CenterFormer: Center-based Transformer for 3D Object Detection [3]

Minji Kim (minji@snu.ac.kr; 2020-28702), Dept. of Electrical and Computer Engineering, Seoul National University

1. Introduction

Inspired by DEtection Transformer (DETR) [1], this paper investigates a novel center-based Transformer for 3D object detection named CenterFormer. In DETR-style encoder-decoder transformer network, the computational complexity grows quadratically when input size increases. Thus, directly using the Transformer encoder on voxel or Bird Eye View (BEV) feature map is impractical due to its large size. For the efficiency and effectiveness in learning the DETR-style networks, this paper proposes to use the center feature as the initial query embedding to capture the object-centric information. Furthermore, the cross-attention learning based on a small multi-scale window is also helpful for reducing the computational complexity. Results on Waymo Open Dataset shows that CenterFormer reaches state-of-the-art performance with multi-frame extension.

2. CenterFormer

Fig. 1 shows the overall architecture of CenterFormer. The network consists of four parts: a voxel feature encoder that encodes the raw point cloud into a BEV feature representation, a multi-scale center proposal network (CPN), the center-based transformer decoder, and a regression head that predicts the bounding box. First, the point cloud is encoded into a BEV function representation using a standard voxel-based backbone network. Next, the feature map is transformed to different scales, followed by a multi-scale center proposal network to predict the initial center position. The proposed center feature is utilized as the input for the transformer encoder as query embedding. Like Deformable DETR [4], a deformable cross attention layer is adopted in each Transformer module for the effective aggregation of multi-scale feature maps. Finally, the output object representation is regressed to other object properties to produce the final object prediction.

3. Experiments

Results Experimental results are summarized in Fig. 2, surpassing other methods in terms of Mean L2, although using multi-frame settings in comparison with other methods with single frame setting seems unfair for me.

Difference with Deformable DETR This paper is largely inspired by Deformable DETR [4]. However, the difference between CenterFormer and Deformable DETR can be summarized as follows:

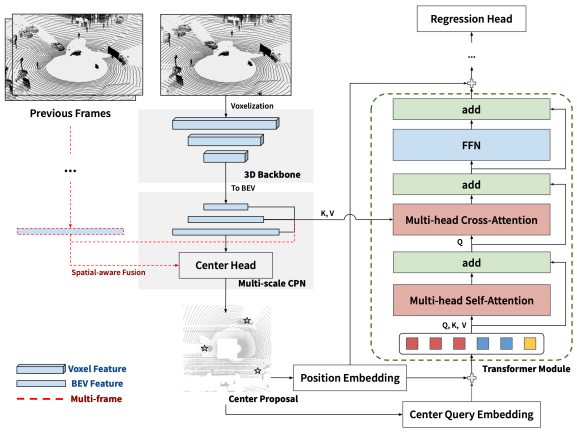

Figure 1. The overall pipeline of CenterFormer.

Method	Frame	Vehicle L1 (mAP/APH)	Vehicle L2 (mAP/APH)	Pedestrain L1 (mAP/APH)	Pedestrain L2 (mAP/APH)	Cyclist L1 (mAP/APH)	Cyclist L2 (mAP/APH)	Mean L2 mAPH
StarNet [28]	1	55.1/54.6	48.7/48.3	68.3/60.9	59.3/52.8	- / -	- / -	-
SECOND† [49]	1	72.3/71.7	63.9/63.3	68.7/58.2	60.7/51.3	60.6/59.3	58.3/57.1	57.2
LiDAR R-CNN [20]	1	73.5/73.0	64.7/64.2	71.2/58.7	63.1/51.7	68.6/66.9	66.1/64.4	60.1
Part-A2† [37]	1	77.1/76.5	68.5/68.0	75.2/66.9	66.2/58.6	68.6/67.4	66.1/64.9	63.8
3D-MAN [50]	16	74.5/74.0	67.6/67.1	71.7/67.7	62.6/59.0	- / -	- / -	-
PV-RCNN++ [35]	1	79.1/78.6	70.3/69.9	80.6/74.6	71.9/66.3	73.5/72.4	70.7/69.6	68.6
CenterPoint [53]	1	- / -	- /67.9	- / -	- /65.6	- / -	- /68.6	67.4
CenterPoint [53]	2	- / -	- /69.7	- / -	- /70.3	- / -	- /70.9	70.3
CenterFormer	1	75.0/74.4	69.9/69.4	78.0/72.4	73.1/67.7	73.8/72.7	71.3/70.2	69.1
CenterFormer†	1	75.2/74.7	70.2/69.7	78.6/73.0	73.6/68.3	72.3/71.3	69.8/68.8	69.0
CenterFormer	2	77.1/76.6	72.2/71.7	80.9/77.6	76.2/73.0	76.0/75.1	73.6/72.7	72.5
CenterFormer†	2	77.0/76.5	72.1/71.6	81.4/78.0	76.7/73.4	**76.6/75.7**	**74.2/73.3**	72.8
CenterFormer†	4	78.1/77.6	73.4/72.9	81.7/78.6	77.2/74.2	75.6/74.8	73.4/72.6	73.2
CenterFormer†	8	**78.8/78.3**	**74.3/73.8**	**82.1/79.3**	**77.8/75.0**	75.2/74.4	73.2/72.3	**73.7**

Figure 2. The overall pipeline of CenterFormer.

- CenterFormer only adopts decoder part of Transformers.
- Center feature is used as the query embedding instead of the learnable parameters.
- The set matching strategy is not adopted since it is hard to converge [2]

These features help the CenterFormer to train and test effectively and robustly in the task of 3D object detection.

References

[1] N. Carion, et al. End-to-end object detection with transformers. In *ECCV*, 2020. 1

[2] T. Yin, et al. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021. 1

[3] Z. Zhou, et al. Centerformer: Center-based transformer for 3d object detection. In *ECCV*, 2022. 1

[4] X. Zhu, et al. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1