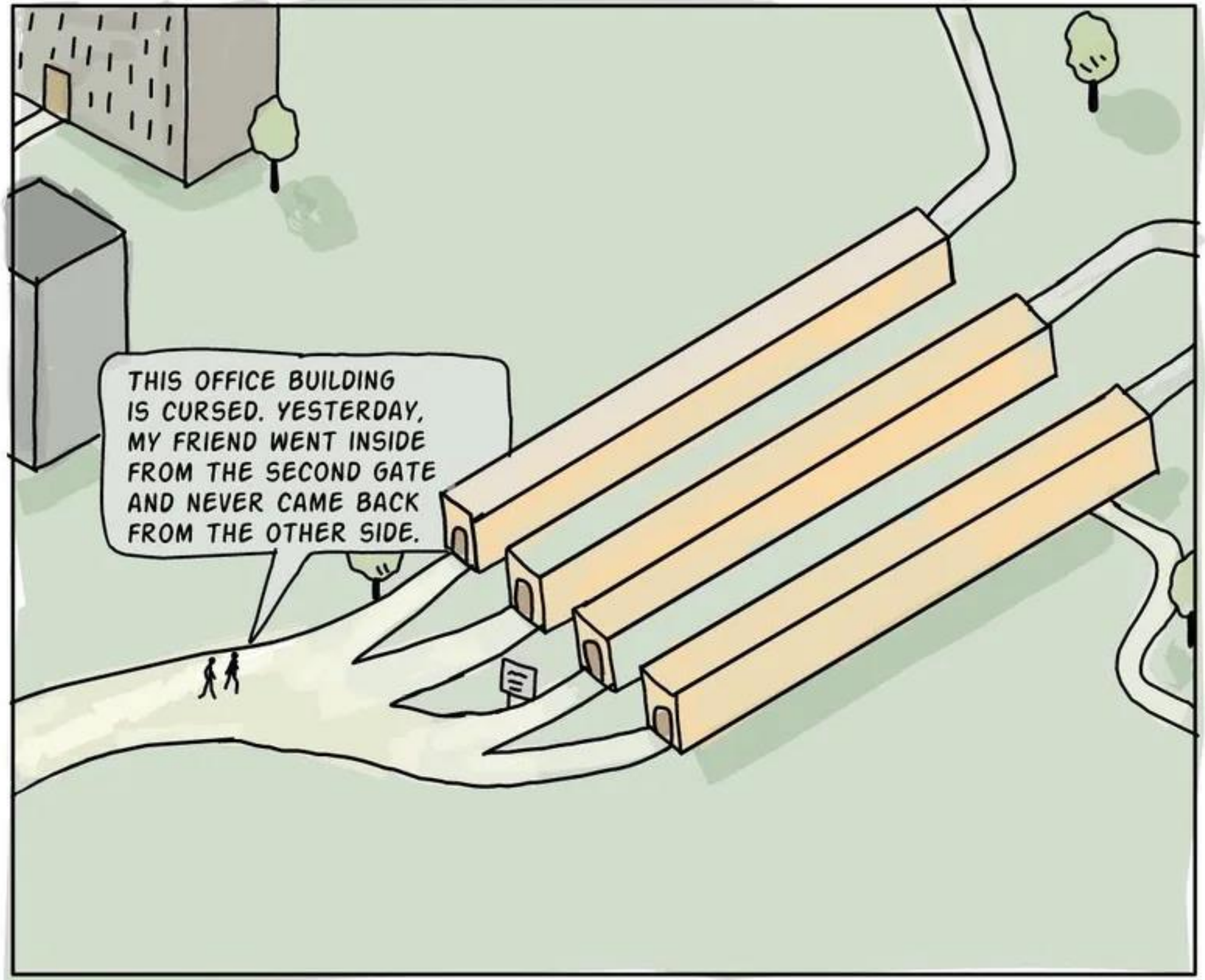


https://en.wikipedia.org/wiki/Impossible_trident

3D Point Processing

James Hays

Recap: Self Supervised Learning


I'd love to go back in time and blow people's minds with my incredible knowledge!
But how do you make this
Large Pretrained Model?
I don't know


I'd love to go back in time and blow people's minds with my incredible knowledge!
But how do you make this
Large Pretrained Model?
Some combination of self supervision from missing patch prediction (MAE) and invariance to augmentations (SimCLR) works without any human labels.

Dino v3

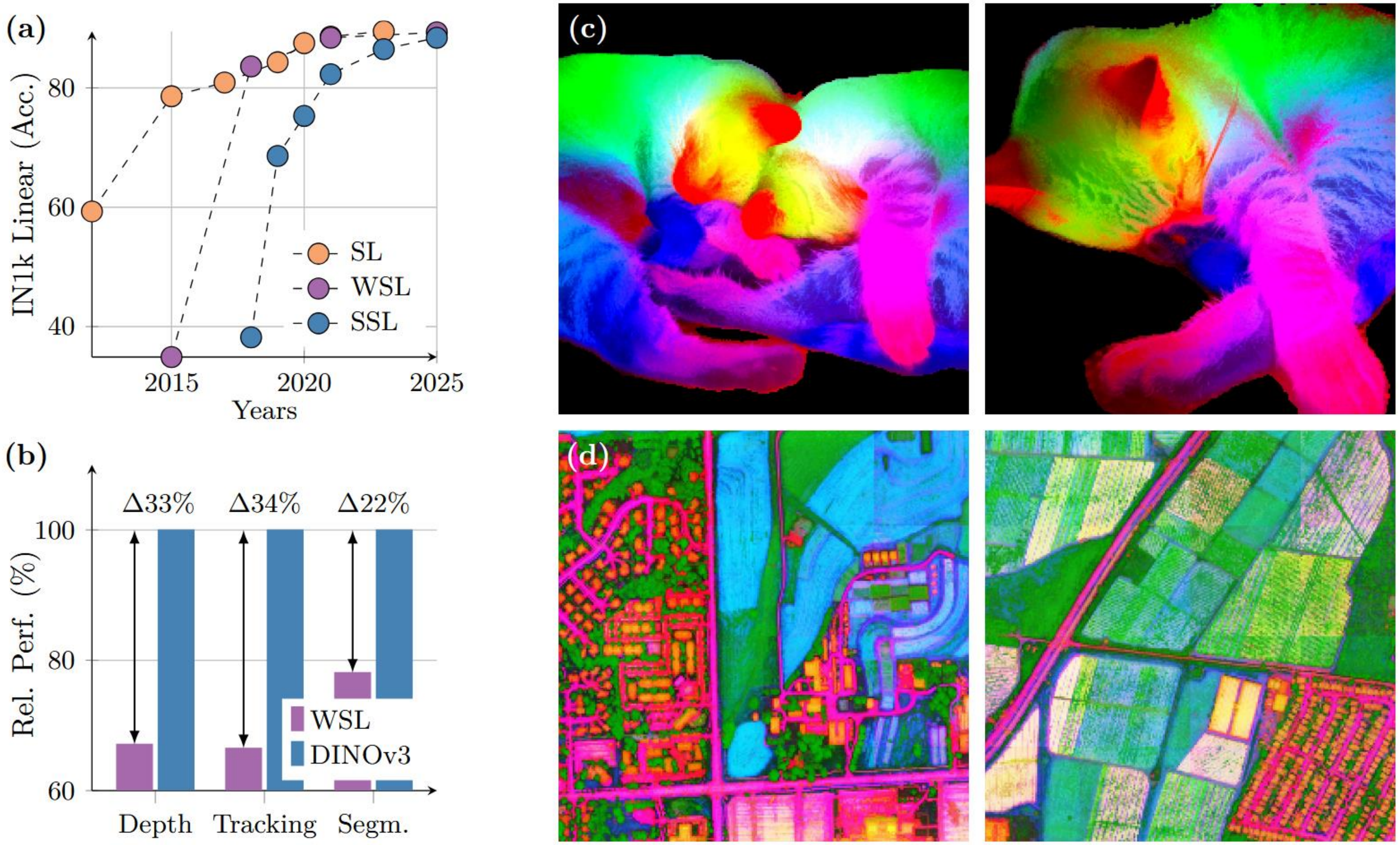
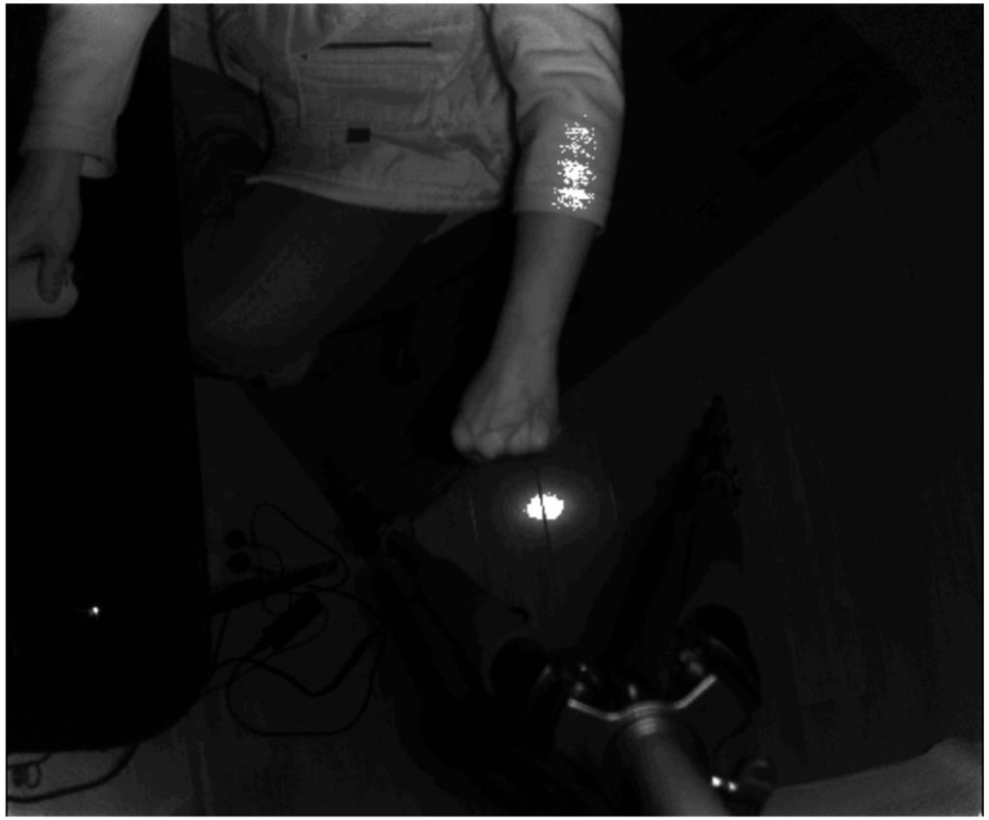


Figure 1: (a) Evolution of linear probing results on ImageNet1k (IN1k) over the years, comparing fully- (SL), weakly- (WSL) and self-supervised learning (SSL) methods. Despite coming into the picture later, SSL has quickly progressed and now reached the Imagenet accuracy plateau of recent years. On the other hand, we demonstrate that SSL offers the unique promise of high-quality dense features. With DINOv3, we markedly improve over weakly-supervised models on dense tasks, as shown by the relative performance of the best-in-class WSL models to DINOv3 (b). We also produce PCA maps of features obtained from high resolution images with DINOv3 trained on natural (c) and aerial images (d).

3D Point Processing Outline

- How do we measure 3D points?
- How do we make decisions about point clouds?
    - PointNet – orderless point processing
    - VoxelNet – voxel-based point processing
    - PointPillars – bird’s eye view point processing
        - Exploiting Visibility for 3D Object Detection
    - Range view object detection

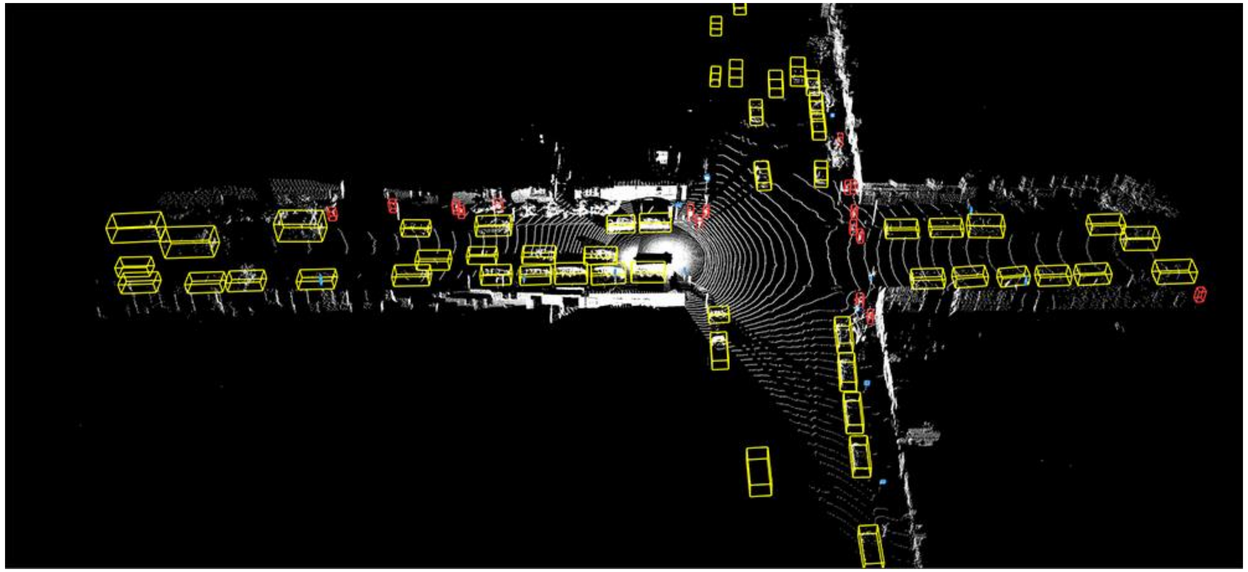
Kinect V1 and V2



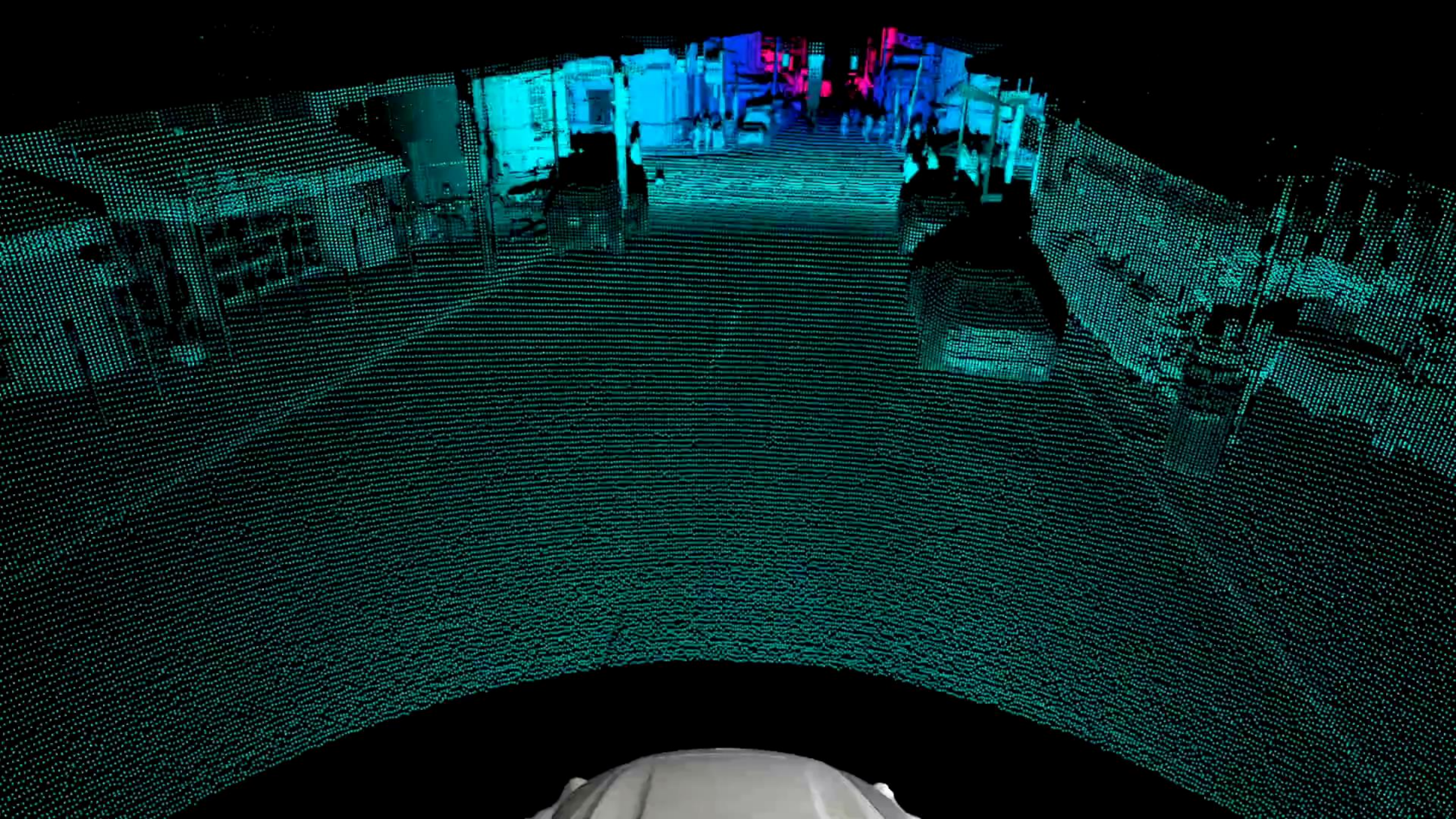
Infrared images of Kinect V1 structured light pattern and Kinect V2 time of flight pattern. Credit "Lightweight Algorithms for Depth Sensor Equipped Embedded Devices" by Henry Zhong

Lidar overview

Lidar overview



Source: Waymo Open Dataset



Some datasets are entirely synthetic, though

ModelNet – CAD models selected to match common categories in the SUN dataset

Figure 5: **ModelNet Dataset.** Left: word cloud visualization of the ModelNet dataset based on the number of 3D models in each category. Larger font size indicates more instances in the category. Right: Examples of 3D chair models.

Outline

- What is lidar?
- How do we make decisions about point clouds?
  - PointNet – orderless point processing
  - VoxelNet – voxel-based point processing
  - PointPillars – bird's eye view point processing
    - Exploiting Visibility for 3D Object Detection
  - Range view object detection

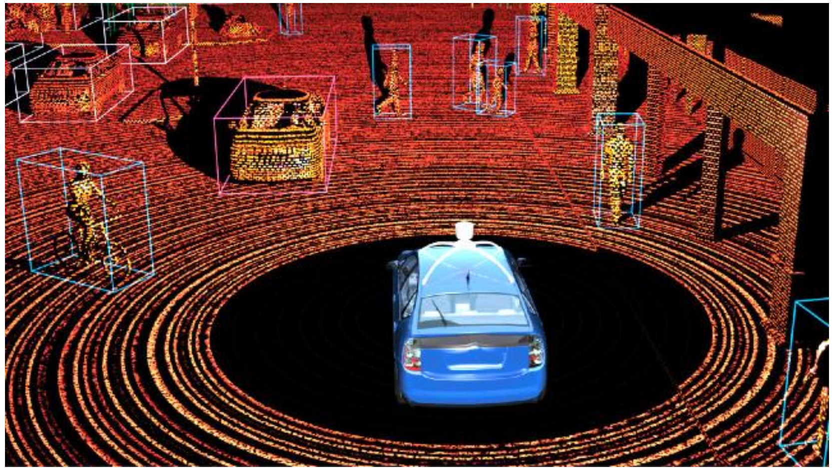
PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation

Charles R. Qi* Hao Su* Kaichun Mo Leonidas J. Guibas

Stanford University

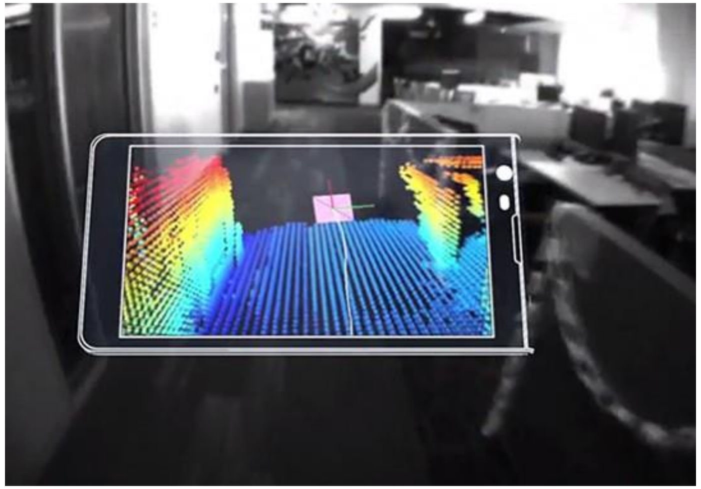
Big Data + Deep Representation Learning

Robot Perception



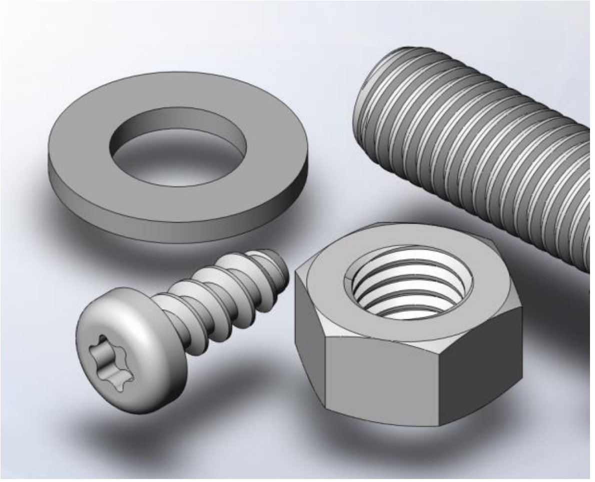
source: Scott J Grunewald

Augmented Reality



source: Google Tango

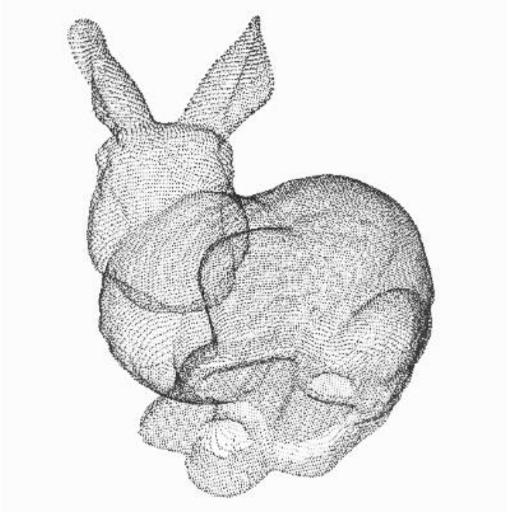
Shape Design



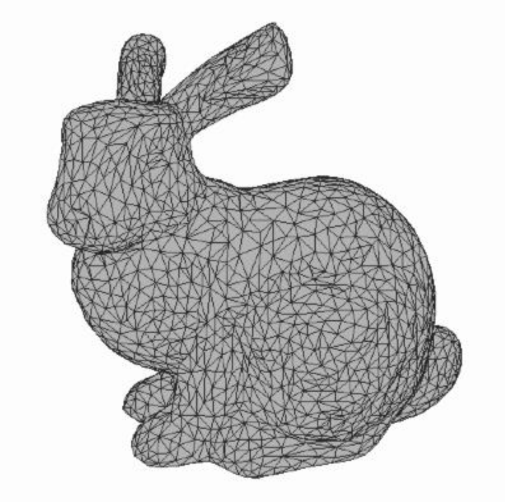
source: solidsolutions

Need for 3D Deep Learning!

