APPLIED SPATIAL SENSING AND REASONING

CASE STUDIES

Dr TIAN Jing

tianjing@nus.edu.sg

RGB-D datasets

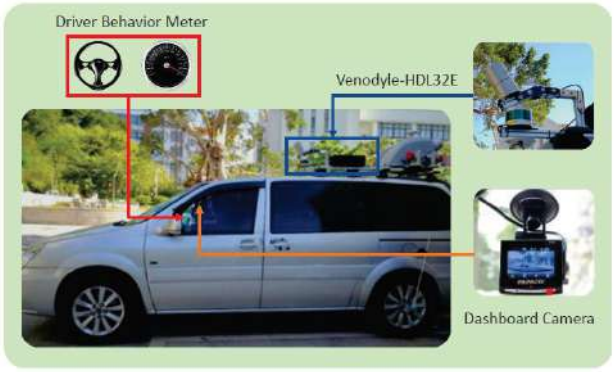
- Seven scene dataset, Microsoft, https://www.microsoft.com/en-us/research/project/rgb-d-dataset-7-scenes/
- NTU RGB+D dataset contains 60 action classes and 56,880 video samples. http://rose1.ntu.edu.sg/Datasets/actionRecognition.asp
- Major datasets are reviewed in “RGB-D Datasets: Past, Present and Future,” https://arxiv.org/pdf/1604.00999.pdf

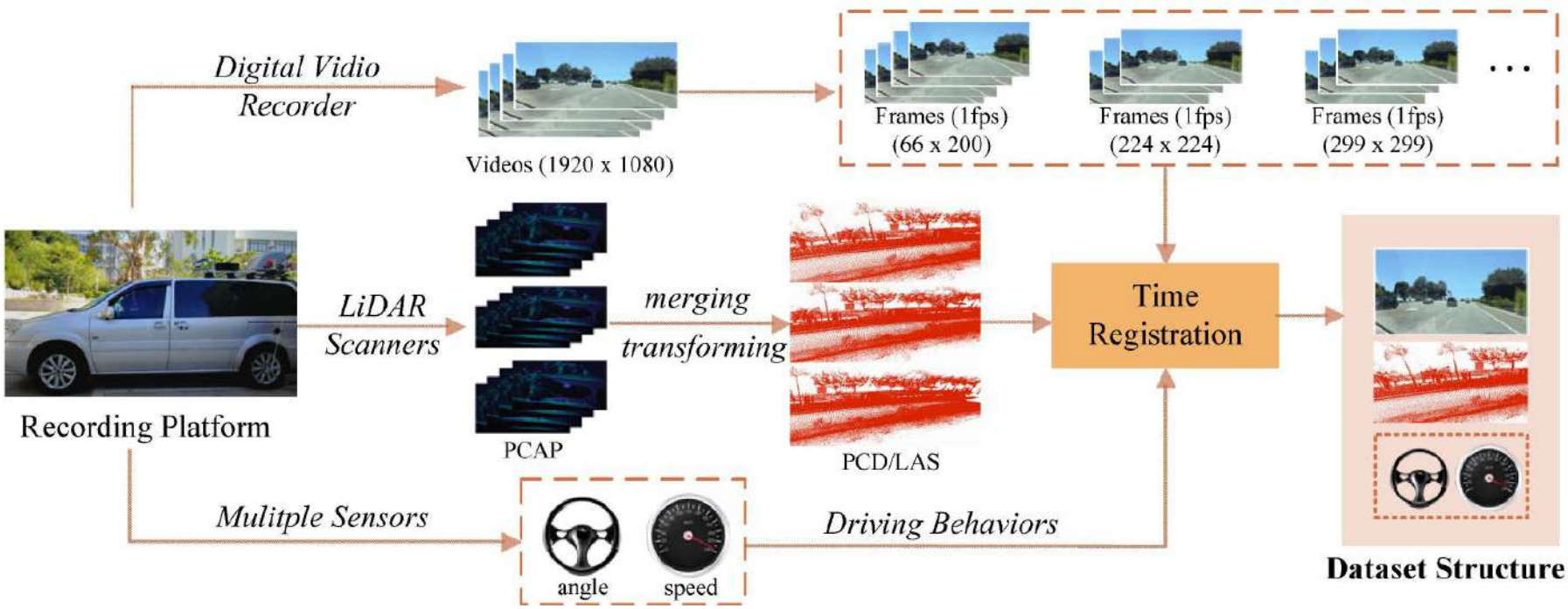
Name	Year	Labeled size	Classes	Resolution	Annotations
Cornell-RGBD-Dataset [1]	2011	550 frames from 52 scenes	17	N/A	per-point clouds annotations
RGB-D Object Dataset [2]	2011	250,000 frames of 300 objects	51	640×480	object annotations
NYU Depth v1 [3]	2011	2347 frames from 64 scenes	13	640×480	dense pixel annotations
NYU Depth v2 [4]	2012	1449 frames from 464 scenes	4/13/40	640×480	dense pixel annotations
SUN3D [5]	2013	415 sequences from 254 scenes	33	640×480	object polygons annotations
Berkeley B3DO [6]	2013	849 frames from 75 scenes	over 50	640×480	bounding box annotations
Kinect RGBD Dataset for Category Modeling [7]	2013	900 frames from 264 scenes	7	640×480	object annotations
SUN RGB-D [8]	2015	10,335 frames across 47 scene classes	37	variable	dense pixel annotations

