Polarimetric Imaging for Perception

Michael Baltaxe
michael.baltaxe@gm.com
Tomer Pe'er
tomer.1.peer@gm.com
Dan Levi
dan.levi@gm.com

General Motors
Hertzliya, Israel

Abstract

Autonomous driving and advanced driver-assistance systems rely on a set of sensors and algorithms to perform the appropriate actions and provide alerts as a function of the driving scene. Typically, the sensors include color cameras, radar, lidar and ultrasonic sensors. Strikingly however, although light polarization is a fundamental property of light, it is seldom harnessed for perception tasks. In this work we analyze the potential for improvement in perception tasks when using an RGB-polarimetric camera, as compared to an RGB camera. We examine monocular depth estimation and free space detection during the middle of the day, when polarization is independent of subject heading, and show that a quantifiable improvement can be achieved for both of them using state-of-the-art deep neural networks, with a minimum of architectural changes. We also present a new dataset composed of RGB-polarimetric images, lidar scans, GNSS / IMU readings and free space segmentations that further supports developing perception algorithms that take advantage of light polarization. The dataset can be downloaded here.

1 Introduction

Advanced driver-assistance systems (ADAS) and autonomous vehicles need to interpret the surrounding environment to plan and act appropriately on the road. To do so, modern vehicles are equipped with a set of sensors and algorithms that carry out a variety of perception tasks such as free space detection [8, 11, 23], lane detection [14, 27], object detection [26, 28, 29], 3D pose estimation [30, 33], depth estimation [13, 16], etc.

The quality of the perception output is a function of the quality of the sensor suite installed in the vehicle. Currently, RGB cameras, ultrasonic sensors and radars are standard equipment in production vehicles, and in the near future lidars will also be readily available thanks to the steady decrease in their size and price.

RGB-polarimetric cameras are sensors that measure light polarization, in addition to light intensity and color. These cameras are already in use in several industrial applications; for example, to detect defects in surfaces and to improve image quality by removing reflections. However, to the best of our knowledge, they have not been used in the automotive domain. In this work we explore the potential of using RGB-polarimetric cameras to improve the performance of algorithms employed in common autonomous driving perception tasks.

Since light polarization is a complex function of light properties, scene properties and viewing direction, we limited this study to the case when the sun was high in the sky (around midday). This way the polarization state of collected light was independent of vehicle heading, and consistent readings were made in all driving directions.

We focus on free space detection and depth estimation. Both are of paramount importance for automated driving and ADAS systems, but are also useful for viewing systems when creating surround and bowl-views of the environment.

The purpose of free space detection is to segment the input image such that all pixels where the vehicle can drive are labeled '1' and all the other pixels in the image are labeled '0', thus creating a mask indicating the "drivable" space which enables planning and navigation.

By contrast, depth estimation calculates the depth from the camera sensor to the imaged object in each pixel, thus yielding a dense distance map of the scene. There are currently several methods for depth estimation, many of which use either multiple view geometry or active imaging. Here, we examined monocular depth estimation (monodepth), where the depth is estimated from a single image taken with a passive camera (i.e., without active illumination). In our case, we used polarimetric information in addition to the RGB image to estimate depth.

Note that specialized hardware is used in many applications to achieve depth estimation. Lidar, for example, has been the prime choice to enrich the sensing suite and achieve high quality reliable sensing. Compared to this solution, the RGB-polarimetric camera has the desirable properties of being cheaper and providing a dense map that is readily aligned with the RGB image, without the need for complex alignment procedures.

The development and evaluation of perception algorithms calls for a dataset that supports the target tasks. Modern perception methods typically use convolutional neural networks (CNN) [20] and other deep learning methods [21], which require a large amount of data for training. Although there are several available public-domain automotive datasets [4, 6, 9, 15, 25, 31, 34], few have polarimetric data. To the best of our knowledge, the only automotive dataset that includes polarimetric information is [4]. Its major drawback is that the RGB and polarimetric images were obtained from different devices, so that the synchronization between modalities is only partial. To overcome this hurdle, we built a dataset of RGB-polarimetric data composed of RGB images, polarimetric images (angle of linear projection and degree of linear polarization), lidar point clouds, GPS and inertial measurements.

