EXHIBIT 1029

This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Pseudo-LiDAR from Visual Depth Estimation: Bridging the Gap in 3D Object Detection for Autonomous Driving

Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q. Weinberger
Cornell University, Ithaca, NY
{yw763, wc635, dg595, bh497, mc288, kqw4}@cornell.edu

Abstract

3D object detection is an essential task in autonomous driving. Recent techniques excel with highly accurate detection rates, provided the 3D input data is obtained from precise but expensive LiDAR technology. Approaches based on cheaper monocular or stereo imagery data have, until now, resulted in drastically lower accuracies — a gap that is commonly attributed to poor image-based depth estimation. However, in this paper we argue that it is not the quality of the data but its representation that accounts for the majority of the difference. Taking the inner workings of convolutional neural networks into consideration, we propose to convert image-based depth maps to pseudo-LiDAR representations — essentially mimicking the LiDAR signal. With this representation we can apply different existing LiDAR-based detection algorithms. On the popular KITTI benchmark, our approach achieves impressive improvements over the existing state-of-the-art in image-based performance — raising the detection accuracy of objects within the 30m range from the previous state-of-the-art of 22% to an unprecedented 74%. At the time of submission our algorithm holds the highest entry on the KITTI 3D object detection leaderboard for stereo-image-based approaches.

1. Introduction

Reliable and robust 3D object detection is one of the fundamental requirements for autonomous driving. After all, in order to avoid collisions with pedestrians, cyclist, and cars, a vehicle must be able to detect them in the first place.

Existing algorithms largely rely on LiDAR (Light Detection And Ranging), which provide accurate 3D point clouds of the surrounding environment. Although highly precise, alternatives to LiDAR are desirable for multiple reasons. First, LiDAR is expensive, which incurs a hefty premium for autonomous driving hardware. Second, over-reliance on a single sensor is an inherent safety risk and it would be advantageous to have a secondary sensor to fall-back onto in case of an outage. A natural candidate are images from stereo or monocular cameras. Optical cameras are highly affordable (several orders of magnitude cheaper than LiDAR), operate at a high frame rate, and provide a dense depth map rather than the 64 or 128 sparse rotating laser beams that LiDAR signal is inherently limited to.

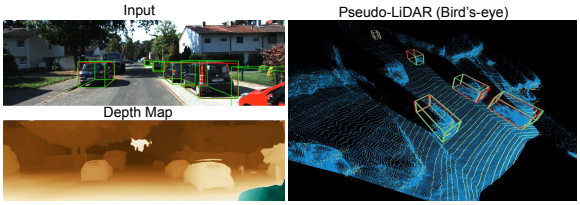

Figure 1: **Pseudo-LiDAR signal from visual depth estimation.** Top-left: a KITTI street scene with super-imposed bounding boxes around cars obtained with LiDAR (red) and pseudo-LiDAR (green). Bottom-left: estimated disparity map. Right: pseudo-LiDAR (blue) vs. LiDAR (yellow) — the pseudo-LiDAR points align remarkably well with the LiDAR ones. Best viewed in color (zoom in for details.)

Several recent publications have explored the use of monocular and stereo depth (disparity) estimation [13, 19, 32] for 3D object detection [5, 6, 22, 30]. However, to-date the main successes have been primarily in supplementing LiDAR approaches. For example, one of the leading algorithms [17] on the KITTI benchmark [11, 12] uses sensor fusion to improve the 3D average precision (AP) for cars from 66% for LiDAR to 73% with LiDAR and monocular images. In contrast, among algorithms that use only images, the state-of-the-art achieves a mere 10% AP [30].

One intuitive and popular explanation for such inferior performance is the poor precision of image-based depth estimation. In contrast to LiDAR, the error of stereo depth estimation grows quadratically with depth. However, a visual comparison of the 3D point clouds generated by LiDAR and a state-of-the-art stereo depth estimator [3] reveals a high quality match (cf. Fig. 1) between the two data modalities — even for faraway objects.

Ex.-1029 - ADB Safegate et al.
ADB Safegate et al. v. Honeywell International Inc. – PGR2023-00053
Page 2

In this paper we provide an alternative explanation with significant performance implications. We posit that the major cause for the performance gap between stereo and LiDAR is not a discrepancy in depth accuracy, but a poor choice of representations of the 3D information for ConvNet-based 3D object detection systems operating on stereo. Specifically, the LiDAR signal is commonly represented as 3D point clouds [23] or viewed from the top-down "bird's-eye view" perspective [33], and processed accordingly. In both cases, the object shapes and sizes are invariant to depth. In contrast, image-based depth is densely estimated for each pixel and often represented as additional image channels [6, 22, 30], making far-away objects smaller and harder to detect. Even worse, pixel neighborhoods in this representation group together points from far-away regions of 3D space. This makes it hard for convolutional networks relying on 2D convolutions on these channels to reason about and precisely localize objects in 3D.

To evaluate our claim, we introduce a two-step approach for stereo-based 3D object detection. We first convert the estimated depth map from stereo or monocular imagery into a 3D point cloud, which we refer to as *pseudo-LiDAR* as it mimics the LiDAR signal. We then take advantage of existing LiDAR-based 3D object detection pipelines [16, 23], which we train directly on the pseudo-LiDAR representation. By changing the 3D depth representation to pseudo-LiDAR we obtain an unprecedented increase in accuracy of image-based 3D object detection algorithms. Specifically, on the KITTI benchmark with IoU (intersection-over-union) at 0.7 for "moderately hard" car instances — the metric used in the official leaderboard — we achieve a 45.3% 3D AP on the validation set: almost a 350% improvement over the previous state-of-the-art image-based approach. Furthermore, we halve the gap between stereo-based and LiDAR-based systems.

