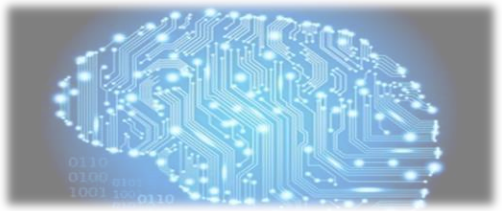
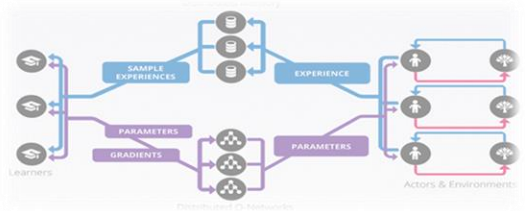

Survey: 3D Object Detection Methods for Autonomous Driving Applications

AutoML 연구실
석사과정 문아성

21.01.27

3D Object Detection

Recently, 2D vehicle detectiors achieved more than 90% Average Precision

whereas average precision of 3D vehicle detection achieved only 70%

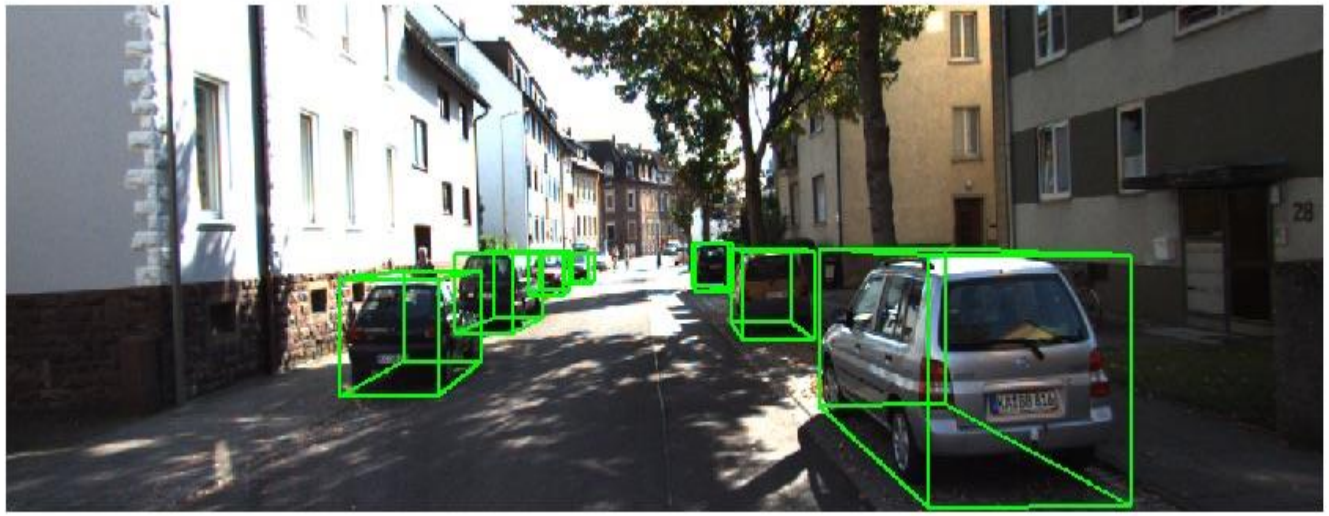

Fig. 1: Example 3D detection result from the KITTI validation set projected onto an image.

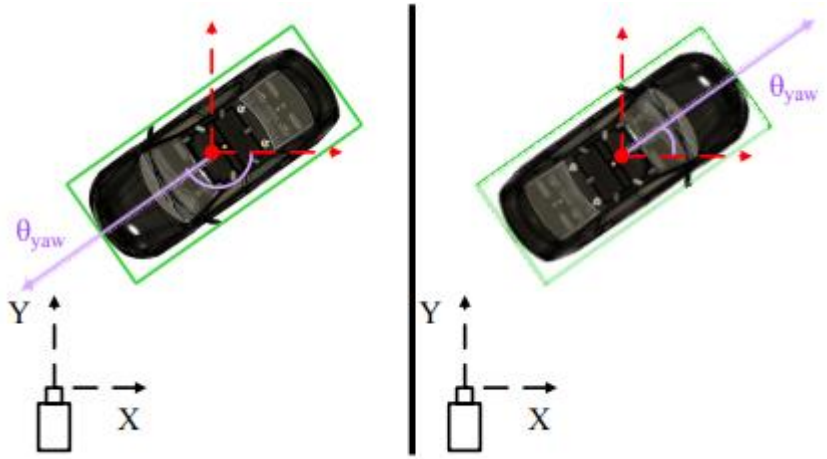

Fig. 2: A visual representation of the 3D detection problem from Bird's Eye View (BEV).

