejiang University
[3]Xi'an Jiaotong-Liverpool University [4]Hangzhou Dianzi University
tma184@connect.hkust-gz.edu.cn mengmengwang@zju.edu.cn
jimin.xiao@xjtlu.edu.cn whf@hdu.edu.cn yongliu@iipc.zju.edu.cn

1. Decoder Network and Losses

With the encoded point-wise features of various scales, a multi-scale feature fusion module is adopted to fuse the search region features only and output the features with the same origin input size, feeding into the decoder part for final predictions. Following the V2B [1], the features are voxelized as a volumetric representation and 3D convolutions are utilized on the encoded features. To ensure the features with high response to the target can be distinguished from all features, max-pooling operation along the z-axis is adopted to acquire the BEV feature maps for regression. Afterward, layers of 2D convolution blocks (2D convolution, batch normalization and ReLU activation) are leveraged to aggregate the features from dense BEV feature maps, thus the local representations can be captured for the potential target. The decoding process is anchor-free and enjoys the accurate localization due to the perspective of BEV.

Focal loss [2] and L1 loss are leveraged for classification and regression, respectively. Following the V2B [1], the 2D target center (c_x, c_y) can be parameterized as $(\frac{x-x_{min}}{v_x}, \frac{y-y_{min}}{v_y})$, where x_{min} and y_{min} are the lower limit of x and y dimension in search area, and v_x, v_y are the voxel size in $x-y$ plane. The discrete 2D center is defined by $\hat{c_x} = \lfloor c_x \rfloor$ and $\hat{c_y} = \lfloor c_y \rfloor$. For the pixel (i, j) in the 2D bounding box, if (i, j) is the center of target, then the ground truth classification $\mathcal{G}_{cls}$ is 1, otherwise $\frac{1}{\gamma+1}$, where γ is the Euclidean distance between (i, j) and the target center. $\mathcal{G}_{cls}$ equals 0 if a pixel is outside of the bounding box. Based on that, Focal loss is adopted for the classification. For the offset head, the ground truth is $\mathcal{G}_{reg} \in \mathbb{R}^{3\times r\times r}$, and the r is the radius of the object center. The regression target is $[c_x - \hat{c_x}, c_y - \hat{c_y}, \theta]$, where θ is the rotation angle. Also, the ground truth of z-axis $\mathcal{G}_z$ is also considered. Therefore, L1 loss is utilized for both the offset regression and z-axis regression. The coefficient is $1, 1, 2$ for the Focal loss, offset L1 loss and z-axis L1 loss, respectively.

References

[1] Le Hui, Lingpeng Wang, Mingmei Cheng, Jin Xie, and Jian Yang. 3d siamese voxel-to-bev tracker for sparse point clouds. *Advances in Neural Information Processing Systems*, 34, 2021. 1

[2] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1

*Equal Contribution
†Corresponding Author