CANDIDATES DECLARATION

I hereby certify that the work, which is being presented in the report, entitled **3D Point Cloud for Object Segmentation, Localization and Recognition**, in partial fulfillment of the requirement for the award of the Degree of **Integrated Masters of Technology** and submitted to the institution is an authentic record of my own work carried out during the period *May 2022* to *September 2022* under the supervision of **Dr Sunil Kumar**. I have also cited the references about the text(s)/figure(s)/table(s) from where they have been taken.

Date: Signatures of the Candidates

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Date: Signatures of the Research Supervisor

ABSTRACT

The detection of pedestrians cars and other object in 3 dimension is an important method in autonomous driving. Though Li-DAR sensors can provide accurate 3D point cloud estimates of the environment, they are very expensive for practical use. The new existing methods used expensive Lidar sensor to calculate depth information. We have used stereo images to create Pseudo Lidar which is relatively cheaper as compared to Lidar. Neural networks are used by Pseudo Lidar for 3D depth estimation for 3D object detection. It is done by converting 2D depth map outputs to 3D point cloud . In this paper we have used Pseudo Lidar for depth estimation for object segmentation, localization and estimation. We have used Point RCNN techniques which improves the PL consistently across all benchmarks. In this paper, we have used SDN and GDC techniques to create a accurate and consistent model. We then have used AVOD, P-RCNN techniques to train our model.

ACKNOWLEDGEMENTS

Dear **Dr. Sunil Kumar**, Thank you for the great skills and knowledge you've imparted to me. I greatly appreciate your contribution to my success. I was lucky to have you as a teacher, and then I was even more fortunate to have you as a mentor. Thank you for everything that you've invested in me. As a teacher, you inspired me, and as a mentor, you motivated and invested in me. I can never repay you for everything I've gained from your instruction. My mentor always answered myriad of my doubts with smiling graciousness and prodigious patience, never letting me feel that I am novices by always lending an ear to my views, appreciating and improving them and by giving me a free hand in my report. It's only because of his overwhelming interest and helpful attitude, the present work has attained the stage it has.

Dated: Akash Verma

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

PL	Pseudo Lidar
RCNN	Region based Convolutional Neural Network
Lidar	Light Detection and Ranging
SDN	Stereo Depth Network
GDC	Graph Based Depth Correction
AVOD	Aggregate View Object Detection
ROIS	Region of Interests

CHAPTER 1

Introduction

1.1 Context

The detection of cars, pedestrians and other object in 3 dimension is an important method in autonomous driving. Using Li-DAR sensors accurate 3D point cloud assessments of the environment can be provided by sensors,but they are very expensive for practical use. The new existing methods used expensive Lidar sensor to calculate depth information. We have used stereo images to create Pseudo Lidar which is relatively cheaper as compared to Lidar.

1.2 Problem

3D object detection is one of the major components of autonomous driving. A self driving car must accurately segment, localize and detect objects. These objects can be cars, pedestrians walking on the road. Our model is reliable and efficient in order to maintain path safety and avoid collision. To this end, Lidar(Light Detection and Ranging) only provides precise 3D point cloud of the surrounding environment. But the major drawback of Lidar is it is very expensive. It is not viable to employ a 64 beam model since it costs more than a car does.

One solution is to use sensors like stereo cameras. Stereo camera act as a Pseudo - Lidar, which converts stereo images into 3D point cloud. Taking advantage of the pseudo-Lidar state of art algorithms [**?**,]]24 [**?**,]]2 [**?**,]]4. Our pseudo-Lidar achieves the accuracy of 34.1% and 42.4% in KITTI leaderboard.

Pseudo lidar consists of two system:

- a depth estimator that was learned using the stereo pictures
- object detector trained upon point cloud data created from the depth estimates.

To achieve the ultimate goal of maximising detection accuracy, the two goals should be perfectly linked. We discovered that the stereo approach identified the items and assessed the depth of the object from a very great distance or very close distance. We de-bias these depth estimates to precisely localise distant objects in 3D without incurring astronomical expenditures. The disparity error, for instance, is 10 cm in depth for an item that is 5 metres distant and 5.8 m in depth for an object that is 50 metres away. The stereo network architecture and loss function for direct depth estimation are thus what we have suggested. On the grid of depth, we built the cost volume. This was previously applied on the disparity values. This enables loss function and the 3D convolutions to perform on the correct scale of depth prediction and estimation.

