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

Fan Bu[1], Trinh Le[2], Xiaoxiao Du[2], Ram Vasudevan[1], and Matthew Johnson-Roberson[2]

***Abstract*—Pedestrian detection is an important task for human-robot interaction and autonomous driving applications. Most previous pedestrian detection methods rely on data collected from three-dimensional (3D) Light Detection and Ranging (LiDAR) sensors in addition to camera imagery, which can be expensive to deploy. In this paper, we propose a novel Pedestrian Planar LiDAR Pose Network (PPLP Net) based on two-dimensional (2D) LiDAR data and monocular camera imagery, which offers a far more affordable solution to the oriented pedestrian detection problem. The proposed PPLP Net consists of three sub-networks: an orientation detection network (OrientNet), a Region Proposal Network (RPN), and a PredictorNet. The OrientNet leverages state-of-the-art neural-network-based 2D pedestrian detection algorithms, including Mask R-CNN and ResNet, to detect the Bird's Eye View (BEV) orientation of each pedestrian. The RPN transfers 2D LiDAR point clouds into occupancy grid map and uses a frustum-based matching strategy for estimating non-oriented 3D pedestrian bounding boxes. Outputs from both OrientNet and RPN are passed through the PredictorNet for a final regression. The overall outputs of our proposed network are 3D bounding box locations and orientation values for all pedestrians in the scene. We present oriented pedestrian detection results on two datasets, the CMU Panoptic Dataset and a newly collected FCAV M-Air Pedestrian (FMP) Dataset, and show that our proposed PPLP network based on 2D LiDAR and monocular camera achieves similar or better performance to previous state-of-the-art 3D-LiDAR-based pedestrian detection methods in both indoor and outdoor environments.**

***Index Terms*—Human Detection and Tracking, Computer Vision for Automation, Recognition**

I. INTRODUCTION

WHEN mobile robots (e.g. autonomous vehicles) interact with pedestrians, it is essential to accurately detect pedestrian location and orientation for pedestrian intent recognition and collision-free navigation. LiDAR and camera sensor data can provide depth and color information and are widely used in combination for pedestrian localization and detection [1]–[5].

Manuscript received: August, 15, 2019; Revised November, 14, 2019; Accepted December, 6, 2019.

This paper was recommended for publication by Editor Tamim Asfour upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by a grant from Ford Motor Company via the Ford-UM Alliance under award N022977.

[1]F. Bu and R. Vasudevan are with the Department of Mechanical Engineering, University of Michigan, Ann Arbor, MI 48109 USA {fanbu, ramv}@umich.edu

[2]T. Le, X. Du and M. Johnson-Roberson are with the Department of Naval Architecture and Marine Engineering, University of Michigan, Ann Arbor, MI 48109 USA {trle, xiaodu, mattjr}@umich.edu

Digital Object Identifier (DOI): see top of this page.

Three-dimensional (3D), multi-layer LiDARs are one of the primary sensors in a typical perception system installed on an autonomous vehicle today. 3D LiDAR scans the environment with many beams of light in all directions and returns 3D point cloud data of objects and environments. However, 3D LiDARs can be cost prohibitive for many applications with sensors costing upwards of $100,000 for a Velodyne Alpha Puck [6], for example. In addition, 3D LiDAR point cloud data is computationally intensive to process due to its density and high resolution.

In contrast, a two-dimensional (2D) LiDAR, also known as planar LiDAR, is relatively inexpensive. For example, a Hokuyo UTM-30LX-EW planar LiDAR used in our experiments costs around $4,500, which enables much less expensive deployment. Planar LiDAR also conserves memory and computation resources as the relative volume of data is much smaller. Prior art in this field [7]–[9] has looked at using planar LiDAR for pedestrian detection. However, these works mainly focus on 2D pedestrian tracking and image classification applications and did not address 3D pedestrian location detection and orientation estimation.

In this paper, we propose a novel Pedestrian Planar LiDAR Pose Network (PPLP Net*) that can perform 3D oriented pedestrian detection based on 2D LiDAR data and monocular camera imagery. To our knowledge, our work is the first that achieves accurate 3D pedestrian detection and orientation estimation based solely on 2D LiDAR and monocular images. We envision our work providing an inexpensive alternative to applications where cost, size, weight, and processing power constraints limit the application of other 3D-LiDAR-based approaches. The main contributions of this paper are summarized as follows:

- We propose a novel end-to-end deep neural network architecture for oriented 3D pedestrian detection based on 2D LiDAR data and monocular camera imagery. Our method does not require 3D LiDAR, which can potentially reduce the cost of sensor setups when deploying robots in real applications.
- We propose OrientNet, an image-based orientation detection method to estimate the Bird's Eye View (BEV) orientation of pedestrians directly from 2D monocular RGB images based on an intermediate silhouette representation inspired by [10], which significantly improved orientation estimation accuracy.

*The PPLP code is available at https://github.com/BoomFan/PPLP.

- We provide quantitative results of our method on the CMU Panoptic Dataset [11]–[13] for indoor environment performance evaluation. We also collected an additional public dataset, the *FCAV M-Air Pedestrian (FMP) Dataset*, at an outdoor environment, and provide benchmark results for pedestrian detection using 2D LiDAR and monocular camera from this new dataset.
- We add occupancy grid encoding to previous state-of-the-art method AVOD and show in an ablation study that our method, with added occupancy grid encoding, can detect pedestrians even under heavy occlusion.

II. RELATED WORK

The oriented pedestrian detection pipeline consists of two parts, pedestrian detection and orientation estimation. In this section, we review pedestrian detection methods based on 2D and 3D LiDAR and related feature-based orientation estimation methods.

Pedestrian Detection based on 2D/3D LiDAR: 3D LiDAR provides dense point cloud returns and can help detect and localize pedestrians reliably. Thus, 3D LiDAR has been used predominantly for pedestrian detection in the literature [14]–[18]. Recent methods, such as AVOD [1], MV3D [19], Frustum PointNets [2], and PointPillars [4], employ deep neural network architectures and can achieve high detection performance by learning complex features of objects from 3D point clouds. The drawback, as discussed in the Introduction, is that 3D LiDAR can be expensive to purchase, and their data can be expensive to process and store. On the other hand, 2D LiDAR provides single-layer point cloud and is relatively inexpensive to set up. However, few works exist that exploit 2D LiDAR for detection tasks. Arras *et al.* [20] applied the AdaBoost algorithm to learn a robust classifier to detect people in 2D range data from LiDAR in a cluttered office environment. Shao *et al.* [21] used laser range scanners to track the feet movement of multiple pedestrians. Other work such as [7], [9], [22], [23] fuse information from both 2D LiDAR and RGB camera images to perform object detection. However, these methods focus mostly on 2D pedestrian classification and cannot handle 3D pedestrian detection and localization. In this work, we seek to fill in the gap and perform 3D pedestrian detection while relying on 2D LiDAR as a simpler alternative.