This dataset is composed of 12,627 images from 6 different locations. The data were collected around noon in fair weather. The intrinsic parameters of the camera (focal distance and principal point) and extrinsic parameters (translation and rotation) between all the hardware components were calibrated, thus providing full alignment between the camera and all other sensing elements. Note that alignment between the RGB and polarimetric images was easy to achieve since a single sensor was used to create both images (see section 3.1).

The contribution of this work is twofold. First, it presents a dataset of RGB-polarimetric data with naturalistic driving scenarios useful for several perception tasks. Then, we present a detailed analysis of the performance of key perception algorithms using RGB-polarimetric data which are compared to the performance obtained when only RGB data are available.

The remainder of this paper is organized as follows. Section 2 presents related work concerning perception in the automotive domain and the use of light polarization for perception. Section 3 presents the specifics of the dataset and the approach used to include polarimetric data in two perception tasks. Section 4 reports the experiments and section 5 concludes.

2 Related Work

2.1 Shape from Polarization

Shape from polarization aims to recreate surface shapes from light polarization measurements. The method presented in [1] was one of the first to recover surface normals from polarimetric images of objects with diffusive-reflecting materials using analytical methods. Newer methods such as [2, 22] use deep neural networks to cope with the fact that real world objects exhibit specular and diffusive behavior. The method in [3] used a CNN to estimate depth from polarimetric and grayscale images rather than estimating surface normals.

2.2 Monocular Depth Estimation

Monocular depth estimation is crucial to computer vision. Supervised methods such as in [13, 19] were trained to learn a direct distance metric for each pixel from ground truth data collected with specialized depth sensors. By contrast, self-supervised methods such as [16, 17, 35, 36] are much more data efficient because they take advantage of the geometric relationships within a scene when the camera moves in space. For this purpose, pairs of consecutive frames are used to learn pose and depth estimation networks, while minimizing the reprojection error.

2.3 Free Space Detection

Free space detection has been widely studied by the autonomous driving community. In this task, the system outputs a segmentation of the environment in which the vehicle can drive, usually corresponding to a road. Several approaches exclusively use camera information, while others also use lidar point clouds. Early work introduced in [23] developed a method that takes an RGB image as input and uses a CNN to extract *stixels*, a compact representation of free space. More recent works such as [7, 11, 32] use lidar point clouds to extract surface normals which are fed along with the RGB image into a fully convolutional network. Similarly, [8] used a two-stream neural network to process an RGB image together with an *altitude difference* image extracted from the lidar point cloud. Yet another approach was taken by [24], which used multiple cameras and a vision transformer to yield a precise segmentation.

2.4 Datasets for Driving Perception Tasks

There are several specialized open datasets for driving perception tasks. Probably the best known are described in [6, 9, 15, 25, 31, 34]. All these datasets include RGB images, and several also provide lidar point clouds. The annotation level covers 2D object bounding boxes, 3D object bounding boxes, drivable area delineation, object tracking, instance segmentation and segmantic segmentation for the image or point cloud modalities. In [4], a dataset of RGB and polarimetric images was used for object detection. In this dataset, two different cameras were used to capture the two modalities but no extrinsic calibration was calculated, so that the RGB and polarization image pairs were not perfectly aligned or synchronized.

3 Method

3.1 Dataset

Since our perception methods rely on deep learning techniques, we needed a large dataset with polarimetric information. We built a custom setup to gather data for our experiments. The setup included the following hardware:

1. RGB-polarimetric camera (Lucid Vision TRI050S-QC with Sony IMX250MYR CMOS color sensor).
2. Lidar (Velodyne Alpha Prime).
3. GNSS / INS (OxTS RT3000).