We evaluate multiple combinations of stereo depth estimation and 3D object detection algorithms and arrive at remarkably consistent results. This suggests that the gains we observe are because of the *pseudo-LiDAR* representation and are *less dependent* on innovations in 3D object detection architectures or depth estimation techniques.

In sum, the contributions of the paper are two-fold. First, we show empirically that a major cause for the performance gap between stereo-based and LiDAR-based 3D object detection is not the quality of the estimated depth but its *representation*. Second, we propose *pseudo-LiDAR* as a new recommended representation of estimated depth for 3D object detection and show that it leads to state-of-the-art stereo-based 3D object detection, effectively *tripling* prior art. Our results point towards the possibility of using stereo cameras in self-driving cars — potentially yielding substantial cost reductions and/or safety improvements.

2. Related Work

LiDAR-based 3D object detection. Our work is inspired by the recent progress in 3D vision and LiDAR-based 3D object detection. Many recent techniques use the fact that LiDAR is naturally represented as 3D point clouds. For example, frustum PointNet [23] applies PointNet [24] to each frustum proposal from a 2D object detection network. MV3D [7] projects LiDAR points into both bird-eye view (BEV) and frontal view to obtain multi-view features. VoxelNet [34] encodes 3D points into voxels and extracts features by 3D convolutions. UberATG-ContFuse [17], one of the leading algorithms on the KITTI benchmark [12], performs continuous convolutions [27] to fuse visual and BEV LiDAR features. All these algorithms assume that the precise 3D point coordinates are given. The main challenge there is thus on predicting point labels or drawing bounding boxes in 3D to locate objects.

Stereo- and monocular-based depth estimation. A key ingredient for image-based 3D object detection methods is a reliable depth estimation approach to replace LiDAR. These can be obtained through monocular [10, 13] or stereo vision [3, 19]. The accuracy of these systems has increased dramatically since early work on monocular depth estimation [8, 15, 26]. Recent algorithms like DORN [10] combine multi-scale features with ordinal regression to predict pixel depth with remarkably low errors. For stereo vision, PSMNet [3] applies Siamese networks for disparity estimation, followed by 3D convolutions for refinement, resulting in an outlier rate less than 2%. Recent work has made these methods mode efficient [28], enabling accurate disparity estimation to run at 30 FPS on mobile devices.

Image-based 3D object detection. The rapid progress on stereo and monocular depth estimation suggests that they could be used as a substitute for LiDAR in image-based 3D object detection algorithms. Existing algorithms of this flavor are largely built upon 2D object detection [25], imposing extra geometric constraints [2, 4, 21, 29] to create 3D proposals. [5, 6, 22, 30] apply stereo-based depth estimation to obtain the true 3D coordinates of each pixel. These 3D coordinates are either entered as additional input channels into a 2D detection pipeline, or used to extract hand-crafted features. Although these methods have made remarkable progress, the state-of-the-art for 3D object detection performance lags behind LiDAR-based methods. As we discuss in Section 3, this might be because of the depth representation used by these methods.

3. Approach

Despite the many advantages of image-based 3D object recognition, there remains a glaring gap between the state-of-the-art detection rates of image and LiDAR-based ap-

Ex.-1029 - ADB Safegate et al.
ADB Safegate et al. v. Honeywell International Inc. – PGR2023-00053
Page 3

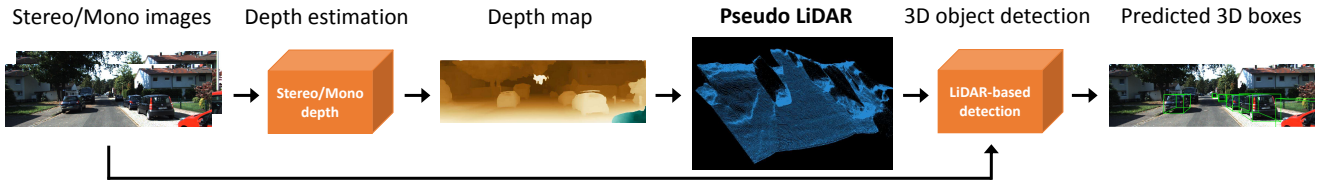

Figure 2: **The proposed pipeline for image-based 3D object detection.** Given stereo or monocular images, we first predict the depth map, followed by back-projecting it into a 3D point cloud in the LiDAR coordinate system. We refer to this representation as *pseudo-LiDAR*, and process it exactly like LiDAR — any LiDAR-based detection algorithms can be applied.

proaches (see Table 1 in Section 4.3). It is tempting to attribute this gap to the obvious physical differences and its implications between LiDAR and camera technology. For example, the error of stereo-based 3D depth estimation grows quadratically with the depth of an object, whereas for Time-of-Flight (ToF) approaches, such as LiDAR, this relationship is approximately linear.

Although some of these physical differences do likely contribute to the accuracy gap, in this paper we claim that a large portion of the discrepancy can be explained by the data representation rather than its quality or underlying physical properties associated with data collection.

In fact, recent algorithms for stereo depth estimation can generate surprisingly accurate depth maps [3] (see figure 1). Our approach to "close the gap" is therefore to carefully remove the differences between the two data modalities and align the two recognition pipelines as much as possible. To this end, we propose a two-step approach by first estimating the dense pixel depth from stereo (or even monocular) imagery and then back-projecting pixels into a 3D point cloud. By viewing this representation as *pseudo-LiDAR* signal, we can then apply *any* existing LiDAR-based 3D object detection algorithm. Fig. 2 depicts our pipeline.