Orientation Estimation: In addition to location detection, it is important to estimate pedestrian orientation to facilitate applications such as intent recognition, trajectory prediction, and social interaction. A common approach for estimating the pedestrian orientation from a single-frame image is to treat it as a multi-classification problem by discretizing the orientations into fixed number of bins [24], [25]. Another approach is to address the task as a continuous regression problem [26]–[28]. However, these methods are all based on hand-crafted image features. Some methods, such as [2], [29], take a hybrid approach of first classifying angles into a discretized set of bins and then regressing them to ground-truth values within each bin for further angle refinement. Deep learning methods such as [30] can directly regress bounding box orientations of objects together with bounding box locations. Most recently, the SilhoNet method [10] has been shown to achieve top performance for object pose estimation from monocular camera images. SilhoNet first predicts an intermediate silhouette representation based on monocular camera images and its translation vectors, and then regress the 3D object orientation based on the predicted silhouette representation. Our work follows a similar architecture to [10]. However, instead of just using object silhouettes, our network also uses pixel-level feature masks from Mask R-CNN [31] as an additional input to further improve the accuracy of orientation estimation.

III. PPLP NETWORK

The proposed PPLP Net consists of three sub-networks, an Orientation Network (OrientNet) for orientation estimation, a Regional Proposal Network (RPN) for generating non-oriented pedestrian bounding box proposals, and a PredictorNet for regressing over ground truth bounding boxes and making final predictions of oriented pedestrian bounding boxes. Sections III-A, III-B, and III-C describe the three subnets in detail. Fig. 1 shows the complete flowchart for our proposed network.

A. OrientNet

In the OrientNet, RGB images from monocular cameras were first resized to 1024×1024 and were passed through a pre-trained[†] Mask R-CNN network [31]. We chose Mask R-CNN network as it can simultaneously output 2D bounding boxes, pixel-level masks for pedestrian locations, and feature maps (predefined 17 body key-points). Then, the cropped RGB image with masks generated from Mask R-CNN were passed into a ResNet-18 [32] architecture, followed by a fully connected layer, to predict the orientation of each image crop in quaternion form. We use the quaternion representation for the orientation angles following [10] since it does not suffer from gimbal lock like the Euler angle representation.

As discussed in [10], true object orientation can vary depending on the camera viewpoints. That is to say, the orientation value of a pedestrian will be different if the pedestrian is at the center of the camera image versus at the edge. In order to prevent such visual ambiguity and to have a consistent quaternion orientation value for all viewpoints, we transformed all orientations into an "apparent coordinate system", where the apparent orientation is estimated as though the image crop (also referred to as region of interest/ROI in [10]) were extracted from the center of the image[‡].

With the help of the above transformation, the OrientNet can be trained effectively for all pedestrians at different 3D locations using the same loss function, defined as

$$Loss_{orient} = Loss(\tilde{q}, q) = log(\epsilon + 1 - |\widetilde{q} \cdot q|), \qquad (1)$$

where q and $\tilde{q}$ are the pedestrian quaternion in the apparent frame and the predicted quaternion for each pedestrian, respectively. The loss defined above is a negatively decreasing

[†]The pre-trained model is available at https://github.com/Superlee506/Mask_RCNN_Humanpose/releases.

[‡]Details about this transformation between camera inertial coordinates and apparent coordinates are provided in the supplementary file.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LRA.2019.2962358, IEEE Robotics and Automation Letters

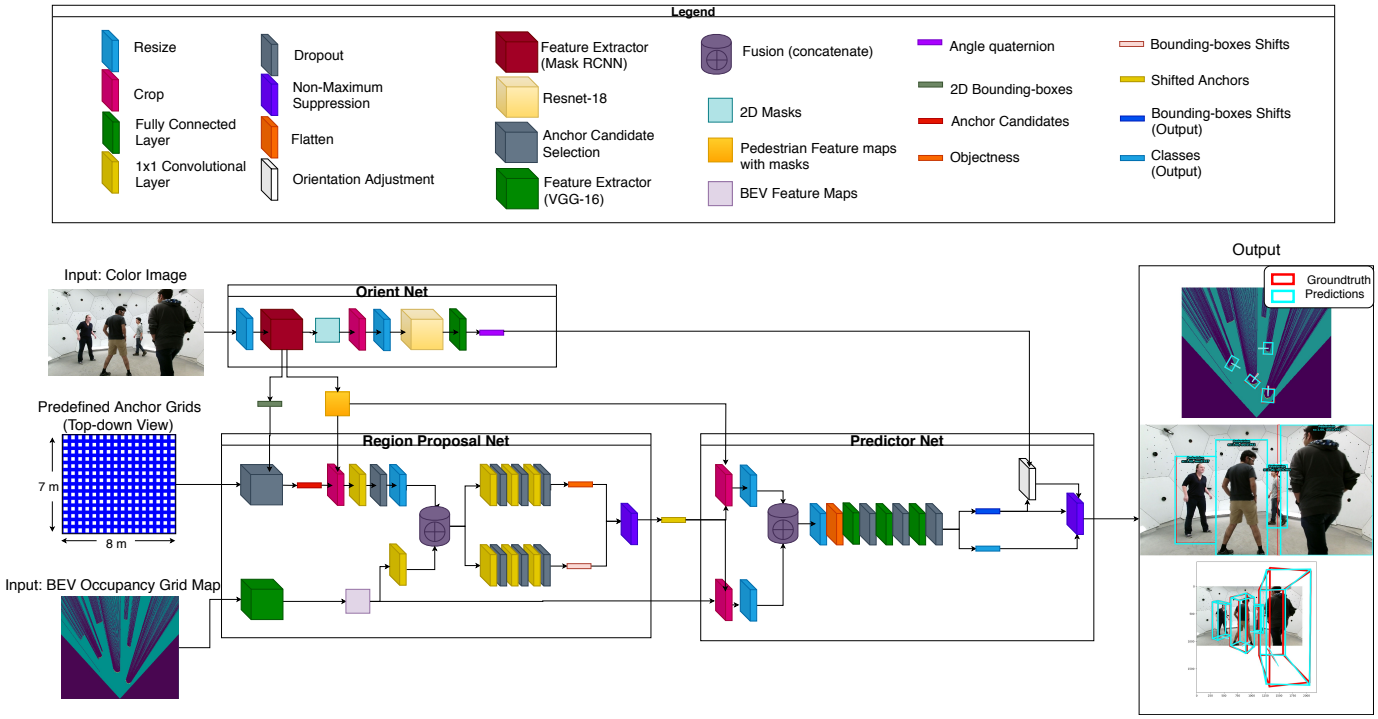


Fig. 1: Overall architecture of our PPLP network. The left column are inputs: RGB images from monocular camera, predefined anchor grids, and BEV occupancy grid map generated from 2D LiDAR signal. The center of this figure shows the three sub-networks: OrientNet for orientation estimation (Section III-A), Region Proposal Net (RPN) for proposal generation (Section III-B) and PredictorNet for final regression (Section III-C). The right column shows a set of example outputs: estimated 3D bounding boxes with location and orientation (in BEV and side view) for all pedestrians in the scene.

loss that is always smaller than zero. The ϵ parameter is a small value to prevent the loss from going to negative infinity. We set $\epsilon = 10^{-4}$ in our experiments, following [10].

