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/ACCESS.2021.3063692, IEEE Access

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 000000000000

ISETAuto: Detecting vehicles with depth and radiance information

ZHENYI LIU[1], JOYCE FARRELL[2] AND BRIAN WANDELL[2]
[1]State Key Laboratory of Automotive Simulation and Control, Jilin University (e-mail: zhenyiliu27@gmail.com)
[2]Stanford University (e-mail: jefarrel, wandell@stanford.edu)

Corresponding author: Zhenyi Liu (e-mail: zhenyiliu27@gmail.com)

Supported by Jilin University. We thank Boyd Fowler at Omnivision and Sergio Goma at Qualcomm for drawing our attention to prior work on RGB-D sensor technology.

ABSTRACT Autonomous driving applications use two types of sensor systems to detect vehicles - depth sensing LiDAR and radiance sensing cameras. We compare the performance (average precision) of a ResNet for vehicle detection in complex, daytime, driving scenes when the input is a depth map [D = d(x,y)], a radiance image [L = r(x,y)], or both [D,L]. (1) When the spatial sampling resolution of the depth map and radiance image are both equal to typical camera resolutions, a ResNet detects vehicles at higher average precision from depth than radiance. (2) When the spatial sampling of the depth map matches the range of current LiDAR devices, the average precision is higher for radiance than depth. (3) A hybrid system that combines depth and radiance has higher average precision than systems using depth or radiance alone. We confirm these observations in both simulation and real-world data. We explain the advantage of combining depth and radiance by noting that the two types of information have complementary weaknesses. The radiance data are limited by dynamic range, motion blur and illumination variation; the LiDAR data have low spatial resolution. The ResNet effectively combines the two data sources to improve vehicle detection.

INDEX TERMS LiDAR, Camera, Sensor fusion, Autonomous driving, Object detection, Convolutional neural network.

I. INTRODUCTION

PEOPLE can detect vehicles using only one eye (monocular), and people with monocular vision are permitted to drive. Moreover, people can accurately recognize objects from 2D images that contain no stereo information. These simple observations raise a practical question: Given the high accuracy of vehicle detection using monocular information or 2D images, how much will explicit depth information improve accuracy?

The reverse formulation of this question is also interesting. Bats, the second largest order of mammals after rodents, include many species that navigate through complex environments using depth sensing [14], [15]. There are several clear advantages to using depth sensing. Under low illumination conditions, such at night and in caves, radiance data are unreliable. Further, depth sensing avoids some of the challenging aspects of radiance measurements, such as non-uniform illumination and high dynamic range. How well can a system perform using only depth information?

This paper assesses the value of explicit depth information such as one might obtain from LiDAR systems in an automotive application. This assessment has practical significance because obtaining and using accurate depth information can be expensive. At present, LiDAR sensors for automotive imaging are a significant part of the cost of an automotive object detection system and has increases the complexity of the system. Before committing to explicit depth measurements from LiDAR, it seems sensible to quantify the benefits.

We refer to the depth information from LiDAR as explicit to distinguish it from the wealth of implicit depth information in monocular images. The implicit depth information is present in the form of object size, occlusion, and texture gradients, and this information is routinely used in human visual perception [6]. The implicit depth information may be discovered by deep networks that detect objects, such as vehicles, from images.

The simulations and analyses of real-world data reveal great value in explicit depth information for detecting vehicles. This does not suggest that radiance data are not important: some critical driving information (road markings, traffic light status) is only available through radiance sensors. Therefore, we investigated a simple network architecture that

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/ACCESS.2021.3063692, IEEE Access

combines radiance and depth information, and we quantify the network's performance comparing radiance, depth and the fused sensor data.

II. RELATED WORK

A number of authors have explored the value of depth information alone or in combination with radiance data. These groups have used a variety of neural network architectures. Some authors input explicit depth and imaging data on independent channels that are processed separately through several network layers [4], [18], [29], [34]. After some processing, the depth and radiance channels are fused. This design makes it possible to use either imaging or non-imaging representations of the depth data, such as point clouds.

This design raises the question of which layer is best for merging the independent channels; one might expect that the answer depends on both the network and the data. Using a YOLOv2 network, [29] explored how variations in the merged layer influenced performance. The portions of their analysis most relevant to our work detected vehicles using data obtained from the KITTI database. The average precision performance using radiance and "moderately difficult" vehicles (77%) was better than depth alone (64%). For their case, performance for the optimal combination was (80%).

Caltagirone et al. [3] used a more complex network architecture involving the cross-fusion of two different input channels for detecting the road surface. They converted explicit depth information from point clouds into depth maps and compared various fusion strategies. Performance on depth maps was slightly higher than performance on the radiance (RGB) image data, though in this task performance was nearly at the ceiling in most cases.

Additional papers examining architectures that integrate depth and RGB data streams have been explored by multiple authors [17], [20], [39]. These authors report that explicit depth information significantly improves object detection performance.

An alternative conceptualization is provided by [37]. These authors focus on the representation of the depth information, showing that a transformation of the input from depth map images into 3D point cloud representations substantially improves 3D vehicle detection, which is more challenging than 2D image location. They report performance vehicle detection on the KITTI dataset improves from 22% to 74%.

III. CONTRIBUTIONS

Our contributions are:

First, we provide an open-source, freely available image systems simulation toolbox that models camera images and LiDAR images in relatively complex 3D automotive scenes. We use the image system simulation to sweep out a much larger range of system designs [26] and create datasets that generalize better than the widely-used KITTI data sets [25].

Second, we provide a novel assessment of the contributions from radiance, depth or their combination on ResNet [19] performance, making specific measurements of the dependence on the spatial sampling of both depth and radiance information. We further quantify the value of explicit depth information either in isolation or combined with radiance data.

Third, we propose a system architecture that uses multimodal (RGD) input to a ResNet. The simplicity, performance, and widespread use of ResNet, combined with the RGD input format, is a practical design for integrating radiance and depth data in applications.

Fourth, we report a close agreement between the simulations and real-world data analyses. Specifically, we validate network performance trained using image systems simulations data with respect to a network trained using real-world data.