REF: https://www.groundai.com/project/mvx-net-multimodal-voxelnet-for-3d-object-detection/1

3D Object Detection Methods for Autonomous Driving Applications

3D Object Detection Methods

Category		Methodology/Advantages	Limitations	Method
Monocular		Use single RGB images to predict 3D object bounding boxes. Predicts 2D bounding boxes on the image plane then extrapolate them to 3D through reprojection constraints or bounding box regression.	The lack of explicit depth information on the input format limits the accuracy of localization performance.	Mono3D 3DVP SubCNN DeepManta
Point-cloud	**Projection**	Project point clouds into a 2D image and use established architectures for object detection on 2D images with extensions to regress 3D bounding boxes.	Projecting the point cloud data inevitably causes information loss. It also prevents the explicit encoding of spatial information as in raw point cloud data.	VeloFCN C-YOLO
	Volumetric	Generates a 3 dimensional representation of the point cloud in a voxel structure and uses FCNs to predict object detections. Shape information is encoded explicitly.	Expensive 3D convolutions increase models inference time. The volumetric representation is sparse and computationally inefficient.	3DFCN Vote3Deep
	PointNet	Uses feed-forward networks consuming raw 3d point clouds to generate predictions on class and estimated bounding boxes.	Considering the whole point cloud as input can increase run-time. Difficult establishing region proposals considering raw point inputs.	VoxelNet
Fusion		Fuses both front view images and point clouds to generate a robust detections. Architectures usually consider multiple branches, one per modality, and rely on region proposals. Allows modalities to interact and complement each other.	Requires calibration between sensors, and depending on the architecture can be computationally expensive.	MV3D AVOD

Table 1. Comparison of 3D object detection methods by category

Monocular-based Methods: Mono3D

X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, “Monocular 3D object detection for autonomous driving,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 2147–2156.

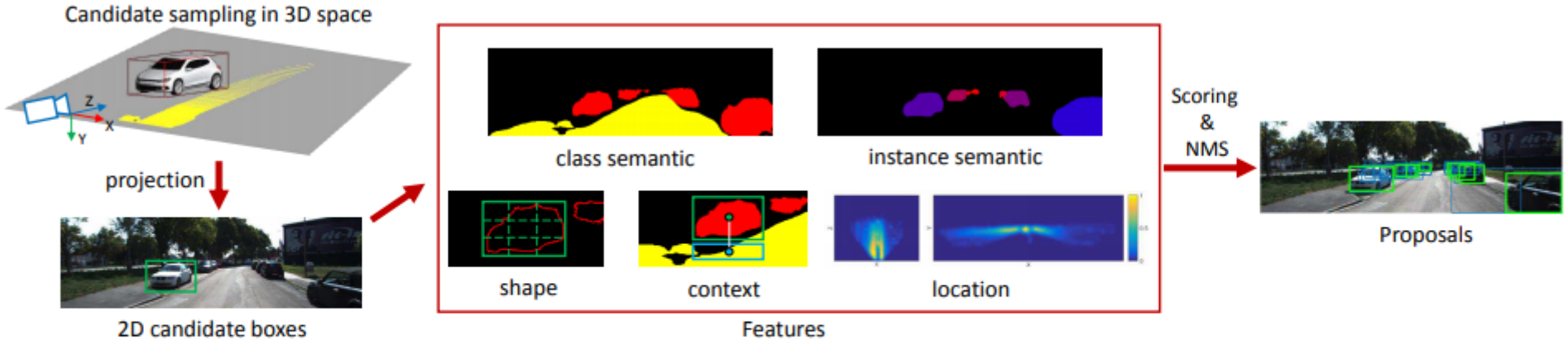

Figure 1: **Overview of our approach:** We sample candidate bounding boxes with typical physical sizes in the 3D space by assuming a prior on the ground-plane. We then project the boxes to the image plane, thus avoiding multi-scale search in the image. We score candidate boxes by exploiting multiple features: class semantic, instance semantic, contour, object shape, context, and location prior. A final set of object proposals is obtained after non-maximum suppression.

Monocular-based Methods: 3D Voxel Patterns (3DVP)

Y. Xiang, W. Choi, Y. Lin, and S. Savarese, "Data-driven 3D voxel patterns for object category recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2015, pp. 1903–1911.

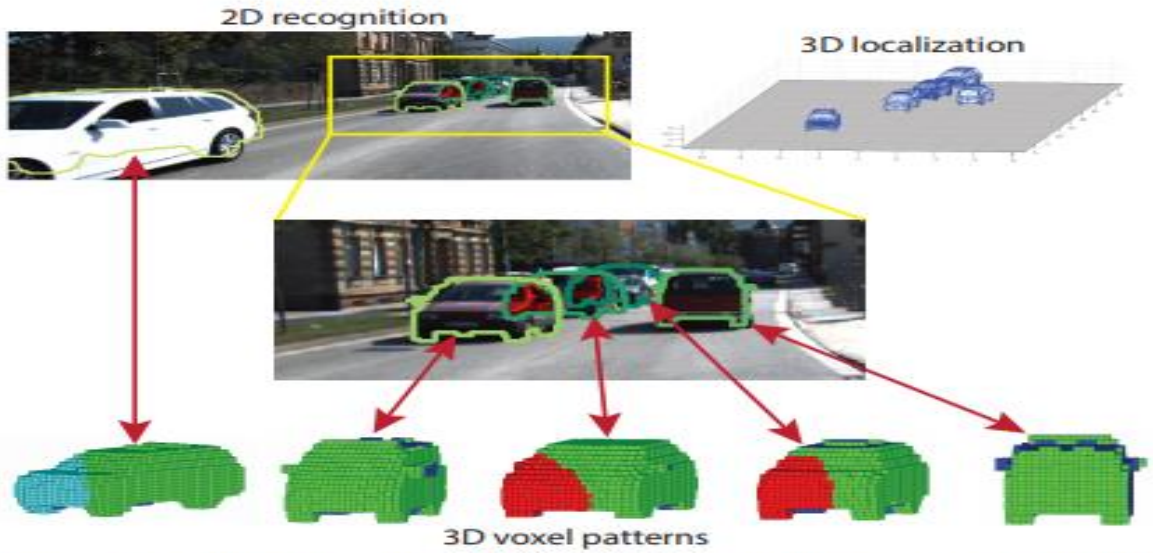

Figure 1. By introducing the 3D voxel patterns, our recognition framework is able to not only detect objects in images, but also segment the detected objects from the background, estimate the 3D poses and 3D shapes, localize them in the 3D space, and even infer the occlusion relationship among them. Green, red and cyan voxels are visible, occluded and truncated respectively.