Depth estimation. Our approach is agnostic to different depth estimation algorithms. We primarily work with stereo disparity estimation algorithms [3, 19], although our approach can easily use monocular depth estimation methods.

A stereo disparity estimation algorithm takes a pair of left-right images I_l and I_r as input, captured from a pair of cameras with a horizontal offset (i.e., baseline) b, and outputs a disparity map Y of the same size as either one of the two input images. Without loss of generality, we assume the depth estimation algorithm treats the left image, I_l, as reference and records in Y the horizontal disparity to I_r for each pixel. Together with the horizontal focal length f_U of the left camera, we can derive the depth map D via the following transform,

$$D(u,v) = \frac{f_U \times b}{Y(u,v)}. \tag{1}$$

Pseudo-LiDAR generation. Instead of incorporating the depth D as multiple additional channels to the RGB images, as is typically done [30], we can derive the 3D location (x, y, z) of each pixel (u, v), in the left camera's coordinate system, as follows,

$$\text{(depth)} \qquad z = D(u,v) \tag{2}$$

$$\text{(width)} \qquad x = \frac{(u - c_U) \times z}{f_U} \tag{3}$$

$$\text{(height)} \qquad y = \frac{(v - c_V) \times z}{f_V}, \tag{4}$$

where (c_U, c_V) is the pixel location corresponding to the camera center and f_V is the vertical focal length.

By back-projecting all the pixels into 3D coordinates, we arrive at a 3D point cloud $\{(x^{(n)}, y^{(n)}, z^{(n)})\}_{n=1}^{N}$, where N is the pixel count. Such a point cloud can be transformed into any cyclopean coordinate frame given a reference viewpoint and viewing direction. We refer to the resulting point cloud as *pseudo-LiDAR* signal.

LiDAR vs. pseudo-LiDAR. In order to be maximally compatible with existing LiDAR detection pipelines we apply a few additional post-processing steps on the pseudo-LiDAR data. Since real LiDAR signals only reside in a certain range of heights, we disregard pseudo-LiDAR points beyond that range. For instance, on the KITTI benchmark, following [33], we remove all points higher than 1m above the fictitious LiDAR source (located on top of the autonomous vehicle). As most objects of interest (e.g., cars and pedestrians) do not exceed this height range there is little information loss. In addition to depth, LiDAR also returns the reflectance of any measured pixel (within [0,1]). As we have no such information, we simply set the reflectance to 1.0 for every pseudo-LiDAR points.

Fig 1 depicts the ground-truth LiDAR and the pseudo-LiDAR points for the same scene from the KITTI dataset [11, 12]. The depth estimate was obtained with the pyramid stereo matching network (PSMNet) [3]. Surprisingly, the pseudo-LiDAR points (blue) align remarkably well to true LiDAR points (yellow), in contrast to the common belief that low precision image-based depth is the main cause of inferior 3D object detection. We note that a LiDAR can capture $> 100,000$ points for a scene, which is of the same order as the pixel count. Nevertheless, LiDAR points are distributed along a few (typically 64 or 128) horizontal

Ex.-1029 - ADB Safegate et al.
ADB Safegate et al. v. Honeywell International Inc. – PGR2023-00053
Page 4

beams, only sparsely occupying the 3D space.

3D object detection. With the estimated pseudo-LiDAR points, we can apply *any* existing LiDAR-based 3D object detectors for autonomous driving. In this work, we consider those based on multimodal information (i.e., monocular images + LiDAR), as it is only natural to incorporate the original visual information together with the pseudo-LiDAR data. Specifically, we experiment on AVOD [16] and frustum PointNet [23], the two top ranked algorithms with open-sourced code on the KITTI benchmark. In general, we distinguish between two different setups:

a) In the first setup we treat the pseudo-LiDAR information as a *3D point cloud*. Here, we use frustum PointNet [23], which projects 2D object detections [18] into a frustum in 3D, and then applies PointNet [24] to extract point-set features at each 3D frustum.

b) In the second setup we view the pseudo-LiDAR information from a *Bird's Eye View (BEV)*. In particular, the 3D information is converted into a 2D image from the top-down view: width and depth become the spatial dimensions, and height is recorded in the channels. AVOD connects visual features and BEV LiDAR features to 3D box proposals and then fuses both to perform box classification and regression.

Data representation matters. Although pseudo-LiDAR conveys the same information as a depth map, we claim that it is much better suited for 3D object detection pipelines that are based on deep convolutional networks. To see this, consider the core module of the convolutional network: 2D convolutions. A convolutional network operating on images or depth maps performs a sequence of 2D convolutions on the image/depth map. Although the filters of the convolution can be learned, the central assumption is two-fold: (a) local neighborhoods in the image have meaning, and the network should look at local patches, and (b) all neighborhoods can be operated upon in an identical manner.

These are but imperfect assumptions. First, local patches on 2D images are only coherent physically if they are entirely contained in a single object. If they straddle object boundaries, then two pixels can be co-located next to each other in the depth map, yet can be very far away in 3D space. Second, objects that occur at multiple depths project to *different scales* in the depth map. A similarly sized patch might capture just a side-view mirror of a nearby car or the entire body of a far-away car. Existing 2D object detection approaches struggle with this breakdown of assumptions and have to design novel techniques such as feature pyramids [18] to deal with this challenge.

In contrast, 3D convolutions on point clouds or 2D convolutions in the bird's-eye view slices operate on pixels that are *physically* close together (although the latter do pull together pixels from different heights, the physics of the world implies that pixels at different heights at a particular spatial location usually do belong to the same object). In addition, both far-away objects and nearby objects are treated exactly the same way. These operations are thus inherently more physically meaningful and hence should lead to better learning and more accurate models.