1.3 Motivation

We are imagining a world where all the car will be autonomous. There will be no need for the driver and every car will be operating on its own. For this purpose we have created a model for object segmentation, localization and recognition. Reliable and accurate 3D detection of object is critical task for safe autonomous driving. There is still the performance gap between stereo and Lidar based method. We aim to make our results as close as lidar based method for object detection. There are many drawbacks of stereo-based method as their is much high variance in depth estimation accuracy. Our model aim to reduce the high variance so that we can get a better, reliable and accurate depth estimation. This would overall increase models efficiency and accuracy.

1.4 Objectives

The main objectives is to create a model with better accuracy and reliability.

- Increase the efficiency of depth estimation

- Increase the quality of depth map so that we can identify even the occluded objects.
- Create a sparse point cloud representation of stereo image.
- Create a 3D box around the object and recognize it.

CHAPTER 2

Literature review

2.1 Background/Key Related Research

Mostly 3D Lidar Point clouds are used for the 3D object detection works ,[**?**,]]21[**?**,]]22[**?**,]]23. There are two ways to process point cloud.

- mostly by applying PointNet and 3D convolution over neighbours to an unordered point cloud in 3D.
- applied to 3D tensor data, which is created by discretizing point cloud data into some predetermined grids.

Beside from stereo image Lidar based models, there are some works done in 2D monofocal image, but most of them are not quite effective.
Chang et al [reference] has proposed pyramid pooling model followed by stacked 3D CNN.

- **3D object detection based on Lidar**- Lidar give the accurate point clouds of object depth and shape. They use voxelization [reference],or PointNet[**?**,]]20,[**?**,]]3 or either both of them [**?**,]]21,[**?**,]]22,[**?**,]]23. We can depict the similar performance using stereo-based detectors.

- **3D object detection based on Stereo**- In order to produce a 3D bounding box and align the sparse point cloud with it, stereo RCNN[**?**,]]7 utilised 2D input from the left and right images. To obtain a pair of RoIs,[**?**,]]24 has employed TLNet, the anchor box to stereo pictures. A better depth map was obtained using RT3DStereo, which combines disparity and semantic data. Citation technique OC-Stereo was utilised in [**?**,]]6 to associate 2D bounding box area. Our method suggests creating pseudo-lidar from stereo pictures and then identifying an object using a point cloud representation of the object. In order to overcome the problem of stereo 3D object detection, the most recent state-of-the-art technique, DSGN [**?**,]]24, suggests using a differentiable 3D volumetric representation of the environment.

64 beam lidar system is very expensive.[**?**,]]4 have used 4 beam lidar system which is comparatively less costly and give very good results in measuring distance from the object.

2.2 Analysis

There are many state of the art models[**?**,]]4,[**?**,]]24 for creating point cloud from depth maps. We created better depth map using the depth cost volume instead of disparity cost volume.

2.3 Research gaps

There are couple of research gaps we could find. These gaps were based on the construction of depth map from left and right image.

- **Depth Loss:** We propose two changes so that we can find direct depth estimation. Instead of disparity loss we consider depth loss. This helps in creating strong emphasis on tiny depth error of close objects.
- **Depth Cost Volume:** We wanted to facilitate more on depth learning rather than disparity. We need to create depth estimation pipeline where convolution is applied within the 4D disparity cost volume.

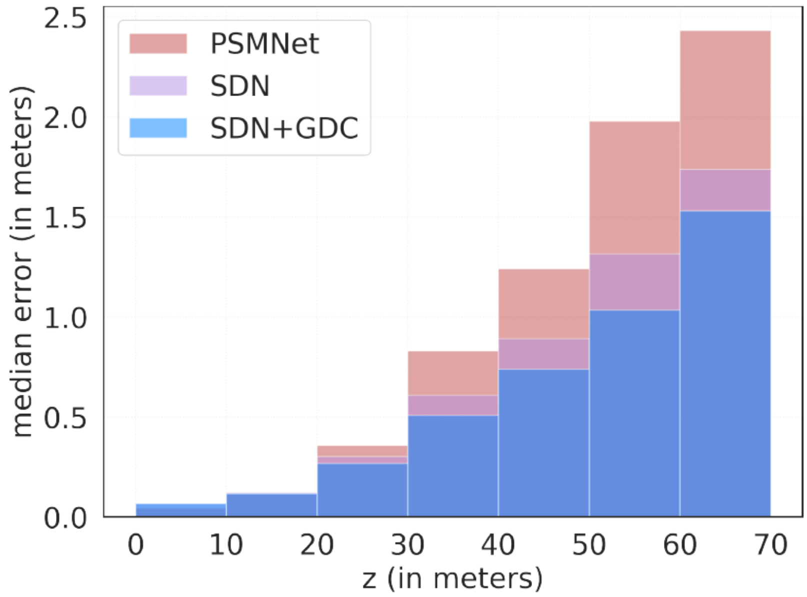


Figure 2.1: Depth Estimation vs Error

Figure 2 shows that as the distance z increases the median error also increases. Our SDN + GDC model have performed well as compared to simple Stereo depth network and stereo disparity network.

