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 \(\left(\frac{x_{\text{min}} - x}{v_x}, \frac{y_{\text{min}} - y}{v_y}\right)\), where \(x_{\text{min}}\) and \(y_{\text{min}}\) are the lower limit of \(x\) and \(y\) dimension in search area, and \(v_x\) and \(v_y\) 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 \(G_{\text{cls}}\) is 1, otherwise \(\gamma + 1\), where \(\gamma\) is the Euclidean distance between \((i, j)\) and the target center. \(G_{\text{cls}}\) 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 \(G_{\text{reg}} \in \mathbb{R}^{3 \times r \times r}\), 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 \(\theta\) is the rotation angle. Also, the ground truth of \(z\)-axis \(G_z\) 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