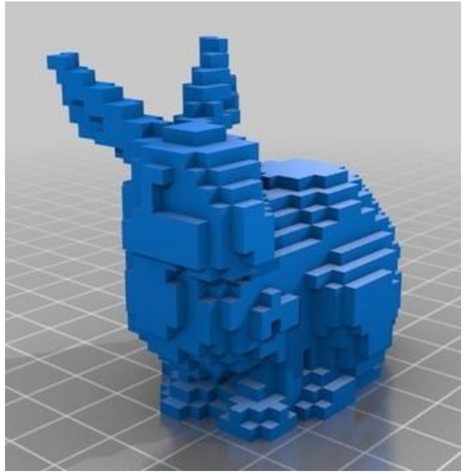
3D Representations



Point Cloud



Mesh



Volumetric

Projected View
RGB(D)

...

Previous Works

Prior to PointNet, there were many **handcrafted** Point Cloud Features

Feature Name	Supports Texture / Color	Local / Global / Regional	Best Use Case
PFH	No	L	
FPFH	No	L	2.5D Scans (Pseudo single position range images)
VFH	No	G	Object detection with basic pose estimation
CVFH	No	R	Object detection with basic pose estimation, detection of partial objects
RIFT	Yes	L	Real world 3D-Scans with no mirror effects. RIFT is vulnerable against flipping.

Source: https://github.com/PointCloudLibrary/pcl/wiki/Overview-and-Comparison-of-Features

Previous Works

Point cloud is **converted to other representations** before it's fed to a deep neural network

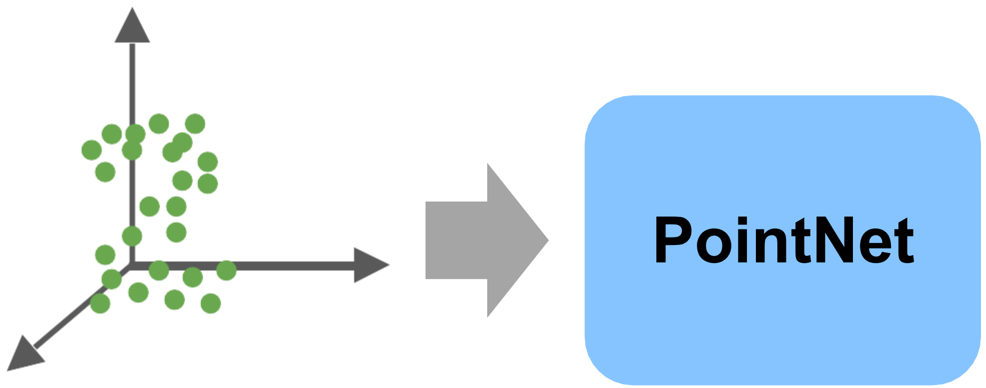
Conversion	Deep Net
Voxelization	3D CNN
Projection/Rendering	2D CNN
Feature extraction	Fully Connected

Research Question:

Can we achieve effective **feature learning directly** on point clouds?

Our Work: PointNet

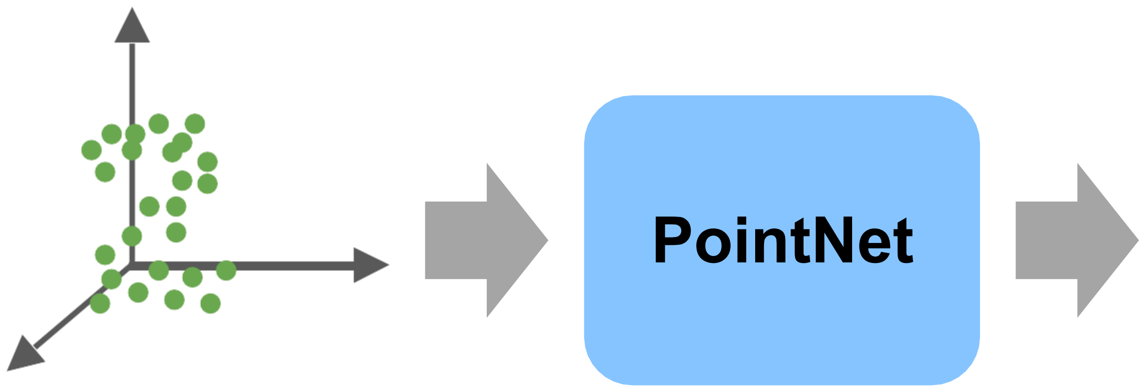
End-to-end learning for **scattered, unordered** point data



Our Work: PointNet

End-to-end learning for **scattered, unordered** point data

Unified framework for various tasks



Object Classification

Object Part Segmentation

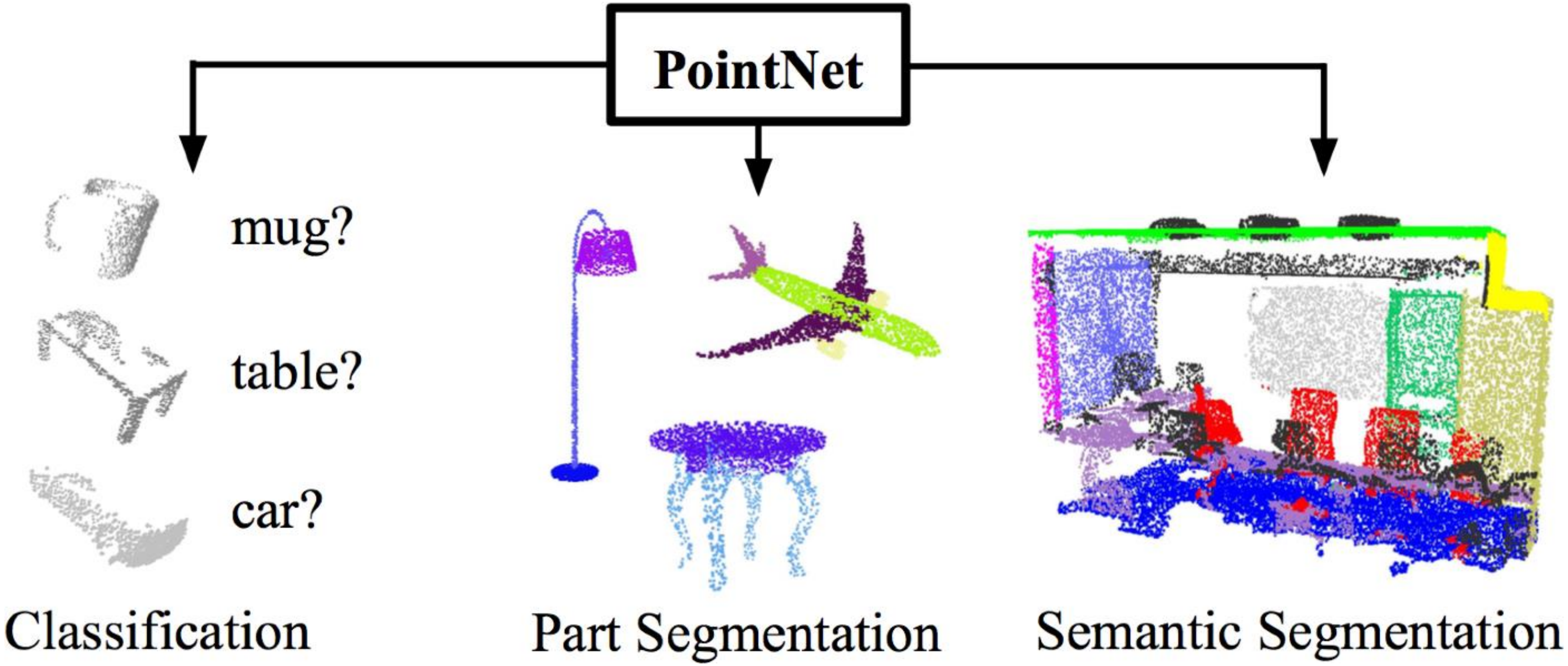
Semantic Scene Parsing

...

Our Work: PointNet

End-to-end learning for **scattered, unordered** point data

Unified framework for various tasks



Challenges

Unordered point set as input

Model needs to be invariant to N! permutations.

Invariance under geometric transformations

Point cloud rotations should not alter classification results.

Challenges

<u>Unordered</u> point set as input

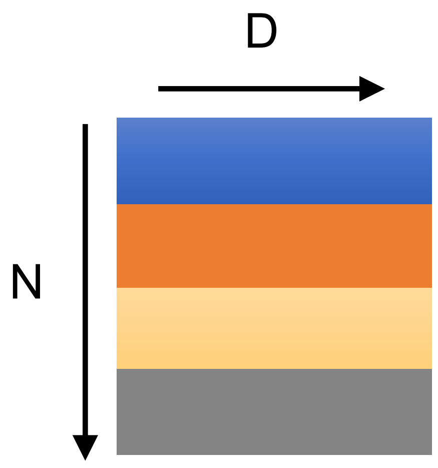
Model needs to be invariant to N! permutations.

Invariance under <u>geometric transformations</u>

Point cloud rotations should not alter classification results.

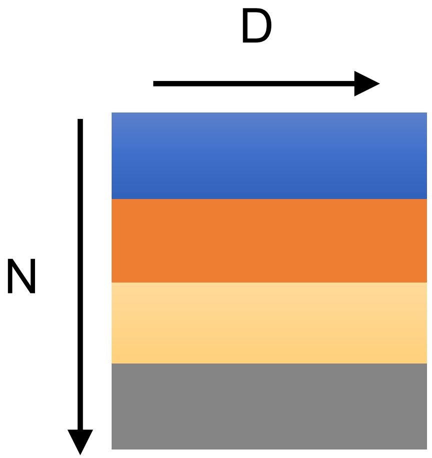
Unordered Input

Point cloud: N **orderless** points, each represented by a D dim vector

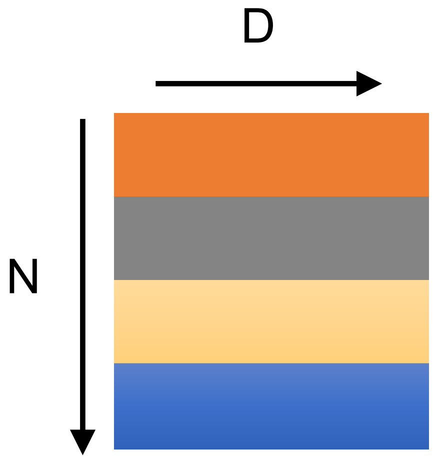


Unordered Input

Point cloud: N **<u>orderless</u>** points, each represented by a D dim vector

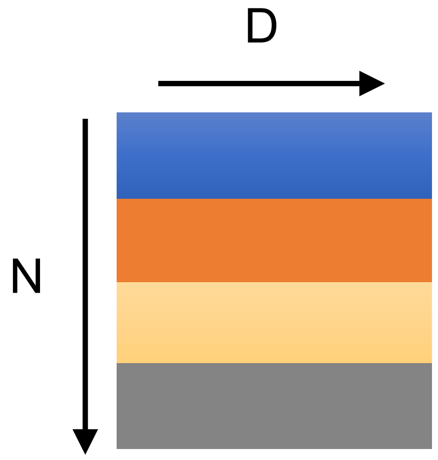


represents the same **set** as

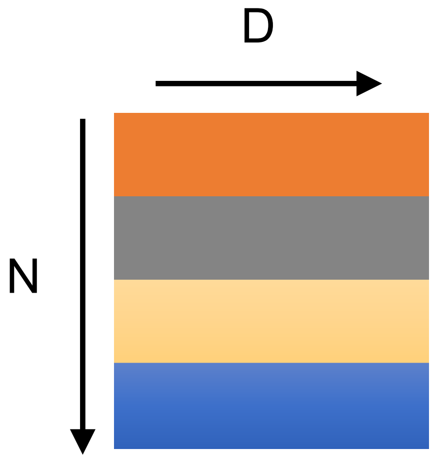


Unordered Input

Point cloud: N **orderless** points, each represented by a D dim vector



represents the same **set** as

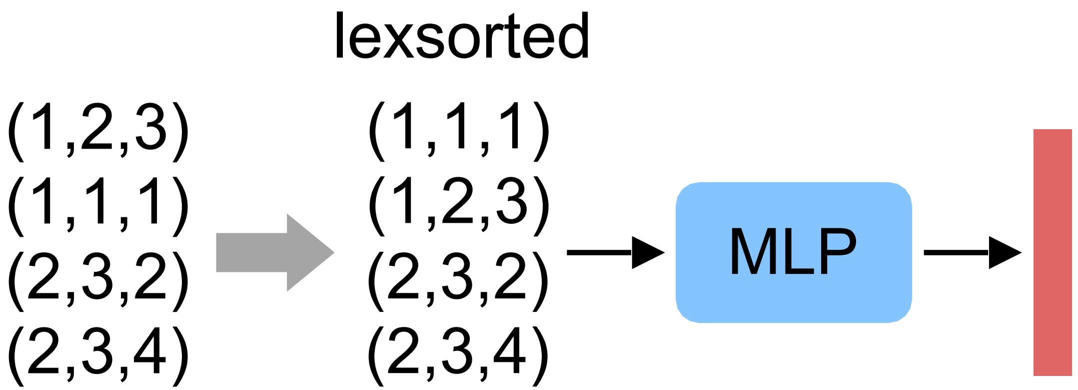


Model needs to be invariant to N! permutations

Permutation Invariance: How about Sorting?

"Sort" the points before feeding them into a network.

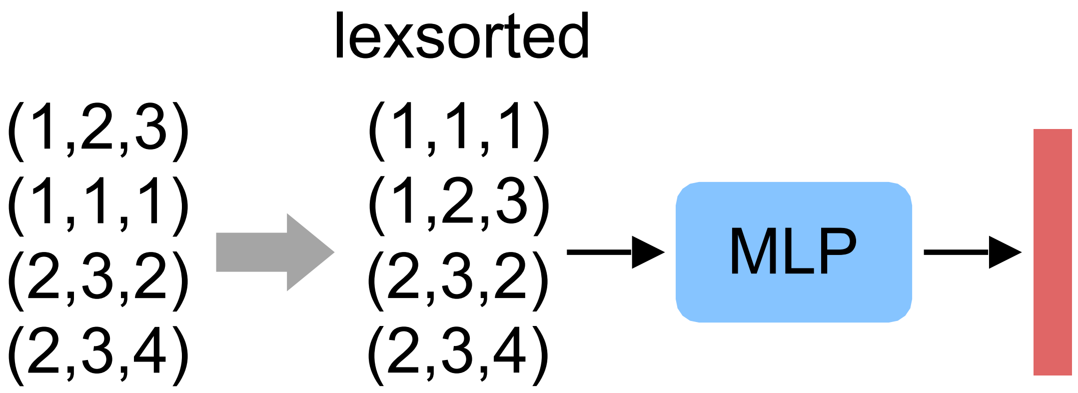
Unfortunately, there is no canonical order in high dim space.



Permutation Invariance: How about Sorting?

“Sort” the points before feeding them into a network.

Unfortunately, there is no canonical order in high dim space.



Multi-Layer Perceptron
(ModelNet shape classification)

	Accuracy
Unordered Input	12%
Lexsorted Input	40%
PointNet (vanilla)	**87%**

Point Transformer 3

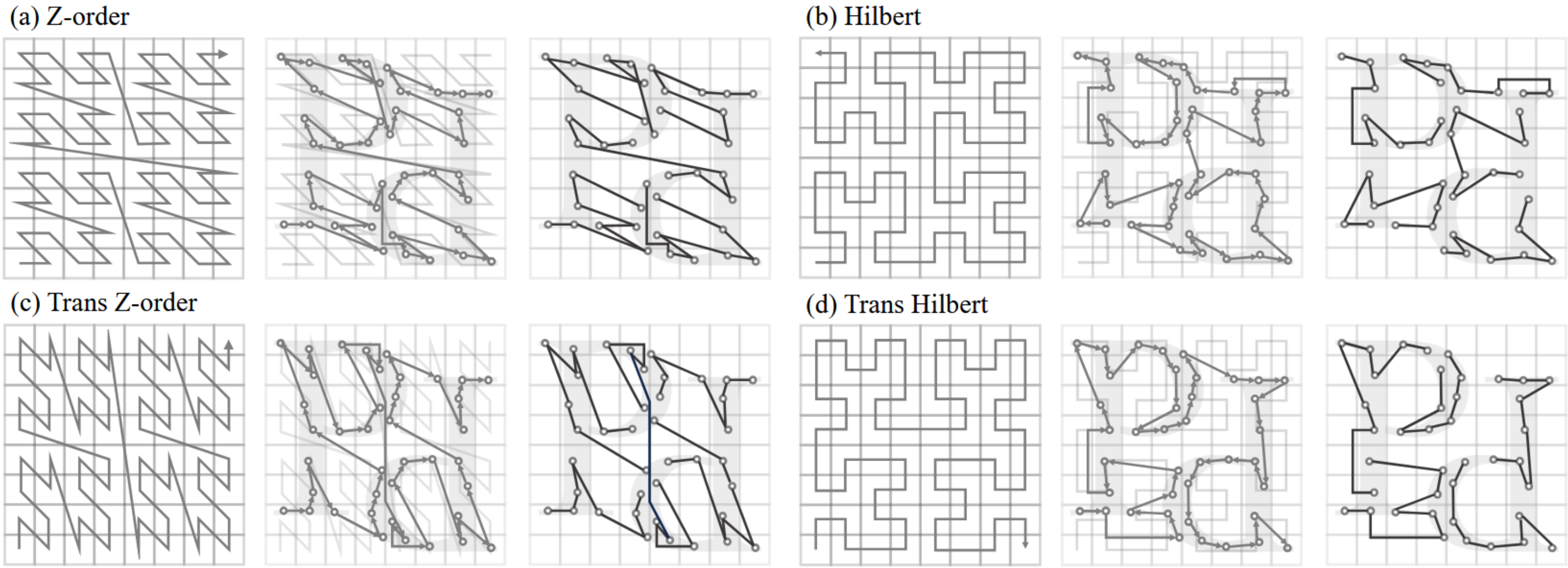
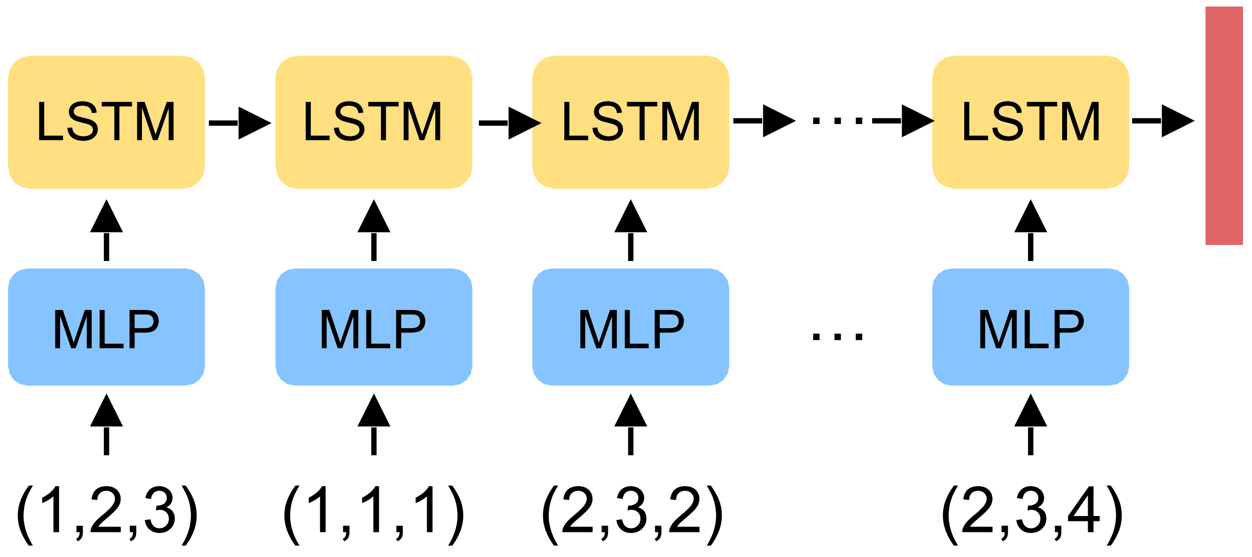


Figure 3. **Point cloud serialization.** We show the four patterns of serialization with a triplet visualization. For each triplet, we show the space-filling curve for serialization (left), point cloud serialization var sorting order within the space-filling curve (middle), and grouped patches of the serialized point cloud for local attention (right). Shifting across the four serialization patterns allows the attention mechanism to capture various spatial relationships and contexts, leading to an improvement in model accuracy and generalization capacity.

