Pyramid R-CNN: Towards Better Performance and Adaptability for 3D Object Detection

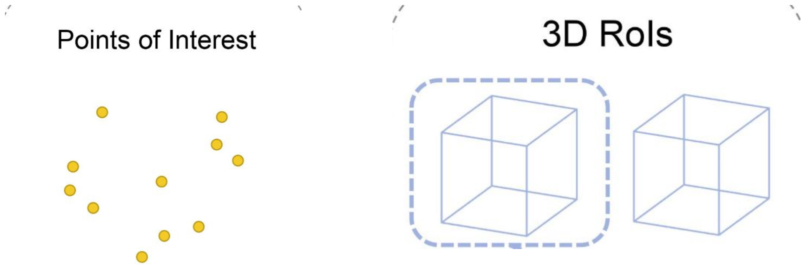
Authors: Jiageng Mao[1], Minzhe Niu[2], Haoyue Bai[3], Xiaodan Liang[4], Hang Xu[2], Chunjing Xu[2]
Institutes: [1] The Chinese University of Hong Kong [2] Huawei Noah's Ark Lab [3] HKUST [4] Sun Yat-Sen University
Conference: :In Proceedings of the IEEE/CVF International Conference on Computer Vision 2021
Citations: 64 (Google scholar)

Presentation by:
Sophia Damilola Lawal
310882

Contents

Terminologies

- ➔ **Points of Interest (PoIs)**:refers to a specific location in the 3D space, such as a corner or a vertex of an object; used to define objects' shape and structure
- ➔ **Regions of Interest (RoIs)**: are defined areas or volumes that enclose the objects



J. Mao, M. Niu, H. Bai, X. Liang, H. Xu, and C. Xu, ‘Pyramid R-CNN: Towards Better Performance and Adaptability for 3D Object Detection’, in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, Oct. 2021, pp. 2703–2712, doi: 10.1109/ICCV48922.2021.00272 [Online]. Available: https://ieeexplore.ieee.org/document/9709959/. [Accessed: Dec. 31, 2022]

Introduction

- Two-stage detectors are more accurate than Single-stage detectors because of ROI refinement(second stage)
- 3D detectors perform different types of RoI feature extraction on points of interest like RegionPooling, sparse convolution, RoI grid pooling
- **Problem:** Points of Interest (PoIs) are affected by sparsity and non-uniform distribution of the input point clouds, which can lead to difficulty detecting objects further away

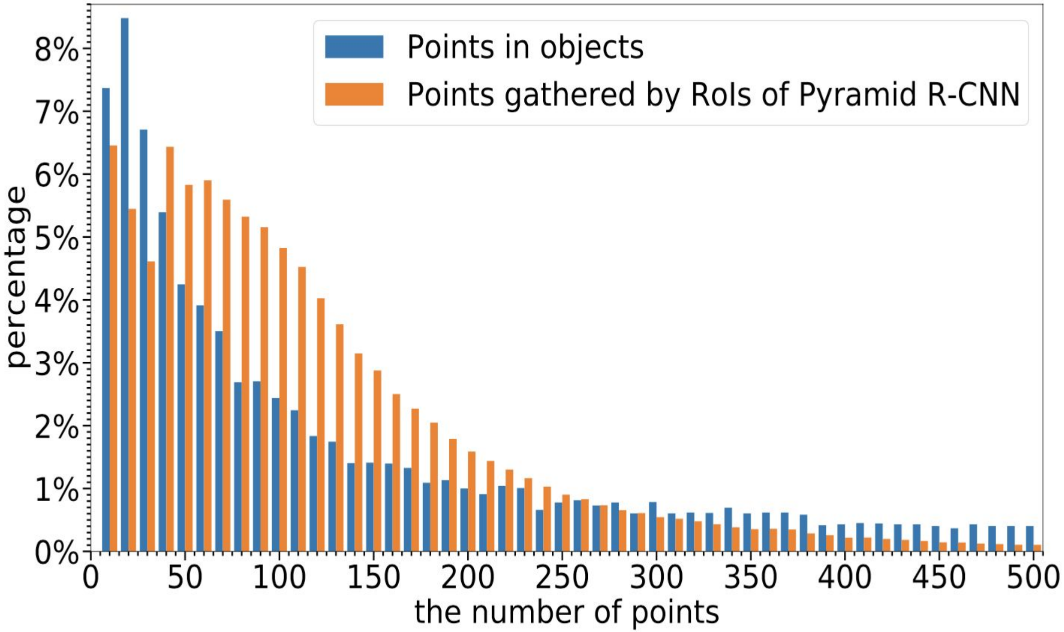


Figure 1. Statistical results on the KITTI dataset.

Jiageng Mao, Minzhe Niu, Haoyue Bai, Xiaodan Liang, Hang Xu, and Chunjing Xu. 2021. Pyramid R-CNN: Towards Better Performance and Adaptability for 3D Object Detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Montreal, QC, Canada, 2703–2712. DOI:https://doi.org/10.1109/ICCV48922.2021.00272

Introduction

- **Contribution**: Proposed a second stage module called the Pyramid R-CNN which consist of three main components
    - RoI-grid Pyramid
    - RoI-grid Attention
    - Density Aware Radius Prediction (DARP)
- The Pyramid R-CNN module can be applied to different two-stage detector backbones
- The Pyramid-PV ranked 1st on the waymo dataset leaderboard for detecting vehicle using lidar only

Jiageng Mao, Minzhe Niu, Haoyue Bai, Xiaodan Liang, Hang Xu, and Chunjing Xu. 2021. Pyramid R-CNN: Towards Better Performance and Adaptability for 3D Object Detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Montreal, QC, Canada, 2703–2712. DOI:https://doi.org/10.1109/ICCV48922.2021.00272

Related Work

Single-stage 3D Object Detection

Single-stage detectors for 3D object detection can be divided into three categories **based on the type of input representation they use**: point-based, voxel-based, and pillar-based

- **Point-based methods**: These methods operate directly on point clouds
  - **3DSSD** (Yang et al., 2020)**,**
  - **Point-GNN**(Shi & Rajkumar, 2020)
- **Voxel-based methods** operate on voxel grids, three-dimensional grids of cubic cells that divide the 3D space into discrete volumes
  - **VoxNet**(Zhou & Tuzel, 2018)**,**
  - **SECOND**(Yan et al., 2018)**,**
  - **CenterPoint**(Yin et al., 2020)
- **Pillar-based approaches** involve changing the original 3D input point clouds into 2D simulations of an aerial view called Bird-Eye-View (BEV) pillars.
  - **PointPilar**(Lang et al., 2019)**,**
  - **Pilar-based Network**(Wang et al., 2020)

Two-stage 3D object detection

Two-stage approaches for 3D object detection can be divided into three categories **based on the representation of points of interest (POIs):** point-based, voxel-based, and point-voxel-based

- **Point-based methods**: These methods operate on sample input point clouds, as Points of Interest
  - **PointRCNN**(Shi et al., 2019)**,**
  - **STD**(Yang et al., 2019)
- **Voxel-based methods** uses voxel points from 3D CNN as Points of Interest (PoIs)
  - **Part-A^2 Net**(Shi et al., 2020)**,**
  - **Voxel R-CNN**(Deng et al., 2021)
- **Point-voxel-based methods:** use a set of points, called "keypoints", that represent the entire 3D scene as PoIs.
  - **PV-RCNN**(Shi et al., 2020)**,**
  - **PV-RCNN++**(Shi et al., 2022)*

Jiageng Mao, Minzhe Niu, Haoyue Bai, Xiaodan Liang, Hang Xu, and Chunjing Xu. 2021. Pyramid R-CNN: Towards Better Performance and Adaptability for 3D Object Detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Montreal, QC, Canada, 2703–2712. DOI:https://doi.org/10.1109/ICCV48922.2021.00272