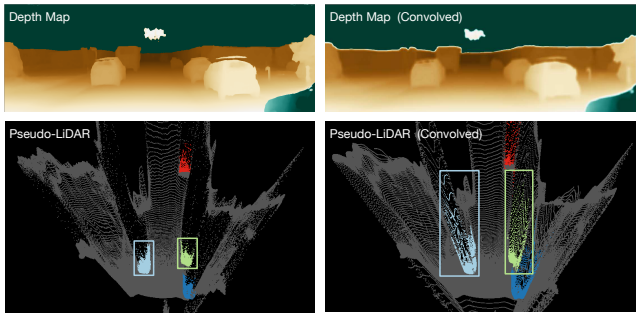

Figure 3: We apply a single 2D convolution with a uniform kernel to the frontal view depth map (top-left). The resulting depth map (top-right), after back-projected into pseudo-LiDAR and displayed from the bird's-eye view (bottom-right), reveals a large depth distortion in comparison to the original pseudo-LiDAR representation (bottom-left), especially for far-away objects. We mark points of each car instance by a color. The boxes are super-imposed and contain all points of the green and cyan cars respectively.

To illustrate this point further, in Fig. 3 we conduct a simple experiment. In the left column, we show the original depth-map and the pseudo-LiDAR representation of an image scene. The four cars in the scene are highlighted in color. We then perform a single 11×11 convolution with a box filter on the depth-map (top right), which matches the receptive field of 5 layers of 3×3 convolutions. We then convert the resulting (blurred) depth-map into a pseudo-LiDAR representation (bottom right). From the figure, it becomes evident that this new pseudo-LiDAR representation suffers substantially from the effects of the blurring. The cars are stretched out far beyond their actual physical proportions making it essentially impossible to locate them precisely. For better visualization, we added rectangles that contain all the points of the green and cyan cars. After the convolution, both bounding boxes capture highly erroneous areas. Of course, the 2D convolutional network will learn to use more intelligent filters than box filters, but this example goes to show how some operations the convolutional network might perform could border on the absurd.

4. Experiments

We evaluate 3D-object detection with and without pseudo-LiDAR across different settings with varying approaches for depth estimation and object detection.

Ex.-1029 - ADB Safegate et al.
ADB Safegate et al. v. Honeywell International Inc. – PGR2023-00053
Page 5

Throughout, we will highlight results obtained with *pseudo-LiDAR* in blue and those with actual LiDAR in gray.

4.1. Setup

Dataset. We evaluate our approach on the KITTI object detection benchmark [11, 12], which contains 7,481 images for training and 7,518 images for testing. We follow the same training and validation splits as suggested by Chen et al. [5], containing 3,712 and 3,769 images respectively. For each image, KITTI provides the corresponding Velodyne LiDAR point cloud, right image for stereo information, and camera calibration matrices.

Metric. We focus on 3D and bird's-eye-view (BEV)[1] object detection and report the results on the *validation set*. Specifically, we focus on the "car" category, following [7, 31]. We follow the benchmark and prior work and report average precision (AP) with the IoU thresholds at 0.5 and 0.7. We denote AP for the 3D and BEV tasks by AP_{3D} and AP_{BEV}, respectively. Note that the benchmark divides each category into three cases — easy, moderate, and hard — according to the bounding box height and occlusion/truncation level. In general, the easy case corresponds to cars within 30 meters of the ego-car distance [33].

Baselines. We compare to MONO3D [4], 3DOP [5], and MLF [30]. The first is monocular and the second is stereo-based. MLF [30] reports results with both monocular [13] and stereo disparity [19], which we denote as MLF-MONO and MLF-STEREO, respectively.

4.2. Details of our approach

Stereo disparity estimation. We apply PSMNET [3], DISPNET [19], and SPS-STEREO [32] to estimate dense disparity. The first two approaches are learning-based and we use the released models, which are pre-trained on the Scene Flow dataset [19], with over 30,000 pairs of synthetic images and dense disparity maps, and fine-tuned on the 200 training pairs of KITTI stereo 2015 benchmark [12, 20]. We note that, MLF-STEREO [30] also uses the released DISPNET model. The third approach, SPS-STEREO [32], is non-learning-based and has been used in [5, 6, 22].

DISPNET has two versions, without and with correlations layers. We test both and denote them as DISPNET-S and DISPNET-C, respectively.

While performing these experiments, we found that the 200 training images of KITTI stereo 2015 overlap with the validation images of KITTI object detection. That is, the released PSMNET and DISPNET models actually used some validation images of detection. We therefore train a version of PSMNET using Scene Flow followed by finetuning on the 3,712 training images of detection, instead of the 200 KITTI stereo images. We obtain *pseudo disparity* ground truth by projecting the corresponding LiDAR points into the 2D image space. We denote this version PSMNET$\star$. Details are included in the Supplementary Material.

The results with PSMNET$\star$ in Table 3 (fined-tuned on 3,712 training data) are in fact better than PSMNET (fine-tuned on KITTI stereo 2015). We attribute the improved accuracy of PSMNET$\star$ on the fact that it is trained on a larger training set. Nevertheless, future work on 3D object detection using stereo must be aware of this overlap.

Monocular depth estimation. We use the state-of-the-art monocular depth estimator DORN [10], which is trained by the authors on 23,488 KITTI images. We note that some of these images may overlap with our validation data for detection. Nevertheless, we decided to still include these results and believe they could serve as an upper bound for monocular-based 3D object detection. Future work, however, must be aware of this overlap.

Pseudo-LiDAR generation. We back-project the estimated depth map into 3D points in the Velodyne LiDAR's coordinate system using the provided calibration matrices. We disregard points with heights larger than 1 in the system.