IV. METHODS

We generated a collection of complex scenes for vehicle detection using open source software, ISET3d [26]. This software allows users to create scene spectral radiance from pre-built or user-assembled three-dimensional scenes rendered with physically based rendering techniques [30]. Lights and surface reflectances are represented as spectral quantities, and the size, position, and movement of assets are defined in physical units based on a traffic flow simulator [2].

We created scenes by randomly sampling assets (e.g. vehicles, buses, trucks, pedestrians, buildings, roads, trees, etc.) from a collection we maintain on a cloud-based database, Flywheel [1]. The asset positions were defined by models of street scenes and the random movements in the driving simulator. Each scene is unique and the collection is designed to maximize the diversity of the images of daytime driving scenes. In previous work, we quantified how well training on the ISETAuto dataset generalizes to real-world datasets, including KITTI, CityScape, Baidu-Apollo, and Berkeley Deep Drive [25]. In this paper we add new comparisons with a Waymo dataset.

A. SIMULATED CAMERA (RADIANCE) DATASET

We use open-source software, ISETCam [9], to convert the scene spectral radiance to sensor voltages. We simulate an automotive sensor - MT9V024 sensor manufactured by ON Semiconductor- that provides different color filter array options for automotive vision applications, e.g. monochrome, RGB Bayer, and RCCC color filter arrays. The MT9V024 sensor has 2.5 μm pixels with relatively high light sensitivity, high signal-to-noise and a dynamic range of 55 dB.

B. SIMULATED LIDAR (DEPTH) DATASET

We generated scene depth maps by tracing rays from each camera pixel into the 3D world. Rays are traced from each pixel through the principal point of the lens. When the ray reaches the surface of an asset, or the environment map, we record the [x, y, z] value. The depth information is stored for every pixel to form the depth map.

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/ACCESS.2021.3063692, IEEE Access

C. ISETAUTO

We refer to the combination of ISETCam and ISET3d and the automotive assets used here as ISETAuto. We developed ISETAuto because other simulators are not designed for the end-to-end physically accurate scene to sensor simulations [5] [32] [33]. ISETAuto enables one to quantify scene spectral radiance and depth, label objects and materials, model multi-element lenses that generate the sensor irradiance, and specify sensor geometry and electrical properties. These capabilities are described in more detail in prior work [24] [25] [26].

D. REAL-WORLD DATASET (WAYMO)

The Waymo dataset [35] consists of 1150 video sequences (20 sec) of different scenes. It comprises well registered LiDAR and camera data. Waymo provides 5 different camera views (Front, front right, front left, side right, side left). In this paper, we collected images from the front camera as the dataset for network training and evaluation. The original spatial resolution of the images is 1920x1280. To match the vertical field of view of the camera and LiDAR data, we cropped the camera images to 1920x743.

The Waymo LiDAR data have a maximum distance of 76 meters. The RGB images encode radiance over a much larger distance. In the following experiments, we analyzed the results based on the labeled objects up to 76 meters. We selected one of every 100 images in the Waymo video sequences to create a diverse dataset consisting of 3700 images, 3000 images for training, 350 held-out images for testing, and another 350 held-out images for evaluation.

E. OBJECT DETECTION NETWORK & METRICS

We use Mask R-CNN [12] with a ResNet50 as the backbone for vehicle detection. The performance of a ResNet network is higher than other networks we have tested [31]. The simplicity and widespread use of the ResNet model make it an attractive and practical system for integrating RGB and depth data. Mask R-CNN includes a region proposal network (RPN) that specifies different regions in an image where an object might be found; The fully connected layers is used for bounding box classification and regression.

A detection is considered correct when the area of the intersection of the labeled vehicle bounding box with the proposed region is greater than 50% of the area of the union of the proposed region and the bounding of the vehicle (intersection over union, IoU). Combining the hits and false alarms from this measure, we obtain the average precision of the IoU, a metric that is widely used in machine-learning [7]. Unless indicated otherwise, we use the shorthand AP to describe AP@0.5IoU.

We trained all models from scratch. Training was based on 3000 images which were presented to the network with a batch size of 8 images per training step; model weights were updated after each batch. For example, for the case of 40,000 training steps, a total of 320,000 images were presented and the training set of 3000 images was presented about 106 times (epochs). The model was evaluated and the AP values saved at 16 checkpoints. Model performance was evaluated based on 350 images that were not used in training (held out). We tested the model on another 350 held-out images to report average precision results. We trained and evaluated the model for vehicles and pedestrians using 4 Nvidia P100 GPUs. We performed the same training and evaluation strategy for all experiments in this paper.

V. RESULTS

A. EQUATED FOR SPATIAL RESOLUTION, DEPTH IS BETTER THAN RADIANCE FOR VEHICLE DETECTION IN COMPLEX SCENES.

How much information about the presence or absence of a vehicle is contained in high-quality depth measurements? We addressed this question by simulating pixel-wise depth maps from 3000 different driving scenes. We converted these depth maps into linear gray-scale image values (i.e., a depth map). Examples of a simulated RGB image, monochrome image, and depth map are shown in Figure 1. The simulated depth maps have high spatial resolution, beyond what is typically measured by LiDAR and also beyond the accuracy of explicit depth information that can be obtained from even the best stereo algorithms.

We trained the ResNet using each of these types of simulated inputs and compared the AP@0.5IoU on held-out data after training. The AP based on depth alone is quite high, a little more than 90%, and substantially higher than the AP from either the RGB or monochrome cameras at matched resolution (Figure 1, bar plot). This simulation and many others in this paper show that a high resolution and noise-free depth map contains a great deal of information that can be used to detect the positions of the vehicles.

B. AT LIDAR SPATIAL RESOLUTION, DEPTH IS LESS EFFECTIVE THAN RADIANCE

The scene spatial sampling from a LiDAR sensor is typically lower than the spatial sampling by a camera. In addition, the LiDAR samples are typically matched to the angular resolution that is natural for a beam that sweeps across the scene, which corresponds to the pixel spatial sampling of a pinhole camera. The spatial sampling in a camera depends on the geometric transformation imposed by the camera optics. The examples here used a lens model with relatively little geometric distortion.