Monocular-based Methods: SubCNN

Y. Xiang, W. Choi, Y. Lin, and S. Savarese, “Subcategory-aware convolutional neural networks for object proposals and detection,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Mar. 2017, pp. 924–933.

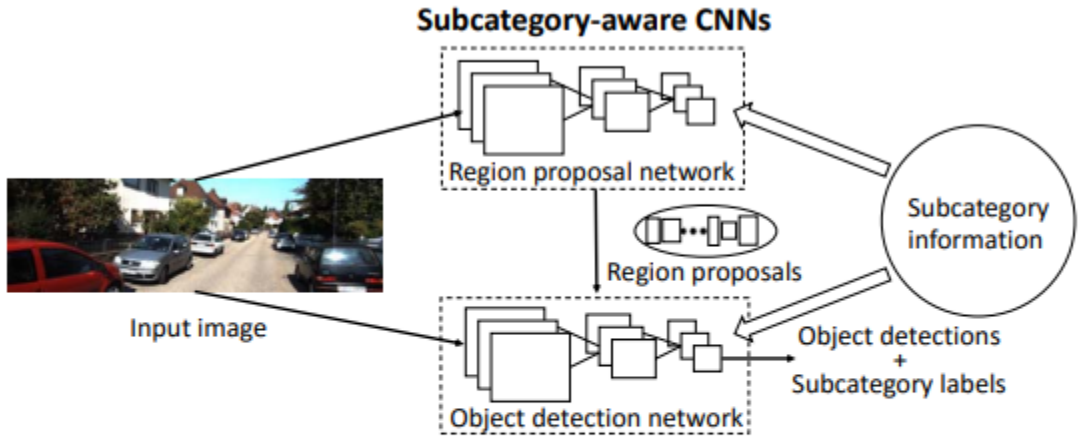

Figure 1. Overview of our object detection framework. By exploiting subcategory information, we propose a new CNN architecture for region proposal and a new object detection network for joint detection and subcategory classification.

Monocular-based Methods: DeepManta

F. Chabot, M. Chaouch, J. Rabarisoa, C. Teulière, and T. Chateau, “Deep MANTA: A coarse-to-fine many-task network for joint 2D and 3D vehicle analysis from monocular image,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1827–1836.