Architecture

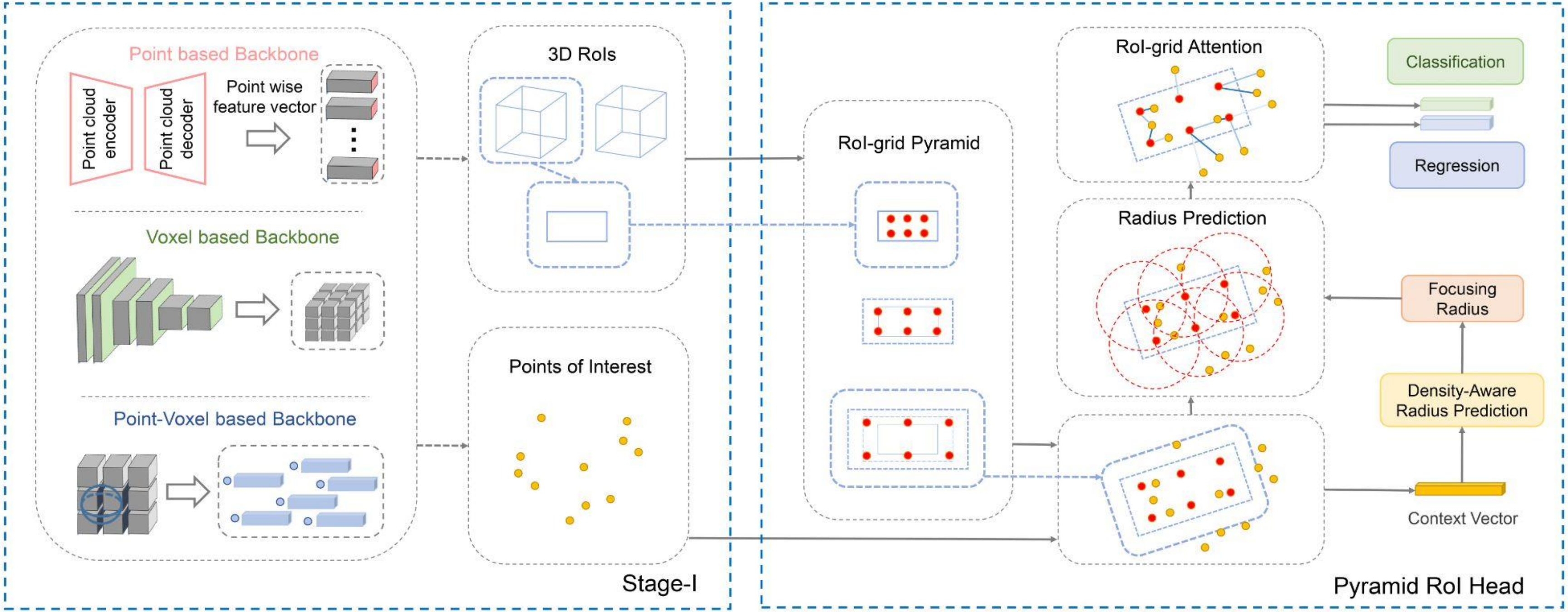


Figure 2. The overall architecture. Our Pyramid R-CNN can be plugged on diverse backbones (e.g. point-based, voxel-based and point- voxel-based networks)

J. Mao, M. Niu, H. Bai, X. Liang, H. Xu, and C. Xu, 'Pyramid R-CNN: Towards Better Performance and Adaptability for 3D Object Detection', in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, Oct. 2021, pp. 2703–2712, doi: 10.1109/ICCV48922.2021.00272 [Online]. Available: https://ieeexplore.ieee.org/document/9709959/. [Accessed: Dec. 31, 2022]

RoI-grid Pyramid

- RoI feature extraction creates an RoI-grid for each RoI
  - RoI-grid is made up of individual points that collect features from neighbouring PoI
  - Individual points in the RoI-grid are called **RoI-grid Points**
- RoI-grid point location p_{grid}^{ijk} can be computed as :

$$p_{\text{grid}}^{ijk} = \overbrace{\Bigg(\underbrace{\frac{W}{N_w}, \frac{L}{N_l}, \frac{H}{N_h}}_{\text{Grid sizes}}\Bigg)}^{\text{RoI width, length, and height}} \cdot \overbrace{(0.5 + (i, j, k))}^{\text{RoI-grid point location}} + \overbrace{(x_c, y_c, z_c)}^{\text{the bottom left corner in RoI-grid}}$$

- Grid points are generated inside RoIs
- Points of Interest (PoIs) are affected by sparsity and non-uniform distribution inside RoIs, which can result in difficulty defining object shape and structure (i.e incomplete shape)

Jiageng Mao, Minzhe Niu, Haoyue Bai, Xiaodan Liang, Hang Xu, and Chunjing Xu. 2021. Pyramid R-CNN: Towards Better Performance and Adaptability for 3D Object Detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Montreal, QC, Canada, 2703–2712. DOI:https://doi.org/10.1109/ICCV48922.2021.00272

RoI-grid Pyramid

- RoI-grid pyramid mitigate the above problem by capturing more Points of interest outside RoI
- RoI-grid pyramid balances information between fine grained and context(helps in the identification of incomplete objects)
- RoI-grid point location p_{grid}^{ijk} for a pyramid level can be computed as :

$$p_{\text{grid}}^{ijk} = \overbrace{\left(\underbrace{\frac{\rho_w W}{N'_w}, \frac{\rho_l L}{N'_l}, \frac{\rho_h H}{N'_h}}_{\text{Grid sizes}} \right)}^{\text{RoI width, length, and height}} \cdot \overbrace{(0.5 + (i, j, k))}^{\text{RoI-grid point location}} + \overbrace{(x_c, y_c, z_c)}^{\text{the bottom left corner in RoI-grid}}$$

ϱ : Determines how large we increase the original RoI size, it start from 1 at the base level
N' decreases as the pyramid level increases , initially, N = N'
The base of the pyramid captures fine grained information while at the top, it captures large context information
Each pyramid level have features of grid points which is aggregated using RoI-grid Attention

Jiageng Mao, Minzhe Niu, Haoyue Bai, Xiaodan Liang, Hang Xu, and Chunjing Xu. 2021. Pyramid R-CNN: Towards Better Performance and Adaptability for 3D Object Detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Montreal, QC, Canada, 2703–2712. DOI:https://doi.org/10.1109/ICCV48922.2021.00272

RoI-grid Pyramid

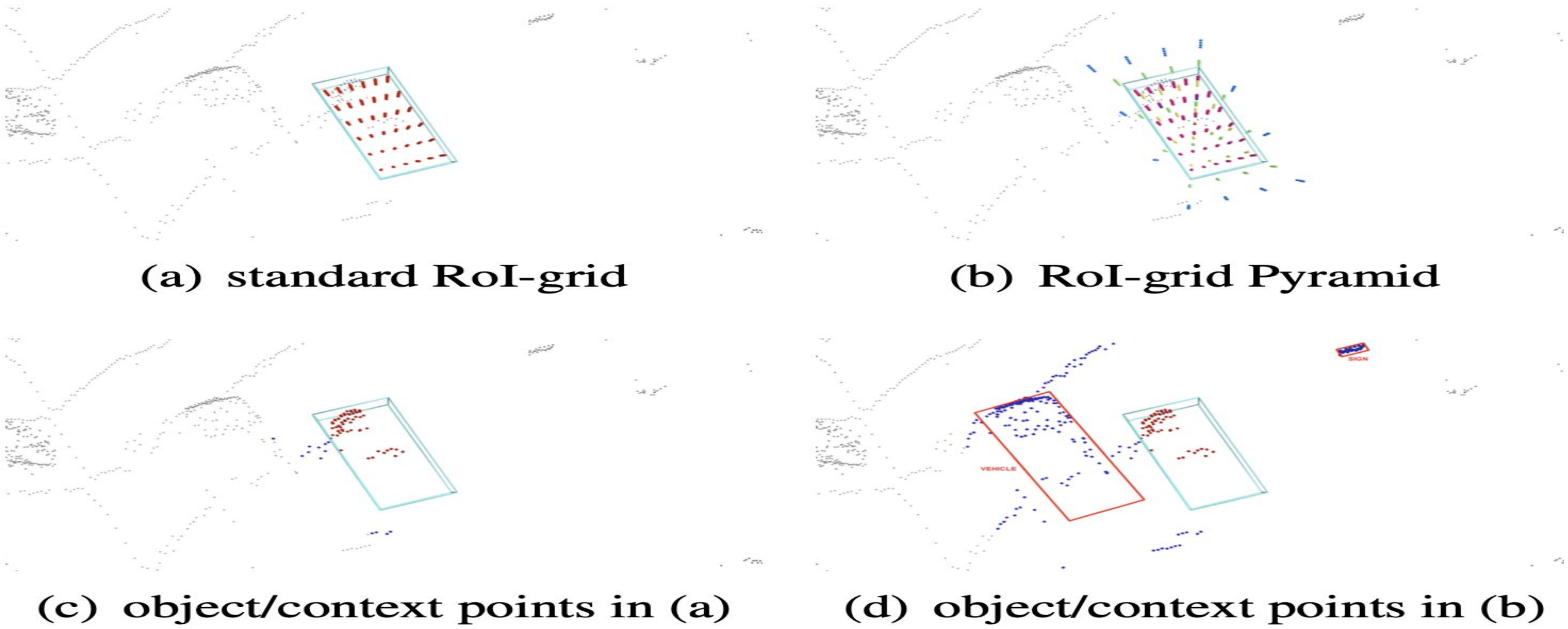


Figure 3. Illustration of the RoI-grid Pyramid

Jiageng Mao, Minzhe Niu, Haoyue Bai, Xiaodan Liang, Hang Xu, and Chunjing Xu. 2021. Pyramid R-CNN: Towards Better Performance and Adaptability for 3D Object Detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Montreal, QC, Canada, 2703–2712. DOI:https://doi.org/10.1109/ICCV48922.2021.00272

RoI-grid Attention

- **Pooling-based Operators:**

$$f^{\text{pool}}_{\text{grid}} = \underset{i \in \Omega(r)}{\text{maxpool}} \overbrace{\left(MLP\left(\left[f_i, p_i - p_{\text{grid}}\right]\right)\right)}^{V^i}$$

- **Graph-based Operators:**

$$f^{\text{graph}}_{\text{grid}} = \sum\nolimits_{i \in \Omega(r)} W \left(\underbrace{\text{Linear}\left(p_i - p_{\text{grid}}\right)}_{Q^i_{pos}}\right) \odot \underbrace{MLP\left(f_i\right)}_{V^i}$$

- **Attention-based Operators:**

$$f^{\text{atten}}_{\text{grid}} = \sum\nolimits_{i \in \Omega(\text{r})} W\left(Q^i_{\text{pos}}, K^i\right) \odot V^i$$