DBNet

DBNet: A large-scale dataset for driving behavior learning, http://www.dbehavior.net/

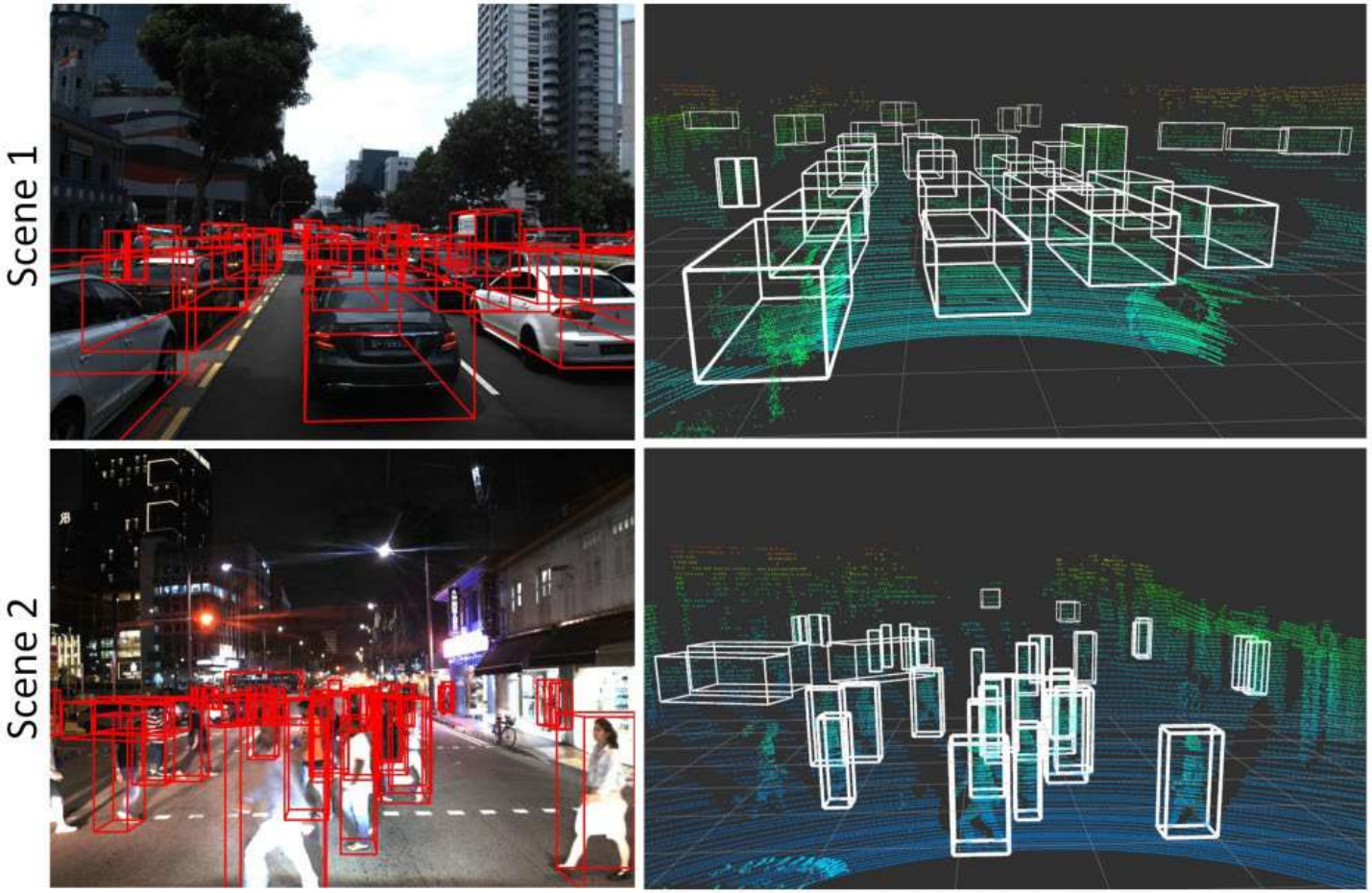
- RGB video
- LiDAR
- Driving behaviors





RGB-LiDAR dataset

- 230K human-labeled 3D object annotations in 39,179 LiDAR point cloud frames and corresponding frontal-facing RGB images.
