Supplementary File:
Pedestrian Planar LiDAR Pose (PPLP) Network for Oriented Pedestrian Detection Based on Planar LiDAR and Monocular Images

Fan Bu, Trinh Le, Xiaoxiao Du, Ram Vasudevan, and Matthew Johnson-Roberson*

I. Coordinate Systems

We define a camera inertial coordinate system shown in Fig 1 with blue arrows. The origin of the camera inertial coordinate system is the top left corner of the camera image, the x-axis is pointing right along the image width direction, the y-axis is pointing down along the image height direction, and the z-axis is defined by the right-hand rule. We then define an apparent coordinate system for each pedestrian based on the location of their center of mass, as shown in Fig 1 with green arrows. During training, to reduce the error caused by different viewpoints, we transform all ground-truth orientations to the apparent coordinate system using

$$V_A = (R_A^I(\alpha, \beta, \gamma))^{-1} \cdot V_I, \tag{1}$$

where $R_A^I(\alpha, \beta, \gamma)$ is the rotation matrix that rotates the camera inertial coordinate frame into the apparent coordinate frame; α, β, $\gamma = 0$ are the corresponding yaw, pitch, and roll angles; V_A is the ground-truth orientation vector in the apparent frame; and V_I is the ground-truth orientation vector in the camera inertial frame.

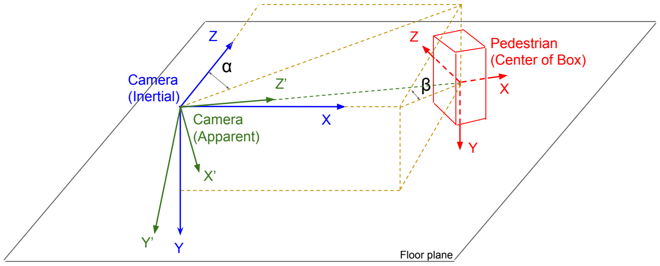

Fig. 1: Relationship between camera inertial coordinates and apparent coordinates.

II. Dataset Specifics

A. CMU Panoptic Dataset

In the CMU Panoptic Dataset, all coordinates (point clouds and label data) were translated to Kinect camera reference frame. From a total of 65 sequences (5.5 hours), five sequences were used in our experiments: "160422_ultimatum1", "160226_haggling1", "160422_haggling1", "160224_haggling1" and "171204_pose3"[†]. These are the only sequences that have ground truth information from body keypoints, high-resolution Kinect RGB-D data available. They also contain the most people in the scene. The "171204_pose3" sequence only has a single person performing a range of actions and we added it to enrich data variety in the training set. Within those five sequences, "160224_haggling1" (8525 frames) is reserved as a testing set. Frames from the remaining four sequences (62,731 frames) are randomly shuffled for training and validation with 3:1 ratio.

B. FCAV M-Air Pedestrian (FMP) Dataset

This dataset contains four short videos with a total recording time of 10 minutes and we used 3,934 frames with good quality groundtruth. In each frame, there are up to two pedestrians walking in the scene, socially interacting and sometimes occluding each other. The last video clip (810 frames) is selected as the test set. Similar to the CMU Panoptic Dataset, frames from the remaining three sequences (3,124 frames) are randomly shuffled for training and validation with 3:1 ratio.

C. Training Parameters

In the CMU Panoptic dataset, all methods were trained with the Adam optimizer at an exponential-decay learning rate. The initial learning rate is 0.0001, decay steps is $30,000$, and the decay factor is 0.8. We validate the models for every $10,000$ steps, and choose the one who performs the best on the validation set to test on the test set. The 3D AVOD method was trained for $240,000$ steps and the 2D AVOD method was trained for $140,000$ steps. In our PPLP network, the OrientNet was trained for $300,000$ steps, and the RPN and the PredictorNet were trained for $120,000$ steps.

When fine-tuning for the FMP dataset, the initial learning rate is doubled for each model from their stopping point while training on the CMU Panoptic dataset. Other training parameters remain the same. The 3D AVOD method was fine-tuned for $20,000$ steps, and the 2D AVOD method was fine-tuned for $30,000$ steps. Due to the lack of full pointclouds in the FMP Dataset, we are not able to generate quaternion groundtruths for the OrientNet in occlusion scenarios. Thus, we fine-tuned the OrientNet with un-occluded image crops for $7,000$ steps. The RPN and the PredictorNet was fine-tuned for $40,000$ steps. The OrientNet was trained for 300,000 steps with a batch size of 1.

*{fanbu, trle, xiaodu, ramv, mattjr}@umich.edu

This work is supported by the Ford Motor Company via the Ford-UM Alliance under award N022977 and by the Office of Naval Research under Award Number N00014-18-1-2575.

[†] Dataset available at http://domedb.perception.cs.cmu.edu/dataset.html.