ρ_{grid}: Coordinate of an RoI-grid point

ρ_i : coordinate of the ith PoI near ρ_{grid}

f_i : feature vector of the ith PoI near ρ_{grid}

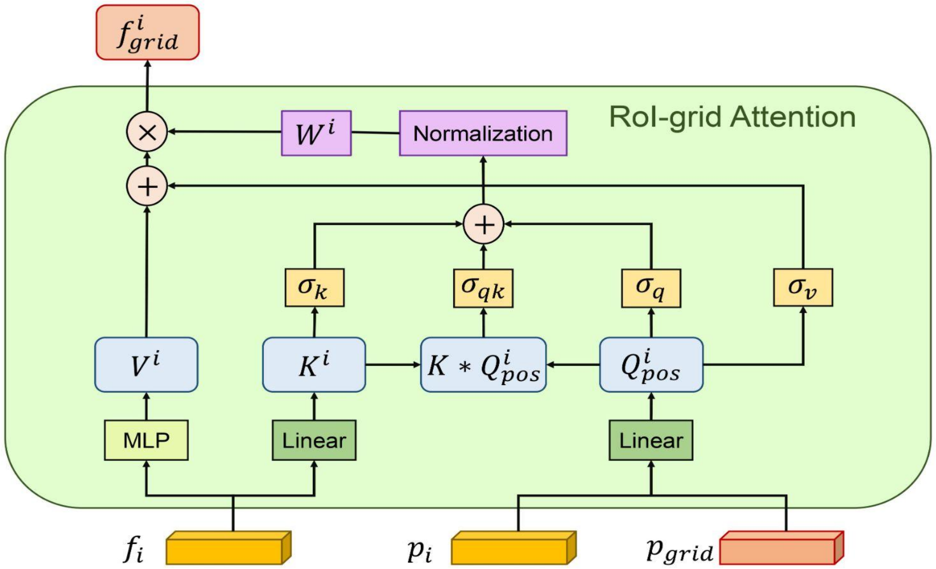
V_i: transformed feature vector

Ω(r): PoI within the fixed radius r of the RoI point ρ_{grid}

Q^i_{pos}: edge,linear projection of relative location

Q^i_{pos},K^i, V^i: query, key and value embedding

Jiageng Mao, Minzhe Niu, Haoyue Bai, Xiaodan Liang, Hang Xu, and Chunjing Xu. 2021. Pyramid R-CNN: Towards Better Performance and Adaptability for 3D Object Detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Montreal, QC, Canada, 2703–2712. DOI:https://doi.org/10.1109/ICCV48922.2021.00272



- RoI-grid Attention: combines graph-based operation and attention based operations
- RoI-grid attention is able geometric information Q_{pos} and Semantic information K as well as their combination $Q_{pos}K$ adaptively.

RoI-grid Attention:

$$f_{\text{grid}} = \sum_{i\in\Omega(r)} W\left(\sigma_k K^i + \sigma_q Q^i_{\text{pos}} + \sigma_{qk} Q^i_{\text{pos}} K^i\right) \odot \left(V^i + \sigma_v Q^i_{\text{pos}}\right)$$

$\boldsymbol{\sigma}_*$: learnable gated function, linear projection of embeddings with sigmoid activation output

Jiageng Mao, Minzhe Niu, Haoyue Bai, Xiaodan Liang, Hang Xu, and Chunjing Xu. 2021. Pyramid R-CNN: Towards Better Performance and Adaptability for 3D Object Detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Montreal, QC, Canada, 2703–2712. DOI:https://doi.org/10.1109/ICCV48922.2021.00272

Density-Aware Radius Prediction

From the Roi-grid Attention, we noticed the term **Ω**(r)

- r is an hyperparameter that determines the neighborhood of PoIs that participates in feature extraction.
- Fixed r are not adaptive and may result in an empty spherical range.
- What if we learn the radius r?
  - This is called Density Aware Radius Prediction module (DARP)
- In ROI-grid Attention, select PoIs within a radius r; perform a weighted combination of these points
- RoI grid attention can be reformulated as a probability.

Sampling from a conditional distribution

$$p(i \mid r)=\begin{cases}0 & \left\|p_i-p_{\text{grid}}\right\|_2>r \\ 1 & \left\|p_i-p_{\text{grid}}\right\|_2 \leq r\end{cases}$$

Probabilistic Expectation

$$f_{\text{grid}}=\mathbb{E}_{i \sim p(i \mid r)}\left[W^i \odot V^i\right]$$

Jiageng Mao, Minzhe Niu, Haoyue Bai, Xiaodan Liang, Hang Xu, and Chunjing Xu. 2021. Pyramid R-CNN: Towards Better Performance and Adaptability for 3D Object Detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Montreal, QC, Canada, 2703–2712. DOI:https://doi.org/10.1109/ICCV48922.2021.00272

Density-Aware Radius Prediction

- DARP proposed a new distribution s(i|r) similar to p(i|r)

$$s(i \mid r) = 1 - \text{sigmoid}\left(\frac{\left\|p_i - p_{\text{grid}}\right\|_2 - r}{\tau}\right)$$

$$sigmoid(x) = (1 + e^{-x})^{-1}$$

- The new formulation of RoI-grid Attention is:

$$f_{\text{grid}} = \sum_{i \in U(\epsilon)} W\left(\sigma_k K^i + \sigma_q Q^i_{\text{pos}} + \sigma_{qk} Q^i_{\text{pos}} K^i\right) \odot \left(V^i + \sigma_v Q^i_{\text{pos}}\right) \cdot s(i \mid r)$$

Jiageng Mao, Minzhe Niu, Haoyue Bai, Xiaodan Liang, Hang Xu, and Chunjing Xu. 2021. Pyramid R-CNN: Towards Better Performance and Adaptability for 3D Object Detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Montreal, QC, Canada, 2703–2712. DOI:https://doi.org/10.1109/ICCV48922.2021.00272

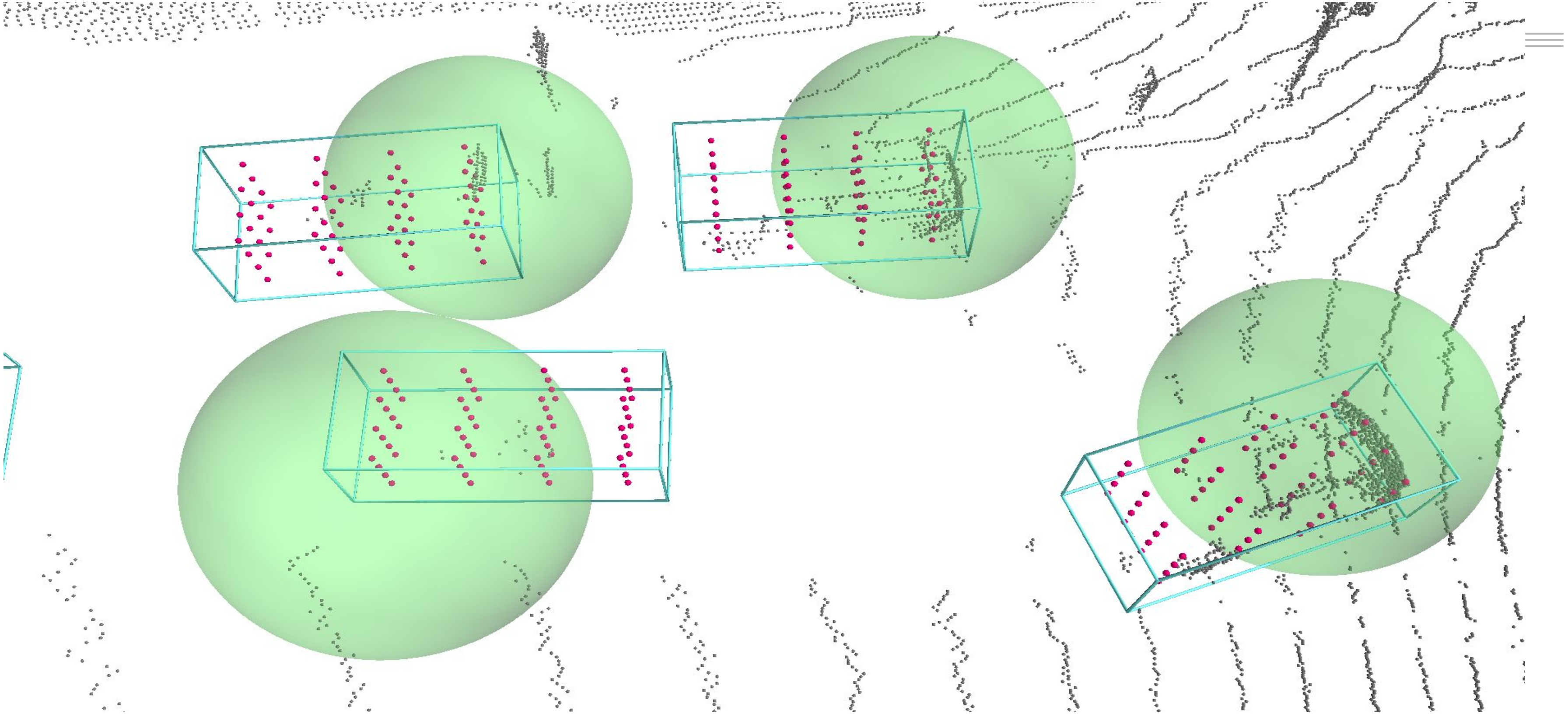


Figure 5. Illustration of dynamic radius predicted by the Density- Aware Radius Prediction module.