Like all physical devices, the LiDAR depth estimates include noise. The amount of noise depends on factors including the distance to the object, the incident angle of rays and the reflectance of the target. In this section, we consider the consequences of obtaining LiDAR data that are at different spatial sampling resolutions and with measurement error in the depth estimate.

1) The dependence on spatial sampling

We calculate the average precision for vehicle detection using spatial sampling patterns that are typical of commercial

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/ACCESS.2021.3063692, IEEE Access

FIGURE 1. A simulated scene captured with different types of sensors: RGB, monochrome, and depth. The radiance sensor parameters are modeled on the MT9V024 sensor by ON Semiconductor, which has a $2.5\,\mu\mathrm{m}$ pixel, 1920 x 650 resolution for a 64 x 21 deg field of view. The depth data are calculated from the simulated scene - LiDAR maps do not produce such high resolution maps. The bar chart shows the average precision (AP) for vehicle detection on networks trained using each data type. High spatial resolution depth information outperforms the RGB or monochrome sensor data by about 10%.

LiDAR systems [28].

The panels in Figure 2A-C contain images from a simulated camera along with superimposed spatial sampling patterns for LiDAR systems with different horizontal and vertical resolutions. The image was modeled as if taken by a wide-angle lens; the superimposed LiDAR sample points are those we expect to obtain assuming the LiDAR images correspond to the sampling through pinhole optics model of the same scene. The lens introduces a small geometric distortion, so there will be a small (1-2 pixel) misalignment between the spatial samples of the LiDAR device and the pixel responses in the camera image.

The chart in Figure 3 compares the AP when networks are trained and then evaluated using only LiDAR data with different spatial sampling resolutions. The densities range from the very highest LiDAR sampling models (AP $\approx$ 90%) to a more typical density with only 0.2% as many samples (AP $\approx$ 64%). The chart shows that when the LiDAR sampling rate is reduced to only 1.5% of the image sampling density (horizontal 0.2 degree/sample, vertical 0.33 degree/sample), performance remains quite high (AP $\approx$ 87%).

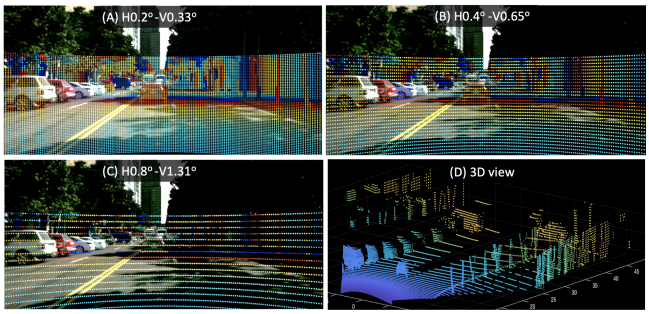


FIGURE 2. Illustration of the different LiDAR spatial resolutions used in the simulations. Panels A-C show the sampling resolution for different horizontal and vertical resolutions (degree/sample). Notice that the LiDAR sampling does not extend into upper angles because, at present, there are no flying vehicles. Panel D is a point cloud representation of the view in Panel (A). The point cloud is useful for human visualization; it contains the same information as the depth map.

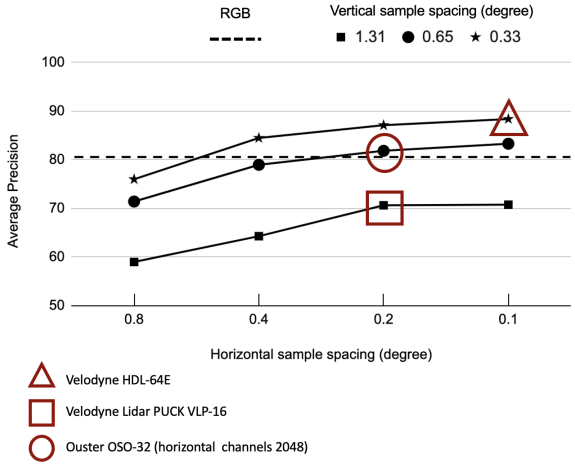


FIGURE 3. The average precision for detecting vehicles from depth data at different spatial resolutions. The lines connect simulations at a common vertical resolution, and the x-axis measures horizontal resolution. The red symbols indicate the spatial resolution of different commercial sensors. The performance using data from an RGB camera (1920 x 650) is shown by the horizontal dashed line.

2) Depth measurement error

Next, we considered the impact of depth-measurement noise on system performance. First, we discuss the noise model, and then we report on the performance when the simulations include noise in the depth data.

LiDAR systems typically send a sequence of rays with a fixed level of energy. As the ray travels from the light source to the object and returns, multiple factors reduce the returned energy level. The factors include transmission loss in the medium, surface reflectivity, geometry relating the incident ray angle, the surface normal, and the surface bidirectional reflectance distribution function. The main source of noise in the LiDAR signal can be attributed to energy reduction. In many cases, the returned ray is not detected by the LiDAR sensor, which is dropout noise.

This noise can be modeled in three different ways. First, the noise can be simulated by randomly deleting points from

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/ACCESS.2021.3063692, IEEE Access

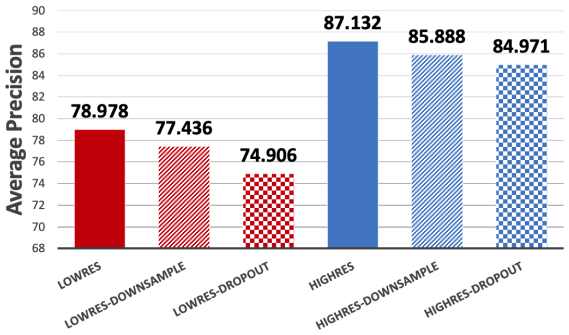


FIGURE 4. AP for vehicle detection when using different depth spatial sampling strategies. Simulations were carried at using low (red) spatial sampling (horizontal 0.4 degree/sample, vertical 0.65 degree/sample) and high (blue) spatial sampling (horizontal 0.2 degree/sample, vertical 0.33 degree/sample). The sampling density was reduced by 20% either by uniformly downsampling (cross-hatched) or by randomly deleting samples (checkerboard).

the simulated sensor [8]. Second, in principle the noise can be simulated by quantitatively accounting for each of the factors that impact the ray transmission, keeping only the rays whose returned energy exceeds a threshold (physical simulation). This approach is somewhat impractical because so many factors are currently unspecified. Third, some investigators have trained a neural network to delete sensor data based on a set of examples of real data [27].