The RGB-polarimetric camera outputs RGB values with polarization filters at four different angles, from which intensity (I), angle of linear polarization (AoLP) and degree of linear polarization (DoLP) images can be calculated, as explained below. The camera has each color pixel sub-divided into four regions, each of which has a different polarization filter, thus, no rotation of a single polarization filter is involved and synchronization is automatic. The resolution of the restored images was 1.25 megapixels, with a field of view of 60°.

Let $P_0, P_{45}, P_{90}, P_{135}$ be the intensity of the polarization images obtained by the camera, where the subscripts indicate the orientation angle of the polarization filter. Then, following standard practice [3, 22], we calculated the intensity, AoLP and DoLP as follows:

$$I = \frac{(P_0 + P_{45} + P_{90} + E_{135})}{2} \tag{1}$$

$$DoLP = \frac{\sqrt{(P_0 - P_{90})^2 + (P_{45} - P135)^2}}{I} \tag{2}$$

$$AoLP = \frac{1}{2} \arctan\left(\frac{P_{45} - P_{135}}{P_0 - P_{90}}\right). \tag{3}$$

The lidar camera system was synchronized temporally with the lidar used as the trigger for the camera, providing a trigger signal each time the revolving head reached the 0° mark. The camera and lidar data were collected at a frame rate of 10 Hz, and the GNSS / INS was sampled at 100 Hz.

The camera's intrinsic parameters were calibrated using the standard chessboard method implemented in OpenCV [5]. The extrinsic parameters between the camera and the lidar (translation and rotation) were calibrated by calculating the rigid transformation that achieved the smallest distance between several planes in the three main directions (in terms of least squares), extracted independently with the lidar and the camera. In this case, the camera planes were extracted by capturing images of a chessboard.

Figure 1 presents a few examples of the collected data. The cyclic color coding in the AoLP image shifts from red for 0° to magenta for 179°. In the DoLP image, 0 corresponds to black and 1 corresponds to yellow.

Windshields tend to light up in the DoLP image because they have a high degree of linear polarization due to the smoothness of the glass. In addition, the road and other horizontal elements tend to have an AoLP close to 0° (purple) because their normal is aligned upwards and the electric field of light oscillates perpendicularly. Note as well that the AoLP depends

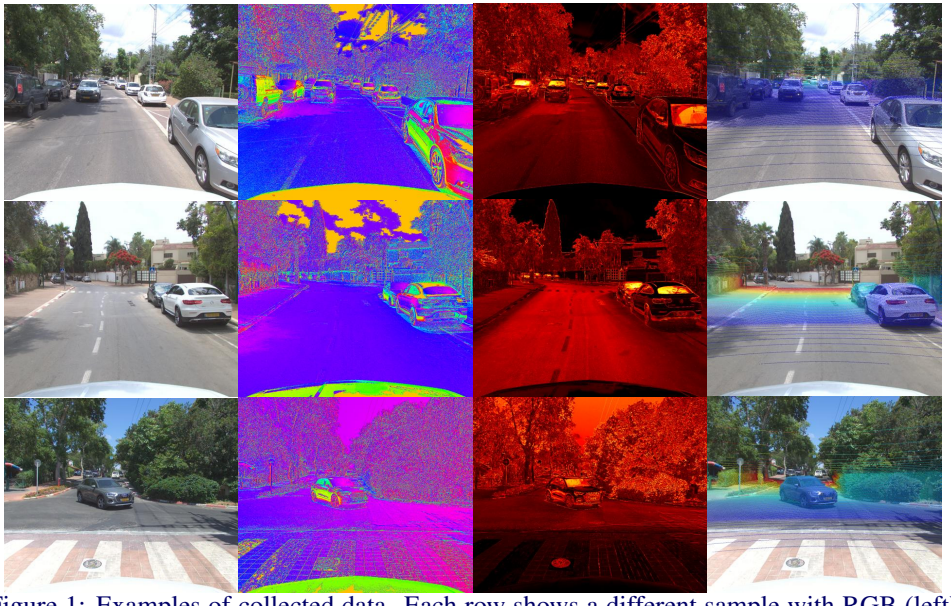

Figure 1: Examples of collected data. Each row shows a different sample with RGB (left), AoLP (middle left), DoLP (middle right) and lidar projected on RGB (right). The cyclic color map in the AoLP images goes from red for 0° to magenta for 179°. In the DoLP images black corresponds to 0 and yellow to 1.