Jiageng Mao, Minzhe Niu, Haoyue Bai, Xiaodan Liang, Hang Xu, and Chunjing Xu. 2021. Pyramid R-CNN: Towards Better Performance and Adaptability for 3D Object Detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Montreal, QC, Canada, 2703–2712. DOI:https://doi.org/10.1109/ICCV48922.2021.00272

Experiment Setup

Waymo Open Dataset

- 1000 sequence
  - 798 sequence for training (158k point cloud samples)
  - 202 sequence for validation (40k point cloud samples)
- Evaluation Metrics
  - 3D mean Average Precision (mAP)
  - mAP weighted by heading accuracy (mAPH)
  - IoU threshold 0.7 for vehicles and 0.5 for other categories

KITTI Dataset

- Evaluation Metrics
  - Mean Average Precision (mAP)
  - Rotated IoU threshold 0.7 for cars
  - 11 recall position
- Test set
  - Mean Average Precision (mAP)
  - 40 recall position

Jiageng Mao, Minzhe Niu, Haoyue Bai, Xiaodan Liang, Hang Xu, and Chunjing Xu. 2021. Pyramid R-CNN: Towards Better Performance and Adaptability for 3D Object Detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Montreal, QC, Canada, 2703–2712. DOI:https://doi.org/10.1109/ICCV48922.2021.00272

Experiment Setup

Waymo Open Dataset

Test Set are divided into two categories

- Distance of object to sensor
  - 0-30m
  - 30-50m
  - >50m
- According to difficulty level
  - **LEVEL1**: boxes with >5 lidar point
  - **LEVEL2:** boxes with at least 1 lidar point

Backbone Architecture

- **PointRCNN**:
  - Replaced the canonical 3D box refinement module with pyramid RoI head
  - The resulting architecture is called **Pyramid-P.**
- **Part-A^2 Net:**
  - Replaced 3D sparse convolution head with pyramid RoI head
  - The resulting architecture is called **Pyramid-V**
- **PV-RCNN**:
  - Replaced the RoI-grid pooling with pyramid RoI head
  - The resulting architecture is called **Pyramid-PV**

Jiageng Mao, Minzhe Niu, Haoyue Bai, Xiaodan Liang, Hang Xu, and Chunjing Xu. 2021. Pyramid R-CNN: Towards Better Performance and Adaptability for 3D Object Detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Montreal, QC, Canada, 2703–2712. DOI:https://doi.org/10.1109/ICCV48922.2021.00272

Experiments - Implementation Details

- RoI-grid pyramid consist of :
  - 5 levels
  - With different number of grid points for each pyramid levels
  - 6^3, 4^3, 4^3, 4^3, 1 configuration respectively for each pyramid levels
  - Enlarging ratio ϱ_w and ϱ_l have the configuration 1,1,1.5,2,4 respectively for each pyramid levels
  - ϱ_h is set to a constant value 1, for all pyramid levels
- RoI-grid Attention:
  - 4 attention heads , with 16 feature channels
  - For each grid point, the maximum number of points used for RoI-grid Attention is 8, 16, 16, 16, 32 for pyramid level

Jiageng Mao, Minzhe Niu, Haoyue Bai, Xiaodan Liang, Hang Xu, and Chunjing Xu. 2021. Pyramid R-CNN: Towards Better Performance and Adaptability for 3D Object Detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Montreal, QC, Canada, 2703–2712. DOI:https://doi.org/10.1109/ICCV48922.2021.00272

Results : Comparisons on the Waymo Open Dataset

Methods	LEVEL_1 3D mAP/mAPH	LEVEL_2 3DmAP/mAPH	LEVEL_1 3D mAP/mAPH by Distance		
			0-30m	30-50m	50m-Inf
PointPillars(Lang et al., 2019)	63.3/62.7	55.2/54.7	84.9/84.4	59.2/58.6	35.8/35.2
MVF (Zhou et al., 2019)	62.93/-	-	86.30/-	60.02/-	36.02/-
Pillar-OD (Wang et al., 2020)	69.8/-	-	88.5/-	66.5/-	42.9/-
AFDet (Ge et al., 2020)	63.69/-	-	87.38/-	62.19/-	29.27/-
LaserNet (Meyer et al., 2019)	52.1/50.1	-	70.9/68.7	52.9/51.4	29.6/28.6
CVCNet (Chen et al., 2020)	65.2/-	-	86.80/-	62.19/-	29.27/-
StarNet(Ngiam et al., 2019)	64.7/56.3	45.5/39.6	83.3/82.4	58.8/53.2	34.3/25.7
RCD (Bewley et al., 2020)	69.0/68.5	-	87.2/86.8	66.5/66.1	44.5/44.0
Voxel R-CNN (Deng et al., 2021)	75.59/-	66.59/-	92.49/-	74.09/-	53.15/-
PointRCNN*(Shi et al., 2019)	45.05/44.25	37.41/36.74	72.24/71.31	31.21/30.41	23.77/23.15
Pyramid-P (ours)	**47.02/46.58**	**39.10/38.76**	**74.24/73.78**	**32.49/31.96**	**25.68/25.24**
Part-A^2 Net* (Shi et al., 2020)	71.69/71.16	64.21/63.70	91.83/91.37	69.99/69.37	46.26/45.41
Pyramid-V (ours)	**75.83/75.29**	**66.77/66.28**	**92.63/92.20**	**74.46/73.84**	**53.40/52.44**
PV-RCNN (Shi et al., 2020)	70.3/69.7	65.4/64.8	91.9/91.3	69.2/68.5	42.2/41.3
Pyramid-PV (ours)	**76.30/75.68**	**67.23/66.68**	**92.67/92.20 5**	**74.91/74.21**	**54.54/53.45**

Table 1:Performance comparison on the Waymo Open Dataset with 202 validation sequences for the vehicle detection. *: re-implemented by ourselves with the official code.

(Mao et al., 2021)

Results : Comparisons on the Waymo Open Dataset

Methods	LEVEL_1 3D mAP/mAPH	LEVEL_2 3DmAP/mAPH	LEVEL_1 3D mAP/mAPH by Distance		
			0-30m	30-50m	50m-Inf
CenterPoint* (Yin et al., 2020)	81.05/80.59	73.42/72.99	92.52/92.13	79.94/79.43	61.06/60.,42
PV-RCNN* (Shi et al., 2020)	81.06/80.57	73.69/73.23	93.40/92.98	80.12/79.57	61.22/60.47
Pyramid-PV^{+} (ours)	**81.77/81.32**	**74.87/74.43**	**93.19/92.80**	**80.53/80.04**	**64.55/63.84**

Table2: Performance comparison on the Waymo Open Dataset test leaderboard for the vehicle detection. *: test submissions are the modified version of original architectures.

Jiageng Mao, Minzhe Niu, Haoyue Bai, Xiaodan Liang, Hang Xu, and Chunjing Xu. 2021. Pyramid R-CNN: Towards Better Performance and Adaptability for 3D Object Detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Montreal, QC, Canada, 2703–2712. DOI:https://doi.org/10.1109/ICCV48922.2021.00272

Results:

Comparisons on the KITTI Dataset

Table3: Performance comparison on the KITTI test set with AP calculated by 40 recall positions for the car category.

(Mao et al., 2021)

Methods	Modality	AP_{3D}(%)		
		Easy	Medium	Had
MV3D (Chen et al., 2017)	R+L	74.97	63.63	54.0
AVOD-FPN (Ku et al., 2018)	R+L	83.07	71.76	65.73
F-PointNet (Qi et al., 2018)	R+L	82.19	69.79	60.59
MMF (Liang et al., 2019)	R+L	88.40	77.43	70.22
3D-CVF (Yoo et al., 2020)	R+L	89.20	80.05	73.11
CLOCs (Pang et al., 2020)	R+L	88.94	80.67	77.15
ContFuse (Liang et al., 2018)	R+L	83.68	68.78	61.67
VoxelNet (Zhou & Tuzel, 2018)	L	77.47	65.11	57.73
PointPillars (Lang et al., 2019)	L	82.58	74.31	68.99
SECOND (Yan et al., 2018)	L	84.65	75.96	68.71
STD (Yang et al., 2019)	L	87.95	79.71	75.09
Patches (Lehner et al., 2019)	L	88.67	77.20	71.82
3DSSD (Yang et al., 2020)	L	88.36	79.57	74.55
SA-SSD (He et al., 2020)	L	88.75	79.79	74.16
TANet (Liu et al., 2020)	L	85.94	75.76	68.32
Voxel R-CNN (Deng et al., 2021)	L	90.90	81.62	77.06
HVNet (Ye et al., 2020)	L	87.21	77.58	71.79
PointGNN (Shi & Rajkumar, 2020)	L	88.33	79.47	72.29
PointRCNN (Shi et al., 2019)	L	86.96	75.64	70.70
Pyramid-P (ours)	L	**87.03**	**80.30**	**76.48**
Part- A^2 Net (Shi et al., 2020)	L	**87.81**	78.49	73.51
Pyramid-V (ours)	L	87.06	**81.28**	**76.85**
PV-RCNN (Shi et al., 2020)	L	**90.25**	81.43	76.8
Pyramid-PV (ours)	L	88.39	**82.08**	**77.49**

Results- Comparisons on the KITTI Dataset

Methods	AP_{3D} (%)		
	Easy	Medium	Hard
PointRCNN (Shi et al., 2019)	88.88	78.63	77.38
Pyramid-P (ours)	**88.47**	**83.10**	**78.44**
Part-A^2 Net (Shi et al., 2020)	89.47	79.47	78.54
Pyramid-V (ours)	**88.44**	**83.141**	**78.61**
PV-RCNN (Shi et al., 2020)	89.35	83.69	78.70
Pyramid-PV (ours)	**89.37**	**84.38**	**78.84**

Table4: Performance comparison on the KITTI val split with AP calculated by 11 recall positions for the car category

Jiageng Mao, Minzhe Niu, Haoyue Bai, Xiaodan Liang, Hang Xu, and Chunjing Xu. 2021. Pyramid R-CNN: Towards Better Performance and Adaptability for 3D Object Detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Montreal, QC, Canada, 2703–2712. DOI:https://doi.org/10.1109/ICCV48922.2021.00272

Ablation Studies

Methods	R.P	D.A.R.P	R.A	LEVEL_1 mAP
PV-RCNN				70.30
PV-RCNN *				74.06
(a)	✓			75.26
(b)	✓	✓		75.63
(c)	✓		✓	75.77
(d)	✓	✓	✓	**76.30**

Table5: Effects of different components in Pyramid-PV on the Waymo dataset.

