DetZero: Rethinking Offboard 3D Object Detection with Long-term Sequential Point Clouds

Tao Ma¹,²*, Xuemeng Yang², Hongbin Zhou², Xin Li³,²*, Botian Shi², Junjie Liu⁴*, Yuchen Yang⁵,²*, Zhizheng Liu⁶*, Liang He³, Yu Qiao², Yikang Li²,³, Botian Shi², Junjie Liu⁴*, Yuchen Yang⁵,²*, Zhizheng Liu⁶*, Liang He³

¹Multimedia Laboratory, The Chinese University of Hong Kong
²Shanghai Artificial Intelligence Laboratory
³East China Normal University
⁴South China University of Technology
⁵Fudan University
⁶ETH Zurich
⁷CPII

Corresponding author
https://github.com/PJLab-ADG/DetZero

Abstract

Existing offboard 3D detectors always follow a modular pipeline design to take advantage of unlimited sequential point clouds. We have found that the full potential of offboard 3D detectors is not explored mainly due to two reasons: (1) the onboard multi-object tracker cannot generate sufficient complete object trajectories, and (2) the motion state of objects poses an inevitable challenge for the object-centric refining stage in leveraging the long-term temporal context representation. To tackle these problems, we propose a novel paradigm of offboard 3D object detection, named DetZero. Concretely, an offline tracker coupled with a multi-frame detector is proposed to focus on the completeness of generated object tracks. An attention-mechanism refining module is proposed to strengthen contextual information interaction across long-term sequential point clouds for object refining with decomposed regression methods. Extensive experiments on Waymo Open Dataset show our DetZero outperforms all state-of-the-art onboard and offboard 3D detection methods. Notably, DetZero ranks 1st place on Waymo 3D object detection leaderboard¹ with 85.15 mAPH (L2) detection performance. Further experiments validate the application of taking the place of human labels with such high-quality results. Our empirical study leads to rethinking conventions and interesting findings that can guide future research on offboard 3D object detection.

¹Work performed during internship at Shanghai Artificial Intelligence Laboratory.


1. Introduction

Autonomous driving has rapidly advanced with promising progress in both industry and academia. A crucial component of this development is offboard 3D object detection, which can utilize entire sequence data from sensors (video or sequential point cloud) with few constraints on model capacity and inference speed. Therefore, some approaches [33, 50] are dedicated to developing high-quality “auto labels”, aiming to reduce manual labor in point cloud annotation.

Subsequently, many online detectors [54, 8, 56] are in-
introduced with the majority focusing on developing sophisticated modules to better utilize temporal context. As shown in Fig. 1, these newly proposed methods outperform both online [61, 20, 49, 37, 58, 36, 38, 22, 18, 15] and offboard 3D detectors [33, 50, 48] by a large margin, leaving the impression that current architecture and pipeline of offboard 3D detectors are too weak to learn the complex representation over long-term sequential point clouds. Therefore, in this paper, we revisit state-of-the-art (SOTA) offboard 3D detectors (see Sec. 3.1 for the pipeline) and identify two main factors hindering the full potential: (1) the onboard multi-object tracker can’t generate sufficient complete object trajectories, and (2) the motion state of objects poses an inevitable challenge for the object-centric refining stage to leverage the long-term temporal context representation.

Specifically, prevailing online 3D detectors achieve promising performance but easily generate severe fragment trajectories, ID switches, and false positives when coupled with a tracking-by-detection multi-object tracking algorithm. As shown in Fig. 2, this phenomenon may prevent the generation of complete temporal context features. Therefore, we adopt an upstream module comprising a multi-frame 3D detector and offline tracker that ensures the completeness and continuity of object tracking while maintaining high recall. Moreover, the sliding-window-based auto labeling model [33, 50] hinders the exploitation of the commonality of object features, as shown in Fig. 3. We notice that the size of the objects remained consistent over time. By capturing data from various viewpoints, we can enhance the point cloud of an object, allowing for more precise size estimation. Furthermore, the object trajectory is independent of its size, and should always follow the kinematic constraints in continuous time, which is manifested by the smoothness of the trajectory. These characteristics serve as the foundation for leveraging the long-term sequential point clouds in a decomposed paradigm: refine the geometry size, smooth the trajectory position, and update the confidence score.