We use the first approach: we delete a random selection of 20% of the data points from the depth map. We use this approach because we do not have enough baseline data to characterize the different sources of physical signal loss (physical simulation), and we do not have a large amount of training data that can be used to train a neural network for the conditions we are simulating. Note that randomly deleting samples reduces the spatial resolution. Hence, it is possible to uniformly, rather than randomly, downsample the data and create a data set with the same number of samples.

We trained and evaluated the ResNet on noise-free data at high and low spatial sampling resolution (Figure 4, solid bars), and with two types of reduced spatial resolution: uniform downsampling and random downsampling to simulate noise (20%). The downsampling was matched between the two methods. Performance is slightly higher with uniform downsampling than random downsampling at both spatial resolutions (Figure 4, cross-hatched and checkers). The uniform downsampling decreases the AP by 1-3 percent at both low and high resolution. The decrease in AP with the random downsampling is 3-4 percent.

C. COMBINED RADIANCE AND DEPTH OUTPERFORMS EITHER ALONE

Radiance and depth data have complementary strengths. Depth information is particularly helpful under very low illumination or when the image is not properly exposed; this can easily happen in a high dynamic range scene when some objects are in direct light and the rest are in shadow. An extreme case is driving through a tunnel. In such conditions, the radiance data captured by the camera may not effectively represent both the bright and dark regions. Depth information is largely invariant to such differences in ambient illumination, making depth helpful for filling in information missed by a poorly exposed camera image.

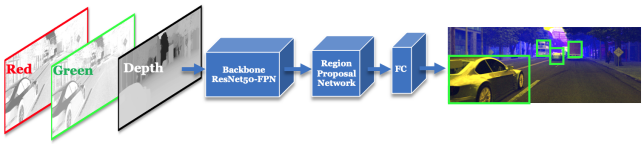


FIGURE 5. Visualization of the ResNet network architecture. The red, green and depth images are input to a ResNet backbone, followed by a region proposal network, and a fully-connected output layer. The final image shows regions where the network detects vehicles (green rectangles). The rectangles are superimposed on the merged RGD image.

Furthermore, radiance data are often obtained using exposure durations that are far longer than the nearly instantaneous temporal point sampling of the LiDAR detector. When measuring nearby moving targets, the radiance data can include a significant amount of motion blur. An advantage of radiance data is that they are easily obtained at much higher spatial resolution than LiDAR, and the data can be acquired at higher frame rates. The radiance data is effective for detecting distant objects that have a small angular extent. Radiance data is also essential for tasks such as finding road markings or identifying traffic light status.

The complementary strengths of the two types of information suggest that a system that combines radiance and depth data may be effective. In the "Related Work" section, we describe a number of papers that explore systems that integrate radiance and depth data. The analyses we have performed suggest that these two types of information can be combined by entering the radiance and depth map in different input channels of a standard CNN (Convolutional Neural Network) (Figure 5). There is no additional computational burden for networks trained using RGD rather than RGB inputs.

1) Creating RGD inputs

We created images by combining the R and G channels from a simulated camera with the depth channel from a simulated LiDAR device. The depth channel is normalized (0-1) and converted to 8-bit integers (0-255). These RGD images have two channels at relatively high spatial resolution (RG) and one channel at lower spatial resolution (D). In separate experiments, we compared filling in the missing values in the D channel with zeroes and linear interpolation. There were no significant differences between these methods.

The RGD data are rendered as color images in Figure 6. The top image shows a simulated driving scene representing using radiance only (RGB), the middle shows a high resolution RGD (horizontal 0.2 degree/sample and vertical 0.65 degree/sample), and the bottom shows a low resolution RGD (horizontal 0.8 degree/sample and vertical 1.31

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/ACCESS.2021.3063692, IEEE Access

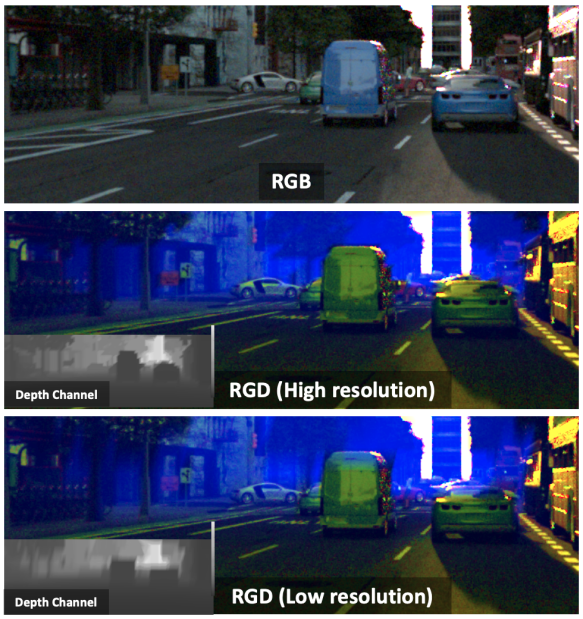


FIGURE 6. Simulated scenes illustrating the radiance image (RGB, top row) or a combined radiance and depth map (RGD, middle and bottom row). In the RGD representation the intensity of the nominal blue channel codes depth. The depth map, sampled at a lower resolution than the image, is shown by the monochrome insets at the lower left of each image. For display we linearly interpolate the low resolution depth data to the RGB resolution. Because the visual system has low spatial resolution to short-wavelength light, the RGD images appear similar to one another despite the difference in depth sampling density.

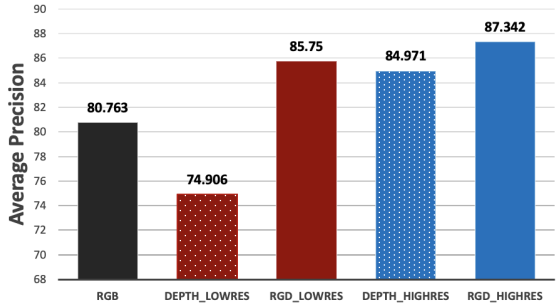


FIGURE 7. Average precision for vehicle detection using radiance alone (RGB), depth alone, or a combination of radiance and depth (RGD). The simulations were performed using low (red) or high (blue) depth spatial sampling resolution. The RGB spatial sampling (black) was 1920 x 650. When RGB resolution is reduced to that of the high resolution depth data, AP falls to 67%.