CHAPTER 3

Methodology

This section introduces the hypothesis and the analytical validation of the proposed solution.

3.1 Stereo camera working

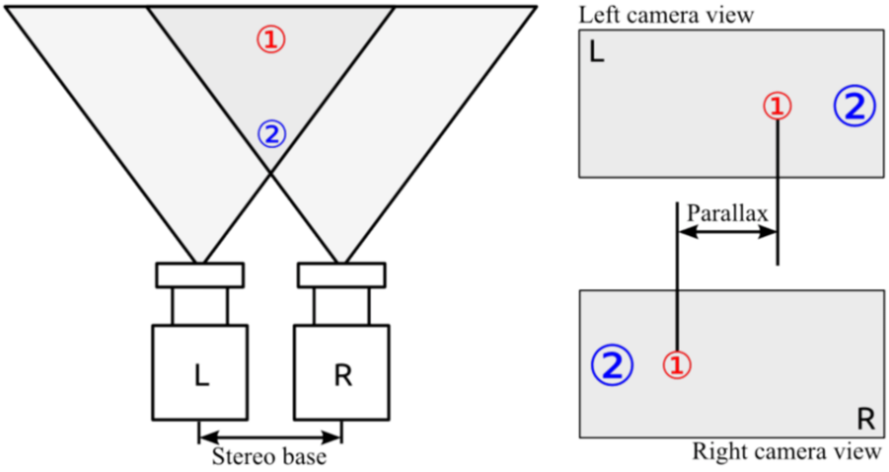


Figure 3.1: Depth Estimation vs Error

A stereo camera works similar to our eyes to give use accurate, real time depth perception. The brain uses this binocular disparity to extract depth information from the two-dimensional retinal images which are known as stereo-psis. It is achieved by using two sensors a set distance apart to triangulate similar pixels from both 2D planes. This Geometric approach is called Triangulation.

3.2 Disparity

We notice that points in the right hand image are shifted to the the left and this shift,i.e the horizontal displacement is referred to as disparity. Shift is less for points that are further away from the camera. This is a very simple relation between the disparity is inversely proportional to the distance(Z) from the object. Similarly distance (Z) is directly proportional to the focal length of the camera used. So if we know the disparity and the focal length of the camera used , we can easily estimate Z. This is the fundamental principle of robotic stereo vision. add an image

Lidar framework was first proposed by [**?**,]]1. Pseudo-LiDAR is used to predict the depth Z(u, v) of every image pixel (u, v) or that we can get accurate depth estimation. A pixel (u, v) will be converted to (x, y, z) in 3D by:

$$z = Z(u, v), x = \frac{(u - c_u) * z}{f_u}, y = \frac{(v - c_v) * z}{f_v} \qquad (1)$$

where f_u and f_v are horizontal and vertical focal length and c_u and c_v is the camera center. This is how depth map is created which further contributes in creation of point cloud. Using the equation 1 we can derive the depth map Z using the following transform,

$$Z(u, v) = \frac{f_u * b}{D(u, v)} \qquad (2)$$

where f_u is horizontal focal length and b is the horizontal offset(i.e. baseline) and D is the disparity map.

3.3 Stereo Depth Network(SDN)

A stereo network is designed and trained to minimize disparity error, which can overemphasize nearby objects with less depth and therefore does not perform well in estimating depths for distant objects.

$$Z \propto \frac{1}{D} \quad \rightarrow \quad \delta \propto \frac{1}{D^2}\delta D \quad \rightarrow \quad \delta Z \propto Z^2 \delta D \qquad (3)$$

The equation 3 is obtained by differentiating Z(D) w.r.t. D. The disparity error of a 2 meter away object is 5cm error in depth. On the other hand the disparity error for 50m away object is 6 meters depth error.