By focusing on these main issues, we propose a new paradigm of offboard 3D object detection named DetZero. A tenet is underscored here: emphasizing high-recall detection and tracking during the upstream, meticulous high-accuracy refining with long-term temporal context during the downstream. Comprehensive empirical studies and evaluations on the Waymo Open Dataset (WOD) demonstrate that DetZero significantly improves perception by fully utilizing long-term sequential point clouds. Notably, we rank 1st place on WOD 3D object detection leaderboard with 85.15 mAPH (L2). Extensive ablation studies and generalization experiments show that our method performs well with different quality upstream inputs and stricter metrics. Semi-supervised experiments further demonstrate that our method can provide high-quality auto labels for onboarding 3D object detection models, which are already on par or even slightly higher than human labels.

The main contributions of our work are summarized as follows:

- We introduce DetZero, a new paradigm of offboard 3D object detection, to activate the potential of long-term sequential point clouds.
- Our proposed multi-frame object detection and offline tracking module generates accurate and complete object tracks, which is crucial for downstream refinement.
- An attention-mechanism based refining module is proposed to leverage the long-term temporal contextual information for objects’ attribute predictions.
- We achieve state-of-the-art 3D object detection performance and tracking performance on the challenging WOD with remarkable margins.

2. Related Work

3D Object Detection. Current 3D object detectors usually process the point cloud in different manners: grid-based and point-based. Different grid-split schemes are designed to transform the point cloud into 3D voxels [61, 49, 37, 58], pillars [45, 20] and bird-eye view maps [51] representation. Point-based methods [37, 38, 53, 52, 39, 29] often employ PointNet [31, 32] as a base feature extractor. The hybrid strategy [36, 60, 53, 12] is also utilized to leverage both advantages. Besides, transformer networks make a success to extract point clouds feature by attention mechanism [26, 62, 43, 35, 11, 63], which have shown great potential.
Appendix.

3.1. Preliminary

We select 3DAL [33] as our baseline, which is a SOTA offboard 3D detector consisting of four modules to process a sequence of point clouds. Specifically, the first detection module takes as input $N$ frames of point clouds $\{P_i \in \mathbb{R}^{N_i \times (3+C)} \mid i = 1,2,\ldots,N\}$ ($N_i$ points for each frame, $C$ is the additional feature for each point such as intensity and elongation) and outputs frame-level 3D bounding boxes $\hat{b}_i \in \mathbb{R}^{n_i \times 7}$ and categories. Then, a multi-object baseline tracker links the detected objects across frames as continuous object tracks $\{T_j \in \mathbb{R}^{L_j \times 7}, j \in N_{obj}\}$ ($L$ is track length) with the unique object IDs. For each object track $T_j$, the object-specific LiDAR points are extracted by cropping original point clouds within corresponding bounding boxes, which are then merged together by eliminating ego-motion with frame poses $\{M_i = [R_i|t_i] \in \mathbb{R}^{3\times4}\}$. The third motion classification module is utilized to determine an object’s motion state (static or dynamic) based on its trajectory features. In the final step, the object-centric auto labeling models extract the object’s temporal representation separately based on its predicted motion state, to predict precise boxes. The refined boxes are eventually transferred back to each frame with the inverse frame pose. Please refer to the original paper for more details [33].

There are two factors that affect object-specific temporal context learning: incomplete tracks from upstream module and motion-state-based auto labeling models that ignore common object characteristics. Incomplete object tracks hinder the generation of effective object-specific temporal point cloud data, illustrated in Fig. 2. The sliding-window-based dynamic object refining mechanism fails to use complete temporal contexts, such as the relation between local position and global trajectory and object geometry consistency, as depicted in Fig. 3.

