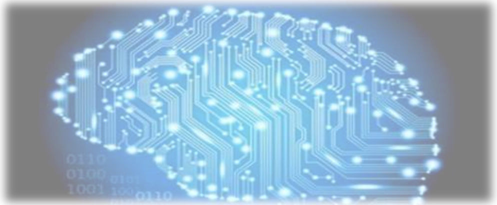
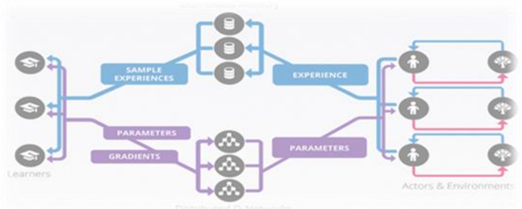
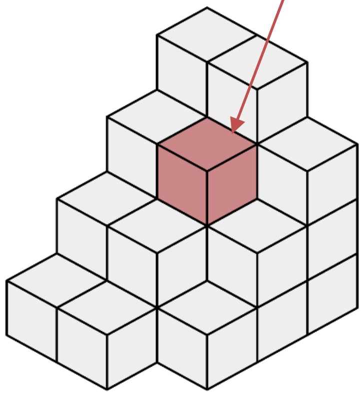




VoxelNet: End-to-end Learning for Point Cloud Based 3D Object Detection

AutoML 연구실
석사과정 문아성

21.02.03



VoxelNet: End-to-end Learning for Point Cloud Based 3D Object Detection

Zhou, Yin, and Oncel Tuzel. "Voxelnet: End-to-end learning for point cloud based 3d object detection."
Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