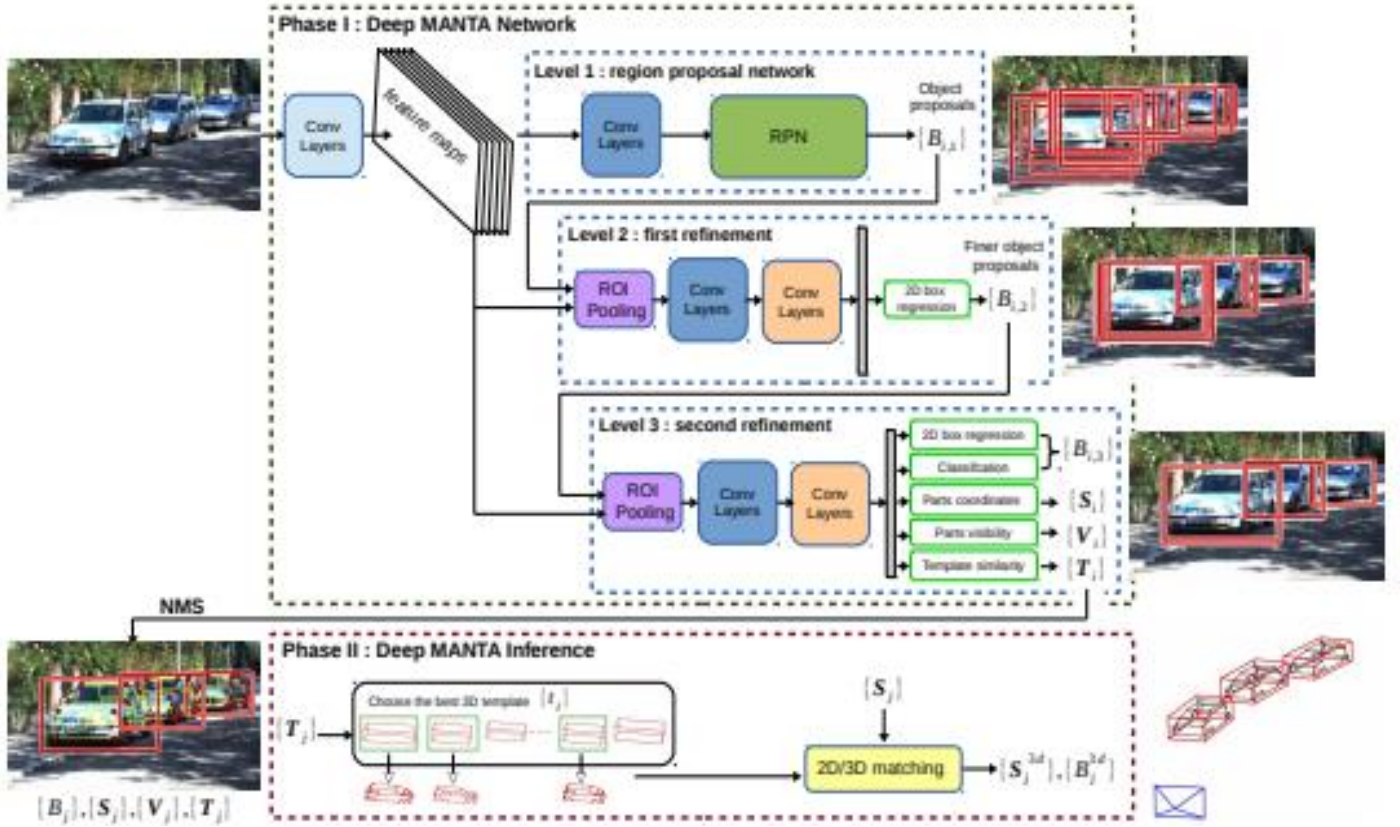

Fig 1. Overview of the Deep MANTA approach.
The entire input image is forwarded inside the Deep MANTA network.
Conv layers with the same color share the same weights. Moreover, these three convolutional blocks correspond to the split of existing CNN architecture.
The inference step allows to choose the best corresponding 3D template using template similarity Tj and then performs 2D/3D pose computation using the associated 3D shape.

Point cloud Projection-based Methods: VeloFCN

B. Li, T. Zhang, and T. Xia, “Vehicle detection from 3D Lidar using fully convolutional network,” in Proc. Robot., Sci. Syst. XII, AnnArbor, MI, USA, Jun. 2016. [Online]. Available: http://www.roboticsproceedings.org/rss12/

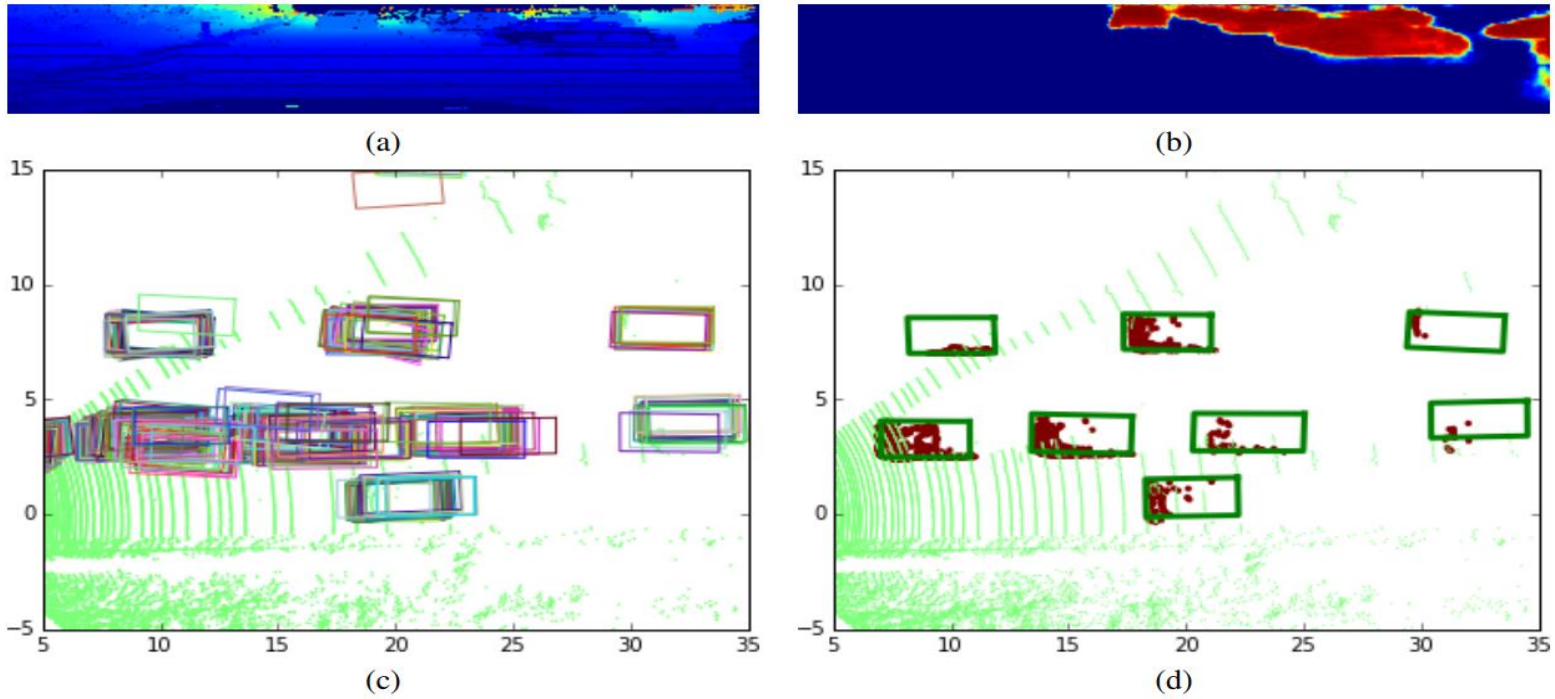

Fig. 1. Data visualization generated at different stages of the proposed approach. (a) The input point map, with the d channel visualized. (b) The output confidence map of the objectness branch at $\mathbf{o}_{\mathbf{p}}^{a}$. Red denotes for higher confidence. (c) Bounding box candidates corresponding to all points predicted as positive, i.e. high confidence points in (b). (d) Remaining bounding boxes after non-max suppression. Red points are the groundtruth points on vehicles for reference.