These observations prompt reconsideration of current offboard 3D object detection conventions. As is illustrated in Fig. 4, our evolution focuses on two main aspects: (1) using a multi-frame 3D detector and offline tracker to provide sufficient accurate and complete object tracks, and (2) modernizing the attention-mechanism refining module to reason about object attribute representations in long-term sequential point clouds. Both aspects significantly impact the model’s performance, yet were not thoroughly investigated in prior studies.

3.2. Complete Object Tracks Generation

Our upstream object detection and tracking module aims to generate accurate and complete object tracks, which is essential as the entry point to the whole pipeline.

Object Detection. The competitive CenterPoint [58] is
adopted as our base detector because the anchor-free design would predict dense and redundant bounding boxes. To provide accurate prediction results as much as possible, we strengthen it in three aspects: (1) a combination of five frames of point cloud serves as the input to maximize the contributions rather than performance diminishes [33, 8]; (2) a point density aware module is designed to leverage the raw point features and voxel feature for precise refinement [13]; (3) to improve the adaption towards complex surroundings, test time augmentation [19] (TTA) for point cloud data and multi-model ensemble (different resolution, network structure and capacity) are utilized to boost the detection performance. Please see the details in Appendix.

**Offline Tracking.** Recent multi-object trackers [46], taking the tracking-by-detection path, always struggle in redundant detected bounding boxes which focus much on box-level detection metrics. Taking inspiration from [28, 44], our multi-object tracker utilizes a two-stage data association strategy to mitigate the possibility of false matching. Concretely, the detected boxes are partitioned into two distinct groups based on their confidence scores. The pre-existing object tracks initially engage in data association solely with the high-score group, and subsequently, successfully associated boxes are utilized to update the existing tracks. The un-updated tracks are further associated with the low-score group, and the un-associated boxes are deprecated. In addition, the life cycle of an object is allowed to persist immortally until the sequence terminates, after which any redundant boxes that have not been updated are removed. This operation benefits the re-connection of truncated object tracks and effectively prevents ID switches.

Post-processing is also crucial to generate good-quality tracks. We re-execute our tracking method following reverse time order to generate another group of tracks $T_{1:j}^{wm}$. These tracks are then associated together by a location-aware similarity matching score. Finally, we fuse the paired tracks with the WBF [40] strategy to further ameliorate the missing boxes and stable the motion state, which is called **forward and reverse order tracking fusion.** Besides, we do not run the downstream modules for too short tracks. The boxes of those short tracks, together with the redundant boxes that have not been updated are merged directly into the final auto labels.

**Object Data Preparation.** Given an object track (identified by the unique object ID), we first slightly scale up the RoI area of the tracked boxes by a parameter $\alpha$ along three dimensions, which compensates for abundant contextual information. Then, points that lie within the regions bounded by these enlarged boxes are taken out. We denote this object-specific LiDAR points sequence as $\{P_{j,i}\}$ for object $j$ with length $L_j$ at $i$-th frame of the original sequence, as well as its corresponding tracked box sequence $\{b_{j,i}\}$ and confidence scores $\{S_{j,i}\}$. **3.3. Attribute-based Refining Module**

Previous object-centric auto labeling methods have employed a state-based strategy to refine the proposals generated by upstream modules. This approach not only results in the propagation of misclassification but also disregards the potential similarities between objects. However, it has been observed that, for rigid objects, the geometric shape of an object does not vary significantly over a continuous
period of time, regardless of its motion state. Furthermore, an object’s motion state typically exhibits regular patterns and strong consistency with neighboring moments. Based on these observations, we propose a novel approach that decomposes the traditional bounding box regression task into three distinct modules that predict an object’s geometry, position, and confidence attributes, respectively.

3.3.1 Geometry Refining Model

Geometry-aware Points Generation. The acquisition of complementary information regarding the appearance and shape of an object can be facilitated by obtaining multiple viewpoints of an object. To obtain a cohesive object track, a local coordinate transform operation, involving translation and rotation, similar to the method presented in [37, 30], is initially applied to align the object points to a local box coordinate at various locations. Subsequently, points from different frames are amalgamated irrespective of their original origin. Henceforth, we randomly sample a set of points $\{P_j = \{p_1, ..., p_n\} \in \mathbb{R}^{n \times (3+C)}, n = 4096\}$ for further processing.