B. Region Proposal Network (RPN)

The RPN takes Mask R-CNN feature maps from the OrientNet as well as 2D LiDAR data as inputs and generates non-oriented pedestrian bounding boxes ("proposals") for pedestrian location detection.

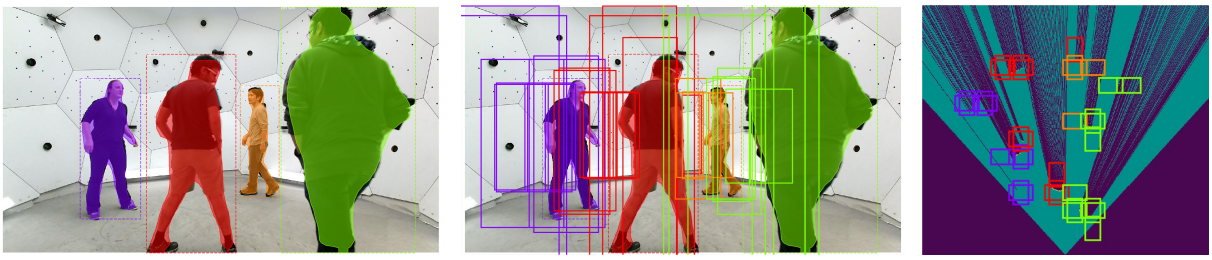
1) Predefined Anchor Grids: We first generate a set of predefined 3D anchor grids to cover all detection areas ($8m \times 7m$) in the scene. Each grid is half a meter apart as suggested in [1] and represents a candidate proposal. Each 3D anchor has a fixed axis-aligned size of (d_x, d_y, d_z) determined from the pedestrian ground truth bounding boxes in the training data, following the discussion in Section III-C in [1].

2) Occupancy Grid Map: We use an occupancy grid map encoding for the 2D LiDAR signals. We evenly divide our $8m \times 7m$ detecting area into $0.01m \times 0.01m$ cells, resulting in a 700×800 BEV grid map. Each cell in the BEV grid map is encoded using one of the three states: free (cell value equals "0"), occupied ("1"), or occluded ("-1"). The occupied state means the cell is occupied by a LiDAR signal (indicating the presence of objects), free/unoccupied state shows there is no pedestrian or other objects, and occluded state indicates there may or may not be an object due to occlusion (such as a pedestrian behind occluding objects or another pedestrian). The Bresenham's line algorithm [33] was used to find all occluded cells where LiDAR ray scans cannot reach. Then, the encoded occupancy grid map was passed into a modified version of the VGG-16 network [34] to extract feature maps for bounding box proposal generation, similar to [1].

3) Anchor Candidate Selection: To select the potential anchors from all predefined 3D anchors, we use a frustum-based selecting strategy inspired by [2] in our RPN. For each 2D bounding box detected by Mask R-CNN, our selecting strategy projects all predefined 3D anchors onto the camera image and the occupancy grid map defined above, assuming the camera parameters are known. If the projection of a 3D anchor on the camera image overlaps with the Mask R-CNN 2D bounding box, and if its projection on the occupancy grid map contains at least one occupied cell, this 3D anchor will be selected as a matched candidate for that Mask R-CNN 2D bounding box. Figure 2a shows an example of four Mask R-CNN 2D bounding boxes and Figure 2b shows the projected boxes of their matched 3D anchor candidates. Figure 2c shows the same 3D anchors as Figure 2b but projected on the BEV occupancy grid map.

4) Fusion (Concatenate): Next, in the regions where all anchor-candidates are located, the BEV feature maps and Mask R-CNN feature maps are cropped, resized and concatenated to form a multi-view feature map vector. To eliminate noise from background or neighboring pedestrians, the background (non-pedestrian) pixel values of Mask R-CNN feature maps are set to zero.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LRA.2019.2962358, IEEE Robotics and Automation Letters



(a) (b) (c)

Fig. 2: An example showing the RPN anchor candidate selecting strategy. (a) 2D bounding boxes (dashed color line) and pixel-wise masks (solid colors) on an RGB image detected by Mask R-CNN. (b) The projection of all matched 3D anchor candidates (solid color boxes) on the same RGB image. (c) The Bird's Eye View of all matched 3D anchor candidates (solid color boxes). In this example, all the occupied anchors on the occupancy grid map are selected as final candidates. Colors of the pedestrian and their anchor candidates are matched. Some of the anchors may be used more than once if their projected 2D boxes overlap with more than one Mask R-CNN result.

5) RPN proposals: Finally, the fused feature vector from previous step is passed through three convolutional layers followed by two fully-connected branches, one for determining objectiveness (pedestrian or not) and the other to calculate the relative shifts in bounding box location of each anchor with respect to their predefined centroid coordinates (the bottom center point of the predefined 3D anchors).

For each pedestrian in the RGB image, the fully connected layer generates a set of potential 3D bounding boxes (proposals) and returns their associated objectness scores. To avoid a large number of proposals, the proposed RPN reduces the number of proposals by performing Non-Maximum Suppression (NMS) [35]. The NMS selects only the proposal bounding boxes that have highest objectness score (highest probability of being a pedestrian) and prune all other boxes with high IoU (Intersection over Union). An IoU threshold of 0.8 in the top-down-view is used in our model.

C. PredictorNet

The pedestrian feature map obtained from OrientNet and the BEV feature map obtained from RPN are passed into the PredictorNet for final prediction. Both feature maps are cropped and resized into 14×14, based on the new anchor proposals from RPN, and concatenated pair by pair. These new feature pairs are then flattened as an one-dimensional vector followed by three fully connected layers. The fully connected layers generate two outputs, one is the class label (pedestrian or non-pedestrian) for each anchor candidate, and the other is the relative shift value in bounding box locations and height, as described in Section III-C of [1]. Dropout layers are applied at a rate of 0.5 between each linear layers to prevent over-fitting during training.

The PredictorNet is trained together with the RPN. The total loss is the summation of RPN loss and PredictorNet loss. Both loss function consists of classification loss and regression loss. The classification loss computes the cross-entropy between predicted classification logits and the ground truth classification vectors. Similar to [1], the regression loss is a smooth L_1 loss for all relative location shifts, following Eq. (3) in [36]. Our network only calculates the average loss for candidate anchors that have at least 0.55 IoU in BEV with the ground truth bounding boxes.

