UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Object Detection and Localization in 2D & 3D Environment

by Zhihao Cui

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Forest Zhu

University of Technology Sydney
Faculty of Engineering and IT

26th August 2021

Certificate of Authorship/Originality

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, *zhihao cui* declare that this thesis, is submitted in fulfilment of the requirements for the award of *Doctor of Philosophy*, in the *FEIT* at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note:
Signature removed prior to publication.

Date: 23/8/2021

ABSTRACT

OBJECT DETECTION AND LOCALIZATION IN 2D & 3D ENVIRONMENT

Computer vision is a science that studies how to make machines "see." It refers to utilizing vision sensors and computers to identify, locate, and track objects. Under this topic, this thesis proposed three frameworks to improve 2D and 3D object detection and localization performance. In the first 3D object detection framework, we investigated the bilateral convolution layers' feasibility to alternate the widely used point cloud voxelization process. The second framework explored the voxel-wise and point-wise proposal fusions method to improve 3D object detection performance. For the 2D instance segmentation, the framework formed an NMS-free and anchor-free detector designed explicitly for the eye-to-hand robotic system.

In existing works, most of the state-of-the-art 3D object detection approaches are based on the point clouds' voxelization method to sample the point cloud into a subdivide voxel space. Although it provides an efficient way to process point cloud data, its lack of feature relationship on voxel-level limits the model's detection accuracy. Furthermore, the voxel sizes hyperparameters tuning increased the model complexity, resulting in a fluctuated model performance. To this end, we aim to simplify the process by re-projecting the point cloud data onto a lattice hyper-plane that saves point cloud processing time while maintaining the model accuracy. The proposed framework Bilateral Lattice Point Network (BLPNet) is provided in chapter three.

In the second framework, Point and Voxel Fusion Net (PVF-Net) is proposed to further push the 3D object detection performance forward. In two-stage approaches, increasing the first stage proposals recall rate positively influences the model final prediction performance. Therefore, in the PVF-Net, we proposed a twofold proposal fusion architecture to extract and fuse the voxel-level and point-level features of the point clouds. The model details are in chapter four, mainly consisting of two novel modules: the Twofold Proposal Fusion (TPF) module and the ROI Deep Fusion (RDF) module.

Lastly, it is well-known that 3D and 2D sensors jointly depict the real world. In chapter five, 2D object detection will become the next goal for improvement. So far, the existing 2D instance segmentation algorithms developed significantly and reached a saturated performance. However, there is no solid solution for heavy occluded or diagonally arranged objects, especially in the vision-guided robot picking system. To solve the problem above, we proposed a real-time occlusion and oblique friendly instance segmentation framework, terms as Keypoint-Mask, assisting the robotic system to handle the complicated detection scenario.