Proposal-to-Point Encoding. It is of paramount importance to effectively utilize the geometric information by encoding proposals into object points rather than discarding them after object points extraction [35]. Specifically, for each point $p_k$ of $P_j$ and its corresponding box $b_{j,i}$, we use a point-to-surface approach to compute the projection distance between $p_k$ and the six surfaces of $b_{j,i}$, denoted as $\Delta p^{\text{sf}}_k$. The newly generated point features can be viewed as a better representation of proposal information, which can be expressed as $[p_k, \Delta p^{\text{sf}}_k, \cdots, \Delta p^{\text{sf}}_k]$, where $\cdot$ denotes the concatenate operation.

Attention-based Geometry Interaction across Views. It has been investigated in 3D object detection that a better initialization of object queries would benefit the convergence of the transformer network [4]. Inspired by this observation, we propose to initialize the geometry query features based on object-specific points. Firstly, we randomly select $t$ samples from the whole object track. Each sample has corresponding 256 randomly-selected points. Each point is also augmented by our proposed proposal-to-point encoding approach, and besides, the corresponding confidence scores. Afterwards, a PointNet-structure encoder $\text{ENC}_1$ is adopted to extract features for each selected sample, which is used to initialize as the geometry queries $Q^{\text{geo}} \in \mathbb{R}^{t \times D}$. Then we utilize another encoder $\text{ENC}_2$ to take as input $P_j$ and extract dense point features, which are served as $K^{\text{geo}}$ and $V^{\text{geo}} \in \mathbb{R}^{n \times D}$.

The generated geometry queries are first fed into the multi-head self-attention layer, to encode rich contextual relationships among selected samples and feature dependencies for refining geometry information. The following cross attention between geometry queries and the point features aggregates relevant context onto the object candidates, which reasons pairwise differences to compensate the point features of supplementary views for each geometry query. At last, a feed-forward network (FFN) independently decodes $t$ geometry queries into $t$ geometry sizes, which are then averaged as the final predicted size.

For better residual target regression, we map the proposals’ size to $D$-dim embeddings with a linear projection layer. They are element-wisely summed with the query features. Details of the network architecture are shown in Appendix.

3.3.2 Position Refining Model

Position-aware Points Generation. For $j$-th object, we randomly select the position of a box from its tracked box sequence $\{b_{j,i}\}$ as a new local coordinate system, and subsequently, the other boxes are transformed to this coordinate, as well as the corresponding object-specific points $\{P_{j,i}\}$. Then, a fixed number of points are randomly selected from $\{P_{j,i}\}$ for each frame.

For each point, in addition to calculating the distance to the proposal’s center, we also compute the relative coordinates between each point and eight corners of the corresponding tracked box as $\Delta p^{\text{co}}_k = p_k - p^{\text{co}}$, which results in a 27-dim feature vector. The final position-aware point features can be expressed as $f^{\text{pos}} = [p_k, \Delta p^{\text{co}}_k, \Delta p^{\text{co}}_1, \cdots, \Delta p^{\text{co}}_8]$. To facilitate training, all object tracks are padded to the same length with zeros.

Attention-based Local-to-Global Position Interaction. For an object track, we utilize the same structured query encoder as $\text{ENC}_1$ in Geometry Refining Model (GRM) to generate position queries $Q^{\text{pos}} \in \mathbb{R}^{L \times D}$ for $L$ frame, whose features consist of position-aware features $f^{\text{pos}}$ and confidence scores. Simultaneously, we extract the point features of the entire object track using another encoder that takes $f^{\text{pos}}$ as input. These features serve as $K^{\text{pos}}$ and $V^{\text{pos}} \in \mathbb{R}^{n^{\text{pos}} \times D}$ for subsequent computation, where $n^{\text{pos}}$ is the number of sampled points.