Methods	Inference speed (Hz)
PointRCNN (Shi et al., 2019)	10.08
Pyramid-PV (ours)	8.92
Part-A^2 Net (Shi et al., 2020)	11.75
Pyramid-PV (ours)	9.68
PV-RCNN (Shi et al., 2020)	9.25
Pyramid-PV (ours)	7.86

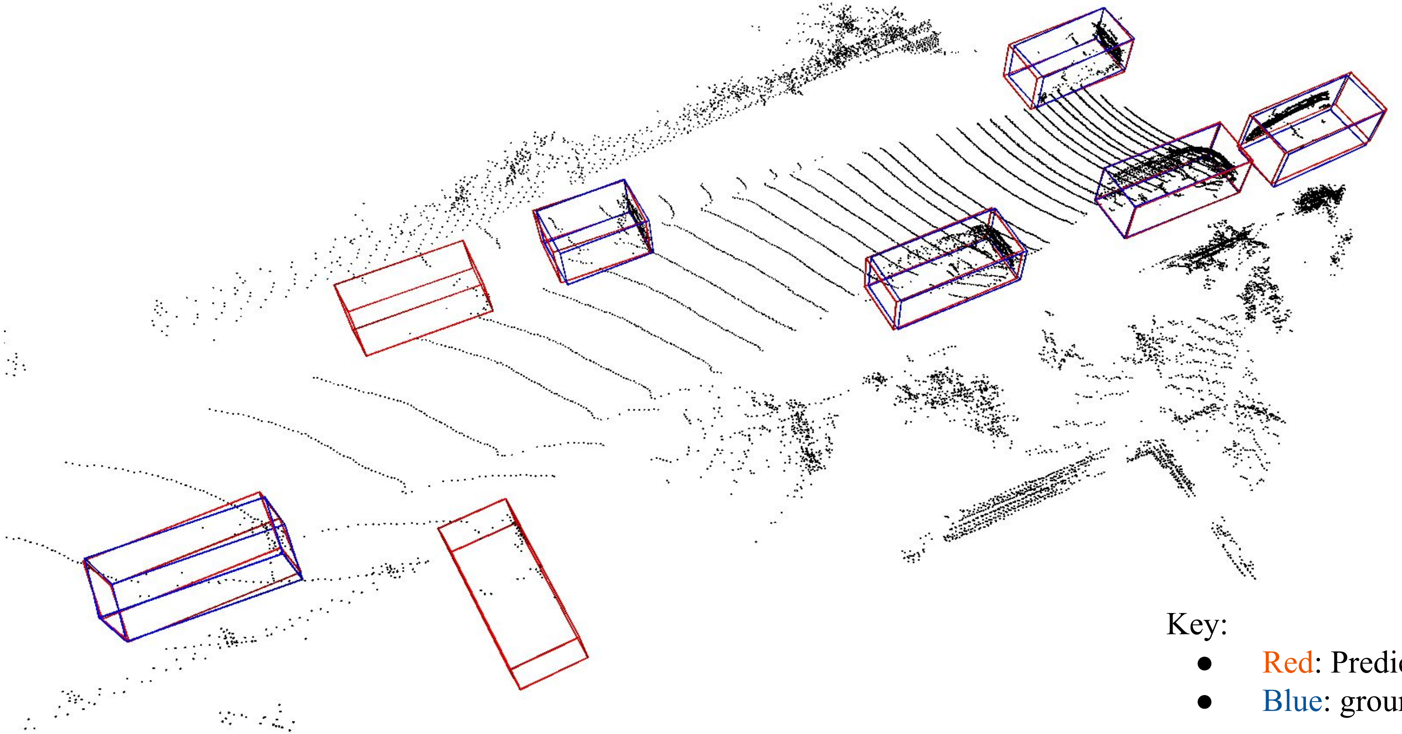
Table 6: Comparisons on the inference speeds of different detection models on the KITTI dataset.

Methods	grid size	ϱ_w, ϱ_l	LEVEL_1 mAP
PV-RCNN	[6, 6]	[1, 1]	74.06
(a)	[6,4,4]	[1,1,2]	74.55
(b)	[6,4,4,4]	[1,1,2,4]	74.71
(c)	[6,4,4,4,1]	[1,1,1.5,2,4]	**75.26**

Table 6:Effects of different RoI pyramids in Pyramid-PV on the Waymo dataset. Each element in [·] stands for the respective parameter of a pyramid level.

Jiageng Mao, Minzhe Niu, Haoyue Bai, Xiaodan Liang, Hang Xu, and Chunjing Xu. 2021. Pyramid R-CNN: Towards Better Performance and Adaptability for 3D Object Detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Montreal, QC, Canada, 2703–2712. DOI:https://doi.org/10.1109/ICCV48922.2021.00272

Results- Visualization on KITTI dataset

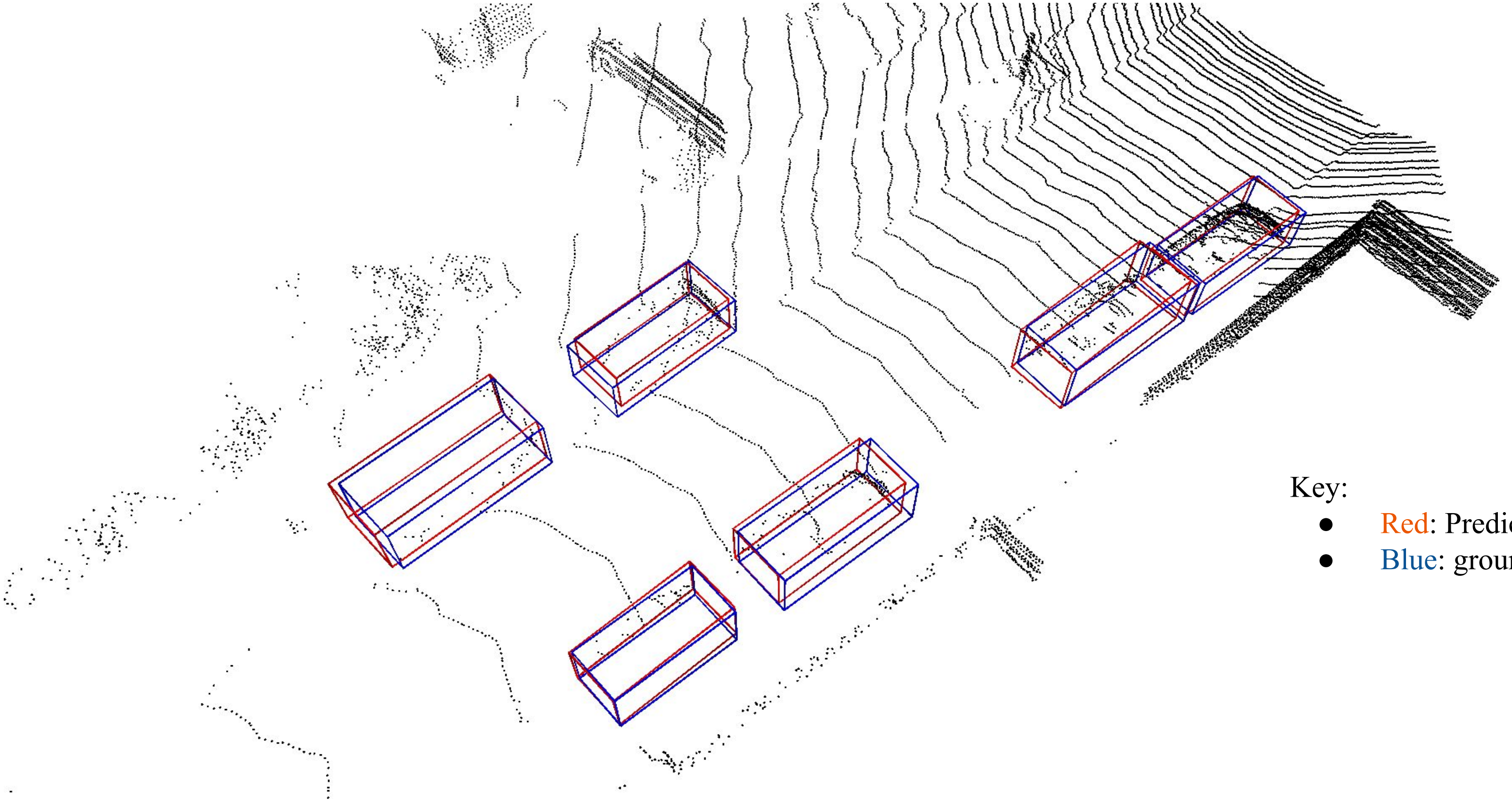


Key:

