TraIL-Det: Transformation-Invariant Local Feature Networks for 3D LiDAR Object Detection with Unsupervised Pre-Training

Supplementary Material

Li Li[1]
li.li4@durham.ac.uk
Tanqiu Qiao[1]
tanqiu.qiao@durham.ac.uk
Hubert P. H. Shum[1]
hubert.shum@durham.ac.uk
Toby P. Breckon[1,2]
toby.breckon@durham.ac.uk

[1] Department of Computer Science
Durham University
Durham, UK

[2] Department of Engineering
Durham University
Durham, UK

In this documentation, we supplement additional materials to support our findings, observations, and experimental results. Specifically, it is organized as follows:

- Sec. A supplements more details on the 3D object detection benchmarks we are using.
- Sec. B supplements more details on Transformation-Invariant Local Feature (TraIL).
- Sec. C acknowledges the public resources used during the course of this work.
- Sec. D attaches additional qualitative results, *i.e.*, the 3D object detection visualizations.

A Details on 3D Object Detection Benchmark

The KITTI 3D object detection task (Sec. A.1) trains detectors to identify classes such as `Car`, `Pedestrian`, and `Cyclist`, requiring both 2D and 3D bounding boxes along with confidence scores. The KITTI dataset emphasizes accurate projections and filtering of objects not visible in image planes. The Waymo Open Dataset (Sec. A.2) enhances autonomous vehicle technologies with high-resolution images and detailed 3D data from multiple LiDARs, capturing objects with unique tracking IDs and specific bounding box criteria. It also includes "No Label Zones" to indicate areas without labels, focusing on detailed spatial awareness and precision.

A.1 KITTI Dataset

The goal in the KITTI 3D object detection task is to train object detectors for the classes `Car`, `Pedestrian`, and `Cyclist`. The object detectors must provide BOTH the 2D 0-based bounding box in the image as well as the 3D bounding box (in the format specified above, *i.e.*, 3D dimensions and 3D locations) and the detection score/confidence. Note that the 2D bounding box should correspond to the projection of the 3D bounding box - this is required to filter objects larger than 25 pixel (height). We also note that not all objects in the point clouds have been labeled. To avoid false positives, detections not visible on the image plane

should be filtered (the evaluation does not take care of this). Similar to the 2D object detection benchmark, we do not count `Van` as false positives for `Car` or `Sitting Person` as false positive for `Pedestrian`. Evaluation criterion follows the 2D object detection benchmark (using 3D bounding box overlap).

A.2 Waymo Open Dataset (WOD)

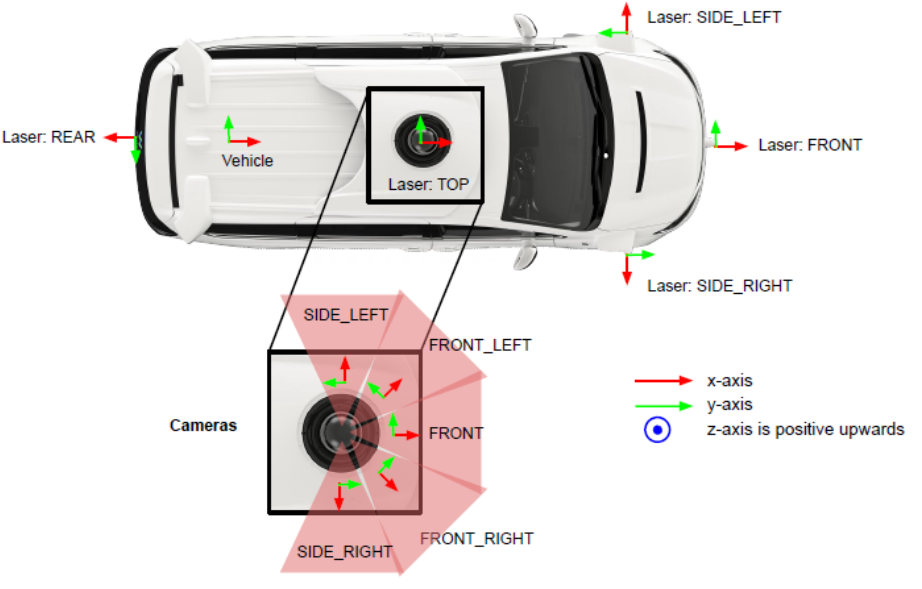

Figure A1: **The sensor setup and configuration on Waymo's autonomous vehicle**. The positions of various laser sensors (REAR, TOP, SIDE_LEFT, SIDE_RIGHT, FRONT) and camera coverage areas (FRONT_LEFT, FRONT, FRONT_RIGHT, SIDE_LEFT, SIDE_RIGHT) are shown. The coordinate systems are shown with red arrows indicating the x-axis, green arrows indicating the y-axis, and a blue circle indicating the z-axis, which is positive upwards.