The position queries are first fed into the self-attention module, to capture the relative distance between itself and others. Additionally, we apply a 1D mask near the position of each query to weigh the self-attention. Subsequently, the local position queries $Q^{\text{pos}}$ and global point trajectory features $K^{\text{pos}}, V^{\text{pos}}$ are fed into the cross-attention module to model the local-to-global position contextual relations. Finally, we predict the offsets between each ground-truth center and the corresponding initial center under the local coordinate system, as well as the bin-based heading angle.
4. Experiments

In this section, we first introduce the dataset details and evaluation metrics used in our experiments. We then provide a detailed performance comparison between our DetZero and other SOTA 3D detectors in Sec. 4.2. Then, we validate whether such high-quality “auto labels” by DetZero could play the same role as human labels in Sec. 4.3. In Sec. 4.4, we present the ablation studies and analysis for convincing each component of our entire approach. Please refer to Appendix for more detailed experiments and ablation results.

4.1. Dataset

We conduct experiments on the challenging Waymo Open Dataset [41], which is one of the largest dataset containing total 1150 LiDAR scenes with 798 for training, 202 for validation and 150 for testing. The dataset provides 20-second point clouds data for each scene with a sampling frequency at 10Hz, and 3D annotations for 4 object categories in 360 degree field of view. We follow the evaluation protocol with the official metrics, i.e., average precision (AP) and average precision weighted by heading (APH), and report the results on both LEVEL 1 (L1) and LEVEL 2 (L2) difficulty levels. The L1 evaluation includes objects with more than five LiDAR points and L2 evaluation only includes 3D labels with at least one and no more than five LiDAR points. Note that mAPH (L2) is the main metric for ranking in the Waymo 3D detection challenge.

4.2. Comparing with State-of-the-art Detectors

We present a comprehensive comparison of our DetZero with various state-of-the-art 3D detectors.

As shown in Table 1, our DetZero achieves the best results on Waymo 3D detection leaderboard with 85.15 mAPH (L2) detection performance. For comparisons among methods processing long-term sequential point clouds (at least 100 frames), DetZero surpasses 3DAL [33] with 5.93 (L1) and 9.51 (L2) mAPH on Vehicle, surpasses INT [48] with 6.16 (L1) and 7.69 (L2) mAPH on Vehicle.
4.3. Comparising with Human Labels

mAPH for object tracks generated by our upstream module, our full results further highlight the great potential of the point cloud to denote the entries using TTA or model ensemble techniques.

Vehicle on both frame and multi-frame based methods with a huge margin the validation set of WOD. We outperform other single-tween SOTA 3D detectors and our internal components on performance in Appendix.

MOTA (L2) by a 2 tors [21, 25, 23], DetZero also yields a strong performance compared to state-of-the-art multi-modal fusion 3D detec-
tors and human’s AP computing method. refer to the appendix for more details about the sequences’ IDs for Vehicle 3D and BEV AP (L1 difficulty) under 0.7 and 0.8 IoU threshold.

Table 3. Comparing human labels and auto labels. The results are 3D and BEV AP (L1 difficulty) under 0.7 and 0.8 IoU threshold for Vehicle on 5 sequences selected from WOD val set. Please refer to the appendix for more details about the sequences’ IDs and human’s AP computing method.

Table 4. Intra-domain semi-supervised learning results.

Human performance is measured by the consistency between the humans’ single-frame-based re-labeling and released multi-frame-based ground-truth labels.

We follow their experimental setup to report the mean AP of our DetZero across the 5 selected sequences. In Table 3, we demonstrate superior performance compared to human and 3DAL in particular. With the common 3D AP@0.7 metric, we achieve 3.79 and 4.87 points gains, while the gap is larger in more strict 3D AP@0.8 metric. We obtain similar gains with the BEV AP by ignoring height. To the best of our knowledge, this is the first time that the offboard 3D detector model can outperform the average human labels.

To better study whether such high-quality auto labels could replace human labels for onboard model training, we conduct another intra-domain semi-supervised learning experiment. We choose the single-stage CenterPoint [58] as our student model. Note that the student model takes as input a single frame, and the GT-Paste data augmentation is not used during training. We first randomly select 10% sequences (79 ones) in the WOD training set to train our entire DetZero pipeline. Next, we can generate “auto labels” for the rest 90% sequences (719 ones) in the training set. Afterwards, the student model is trained with different combinations of human labels and “auto labels”.