3.4 Proposed Changes

We have proposed following changes:

- **Depth Loss:** We propose two changes so that we can find direct depth estimation. Instead of disparity loss we consider depth loss. This helps in creating strong emphasis on tiny depth error of close objects.
- **Depth Cost Volume:** We wanted to concentrate more on depth learning over the disparity. We need to create depth estimation pipeline where convolution is applied in the 4D disparity cost volume.

3 dimension convolutions are used within 4D disparity cost volume. The same kernel is applied on the entire cost volume which is the main cause of the error. It is assumed implicitly that the effect of a convolution is homogeneous throughout. This is clearly violated in the equation 3. As an example it is easy to smooth two near by pixels with disparity 97 and 98, whereas applying kernel for two pixel with disparity 7 and 8 results in moving 3d points by more than 15m or more.

In this research, we propose to construct the depth cost volume C_{depth}, where C_{depth}(u, v, z) will encode features describing how likely the depth Z(u, v) of pixel (u, v) is z. Then, subsequent 3D convolutions will work on the depth grid rather than the disparity, impacting neighbouring depths uniformly regardless of where they are. The pixel depth is then predicted using the final 3D tensor $Sdepth$ in a manner similar to Equation 3.

$$Z(u,v) = \sum_z \text{softmax}(-(S_{depth})(u,v,z))*z \qquad [4]$$

We construct the new depth volume, C_{depth}, based on the intuition that C_{depth}(u, v, z) and C_{disp} $(u, v, \frac{f_u * b}{z})$ should lead to equivalent "cost". Uptill now we applied a bilinear interpolation to construct C_{depth} from C_{disp} using depth to disparity transform in equation 2.

3.5 Depth Correction

SDN greatly improves depth estimations and accurately and precisely renders the object contours. Fundamental Limitation of stereo - the discrete nature of the pixel while the depth is continuous. We therefore use cheap 4 beam lidar. This is used for creating sparse point cloud.

We introduce a graph-based depth correction technique that combines sparse accurate Lidar data with dense stereo depth. Two point clouds from Lidar and pseudo-Lidar are the inputs that we use. Using a directed KNN graph in the PL point cloud, which links each 3D point with its KNN with the proper weight, we first define the local forms.

3.6 Sparse Point Cloud

We used the velodyne data set(64 beam Lidar Spec). In order to convert 64 beam Lidar SPEC to 4 beam Lidar by quantizing the vertical angles into 64 levels with an interval of 0.4 deg

3.6.1 Conclusion

Using GDC and SDN improves the accuracy and preciseness of our model. It is observed that 4 beam Lidar performs well on locating distant objects but cannot detect close object. Further our GDC method combines the merits of two signals, to achieve better performance rather than using them alone.

CHAPTER 4

Experiments and results

This section discusses the various experiments pertaining to the proposed hypothesis and their findings.

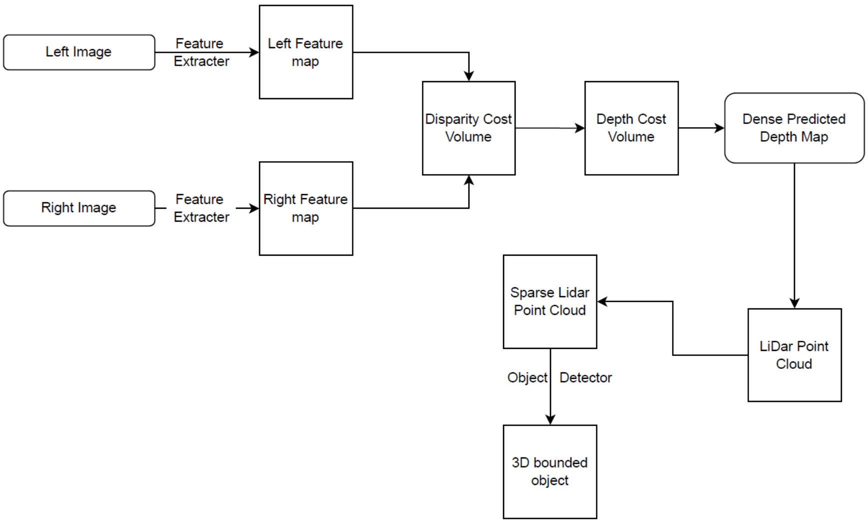


Figure 4.1: Model Pipeline for 3D object detection

4.1 Experiment 1

4.1.1 Dataset Creation

Our model was trained and assessed using the KITTI dataset. KITTI offers the equivalent right picture, 64-beam Velodyne LiDAR point cloud, camera