Point cloud Projection-based Methods: C-YOLO

M. Simon, S. Milz, K. Amende, and H. Gross. (Mar. 2018). “ComplexYOLO: Real-time 3D object detection on point clouds.” [Online].

Available: https://arxiv.org/abs/1803.06199

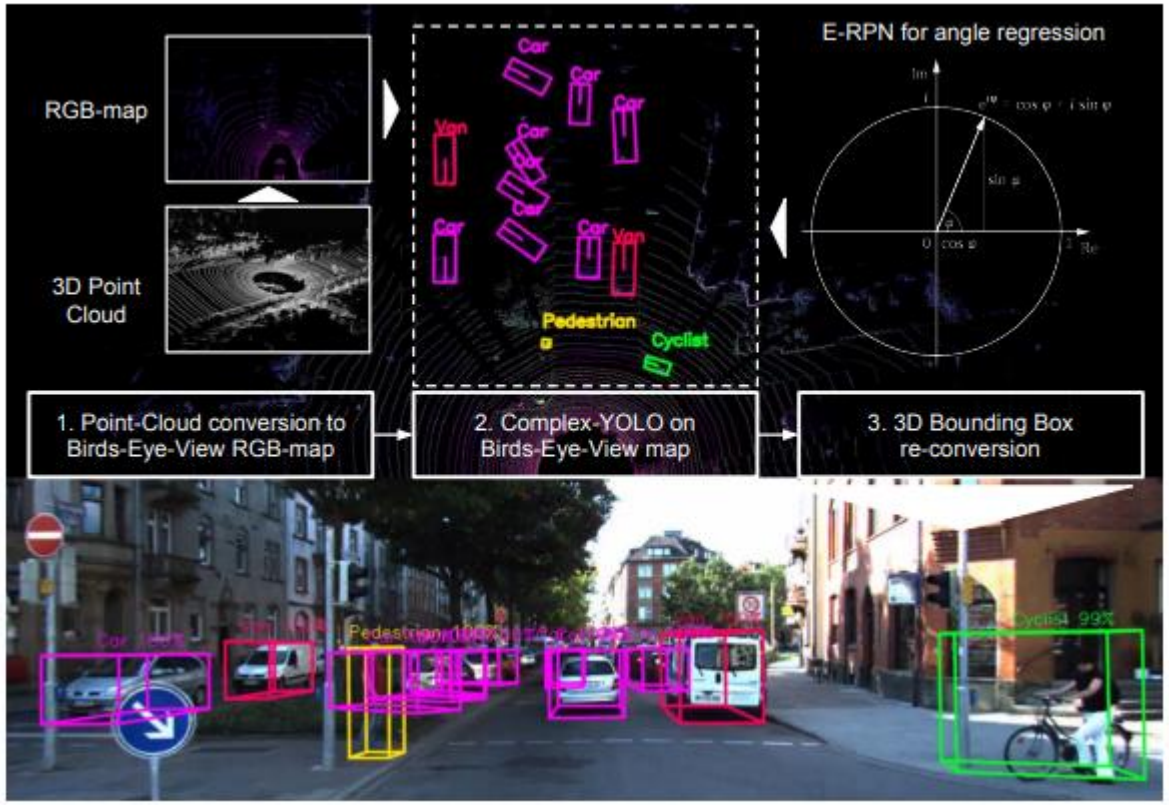

Fig. 1. Complex-YOLO is a very efficient model that directly operates on Lidar only based birds-eye-view RGB-maps to estimate and localize accurate 3D multiclass bounding boxes. The upper part of the figure shows a bird view based on a Velodyne HDL64 point cloud such as the predicted objects. The lower one outlines the re-projection of the 3D boxes into image space. Note: Complex-YOLO needs no camera image as input, it is Lidar based only.

Point cloud Volumetric-based Methods: 3DFCN

B. Li, “3D fully convolutional network for vehicle detection in point cloud,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Sep. 2017, pp. 1513–1518.