Permutation Invariance: How about RNNs?

Train RNN with permutation augmentation.

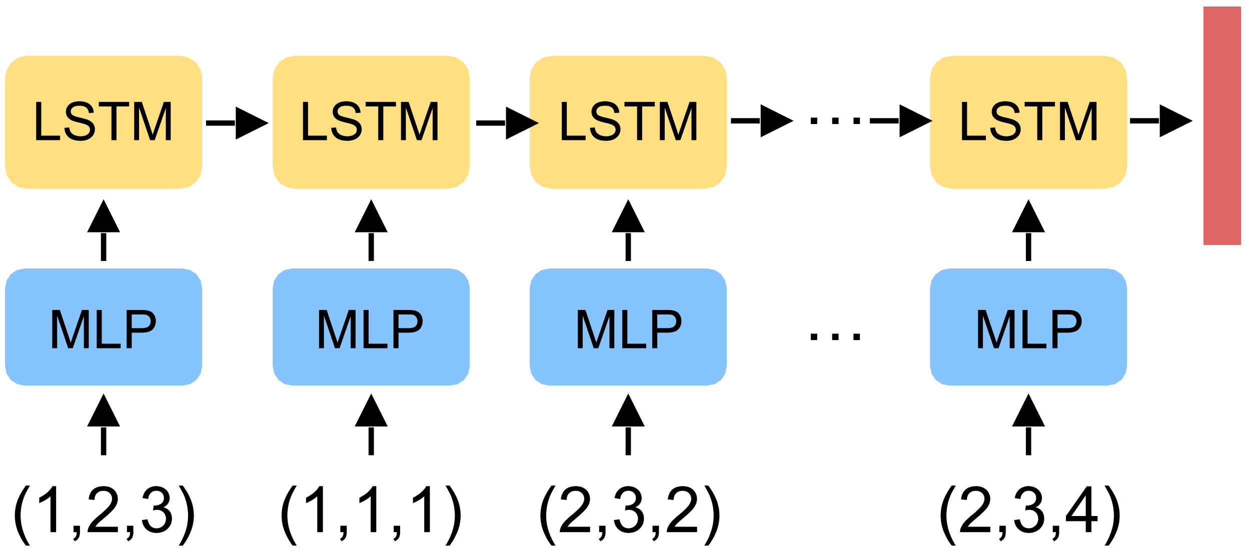
However, RNN forgets and order matters.



Permutation Invariance: How about RNNs?

Train RNN with permutation augmentation.

However, RNN forgets and order matters.



LSTM Network
(ModelNet shape classification)

	Accuracy
LSTM	75%
PointNet (vanilla)	**87%**

Permutation Invariance: Symmetric Function

$$f(x_1, x_2, \ldots, x_n) \equiv f(x_{\pi_1}, x_{\pi_2}, \ldots, x_{\pi_n}), \quad x_i \in \mathbb{R}^D$$

Permutation Invariance: Symmetric Function

$$f(x_1, x_2, \ldots, x_n) \equiv f(x_{\pi_1}, x_{\pi_2}, \ldots, x_{\pi_n}), \quad x_i \in \mathbb{R}^D$$

Examples:

$$f(x_1, x_2, \ldots, x_n) = \max\{x_1, x_2, \ldots, x_n\}$$

$$f(x_1, x_2, \ldots, x_n) = x_1 + x_2 + \ldots + x_n$$

…

Permutation Invariance: Symmetric Function

$$f(x_1, x_2, \ldots, x_n) \equiv f(x_{\pi_1}, x_{\pi_2}, \ldots, x_{\pi_n}), \quad x_i \in \mathbb{R}^D$$

Examples:

$$f(x_1, x_2, \ldots, x_n) = \max\{x_1, x_2, \ldots, x_n\}$$

$$f(x_1, x_2, \ldots, x_n) = x_1 + x_2 + \ldots + x_n$$

…

How can we construct a family of symmetric functions by neural networks?

Permutation Invariance: Symmetric Function

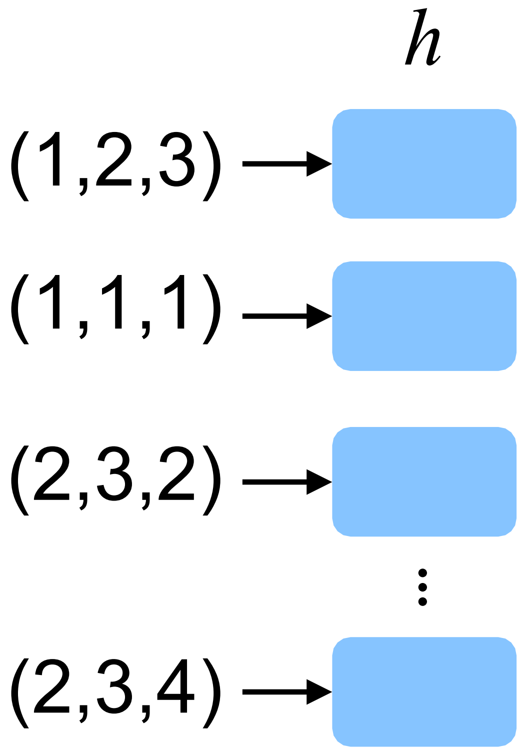
Observe:

$f(x_1, x_2, \ldots, x_n) = \gamma \circ g(h(x_1), \ldots, h(x_n))$ is symmetric if g is symmetric

Permutation Invariance: Symmetric Function

Observe:

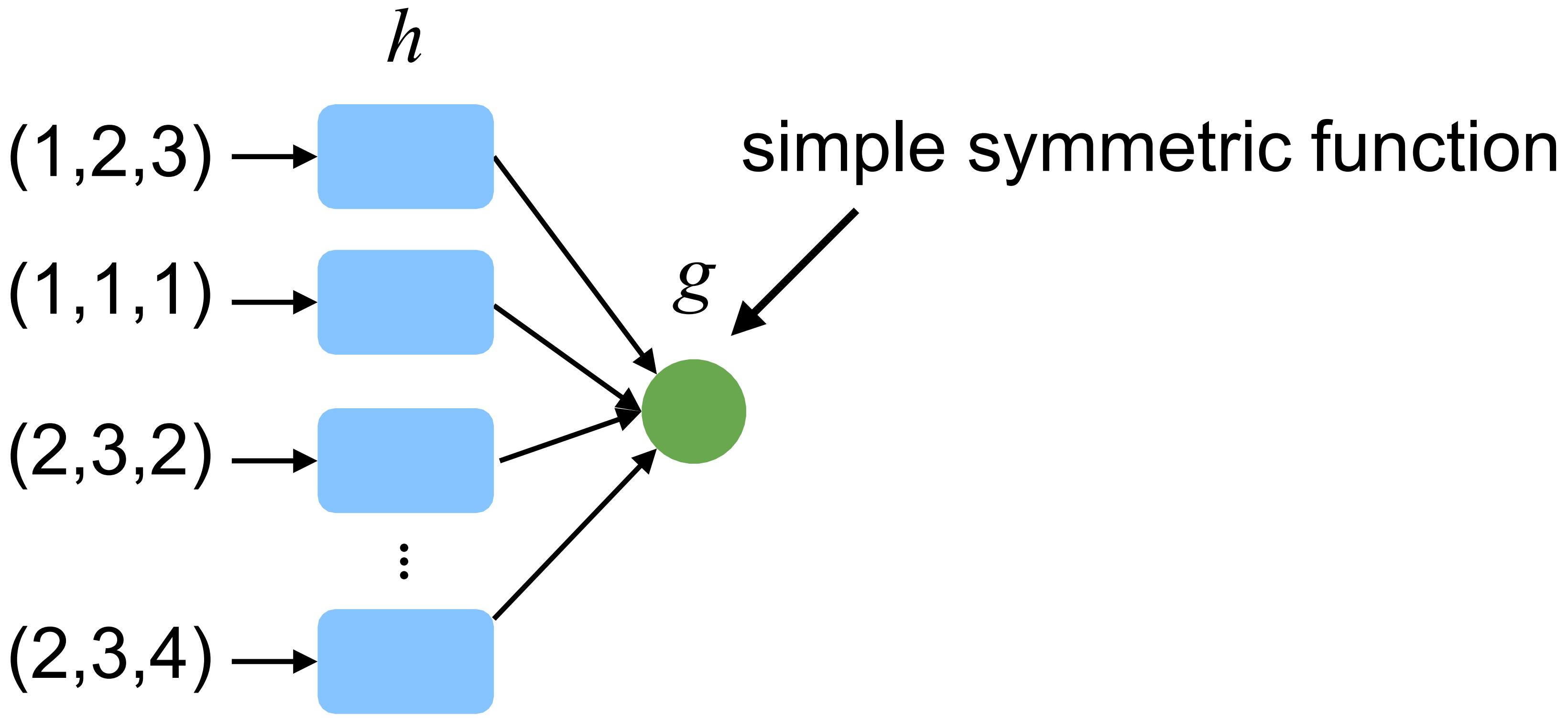
$f(x_1, x_2, \ldots, x_n) = \gamma^o g(h(x_1), \ldots, h(x_n))$ is symmetric if g is symmetric



Permutation Invariance: Symmetric Function

Observe:

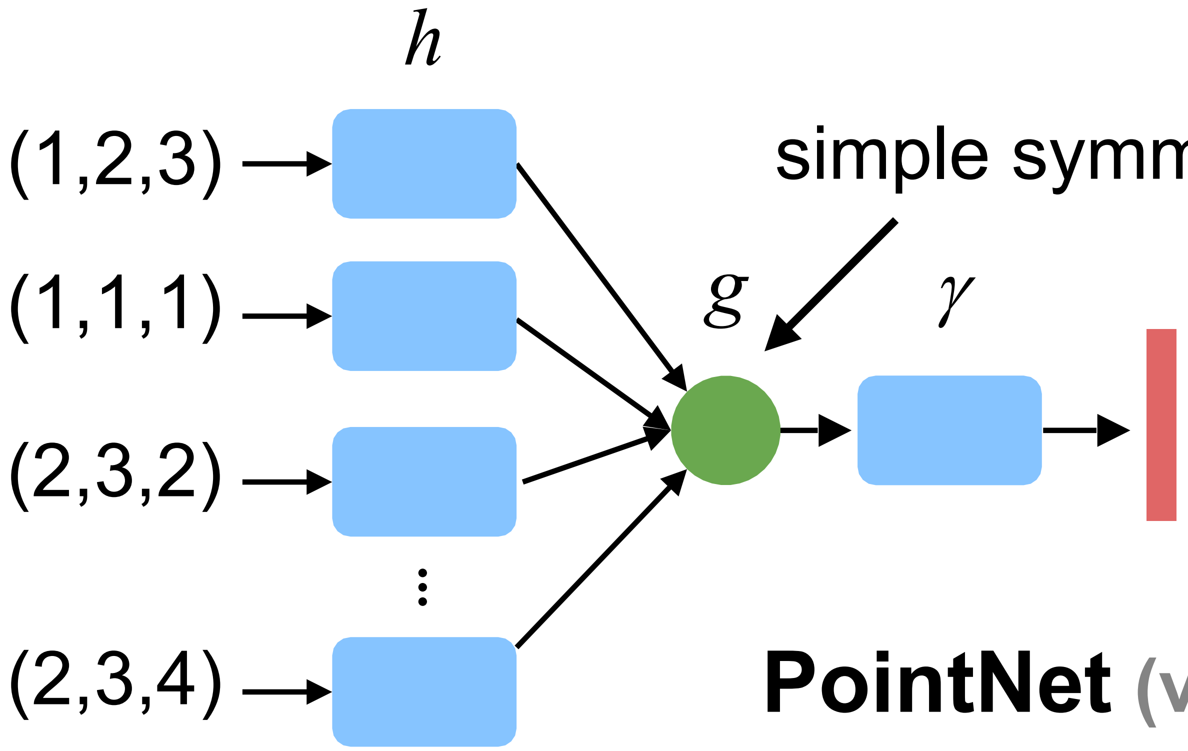
$f(x_1, x_2, \ldots, x_n) = \gamma \circ g(h(x_1), \ldots, h(x_n))$ is symmetric if g is symmetric



Permutation Invariance: Symmetric Function

Observe:

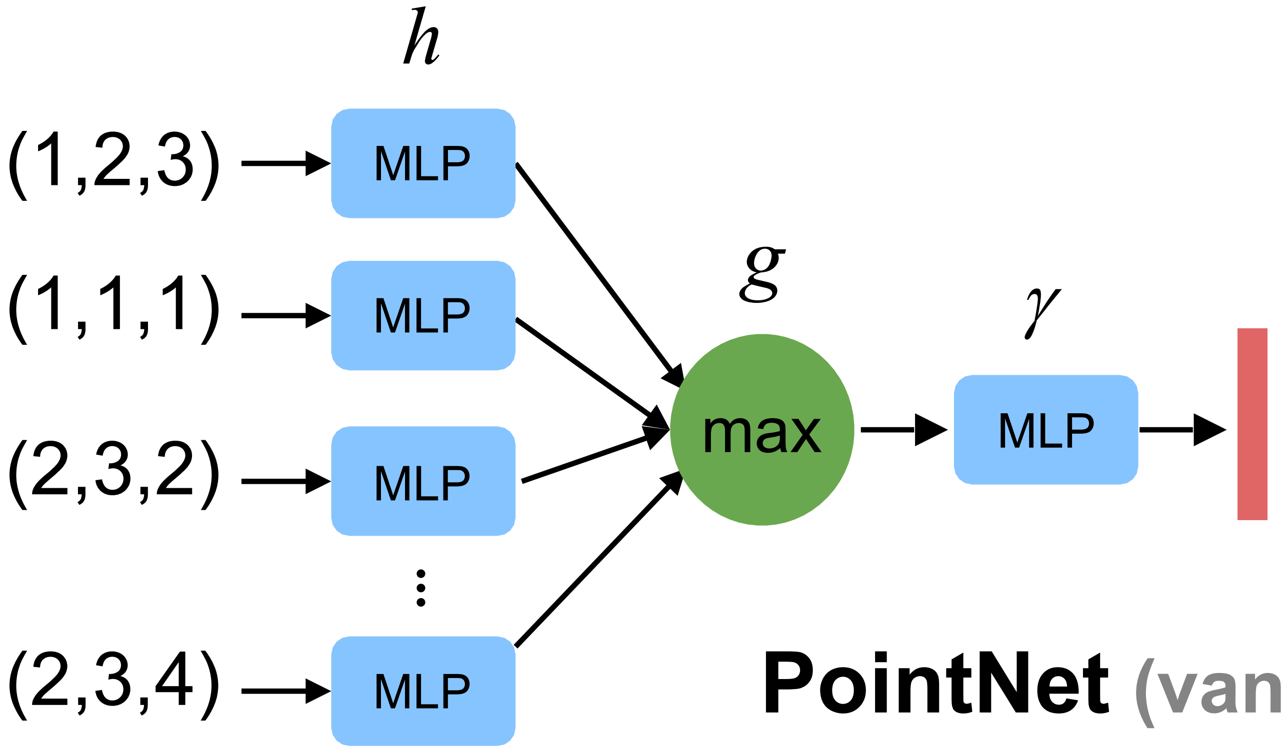
$f(x_1, x_2, \ldots, x_n) = \gamma \circ g(h(x_1), \ldots, h(x_n))$ is symmetric if g is symmetric



PointNet **(vanilla)**

Basic PointNet Architecture

Empirically, we use **multi-layer perceptron (MLP)** and **max pooling**:



PointNet (vanilla)

Challenges

Unordered point set as input

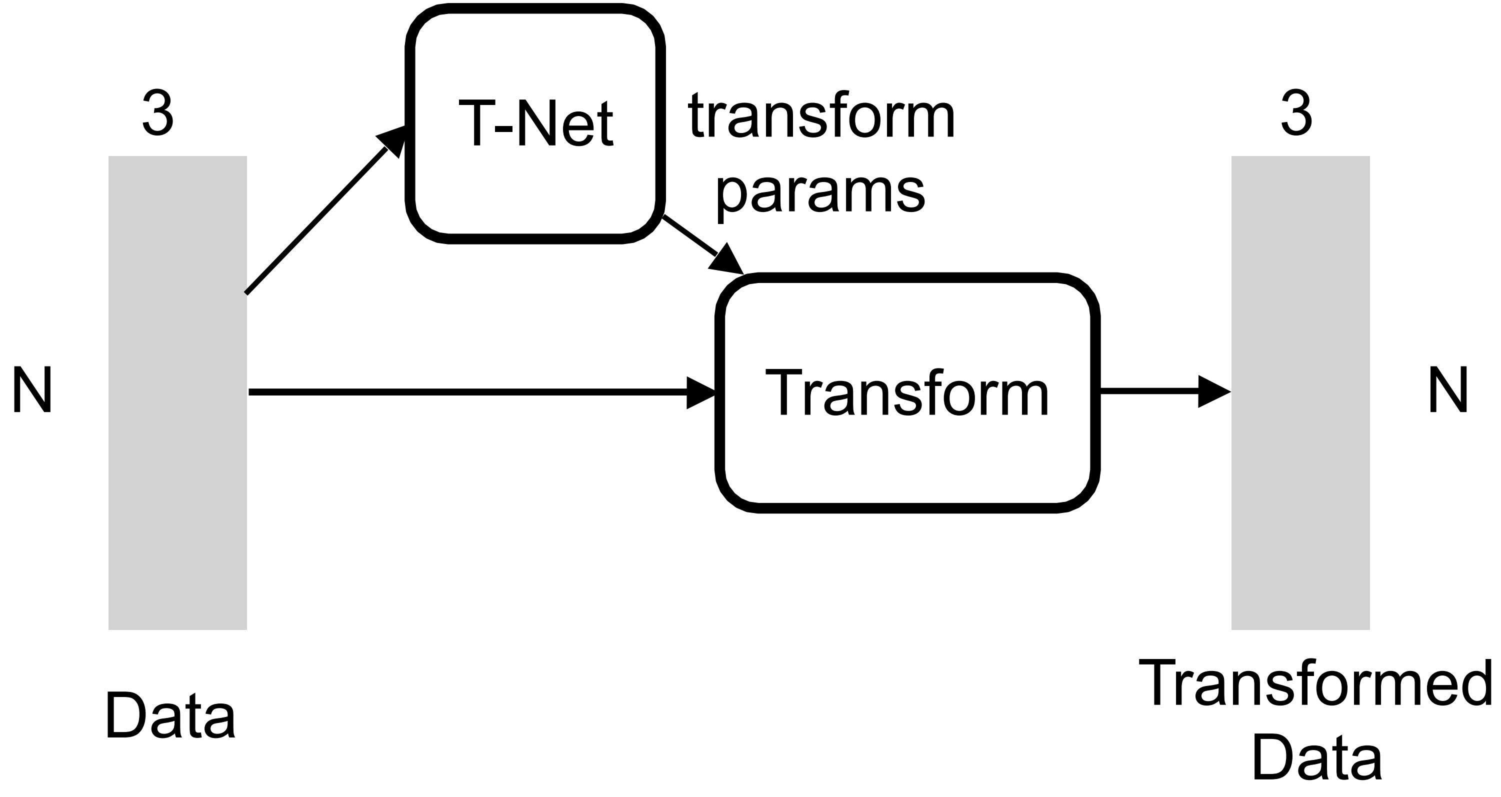
Model needs to be invariant to N! permutations.

Invariance under geometric transformations

Point cloud rotations should not alter classification results.