At the end of PredictorNet, we apply another NMS layer to prune the bounding box proposals, similar to the end of Section III-B. The NMS layer with IoU threshold of 0.01 was executed at the end of PredictorNet for final adjustment, following [1]. After generating the location for each predicted pedestrian, the PredictorNet adjusts the orientations that are fed from OrientNet. As discussed in Section III-A, to revert the orientation error caused by camera viewpoint, the 3D orientations for each pedestrian should be rotated from their apparent frame to inertial frame based on the viewpoint decided by their location predictions.

The detection outputs from the NMS module and the adjusted orientations are our final outputs of PPLP Network, which are the estimated oriented 3D bounding boxes for each pedestrian in the scene.

IV. Datasets and Experimental Setup

In this paper, we conduct experiments on the CMU Panoptic Dataset [11]–[13] as it contains annotated pedestrian camera imagery as well as Kinect point cloud data, which is dense enough to extract 2D-LiDAR-like signals reflected from each pedestrian at the height of their waists. We also collected an additional 2D LiDAR and monocular camera dataset from a mobile robot in an outdoor environment for further evaluation.

A. CMU Panoptic Dataset

The CMU Panoptic Dataset [11]–[13] provides RGB-D data from a massive multi-view system of 480 VGA cameras, 31 HD cameras, and 10 Kinect v2 RGB-D sensors inside a Panoptic Studio [13]. This dataset contains annotated 3D body pose, 3D hands, and facial key-point markers of multiple groups of people with varying poses and occlusion levels. In this experiment, we chose "Kinect 50_01" to be our source data as it is horizontally well-aligned with the ground surface and it has the best view of the entire human body.A one-centimeter horizontal slice around human belly/hip position

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LRA.2019.2962358, IEEE Robotics and Automation Letters

is extracted from the 3D point clouds to synthesize planar-LiDAR-like signals. From the available 3D body pose data, a 3D bounding box for each pedestrian is generated as ground-truth, parallel to the ground in the direction the two shoulders are heading. The number of pedestrians varies from two to seven in a frame in training and validation data. A sample image from this dataset can be seen in Fig. 3[§].

B. FCAV M-Air Pedestrian (FMP) Dataset [¶]

The CMU Panoptic dataset was collected in an simulated indoor setting. To evaluate how well our model performs in an outdoor environment, we built a mobile robot platform and collected an additional dataset in an outdoor environment using the "M-Air" facility[‖] at the University of Michigan campus in Ann Arbor, MI, USA in January 2019. This dataset was recorded using an HD camera and a Hokuyo UTM-30LX-EW planar LiDAR mounted on a ROS-enabled Segway robot. A Qualisys Motion Capture system was used to record shoulder key-points of pedestrians for ground truth data. A sample image from this dataset can be seen in Fig. 5.

V. EXPERIMENTAL EVALUATION

This section presents our experimental results on both the CMU Panoptic dataset and the FMP dataset.

A. Baselines: AVOD

We compared our method with three variations on the state-of-the-art 3D object detection algorithm, AVOD [1]. Note that the AVOD method was originally designed to work with 3D LiDARs signals, where it uses a five-layer tensor to represent the height map of the full 3D point clouds and an additional layer for the density map (see Section III-A in [1]). In the CMU dataset, the full 3D point clouds from the depth camera were available, so we were able to train and test the original 3D AVOD model using 3D point clouds. We call this method AVOD(3D/3D). We also trained an AVOD model using 3D point clouds and tested the 3D AVOD model on the extracted 2D-LiDAR-like signals, called the AVOD(3D/2D) method. For the testing data in this AVOD(3D/2D) method, we fit the 2D LiDAR point clouds into the five-layer height map with most layers other than the given 2D signal slice being padded with zero values. In addition, to train the AVOD method for 2D planar LiDAR signals, we adapted the original AVOD five-layer height map tensor to a one-layer tensor while keeping the density map the same. This adapted 2D AVOD model can be trained and tested with 2D LiDAR point clouds directly and we call this method AVOD(2D/2D).

[§]The specific sequences used and data statistics are provided in the supplementary file.

[¶]The FMP dataset is available at https://github.com/umautobots/FMP-dataset.

[‖]M-Air: https://robotics.umich.edu/mair/.

B. Evaluation Metrics

We evaluated our performance using following three metrics, same as the KITTI dataset [30]: 2D Average Precision (AP), BEV AP, and BEV Average Orientation Similarity (AOS). 2D AP and BEV AP evaluates the location detection performance only, and AOS evaluates both orientation and detection accuracy. The mathematical definition of AOS can be found in [30] and it reflects the overall Recall scores while each True Positive detection is weighted by the precision of its orientation. In our experiments, the IoU threshold of 2D AP is set to be 0.5 and the IoU threshold of BEV AP is 0.25, following [1].

C. Results on the CMU Panoptic Dataset

Table I provides a quantitative comparison of the detection performance across all methods on the CMU dataset. Figure 3 provides a visual sample result for our proposed PPLP method and all baseline AVOD models. Figure 3 shows all detection results with classification scores greater than 0.3.

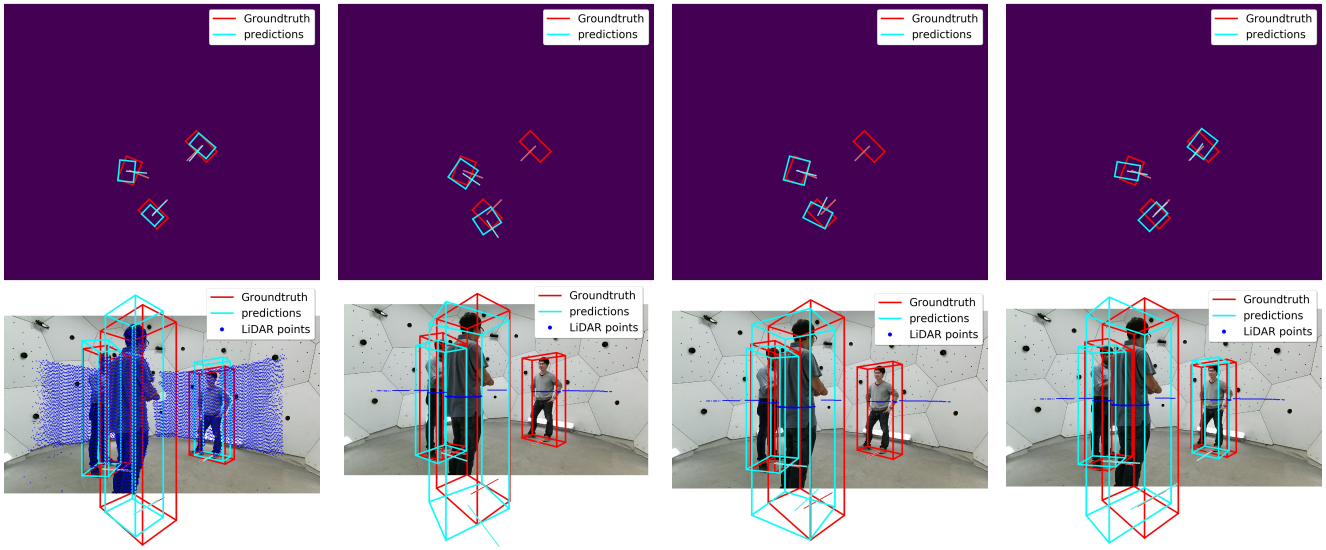
Figure 3a and 3b and the second and third columns in Table I show the comparison between AVOD(3D/3D) and AVOD(3D/2D) models. They were both trained with full 3D point clouds provided in the original CMU dataset and tested on full 3D LiDAR versus extracted 2D LiDAR signals, respectively. As can be seen in Figure 3b, the AVOD(3D/2D) model failed to detect the pedestrian on the right because both height and density information in the point clouds drastically decreased when using 2D LiDAR data. The orientation estimation of the pedestrian in the middle by the AVOD(3D/2D) model was also less accurate because of the lack of 3D information.

