PolarFormer: Multi-Camera 3D Object Detection with Polar Transformer

Yanqin Jiang$^{1,4,*}$, Li Zhang$^{2,1}$, Zhenwei Miao$^5$, Xiatian Zhu$^6$, Jin Gao$^{1,4}$, Weimin Hu$^{1,4,7}$, Yu-Gang Jiang$^3$

$^1$NLPR, Institute of Automation, Chinese Academy of Sciences  
$^2$School of Data Science, Fudan University  
$^3$School of Computer Science, Fudan University  
$^4$School of Artificial Intelligence, University of Chinese Academy of Sciences  
$^5$Alibaba DAMO Academy  
$^6$Surrey Institute for People-Centred Artificial Intelligence, CVSSP, University of Surrey  
$^7$School of Information Science and Technology, ShanghaiTech University

Abstract

3D object detection in autonomous driving aims to reason "what" and "where" the objects of interest present in a 3D world. Following the conventional wisdom of previous 2D object detection, existing methods often adopt the canonical Cartesian coordinate system with perpendicular axis. However, we conjecture that this does not fit the nature of the ego car’s perspective, as each onboard camera perceives the world in shape of wedge intrinsic to the imaging geometry with radical (non-perpendicular) axis. Hence, in this paper we advocate the exploitation of the Polar coordinate system and propose a new Polar Transformer (PolarFormer) for more accurate 3D object detection in the bird’s-eye-view (BEV) taking as input only multi-camera 2D images. Specifically, we design a cross-attention based Polar detection head without restriction to the shape of input structure to deal with irregular Polar grids. For tackling the unconstrained object scale variations along Polar’s distance dimension, we further introduce a multi-scale Polar representation learning strategy. As a result, our model can make best use of the Polar representation rasterized via attending to the corresponding image observation in a sequence-to-sequence fashion subject to the geometric constraints. Thorough experiments on the nuScenes dataset demonstrate that our PolarFormer outperforms significantly state-of-the-art 3D object detection alternatives.

Introduction

3D object detection is an enabling capability of autonomous driving in unconstrained real-world scenes (Wang et al. 2022b, 2021). It aims to predict the location, dimension and orientation of individual objects of interest in a 3D world. Despite favourable cost advantages, multi-camera based 3D object detection (Wang et al. 2021, 2022a,b; Zhou, Wang, and Krähenbühl 2019) remains particularly challenging. To obtain 3D representation, dense depth estimation is often leveraged (Philion and Fidler 2020), which however is not only expensive in computation but error prone. To bypass depth inference, more recent methods (Wang et al. 2022b; Li et al. 2022b) exploit query-based 2D detection (Carion et al. 2020) to learn a set of sparse and virtual embedding space. This is largely restricted by convolution based models used. In contrast, the physical world perceived under each camera in the ego car’s perspective is in shape of wedge intrinsic to the camera imaging geometry with radical non-perpendicular axis (Figure 5). Bearing this imaging property in mind, we conjecture that the Polar coordinate system should be more appropriate and natural than the often adopted Cartesian counterpart for 3D object detection. Indeed, the Polar coordinate has been exploited in a few LiDAR-based 3D perception methods (Zhang et al. 2020; Bewley et al. 2020; Rapoport-Lavie and Raviv 2021; Zhu et al. 2021). However, they are limited algorithmically due to the adoption of convolutional networks restricted to rectangular grid structure and local receptive fields.

Motivated by the aforementioned insights, in this work a novel Polar Transformer (PolarFormer) model for multi-camera 3D object detection in a Polar coordinate system is introduced (Figure 3). Specifically, we first learn the representation of polar rays corresponding to image regions in a sequence-to-sequence cross-attention formulation. Then we rasterize a BEV Polar representation consisting of a set of Polar rays evenly distributed around 360 degrees. To deal with irregular Polar grids as suffered by conventional LiDAR based solutions (Zhang et al. 2020; Bewley et al. 2020; Rapoport-Lavie and Raviv 2021; Zhu et al. 2021), we propose a cross-attention based decoder head design without restriction to the shape of input structure. For tackling the challenge of unconstrained object scale variation along Polar’s
distance dimension, we resort to a multi-scale Polar BEV representation learning strategy.