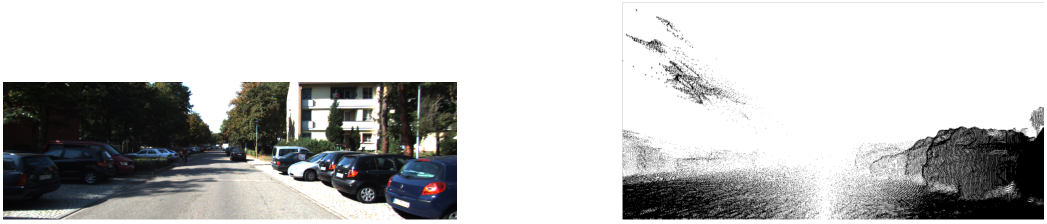
calibration matrices, and bounding boxes for each (left) image. 15000 photos total, broken down into 2 groups (automobiles and pedestrians/cyclists) make up the dataset. The data was split into training and testing groups. Approximately 7481 photos make up training, whereas 7418 images make up testing.
The 7418 testing picture was divided into 3712 training images and 3769 evaluating images. For objects with 2D box heights less than or occlusion/truncation levels more than specific thresholds, KITTI specifies the easy, moderate, and hard settings, respectively.

4.1.2 Stereo depth network

Our stereo depth estimation network is built on PSMNET (Chang Chen, 2018). (SDN). We pre-train SDN using the synthetic Scene Flow dataset and fine-tune it using the 3,712 training pictures of KITTI, as per Wang et al. We constructed depth ground truth by projecting Lidar points onto the photos after acquiring the depth maps of the photographs.

4.1.3 Point Cloud

Projecting the associated LiDAR points onto the photos allows us to determine the depth ground truth. For contrast, we train a PSMNET in the same manner to reduce disparity error.



(a) Real image

(b) Point Cloud representation

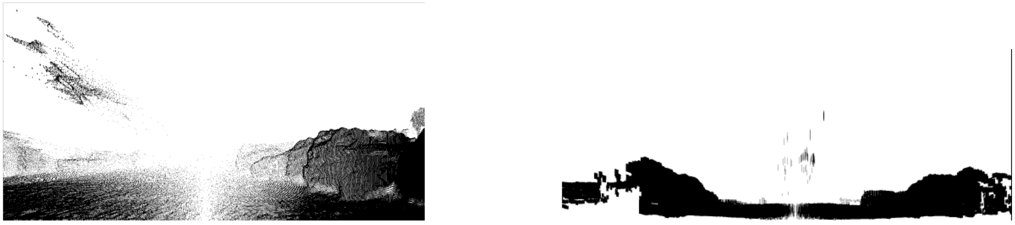
Figure 4.3: Conversion of Real image to Point cloud

4.1.4 3D Object Detection

Three algorithms—AVOD, PIXOR, and P-RCNN—were employed. They all made use of data from Lidar point clouds. More than 3000 stereo pictures were used to train our model from scratch.

4.1.5 Sparser Lidar

Using velodyne data, we triggered sparser Lidar. The vertical angles were initially quantized into 64 levels with 0.4 degree intervals. To replicate the sparser signal, we maintain the spots that have fallen in the beam subset.



(a) Point Cloud representation

(b) Sparse Point Cloud Representation

Figure 4.4: Conversion of Point Cloud to Sparse Point cloud

4.1.6 Results

Our Psuedo Lidar(GDC+SDN) perfromed very well over simple SDN. We perfromed very close to Lidar's performance. Also Psuedo Lidar with GDC involved performed very close to real Lidar sensor.We mainly experimented with P-RCNN. AVOD and PIXOR performed similarly and showed the similar trends.

Table 4.1: **Results on car category on test set.** Results show BEV/3D

Input signal	**Easy**	**Moderate**	**Hard**
PL (SDN)	76.6 / 60.5	51.3 / 42.9	52.5 / 34.4
PL++ (SDN + GDC)	**85.8 / 66.8**	74.5 / 56.8	**67.6 / 51.2**
LiDAR	90.5 / 86.9	87.7 / 75.8	80.1 / 70.3

Table 4.2: **Results on pedestrian(top) and cyclist(bottom) category on test set.** Results show BEV/3D