Fig. A1 shows Waymo sensor setup and sensor configuration on Waymo's autonomous vehicle. Top LiDAR covers a vertical field of view (VFOV) from -17.6 to 2.4 degrees, and its range is 75 meters and covers 360 degrees horizontally. Front, side left, side right, and rear LiDARs covers a relatively smaller area than the top LiDAR. They all include a vertical field of view (VFOV) from -90 to 30 degrees, and their range is 20 meters, which is smaller than the top LiDAR. The following objects have 3D labels: `vehicles`, `pedestrians`, `cyclists`, `signs`. 3D bounding box labels in LiDAR data. The LiDAR labels are 3D 7-DOF bounding boxes in the vehicle frame with globally unique tracking IDs. The bounding boxes have zero pitch and zero roll. Heading is the angle (in radians, normalized to $[-\pi, \pi]$) needed to rotate the vehicle frame +X axis about the Z axis to align with the vehicle's forward axis. Each scene may include an area that is not labeled, which is called a "No Label Zone" (NLZ). NLZs are represented as polygons in the global frame. These polygons are not necessarily convex. In addition to these polygons, each LiDAR point is annotated with a boolean to indicate whether it is in an NLZ or not.

The dataset contains data from five LiDARs (TOP = 1; FRONT = 2; SIDE_LEFT = 3; SIDE_RIGHT = 4; REAR = 5) - one mid-range LiDAR (top) and four short-range LiDARs (front, side left, side right, and rear). The point cloud of each LiDAR is encoded as a range image. Two range images are provided for each LiDAR, one for each of the two strongest returns. It has 4 channels:

- **Channel 0**: range (see spherical coordinate system definition)
- **Channel 1**: LiDAR intensity channel
- **Channel 2**: LiDAR elongation
- **Channel 3**: is_in_nlz (1 = in, -1 = not in)

B Details of Transformation-Invariant Local Feature (TraIL)

Given a fixed integer $k > 0$ denoting the number of nearest point neighbors, and a point cluster P containing at least k points, the Transformation-Invariant Local Feature (TraIL) is a $u \times k$ matrix preserving spatial distances among points in P.

The k-point TraIL matrix is formally defined as follows:

$$\mathrm{TraIL}(P;k) = \mathrm{sort}\left(\left[\mathrm{sort}\left(\boldsymbol{\rho}_{j,1},\ldots,\boldsymbol{\rho}_{j,k}\right)\right]_{j=1}^{u}\right), \tag{B1}$$

where $\boldsymbol{\rho}_{j,i}$ represents the distances from the j-th point in P to its k nearest neighbors. Each row of the TraIL matrix corresponds to one point in P and contains the distances to its k nearest neighbors. For convenience to facilitate comparison of various TraIL matrices, we arrange TraIL lexicographically by sorting Eq. (B1), where $\mathrm{sort}(\cdot)$ on the inner and outer brackets sorts the elements $\boldsymbol{\rho}_{j,l}$ within each row j, and the sorted rows based on their first differing elements, both in ascending order.

The distance between points is defined as:

$$\boldsymbol{\rho}_{j,l} = \left\|\boldsymbol{p}_j - \boldsymbol{p}_{j,l}\right\|_2, \quad \forall l \in \{1,\ldots,k\}, j \in \{1,\ldots,u\}, \tag{B2}$$

where $\boldsymbol{p}_j$ and $\boldsymbol{p}_{j,l}$ denote the 3D coordinates of the j-th point and its l-th nearest neighbor within P, respectively. $\|\cdot\|_2$ is the Euclidean norm to compute the spatial distance.

C Public Resources Used

We acknowledge the use of the following public resources, during the course of this work:

- nuScenes[1] ..... CC BY-NC-SA 4.0
- nuScenes-devkit[2] ..... Apache License 2.0
- The KITTI Vision Benchmark Suite[3] ..... CC BY-NC-SA 4.0
- ProposalContrast[4] ..... MIT License
- VoxSeT[5] ..... MIT License

[1] https://www.nuscenes.org/nuscenes.
[2] https://github.com/nutonomy/nuscenes-devkit.
[3] https://www.cvlibs.net/datasets/kitti/.
[4] https://github.com/yinjunbo/ProposalContrast.
[5] https://github.com/skyhehe123/VoxSeT.

- SpConv[6] . . . . . Apache License 2.0
- Average-Minimum-Distance[7] . . . . . CC BY-NC-SA 4.0
- PyTorch-Lightning[8] . . . . . Apache License 2.0

D More Qualitative Results

TraIL Visualization of 3D Object Detection: We present supporting qualitative results on 3D Object Detection in Figs. D1 and D2. As shown in Figs. D1 and D2, our approach achieves excellent performance in 3D object detection. Although some vehicles are occluded in the RGB images, our method can still rely on the TraIL features from the point cloud to address the issue of occlusion to some extent.

References

(***NOT THE END****; visualization images follow*)

[6] https://github.com/traveller59/spconv.
[7] https://github.com/dwiddo/average-minimum-distance.
[8] https://github.com/Lightning-AI/lightning.

Figure D1: Qualitative results of 3D object detection with our TraIL-Det on KITTI dataset. The predicted 3D bounding box is labeled in the LiDAR point cloud, while its corresponding 2D bounding box is labeled in the RGB image. In the point cloud, white points represent points within the camera field of view (FOV), and purple points indicate points outside the camera FOV. Best viewed in color.

Figure D2: Qualitative results of 3D object detection with our TraIL-Det on KITTI dataset. The predicted 3D bounding box is labeled in the LiDAR point cloud, while its corresponding 2D bounding box is labeled in the RGB image. In the point cloud, white points represent points within the camera field of view (FOV), and purple points indicate points outside the camera FOV. Best viewed in color.