3D Object detection. We consider two algorithms: Frustum PointNet (F-POINTNET) [23] and AVOD [16]. More specifically, we apply F-POINTNET-v1 and AVOD-FPN. Both of them use information from LiDAR and monocular images. We train both models on the 3,712 training data from scratch by replacing the LiDAR points with pseudo-LiDAR data generated from stereo disparity estimation. We use the hyper-parameters provided in the released code.

We note that AVOD takes image-specific ground planes as inputs. The authors provide ground-truth planes for training and validation images, but do not provide the procedure to obtain them (for novel images). We therefore fit the ground plane parameters with a straight-forward application of RANSAC [9] to our pseudo-LiDAR points that fall into a certain range of road height, during evaluation. Details are included in the Supplementary Material.

4.3. Experimental results

We summarize the main results in Table 1. We organize methods according to the input signals for performing detection. Our stereo approaches based on pseudo-LiDAR significantly outperform all image-based alternatives by a large margin. At IoU = 0.7 (moderate) — the metric used to rank algorithms on the KITTI leaderboard — we achieve *double* the performance of the previous state of the art. We also observe that pseudo-LiDAR is applicable and highly beneficial to two 3D object detection algorithms with very different architectures, suggesting its wide compatibility.

One interesting comparison is between approaches using pseudo-LiDAR with monocular depth (DORN) and stereo depth (PSMNET$\star$). While DORN has been trained with

[1] The BEV detection task is also called 3D localization.

Ex.-1029 - ADB Safegate et al.
ADB Safegate et al. v. Honeywell International Inc. – PGR2023-00053
Page 6

Table 1: 3D object detection results on the KITTI validation set. We report AP_{BEV} / AP_{3D} (in %) of the **car** category, corresponding to average precision of the bird's-eye view and 3D object box detection. Mono stands for monocular. Our methods with *pseudo-LiDAR* estimated by PSMNET⋆ [3] (stereo) or DORN [10] (monocular) are in blue. Methods with LiDAR are in gray. Best viewed in color.

		IoU = 0.5			IoU = 0.7		
Detection algorithm	Input signal	Easy	Moderate	Hard	Easy	Moderate	Hard
MONO3D [4]	Mono	30.5 / 25.2	22.4 / 18.2	19.2 / 15.5	5.2 / 2.5	5.2 / 2.3	4.1 / 2.3
MLF-MONO [30]	Mono	55.0 / 47.9	36.7 / 29.5	31.3 / 26.4	22.0 / 10.5	13.6 / 5.7	11.6 / 5.4
AVOD	Mono	61.2 / 57.0	45.4 / 42.8	38.3 / 36.3	33.7 / 19.5	24.6 / 17.2	20.1 / 16.2
F-POINTNET	Mono	70.8 / 66.3	49.4 / 42.3	42.7 / 38.5	40.6 / 28.2	26.3 / 18.5	22.9 / 16.4
3DOP [5]	Stereo	55.0 / 46.0	41.3 / 34.6	34.6 / 30.1	12.6 / 6.6	9.5 / 5.1	7.6 / 4.1
MLF-STEREO [30]	Stereo	-	53.7 / 47.4	-	-	19.5 / 9.8	-
AVOD	Stereo	89.0 / 88.5	77.5 / 76.4	68.7 / 61.2	74.9 / 61.9	56.8 / 45.3	49.0 / 39.0
F-POINTNET	Stereo	89.8 / 89.5	77.6 / 75.5	68.2 / 66.3	72.8 / 59.4	51.8 / 39.8	44.0 / 33.5
AVOD [16]	LiDAR + Mono	90.5 / 90.5	89.4 / 89.2	88.5 / 88.2	89.4 / 82.8	86.5 / 73.5	79.3 / 67.1
F-POINTNET [23]	LiDAR + Mono	96.2 / 96.1	89.7 / 89.3	86.8 / 86.2	88.1 / 82.6	82.2 / 68.8	74.0 / 62.0

Table 2: Comparison between frontal and *pseudo-LiDAR* representations. AVOD projects the pseudo-LiDAR representation into the bird-eye's view (BEV). We report AP_{BEV} / AP_{3D} (in %) of the **moderate car** category at IoU = 0.7. The best result of each column is in bold font. The results indicate strongly that the data representation is the key contributor to the accuracy gap.

Detection	Disparity	Representation	AP_{BEV} / AP_{3D}
MLF [30]	DISPNET	Frontal	19.5 / 9.8
AVOD	DISPNET-S	Pseudo-LiDAR	36.3 / 27.0
AVOD	DISPNET-C	Pseudo-LiDAR	36.5 / 26.2
AVOD	PSMNET⋆	Frontal	11.9 / 6.6
AVOD	PSMNET⋆	Pseudo-LiDAR	**56.8 / 45.3**

almost ten times more images than PSMNET⋆ (and some of them overlap with the validation data), the results with PSMNET⋆ dominate. This suggests that stereo-based detection is a promising direction to move in, especially considering the increasing affordability of stereo cameras.

In the following section, we discuss key observations and conduct a series of experiments to analyze the performance gain through pseudo-LiDAR with stereo disparity.