Results\Datasets (Average Precision)	Radiance (80.76)	Depth (84.97)	RGD (87.34)
Number of vehicles	1749	1749	1749
Number of misses	248	239	183
Number of false alarms	4570	1938	2078

FIGURE 8. Vehicle detection examples on a ResNet trained with radiance, depth or both. The three images show the same scene with ResNet labeled vehicles (green) and ground-truth (purple). The top image shows the network trained on radiance alone (RGB), the middle trained on depth alone (horizontal 0.2 degree/sample, vertical 0.33 degree/sample), and the bottom trained on combined radiance and depth data (RGD). Vehicles that are missed in the RGB trained (purple, top) network differ from the vehicles missed in the depth map (purple middle). The ResNet, trained on the combination of radiance and depth succeeds in both cases. The table shows the statistics of the detection results from radiance, depth and RGD networks.

degree/sample). The images represent the same scene, but the B color channel is replaced with data from the depth map. When depth is small, the images appears yellow (object is near) and when depth is large blue is large (object is far). Consequently, this image appears to be a gradation of more yellow to less yellow as a function of distance. The RGD images are a convenient representation to use as the input to the ResNet, which is designed for a radiance camera (RGB).

2) Evaluating AP with RGD inputs

The data in Figure 7 show the effect on average precision when the ResNet was trained with simulated data either for a conventional radiance camera (RGB), depth data (D), or the combination of radiance and depth data (RGD). The average precision was evaluated at two different depth resolutions. In both cases combining radiance and depth outperforms radiance or depth alone. The improvement is particularly significant when combining low-resolution depth data with the radiance image. In that case the depth alone AP is about 75%, the RGB alone is about 81%, and the combination is about 86%.

Figure 8 shows specific examples in which the RGD sensor achieves better results than either radiance or depth alone. Notice that the image includes a nearby vehicle that is moving, and thus its radiance image is blurred. The vehicle is missed in the radiance image (RGB), but it is detected correctly in the depth map. The distant vehicles at small resolution span very few sample points in the depth map and are missed in the depth image, but the high resolution RGB data detect the distant vehicles well. The vehicles are correctly detected in both cases when using the combined (RGD) data. The table confirms that the performance when

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/ACCESS.2021.3063692, IEEE Access

FIGURE 9. The two images at the top are radiance images from the Waymo dataset. The data set includes corresponding depth data. We combined the radiance and depth information into the two RGD images rendered at the bottom.

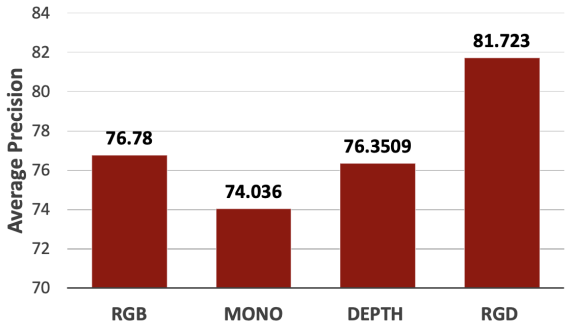


FIGURE 10. Calculations using the Waymo dataset confirm the simulations: AP on the combined radiance and depth data is higher than either alone. Training with the RGB data (1920x743) reached an AP of nearly 77%. Simplifying the radiance data to a single luminance channel (MONO) decreases the AP by a little more than 2%. The AP of a network trained on depth alone is similar to the AP of the RGB data, even though the depth data contain many fewer spatial samples. The ResNet trained on the combined radiance and depth information (RGD) exceeds that of the RGB and the depth by about 5%.

trained with depth data exceeds performance compared to training with radiance data, largely because there are many more false alarms in the radiance condition. Combining the radiance and depth reduces the number of misses compared to depth alone by about 25% but increases the number of false alarms by less than 10%. Thus, the RGD trained network is significantly better than the radiance or depth networks.

The vehicles that are missed by the radiance- and depth-trained networks are not the same. The network trained on depth detects 97 vehicles that were missed by the network trained on radiance. Conversely, the network trained on radiance data detects 94 vehicles that were missed by the network trained on depth.

Combining radiance and depth data in a multi-modal representation has significant advantages. The network trained with RGD data detects 115 vehicles that were missed when trained on radiance alone, 116 vehicles that were missed when trained on depth alone, and 43 vehicles that were missed by both radiance and depth. The multi-modal data extracts information that is not available using radiance or depth separately.

D. SIMULATIONS ARE VALIDATED USING REAL-WORLD DATA

The simulation results are clear: combining radiance and depth information outperforms radiance or depth alone. In this section, we ask whether we find the same pattern of results using publicly available radiance and depth data provided by Waymo [35]. Figure 9 shows renderings of the combined radiance and depth data (RGD) that we constructed from that dataset.

Analyses using the Waymo dataset confirm the simulation findings. Training the ResNet on the RGB radiance data, on a monochrome channel alone, or the depth map alone, resulted in an average precision of around 76% (Figure 10). Combining the radiance and depth information (RGD) increased the average precision to 81%, higher than either radiance or depth alone.

Similarly, as we observed in simulation, the depth map performs well even when its spatial density is relatively low. In the Waymo dataset the spatial samples of the depth map comprise only 1% of the spatial samples of the radiance (RGB) data. Yet, the average precision levels using the two types of data are almost equal.

The examples in Figure 11 confirm that the increase in performance based on the RGD input arises because the combined ResNet can be trained to detect information in both types of data, radiance and depth, taking advantage of the complementary strengths of the two types of measurements.

The table in Figure 11 shows the statistics of the missed and false alarms from the networks trained by three types of data. Comparing with the results in Figure 8, the spatial resolution and noise distribution - largely in the form of missing sample depths - of real-world data is different from the simulated data. This reduces the performance when training a network using the Waymo depth data. But they have enough similarities so that we can see the main observation - RGD performance exceeds Radiance or Depth alone - still holds.