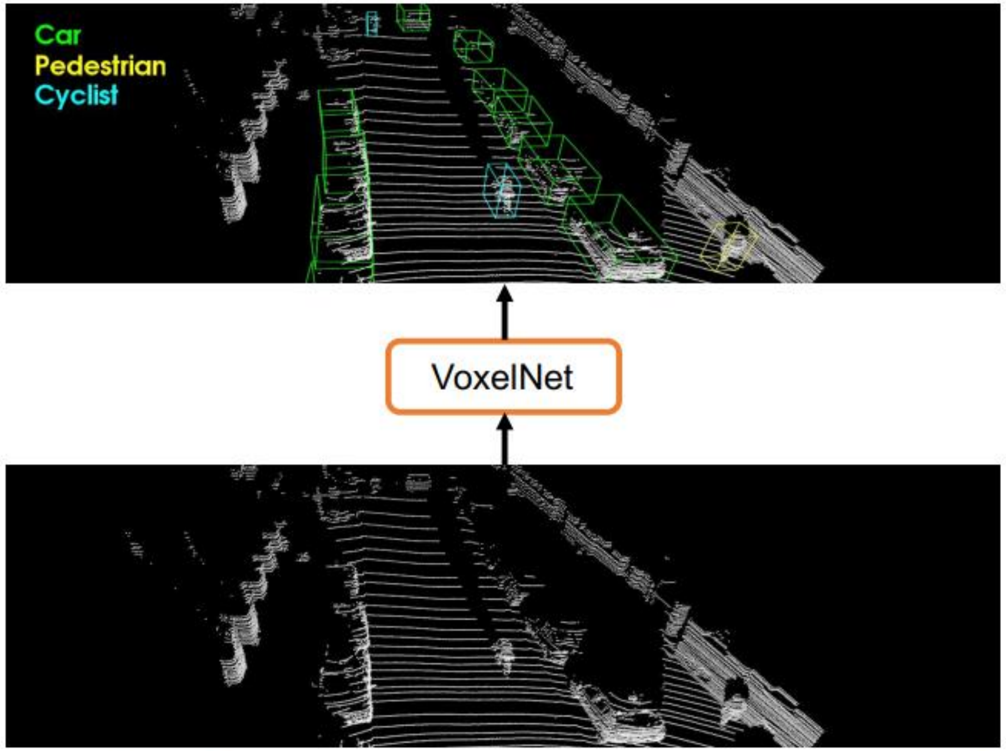
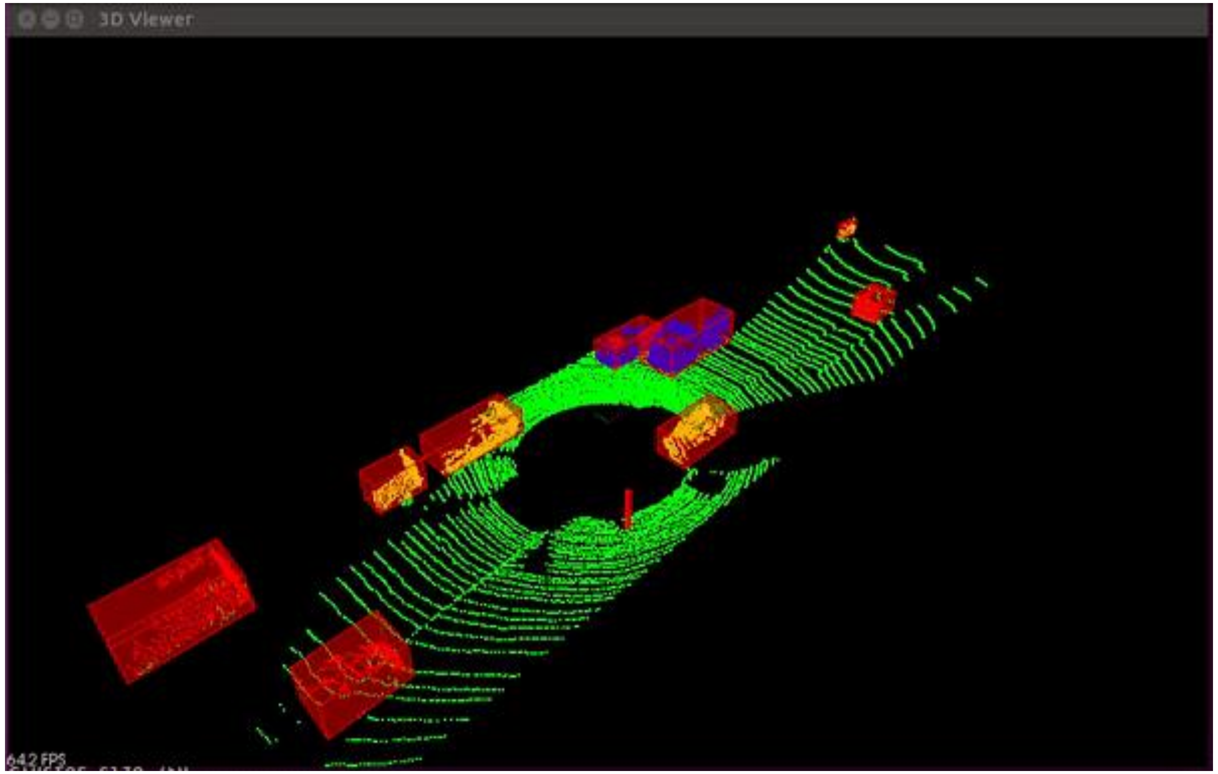


Figure 1. VoxelNet directly operates on the raw point cloud (no need for feature engineering) and produces the 3D detection results using a single end-to-end trainable network.



REF: https://laptrinhx.com/lidar-point-cloud-based-3d-object-detection-implementation-with-colab-part-1-of-2-2816461639/