Impact of data representation. When comparing our results using DISPNET-S or DISPNET-C to MLF-STEREO [30] (which also uses DISPNET as the underlying stereo engine), we observe a large performance gap (see Table. 2). Specifically, at IoU= 0.7, we outperform MLF-STEREO by at least 16% on AP_{BEV} and 16% on AP_{3D}. The later is equivalent to a 160% relative improvement. We attribute this improvement to the way in which we represent the resulting depth information. We note that both our approach and MLF-STEREO [30] first back-project pixel depths into 3D point coordinates. MLF-STEREO construes the 3D coordinates of each pixel as additional feature maps in the frontal view. These maps are then concatenated with RGB channels as the input to a modified 2D object detection pipeline based on Faster-RCNN [25]. As we point out earlier, this has two problems. Firstly, distant objects become smaller, and detecting small objects is a known hard problem [18]. Secondly, while performing local computations like convolutions or ROI pooling along height and width of an image makes sense to 2D object detection, it will operate on 2D pixel neighborhoods with pixels that are far apart in 3D, making the precise localization of 3D objects much harder (cf. Fig. 3).

By contrast, our approach treats these coordinates as pseudo-LiDAR signals and applies PointNet [24] (in F-POINTNET) or use a convolutional network on the BEV projection (in AVOD). This introduces invariance to depth, since far-away objects are no longer smaller. Furthermore, convolutions and pooling operations in these representations put together points that are physically nearby.

To further control for other differences between MLF-STEREO and our method we ablate our approach to use the same frontal depth representation used by MLF-STEREO. AVOD fuses information of the frontal images with BEV LiDAR features. We modify the algorithm, following [6, 30], to generate five frontal-view feature maps, including 3D pixel locations, disparity, and Euclidean distance to the camera. We concatenate them with the RGB channels while disregarding the BEV branch in AVOD, making it fully dependent on the frontal-view branch. (We make no additional architecture changes.) The results in Table 2 reveal a staggering gap between frontal and pseudo-LiDAR results. We found that the frontal approach struggles with inferring object depth, even when the five extra maps have provided sufficient 3D information. Again, this might be because 2d convolutions put together pixels from far away depths, making accurate localization difficult. This experiment suggests that the chief source of the accuracy improvement is indeed the *pseudo-LiDAR* representation.

Impact of stereo disparity estimation accuracy. We compare PSMNET [3] and DISPNET [19] on pseudo-LiDAR-based detection accuracies. On the leaderboard of KITTI

Ex.-1029 - ADB Safegate et al.
ADB Safegate et al. v. Honeywell International Inc. – PGR2023-00053
Page 7

Table 3: Comparison of different combinations of stereo disparity and 3D object detection algorithms, using *pseudo-LiDAR*. We report AP_{BEV} / AP_{3D} (in %) of the **moderate car** category at IoU = 0.7. The best result of each column is in bold font.

Disparity	Detection algorithm	
	AVOD	F-POINTNET
DISPNET-S	36.3 / 27.0	31.9 / 23.5
DISPNET-C	36.5 / 26.2	37.4 / 29.2
PSMNET	39.2 / 27.4	33.7 / 26.7
PSMNET⋆	**56.8 / 45.3**	**51.8 / 39.8**

stereo 2015, PSMNET achieves 1.86% disparity error, which far outperforms the error of 4.32% by DISPNET-C.

As shown in Table 3, the accuracy of disparity estimation does not necessarily correlate with the accuracy of object detection. F-POINTNET with DISPNET-C even outperforms F-POINTNET with PSMNET. This is likely due to two reasons. First, the disparity accuracy may not reflect the depth accuracy: the same disparity error (on a pixel) can lead to drastically different depth errors dependent on the pixel's true depth, according to Eq. (1). Second, different detection algorithms process the 3D points differently: AVOD quantizes points into voxels, while F-POINTNET directly processes them and may be vulnerable to noise.

By far the most accurate detection results are obtained by PSMNET⋆, which we trained from scratch on our own KITTI training set. These results seem to suggest that significant further improvements may be possible through end-to-end training of the whole pipeline.

We provide results using SPS-STEREO [32] and further analysis on depth estimation in the Supplementary Material.

Comparison to LiDAR information. Our approach significantly improves stereo-based detection accuracies. A key remaining question is, how close the pseudo-LiDAR detection results are to those based on real LiDAR signal. In Table 1, we further compare to AVOD and F-POINTNET when actual LiDAR signal is available. For fair comparison, we retrain both models. For the easy cases with IoU $= 0.5$, our stereo-based approach performs very well, only slightly worse than the corresponding LiDAR-based version. However, as the instances become harder (e.g., for cars that are far away), the performance gaps resurfaces — although not nearly as pronounced as without pseudo-LiDAR. We also see a larger gap when moving to IoU $= 0.7$. These results are not surprising, since stereo algorithms are known to have larger depth errors for far-away objects, and a stricter metric requires higher depth precision. Both observations emphasize the need for accurate depth estimation, especially for far-away distances, to bridge the gap further. A key limitation of our results may be the low resolution of the 0.4 MegaPixel images, which cause far away objects to only consist of a few pixels.

Table 4: 3D object detection on the **pedestrian** and **cyclist** categories on the validation set. We report AP_{BEV} / AP_{3D} at IoU = 0.5 (the standard metric) and compare F-POINTNET with *pseudo-LiDAR* estimated by PSMNET⋆ (in blue) and LiDAR (in gray).

Input signal	Easy	Moderate	Hard
Pedestrian			
Stereo	41.3 / 33.8	34.9 / 27.4	30.1 / 24.0
LiDAR + Mono	69.7 / 64.7	60.6 / 56.5	53.4 / 49.9
Cyclist			
Stereo	47.6 / 41.3	29.9 / 25.2	27.0 / 24.9
LiDAR + Mono	70.3 / 66.6	55.0 / 50.9	52.0 / 46.6

Table 5: 3D object detection results on the **car** category on the *test* set. We compare *pseudo-LiDAR* with PSMNET⋆ (in blue) and LiDAR (in gray). We report AP_{BEV} / AP_{3D} at IoU = 0.7. †: Results on the KITTI leaderboard.