The contributions of this work are summarized as follows: (I) We propose a new Polar Transformer (PolarFormer) model for multi-camera 3D object detection in the Polar coordinate system. (II) This is achieved based on two Polar-tailored designs: A cross-attention based decoder design for dealing with the irregular Polar grids, and a multi-scale Polar representation learning strategy for handling the unconstrained object scale variations over Polar’s distance dimension. (III) Extensive experiments on the nuScenes dataset show that our PolarFormer achieves leading performance for camera-based 3D object detection (Figure 1).

Related Work

Monocular/multi-camera 3D object detection Image-based 3D object detection aims to estimate the object location, dimension and orientation in the 3D space alongside its category given only image input. To solve this ill-posed problem, (Zhou, Wang, and Krähenbühl 2019; Wang et al. 2021) naively build their detection pipelines by augmenting 2D detectors (Zhou, Wang, and Krähenbühl 2019; Tian et al. 2019) in a data driven fashion. PGD (Wang et al. 2022a) further captures the uncertainty and models the relationship of different objects by utilizing geometric prior. Contrast to the above image-only methods, depth-based methods (Xu and Chen 2018; Ding et al. 2020; Wang et al. 2019; You et al. 2019; Ma et al. 2019; Reading et al. 2021) use depth cues as 3D information to mitigate the naturally ill-posed problem. Recently, multi-camera-based 3D object detection emerges. DETR3D (Wang et al. 2022b) considers detecting objects across all cameras collectively. It learns a set of sparse and virtual query embedding, without explicitly building the geometry structure among objects/queries. (Li et al. 2022b) considers detecting objects in BEV, performing end-to-end object detection via object queries. Note that multi-camera setting uses the same amount of training data as the monocular pipelines. Both multi-camera and monocular paradigms share the same evaluation metrics.

Bird’s-eye-view (BEV) representation Recently there is a surge of interest in transforming the monocular or multi-view images from ego car cameras into the bird’s-eye-view coordinate (Roddick, Kendall, and Cipolla 2019; Philion and Fidler 2020; Li et al. 2022a; Roddick and Cipolla 2020; Reading et al. 2021; Saha et al. 2022) followed by specific optimization tasks (e.g., 3D object detection, semantic segmentation). A natural solution (Philion and Fidler 2020; Hu et al. 2021) is to learn the BEV representation by leveraging the pixel-level dense depth estimation. This however is error-prone due to lacking ground-truth supervision. Another line of research aims to bypass the depth prediction and directly leverage a Transformer (Chitta, Prakash, and Geiger 2021; Can et al. 2021; Saha et al. 2022) or a FC layer (Li et al. 2022a; Roddick and Cipolla 2020; Yang et al. 2021) to learn the transformation from camera inputs to the BEV coordinate. A similar attempt as ours is conducted in (Saha et al. 2022) but limited in a couple of aspects: (i) It is restricted to monocular input for a straightforward 2D segmentation task while we consider multiple cameras collectively for more challenging 3D object detection; (ii) We uniquely provide a solid multi-scale Polar BEV transformation to tackle the unconstrained object scale variations and followed by a jointly optimized cross-attention based Polar head.

3D object detection in Polar coordinate 3D object detection in the Polar or Polar-like coordinate system has been attempted in LiDAR-based perception methods. For example, CyliNet (Zhu et al. 2021) introduces range-based guidance for extracting Polar-consistent features. In particular, it adapts a Cartesian heatmap to a Polar version for object classification, whilst learning relative heading angles and velocities. However, CyliNet still lags clearly behind the Cartesian counterpart. Recently, PolarSteam (Chen, Vora, and Be-
PolarFormer consists of an overall architecture of PolarFormer is depicted in Figure 2. With each other in a shared Polar BEV encoder. Multi-scale features with multi-scale feature interaction. Finally, a Polar alignment module then aggregates Polar rays from multiple camera views to generate a structured Polar map. Further, a BEV encoder enhances the Polar features with multi-scale feature interaction. Finally, a \textit{Polar detection head} decodes the Polar map and predicts the objects in the Polar coordinate system. For tackling the unconstrained object scale variation with multi-granularity of details, we consider a multi-scale BEV Polar representation structure. As shown in Figure 4, image features with different scales have unique cross-plane encoders and interact with each other in a shared Polar BEV encoder. Multi-scale Polar BEV maps are then queried by Polar decoder head. An overall architecture of PolarFormer is depicted in Figure 2.