- Red: Prediction
- Blue: ground truth

(Mao et al., 2021)

Results- Visualization on KITTI dataset

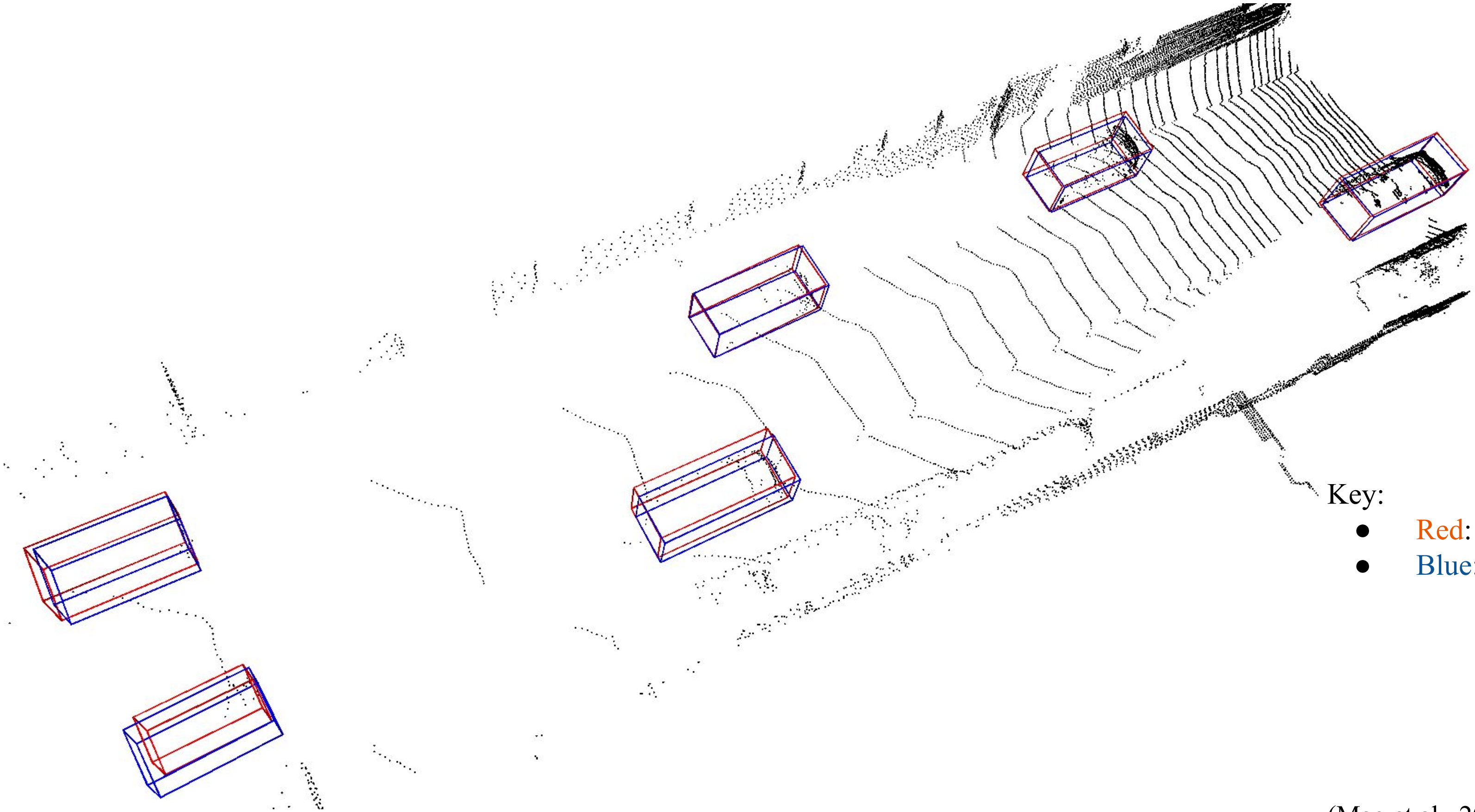


Key:

- Red: Prediction
- Blue: ground truth

(Mao et al., 2021)

Results- Visualization on KITTI dataset



Key:

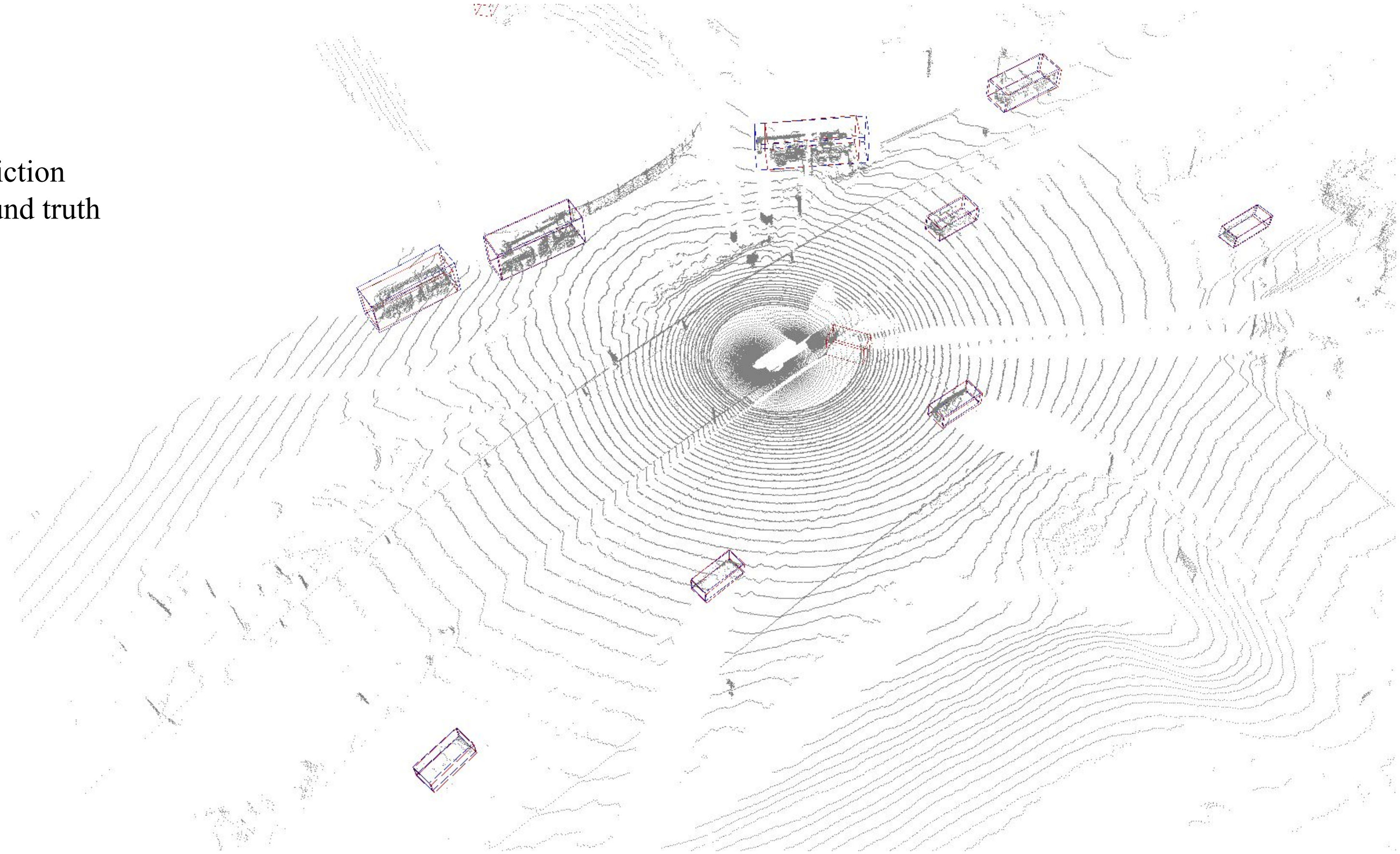
- Red: Prediction
- Blue: ground truth

(Mao et al., 2021)