VoxelNet: End-to-end Learning for Point Cloud Based 3D Object Detection

Architecture

1. Feature Learning Network
2. Convolutional Middle Layers
3. Region Proposal Network

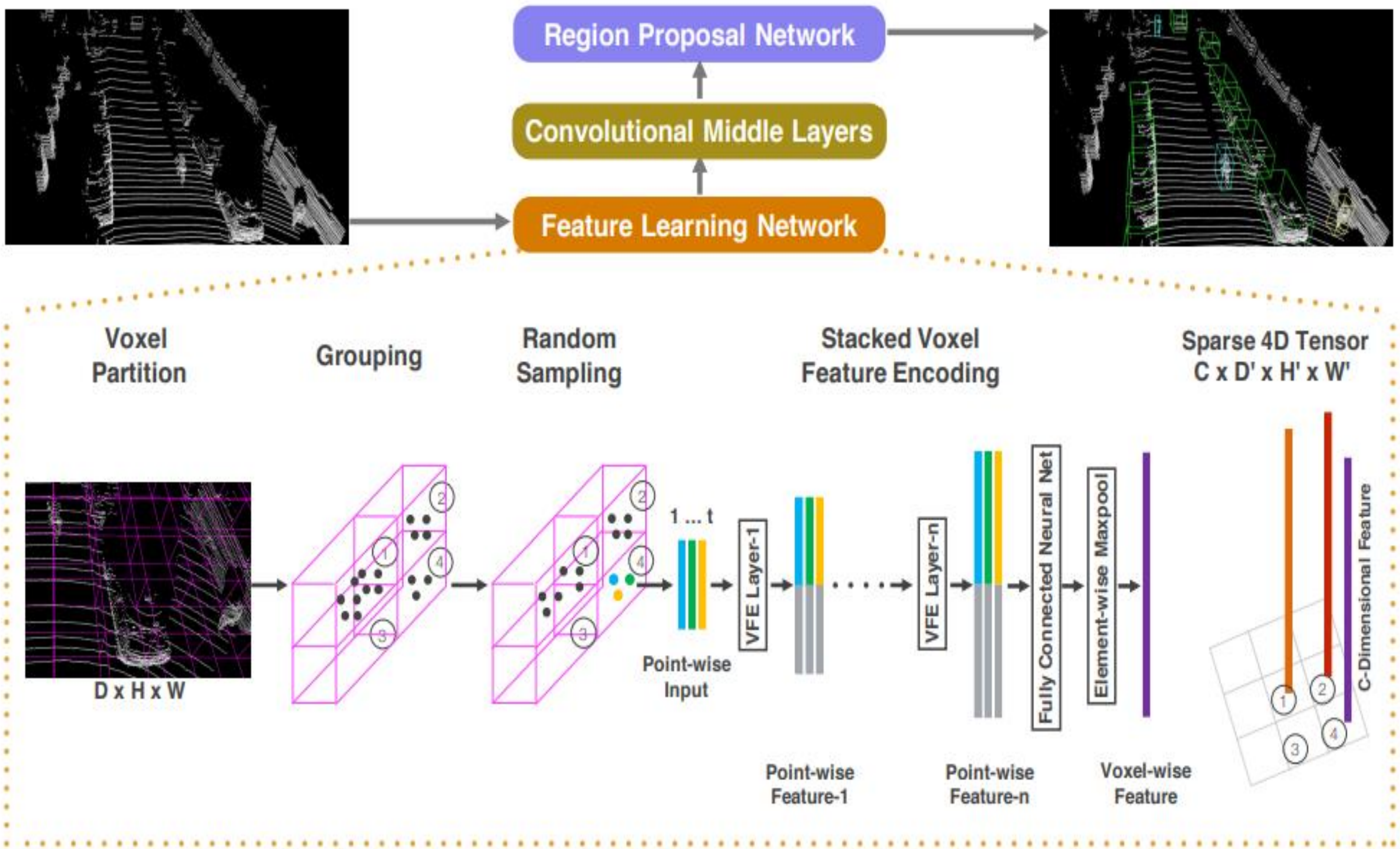
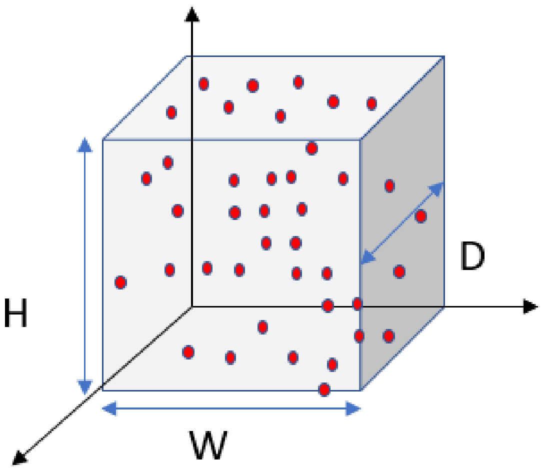