Cross-Plane Encoder

The goal of cross-plane encoder is to associate an image with BEV Polar rays. According to the geometric model of camera, for any camera coordinate \((x(C), y(C), z(C)) \in \mathbb{R}^3\), the transformation to image coordinate \((x(I), y(I)) \in \mathbb{R}^2\) could be described as:

\[
s \begin{bmatrix} x(I) \\ y(I) \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x(C) \\ y(C) \\ z(C) \end{bmatrix},
\]

where \(f_x, f_y, u_0 \) and \(v_0 \) are camera intrinsic parameters in \(\Pi\), \(x(C), y(C)\) and \(z(C)\) are the horizontal, vertical, depth coordinate respectively. \(s\) is the scale factor. For any BEV Polar coordinate \((\rho(P), \phi(P))\), we have:

\[
\rho(P) = \sqrt{(x(C))^2 + (z(C))^2} = z(C) \sqrt{(x(I) - u_0)^2 + 1},
\]

\[
\phi(P) = \frac{x(C)}{z(C)} = \arctan \frac{x(I) - u_0}{f_x},
\]

Eq. (2) suggests that the azimuth \(\phi(P)\) is irrelevant to the vertical value of image coordinate. It is hence natural to build a one-to-one relationship between Polar rays and image columns (Saha et al. 2022). However, we need object depth \(z(C)\) to compute radius \(\rho(P)\), the estimation of which is ill-posed. Instead of explicit depth estimation, we leverage the attention mechanism (Vaswani et al. 2017) to model the relationship between pixels along the image column and positions along the Polar ray.

Let \(f_{n,u,w} \in \mathbb{R}^{H \times W \times C}\) represent the image column from \(n\)th camera, with scale and \(w\)th column, and \(\hat{p}_{n,u,w} \in \mathbb{R}^{N_t \times C}\) denote the corresponding Polar ray query we introduce, where \(H\) and \(R\) are the image feature map's height and Polar map's range. We formulate cross-plane attention as:

\[
p_{n,u,w} = \text{MultiHead}(\hat{p}_{n,u,w}, f_{n,u,w}, f_{n,u,w}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h) W^Q, \tag{4}
\]

where

\[
\text{head}_i = \text{Attention}(\hat{p}_{n,u,w} W^Q_{i,u}, f_{n,u,w} W^K_{i,u}, f_{n,u,w} W^V_{i,u}), \tag{5}
\]

and

\[
f_{n,u,w} = \text{Stack}(p_{n,u,1}, \ldots, p_{n,u,W_u}), \quad \text{dim} = 1 \in \mathbb{R}^{H \times W_u \times C}, \tag{6}
\]