As shown in Table 4, the first two rows give a performance comparison by reducing the human annotations to

cle, and 7.65 (L1) and 9.09 (L2) mAPH on Pedestrian. DetZero shows great ability to leverage the long-term sequential point clouds for offboard perception. Moreover, compared to state-of-the-art multi-modal fusion 3D detectors [21, 25, 23], DetZero also yields a strong performance gain with at least 3.43 (L1) and 4.63 (L2) mAPH on Vehicle, and 2.93 (L1) and 3.54 (L2) mAPH on Pedestrian. These results further highlight the great potential of the point cloud sequences explored by DetZero. We also ranked 1st place on Waymo 3D tracking challenge leadboard [2] with 75.05 MOTA (L2) by a 9.97 point margin, please see the detailed performance in Appendix.

Additionally, in Table 2, we provide a comparison between SOTA 3D detectors and our internal components on the validation set of WOD. We outperform other single-frame and multi-frame based methods with a huge margin on both Vehicle and Pedestrian. Thanks to the high-quality object tracks generated by our upstream module, our full model gets a significant internal improvement; 6.49 (L1) and 7.68 (L2) mAPH for Vehicle, 3.99 (L1) and 4.67 (L2) mAPH for Pedestrian, more analysis is shown in Sec. 4.4.

4.3. Comparising with Human Labels

It has been shown that humans’ capability of recognizing objects in a dynamic 3D scene has minor fluctuations [33].
10%, which decreases the student model’s performance by 8.53 and 8.6 points for Vehicle, 10.38 and 11.5 points for Pedestrian. Surprisingly, when we add other 90% auto labels, the performance increase with 7.56 and 7.63 points for Vehicle which is close to the first row, 11.79 and 12.36 points for Pedestrian which is already higher. Besides, when we remove the 10% human labels (3rd row), the results are predictably slightly lower than the 4th row, still showing 1.06 and 0.23 gain for Pedestrian. These results demonstrate that “auto labels” generated by our DetZero are qualified for training online models. We visualized the results and found that “auto labels” contain fewer pedestrian labels than human labels, such as the hard samples at a far distance. Hence, the student model trained with our “auto labels” would output fewer false positives compared to the model trained with 100% human labels. More detailed analyses are in Appendix.

### 4.4. Ablation Studies and Analysis

We conduct ablation studies on the WOD validation set to verify all the components of our approach, especially under a fair experimental setting regardless of the techniques for the leaderboard. Additional ablations for the network structures and data augmentations are shown in Appendix.

**Effects of each Component.** We enable different combinations of our proposed modules to evaluate the performances. In addition to the commonly used standard IoU threshold, we also report the performance under a higher IoU threshold to more accurately assess the disparity between predictions and the ground-truth labels.

In Table 5, compared to the upstream results (2nd row), when IoU equals 0.7, the 3rd row shows that the GRM gains 1.92 (L1) and 1.93 (L2) points for Vehicle, and 1.64 and 1.7 points for Pedestrian. As a comparison, the 4th row shows that the PRM gains 2.91, 3.31, 1.25 and 1.44 points respectively. When we combine them together, the performance improves a lot, shown by the 5th row. And the CRM also performs well, by re-scoring the samples based on their qualities. When IoU equals 0.8, we get impressive improvements. Specifically, our full downstream refining module boosts the performance by 26.49% and 29.08% for Vehicle, by 11.82% and 13.45% for Pedestrian. This shows that our entire DetZero tries its best to generate high-quality 3D boxes.

**Cross Evaluation.** In order to better verify the effect of our proposed principle, we reproduce the baseline tracker [46] and motion state based object auto labeling model [33] and make a cross-evaluation between their modules and ours by using the same detection results (ours). In Table 6, the first row can be viewed as the 3DAL approach [33] and achieve the lowest performance. Based on this baseline performance, using attribute-based refining modules yields 1.83 and 1.81 point gains for Vehicle, 1.55 and 1.3

### Table 5. Effect of each component in our DetZero on WOD val set. Metrics are 3D APH of both L1 and L2 difficulties for Vehicle and Pedestrian with a standard IoU threshold (0.7 & 0.5) and a higher IoU threshold (0.8 & 0.6).