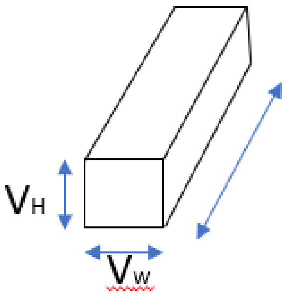
Figure 2. VoxelNet architecture. The feature learning network takes a raw point cloud as input, partitions the space into voxels, and transforms points within each voxel to a vector representation characterizing the shape information. The space is represented as a sparse 4D tensor. The convolutional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

Feature Learning Network

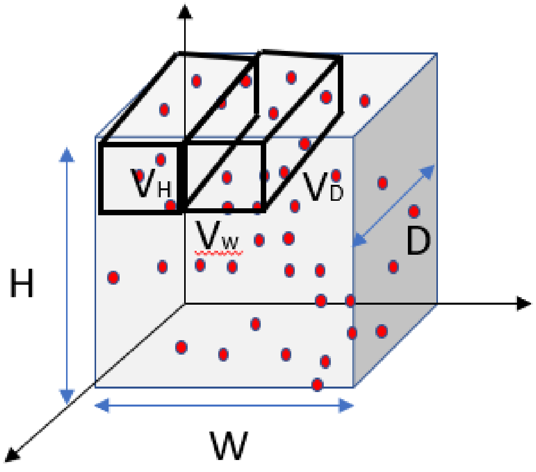
Voxel Partition



3D space with range
D,H,W along Z,Y,X axis

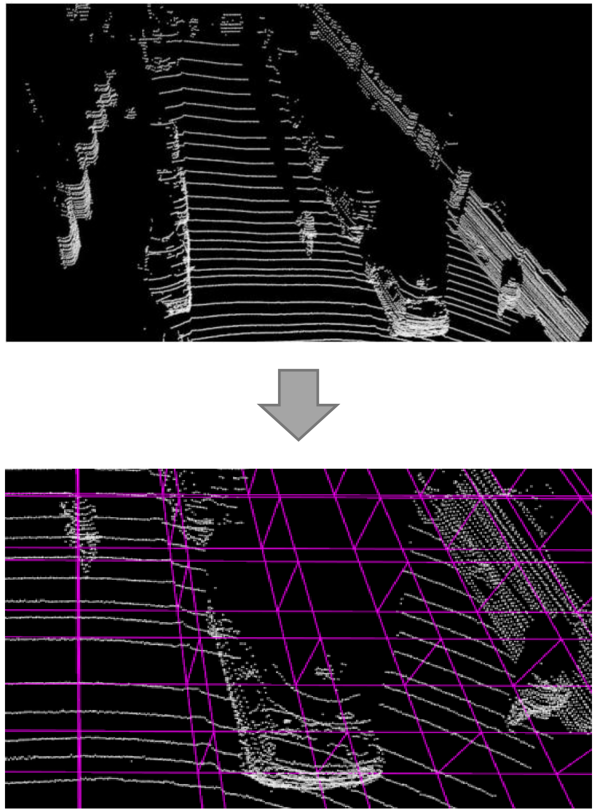


voxel of size
v_D, v_H, v_W



Divides the 3D space (D,H,W) into uniform voxel grids size(V_D^1, V_H^1, V_w^1)

$$V_D^1 = D/V_D, V_H^1 = H/V_H, V_W^1 = W/V_W$$



D x H x W

REF: https://laptrinhx.com/lidar-point-cloud-based-3d-object-detection-implementation-with- colab-part-1-of-2-2816461639/