Polar Alignment across Multiple Cameras

Our Polar alignment module transforms Polar rays from different camera coordinates to a shared world coordinate. Taking multi-view Polar feature maps \(\{\mathbf{p}_{n,u}\}^N_{n=1}\) and camera matrix \(\{\Pi_n, E_n\}^N_{n=1}\) as inputs, it produces a coherent BEV Polar map \(\mathbf{G}_u \in \mathbb{R}^{N_t \times N_u \times C}\) covering all camera views, where \(R_u, N_u\) and \(C\) are the dimensions of radius, azimuth and feature. Concretely, it first generates a set of 3D points in the cylindrical coordinate uniformly, denoted by \(G^{(P)} = \{(\rho_i^{(P)}, \phi_j^{(P)}, z_k^{(P)}) | i = 1, \ldots, R_u; j = 1, \ldots, N_u; k = 1, \ldots, Z_u\}\), where \(Z_u\) is the number of points along \(z\) axis. Since cylindrical coordinate and Polar coordinate share radius and azimuth axis, their superscripts are both denoted with \(P\). The points are then projected to the image plane of
Figure 2: Schematic illustration of our proposed *PolarFormer* for multi-camera 3D object detection. For each image captured by any camera view, our model first extracts the feature maps at multiple spatial scales. Given such a feature map, the cross-plane encoder (a) then transforms all the feature columns to a set of Polar rays in a sequence-to-sequence manner via polar queries based cross-attention. The polar rays from all the cameras are subsequently processed by a Polar alignment module (b) to generate a structured multi-scale Polar BEV map, followed by further enhancement via interactions among different scales using a Polar BEV encoder (c). At last, a specially designed Polar Head decodes multi-scale Polar BEV features for making final predictions in the Polar coordinate.

In the *PolarReference Points* layer, we introduce a multi-scale deformable attention module (Zhu et al. 2020) as:

$$\text{MSDeformAttn}(z_q, x_q, \{G_u\}_{u=1}^U) = \\
\sum_{m=1}^M \sum_{u=1}^U \sum_{k=1}^K A_{muqk} W'_m G_u(z_{uq} + \Delta x_{muqk}) |,$$

where $m$ and $k$ are the index of the attention head and the sampling point. $z_q$ is the query feature. $\Delta x_{muqk}$ and $A_{muqk}$ denote the sampling offset and the attention weight.
of the kth sampling point in ith feature level and mth attention head. The attention weight \( A_{muqk} \) is normalized by \( \sum_{i=1}^{U} \sum_{k=1}^{K} A_{muqk} = 1 \). Sampling offsets \( \Delta x_{muqk} \) are generated by applying MLP layers on query \( q \). Function \( \hat{v} \) generates the sampling offsets and rescales the normalized coordinate \( \hat{x}_i \) to the ith feature scale. \( W_m \) and \( W'_m \) are learnable parameters. Serving as query, each pixel in the multi-scale feature maps exploits the information from both neighbouring pixels and pixels across scales. This enables learning richer semantics across all feature scales.

**Polar BEV Decoder at Multiple Scales**

The Polar decoder decodes the above multi-scale Polar features to make predictions in the Polar coordinate. We construct the Polar BEV decoder with deformable attention (Zhu et al. 2020). Specifically, we query \( q \) in Eq. (9) as learnable parameters.

Unlike 2D reference points in the encoder, here the reference points are in 3D cylindrical coordinate, equal to Polar coordinate when projected to BEV. The classification branch in each decoder layer outputs the confidence score vector \( \mathbf{c} \in \mathbb{R}^{C} \), where \( C \) is the number of categories. The key learning targets of regression branch are in polar coordinate instead of Cartesian coordinate, as illustrated in Figure 3.

For simplicity, superscript \((P)\) is omitted. Reference points \((\rho, \phi, z)\) are iteratively refined in the decoder. With reference points, the regression branch regresses the offsets \( d_\rho, d_\phi \) and \( d_z \). The learning targets for orientation \( \theta \) and velocity \( v \) are relative to azimuths of objects and separated to orthogonal components \( \theta_\phi, \theta_\rho, v_\phi \) and \( v_\rho \), defined by:

\[
\bar{\theta}_\text{ori} = \theta_\text{ori} - \phi, \quad \theta_\phi = \sin \bar{\theta}_\text{ori}, \quad \theta_\rho = \cos \bar{\theta}_\text{ori}, \quad (10)
\]

and

\[
\bar{\theta}_v = \theta_v - \phi, \quad v_\phi = v_{abs} \sin \bar{\theta}_v, \quad v_\rho = v_{abs} \cos \bar{\theta}_v, \quad (11)
\]

Here, \( \theta_\text{ori} \) is the yaw angle of the bounding box. \( v_{abs} \) and \( \theta_v \) are the absolute value and angle of velocity. We regress the object size \( l, w \) and \( h \) as \( \log l, \log w \) and \( \log h \). We adopt Focal loss (Lin et al. 2017) and L1 loss for classification and regression respectively.