Results- Visualization on Waymo Open dataset

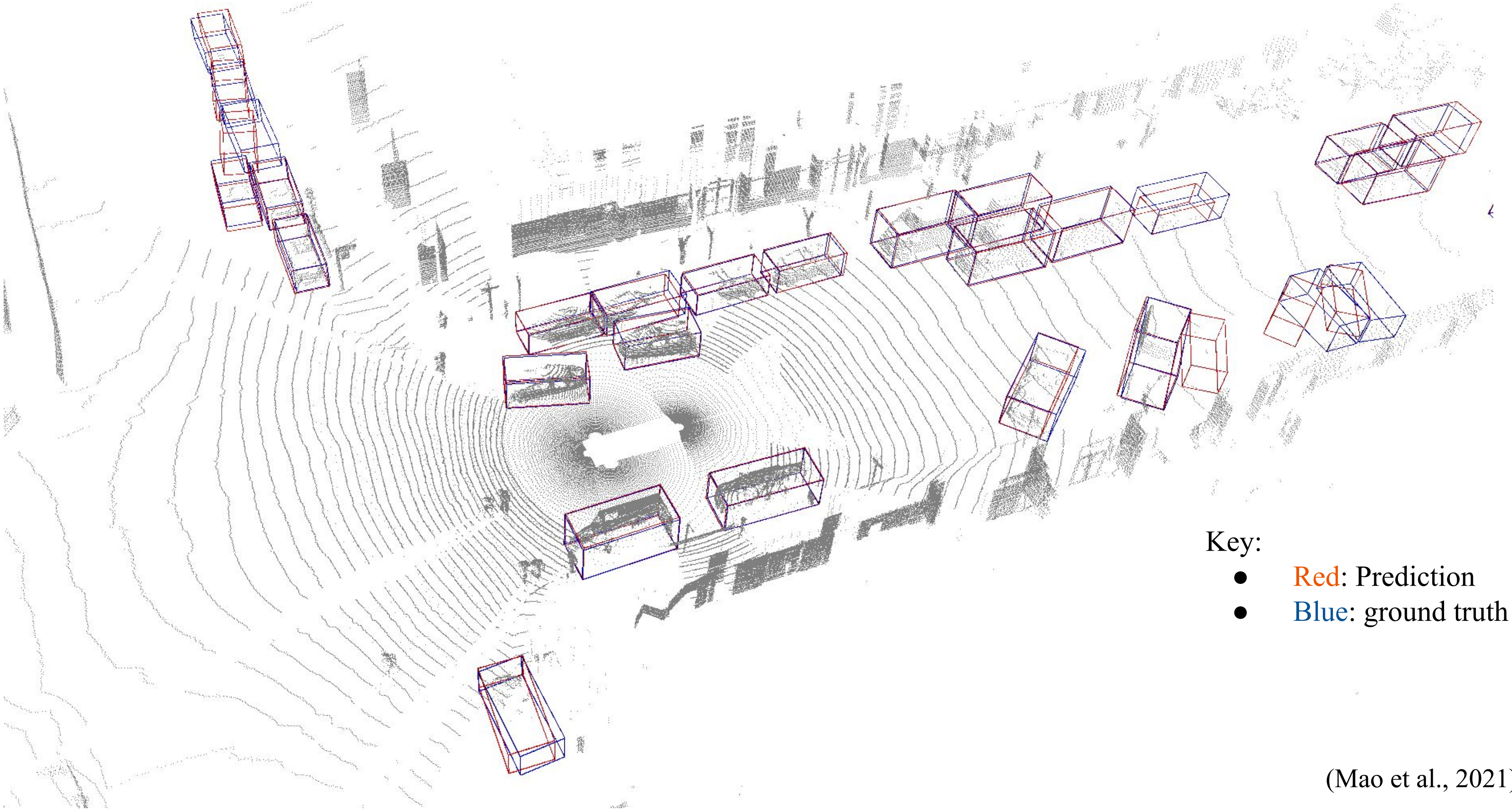
Key:

- Red: Prediction
- Blue: ground truth



(Mao et al., 2021)

Results- Visualization on Waymo Open dataset

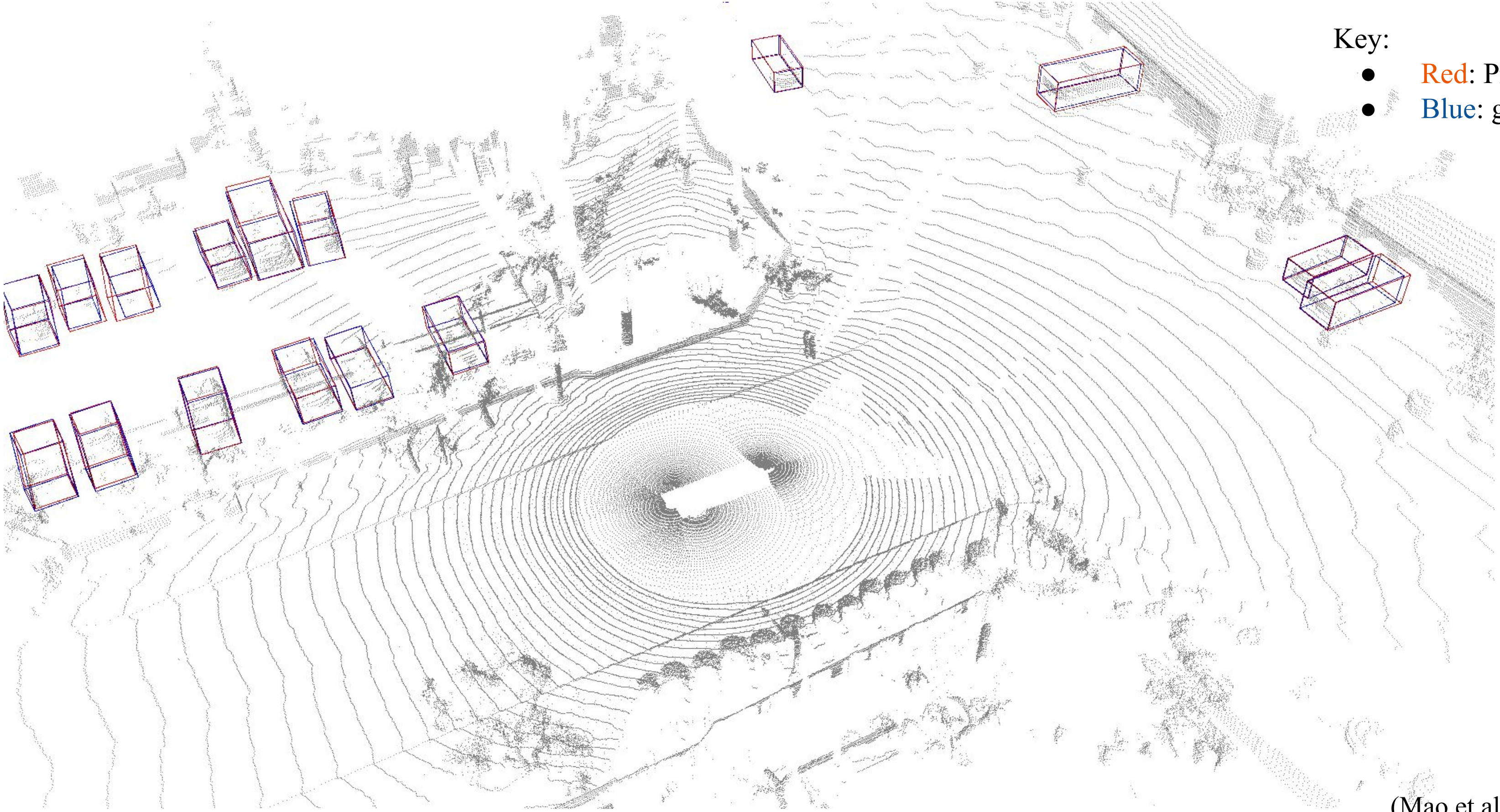


Key:

- Red: Prediction
- Blue: ground truth

(Mao et al., 2021)

Results- Visualization on Waymo Open dataset



Key:

- Red: Prediction
- Blue: ground truth

(Mao et al., 2021)

Conclusion

- The authors proposed a robust novel second-stage feature extraction module called Pyramid R-CNN, which mitigates the sparsity and non-uniform distribution of input point clouds.
- Pyramid R-CNN can be used with different two-stage detectors backbones
- Pyramid-PV achieved a state-of-the-art result on the Waymo open dataset

Jiageng Mao, Minzhe Niu, Haoyue Bai, Xiaodan Liang, Hang Xu, and Chunjing Xu. 2021. Pyramid R-CNN: Towards Better Performance and Adaptability for 3D Object Detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Montreal, QC, Canada, 2703–2712. DOI:https://doi.org/10.1109/ICCV48922.2021.00272

References

- Bewley, A., Sun, P., Mensink, T., Anguelov, D., & Sminchisescu, C. (2020). *Range conditioned dilated convolutions for scale invariant 3d object detection*. https://doi.org/10.48550/ARXIV.2005.09927
- Chen, Q., Sun, L., Cheung, E., & Yuille, A. L. (2020). Every view counts: Cross-view consistency in 3d object detection with hybrid-cylindrical-spherical voxelization. *Advances in Neural Information Processing Systems*, *33*, 21224–21235. https://proceedings.neurips.cc/paper/2020/hash/f2fc990265c712c49d51a18a32b39f0c-Abstract.html
- Chen, X., Ma, H., Wan, J., Li, B., & Xia, T. (2017). Multi-view 3d object detection network for autonomous driving. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6526–6534. https://doi.org/10.1109/CVPR.2017.691
- Deng, J., Shi, S., Li, P., Zhou, W., Zhang, Y., & Li, H. (2021). Voxel r-cnn: Towards high performance voxel-based 3d object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(2), 1201–1209. https://doi.org/10.1609/aaai.v35i2.16207
- Ge, R., Ding, Z., Hu, Y., Wang, Y., Chen, S., Huang, L., & Li, Y. (2020). *Afdet: Anchor free one stage 3d object detection*. https://doi.org/10.48550/ARXIV.2006.12671
- He, C., Zeng, H., Huang, J., Hua, X.-S., & Zhang, L. (2020). Structure aware single-stage 3d object detection from point cloud. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11870–11879. https://doi.org/10.1109/CVPR42600.2020.01189
- Ku, J., Mozifian, M., Lee, J., Harakeh, A., & Waslander, S. L. (2018). Joint 3d proposal generation and object detection from view aggregation. *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1–8. https://doi.org/10.1109/IROS.2018.8594049
- Lang, A. H., Vora, S., Caesar, H., Zhou, L., Yang, J., & Beijbom, O. (2019). Pointpillars: Fast encoders for object detection from point clouds. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12689–12697. https://doi.org/10.1109/CVPR.2019.01298
- Lehner, J., Mitterecker, A., Adler, T., Hofmarcher, M., Nessler, B., & Hochreiter, S. (2019). *Patch refinement—Localized 3d object detection*. https://doi.org/10.48550/ARXIV.1910.04093
- Liang, M., Yang, B., Chen, Y., Hu, R., & Urtasun, R. (2019). Multi-task multi-sensor fusion for 3d object detection. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7337–7345. https://doi.org/10.1109/CVPR.2019.00752