- Captured at different times (day, night) and weathers (sun, cloud, rain).

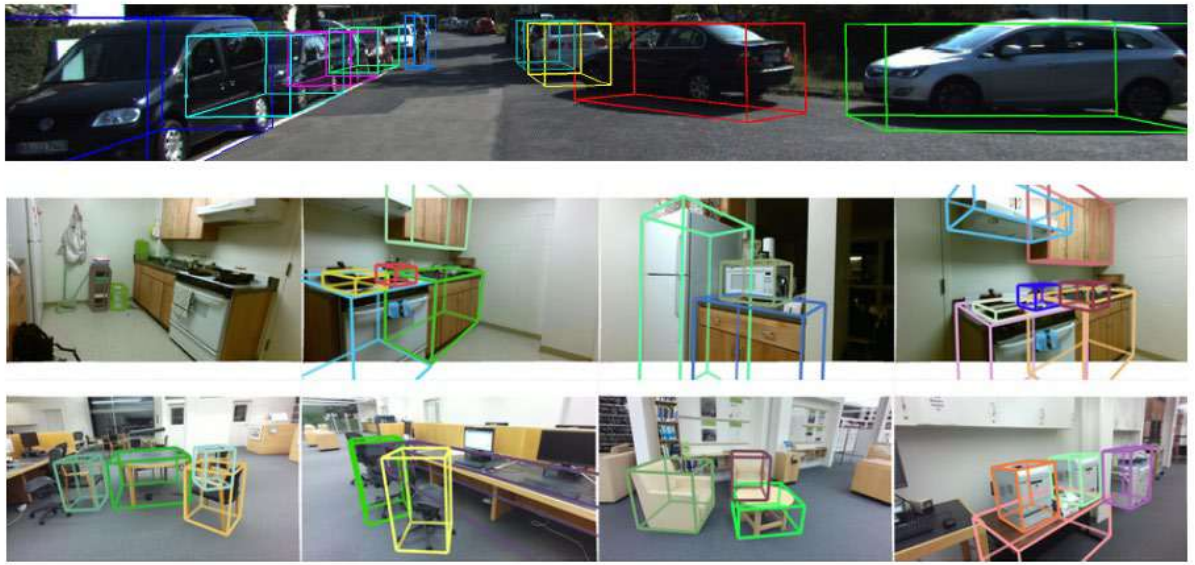


Reference: https://github.com/I2RDL2/ASTAR-3D

Semantic recognition for 3D data

- Input: RGB-D data
- Output: Amodal 3D bounding boxes and semantic class labels for objects in the scene. “D” can be sparse point cloud from LiDAR or dense depth map from indoor depth sensors