**Experiments**

**Dataset** We evaluate the PolarFormer on the nuScenes dataset (Caesar et al. 2020). It provides images with a resolution of \( 1600 \times 900 \) from 6 surrounding cameras (Figure 1). The total of 1000 scenes, where each sequence is roughly 20 seconds long and annotated every 0.5 second, is split

<table>
<thead>
<tr>
<th>Methods</th>
<th>Backbone</th>
<th>mAP†</th>
<th>NDS†</th>
<th>mATE↓</th>
<th>mASE↓</th>
<th>mAOE↓</th>
<th>mAVE↓</th>
<th>mAAME↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCOS3D† (Wang et al. 2021)</td>
<td>R101</td>
<td>35.8</td>
<td>42.8</td>
<td>69.0</td>
<td>24.9</td>
<td>45.2</td>
<td>143.4</td>
<td>12.4</td>
</tr>
<tr>
<td>PGD† (Wang et al. 2022a)</td>
<td>R101</td>
<td>38.6</td>
<td>44.8</td>
<td>62.6</td>
<td>24.5</td>
<td>45.1</td>
<td>150.9</td>
<td>12.7</td>
</tr>
<tr>
<td>Ego3RT† (Lu et al. 2022)</td>
<td>R101</td>
<td>39.8</td>
<td>44.3</td>
<td>59.9</td>
<td>26.8</td>
<td>47.0</td>
<td>116.9</td>
<td>17.2</td>
</tr>
<tr>
<td>BEVFormer-S† (Li et al. 2022b)</td>
<td>R101</td>
<td>40.9</td>
<td>46.2</td>
<td>65.0</td>
<td>26.1</td>
<td>43.9</td>
<td>92.5</td>
<td>14.7</td>
</tr>
<tr>
<td>PolarFormer†</td>
<td>R101</td>
<td>41.5</td>
<td>47.0</td>
<td>65.7</td>
<td>26.3</td>
<td>40.5</td>
<td>91.1</td>
<td>13.9</td>
</tr>
<tr>
<td>BEVFormer (Li et al. 2022b)</td>
<td>V2-99</td>
<td>44.5</td>
<td>53.5</td>
<td>63.1</td>
<td>25.7</td>
<td>40.5</td>
<td>43.5</td>
<td>14.3</td>
</tr>
<tr>
<td>PolarFormer†</td>
<td>R101</td>
<td>45.7</td>
<td>54.3</td>
<td>61.2</td>
<td>25.7</td>
<td>39.2</td>
<td>46.7</td>
<td>12.9</td>
</tr>
<tr>
<td>DETR3D† (Wang et al. 2022b)</td>
<td>V2-99</td>
<td>42.3</td>
<td>47.9</td>
<td>64.1</td>
<td>25.5</td>
<td>39.4</td>
<td>84.5</td>
<td>13.3</td>
</tr>
<tr>
<td>M2BEV*‡ (Xie et al. 2022)</td>
<td>X101</td>
<td>42.9</td>
<td>47.4</td>
<td>58.3</td>
<td>25.4</td>
<td>37.6</td>
<td>10.53</td>
<td>19.0</td>
</tr>
<tr>
<td>Ego3RT† (Lu et al. 2022)</td>
<td>V2-99</td>
<td>42.5</td>
<td>47.9</td>
<td>54.9</td>
<td>26.4</td>
<td>43.3</td>
<td>101.4</td>
<td>14.5</td>
</tr>
<tr>
<td>BEVFormer-S† (Li et al. 2022b)</td>
<td>V2-99</td>
<td>43.5</td>
<td>49.5</td>
<td>58.9</td>
<td>25.4</td>
<td>40.2</td>
<td>84.2</td>
<td>13.1</td>
</tr>
<tr>
<td>PolarFormer†</td>
<td>V2-99</td>
<td>45.5</td>
<td>50.3</td>
<td>59.2</td>
<td>25.8</td>
<td>38.9</td>
<td>87.0</td>
<td>13.2</td>
</tr>
<tr>
<td>BEVFormer (Li et al. 2022b)</td>
<td>V2-99</td>
<td>48.1</td>
<td>56.9</td>
<td>58.2</td>
<td>25.6</td>
<td>37.5</td>
<td>37.8</td>
<td>12.6</td>
</tr>
<tr>
<td>PolarFormer†</td>
<td>V2-99</td>
<td>49.3</td>
<td>57.2</td>
<td>55.6</td>
<td>25.6</td>
<td>36.4</td>
<td>44.0</td>
<td>12.7</td>
</tr>
</tbody>
</table>