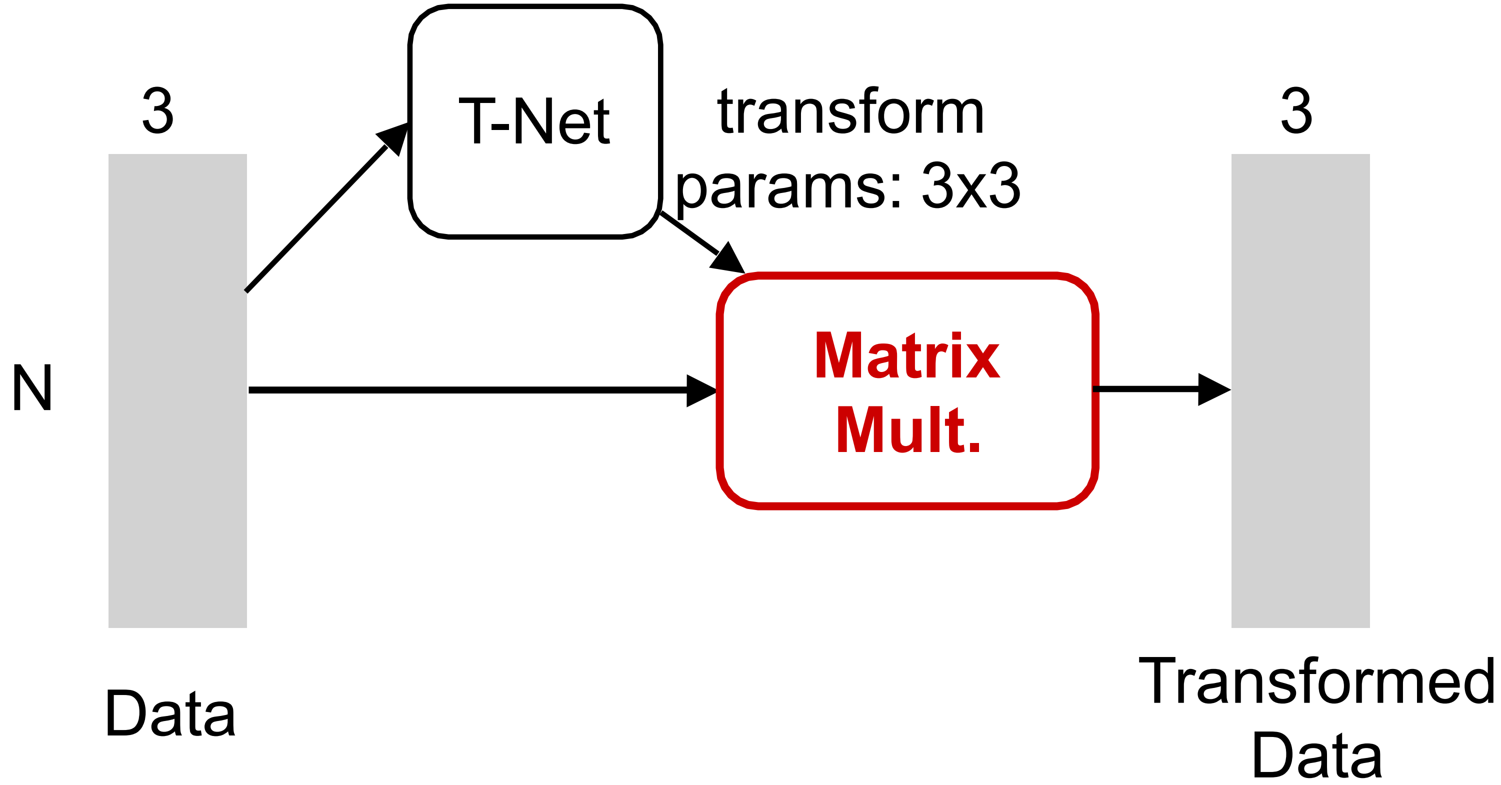
Input Alignment by Transformer Network

Idea: Data dependent transformation for automatic alignment

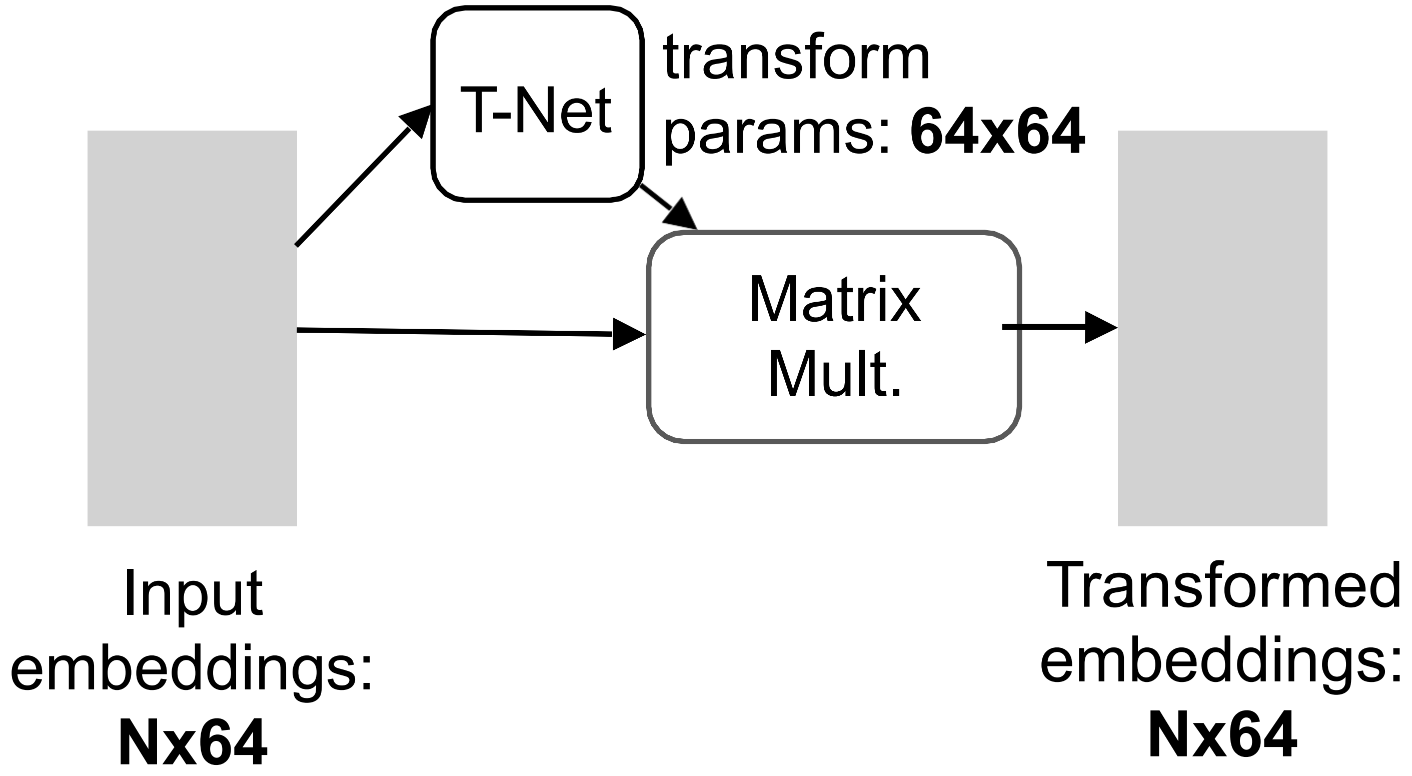


Input Alignment by Transformer Network

The transformation is just matrix multiplication!