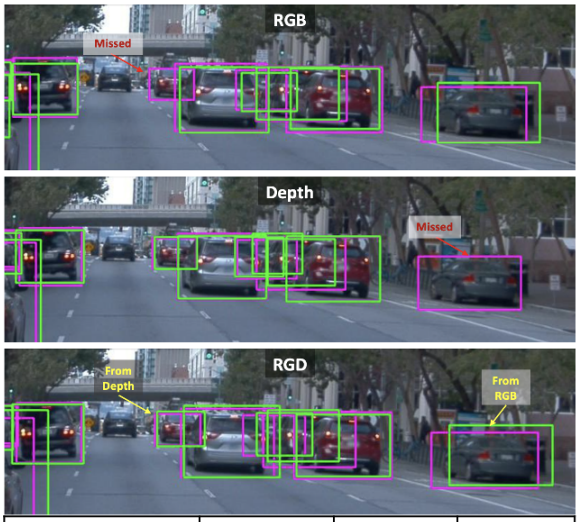
Combining radiance and depth data in a multi-modal representation has significant advantages. The network trained with RGD data detects 88 vehicles that were missed when trained on radiance alone, 68 vehicles when trained on depth alone, and 37 vehicles that were missed by both radiance and depth. The multi-modal data extracts information that is not available using radiance or depth separately.

VI. DISCUSSION

A. SENSOR TECHNOLOGIES

We quantified the increase in average precision using a system that combines radiance and depth information compared to a system with either one alone. The advantages of the combination are significant, making it worth considering the practical challenges of developing an integrated sensor that accurately measures co-registered radiance and depth images under driving conditions. Hardware devices along with

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/ACCESS.2021.3063692, IEEE Access



Results\Datasets (Average Precision)	Radiance (76.78)	Depth (76.35)	RGD (81.72)
Number of vehicles	950	950	950
Number of misses	183	162	114
Number of false alarms	1708	3189	2673

FIGURE 11. Vehicle detection examples from the Waymo dataset on a ResNet trained with radiance, depth or both.match the simulation analyses. Vehicles that were missed in RGB but found in depth, are also found in RGD. Similarly, vehicles that were found in RGB but missed in depth are found in RGD. Box outline colors as in Figure 8. The table shows the statistics of the detection results from radiance, depth and RGD networks.

their limitations are reviewed in [21]. Calibration and co-registration algorithms to fuse LiDAR and radiance images are reviewed in [23].

Time-of-flight. LiDAR systems typically sweep a laser light through the scene and measure the time-of-flight using a small array of avalanche photodetectors, and more recently single-photon avalanche detectors (SPADs). Gated imaging sensor systems also determine depth based on time-of-flight [16]. These systems emit a very brief near-infrared illumination pulse or a periodic temporal pattern and then precisely control (gate) a sequence of very brief electronic exposure times. The duration of the pulse, the distance to the object, and timing of 3-5 pre-determined exposure times produces a pattern of photon absorptions that can be used to estimate the distance to a scene object. For example, Canesta developed a time-of-flight system based on a special purpose CMOS image sensor [13].

The circuit requirements for specialized time-of-flight devices differ significantly from the circuits for radiance measurements. Kim et al. [22] showed that it is possible to integrate both types of circuits in a single sensor, interleaving radiance and gated time-of-flight pixels in a single array. This type of sensor could provide spatially aligned radiance and depth maps at resolutions that are comparable to the density simulated in this paper. It is possible to use image systems simulation to evaluate the performance for different spatial sampling configurations as well as temporal gating configurations that are within the reach of modern technology.

Structured light. Current technologies that simultaneously acquire radiance and depth information often use structured light (e.g., Kinect, Real Sense from Intel; Kinect style RGB-D type cameras) [11]. These systems illuminate a scene with a known spatial pattern and measure the returned image, inferring depth from the difference between the known illumination pattern and the measured image. Such technology is effective in certain contexts, but it is not appropriate for the vehicle detection we analyze in this paper.

Stereo. Depth can also be estimated using a stereo pair or array of cameras. This approach relies only on radiance and would not require integrating radiance and time-of-flight technology. But the simulations demonstrate that a key limit of the radiance data include dynamic range and blur, and these problems are not solved by a system based only on radiance cameras. Furthermore, the simulations show that the spatial pattern of the missing depth information matters; when estimating depth from stereo pairs or arrays the spatial pattern of the missing depth information will be highly structured and very different from the missing information using LiDAR systems.

It will be useful to assess whether the same high level of vehicle detection can be obtained using depth derived from stereo. The quality of stereo depth estimation algorithms continues to improve, and it may be that this can be an effective approach [38]. Similarly, the quality of depth information obtained from SPAD arrays has generally been inferior to the data from standard LiDAR using avalanche detectors. There is promising research that seeks to improve the SPAD depth estimation by combining the time-of-flight data with radiance data ([36]). The simulations in this paper suggest that low resolution SPAD inputs that are properly aligned to the radiance data may be a useful approach to finding vehicles.

B. DEPTH AND RADIANCE REPRESENTATIONS

Investigators have used a variety of approaches to represent radiance and depth information. For example, some investigators convert RGB-Depth information into the format of height/horizontal disparity and angle (HHA [17]). Others keep the depth and radiance information separate through multiple input stages, allowing them to converge only many layers deep in the network [4]. Keeping the two sources of data distinct also permits the system to use very different formats for representing the radiance and depth information.