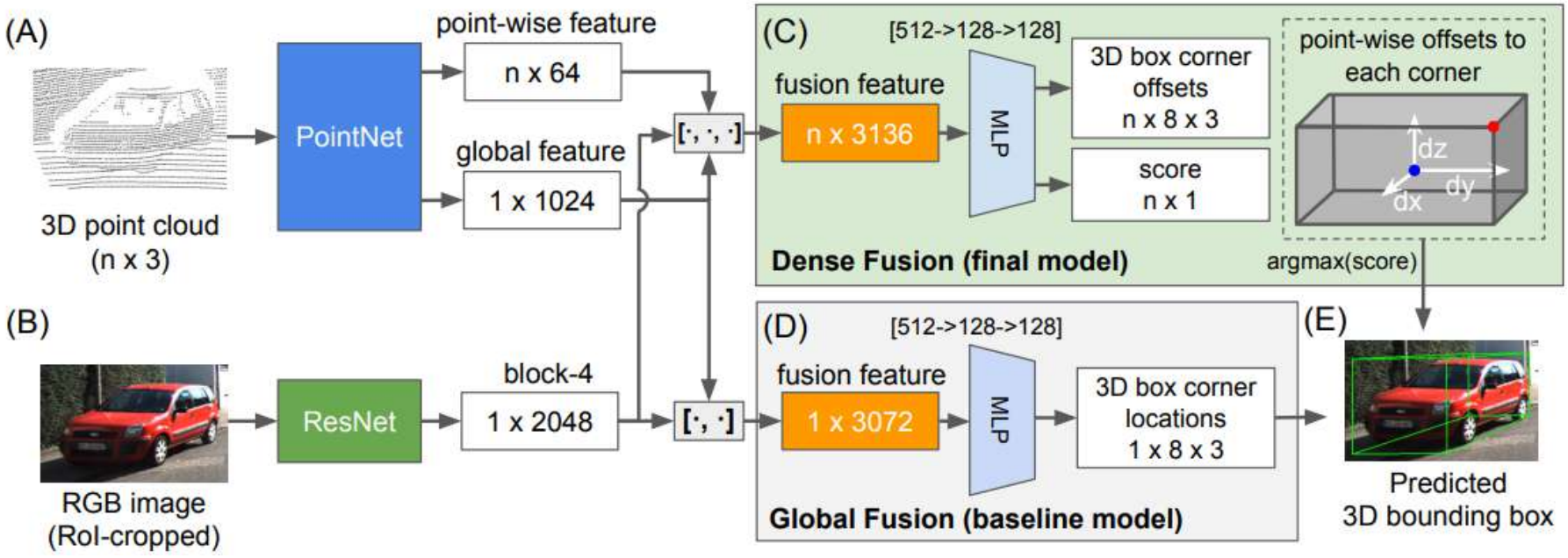
	Advantages	Disadvantages
2D Object Detection	Well established datasets and detection architectures. Usually RGB only input can achieve accurate results in the image plane.	Limited information: lack of object’s pose, occlusion and 3D position information.
3D Object Detection	3D bounding box provides object size and position in world coordinates. These detailed information allows better environment understanding.	Requires depth estimation for precise localization. Extra dimension regression increases model complexity. Scarse 3D labelled datasets.



Reference: E. Arnold, O. Y. Al-Jarrah, M. Dianati, S. Fallah, D. Oxtoby and A. Mouzakitis, "A survey on 3D object detection methods for autonomous driving applications," IEEE Trans. on Intelligent Transportation Systems, http://wrap.warwick.ac.uk/114314/

Point Fusion

PointFusion has two feature extractors: a PointNet variant that processes raw point cloud data, and a CNN that extracts visual features from an input image. We present two fusion network formulations: a vanilla global architecture that directly regresses the box corner locations (D), and a novel dense architecture that predicts the spatial offset of each of the 8 corners relative to each input point, as illustrated in (C): for each input point, the network predicts the spatial offset (white arrows) from a corner (red dot) to the input point (blue), and selects the prediction with the highest score as the final prediction (E).