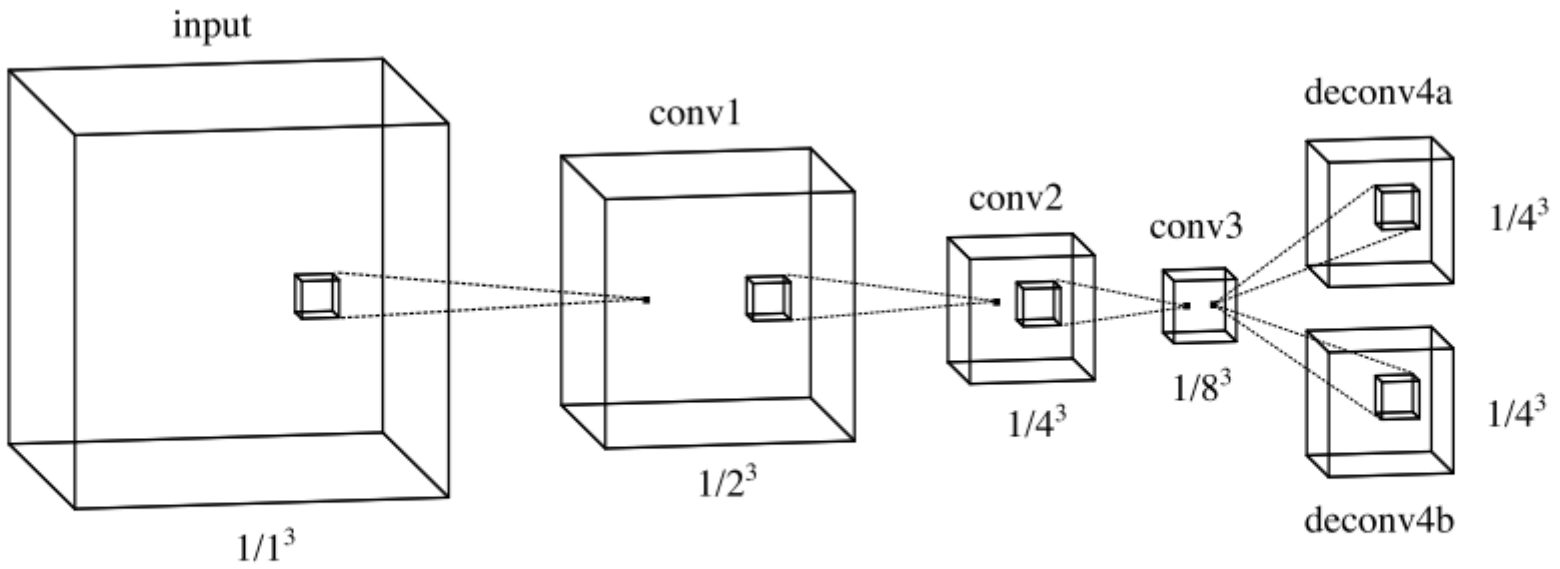

Fig. 2. A sample illustration of the 3D FCN structure used in this paper. Feature maps are first down-sampled by three convolution operation with the stride of $1/2^3$ and then up-samped by the deconvolution operation of the same stride. The output objectness map ($\mathbf{o}^a$) and bounding box map ($\mathbf{o}^b$) are collected from the deconv4a and deconv4b layers respectively.

Point cloud Volumetric-based Methods: Vote3Deep

B. Li, “3D fully convolutional network for vehicle detection in point cloud,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Sep. 2017, pp. 1513–1518.

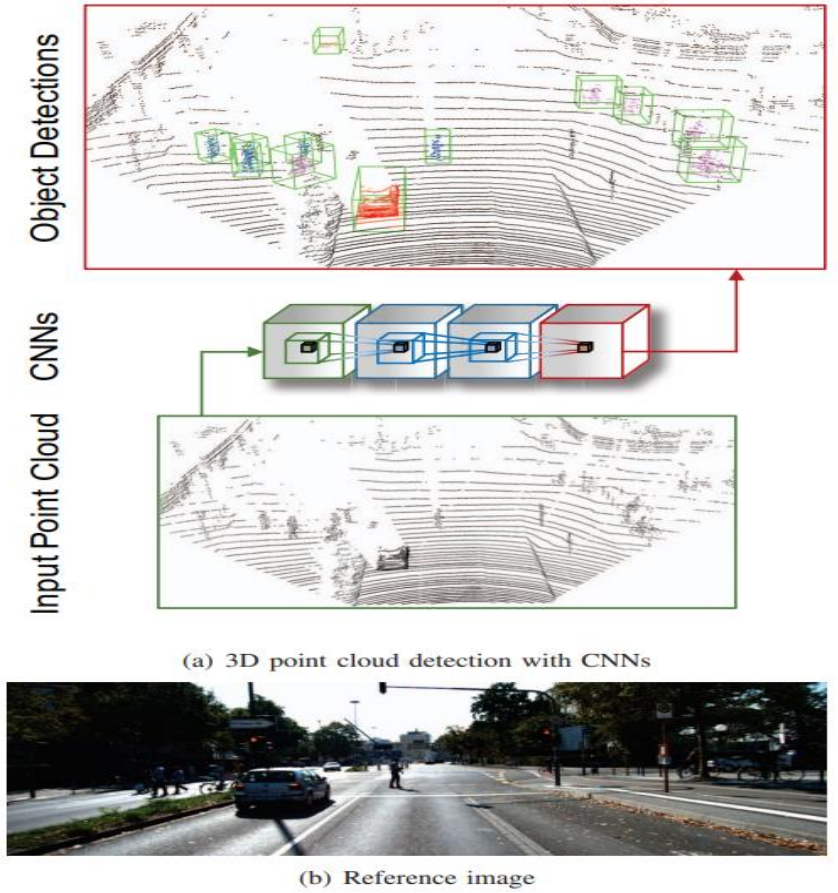

Fig. 1. The result of applying Vote3Deep to an unseen point cloud from the KITTI dataset, with the corresponding image for reference.
The CNNs apply sparse convolutions natively in 3D via voting.
The model detects cars (red), pedestrians (blue), and cyclists (magenta), even at long range, and assigns bounding boxes (green) sized by class.
Best viewed in colour.