Contents

1. **Introduction** 1

1.1 Background 1

1.1.1 2D Image-based Vision 3

1.1.2 3D Point Cloud Vision 8

1.1.3 Practical Application in 2D and 3D Perception 12

1.2 Research Problems 12

1.2.1 Learning Robust Features in Lattice Hyperplane 13

1.2.2 Fusion from Multi-Formats of Point Clouds 14

1.2.3 Special Case Occlusion in Instance Segmentation 15

1.3 Thesis Contribution 16

2. **Literature Review** 18

2.1 3D Object Detection 18

2.1.1 3D Two-stage methods 18

2.1.2 3D Single-stage methods 25

2.2 2D Object Detection 28

2.2.1 2D Two-stage methods 29

2.2.2 2D Single-stage methods 31

2.3 2D Instance Segmentation 32

2.3.1 Semantic segmentation 33

2.3.2 Two-stage Instance Segmentation 34

2.3.3 Single-stage Instance Segmentation 36

3. **Bilateral Lattice Point Network** 38

3.1 Introduction 38

3.1.1 Problem Formulation 40

3.1.2 Motivation 41

3.1.3 Contributions 41

3.2 The Proposed Method 44

3.2.1 Bilateral Convolution Layer 44

3.2.2 Lattice to Feature Map Layer ... 45

3.2.3 Region-based Feature Clustering ... 47

3.2.4 Loss Function ... 48

3.3 Implementation Details ... 49

3.3.1 Data preparation ... 49

3.3.2 Feature Learning Network ... 50

3.3.3 Pattern Recognition Backbone ... 51

3.4 Experiments ... 55

3.4.1 Training Schemes ... 55

3.4.2 Results ... 55

3.5 Ablation Study ... 57

3.5.1 Module Performance ... 57

3.5.2 Effects of Feature Extraction BCL Architecture ... 59

3.5.3 Effects of Region-based Clustering Fusion Methods ... 60

3.5.4 Model Efficiency ... 61

3.6 Summary ... 61

4. Point Voxel Fusion Network ... 62

4.1 Introduction ... 62

4.1.1 Problem Formulation ... 64

4.1.2 Motivation ... 65

4.1.3 Contributions ... 66

4.2 The Proposed Method ... 67

4.2.1 Twofold Proposal Fusion Module ... 68

4.2.2 RoI Deep Fusion Module ... 71

4.2.3 Prediction Network ... 73

4.2.4 Loss Function ... 74

4.3 Experiments ... 75

4.3.1 Network Architecture ... 75

4.3.2 Training schemes ... 76

4.3.3 Results ... 77

4.4 Ablation Study ... 80

4.4.1 Effects of RoI Deep Fusion Module 80

4.4.2 Effects of RoI Features Selection 81

4.4.3 Effects of RoI Deep Fusion Module 82

4.4.4 Efficiency of VDS and PSS module 82

4.5 Summary 83

5. Keypoint-Mask 84

5.1 Introduction 84

5.1.1 Problem Formulation 85

5.1.2 Motivation 86

5.1.3 Contributions 88

5.2 The Proposed Method 89

5.2.1 Trapezoid Net 89

5.2.2 Keypoint-Mask Head 92

5.2.3 Loss Function 93

5.3 Experiments 95

5.3.1 Network Architecture 95

5.3.2 Training schemes 95

5.3.3 Results 96

5.4 Ablation Study 102

5.4.1 Effects of Trapezoid Net 102

5.4.2 Effects of other Backbone Architecture 104

5.4.3 Effects of Center Sensitive RoI Pooling 105

5.5 Summary 105

6. Conclusion and Future Work 107

6.1 Conclusion 107

6.2 Future Work 109

6.2.1 Industry-Level Specific Object 109

6.2.2 Industry-Level Arbitrary Object 111

6.2.3 Difficulties and Challenges 111

Bibliography 113

List of Figures

Figure 1.1 : Illustrate the four most basic tasks in computer vision based on 2D-images. 2

Figure 1.2 : Demonstrate the general pipeline of the 2D-based object detection framework. 4

Figure 1.3 : Examples of 3D point cloud visualization in autonomous driving and indoor modeling. 7

Figure 1.4 : Demonstration of different combinations of point clouds. Conv (i) (ii), (iii) represent three set point cloud dataset. W1-W4 represents the kernel weight in the CNN layer. 8

Figure 1.5 : Demonstration of the diagonally arranged thin objects that are side-by-side..15

Figure 2.1 : The summary of 3D target detection methods. 19

Figure 2.2 : The model architecture of the MV3D, three streams proposed to separately process Lidar Bird View, front view, and 2D image. 3D proposals network and Region-based Fusion Network formed the two-stage detector. 20

Figure 2.3 : The model architecture of the PointRCNN, Top module is the first stage. point-wise proposals generation process. And then, the bottom module is the second. proposal refinement stage to predict the final oriented bounding boxes...21

Figure 2.4 : Illustrate the fundamental idea of the 3D-based object detection in the frustum-based method...23

Figure 2.5 : The voxelization based 3D object detection method VoxelNet...26

Figure 2.6 : The summary of 3D target detection methods. 29

Figure 2.7 : Illustrate the anchor-free based box regression method of the RepPoints. 30

Figure 2.8 : The summary of 2D instance segmentation methods..................32

Figure 2.9 : Demonstrated the architecture of FCN. 33

Figure 3.1 : The overall pipeline of the BLPNet. The feature learning network consists of the BCL feature extraction layer, Region-based feature clustering layer, and point-to-feature map layer. The region proposal network is applied as the pattern recognition backbone in the model. Two fully connected branches are employed as the 3D box classification and prediction head...43

Figure 3.2 : Visual demonstration of the splatting, convolving, and slicing steps in the bilateral convolution layer. The blue lattice hyperplane is composed of the uniformed simplex. 45