Reference: D. Xu, et al., PointFusion: Deep Sensor Fusion for 3D Bounding Box Estimation, https://arxiv.org/abs/1711.10871

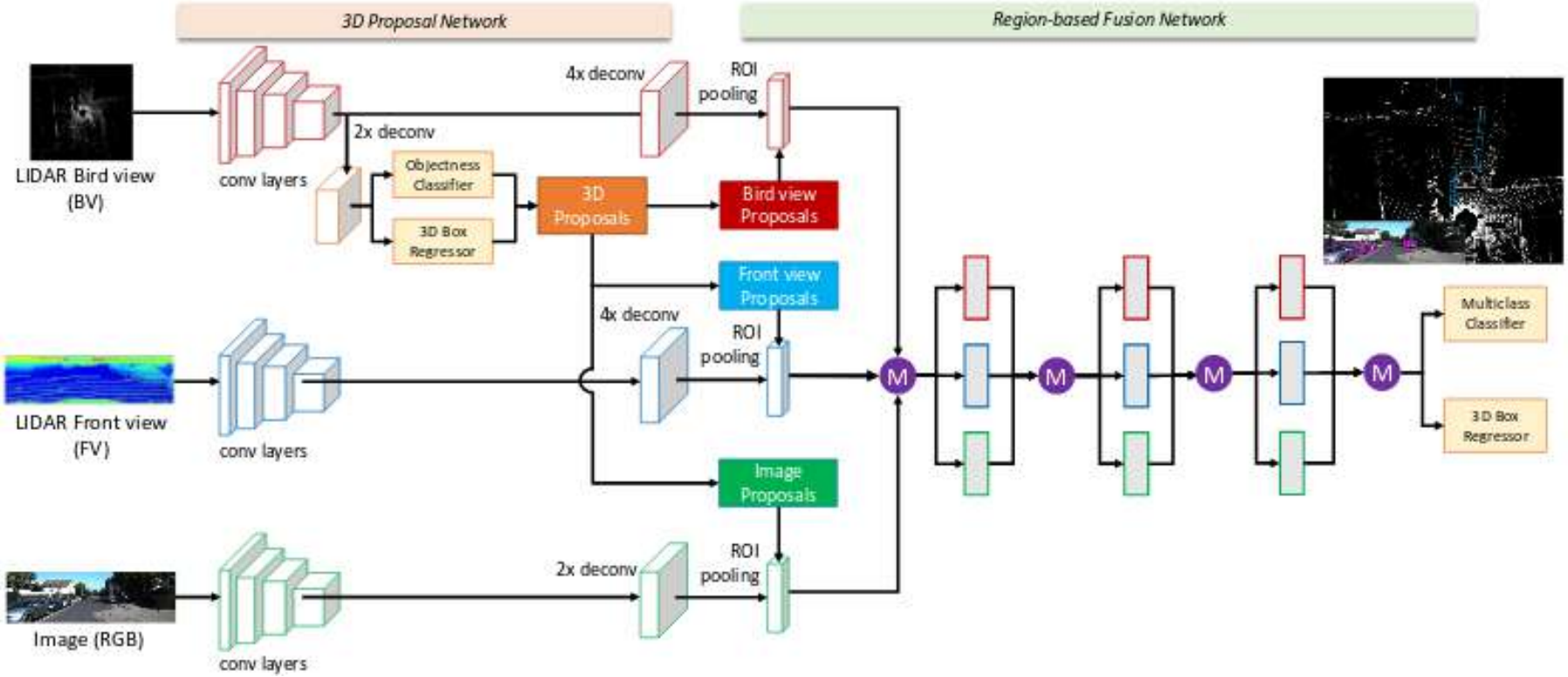
DenseBox

- Divides the point cloud into voxels. So instead of running 2D pixels through a network, we're running 3D voxels.
- Trains an FCN to identify features in the voxel-zed point cloud.
- Upsamples the FCN to produce two output tensors: an objectness tensor, and a bounding box tensor. The bounding box tensor is probably more interesting for perception purposes. It draws a bounding box around cars on the road.