Input signal	Easy	Moderate	Hard
AVOD			
Stereo	66.8 / 55.4	47.2 / 37.2	40.3 / 31.4
†LiDAR + Mono	88.5 / 81.9	83.8 / 71.9	77.9 / 66.4
F-POINTNET			
Stereo	55.0 / 39.7	38.7 / 26.7	32.9 / 22.3
†LiDAR + Mono	88.7 / 81.2	84.0 / 70.4	75.3 / 62.2

Pedestrian and cyclist detection. We also present results on 3D pedestrian and cyclist detection. These are much more challenging tasks than car detection due to the small sizes of the objects, even given LiDAR signals. At an IoU threshold of 0.5, both AP_{BEV} and AP_{3D} of pedestrians and cyclists are much lower than that of cars at IoU 0.7 [23]. We also notice that none of the prior work on image-based methods report results in this category.

Table 4 shows our results with F-POINTNET and compares to those with LiDAR, on the validation set. Compared to the car category (cf. Table 1), the performance gap is significant. We also observe a similar trend that the gap becomes larger when moving to the hard cases. Nevertheless, our approach has set a solid starting point for image-based pedestrian and cyclist detection for future work.

4.4. Results on the test set

We report our results on the car category on the test set in Table 5. We see a similar gap between pseudo-LiDAR and LiDAR as on the validation set, suggesting that our approach does not simply over-fit to the "validation data." *We also note that, at the time we submit the paper, we are at the first place among all the image-based algorithms on the KITTI leaderboard.* Details and results on the pedestrian and cyclist categories are in the Supplementary Material.

4.5. Visualization

We further visualize the prediction results on validation images in Fig. 4. We compare LiDAR (left), stereo

Ex.-1029 - ADB Safegate et al.
ADB Safegate et al. v. Honeywell International Inc. – PGR2023-00053
Page 8

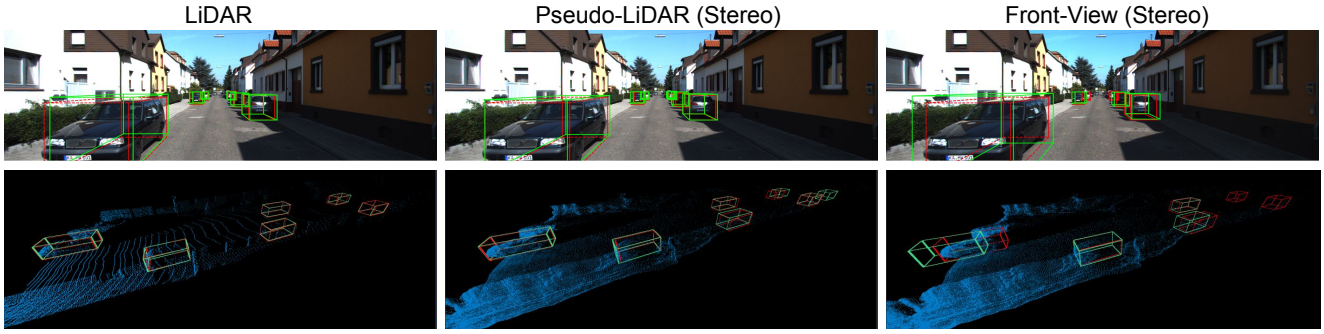

Figure 4: **Qualitative comparison.** We compare AVOD with LiDAR, pseudo-LiDAR, and frontal-view (stereo). Ground-truth boxes are in red, predicted boxes in green; the observer in the pseudo-LiDAR plots (bottom row) is on the very left side looking to the right. The frontal-view approach (*right*) even miscalculates the depths of nearby objects and misses far-away objects entirely. Best viewed in color.

pseudo-LiDAR (middle), and frontal stereo (right). We used PSMNET⋆ to obtain the stereo depth maps. LiDAR and pseudo-LiDAR lead to highly accurate predictions, especially for the nearby objects. However, pseudo-LiDAR fails to detect far-away objects precisely due to inaccurate depth estimates. On the other hand, the frontal-view-based approach makes extremely inaccurate predictions, even for nearby objects. This corroborates the quantitative results we observed in Table 2. We provide additional qualitative results and failure cases in the Supplementary Material.

5. Discussion and Conclusion

Sometimes, it is the simple discoveries that make the biggest differences. In this paper we have shown that a key component to closing the gap between image- and LiDAR-based 3D object detection may be simply the representation of the 3D information. It may be fair to consider these results as the correction of a systemic inefficiency rather than a novel algorithm — however, that does not diminish its importance. Our findings are consistent with our understanding of convolutional neural networks and substantiated through empirical results. In fact, the improvements we obtain from this correction are unprecedentedly high and affect all methods alike. With this quantum leap it is plausible that image-based 3D object detection for autonomous vehicle will become a reality in the near future. The implications of such a prospect are enormous. Currently, the LiDAR hardware is arguably the most expensive additional component required for autonomous driving. Without it, the additional hardware cost for autonomous driving becomes relatively minor. Further, image-based object detection would also be beneficial even in the presence of LiDAR equipment. One could imagine a scenario where the LiDAR data is used to continuously train and fine-tune an image-based classifier. In case of our sensor outage, the image-based classifier could likely function as a very reliable backup. Similarly, one could imagine a setting where high-end cars are shipped with LiDAR hardware and continuously train the image-based classifiers that are used in cheaper models.