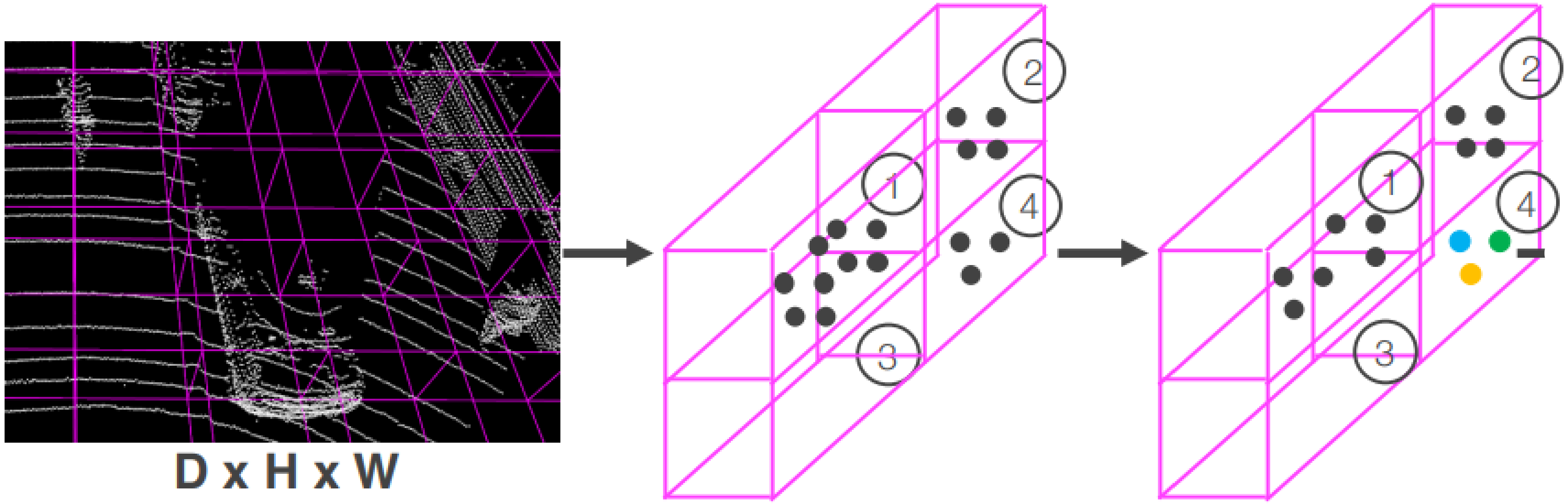
VoxelNet: End-to-end Learning for Point Cloud Based 3D Object Detection

Feature Learning Network

Grouping & Random Sampling

Maximum number of points per voxel: T

ex) T = 6, ①: 9 -> 6 (Random Sampling)



Voxel Partition Image Grouping Random Sampling

VoxelNet: End-to-end Learning for Point Cloud Based 3D Object Detection

Feature Learning Network

Stacked Voxel Feature Encoding

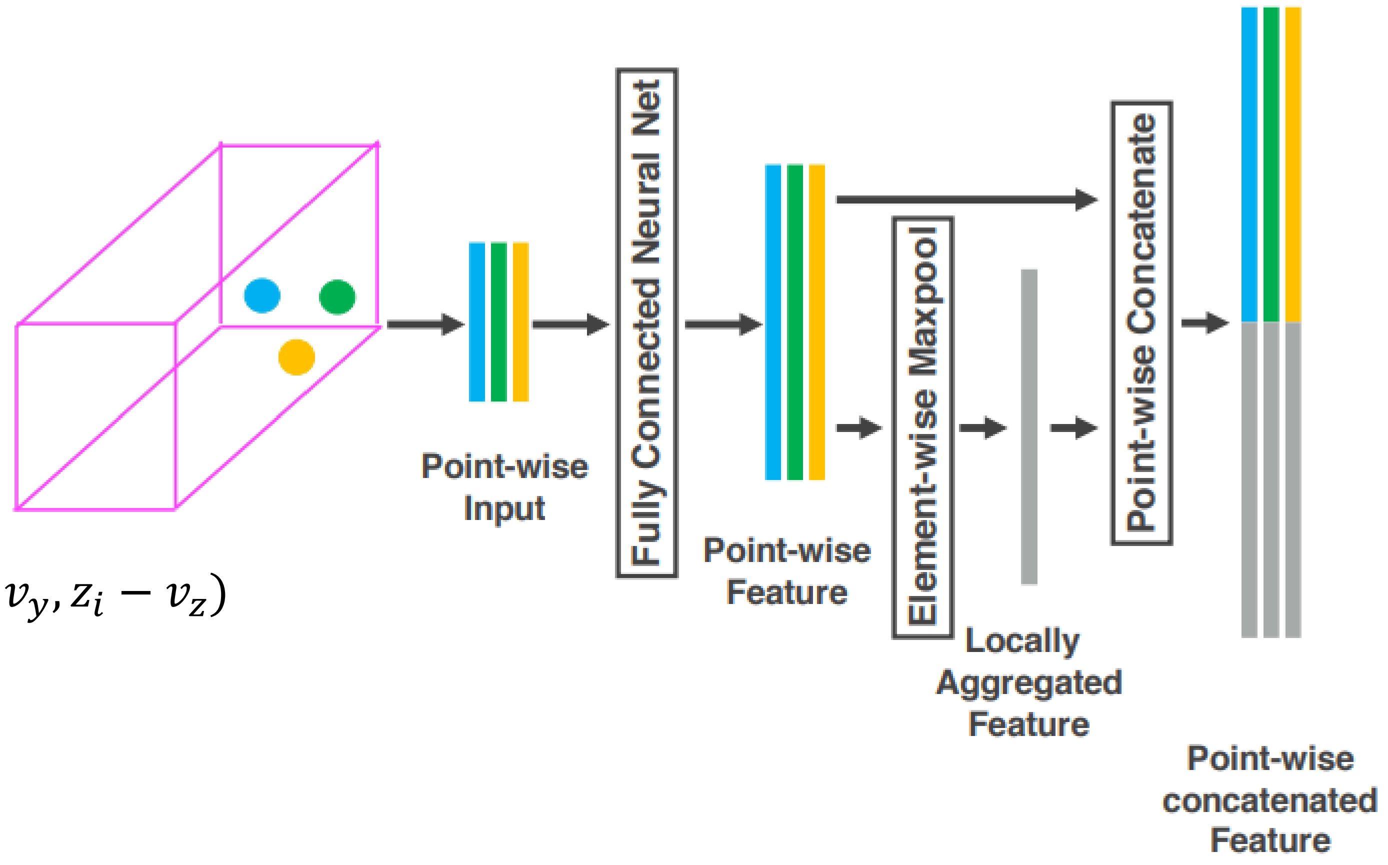
LiDAR DATA: (x, y, z, r) $\quad r: reflectance$