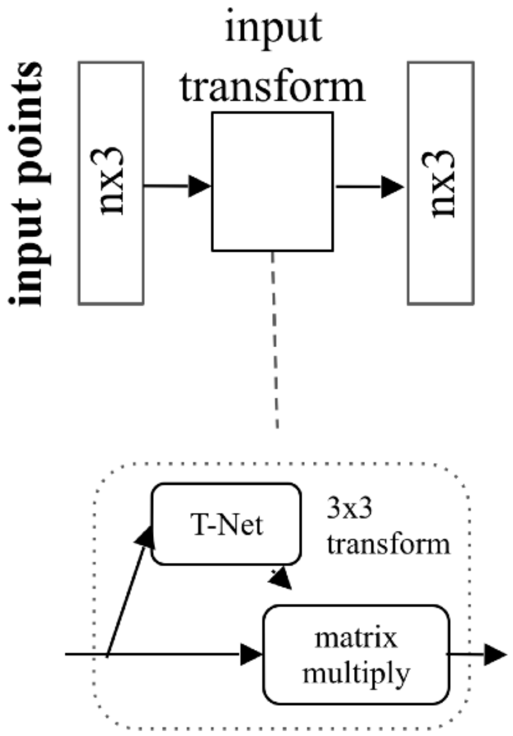


Embedding Space Alignment

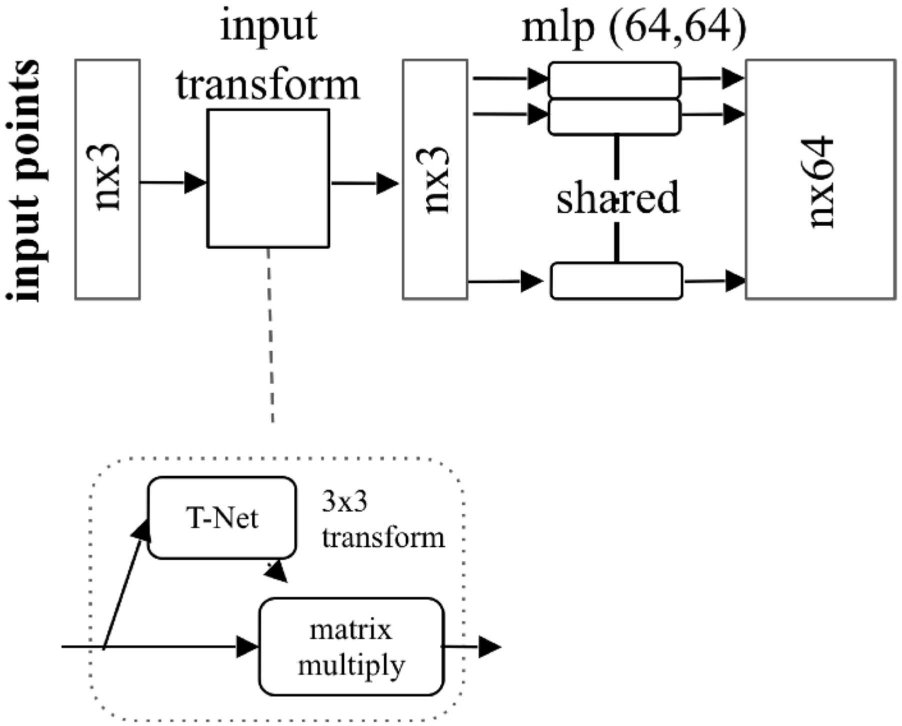


PointNet Classification Network

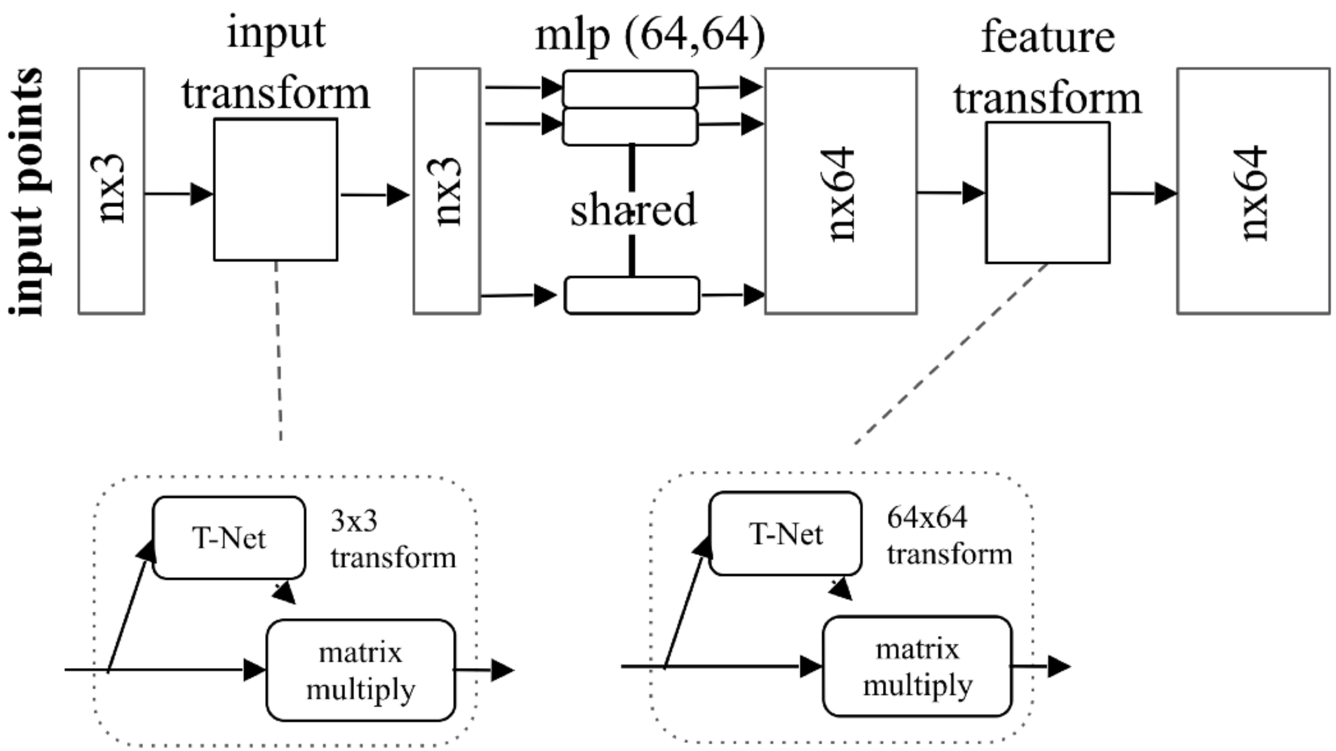
PointNet Classification Network



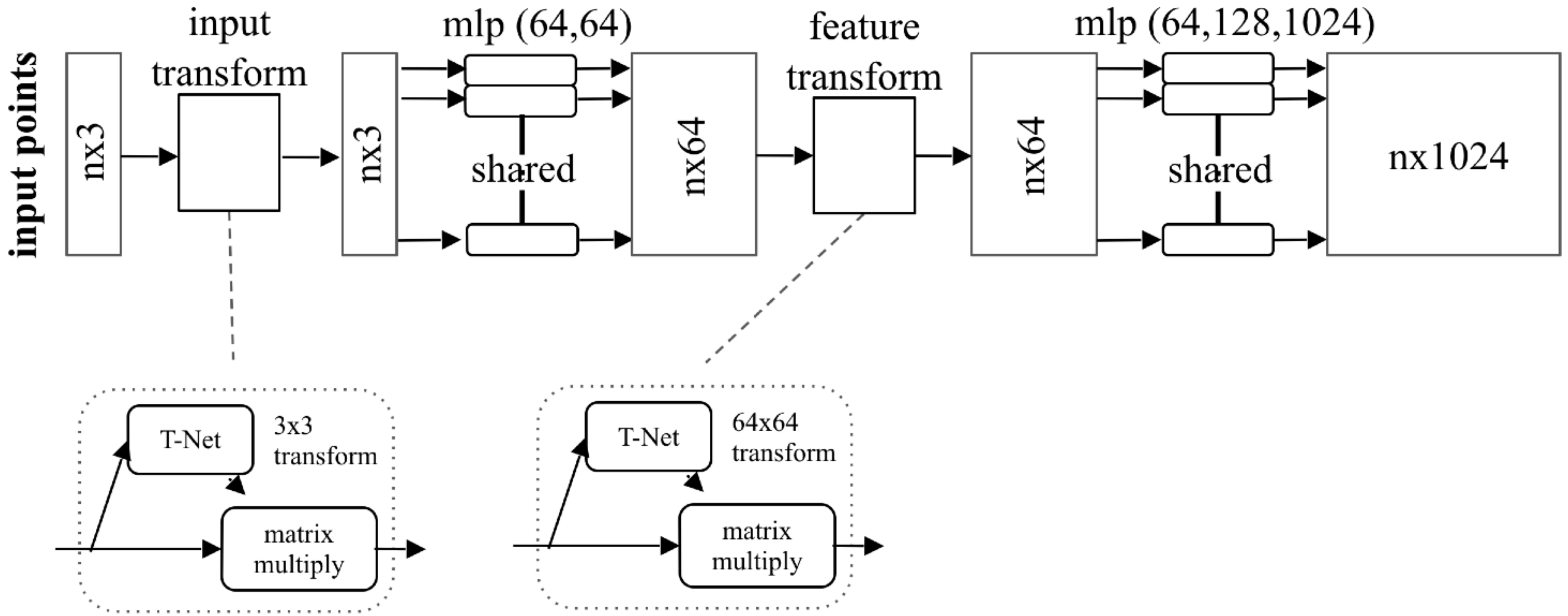
PointNet Classification Network



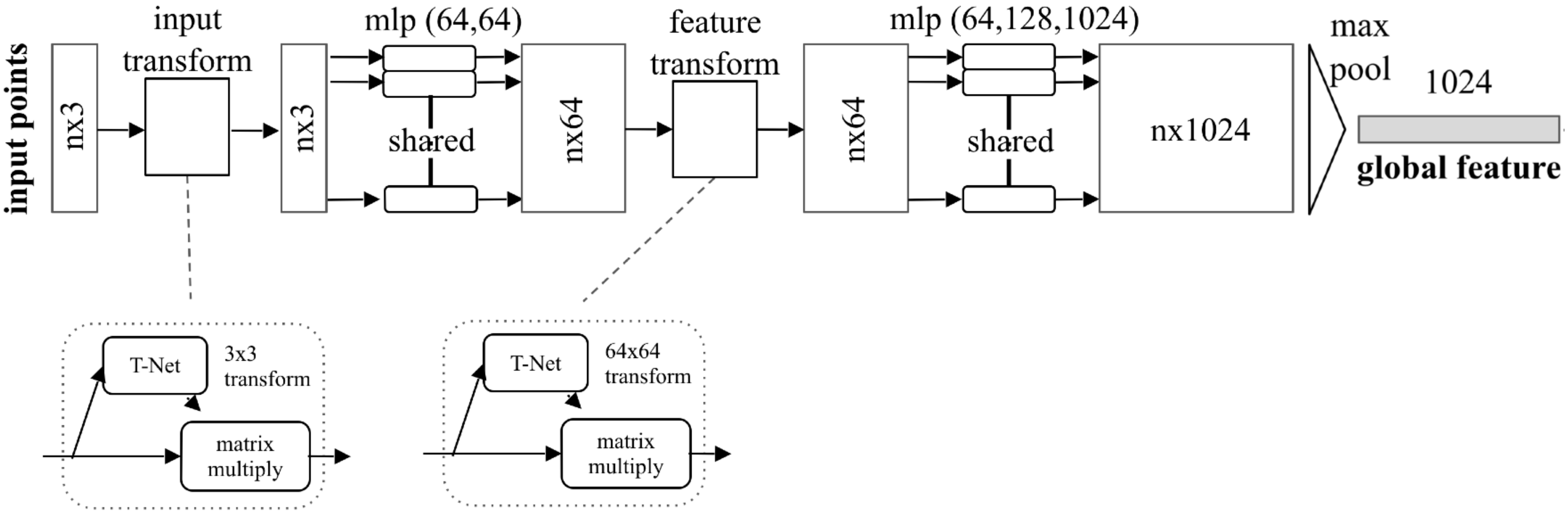
PointNet Classification Network



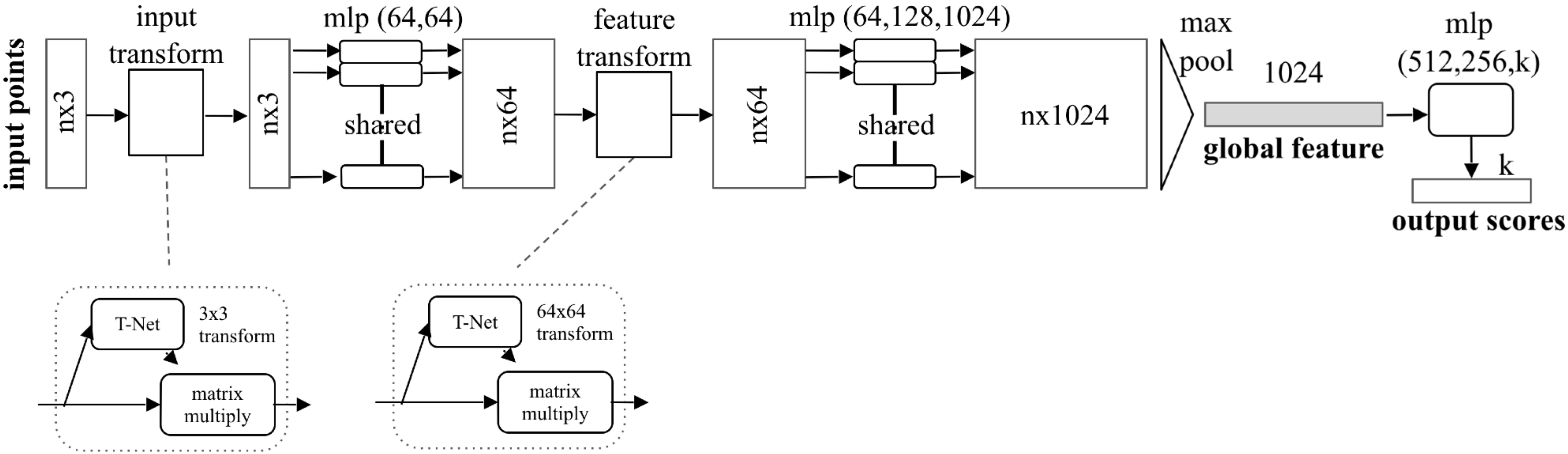
PointNet Classification Network



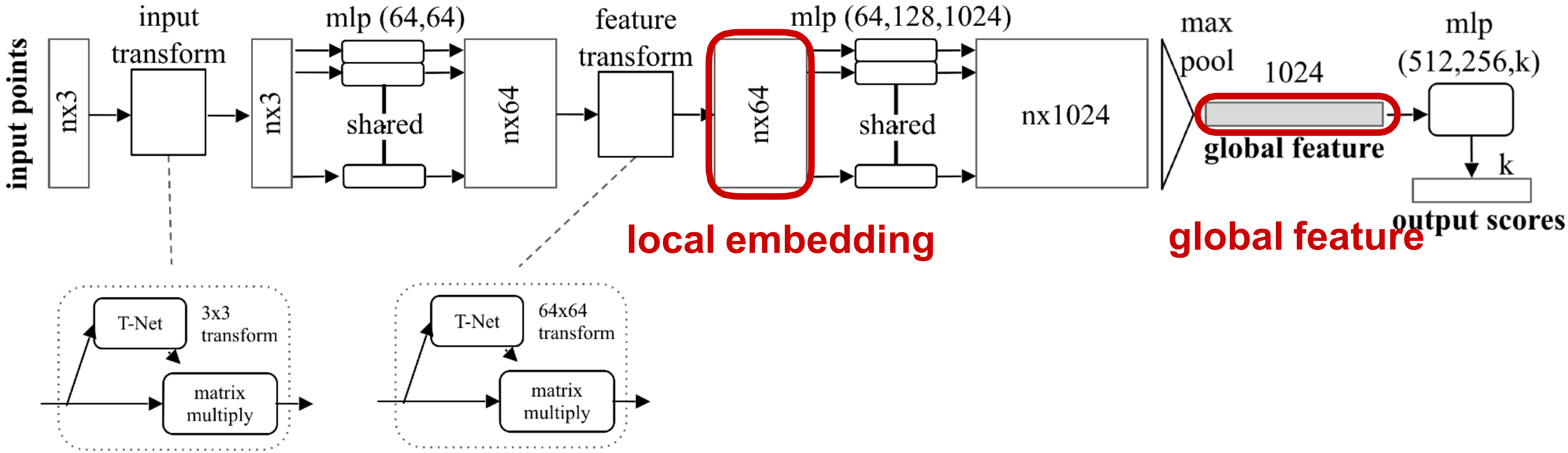
PointNet Classification Network



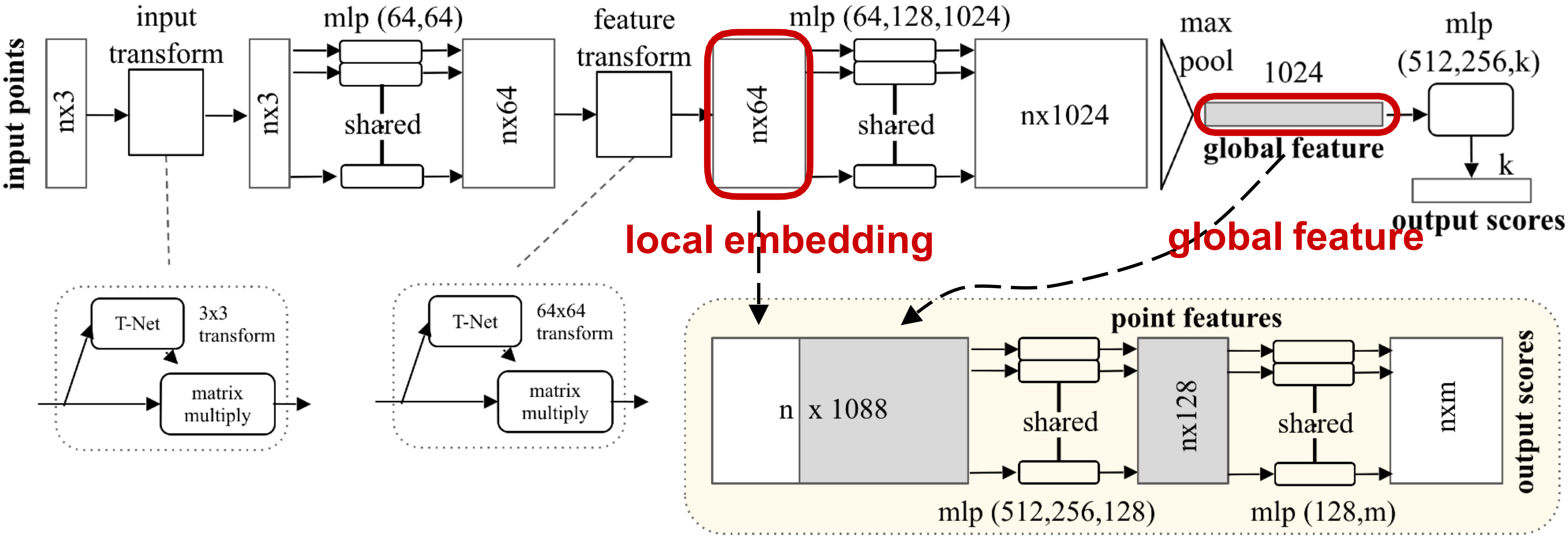
PointNet Classification Network



Extension to PointNet Segmentation Network



Extension to PointNet Segmentation Network



Results

Results on Object Classification

	input	#views	accuracy avg. class	accuracy overall
SPH [12]	mesh	-	68.2	-
3DShapeNets [29]	volume	1	77.3	84.7
VoxNet [18]	volume	12	83.0	85.9
Subvolume [19]	volume	20	86.0	**89.2**
LFD [29]	image	10	75.5	-
MVCNN [24]	image	80	**90.1**	-
Ours baseline	point	-	72.6	77.4
Ours PointNet	point	1	86.2	**89.2**

dataset: ModelNet40; metric: 40-class classification accuracy (%)

Results on Object Classification

	input	#views	accuracy avg. class	accuracy overall
SPH [12]	mesh	-	68.2	-
3DShapeNets [29]	volume	1	77.3	84.7
VoxNet [18]	volume	12	83.0	85.9
Subvolume [19]	volume	20	86.0	**89.2**
LFD [29]	image	10	75.5	-
MVCNN [24]	image	80	**90.1**	-
Ours baseline	point	-	72.6	77.4
Ours PointNet	point	1	86.2	**89.2**

3D CNNs (3DShapeNets, VoxNet, Subvolume)

dataset: ModelNet40; metric: 40-class classification accuracy (%)

Results on Object Cl

	inp	acy / ll
SPH [12]	me	
3DShapeNets [29]	volu	7
VoxNet [18]	volu	9
Subvolume [19]	volu	2
LFD [29]	ima	
MVCNN [24]	ima	
Ours baseline	po	4
Ours PointNet	po	2

3D CNNs

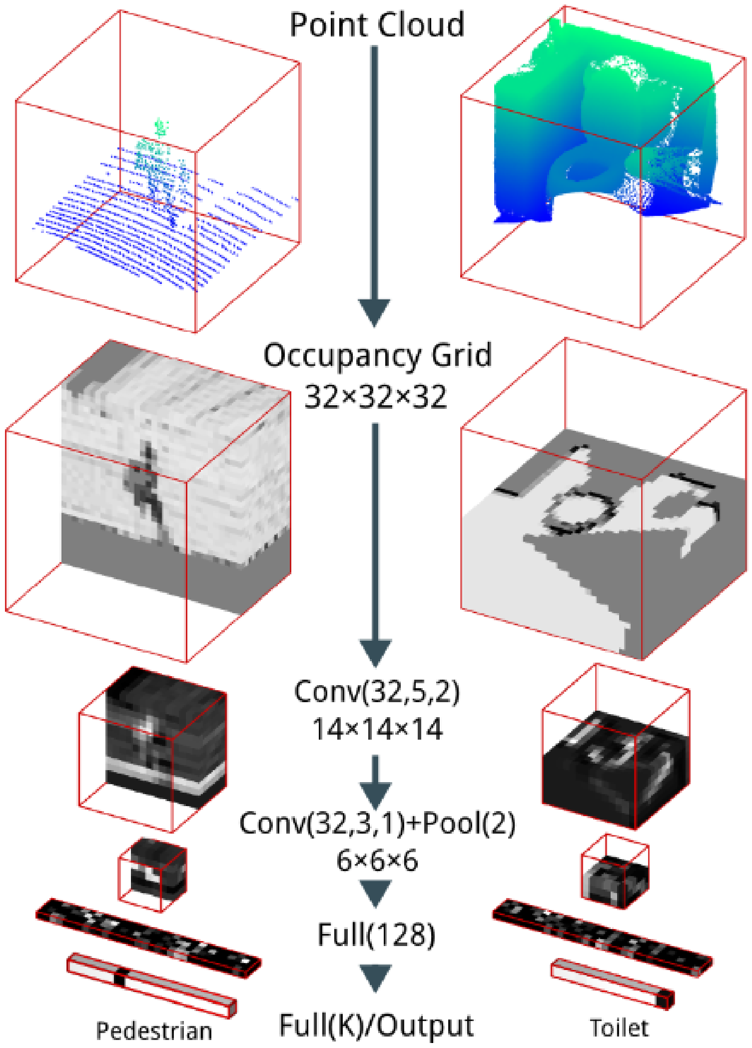


Fig. 1. The VoxNet Architecture. $Conv(f, d, s)$ indicates f filters of size d and at stride s, $Pool(m)$ indicates pooling with area m, and $Full(n)$ indicates fully connected layer with n outputs. We show inputs, example feature maps, and predicted outputs for two instances from our experiments. The point cloud on the left is from LiDAR and is part of the Sydney Urban Objects dataset [4]. The point cloud on the right is from RGBD and is part of NYUv2 [5]. We use cross sections for visualization purposes.

dataset: ModelNet40; metri)

Results on Object Classification

	input	#views	accuracy avg. class	accuracy overall
SPH [12]	mesh	-	68.2	-
3DShapeNets [29]	volume	1	77.3	84.7
VoxNet [18]	volume	12	83.0	85.9
Subvolume [19]	volume	20	86.0	**89.2**
LFD [29]	image	10	75.5	-
MVCNN [24]	image	80	**90.1**	-
Ours baseline	point	-	72.6	77.4
Ours PointNet	point	1	86.2	**89.2**

2D projections of 3D data

dataset: ModelNet40; metric: 40-class classification accuracy (%)

Results on Object Classification

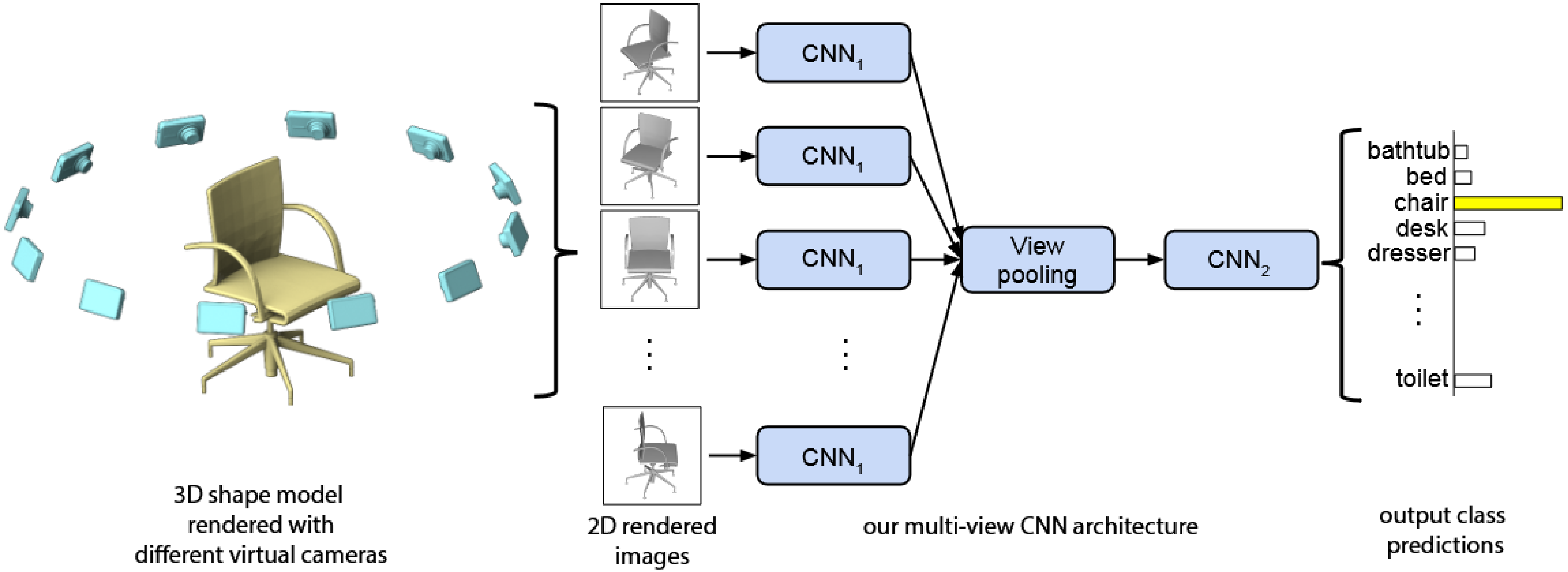


Figure 1. Multi-view CNN for 3D shape recognition (illustrated using the 1st camera setup). At test time a 3D shape is rendered from 12 different views and are passed thorough CNN_1 to extract view based features. These are then pooled across views and passed through CNN_2 to obtain a compact shape descriptor.

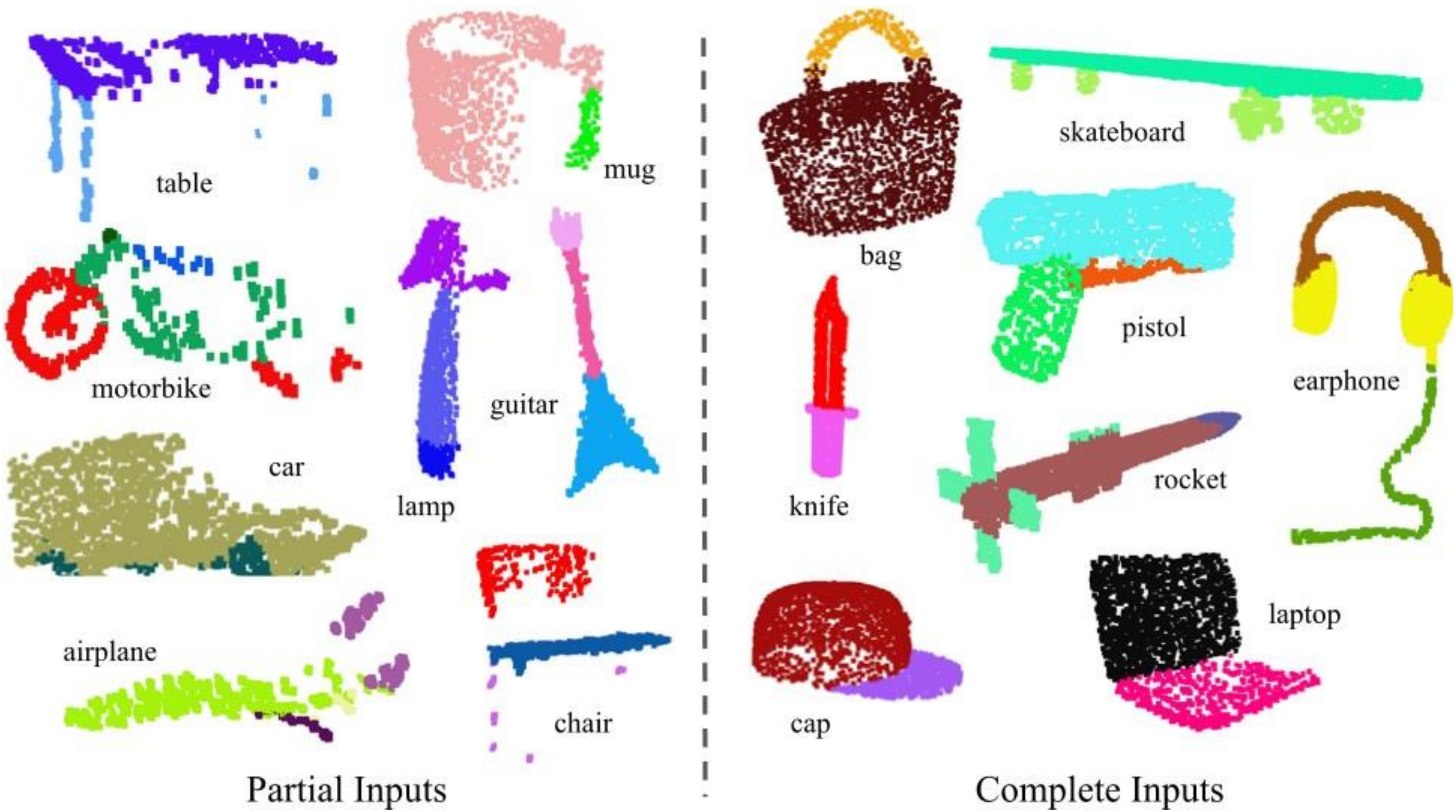
dataset: ModelNet40; metric: 40-class classification accuracy (%)

Results on Object Classification

	input	#views	accuracy avg. class	accuracy overall
SPH [12]	mesh	-	68.2	-
3DShapeNets [29]	volume	1	77.3	84.7
VoxNet [18]	volume	12	83.0	85.9
Subvolume [19]	volume	20	86.0	**89.2**
LFD [29]	image	10	75.5	-
MVCNN [24]	image	80	**90.1**	-
Ours baseline	point	-	72.6	77.4
Ours PointNet	point	1	86.2	**89.2**

dataset: ModelNet40; metric: 40-class classification accuracy (%)

Results on Object Part Segmentation



Results on Object Part Segmentation

	mean	aero	bag	cap	car	chair	ear phone	guitar	knife	lamp	laptop	motor	mug	pistol	rocket	skate board	table
# shapes		2690	76	55	898	3758	69	787	392	1547	451	202	184	283	66	152	5271
Wu [28]	-	63.2	-	-	-	73.5	-	-	-	74.4	-	-	-	-	-	-	74.8
Yi [30]	81.4	81.0	78.4	77.7	**75.7**	87.6	61.9	**92.0**	85.4	**82.5**	**95.7**	**70.6**	91.9	**85.9**	53.1	69.8	75.3
3DCNN	79.4	75.1	72.8	73.3	70.0	87.2	63.5	88.4	79.6	74.4	93.9	58.7	91.8	76.4	51.2	65.3	77.1
Ours	**83.7**	**83.4**	**78.7**	**82.5**	74.9	**89.6**	**73.0**	91.5	**85.9**	80.8	95.3	65.2	**93.0**	81.2	**57.9**	**72.8**	**80.6**

dataset: ShapeNetPart; metric: mean IoU (%)

Results on Semantic Scene Parsing

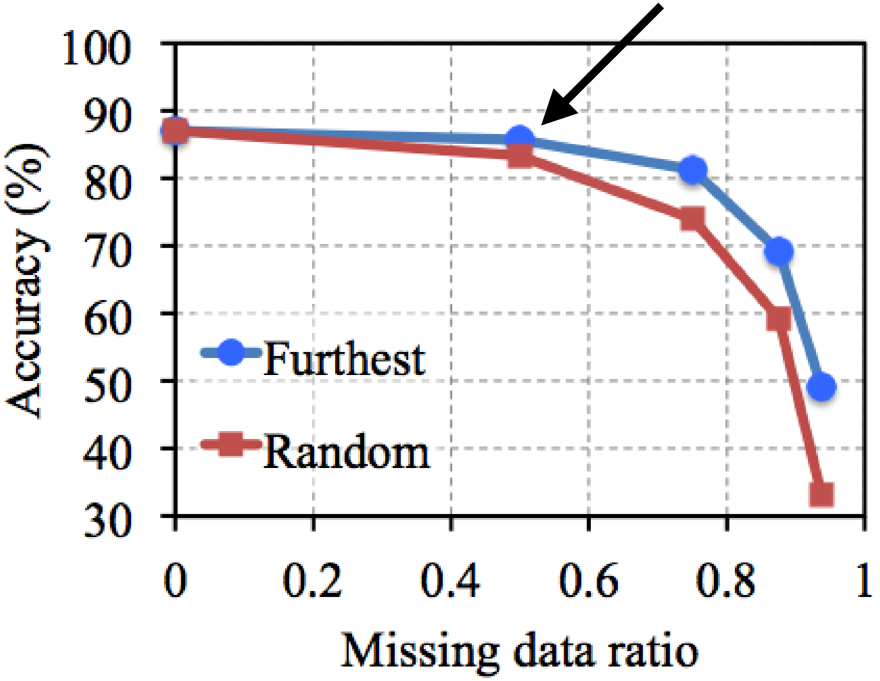
dataset: Stanford 2D-3D-S (Matterport scans)

Robustness to Data Corruption

dataset: ModelNet40; metric: 40-class classification accuracy (%)