Figure 3.3 : The point features extracted from BCL layer are clustered by the region-based feature clustering module. Then, through lattice to feature map layer, the clustered features are projected onto a 2D feature map. Tables on the right demonstrate the overall process. $\mathtt{CN}$ indicates the N-th feature in a cluster. $\mathtt{H1,W1}$ indicates the first feature map index. 46

Figure 3.4 : Visual representation of region-based feature clustering. Rectangles with the same color indicating features in the same region. The element-wise sum is applied to points features with the same color and followed by a multi layered perceptron layer. 47

Figure 3.5 : Demonstration of the qualitative results of BLPNet on the KITTI dataset. The model detection is under different road environments, in which the red bounding box indicates prediction and green indicates the ground truth box. 54

Figure 4.1 : Illustrate the overall architecture of our proposed PVF-NET. In the first stage, the point cloud feed into the VDS and PSS, respectively, followed by the proposals integration layer to integrate the voxel-based and point-wise proposals. The RoI Deep Fusion module is applied to extract the point-wise semantic and multi-scale voxel-based features through the voxel-aware pooling and point-aware pooling layers. The proposal-aware pooling layer further fuses those features, and the fully connected branches are used for proposal classification and regression. 67

Figure 4.2 : Demonstration of the generated proposals from the VDS network and PSS network. The VDS proposals indicate in blue boxes, and the PSS propels in red. 70

Figure 4.3 : Illustration of the RoI deep fusion module. It consists of a voxel-aware pooling layer and point-aware pooling layer to effectively extract the multi-level voxel-base features and point-wise features from the first stage. In the end, the proposal-aware fusion layer is employed to select and aggregate the refined proposal..71

Figure 5.1 : Demonstrate the diagonally arranged thin objects that are side-by-side placed on a scene. The right-hand side is the detection results by Keypoint-Mask. And the left-hand side is the detection results by Mask r-cnn...86

Figure 5.2 : The architecture of Keypoint-Mask, where c denotes the out-channel dimension in each block. Center-Sensitive RoI Pooling predicts the segmentation mask inside of each detected bounding box according to the keypoint center. Feature Transition and Residual connection are deployed to boost the performance further. 90

Figure 5.3: Visualization results of the Keypoint-Mask on the MS COCO and factory dataset. 98

Figure 5.4: Visualization results between the Keypoint-Mask and other methods 99

Figure 5.5: Visualization results of the Keypoint-Mask on more industry datasets. 100

Figure 6.1 : The classic scenario of the vision-guided robot arm picking system 108

Figure 6.2 : The industry-level arbitrary object data collection. The objects are scattered in the bin, and graspable area is labeled as green and yellow. 110

Figure 6.3 : Object poses visualization results. 111

List of Tables

Table 3.1 The BLPNet result compares against other state of the art models on KITTI test set. V2 = Voxelization Two-stage; NV2 = Non-Voxelization Two-stage; V1 = Voxelization Single-stage; NV1 = None-Voxelization Single-stage...53

Table 3.2: The Model performances in applying different layer clustering methods in Car class at IoU = 0.7 on the KITTI validation set. 57

Table 3.3: The experiment results of utilizing different lattices scale and feature dimensions on the KITTI validation split set in easy, moderate, and hard mode. The IoU is above 0.7. 58

Table 3.4 : The experiment results in different method of fusion on region-based feature clustering. 59

Table 3.5 : Model inference speed comparisons...60

Table 4.1 : The comparison of the PVF-NET against the state-of-the-arts 3d object models on the KITTI test set..79

Table 4.2: Experiment of the ratio between the VDS and PSS proposals. 80

Table 4.3: Recall of proposals on the KITTI validation set compared to other methods. 81

Table 4.4: The feature selection experiment in the VDS & PSS network. 82

Table 4.5: The experiment results of the different ROI fusion layers on the KITTI validation set to detect the car class. 83

Table 5.1: The comparison of the Keypoint-Mask against the state of arts instance segmentation results on COCO test-dev 2017.........................101

Table 5.2: The comparison of the Keypoint-Mask against the state of arts instance segmentation results on factory dataset. 101

Table 5.3: Model performance between the different layers in backbones Trapezoid S3 and S4. 102

Table 5.4: The combination effects of different connections. 103

Table 5.5: The Keypoint-Mask performance comparison between different backbones. 104

Table 5.6: The effectiveness of center RoI pooling. 105