Figure 3b and 3c and the third and fourth columns in Table I show the comparison between AVOD(3D/2D) and AVOD(2D/2D) models. The AVOD(3D/2D) was trained on the full 3D point clouds while AVOD(2D/2D) was adapted to fit the one-layer signal. We observed that the AVOD(2D/2D) model obtained better orientation results than the AVOD(3D/2D) model, which makes sense as the AVOD(2D/2D) were trained and tested on the same domain. However, the detection results were still subpar (about 70%). Moreover, we found that the AVOD(3D/2D) model tend to detect more false positives than the AVOD(2D/2D) model, which means the AVOD(3D/2D) model are more sensitive to noise signals, such as the points reflected from the wall.

Figure 3c and 3d and fourth and sixth columns in Table I show the comparison between AVOD(2D/2D) model and our proposed PPLP(2D/2D) method on 2D point cloud data. We observe that our proposed PPLP network achieves higher location detection and orientation estimation accuracy by a large margin (20%-40%) than comparison AVOD (3D/2D) and AVOD (2D/2D) methods (all tested on 2D data). Our PPLP network also correctly detects the person on the right in Figure 3d while both AVOD (3D/2D) and AVOD (2D/2D) methods missed.

Note that the second column (AVOD 3D/3D, greyed) in Table I shows the performance of AVOD method trained on full 3D point clouds and tested with 3D point cloud (not

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LRA.2019.2962358, IEEE Robotics and Automation Letters



(a) AVOD (train/test : 3D/3D) (b) AVOD (train/test : 3D/2D) (c) AVOD (train/test : 2D/2D) (d) PPLP (train/test : 2D/2D)

Fig. 3: Results on CMU Panoptic Dataset. The first row of images are the results in Bird's Eye View within the detection area. The second row of images are the results on the camera view. Blue dots show the synthetic LiDAR signals; Red boxes with red orientation vectors are the ground-truths; Cyan boxes with cyan orientation vectors are the predictions. Same for all following figures.

2D) and thus is not a fair comparison against all other test results on 2D data. Nevertheless, we included this column as an upper bound to show how 3D-point-cloud-based algorithm behaves if provided with full 3D LiDAR signals. Since the 3D point clouds contain more information for pedestrian location and contours than 2D point clouds, we observed, as expected, that this AVOD(3D/3D) method was able to yield high detection accuracy. However, it is worth noting that our proposed PPLP method, trained on 2D LiDAR data only, can achieve comparable (only about 2-5% worse) performance with the AVOD method trained and tested on full 3D LiDAR point clouds.

TABLE I: Experimental results on the CMU Panoptic dataset. The second row shows different training and testing combinations.

Method Name	AVOD			Ours	
				height-density	occupancy (proposed)
Pointclouds (train/test)	$3D/3D$	$3D/2D$	$2D/2D$	$2D/2D$	$2D/2D$
pedestrian 2D detection AP	94.2	58.2	76.7	86.6	**89.3**
pedestrian BEV AP	94.4	58.1	77.0	89.4	**92.2**
pedestrian BEV AOS:	93.0	35.2	72.7	86.3	**90.4**

We also conducted an ablation study to observe the effect of adding occupancy encoding. We compared the proposed method with a version of our method without the occupancy encoding, similar to AVOD's BEV encoding. We call this method "Ours (height-density)". The results without and with occupancy encoding were reported in the last two columns of Table I. We observed that our proposed full version (with occupancy encoding) outperforms the non-occupancy-encoding version by approximately 3% accuracy in terms of BEV AP for location detection, and 4% in terms of BEV AOS for orientation estimation. A visual example was shown in Figure 4, where the leftmost pedestrian (shown with the blue overlay mask) was heavily occluded by the second pedestrian from the left (orange mask). Without the proposed occupancy encoding with occlusion information (using just AVOD's BEV encoding), the detector failed to detect the occluded leftmost pedestrian in blue as shown in Figure 4a. On the other hand, our proposed occupancy encoding method helps generate locations and orientation estimations for all three pedestrians in the scene, as shown in Figure 4b.

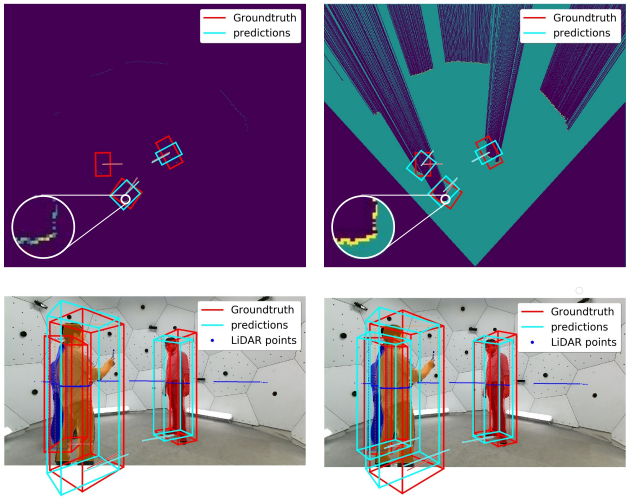
We also conducted additional ablation studies on the effect of OrientNet and results are provided in the supplementary file.

D. Results on the FMP Dataset

The FMP dataset was collected with only a 2D planar LiDAR and a monocular camera without full 3D LiDAR data, so we do not compare with the AVOD(3D/3D) method in this section. We trained the AVOD(3D/2D) model on the CMU Panoptic dataset with 3D point clouds, and trained AVOD(2D/2D) model and our proposed model on the modified CMU Panoptic dataset with 2D point clouds. Then, we fine-tuned all three models with the 2D FMP dataset. We tested all models with the same FMP test dataset.