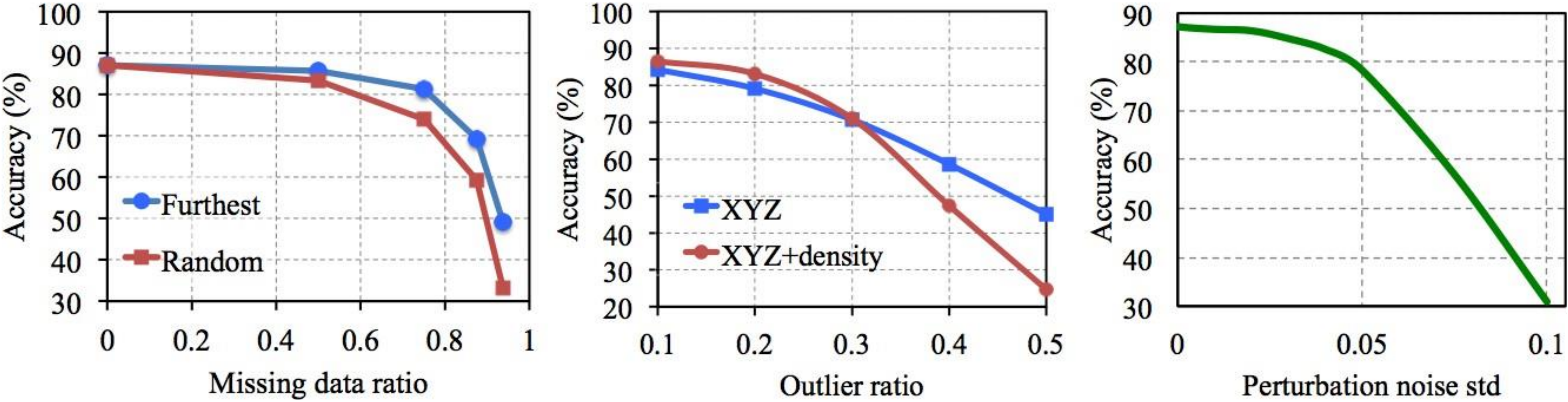
Robustness to Data Corruption

Less than 2% accuracy drop with 50% missing data



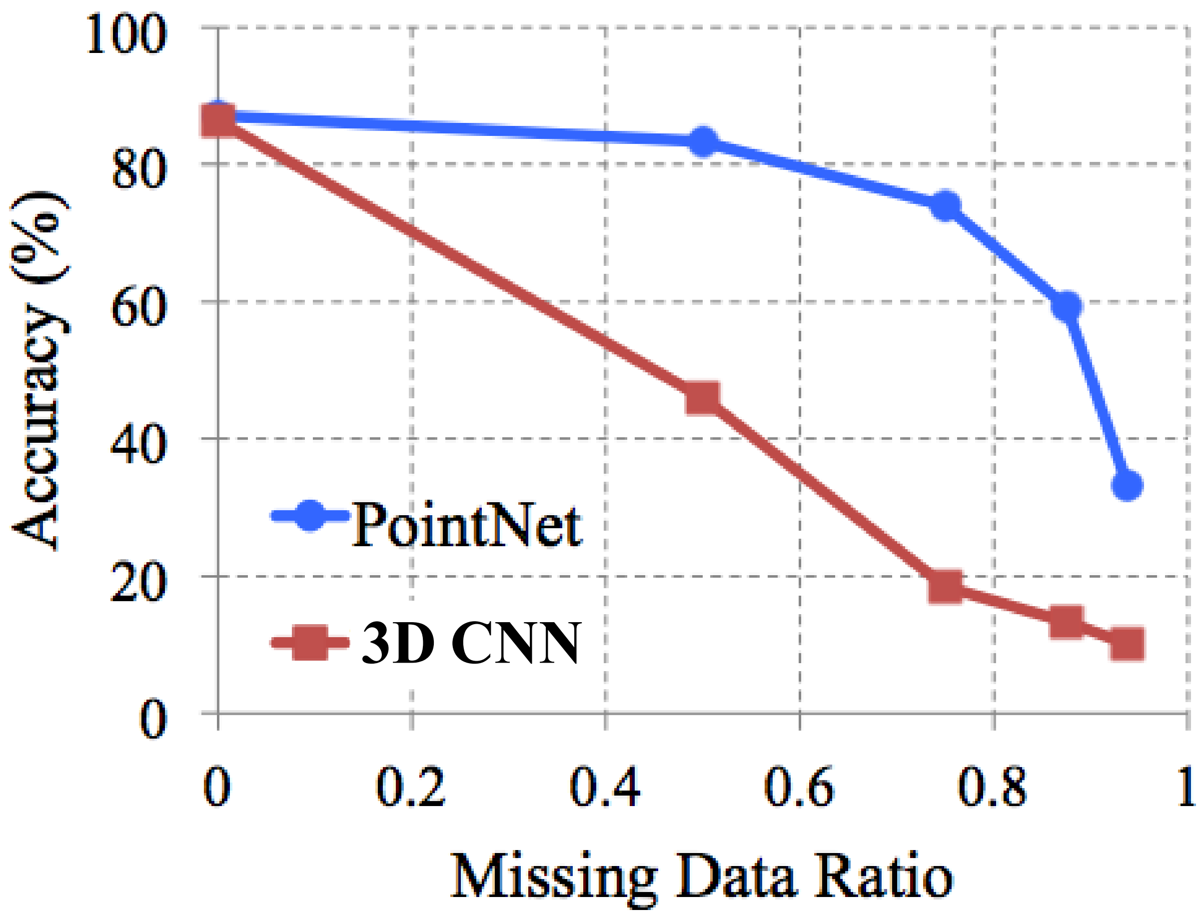
dataset: ModelNet40; metric: 40-class classification accuracy (%)

Robustness to Data Corruption



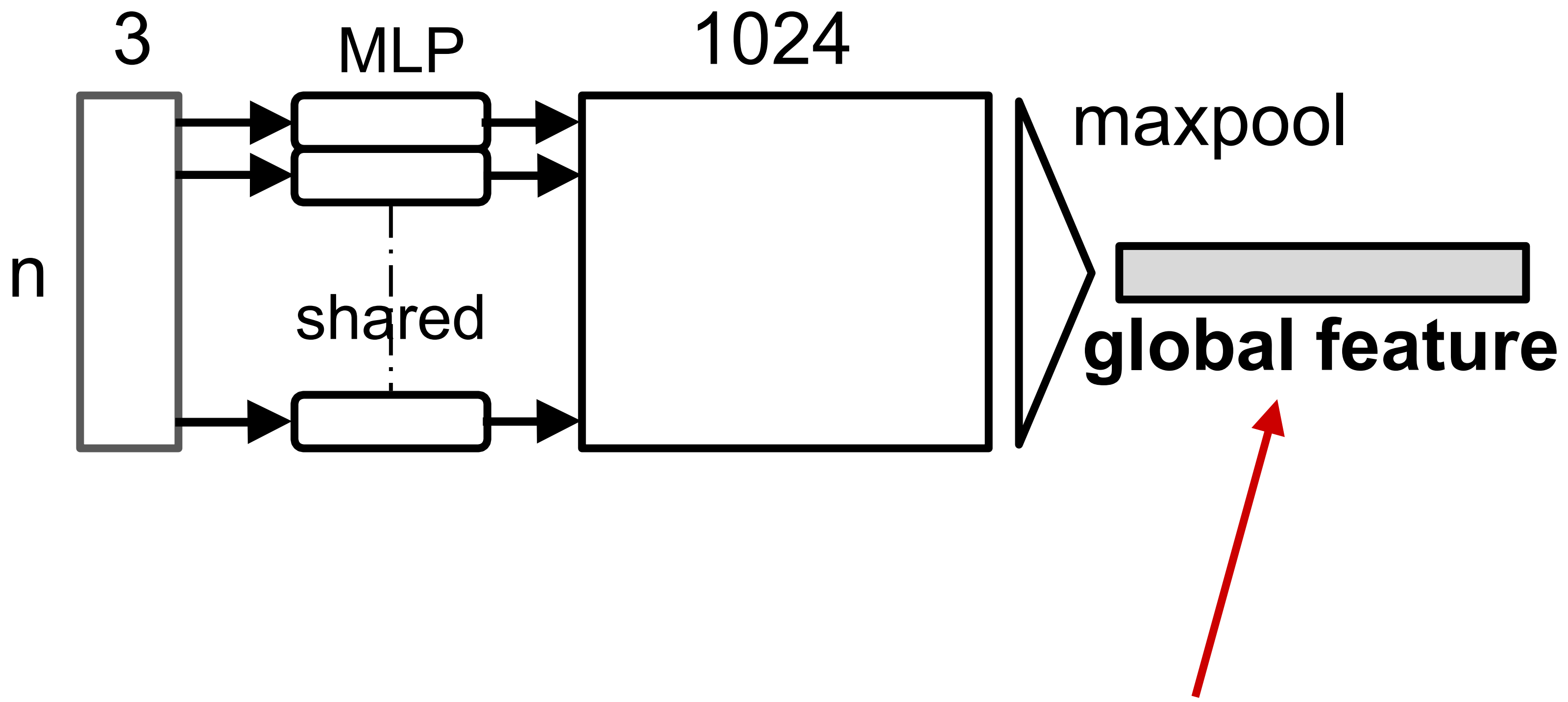
dataset: ModelNet40; metric: 40-class classification accuracy (%)

Robustness to Data Corruption



Why is PointNet so robust to missing data?

Visualizing Global Point Cloud Features



<u>Which input points</u> are contributing to the global feature?

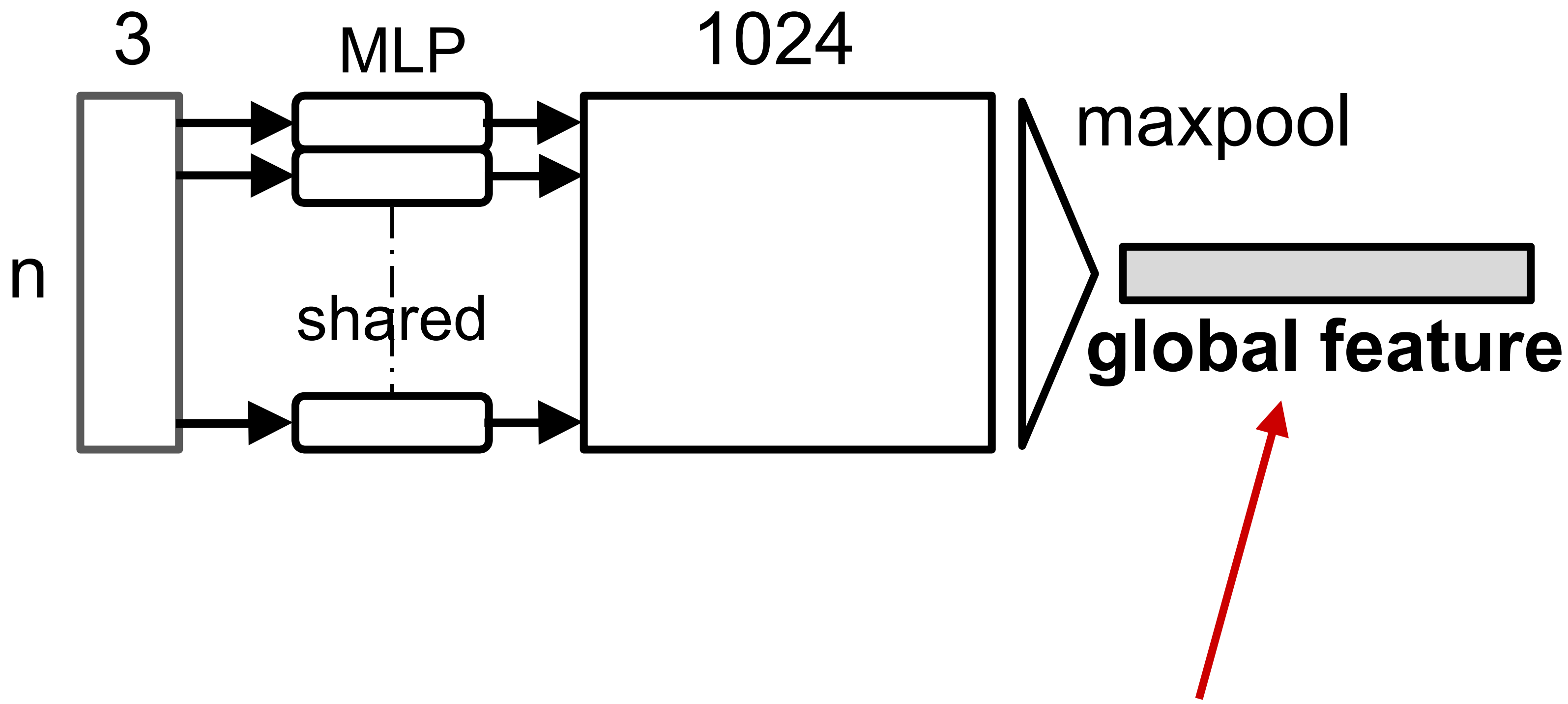
(critical points)

Visualizing Global Point Cloud Features

Original Shape:

Critical Point Sets:

Visualizing Global Point Cloud Features



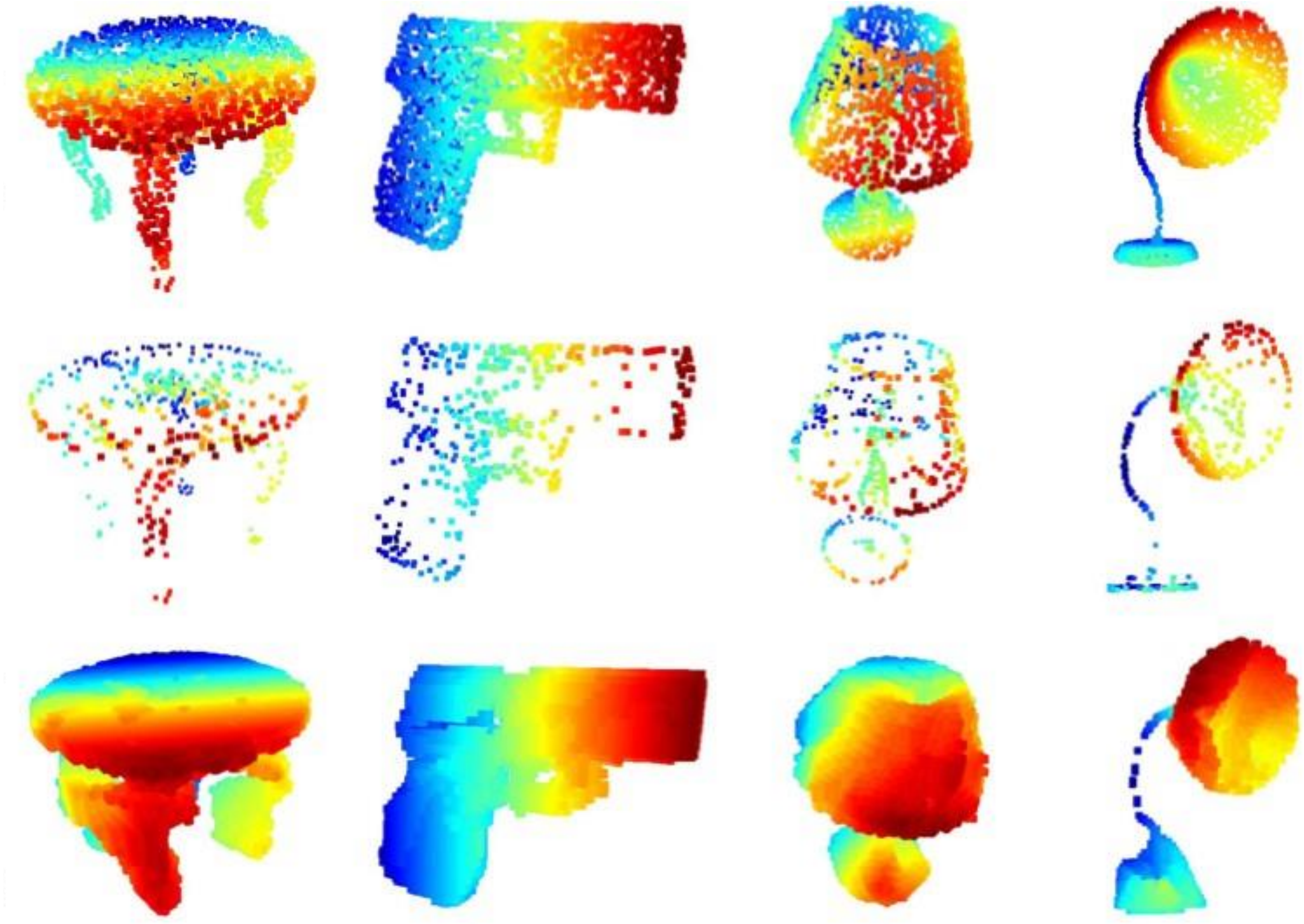
Which points won't affect the global feature?

Visualizing Global Point Cloud Features

Original Shape:

Critical Point Set:

Upper bound set:

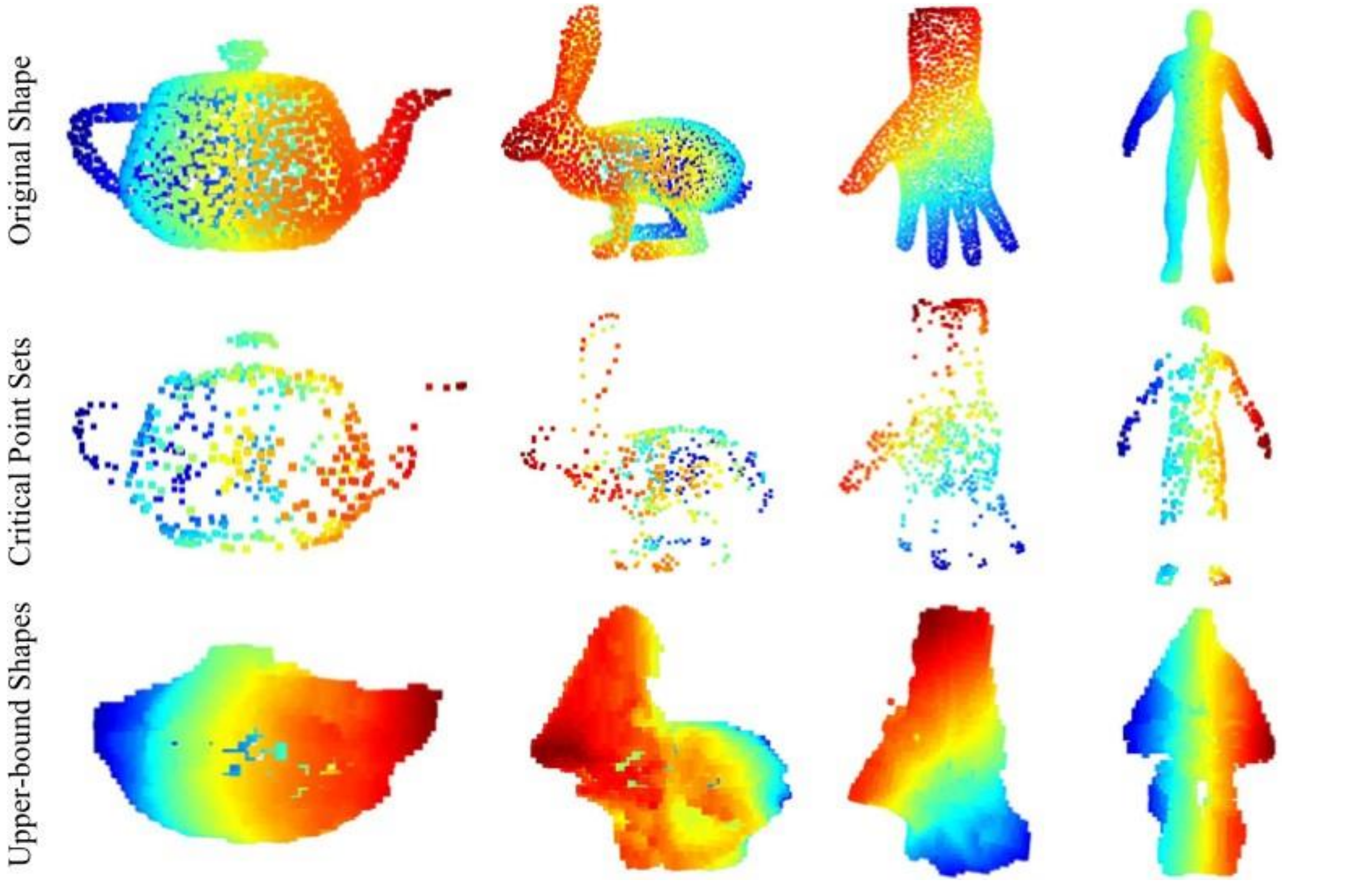


Visualizing Global Point Cloud Features (OOS)