Point cloud PointNet-based Methods: VoxelNet

Y. Zhou and O. Tuzel. (Nov. 2017). “VoxelNet: End-to-end learning for point cloud based 3D object detection.” [Online]. Available: https://arxiv.org/abs/1711.06396

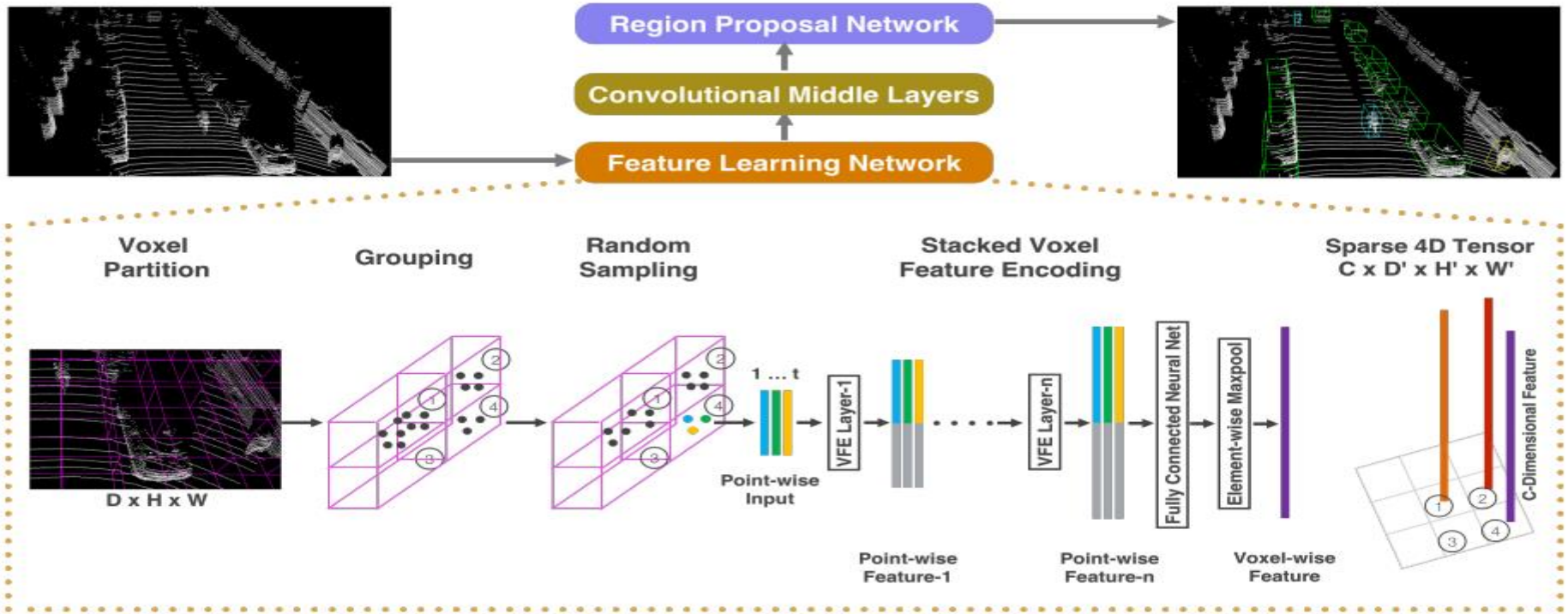

Figure 2. VoxelNet architecture. The feature learning network takes a raw point cloud as input, partitions the space into voxels, and transforms points within each voxel to a vector representation characterizing the shape information. The space is represented as a sparse 4D tensor. The convolutional middle layers processes the 4D tensor to aggregate spatial context. Finally, a RPN generates the 3D detection.

Fusion-based Methods: MV3D

X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3D object detection network for autonomous driving," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 6526–6534.