not only on geometry, but also on the material, as shown for example on the side of the two vehicles in the second row. While both vehicles have the same orientation with their sides located vertically, for the black car the AoLP is close to 90° (green), but for the white car the AoLP is close to 0° (purple). This effect can also be seen on the side window of the white car in the second row, which has the same geometry as the car side, but is made of different material.

The dataset was composed of 12,627 images from 6 different locations. These locations represent typical suburbs where the scenes are not remarkably cluttered, but provide a good distribution of vehicles, pedestrians and buildings. No highways were included in the dataset since we used self-supervised methods for monodepth estimation (section 3.3), which are known to degrade strongly when the dataset contains vehicles that move at speeds similar to the ego-vehicle's speed. Additionally, the dataset includes free space segmentation of 8,141 images. The segmentations were created in a semi-automatic way using the SAM segmentation method [18] followed by manual refinement.

3.2 Free Space Detection

Our objective in this study was to quantify the potential benefits of using an RGB-polarimetric camera for perception tasks. For the free space scenario, we based our method on the SNE-RoadSeg architecture [11], one of the top-scoring methods in the KITTI road benchmark with open-source code. The original network takes an RGB image and a depth image (usually acquired with a lidar) as inputs, and outputs a free space segmentation. This method initially estimates the surface normal from the depth image using the SNE module, and then uses a deep neural network to perform the final segmentation. The SNE-RoadSeg network

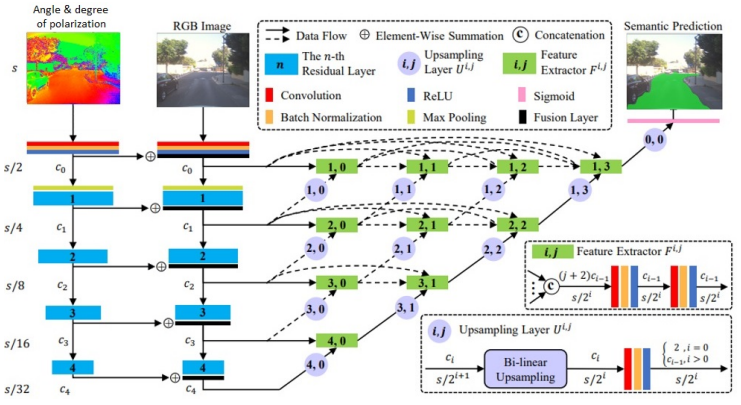

Figure 2: RGBP-RoadSeg architecture.[1]

has two streams: one extracts features from the surface normal channels and the other extracts features from a concatenation of the RGB and the surface normal images.

Surface normals are tightly correlated to polarimetric measurements. Specifically, following [1] (under the assumption that the materials are not ferromagnetic), the specular AoLP (ϕ_s) and DoLP (ρ_s) and the diffusive AoLP (ϕ_d) and DoLP (ρ_d) are related to the surface normal's azimuth (α) and zenith (θ) angles as follows:

$$\phi_s = \alpha - \frac{\pi}{2} \tag{4}$$

$$\rho_s = \frac{2\sin^2(\theta)\cos(\theta)\sqrt{n^2-\sin^2(\theta)}}{n^2-\sin^2(\theta)-n^2\sin^2(\theta)+2\sin^4(\theta)} \tag{5}$$

$$\phi_d = \alpha \tag{6}$$

$$\rho_d = \frac{\left(n-\frac{1}{n}\right)^2\sin^2(\theta)}{2+2n^2-\left(n+\frac{1}{n}\right)^2\sin^2(\theta)+4\cos(\theta)\sqrt{n^2-\sin^2(\theta)}} \tag{7}$$

where n is the refractive index of the material of the object being imaged.