Table II shows the quantitative comparison results on the FMP dataset. As mentioned in Section V-C, the AVOD(3D/2D) model is more sensitive to the noise than the AVOD(2D/2D) model, which makes the AVOD(3D/2D) model behave worse

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LRA.2019.2962358, IEEE Robotics and Automation Letters



(a) AVOD's BEV encoding (b) Our occupancy encoding

Fig. 4: Performance of different encodings under occlusion. The first row of images are the results on the BEV map of two different encodings. In column (a), the top left image shows the density map of the AVOD method. The dark blue color corresponds to zero density value, the white circle shows the zoom-in view of the BEV map around a person, with more yellow color corresponding to higher density. In column (b), the top right image shows the occupancy map of our proposed method. The dark blue color corresponds to occluded space, the green color corresponds to free space, and the yellow points in zoom-in view corresponds to the occupied cells as described in Section III-B2. The second row shows the results of both methods in the camera view.

than the AVOD(2D/2D) model in the CMU Panoptic dataset. However, for the FMP Dataset, there is less noise in the dataset as the data collection area has less obstacles than the CMU Panoptic dataset. For this reason, we observed that the AVOD(3D/2D) model performs better than the AVOD(2D/2D) model in the FMP dataset (second and third columns in Table II). The last column of Table II shows the detection results of our proposed PPLP method against comparison AVOD models. We observed that our proposed method outperforms both the 2D and 3D AVOD models across all evaluation metrics. Figure 5 shows a visual example of the detection results on the FMP dataset. As can be seen, the location and orientation estimation by our method better matches the ground-truth than comparison AVOD methods.

VI. CONCLUSIONS

In this paper, we proposed PPLP, an end-to-end network for 3D pedestrian detection based on 2D LiDAR and monocular imagery. Evaluated on two distinctive datasets, we show that our proposed method, based solely on 2D LiDAR data, can achieve comparable or better results with the state-of-the-art 3D-LiDAR-based AVOD method in both pedestrian location and orientation detection in both indoor and outdoor environments. By proposing OrientNet, a subnet in PPLP leveraging

TABLE II: Experimental results on the FMP dataset. The second row shows different training and testing combinations.

Method Name	AVOD		Ours (proposed)
Pointclouds (train/test)	$3D/2D$	$2D/2D$	$2D/2D$
pedestrian 2D detection AP	90.1	46.1	**96.8**
pedestrian BEV AP	90.1	46.1	**96.7**
pedestrian BEV AOS:	68.9	38.3	**77.3**

the Mask R-CNN [31], we show that a pedestrian's bird's-eye view orientation can be accurately and directly estimated from 2D RGB imagery. Our method also improves the accuracy of pedestrian detection, especially for partially occluded pedestrians, by using a frustum-based candidate selecting strategy in our region proposal network.

One drawback we observed from our results such as Figure 4 is that when a pedestrian is heavily occluded by others, the orientation of the occluded pedestrian could be occasionally incorrectly matched to another person's image mask, thus causing errors in orientation estimation. Future work will include investigating new methods to correct the mask matching problem and further improve the detection performance.

Currently, there are limited datasets for 2D LiDAR and monocular camera. Future work will include collecting and applying our methods across more planar LiDAR and monocular camera datasets to explore in-the-wild pedestrian behavior in varied settings. Our method currently takes approximately 1.2 seconds per frame to compute; we will seek to improve the computation time in our future work. Alternative methods for improving orientation accuracy can also be explored.

ACKNOWLEDGMENT

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. We thank Junming Zhang for helpful discussions. We also thank the authors of the CMU Panoptic dataset for making this dataset available [11].

REFERENCES

[1] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, "Joint 3d proposal generation and object detection from view aggregation," in *IEEE/RSJ Int. Conf. Intell. Robots and Systems (IROS)*, 2018, pp. 1–8.

[2] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 918–927.

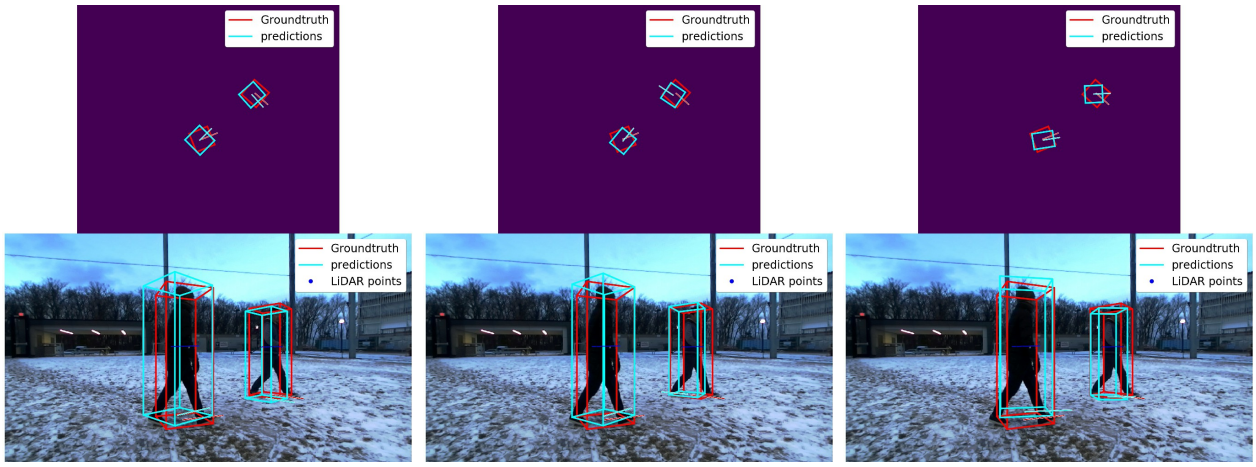
[3] Z. Yang, Y. Sun, S. Liu, X. Shen, and J. Jia, "Ipod: Intensive point-based object detector for point cloud," *arXiv preprint arXiv:1812.05276*, 2018.

[4] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12 697–12 705.

[5] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[6] M. Deutscher, "With new 128-laser sensor, ouster ups the ante on lidar," https://siliconangle.com/2019/01/03/new-128-channel-sensor-ouster-ups-ante-lidar, Jan. 2019, accessed: 2019-06-06.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LRA.2019.2962358, IEEE Robotics and Automation Letters



(a) AVOD (train/test : 3D/2D) (b) AVOD (train/test : 2D/2D) (c) PPLP (train/test : 2D/2D)

Fig. 5: Results on FMP Dataset.

[7] C. Premebida, G. Monteiro, U. Nunes, and P. Peixoto, "A lidar and vision-based approach for pedestrian and vehicle detection and tracking," in *IEEE Intell. Transp. Syst. Conf.*, 2007, pp. 1044–1049.