<table>
<thead>
<tr>
<th>Det.</th>
<th>Tra.</th>
<th>GRM</th>
<th>PRM</th>
<th>CRM</th>
<th>Vehicle (L1 / L2)</th>
<th>Pedestrian (L1 / L2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>IoU=0.7</td>
<td>IoU=0.8</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>82.57 / 75.09</td>
<td>51.34 / 44.77</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>84.49 / 77.17</td>
<td>56.71 / 49.60</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>85.48 / 78.55</td>
<td>56.35 / 49.45</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>87.81 / 81.01</td>
<td>64.53 / 57.15</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td><strong>89.06 / 82.92</strong></td>
<td><strong>64.94 / 57.84</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>83.23 / 76.47</td>
<td>64.12 / 56.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>84.71 / 78.04</td>
<td>68.33 / 60.48</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>84.32 / 77.78</td>
<td>66.53 / 58.99</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>85.94 / 79.48</td>
<td>70.97 / 63.26</td>
</tr>
</tbody>
</table>

### Table 6. Evaluate the function of different upstream and downstream modules. We reproduce 3DAL [33] and use the subscripts 1 and 2 to represent their and our models respectively. Metrics are standard 3D APH of both L1 and L2 difficulties for Vehicle and Pedestrian.

<table>
<thead>
<tr>
<th>Trk1</th>
<th>Trk2</th>
<th>Ref1</th>
<th>Ref2</th>
<th>Vehicle</th>
<th>Pedestrian</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>L1</td>
<td>L2</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>83.74</td>
<td>76.33</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>85.57</td>
<td>78.14</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>84.92</td>
<td>77.06</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td><strong>89.06</strong></td>
<td><strong>82.92</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>83.92</td>
<td>76.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>85.47</td>
<td>78.24</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>84.64</td>
<td>77.52</td>
</tr>
</tbody>
</table>

### Table 7. Tracking performance comparison on val set of WOD. Metrics are standard 3D MOTA and track-level recall (Recall@track) of L2 difficulty. A ground-truth object track is regarded as a track-level TP only if at least 80% boxes are matched (3D IoU=0.7 for Vehicle, 0.5 for Pedestrian) with those of a single predicted track.

<table>
<thead>
<tr>
<th>Trk1</th>
<th>Trk2</th>
<th>Ref2</th>
<th>MOTA</th>
<th>Recall@track</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Vehicle</td>
<td>Pedestrian</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>57.14</td>
<td>61.78</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>58.41</td>
<td>62.50</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>63.24</td>
<td>65.52</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td><strong>71.36</strong></td>
<td><strong>68.78</strong></td>
</tr>
</tbody>
</table>
point gains for Pedestrian. And using offline tracking provides 1.18 and 0.73 point gains for Vehicle, 0.72 and 0.58 point gains for Pedestrian. For this two groups comparison, the attribute-based refiner improves much more than offline tracker, though the object tracks are not that good, our refiner can still leverage the temporal context information. It also shows that complete object tracks are essential to affect the process of long-term sequential point clouds. The reason is revealed in Table 7, our offline tracker yields to affect the process of long-term sequential point clouds. Evaluated on WOD, our method has 16.32 point and 18.56 point gains on Recall@track respectively, while MOTA is slightly higher. Based on the complete tracks, our attribute-based refiner could further boost the performance, as the last row of Table 6 and 7 is shown. These results demonstrate the strong ability of our DetZero.

**Compare with prior trackers.** We replace our proposed offline tracker with several SOTA trackers but maintain all the other modules in our pipeline, and evaluate the trackers’ performance (Recall@track) and the final performance after refining (3D APH). The other trackers lead to degraded final APH performance because our tracker promises the completeness of tracks (Recall@track). Note that the other trackers would update the geometry size and trajectory of objects, while our offline tracker doesn’t at this step.