We hypothesized that a network that is able to extract information from the surface normal should also infer successfully from polarimetric data, when trained properly. For this reason, our architecture was exactly the same as the SNE-RoadSeg, except that the SNE module was dropped and the input surface normal channels were replaced by a concatenation of polarimetric channels as follows:

$$P = [\sin(2\cdot AoLP), \cos(2\cdot AoLP), 2\cdot DoLP - 1]\,. \tag{8}$$

We used the sine and cosine functions on the AoLP to cope with the fact that the AoLP is a cyclic function, where a measurement of 0° is equivalent to 180°. We scaled the DoLP to be in the range [-1, 1], as for the other two features.

The architecture used for the free space detection network, dubbed *RGBP-RoadSeg*, is depicted in Figure 2.

[1] Image adapted from [1] with permission from the authors.

3.3 Monocular Depth Estimation

For the monocular depth estimation we drew on the well-known monodepth v2 framework [16]. The main idea is to take pairs of consecutive frames over time and learn two networks: one for depth estimation and one for estimating the relative camera pose between the two frames. A loss is calculated by warping the estimated depth from one frame to the next by applying the estimated camera pose and projecting back to the image. This is a clever self-supervised method to learn depth, its main strength is that no manual labeling is needed.

The original work used RGB images as input to the networks. In our system, we used the RGB image concatenated with the three polarimetric features described in equation 8 as input. The architecture is the same as the one used in monodepth v2.

4 Experiments

4.1 Free Space Detection

Data: We used the SAM system [18] to create automatic segmentations of the scenes by providing as input prompt a point right in front of the vehicle's hood, which can be expected to be part of the road with high probability. Then, the segmentations were inspected and manually refined. Overall we extracted 8,141 segmentations which were divided into train (6,206 images), validation (856 images) and test (969 images) splits. The train, test and validation data were mutually exclusive geographically. The number of images used in this task was smaller than the full dataset as extracting segmentations is quite expensive.

Evaluated Methods: Our *RGBP-RoadSeg* as described in section 3.2 is compared to other RoadNet incarnations. The closest RGB-only implementation, named *RGB-RoadSeg*, was exactly like the RGBP-RoadSeg, but the left stream in Figure 2 was completely dropped, leaving only the RGB input and the skip connections of the original network. The *P-RoadSeg* dropped the left stream of the network, and used the polarimetric features of equation 8 as input to the right stream. Finally, we evaluated the standard *SNE-RoadSeg* network [11] which used as input RGB images and depth images processed by the SNE module, allowing to compare lidar-based and polarization-based methods.

Metrics: We used the standard metrics of [12]: accuracy, precision, recall, maximum F-score (F_{max}) and average precision (AP). Intersection over union (IoU) was also evaluated. As in [12], all metrics were calculated on bird's-eye view projections of the scenes.

Results: The results of the free space detection are presented in Table 1. First, note that P-RoadSeg yielded mediocre results, implying that polarization alone does not carry enough information for this task. RGB-RoadSeg provided much better results, which tells us that the RGB modality is more suited for free space estimation. The best results, however, were obtained by RGBP-RoadSeg which uses both RGB and polarimetic information suggesting that both modalities are complementary and carry independent information. RGBP-RoadSeg (polarimetric camera) is on a par with SNE-RoadSeg (lidar), although it is important to recall the noon-time constraint on the polarimetric camera.

Figure 3 shows some qualitative results. Note the extent to which low contrast areas were improved by the use of polarization data. This makes sense since color contrast is not always correlated with polarization contrast. For example, the wall and road in the third column have similar colors (yielding poor color contrast), but the wall orientation has a 90° shift with respect to the road, which yields a high contrast in the AoLP image.