Table 1: State-of-the-art comparison on nuScenes test set. † denotes the prototype setting: The model is initialized from a PCOS3D (Wang et al. 2021) checkpoint trained on the nuScenes 3D detection dataset. ‡ denotes the improved setting: A pretrained model from DD3D (Park et al. 2021) is used, which includes external data from DDAD (Guizilini et al. 2020). * denotes backbone is pretrained on COCO (Lin et al. 2014) and nuImage (Caesar et al. 2020).
Table 2: 3D object detection results in different coordinate systems and ablations for the model architecture. PC denotes feature in Polar and prediction in Cartesian.

(a) Ablation study on coordinate system and detection head.

<table>
<thead>
<tr>
<th>Method</th>
<th>Feature</th>
<th>Prediction</th>
<th>mAP↑</th>
<th>NDS↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Centerpoint (Yin, Zhou, and Krähenbühl 2021)</td>
<td>Cartesian</td>
<td>Cartesian</td>
<td>37.8</td>
<td>45.4</td>
</tr>
<tr>
<td>Centerpoint* (Yin, Zhou, and Krähenbühl 2021)</td>
<td>Cartesian</td>
<td>Cartesian</td>
<td>38.5</td>
<td>45.6</td>
</tr>
<tr>
<td>PolarFormer-CC</td>
<td>Polar</td>
<td>Cartesian</td>
<td>38.1</td>
<td>45.5</td>
</tr>
<tr>
<td>PolarFormer-PC</td>
<td>Polar</td>
<td>Cartesian</td>
<td>38.5</td>
<td>45.0</td>
</tr>
<tr>
<td>PolarFormer</td>
<td>Polar</td>
<td>Polar</td>
<td>39.6</td>
<td>45.8</td>
</tr>
</tbody>
</table>

(b) Ablation on positional encoding (PE).

<table>
<thead>
<tr>
<th>Methods</th>
<th>Multi-scale</th>
<th>mAP↑</th>
<th>NDS↑</th>
<th>mAOE↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>PolarFormer-CC-s</td>
<td>X</td>
<td>38.1</td>
<td>44.9</td>
<td>40.0</td>
</tr>
<tr>
<td>PolarFormer-PC-s</td>
<td>X</td>
<td>38.8</td>
<td>45.0</td>
<td>37.6</td>
</tr>
<tr>
<td>PolarFormer-s</td>
<td>X</td>
<td>39.1</td>
<td>45.0</td>
<td>37.3</td>
</tr>
<tr>
<td>PolarFormer-CC</td>
<td>✓</td>
<td>38.1</td>
<td>45.0</td>
<td>37.5</td>
</tr>
<tr>
<td>PolarFormer-PC</td>
<td>✓</td>
<td>38.5</td>
<td>45.5</td>
<td>40.8</td>
</tr>
<tr>
<td>PolarFormer</td>
<td>✓</td>
<td>39.6</td>
<td>45.8</td>
<td>37.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Position encoding</th>
<th>mAP↑</th>
<th>NDS↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D learnable PE</td>
<td>38.8</td>
<td>45.1</td>
</tr>
<tr>
<td>3D PE</td>
<td>38.5</td>
<td>44.9</td>
</tr>
<tr>
<td>Fixed Sine PE</td>
<td>39.6</td>
<td>45.8</td>
</tr>
</tbody>
</table>

(c) Effectiveness of multi-scale polar representation.