<table>
<thead>
<tr>
<th></th>
<th>Recall@track</th>
<th>3D APH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Vehicle</td>
<td>Pedestrian</td>
</tr>
<tr>
<td>AB3DMOT [41]</td>
<td>23.96</td>
<td>26.63</td>
</tr>
<tr>
<td>SimpleTrack [23]</td>
<td>33.86</td>
<td>35.28</td>
</tr>
<tr>
<td>ImmortalTracker [39]</td>
<td>35.34</td>
<td>39.88</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>40.28</strong></td>
<td><strong>45.19</strong></td>
</tr>
</tbody>
</table>

Table 8. Performance (L2) comparison on val set of WOD with different trackers.

**Generalization Ability of Refining Module.** To better verify the generalization ability of our approach, especially the proposed attribute-based refining module, we take as input three upstream results with different qualities for inference. The low group comes from our base detector, while the mid group utilizes the techniques mentioned in Sec. 3.2 to generate high-quality results. In high group, we leverage the image information to further boost the upstream performance. In Table 9, our downstream refining module obtains significant improvements in all three groups. Besides, on both Vehicle and Pedestrian, the improvements of L2 are greater than those of L1. These results further show two strong conclusions: (1) our upstream module can recall hard samples, even if they are not over the IoU threshold of true positive, and (2) our downstream refining module takes advantage of temporal context to optimize these hard samples.

<table>
<thead>
<tr>
<th></th>
<th>Vehicle</th>
<th>Pedestrian</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L1</td>
<td>L2</td>
</tr>
<tr>
<td>upstream (low)</td>
<td>77.86</td>
<td>70.00</td>
</tr>
<tr>
<td>+ Refine</td>
<td>81.43</td>
<td>74.68</td>
</tr>
<tr>
<td><strong>improvement</strong></td>
<td><strong>4.57</strong></td>
<td><strong>4.68</strong></td>
</tr>
<tr>
<td>upstream (mid)</td>
<td>82.57</td>
<td>75.24</td>
</tr>
<tr>
<td>+ Refine</td>
<td>89.06</td>
<td>82.92</td>
</tr>
<tr>
<td><strong>improvement</strong></td>
<td><strong>6.49</strong></td>
<td><strong>7.68</strong></td>
</tr>
<tr>
<td>upstream (high)</td>
<td>83.80</td>
<td>76.99</td>
</tr>
<tr>
<td>+ Refine</td>
<td>89.34</td>
<td>83.57</td>
</tr>
<tr>
<td><strong>improvement</strong></td>
<td><strong>5.54</strong></td>
<td><strong>6.58</strong></td>
</tr>
</tbody>
</table>

Table 9. Verifying generalization ability of our DetZero on WOD val set. Metrics are standard 3D APH of both L1 and L2 difficulties for Vehicle and Pedestrian.

**5. Conclusion**

In this work, we have proposed DetZero, a state-of-the-art offboard 3D detector using long-term sequential point clouds as input. The cores of our success are a multi-frame object detector and offline tracker which generates high-quality complete object tracks, and an attribute-based auto labeling model leveraging the full potential of long-term sequential point clouds. Evaluated on WOD, our method has ranked 1st place, showing remarkable margins over prior art 3D detectors. Moreover, the extensive ablation studies and analysis lead to convincing evaluation and application exploration with such high-quality perception results.

**Acknowledgments**

This project is funded in part by National Key R&D Program of China Project 2022ZD016100, by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission (ITC)’s InnoHK, by General Research Fund of Hong Kong RGC Project 14204021, and by the Science and Technology Commission of Shanghai Municipality (No. 22DZ1100102). Hongsheng Li is a PI of CPII under the InnoHK.

**References**


and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2017. 2


[40] Roman Solovoyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. arXiv preprint, 2019. 4


[49] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. In Sensors, 2018. 2


[52] Zetong Yang, Yanan Sun, Shu Liu, and Jiaya Jia. 3dssd: Point-based 3d single stage object detector. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2020. 2