$point_i : (x_i, y_i, z_i, r_i)$

The centroid of $point_i : (v_x, v_y, v_z)$

$The\ feature\ of\ point_i : (x_i, y_i, z_i, r_i, x_i - v_x, y_i - v_y, z_i - v_z)$

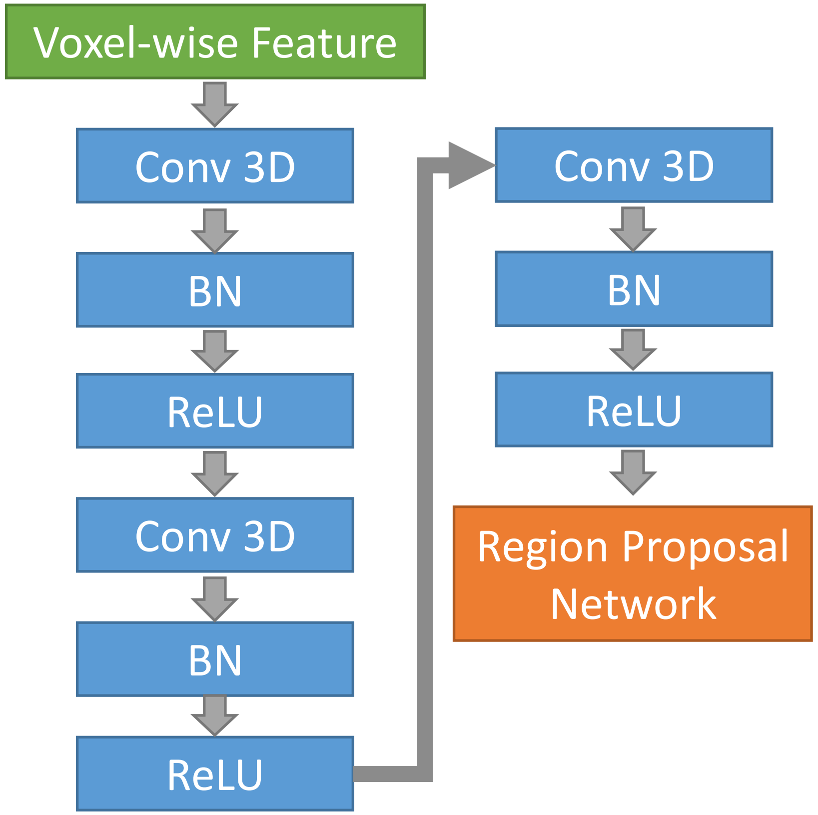
(= Point-wise Input)