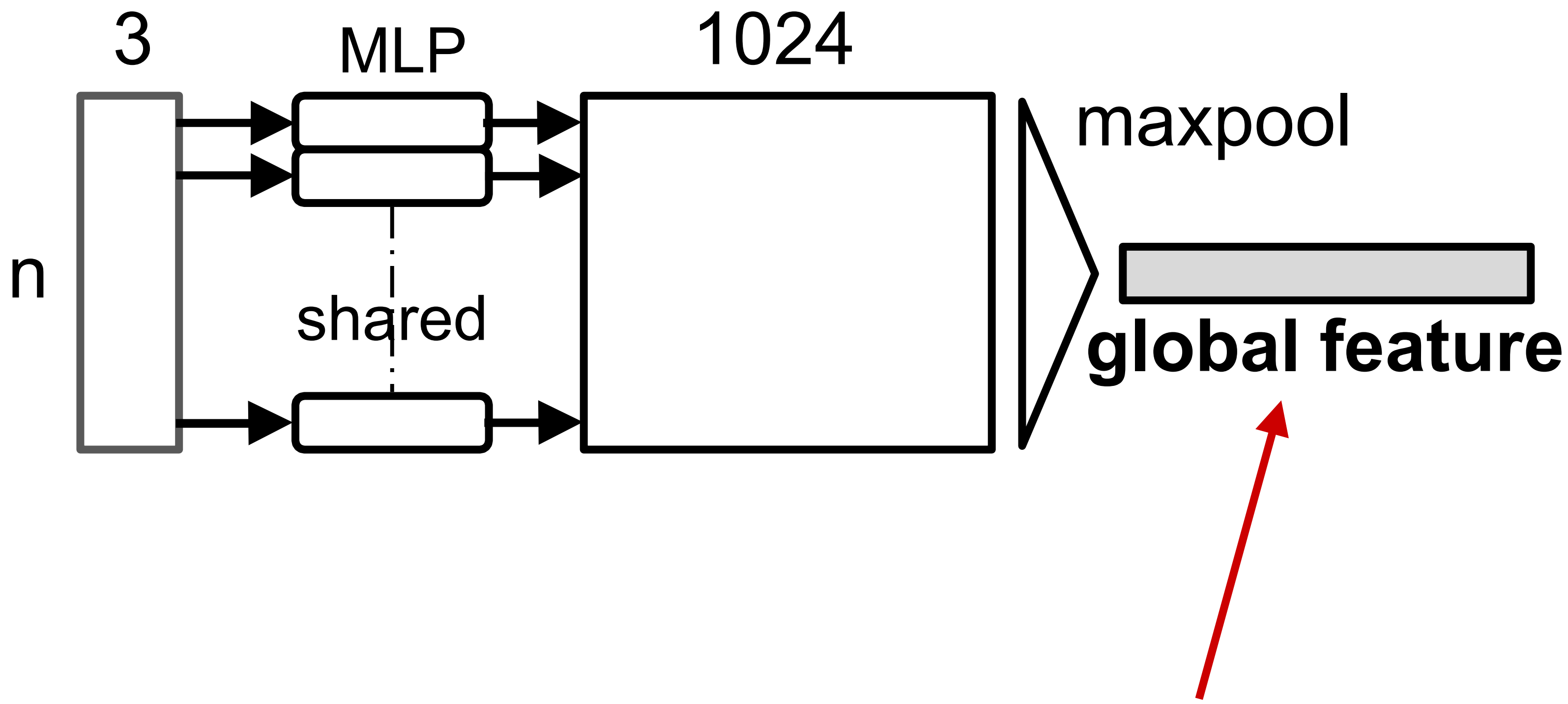
Original Shape:

Critical Point Set:

Upper bound Set:

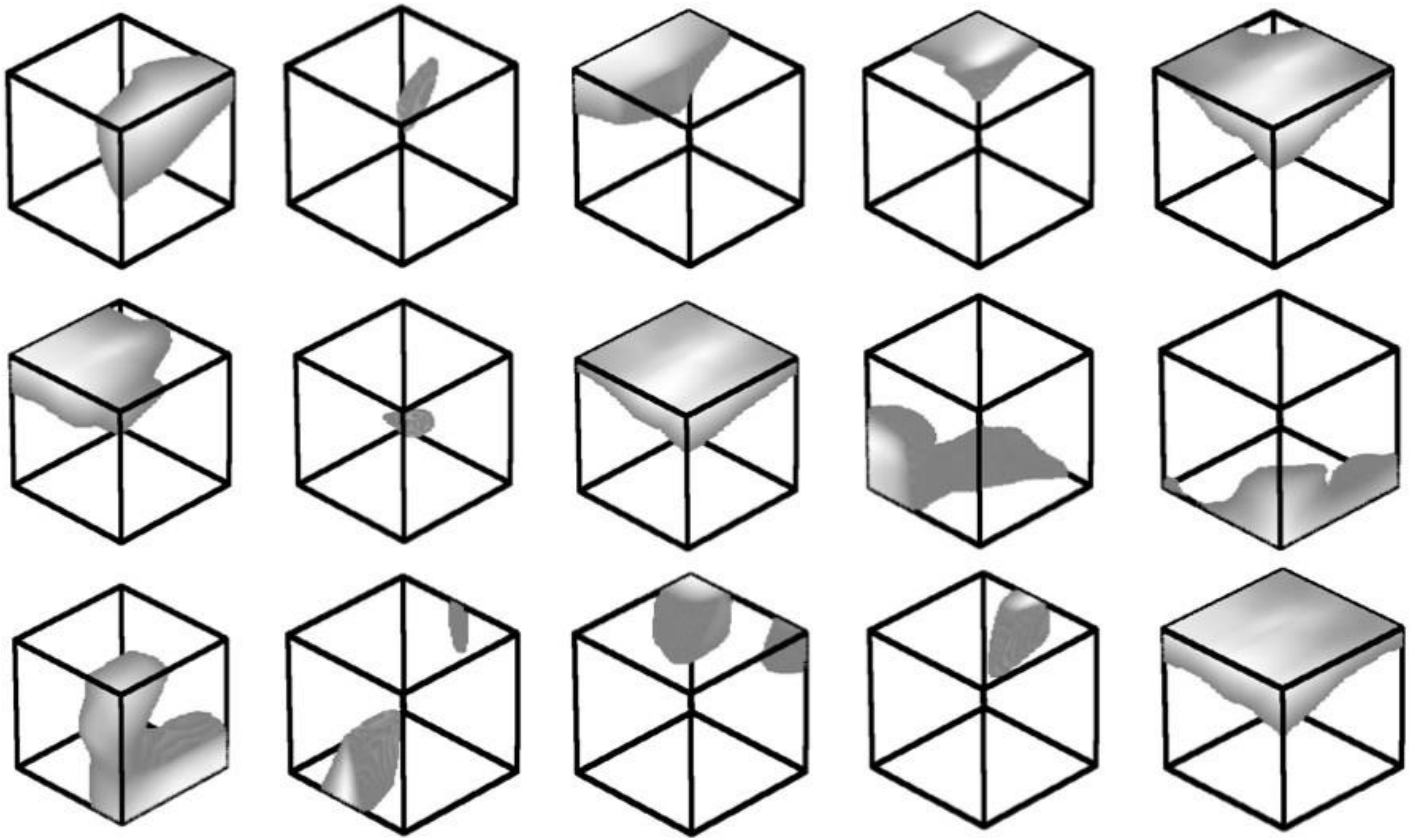


Visualizing Global Point Cloud Features



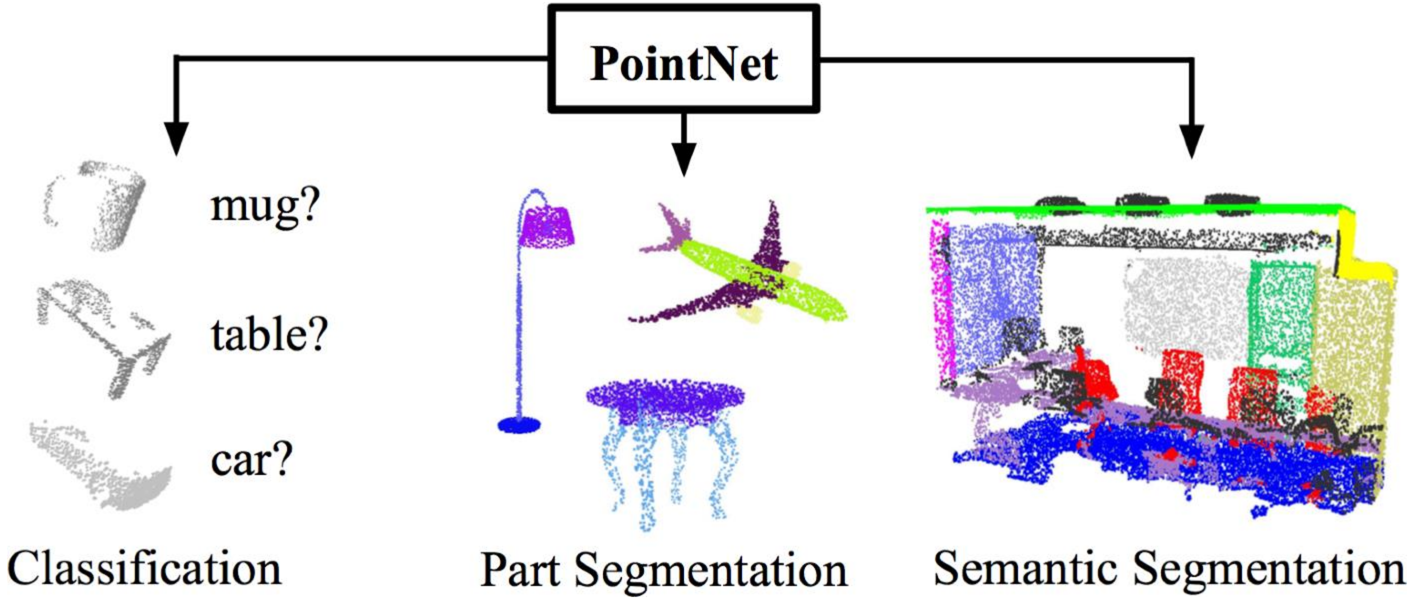
What regions influence each of these dimensions?

Visualizing Point Functions



Conclusion

- PointNet is a novel deep neural network that directly consumes point cloud.
- A unified approach to various 3D recognition tasks.
- Rich theoretical analysis and experimental results.



Code & Data Available!
http://stanford.edu/~rqi/pointnet

Speed and Model Size

	#params	FLOPs/sample
PointNet (vanilla)	0.8M	148M
PointNet	3.5M	440M
Subvolume [16]	16.6M	3633M
MVCNN [20]	60.0M	62057M

Inference time 11.6ms, 25.3ms GTX1080, batch size 8

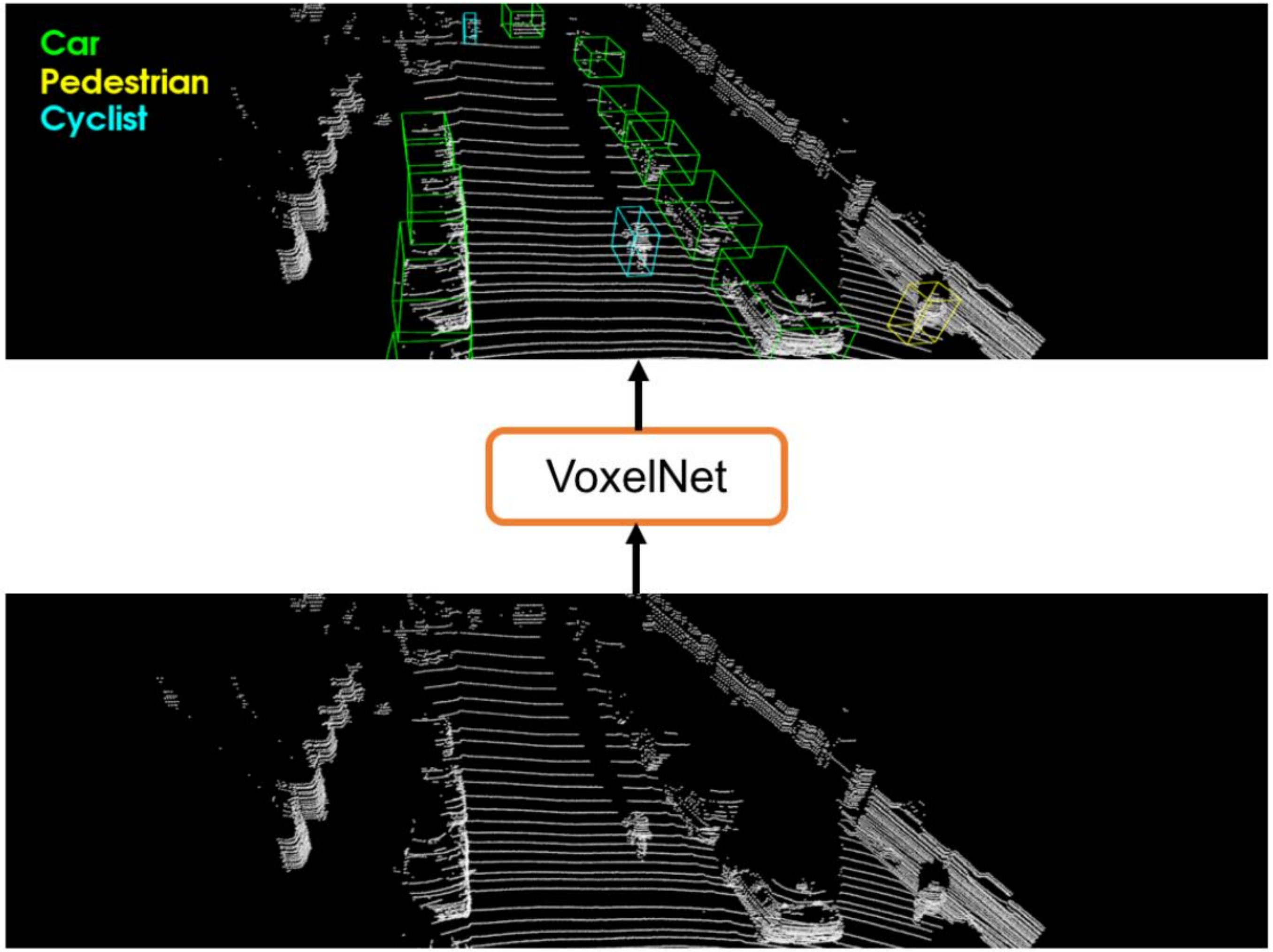
Outline

- What is lidar?
- How do we make decisions about point clouds?
  - PointNet – orderless point processing
  - VoxelNet – voxel-based point processing
  - PointPillars – bird's eye view point processing
    - Exploiting Visibility for 3D Object Detection
  - Range view object detection

Convolutions are powerful

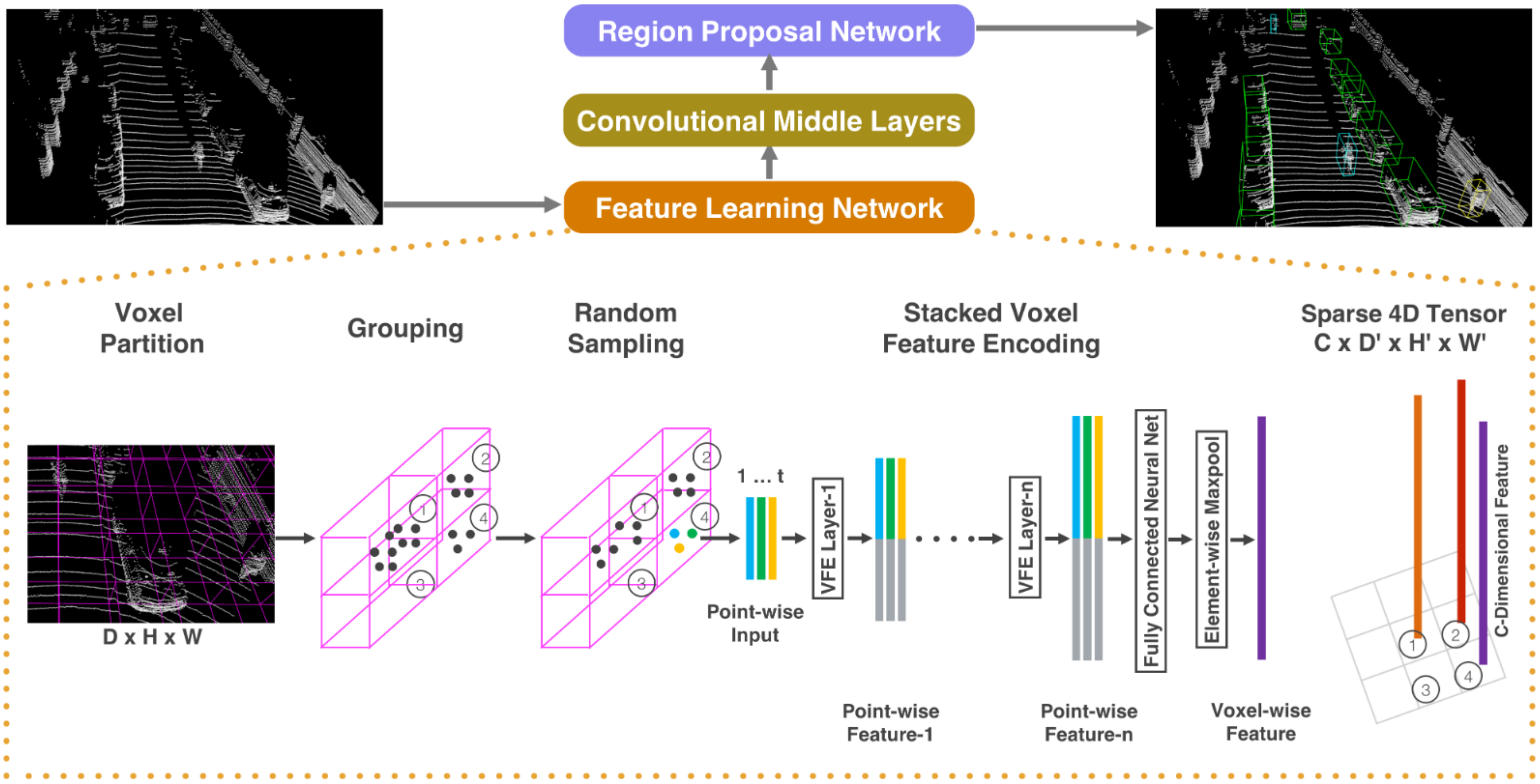
- Convolutions are how networks reason about neighborhoods and spatial relationships.
- PointNet has limited ability to identify things like corners, junctions, straight lines, etc.

VoxelNet

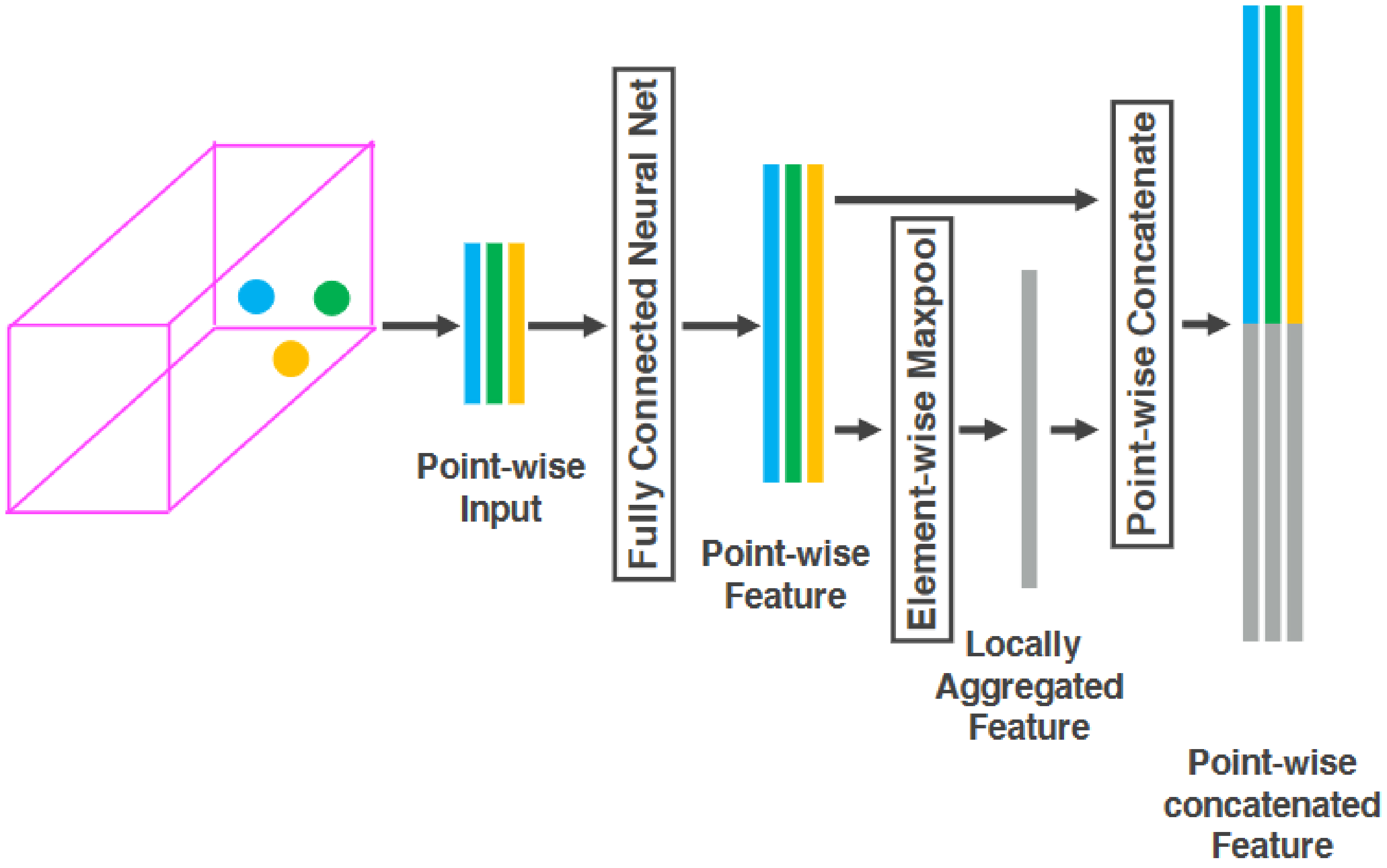


VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection
Yin Zhou and Oncel Tuzel. CVPR 2018