Method	Accuracy	Precision	Recall	F_{max}	IoU	AP
RGB-RoadSeg	0.979	0.949	0.968	0.953	0.902	0.974
P-RoadSeg	0.865	0.845	0.534	0.641	0.467	0.634
RGBP-RoadSeg	**0.986**	**0.966**	**0.972**	**0.968**	**0.939**	**0.994**
SNE-RoadSeg	0.985	0.967	0.967	0.965	0.934	0.993

Table 1: Results for the free space detection task.

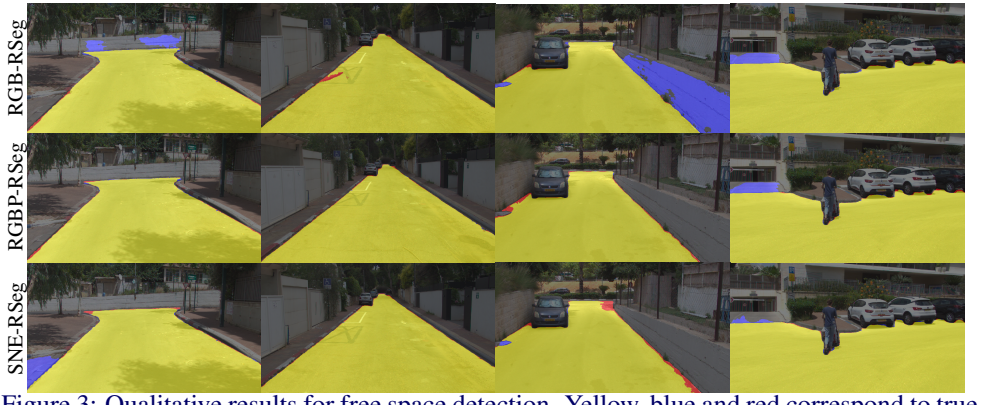

Figure 3: Qualitative results for free space detection. Yellow, blue and red correspond to true positive, false positive and false negative respectively. RGB-RoadSeg bled in low contrast regions and missed painted road areas. SNE-RoadSeg relies on depth and bled to ground outside of the road. The right column is a failure case: semantics are needed to find the edge.

4.2 Monocular Depth Estimation

Data: We divided the dataset into train (6,117 images), validation (779 images) and test (779 images) splits. The number of images used does not include the full dataset since the self-supervised monocular depth paradigm cannot use frames where the vehicle is static. We set a minimum speed of 15 Km/h and use all frames where the vehicle moves at higher speed. The train, validation and test data are divided so that there is no geographical overlapping.

Evaluated Methods: Our baseline method, denoted *RGB-Depth*, simply performed monocular depth estimation using RGB images, as is commonly done. Then, we dropped the RGB images and instead used the polarimetric features in equation 8 as input to the system, we call this *P-Depth*. Next, we analyzed the possibility to use RGB and polarimetric data in a synergistic manner. In this case, we stacked the RGB images with the polarimetric features in equation 8 and used this as input to the monodepth method, we refer to this setup as *RGBP-Depth*. Finally, we pre-trained the model on RGB images and fine-tuned it using the stacking of RGB and polarimetric features (the polarimetric features were initialized randomly). This last method is regarded as *pt-RGBP-Depth*.

Metrics: We used the standard metrics introduced in [10] to quantify both the error and the accuracy of the methods. For details, see [10].

Results: The results of these experiments are presented in Table 2. First of all, note that the RGB results are consistent with the results of the original paper, showing that our dataset is relevant for the task. Using polarimetric data instead of RGB we see an improvement, this

Method	Error metric ↓				Accuracy metric ↑		
	Abs Rel	Sq Rel	RMSE	RMSE Log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
RGB-Depth	0.094	0.838	6.389	0.166	0.904	0.964	0.984
P-Depth	0.091	0.811	6.325	0.164	0.907	0.966	0.985
RGBP-Depth	0.089	0.770	6.172	0.161	0.911	**0.968**	**0.986**
pt-RGBP-Depth	**0.086**	**0.767**	**6.109**	**0.158**	**0.915**	**0.968**	0.985

Table 2: Results for the monodepth estimation task.