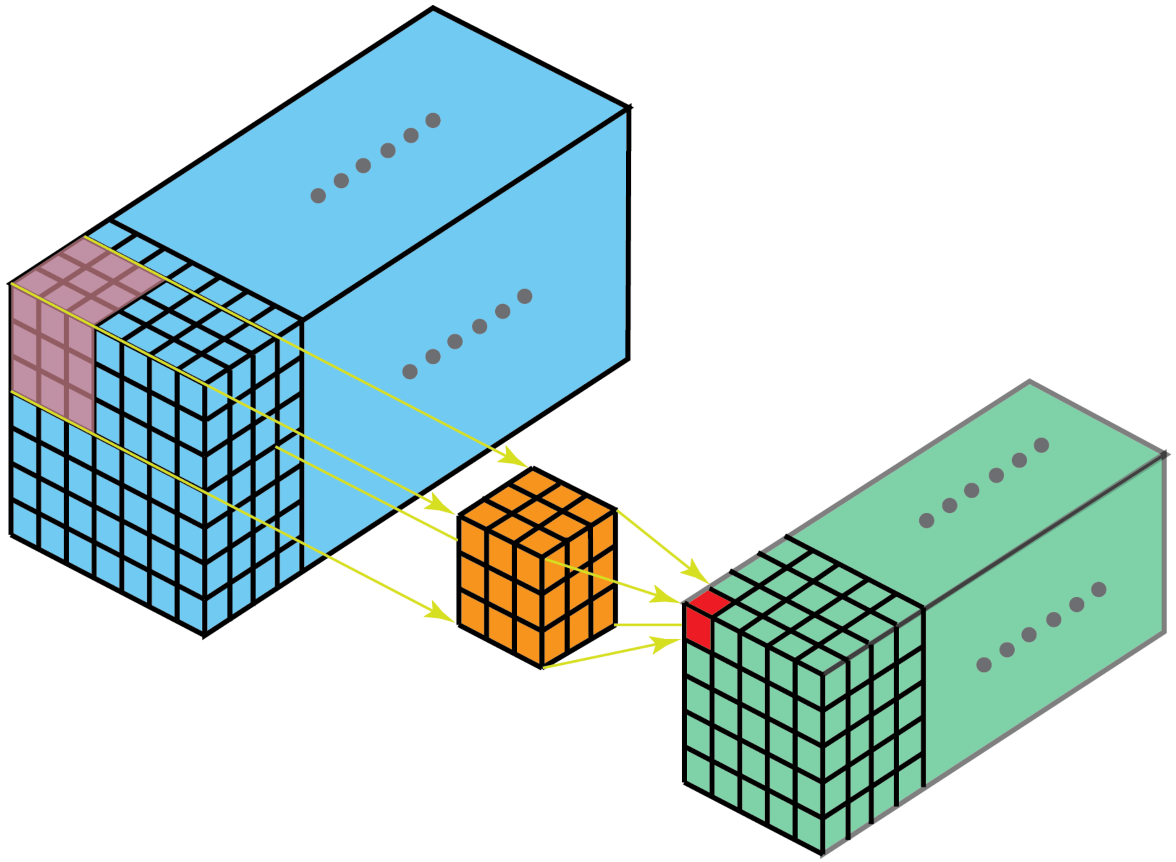


Voxel Feature Encoding Layer

VoxelNet: End-to-end Learning for Point Cloud Based 3D Object Detection

Convolutional Middle Layers



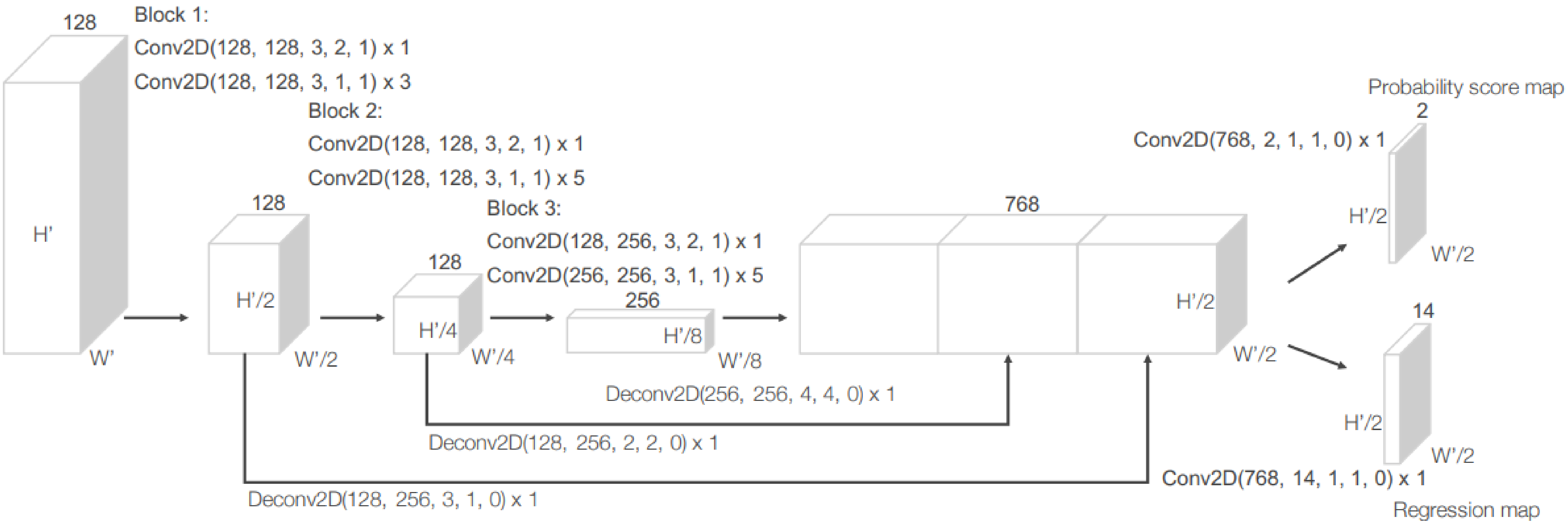


3D Convolution

REF: https://towardsdatascience.com/a-comprehensive-introduction-to-different-types-of-convolutions-in-deep-learning-669281e58215

VoxelNet: End-to-end Learning for Point Cloud Based 3D Object Detection

Region Proposal Network



Region Proposal Network Architecture

Evaluation on KITTI Test Set & Latest 3D Detection Models

VoxelNet Runtime: 0.23s

Method	Modality	Car		
		Easy	Moderate	Hard
Mono3D [3]	Mono	2.53	2.31	2.31
3DOP [4]	Stereo	6.55	5.07	4.10
VeloFCN [22]	LiDAR	15.20	13.66	15.98
MV (BV+FV) [5]	LiDAR	71.19	56.60	55.30
MV (BV+FV+RGB) [5]	LiDAR+Mono	71.29	62.68	56.56