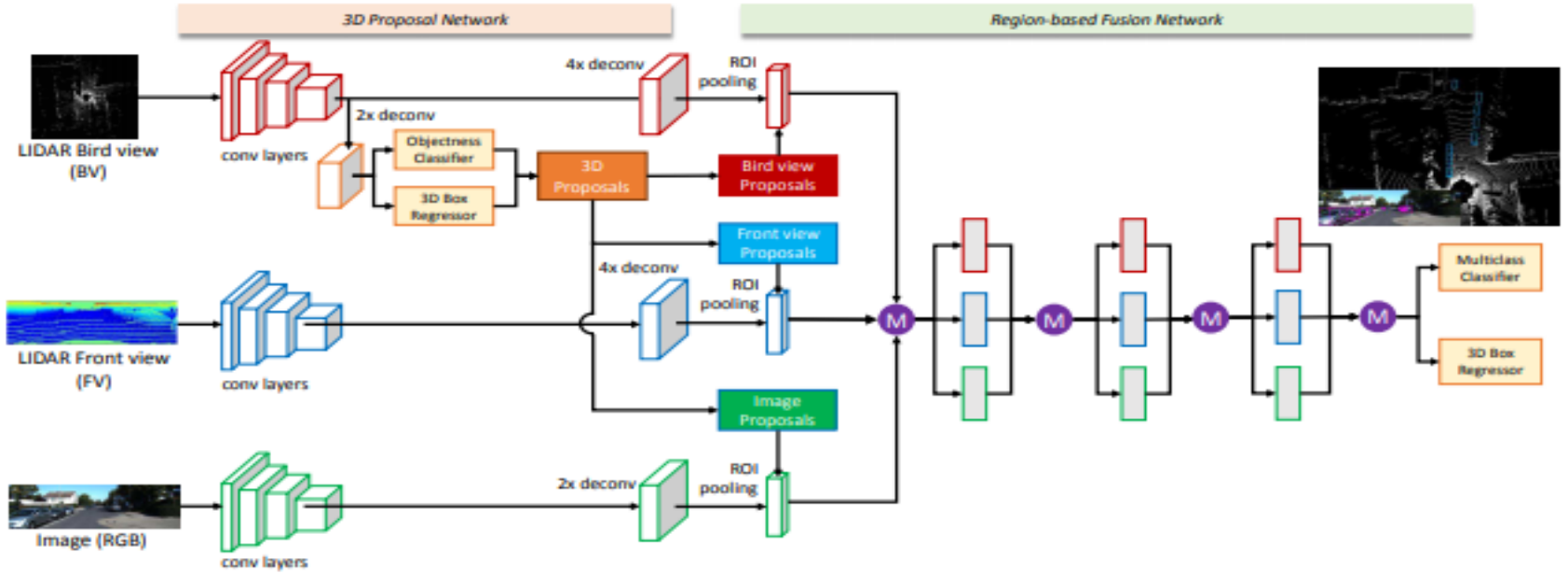

Figure 1: **Multi-View 3D object detection network (MV3D):** The network takes the bird's eye view and front view of LIDAR point cloud as well as an image as input. It first generates 3D object proposals from bird's eye view map and project them to three views. A deep fusion network is used to combine region-wise features obtained via ROI pooling for each view. The fused features are used to jointly predict object class and do oriented 3D box regression.

KITTI Test set Results

TABLE VII

KITTI TEST SET RESULTS ON 2D OBJECT DETECTION FOR CAR CLASS

Method	Modality	AP_{2D}			AOS		
		E	M	H	E	M	H
DeepManta [42]	Mono	96.4	90.1	80.79	96.32	89.91	80.55
SubCNN [45]	Mono	90.81	89.04	79.27	90.67	88.62	78.68
Deep3DBox [40]	Mono	92.98	89.04	77.17	92.9	88.75	76.76
DeepStOP [44]	Stereo	93.45	89.04	79.58	92.04	86.86	77.34
Mono3D [39]	Mono	92.33	88.66	78.96	91.01	86.62	76.84
3DOP [43]	Stereo	93.04	88.64	79.1	91.44	86.1	76.52
3DVP [41]	Mono	87.46	75.77	65.38	86.92	74.9	64.11
F-PointNet [63]	LIDAR+Mono	90.78	90	80.8			
MV3D [64]	LIDAR+Mono	90.53	89.17	80.16			
AVOD-FPN [6]	LIDAR+Mono	89.99	87.44	80.05	89.95	87.13	79.74
VoxelNet [62]	LIDAR	90.3	85.95	79.21			
3DFCN [54]	LIDAR	84.2	75.3	68	84.1	75.2	67.9
Vote3Deep [55]	LIDAR	76.79	68.24	63.23			
VeloFCN [48]	LIDAR	60.3	47.5	42.7	59.1	459	41.1

E, M and H stands for Easy, Moderate and Hard, respectively.

TABLE X

3D OBJECT DETECTION BENCHMARK ON KITTI TEST SET. 3D IOU 0.7

Method	Time(s)	Class	AP_{3D}			AP_{BV}		
			E	M	H	E	M	H
MV3D [64]	0.36	Car	71.09	62.35	55.12	86.02	76.9	68.49
AVOD [6]	0.08		73.59	65.78	58.38	86.8	85.44	77.73
AVOD-FPN [6]	0.1		81.94	71.88	66.38	88.53	83.79	77.9
F-Pointnet [63]	0.17		81.2	70.39	62.19	88.7	84	75.33
Voxelnet [62]	0.23		77.47	65.11	57.73	89.35	79.26	77.39
C-YOLO[1] [30]	0.02		67.72	64	63.01	85.89	77.4	77.33
AVOD [6]	0.08	Ped	38.28	31.51	26.98	42.51	35.24	33.97
AVOD-FPN [6]	0.1		50.8	42.81	40.88	58.75	51.05	47.54
F-Pointnet [63]	0.17		51.21	44.89	40.23	58.09	50.22	47.2
Voxelnet [62]	0.23		39.48	33.69	31.51	46.13	40.74	38.11
C-YOLO[1] [30]	0.02		41.79	39.7	35.92	46.08	45.9	44.2
AVOD [6]	0.08	Cyc	60.11	44.9	38.8	63.66	47.74	46.55
AVOD-FPN [6]	0.1		64	52.18	46.61	68.09	57.48	50.77
F-Pointnet [63]	0.17		71.96	56.77	50.39	75.38	61.96	54.68
Voxelnet [62]	0.23		61.22	48.36	44.37	66.7	54.76	50.55
C-YOLO[1] [30]	0.02		68.17	58.32	54.3	72.37	63.36	60.27

1 The authors did not provide public test set results, only validation set
E, M and H stands for Easy, Moderate and Hard, respectively.

References

- Wu, Junhui, et al. "A Survey on Monocular 3D Object Detection Algorithms Based on Deep Learning." Journal of Physics: Conference Series. Vol. 1518. No. 1. IOP Publishing, 2020.
- Arnold, Eduardo, et al. "A survey on 3d object detection methods for autonomous driving applications." IEEE Transactions on Intelligent Transportation Systems 20.10 (2019): 3782-3795.