Stereo depth	**Easy**	**Moderate**	**Hard**
PSMNET	43.3 / 32.8	35.8 / 26.3	31.2 / 25.6
SDN	42.3 / 41.8	41.1 / 35.9	36.5 / 29.8
SDN + GDC	**64.8 / 51.6**	**55.8 / 43.2**	**45.9 / 37.9**
PSMNET	46.6 / 42.3	30.1 / 26.2	28.1 / 25.0
SDN	46.7 / 45.6	31.6 / 28.7	29.6 / 28.5
SDN + GDC	**62.6 / 63.9**	**42.9 / 41.9**	**43.8 / 37.5**

4.2 Overall conclusion

In this paper we have improved the 3D object detection without using expensive Lidar. We used the novel approach to learn the depth directly instead of disparity estimates using Stereo depth Network. Secondly, We have used cheap sparse 4 beam Lidar instead of high end expensive 64 beam lidar to produce to point cloud. Our system cost less than Rs 1 lakh while on the other hand 64 beam lidar cost over Rs 50 lakhs alone. Our pseudo Lidar cost with both SDN and GDC cost very less and give performance close to real Lidar.

REFERENCES

24

[1] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In CVPR, 2018

[2] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2017, pp. 1907–1915.

[3] S. Shi, X. Wang, and H. Li, "Pointrcnn: 3d object proposal generation and detection from point cloud," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2019, pp. 770–779.

[4] 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," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2019, pp. 8445–8453.

[5] 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," in *International Conference on Learning Representations* (ICLR), 2020.

[6] A. D. Pon, J. Ku, C. Li, and S. L. Waslander, "Object-centric stereo matching for 3d object detection," *Proceedings of the IEEE International Conference on Robotics and Automation* (ICRA), 2020.

[7] P. Li, X. Chen, and S. Shen, "Stereo r-cnn based 3d object detection for autonomous driving," in *Proceedings of the IEEE Con-*

ference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7644– 7652.

[8] X. Wang, W. Yin, T. Kong, Y. Jiang, L. Li, and C. Shen, “Task-aware monocular depth estimation for 3d object detection,” arXiv preprint arXiv:1909.07701, 2019.

[9] J. Ku, A. D. Pon, and S. L. Waslander, “Monocular 3d object detection leveraging accurate proposals and shape reconstruction,” in Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2019, pp. 11 867–11 876

[10] J. Ku, A. D. Pon, and S. L. Waslander, “Monocular 3d object detection leveraging accurate proposals and shape reconstruction,” in Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2019, pp. 11 867–11 876

[11] D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” International journal of computer vision, vol. 47, no. 1-3, pp. 7–42, 2002

[12] E. Tola, V. Lepetit, and P. Fua, “A fast local descriptor for dense matching,” in IEEE conference on computer vision and pattern recognition (CVPR). IEEE, 2008, pp. 1–8.

[13] Y. Ohta and T. Kanade, “Stereo by intra-and inter-scanline search using dynamic programming,” *IEEE Transactions on pattern analysis and machine intelligence*, no. 2, pp. 139–154, 1985.

[14] V. Kolmogorov and R. Zabih, “Computing visual correspondence with occlusions using graph cuts,” in *Proceedings Eighth IEEE International Conference on Computer Vision* (ICCV)., vol. 2. IEEE, 2001, pp. 508–515.

[15] J.-R. Chang and Y.-S. Chen, “Pyramid stereo matching network,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2018, pp. 5410–5418.

[16] F. Zhang, V. Prisacariu, R. Yang, and P. H. Torr, “Ga-net: Guided aggregation net for end-to-end stereo matching,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2019, pp. 185–194

[17] Z. Yin, T. Darrell, and F. Yu, "Hierarchical discrete distribution decomposition for match density estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2019, pp. 6044–6053.

[18] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2018, pp. 4490– 4499.

[19] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," Sensors, vol. 18, no. 10, p. 3337, 2018.

[20] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2018, pp. 918–927.

[21] Yilun Chen, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Fast point r-cnn. In ICCV, 2019.

[22] Xinxin Du, Marcelo H Ang Jr, Sertac Karaman, and Daniela Rus. A general pipeline for 3d detection of vehicles. In ICRA, 2018.

[23] Martin Engelcke, Dushyant Rao, Dominic Zeng Wang, Chi Hay Tong, and Ingmar Posner. Vote3deep: Fast object detection in 3d point clouds using efficient convolutional neural networks. In ICRA, 2017.

[24] Y. Chen, S. Liu, X. Shen, and J. Jia, "Dsgn: Deep stereo geometry network for 3d object detection," *arXiv preprint arXiv:2001.03398*, 2020.

REFERENCES

xx