(d) Ablation on polar resolution.

Table 2: 3D object detection results in different coordinate systems and ablations for the model architecture. PC denotes feature in Polar and prediction in Cartesian.

Implementation details We implement our approach based on the codebase mmdetection3d (Contributors 2020). Following DETR3D (Wang et al. 2022b) and FCOS3D (Wang et al. 2021), a ResNet-101 (He et al. 2016), with 3rd and 4th stages equipped with deformable convolutions is adopted as the backbone architecture. The number of cross-plane encoder layer is set to 3 for each feature scale. The resolution of radius and azimuth for our multiscale Polar BEV maps are (64, 256), (32, 128), (16, 64) respectively. We use 6 Polar BEV encoder and 6 decoder layers. Following DETR3D (Wang et al. 2022b), our backbone is initialized from a checkpoint of FCOS3D (Wang et al. 2021) trained on nuScenes 3D detection task, while the rest is initialized randomly. We use above setting for prototype verification. To fully leverage the sequence data, we further conduct temporal fusion between the current frame and one history sweep in the BEV space. Following BEVDet4D (Huang and Huang 2022), we simply concatenate two temporally adjacent multi-scale Polar BEV maps along the feature dimension and feed to the BEV Polar encoder. We randomly sample a history sweep from [3T; 27T] during training, and sample the frame at 15T for inference. T (≈ 0.083s) refers to the time interval between two sweep frames. We term our temporal version as PolarFormer-T.

Training Following DETR3D (Wang et al. 2022b) we train our models for 24 epochs with the AdamW optimizer and cosine annealing learning rate scheduler on 8 NVIDIA V100 GPUs. The initial learning rate is $2 \times 10^{-4}$, and the weight decay is set to 0.075. Total batch size is set to 48 across six cameras. Synchronized batch normalization is adopted. All experiments use the original input resolution. Note our image variant uses the same amount of training data as the monocular pipelines (Wang et al. 2021) and the multi-camera counterparts (Wang et al. 2022b; Li et al. 2022b). Multi-camera and monocular paradigms share the same evaluation metrics.

Inference We evaluate our model on nuScenes validation set and test server. We do not adopt model-agnostic trick such as model ensemble and test time augmentation.

Comparison with the State of the Art We compare our method with the state of the art on both test and val sets of nuScenes. In addition to the (i) prototype setting mentioned in implementation details, we also evaluate our model in the (ii) improved setting, with VoVNet (V2-99) (Lee et al. 2019) as backbone architecture with a pretrained checkpoint from DD3D (Park et al. 2021) (fine-tuned on extra DDAD15M (Guizilini et al. 2020) dataset) to boost performance.

Table 1 compares the results on nuScenes test set. We observe that our PolarFormer achieves the best performance under both the (i) prototype and (ii) improved setting in terms of mAP and NDS metrics, indicating the superiority of learning representation in the Polar coordinate. With temporal information PolarFormer-T can further boost performance substantially. Additional experiments results on val set and qualitative results are shown in supplementary materials.

Ablation Studies We conduct a series of ablation studies on nuScenes val set to validate the design of PolarFormer. Each proposed component and important hyperparameters are examined thoroughly.

Polar vs. Cartesian Table 2a ablates the coordinate system. We make several observations: (I) Learning the representation and making the prediction both on Cartesian, Centerpoint (Yin, Zhou, and Krähenbühl 2021) gives a strong baseline with 0.378 mAP and 0.454 NDS; After applying circle NMS, Centerpoint* can further improve; (II)
(a) Method Coordinate Near Medium Far
DETR3D (Wang et al. 2022b) Cartesian 49.9/49.3 28.6/39.2 8.1/21.2
BEVFormer-S (Li et al. 2022b) Cartesian 54.2/53.5 31.7/41.3 9.5/21.9
PolarFormer-CC Cartesian 54.7/55.5 32.3/42.2 9.4/22.2
PolarFormer Polar 57.8/55.8 33.6/42.9 9.6/22.3

(b) Figure 5: 3D object detection in (a) Cartesian BEV vs. (b) Polar BEV, and (c) Performance comparison (mAP/NDS) at three distances (Near/Medium/Far). Red and green boxes show the same objects in different coordinates.