Future work. There are multiple immediate directions along which our results could be improved in future work: First, higher resolution stereo images would likely significantly improve the accuracy for faraway objects. Our results were obtained with 0.4 megapixels — a far cry from the state-of-the-art camera technology. Second, in this paper we did not focus on real-time image processing and the classification of all objects in one image takes on the order of 1s. However, it is likely possible to improve these speeds by several orders of magnitude. Recent improvements on real-time multi-resolution depth estimation [28] show that an effective way to speed up depth estimation is to first compute a depth map at low resolution and then incorporate high-resolution to refine the previous result. The conversion from a depth map to pseudo-LiDAR is very fast and it should be possible to drastically speed up the detection pipeline through e.g. model distillation [1] or anytime prediction [14]. Finally, it is likely that future work could improve the state-of-the-art in 3D object detection through sensor fusion of LiDAR and pseudo-LiDAR. Pseudo-LiDAR has the advantage that its signal is much denser than LiDAR and the two data modalities could have complementary strengths. We hope that our findings will cause a revival of image-based 3D object recognition and our progress will motivate the computer vision community to fully close the image/LiDAR gap in the near future.

Acknowledgments

This research is supported in part by grants from the National Science Foundation (III-1618134, III-1526012, IIS-1149882, IIS-1724282, TRIPODS-1740822), the Office of Naval Research DOD (N00014-17-1-2175), and the Bill and Melinda Gates Foundation. We are thankful for generous support by Zillow and SAP America Inc. We thank Gao Huang for helpful discussion.

Ex.-1029 - ADB Safegate et al.
ADB Safegate et al. v. Honeywell International Inc. – PGR2023-00053
Page 9

References

[1] C. Bucilu, R. Caruana, and A. Niculescu-Mizil. Model compression. In *SIGKDD*, 2006. 8

[2] F. Chabot, M. Chaouch, J. Rabarisoa, C. Teulière, and T. Chateau. Deep manta: A coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image. In *CVPR*, 2017. 2

[3] J.-R. Chang and Y.-S. Chen. Pyramid stereo matching network. In *CVPR*, 2018. 1, 2, 3, 5, 6

[4] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun. Monocular 3d object detection for autonomous driving. In *CVPR*, 2016. 2, 5, 6

[5] X. Chen, K. Kundu, Y. Zhu, A. G. Berneshawi, H. Ma, S. Fidler, and R. Urtasun. 3d object proposals for accurate object class detection. In *NIPS*, 2015. 1, 2, 5, 6

[6] X. Chen, K. Kundu, Y. Zhu, H. Ma, S. Fidler, and R. Urtasun. 3d object proposals using stereo imagery for accurate object class detection. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1259–1272, 2018. 1, 2, 5, 6

[7] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia. Multi-view 3d object detection network for autonomous driving. In *CVPR*, 2017. 2, 5

[8] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014. 2

[9] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 5

[10] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, pages 2002–2011, 2018. 2, 5, 6

[11] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1, 3, 5

[12] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 1, 2, 3, 5

[13] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017. 1, 2, 5

[14] G. Huang, D. Chen, T. Li, F. Wu, L. van der Maaten, and K. Q. Weinberger. Multi-scale dense convolutional networks for efficient prediction. *CoRR, abs/1703.09844*, 2, 2017. 8

[15] K. Karsch, C. Liu, and S. B. Kang. Depth extraction from video using non-parametric sampling. In *ECCV*, 2012. 2

[16] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. Waslander. Joint 3d proposal generation and object detection from view aggregation. In *IROS*, 2018. 2, 4, 5, 6

[17] M. Liang, B. Yang, S. Wang, and R. Urtasun. Deep continuous fusion for multi-sensor 3d object detection. In *ECCV*, 2018. 1, 2

[18] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. In *CVPR*, volume 1, page 4, 2017. 4, 6

[19] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016. 1, 2, 3, 5, 6

[20] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *CVPR*, 2015. 5

[21] A. Mousavian, D. Anguelov, J. Flynn, and J. Košecká. 3d bounding box estimation using deep learning and geometry. In *CVPR*, 2017. 2

[22] C. C. Pham and J. W. Jeon. Robust object proposals re-ranking for object detection in autonomous driving using convolutional neural networks. *Signal Processing: Image Communication*, 53:110–122, 2017. 1, 2, 5

[23] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *CVPR*, 2018. 2, 4, 5, 6, 7

[24] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017. 2, 4, 6

[25] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 2, 6

[26] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840, 2009. 2

[27] S. Wang, S. Suo, W.-C. M. A. Pokrovsky, and R. Urtasun. Deep parametric continuous convolutional neural networks. In *CVPR*, 2018. 2

[28] Y. Wang, Z. Lai, G. Huang, B. H. Wang, L. van der Maaten, M. Campbell, and K. Q. Weinberger. Anytime stereo image depth estimation on mobile devices. *arXiv preprint arXiv:1810.11408*, 2018. 2, 8

[29] Y. Xiang, W. Choi, Y. Lin, and S. Savarese. Subcategory-aware convolutional neural networks for object proposals and detection. In *WACV*, 2017. 2

[30] B. Xu and Z. Chen. Multi-level fusion based 3d object detection from monocular images. In *CVPR*, 2018. 1, 2, 3, 5, 6

[31] D. Xu, D. Anguelov, and A. Jain. Pointfusion: Deep sensor fusion for 3d bounding box estimation. In *CVPR*, 2018. 5

[32] K. Yamaguchi, D. McAllester, and R. Urtasun. Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In *ECCV*, 2014. 1, 5, 7

[33] B. Yang, W. Luo, and R. Urtasun. Pixor: Real-time 3d object detection from point clouds. In *CVPR*, 2018. 2, 3, 5

[34] Y. Zhou and O. Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *CVPR*, 2018. 2