HC-baseline	LiDAR	71.73	59.75	55.69
VoxelNet	LiDAR	**81.97**	**65.46**	**62.85**

Table 1. Performance comparison in 3D detection: average precision (in %) on KITTI validation set.

Car

	Method	Setting	Code	Moderate	Easy	Hard	Runtime
1	HIKVISION-ADLab-HZ			82.83 %	89.00 %	76.00 %	0.1 s
2	SE-SSD			82.54 %	91.49 %	77.15 %	0.03 s
3	EA-M-RCNN(BorderAtt)			82.33 %	87.77 %	77.37 %	0.08 s
4	HUAWEI Octopus			82.13 %	88.26 %	77.41 %	0.1 s
5	ADLAB			82.08 %	90.92 %	77.36 %	0.05 s
6	PV-RCNN-v2			81.88 %	90.14 %	77.15 %	0.06 s
7	RangeRCNN-LV			81.85 %	88.76 %	77.18 %	0.1 s
8	PVGNet			81.81 %	89.94 %	77.09 %	0.05 s
9	E^2-PV-RCNN			81.70 %	88.33 %	77.20 %	0.08 s
10	PLNL-3DSSD			81.69 %	88.98 %	74.90 %	0.08 s
11	DomainAdp+PVRCNN			81.66 %	88.64 %	77.08 %	0.09 s

Latest 3D Detection Models