(c) Our PolarFormer-CC (with Cartesian feature and prediction) outperforms the Centerpoint equipped with specially designed CBGS head (Yin, Zhou, and Krienhbuhl 2021); (III) When Polar BEV map is used to feed into a Cartesian decoder head, on-par performance is achieved with the post-processing counterpart; (IV) PolarFormer, our full model that predicts all 10 categories with one head and without any post-processing procedure, exceeds highly optimized Centerpoint by 1.1% in mAP and 0.2% in NDS. This suggests the significance of Polar in both representation learning and exploitation (i.e., decoding).

Visualizations With the quantitative evaluation in Figure 5, (I) it is evident that our model in Polar coordinate yields better results than Cartesian consistently. Specifically, Polar outperforms Cartesian by mAP 3.1% and NDS 0.3% in nearby area, mAP 1.3% and NDS 0.7% in medium area. (II) As shown in Figure 5a and Figure 5b, compared to the Polar map, Cartesian usually downsamples the nearby area (red) with information loss, while upsamples the distant area (green) without actual information added. This would explain the inferiority of Cartesian. (III) Figure 6 shows the attention map of the decoder query in multi-scale Polar BEV features. For better viewing, we resize the multi-scale features into the same resolution. The bottom/top corresponds to the largest/smallest scale features. The y axis represents the radius of the Polar map. It is shown that larger objects represented in the small map (top) are close to the ego car (small radius) whilst small objects in the large map (bottom) distribute through the distant area. This is highly consistent with the geometry structure of raw images (Figure 1), which has shown to be a more effective coordinate for 3D object detection as above.

Architecture (I) We first evaluate three designs of positional embedding (PE): 2D learnable PE, fixed Sine PE, and 3D PE (generated based on a set of 3D points for each Polar ray position). (II) As our cross-plane encoder transforms different levels of feature from FPN into Polar rays independently, we can fuse the multi-level features into a single BEV or naturally shape multiple BEVs with different or same resolutions; Table 2c clearly shows that a model with multi-scale Polar BEVs outperforms the single-scale counterpart under either coordinate. In contrast, little performance gain is achieved from multi-scale features in Cartesian. This suggests that object scale variation is a unique challenge with Polar, but absent with Cartesian. Our design consideration is thus verified. (III) We study the resolution of polar map by adjusting the azimuth $N_1$ and radius $R_1$ (the number of Polar query in the cross-plane encoder); Table 2d shows that the angle with 256 and radius with 64 gives the best performance.

Conclusions We have proposed the Polar Transformer (PolarFormer) for 3D object detection in multi-camera 2D images from the ego car’s perspective. With a rasterized BEV Polar representation geometrically aligned to visual observation, PolarFormer overcomes irregular Polar grids by a cross-attention based decoder. Further, a multi-scale representation learn-
ing strategy is designed for tackling the intrinsic object scale variation challenge. Extensive experiments on the nuScenes dataset validate the superiority of our PolarFormer over previous alternatives on 3D object detection.

Acknowledgments
This work was supported by the National Key R&D Program of China (Grant No. 2018AAA0102803, 2018AAA0102802, 2018AAA0102800), the Natural Science Foundation of China (Grant No. 6210020439, U22B2056, 61972394, 62036011, 62192782, 61721004, 62102417), Lingang Laboratory (Grant No. LG-QS-202202-07), Natural Science Foundation of Shanghai (Grant No. 22ZR1407500), Beijing Natural Science Foundation (Grant No. L223003, JQ22014), the Major Projects of Guangdong Education Department for Foundation Research and Applied Research (Grant No. 2017KZDXM081, 2018KZDXM066), Guangdong Provincial University Innovation Team Project (Grant No. 2020KCGXTD045). Jin Gao was also supported in part by the Youth Innovation Promotion Association, CAS.

References