Reference: L. Huang, Y. Yang, Y. Deng, Y. Yu, "DenseBox: Unifying Landmark Localization with End to End Object Detection," https://arxiv.org/abs/1509.04874

MV3D

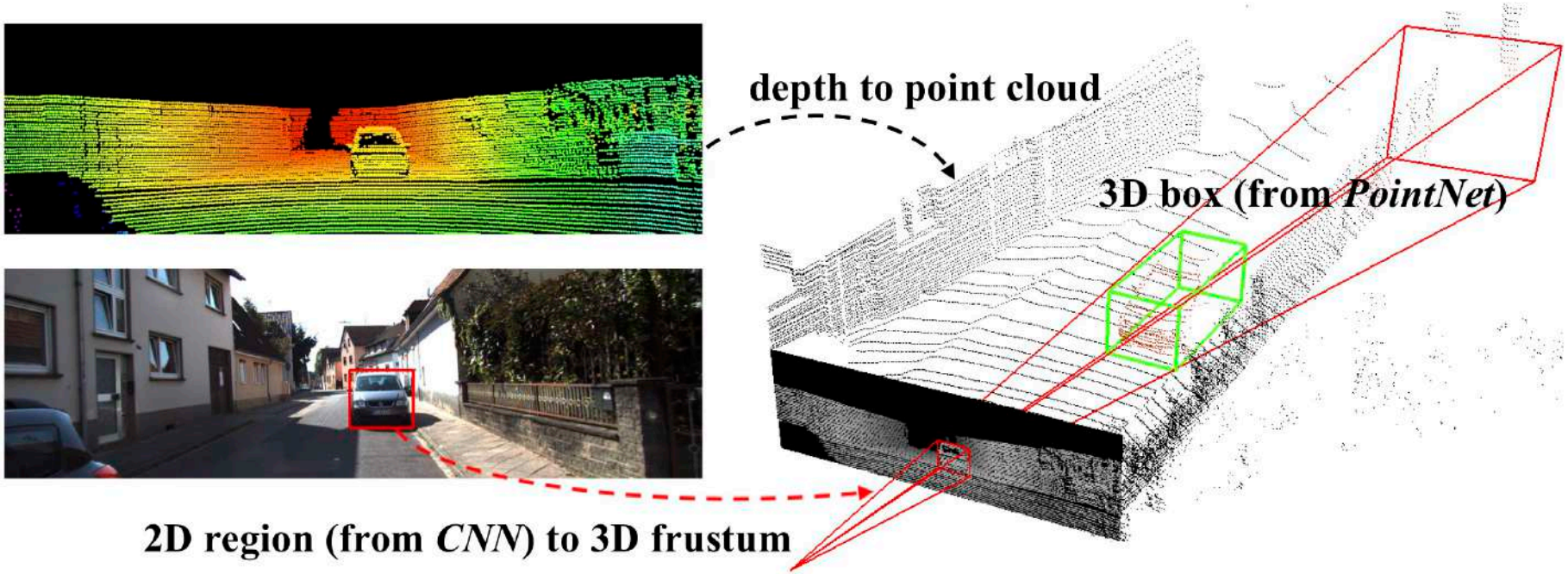
Multi-view V3D (MV3D) simply takes two separate 2D views of the point cloud: one from the front and one from the top (birds' eye). MV3D also uses the 2D camera image associated with each LiDAR scan. That provides three separate 2D images (LiDAR front view, LiDAR top view, camera front view).