Point clouds and depth maps. The depth maps and 3D point cloud formats are equivalent in the sense that there are transformations that convert precisely between them. Yet, for some applications point cloud representations may be advantageous. For example, [37] report that point clouds perform better for determining the 3D bounding box of a vehicle. The task we analyze is detecting 2D bounding boxes, and for this goal the depth map format improves performance significantly. Future experiments should explore how the

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/ACCESS.2021.3063692, IEEE Access

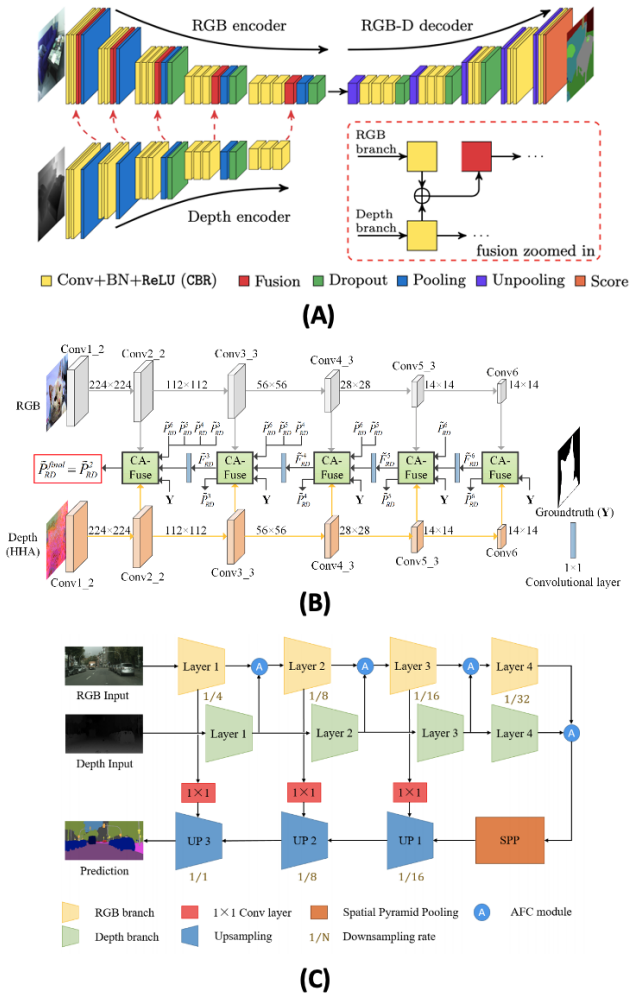


FIGURE 12. Sensor fusion architectures. Panels (A-C) are overviews of the network architectures from three papers that fused radiance and depth images. Panel (D) is the ResNet architecture used in this paper that uses RGD as the fused input. (A) (After Hazirbas et al., 2016) [18] (B) (After Chen et al.2018) [4] (C)(After Sun et al.2020) [34]

effectiveness of the RGD representation for detecting 3D bounding boxes.

C. NEURAL NETWORK ARCHITECTURES FOR SENSOR FUSION

In prior work, authors developed deep learning architectures to combine radiance and depth information [4], [18], [29], [34]. The literature includes numerous innovative approaches for combining depth maps or point clouds with radiance data (Figure 12A-C). Many of these architectures initiate the network by keeping the two modalities in separate channels and combining information from the distinct channels at layers that are deep within the network. One paper systematically examined which network layer would be optimal for combining the two data streams (Fig 13E) [10] and concluded that performance is quite similar if one chooses early, middle, or late fusion architectures.

The simulations we describe here use modern CNN methods [19] that take aligned depth and radiance images as input. The depth maps are combined with the radiance data at the earliest stage: the CNN input channel. Combining the data at the input makes it very straightforward to add a region proposal network [12]. This architecture has been highly optimized; in some cases such a network performs at levels that reach optimal [31]. At this time, we see no reason to use a more complex architecture.

VII. CONCLUSION

The analyses of simulations and real-world data quantify the value of depth information for vehicle detection. The results shows that after equating for spatial resolution, depth information is at least as valuable as radiance information. When depth information is only available at low spatial resolution, combining depth and radiance by inserting the depth map into an image input channel increases the average precision of vehicle detection substantially. We demonstrated this improvement using ISETAuto simulations, and we confirmed the finding using real-world data.

The advantage of combining radiance and depth information can be explained by the fact that the two modalities have complementary weaknesses. The depth information is acquired with extremely short duration exposures that limit the impacts of blur in moving targets. Also, depth information is less vulnerable to the dynamic range limits of cameras. Conversely, cameras have a spatial resolution advantage over LiDAR devices, and they are necessary for measuring some important information such as road markings, signs, and traffic signals. These observations suggest that there may be performance advantages for an integrated sensor that provides aligned radiance and depth images as an input to a CNN for vehicle detection.

REFERENCES

[1] Flywheel • modern informatics platform for biomedical research & collaboration. https://flywheel.io/, Mar. 2017. Accessed: 2021-1-25.
[2] M. Behrisch, L. Bieker, J. Erdmann, and D. Krajzewicz. Sumo–simulation of urban mobility. In The Third International Conference on Advances in System Simulation (SIMUL 2011), Barcelona, Spain, volume 42, 2011.
[3] L. Caltagirone, M. Bellone, L. Svensson, and M. Wahde. LIDAR–camera fusion for road detection using fully convolutional neural networks. Rob. Auton. Syst., 111:125–131, Jan. 2019.
[4] H. Chen and Y. Li. Progressively complementarity-aware fusion network for RGB-D salient object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3051–3060, 2018.
[5] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. Carla: An open urban driving simulator. In Conference on robot learning, pages 1–16. PMLR, 2017.
[6] E. Bruce Goldstein and James R. Brockmole. Sensation and Perception (10th Edition). Cengage Learning, 2017.
[7] M. Everingham, L. Van Gool, C. K. I. Williams, and others. The pascal visual object classes (voc) challenge. International journal of, 2010.
[8] J. Fang, D. Zhou, F. Yan, T. Zhao, F. Zhang, and others. Augmented lidar simulator for autonomous driving. IEEE Robotics and, 2020.
[9] J. E. Farrell, P. B. Catrysse, and B. A. Wandell. Digital camera simulation. Appl. Opt., 51(4):A80–90, Feb. 2012.
[10] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. Feb. 2019.