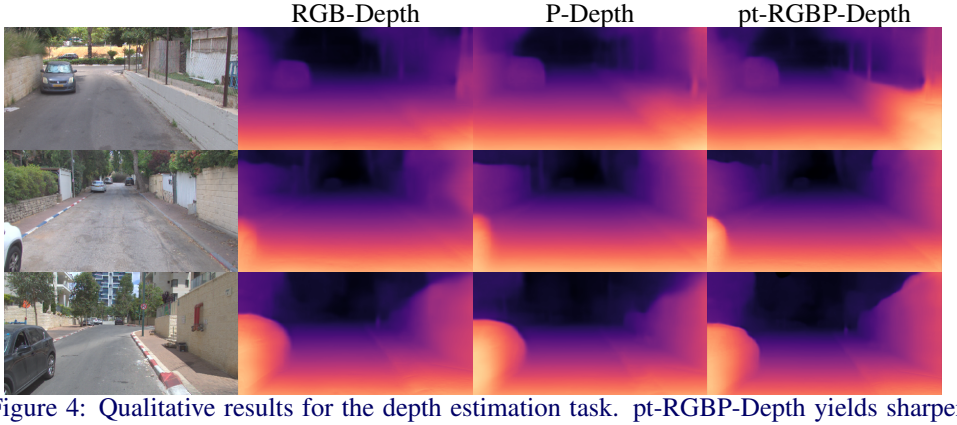

Figure 4: Qualitative results for the depth estimation task. pt-RGBP-Depth yields sharper edges and better recovers all structures.

is probably since the polarimetric modality includes a lot of information that is relevant to the task of depth estimation. Stacking together the RGB and polarimetric data in the RGBP-Depth method yields further improvement, showing that both modalities are complementary and do not carry the same information. Finally, the pt-RGBP-Depth which pre-trains on RGB and uses both modalities for fine tuning reaches the best results. Qualitative results are presented in Figure 4. RGBP-Depth shows sharper edges and fuller structures.

5 Conclusion and Future Work

In this work we examined the advantages of polarimetric imaging. We analyzed the extent to which two perception tasks can be improved when polarization information is used along with standard RGB images.

Our data collection methodology consisted of an RGB-polarimetric camera, a lidar and a GNSS / IMU system. We showed that this setup makes it possible to gather a large database that can serve many perception tasks since all the modalities are aligned and synchronized.

Our evaluation of the free space and monocular depth estimation showed that by using RGB and polarization information we could improve the results as compared to using RGB information alone. Interestingly, this improvement was achieved with only minor architectural changes.

Future work will focus on extending the models to cope with situations where the noon-time constraint does not hold.

Acknowledgments

We thank Shaked Magali, Tal Piterman, Tony Eyal Naim, Eldar Riklis and Gilad Oshri for their help building the data acquisition setup. We also thank Tzvi Philipp, Noa Garnett and Emanuel Mordechai for their useful ideas.

References

[1] Gary A Atkinson and Edwin R Hancock. Recovery of surface orientation from diffuse polarization. *IEEE Trans. Image Process.*, 15(6):1653–1664, 2006.

[2] Yunhao Ba, Alex Gilbert, Franklin Wang, Jinfa Yang, Rui Chen, Yiqin Wang, Lei Yan, Boxin Shi, and Achuta Kadambi. Deep shape from polarization. In *ECCV*, pages 554–571. Springer, 2020.

[3] Marc Blanchon, Désiré Sidibé, Olivier Morel, Ralph Seulin, Daniel Braun, and Fabrice Meriaudeau. P2d: a self-supervised method for depth estimation from polarimetry. In *ICPR*, pages 7357–7364. IEEE, 2021.