- Liang, M., Yang, B., Wang, S., & Urtasun, R. (2018). Deep continuous fusion for multi-sensor 3d object detection. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer Vision – ECCV 2018* (pp. 663–678). Springer International Publishing. https://doi.org/10.1007/978-3-030-01270-0_39
- Liu, Z., Zhao, X., Huang, T., Hu, R., Zhou, Y., & Bai, X. (2020). Tanet: Robust 3d object detection from point clouds with triple attention. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(07), 11677–11684. https://doi.org/10.1609/aaai.v34i07.6837
- Mao, J., Niu, M., Bai, H., Liang, X., Xu, H., & Xu, C. (2021). Pyramid r-cnn: Towards better performance and adaptability for 3d object detection. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2703–2712. https://doi.org/10.1109/ICCV48922.2021.00272
- Meyer, G. P., Laddha, A., Kee, E., Vallespi-Gonzalez, C., & Wellington, C. K. (2019). Lasernet: An efficient probabilistic 3d object detector for autonomous driving. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12669–12678. https://doi.org/10.1109/CVPR.2019.01296
- Ngiam, J., Caine, B., Han, W., Yang, B., Chai, Y., Sun, P., Zhou, Y., Yi, X., Alsharif, O., Nguyen, P., Chen, Z., Shlens, J., & Vasudevan, V. (2019). *Starnet: Targeted computation for object detection in point clouds*. https://doi.org/10.48550/ARXIV.1908.11069
- Pang, S., Morris, D., & Radha, H. (2020). *Clocs: Camera-lidar object candidates fusion for 3d object detection*. https://doi.org/10.48550/ARXIV.2009.00784
- Qi, C. R., Liu, W., Wu, C., Su, H., & Guibas, L. J. (2018). Frustum pointnets for 3d object detection from rgb-d data. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 918–927. https://doi.org/10.1109/CVPR.2018.00102
- Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., & Li, H. (2020). Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10526–10535. https://doi.org/10.1109/CVPR42600.2020.01054
- Shi, S., Wang, X., & Li, H. (2019). Pointrcnn: 3d object proposal generation and detection from point cloud. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–779. https://doi.org/10.1109/CVPR.2019.00086
- Shi, S., Wang, Z., Shi, J., Wang, X., & Li, H. (2020). From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1. https://doi.org/10.1109/TPAMI.2020.2977026

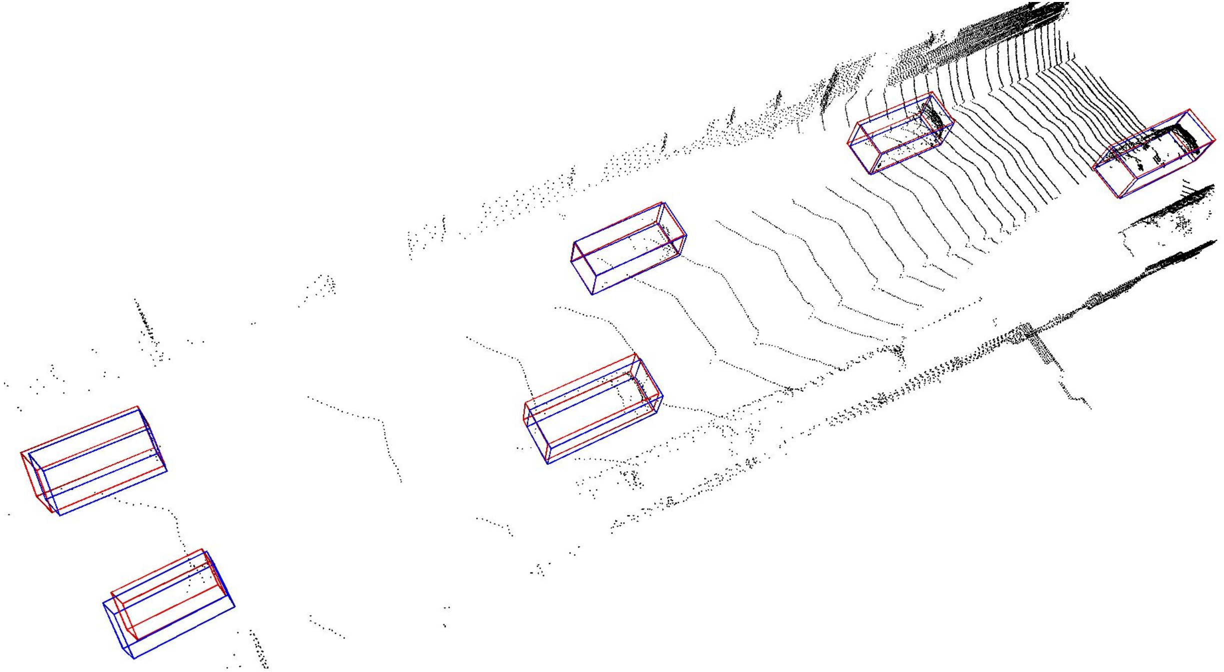
- Shi, W., & Rajkumar, R. (2020). Point-gnn: Graph neural network for 3d object detection in a point cloud. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1708–1716. https://doi.org/10.1109/CVPR42600.2020.00178
- Wang, Y., Fathi, A., Kundu, A., Ross, D. A., Pantofaru, C., Funkhouser, T., & Solomon, J. (2020). Pillar-based object detection for autonomous driving. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer Vision – ECCV 2020* (pp. 18–34). Springer International Publishing. https://doi.org/10.1007/978-3-030-58542-6_2
- Yan, Y., Mao, Y., & Li, B. (2018). Second: Sparsely embedded convolutional detection. *Sensors*, *18*(10), 3337. https://doi.org/10.3390/s18103337
- Yang, Z., Sun, Y., Liu, S., & Jia, J. (2020). 3dssd: Point-based 3d single stage object detector. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11037–11045. https://doi.org/10.1109/CVPR42600.2020.01105
- Yang, Z., Sun, Y., Liu, S., Shen, X., & Jia, J. (2019). Std: Sparse-to-dense 3d object detector for point cloud. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1951–1960. https://doi.org/10.1109/ICCV.2019.00204
- Ye, M., Xu, S., & Cao, T. (2020). Hvnet: Hybrid voxel network for lidar based 3d object detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1628–1637. https://doi.org/10.1109/CVPR42600.2020.00170
- Yin, T., Zhou, X., & Krähenbühl, P. (2020). *Center-based 3d object detection and tracking*. https://doi.org/10.48550/ARXIV.2006.11275
- Yoo, J. H., Kim, Y., Kim, J., & Choi, J. W. (2020). 3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer Vision – ECCV 2020* (pp. 720–736). Springer International Publishing. https://doi.org/10.1007/978-3-030-58583-9_43
- Zhou, Y., Sun, P., Zhang, Y., Anguelov, D., Gao, J., Ouyang, T., Guo, J., Ngiam, J., & Vasudevan, V. (2019). *End-to-end multi-view fusion for 3d object detection in lidar point clouds*. https://doi.org/10.48550/ARXIV.1910.06528
- Zhou, Y., & Tuzel, O. (2018). Voxelnet: End-to-end learning for point cloud based 3d object detection. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4490–4499. https://doi.org/10.1109/CVPR.2018.00472
- Mao, J., Wang, X., & Li, H. (2019). Interpolated convolutional networks for 3d point cloud understanding. arXiv. https://doi.org/10.48550/arXiv.1908.04512

Thank you.

Question 1

What are the major causes of sparsity in point clouds?

- Distance between the sensor and the object:The further an object is from the sensor, the longer it takes the beam to return to the sensor; thus, fewer points are collected.
- Certain object surfaces can cause sparsity in point clouds, e.g., objects with low reflective surfaces or transparent objects are difficult to detect by lidar. Similarly, objects with a very smooth surface are also difficult to detect by lidar.



Question 2

The RoI-grid Attention combines attention-base operation and graph-based operation. So how can we reduce RoI-grid Attention to the individual components/operators it is made from?

$$f_{\text{grid}} = \sum_{i \in \Omega(r)} W \left(\sigma_k K^i + \sigma_q Q^i_{\text{pos}} + \sigma_{qk} Q^i_{\text{pos}} K^i \right) \odot \left(V^i + \sigma_v Q^i_{\text{pos}} \right)$$

- Graph-based Operators: set the values of $\boldsymbol{\sigma}_q \boldsymbol{\sigma}_k \boldsymbol{\sigma}_{qk} \boldsymbol{\sigma}_v$ to 1,0,0,0 respectively
- Attention-based Operators: set the values of $\boldsymbol{\sigma}_q \boldsymbol{\sigma}_k \boldsymbol{\sigma}_{qk} \boldsymbol{\sigma}_v$ to 0,0,1,0 respectively

Question 3

What type(s) of attention is in RoI-grid Attention?

A. Self -attention
B. Cross-attention
C. Mixed- attention