Reference: X. Chen, H. Ma, J. Wan, B. Li, T. Xia, "Multi-View 3D Object Detection Network for Autonomous Driving," https://arxiv.org/abs/1611.07759

Frustum PointNet

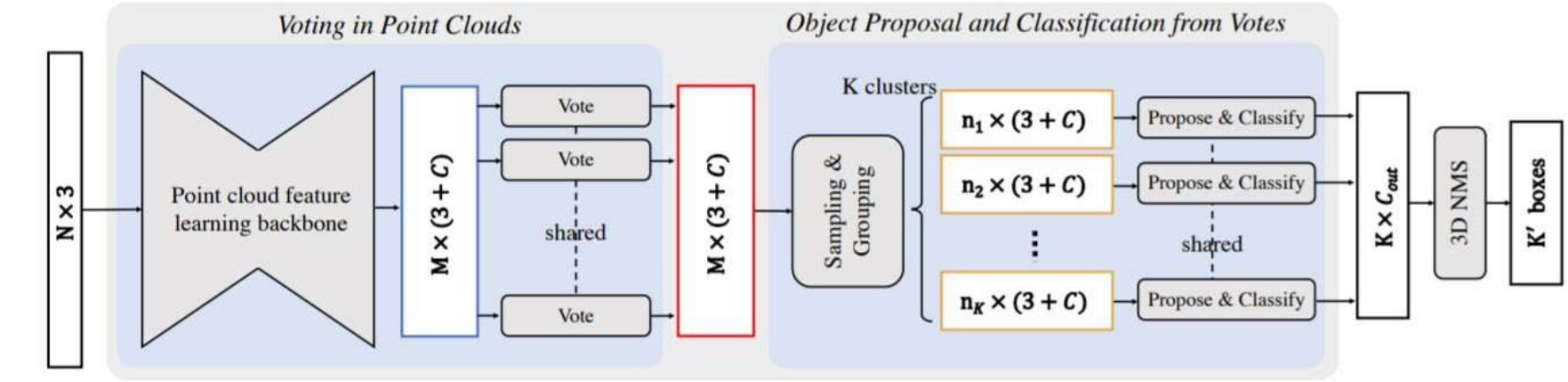
In our pipeline, we firstly build object proposals with a 2D detector running on RGB images, where each 2D bounding box defines a 3D frustum region. Then based on 3D point clouds in those frustum regions, we achieve 3D instance segmentation and amodal 3D bounding box estimation, using PointNet network.

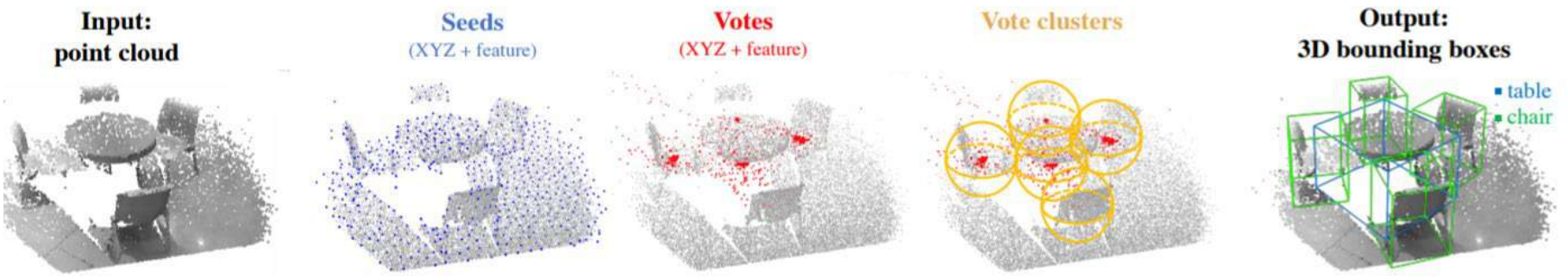


Reference: C. Qi, W. Liu, C. Wu, H. Su, L. Guibas, "Frustum PointNets for 3D Object Detection from RGB-D Data,“ https://github.com/charlesq34/frustum-pointnets

VoteNet

Given an input point cloud of N points with XYZ coordinates, a backbone network (implemented with PointNet) subsamples and learns deep features on the points and outputs a subset of M points but extended by C-dim features. This subset of points are considered as seed points. Each seed independently generates a vote through a voting module. Then the votes are grouped into clusters and processed by the proposal module to generate the final proposals. The classified proposals become the final 3D bounding boxes output.





Reference: Deep hough voting for 3D object detection in point clouds, https://arxiv.org/pdf/1904.09664.pdf

Thank you!

Dr TIAN Jing
Email: tianjing@nus.edu.sg