[8] B. Wu, J. Liang, Q. Ye, Z. Han, and J. Jiao, "Fast pedestrian detection with laser and image data fusion," in *6th Int. Conf. Image and Graphics*. IEEE, 2011, pp. 605–608.

[9] B.-Z. Lin and C.-C. Lin, "Pedestrian detection by fusing 3d points and color images," in *IEEE/ACIS 15th Int. Conf. Comput. Inf. Sci.*, 2016, pp. 1–5.

[10] G. Billings and M. Johnson-Roberson, "SilhoNet: An RGB Method for 6D Object Pose Estimation," *IEEE Robot. Autom. Lett.*, vol. 4, no. 4, pp. 3727–3734, Oct 2019.

[11] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. Godisart, B. Nabbe, I. Matthews, *et al.*, "Panoptic studio: A massively multiview system for social interaction capture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 190–204, 2017.

[12] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1145–1153.

[13] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, "Panoptic studio: A massively multiview system for social motion capture," in *The IEEE Int. Conf. Comput. Vis. (ICCV)*, December 2015.

[14] L. Spinello, K. O. Arras, R. Triebel, and R. Siegwart, "A layered approach to people detection in 3d range data," in *24th AAAI Conf. Artificial Intell.*, 2010.

[15] K. Kidono, T. Miyasaka, A. Watanabe, T. Naito, and J. Miura, "Pedestrian recognition using high-definition lidar," in *IEEE Intell.t Vehicles Symp.*, 2011, pp. 405–410.

[16] K. Li, X. Wang, Y. Xu, and J. Wang, "Density enhancement-based long-range pedestrian detection using 3-d range data," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 5, pp. 1368–1380, 2016.

[17] A. Börcs, B. Nagy, and C. Benedek, "Instant object detection in lidar point clouds," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 7, pp. 992–996, 2017.

[18] Y. Choi, N. Kim, S. Hwang, K. Park, J. S. Yoon, K. An, and I. S. Kweon, "Kaist multi-spectral day/night data set for autonomous and assisted driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 3, pp. 934–948, 2018.

[19] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1907–1915.

[20] K. O. Arras, O. M. Mozos, and W. Burgard, "Using boosted features for the detection of people in 2d range data," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2007, pp. 3402–3407.

[21] X. Shao, H. Zhao, K. Nakamura, K. Katabira, R. Shibasaki, and Y. Nakagawa, "Detection and tracking of multiple pedestrians by using laser range scanners," in *IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, 2007, pp. 2174–2179.

[22] L. Oliveira, U. Nunes, P. Peixoto, M. Silva, and F. Moita, "Semantic fusion of laser and vision in pedestrian detection," *Pattern Recognition*, vol. 43, no. 10, pp. 3648–3659, 2010.

[23] L. Spinello, R. Triebel, and R. Siegwart, "A trained system for multimodal perception in urban environments," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA) Workshop on People Detection and Tracking*, 2009.

[24] M. Ozuysal, V. Lepetit, and P. Fua, "Pose estimation for category specific multiview object localization," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 778–785.

[25] A. Ghodrati, M. Pedersoli, and T. Tuytelaars, "Is 2d information enough for viewpoint estimation?" in *British Mach. Vis. Conf. (BMVC)*, vol. 101, 2014, p. 102.

[26] D. Teney and J. Piater, "Multiview feature distributions for object detection and continuous pose estimation," *Computer Vision and Image Understanding*, vol. 125, pp. 265–282, 2014.

[27] K. Hara and R. Chellappa, "Growing regression forests by classification: Applications to object pose estimation," in *Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2014, pp. 552–567.

[28] ——, "Growing regression tree forests by classification for continuous object pose estimation," *Int. J. Comput. Vis.*, vol. 122, no. 2, pp. 292–312, 2017.

[29] S. Shi, X. Wang, and H. Li, "Pointrcnn: 3d object proposal generation and detection from point cloud," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 770–779.

[30] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3354–3361.

[31] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2980–2988.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[33] R. A. Earnshaw, *Fundamental Algorithms for Computer Graphics: NATO Advanced Study Institute Directed by JE Bresenham, RA Earnshaw, MLV Pitteway*. Springer, 1985.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[35] R. Rothe, M. Guillaumin, and L. Van Gool, "Non-maximum suppression for object detection by passing messages between windows," in *Asian Conf. Comput. Vis.* Springer, 2014, pp. 290–306.

[36] R. Girshick, "Fast r-cnn," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

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.

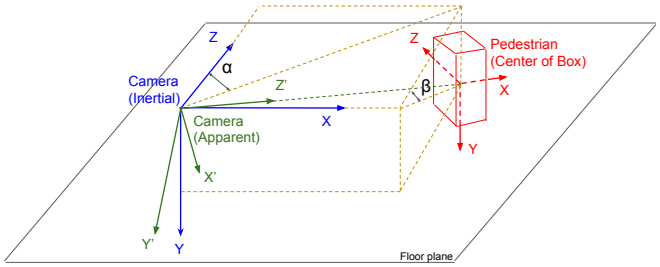


Fig. 1: Relationship between camera inertial coordinates and apparent coordinates.

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 key-points, 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

The FMP dataset was collected from an HD camera and a Hokuyo UTM-30LX-EW planar LiDAR mounted on a ROS-enabled Segway mobile robot platform. The dataset‡ was collected in an outdoor environment using the "M-Air" facility at the University of Michigan campus in Ann Arbor, MI, USA in January 2019. A Qualisys Motion Capture system was used to record shoulder key-points of pedestrians for ground truth data. The FMP dataset contains four short videos with a total recording time of 10 minutes and we used 3,934 frames with good quality ground truth. In each frame, there are up to two pedestrians walking in the scene, interacting with and sometimes occluding each other. The last video clip (810 frames) was 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

*{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.

‡The FMP dataset is available at https://github.com/umautobots/FMP-dataset.

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.

III. Ablation Study on OrientNet

This section provides the results of an ablation study on the effect of OrientNet for orientation estimation.

In our manuscript, we used three metrics to evaluate the performance of our network, which is a set of widely used metrics in pedestrian detection, same as the KITTI dataset [1]: 2D Average Precision (AP), Bird's Eye View (BEV) AP, and BEV Average Orientation Similarity (AOS). The 2D AP and BEV AP only evaluates the location detection performance and AOS evaluates both orientation and location detection accuracy. To calculate BEV AP, all 3D pedestrian anchor boxes in the Bird's Eye View were analyzed. The 3D boxes which overlap more than 50% with the ground truths were counted as True Positives (TP), while the missing (undetected) boxes were counted as False Negatives (FN). Then, the AP was calculated by the recall rate $r = \frac{TP}{TP+FN}$.