[11] J. Geng. Structured-light 3D surface imaging: a tutorial. Adv. Opt. Photon., AOP, 3(2):128–160, June 2011.
[12] G. Gkioxari, J. Malik, and J. Johnson. Mesh R-CNN. June 2019.
[13] S. B. Gokturk, H. Yalcin, and C. Bamji. A Time-Of-Flight depth sensor - system description, issues and solutions. In 2004 Conference on Computer Vision and Pattern Recognition Workshop, pages 35–35, June 2004.
[14] D. R. Griffin, F. A. Webster, and C. R. Michael. The echolocation of flying insects by bats. Readings in the Psychology of Perception, page 21, 1965.
[15] D. W. Griffin. More about bat "radar". Scientific American, 199(1):40–45, 1958.
[16] T. Gruber, F. Julca-Aguilar, M. Bijelic, W. Ritter, K. Dietmayer, and F. Heide. Gated2Depth: Real-time dense lidar from gated images. Feb. 2019.
[17] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik. Learning rich features from RGB-D images for object detection and segmentation. July 2014.
[18] C. Hazırbaş, L. Ma, C. Domokos, and D. Cremers. FuseNet: Incorporating depth into semantic segmentation via Fusion-Based CNN architecture. Nov. 2016.
[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.
[20] J. Hoffman, S. Gupta, J. Leong, S. Guadarrama, and T. Darrell. Cross-modal adaptation for RGB-D detection. In 2016 IEEE International Conference on Robotics and Automation (ICRA), pages 5032–5039, May 2016.
[21] S. Ito, S. Hiratsuka, M. Ohta, H. Matsubara, and M. Ogawa. Small imaging depth LIDAR and DCNN-Based localization for automated guided vehicle. Sensors, 18(1), Jan. 2018.
[22] W. Kim, W. Yibing, I. Ovsiannikov, S. Lee, Y. Park, C. Chung, and E. Fossum. A 1.5mpixel RGBZ CMOS image sensor for simultaneous color and range image capture. In 2012 IEEE International Solid-State Circuits Conference, pages 392–394, Feb. 2012.
[23] G. A. Kumar, J. H. Lee, J. Hwang, J. Park, S. H. Youn, and S. Kwon. LiDAR and camera fusion approach for object distance estimation in Self-Driving vehicles. Symmetry, 12(2):324, Feb. 2020.
[24] Z. Liu, T. Lian, J. Farrell, and others. Soft prototyping camera designs for car detection based on a convolutional neural network. Proc. IEEE, 2019.
[25] Z. Liu, T. Lian, J. Farrell, and B. A. Wandell. Neural network generalization: The impact of camera parameters. IEEE Access, 2020.
[26] Z. Liu, M. Shen, J. Zhang, S. Liu, H. Blasinski, and others. A system for generating complex physically accurate sensor images for automotive applications. Electronic, 2019.
[27] S. Manivasagam, S. Wang, K. Wong, W. Zeng, M. Sazanovich, S. Tan, B. Yang, W.-C. Ma, and R. Urtasun. LiDARsim: Realistic LiDAR simulation by leveraging the real world. June 2020.
[28] H. A. Manufacturer, D. A. R. Li, R. I. Automotive, F. L. Range, E. O. Tracking/Classification, and A. Quality. LiDAR SPECIFICATION COMPARION CHART.
[29] T. Ophoff, K. Van Beeck, and T. Goedemé. Exploring RGB+Depth fusion for Real-Time object detection. Sensors, 19(4), Feb. 2019.
[30] M. Pharr, W. Jakob, and G. Humphreys. Physically Based Rendering: From Theory to Implementation. Morgan Kaufmann, Sept. 2016.
[31] F. Reith and B. Wandell. A convolutional neural network reaches optimal sensitivity for detecting some, but not all, patterns. Nov. 2019.
[32] G. Rong, B. H. Shin, H. Tabatabaee, Q. Lu, S. Lemke, M. Možeiko, E. Boise, G. Uhm, M. Gerow, S. Mehta, et al. Lgsvl simulator: A high fidelity simulator for autonomous driving. In 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), pages 1–6. IEEE, 2020.
[33] S. Shah, D. Dey, C. Lovett, and A. Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In Field and service robotics, pages 621–635. Springer, 2018.
[34] L. Sun, K. Yang, X. Hu, W. Hu, and K. Wang. Real-time fusion network for RGB-D semantic segmentation incorporating unexpected obstacle detection for road-driving images. Feb. 2020.
[35] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. Dec. 2019.
[36] Z. Sun, D. B. Lindell, O. Solgaard, and G. Wetzstein. SPADnet: deep RGB-SPAD sensor fusion assisted by monocular depth estimation. Opt. Express, OE, 28(10):14948–14962, May 2020.
[37] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger. Pseudo-LiDAR from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving. Dec. 2018.
[38] Y. You, Y. Wang, W.-L. Chao, D. Garg, G. Pleiss, B. Hariharan, M. Campbell, and K. Q. Weinberger. Pseudo-LiDAR++: Accurate depth for 3D object detection in autonomous driving. June 2019.
[39] Yuanzhouhan Cao, Chunhua Shen, and Heng Tao Shen. Exploiting depth from single monocular images for object detection and semantic segmentation. IEEE Trans. Image Process., 26(2):836–846, Feb. 2017.

ZHENYI LIU received his MS in Electrical Engineering at Ulsan National Institute of Science and Technology, UNIST (2015), Korea. He is currently a PhD candidate in Automotive Engineering at Jilin University, China (2016-present). Zhenyi was a Visiting Student Researcher at Stanford University (2017-2019). His research interests focus on machine perception systems for autonomous vehicles such as cameras and LiDAR.

JOYCE FARRELL is a Senior Research Engineer and Lecturer in the Department of Electrical Engineering; she is the Executive Director of the Stanford Center for Image Systems Engineering. Dr. Farrell co-founded ImageVal Consulting and has more than 20 years of research and professional experience working at a variety of companies and institutions, including the NASA Ames Research Center, New York University, the Xerox Palo Alto Research Center, Hewlett Packard Laboratories and Shutterfly.

BRIAN A. WANDELL is the first Isaac and Madeline Stein Family Professor. He joined the Stanford Psychology faculty in 1979 and is a member, by courtesy, of Electrical Engineering, Ophthalmology, and the Graduate School of Education. He is Director of Stanford's Center for Cognitive and Neurobiological Imaging and Deputy Director of Stanford's Neurosciences Institute. Wandell's research centers on vision science, spanning topics from visual disorders, reading development in children, to digital imaging devices and algorithms for both magnetic resonance imaging and digital imaging.