[4] Rachel Blin, Samia Ainouz, Stéphane Canu, and Fabrice Meriaudeau. A new multimodal rgb and polarimetric image dataset for road scenes analysis. In *CVPRW*, pages 216–217, 2020.

[5] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[6] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11621–11631, 2020.

[7] Yicong Chang, Feng Xue, Fei Sheng, Wenteng Liang, and Anlong Ming. Fast road segmentation via uncertainty-aware symmetric network. In *ICRA*. IEEE, 2022.

[8] Zhe Chen, Jing Zhang, and Dacheng Tao. Progressive lidar adaptation for road detection. *IEEE/CAA Journal of Automatica Sinica*, 6(3):693–702, 2019.

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016.

[10] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *NeurIPS*, 27, 2014.

[11] Rui Fan, Hengli Wang, Peide Cai, and Ming Liu. Sne-roadseg: Incorporating surface normal information into semantic segmentation for accurate freespace detection. In *ECCV*, pages 340–356. Springer, 2020.

[12] Jannik Fritsch, Tobias Kuehnl, and Andreas Geiger. A new performance measure and evaluation benchmark for road detection algorithms. In *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, pages 1693–1700. IEEE, 2013.

[13] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, pages 2002–2011, 2018.

[14] Noa Garnett, Rafi Cohen, Tomer Pe'er, Roee Lahav, and Dan Levi. 3d-lanenet: end-to-end 3d multiple lane detection. In *ICCV*, pages 2921–2930, 2019.

[15] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11): 1231–1237, 2013.

[16] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, pages 3828–3838, 2019.

[17] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *CVPR*, pages 2485–2494, 2020.

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[19] Yevhen Kuznietsov, Jorg Stuckler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *CVPR*, pages 6647–6655, 2017.

[20] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[21] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553): 436–444, 2015.

[22] Chenyang Lei, Chenyang Qi, Jiaxin Xie, Na Fan, Vladlen Koltun, and Qifeng Chen. Shape from polarization for complex scenes in the wild. In *CVPR*, pages 12632–12641, 2022.

[23] Dan Levi, Noa Garnett, Ethan Fetaya, and Israel Herzlyia. Stixelnet: A deep convolutional network for obstacle detection and road segmentation. In *BMVC*, volume 1, page 4, 2015.

[24] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation fromÂămulti-camera images viaÂăspatiotemporal transformers. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *ECCV*, pages 1–18, 2022.

[25] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022.

[26] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016.

[27] Davy Neven, Bert De Brabandere, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Towards end-to-end lane detection: an instance segmentation approach. In *2018 IEEE intelligent vehicles symposium (IV)*, pages 286–291. IEEE, 2018.

[28] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016.

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015.

[30] Hualian Sheng, Sijia Cai, Na Zhao, Bing Deng, Jianqiang Huang, Xian-Sheng Hua, Min-Jian Zhao, and Gim Hee Lee. Rethinking iou-based optimization for single-stage 3d object detection. In *ECCV*, pages 544–561. Springer, 2022.

[31] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, pages 2446–2454, 2020.

[32] Hengli Wang, Rui Fan, Peide Cai, and Ming Liu. Sne-roadseg+: Rethinking depth-normal translation and deep supervision for freespace detection. In *IROS*, pages 1140–1145. IEEE, 2021.

[33] Honghui Yang, Zili Liu, Xiaopei Wu, Wenxiao Wang, Wei Qian, Xiaofei He, and Deng Cai. Graph r-cnn: Towards accurate 3d object detection with semantic-decorated local graph. In *ECCV*, pages 662–679. Springer, 2022.

[34] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, pages 2636–2645, 2020.

[35] Junsheng Zhou, Yuwang Wang, Kaihuai Qin, and Wenjun Zeng. Unsupervised high-resolution depth learning from videos with dual networks. In *ICCV*, pages 6872–6881, 2019.

[36] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, pages 1851–1858, 2017.