On the other hand, the BEV AOS score depends on both orientation and location. The AOS metric [1] was defined as

$$AOS = \frac{1}{11} \sum_{r \in \{0, 0.1, \ldots, 1\}} \max_{\tilde{r}: \tilde{r} \geq r} s(\tilde{r}), \tag{2}$$

where r is the object detection recall rate defined above. The orientation similarity $s(r) \in [0, 1]$ is a function of the recall rate r, defined as

$$s(r) = \frac{1}{|D(r)|} \sum_{i \in D(r)} \frac{1 + cos\Delta_\theta^{(i)}}{2} \delta_i, \tag{3}$$

where $D(r)$ denotes the set of all object detections at recall rate r and $\Delta_\theta^{(i)}$ is the difference in angle between estimated and ground truth orientation of detection i [1]. This means that the AOS metric cannot evaluate orientation alone. The AOS metric depends on both the orientation performance and the detection results (the recall rate). So far, we could not find a widely-used, well-defined metric to evaluate how good the orientation estimation is by itself regardless of location detection performance.

In this supplementary file, we define a metric based on the AOS to reflect orientation evaluation only and we named this new metric "Pure Orientation Score (POS)". We define POS as

$$POS = \frac{1}{11} \sum_{r \in \{0, 0.1, \ldots, 1\}} \max_{\tilde{r}: \tilde{r} \geq r} p(\tilde{r}), \tag{4}$$

$$p(r) = \frac{1}{|P(r)|} \sum_{i \in P(r)} \frac{1 + cos\Delta_\theta^{(i)}}{2} \delta_i, \tag{5}$$

where $P(r)$ denotes the set of all True Positive detections at recall rate r, and the rest of the notations are the same as in AOS. The difference between POS and AOS is that POS only calculates the score of True Positives (TP), while AOS evaluates all detection. The POS metric essentially gives a weight to each TP based on the orientation angle results (the $\frac{1+cos\Delta_\theta^{(i)}}{2}$ term). This way, since all TP match the location of the ground-truth bounding boxes, the POS metric essentially evaluates the effect of orientation alone.

TABLE I: Experimental results on the CMU Panoptic dataset. The POS row roughly estimates the average angle error.

Method Name	AVOD			Ours		OrientNet
				height-density	occupancy (proposed)	
Pointclouds (train/test)	$3D/3D$	$3D/2D$	$2D/2D$	$2D/2D$	$2D/2D$	N/A
pedestrian 2D detection AP	94.2	58.2	76.7	86.6	**89.3**	91.2
pedestrian BEV AP	94.4	58.1	77.0	89.4	**92.2**	91.2
pedestrian BEV AOS:	93.0	35.2	72.7	86.3	**90.4**	88.6
POS	98.5	60.5	94.4	96.5	**98.0**	97.1
$\Delta(\bar{\theta})$ (degree)	$\pm 14.1^\circ$	$\pm 77.9^\circ$	$\pm 27.4^\circ$	$\pm 21.6^\circ$	$\pm \mathbf{16.3}^\circ$	$\pm 19.6^\circ$

Then, we designed the following ablation study on the effect of orientNet using the POS metric. First, we ran Mask R-CNN on the CMU Panoptic dataset, and Mask R-CNN generated pixel-level RGB masks for each detected pedestrian. Second, we matched the Mask R-CNN masks with the groundtruth bounding box locations using a matching strategy as described in the footnote*. In the CMU dataset, around 91.2% of the Mask R-CNN detections were obtained as the correct matches with the groundtruth labels, and we computed the POS score for these true positive detection results only (assuming their location is correct). Another metric, the average angle error $\Delta(\bar{\theta})$, can also be computed from the POS. Table I shows the extended experimental results including POS and $\Delta(\bar{\theta})$ evaluations across all methods and for OrientNet alone (ablation study). As shown, the POS score of AVOD(3D/3D) is still the highest among all POS scores,

*Notes on the matching strategy: Suppose there are two pedestrians, A and B, standing in the same camera image, and suppose the arm of pedestrian A is occluding the waist of pedestrian B. In this case, Mask R-CNN may generate one whole mask for pedestrian A but may generate two separate masks for pedestrian B (top body and lower body). However, when we try to match the Mask R-CNN result with the groundtruth label, which mask of pedestrian B should be matched with pedestrian B groundtruth? Intuitively, we want to match the mask which occupies the largest part of pedestrian B. With the help of point cloud, for each mask, we extract the point clouds in the mask area and then compute the number of points in those point clouds that actually lie within the groundtruth box in BEV. Only the groundtruth box that contains the highest number of point clouds within is selected. As a result, the mask that occupies smaller part of pedestrian B will have smaller number of point clouds within groundtruth box, hence, is discarded. Under this strategy, it is guaranteed that only the Mask R-CNN detection of pedestrian A and the largest detection of pedestrian B are matched to correct groundtruth labels, and can be passed to the OrientNet to test the accuracy of orientation detection. In our experiments, we observed that such matching strategy could not match pedestrian B to any ground truth box if pedestrian B was occluded too much or if there were not enough point clouds extracted from the mask for algorithm to continue. In these cases, only the Mask R-CNN detection results of pedestrian A were passed to OrientNet for evaluation.

which confirms our observation that dense 3D LiDAR signals can indeed provide more information about pedestrian 3D orientations. In our task where only 2D LiDAR is available, the POS score for OrientNet alone given camera images is higher than the comparison methods AVOD(3D/2D) and the AVOD(2D/2D) given 2D LiDAR data, which shows the effectiveness of OrientNet for orientation estimation and confirms our hypothesis that camera images can be used to effectively estimate 3D pedestrian orientations.

IV. Ablation Study on Mask R-CNN color input for OrientNet

We conducted an additional experiment to compare the performance of using black-and-white (binary) silhouette inputs from [2] and RGB masked inputs from Mask R-CNN. In this experiment, we tried both the black-and-white input and the masked RGB color input on OrientNet, and evaluated their performances using the same metric as we defined in (4) and (5). Table II shows the comparison results. As shown, adding color inputs improved the POS of OrientNet by 9%, and the average angle error has been reduced from $\pm 39.8^\circ$ to $\pm 19.6^\circ$ compared with just using the black-and-white silhouette directly from [2].

TABLE II: Experimental results of OrientNet using different input crops from the CMU Panoptic dataset. The POS row roughly estimates the average angle error.

	OrientNet	
Input format	black-and-white silhouette (SilhoNet)	masked RGB color (Ours)
pedestrian 2D detection AP	**91.2**	**91.2**
pedestrian BEV AP	**91.2**	**91.2**
pedestrian BEV AOS:	80.7	**88.6**
POS	88.4	**97.1**
$\Delta(\theta)$ (degree)	$\pm 39.8^\circ$	$\pm \mathbf{19.6}^\circ$

References

[1] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3354–3361.

[2] G. Billings and M. Johnson-Roberson, "SilhoNet: An RGB Method for 6D Object Pose Estimation," *IEEE Robot. Autom. Lett.*, vol. 4, no. 4, pp. 3727–3734, Oct 2019.