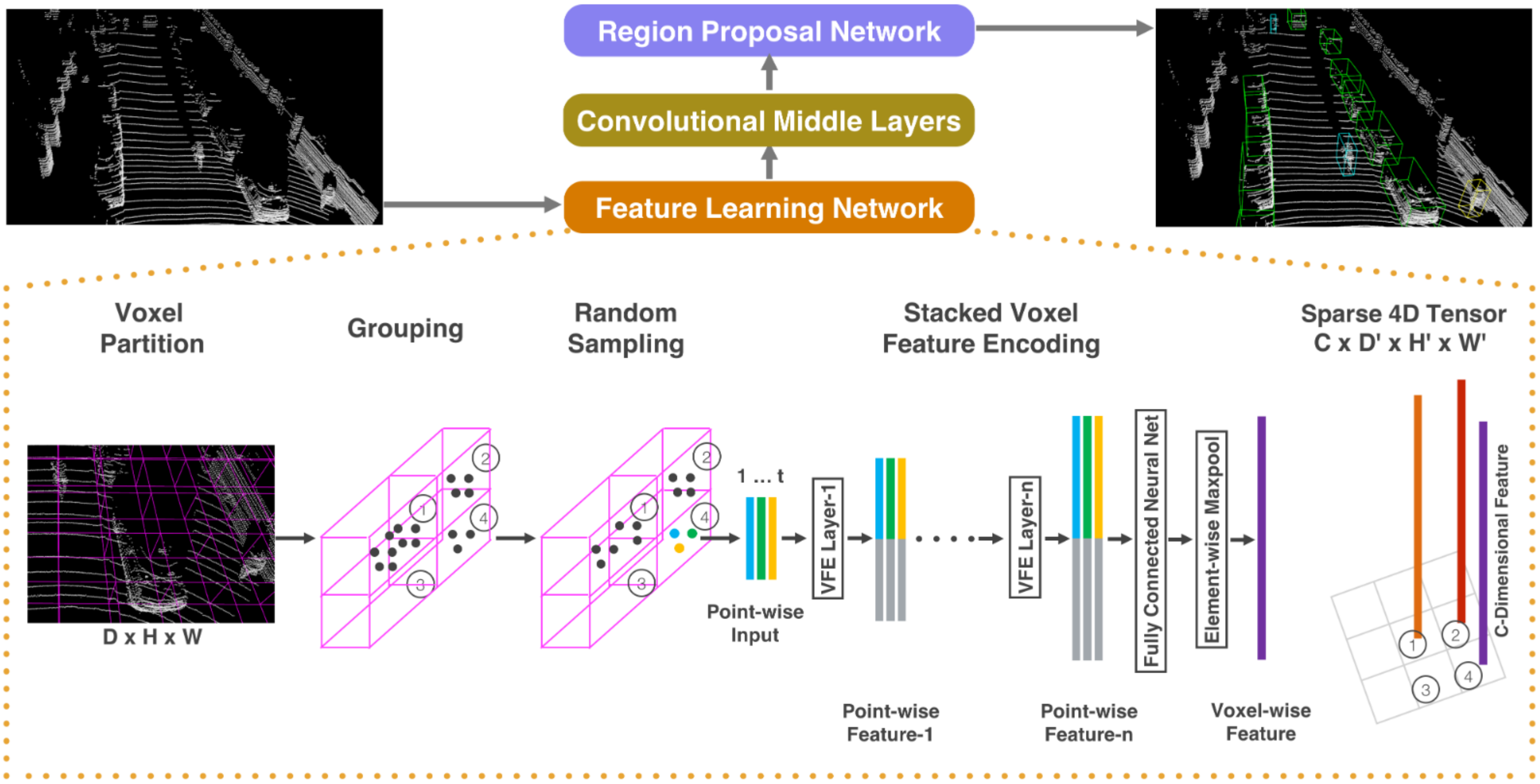
VoxelNet Overview



VoxelNet Voxel encoding looks a lot like PointNet



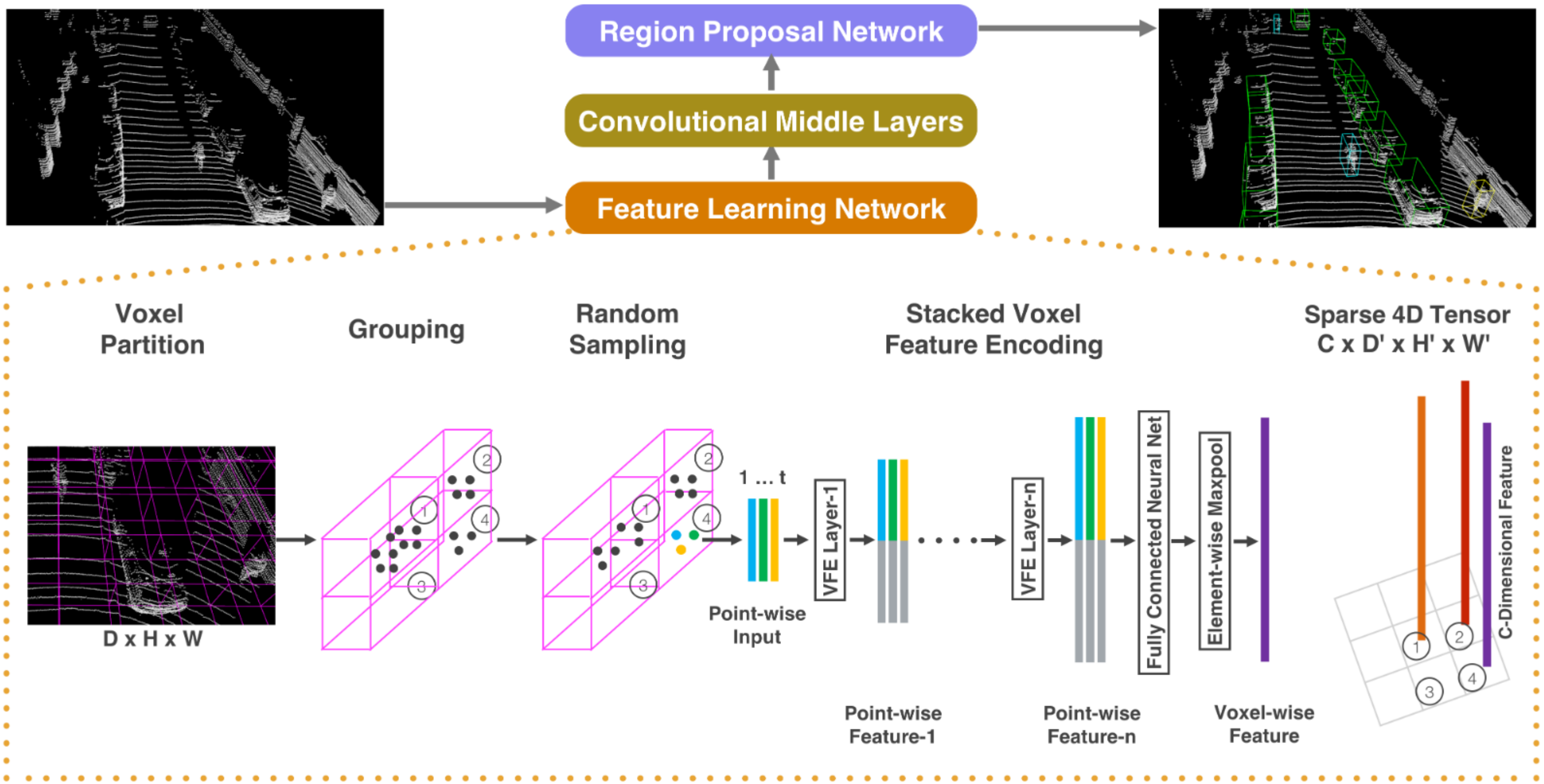
VoxelNet Overview



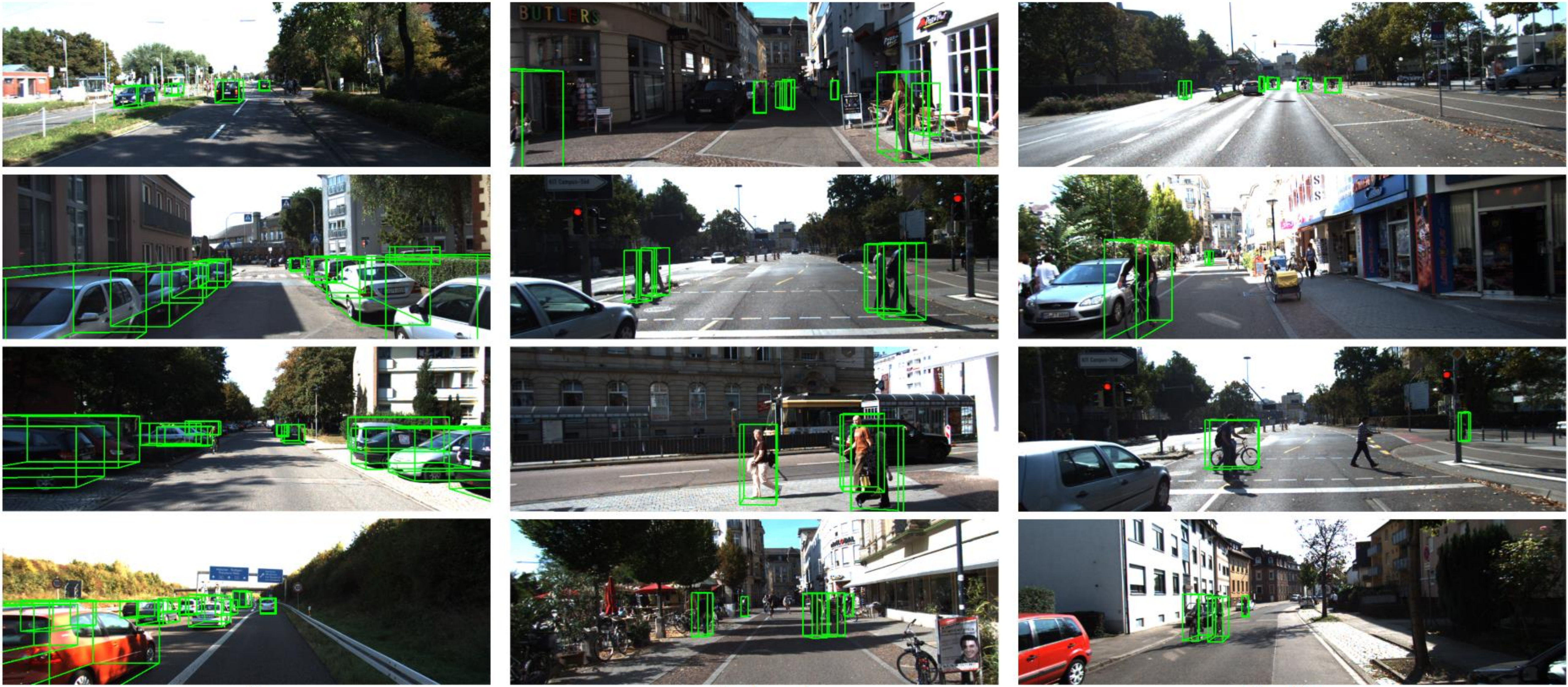
VoxelNet “Convolutional Middle Layers”

- For car detection, divide the world into 10 x 400 x 352 voxels, corresponding to voxels that are 40 cm tall and 20 cm in width/length.
- Uses **3D** convolutions instead of 2D as we’ve seen before.
- The Z / height dimension gets downsampled away after many layers

VoxelNet Overview



VoxelNet qualitative results



Car Pedestrian Cyclist

VoxelNet quantitative results

Method	Modality	Car			Pedestrian			Cyclist		
		Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
Mono3D [3]	Mono	2.53	2.31	2.31	N/A	N/A	N/A	N/A	N/A	N/A
3DOP [4]	Stereo	6.55	5.07	4.10	N/A	N/A	N/A	N/A	N/A	N/A
VeloFCN [22]	LiDAR	15.20	13.66	15.98	N/A	N/A	N/A	N/A	N/A	N/A
MV (BV+FV) [5]	LiDAR	71.19	56.60	55.30	N/A	N/A	N/A	N/A	N/A	N/A
MV (BV+FV+RGB) [5]	LiDAR+Mono	71.29	62.68	56.56	N/A	N/A	N/A	N/A	N/A	N/A
HC-baseline	LiDAR	71.73	59.75	55.69	43.95	40.18	37.48	55.35	36.07	34.15
VoxelNet	LiDAR	**81.97**	**65.46**	**62.85**	**57.86**	**53.42**	**48.87**	**67.17**	**47.65**	**45.11**

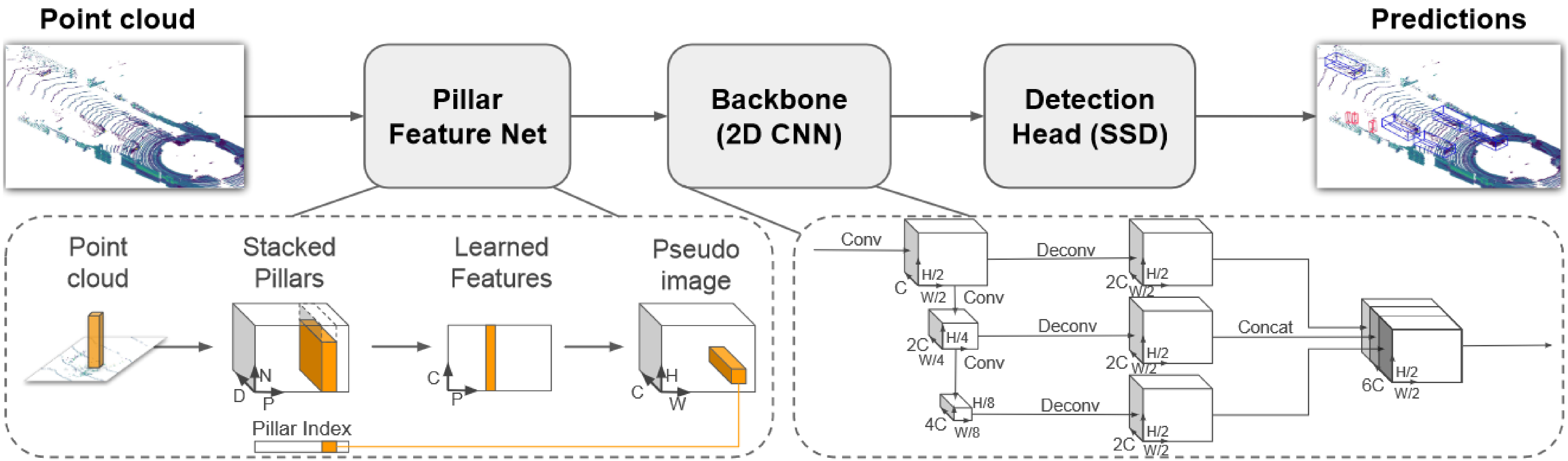
Evaluation on KITTI according to 3D IoU

Outline

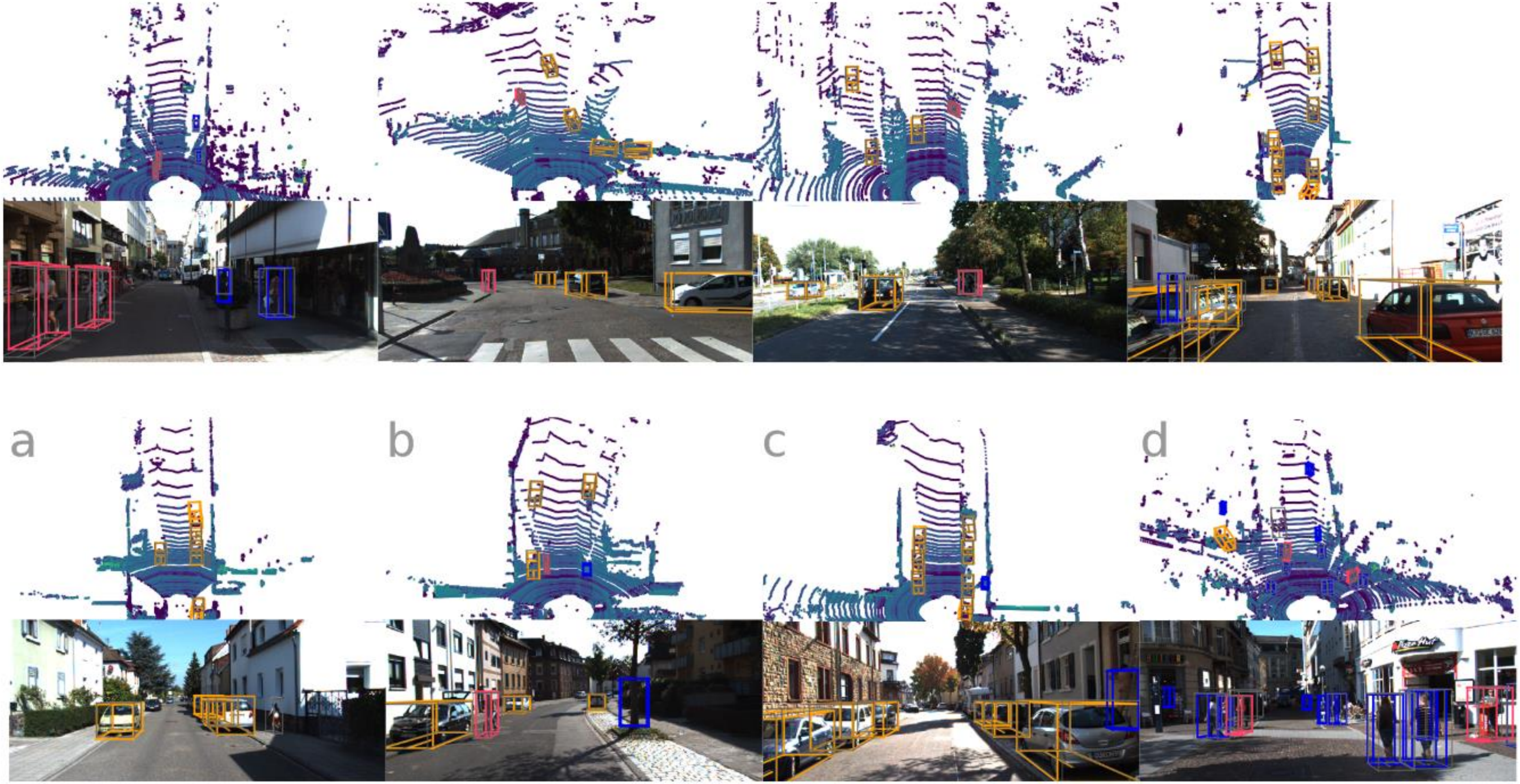
- What is lidar?
- How do we make decisions about point clouds?
  - PointNet – orderless point processing
  - VoxelNet – voxel-based point processing
  - PointPillars – bird's eye view point processing
    - Exploiting Visibility for 3D Object Detection
  - Range view object detection

The X, Y, and Z directions aren’t the same

PointPillars



PointPillars: Fast Encoders for Object Detection from Point Clouds
Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, Oscar Beijbom. CVPR 2019


a
b
c
d

Outline

- What is lidar?
- How do we make decisions about point clouds?
  - PointNet – orderless point processing
  - VoxelNet – voxel-based point processing
  - PointPillars – bird's eye view point processing
    - Exploiting Visibility for 3D Object Detection
  - Range view object detection

What You See Is What You Get
Exploiting Visibility for 3D Object Detection

Peiyun Hu, Jason Ziglar, David Held, Deva Ramanan

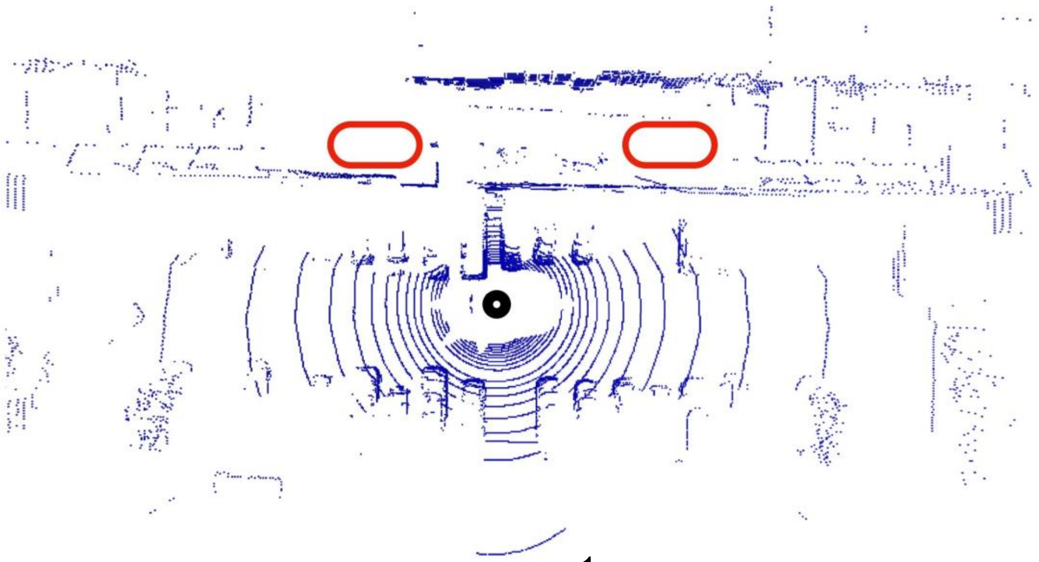
Carnegie Mellon University Argo AI

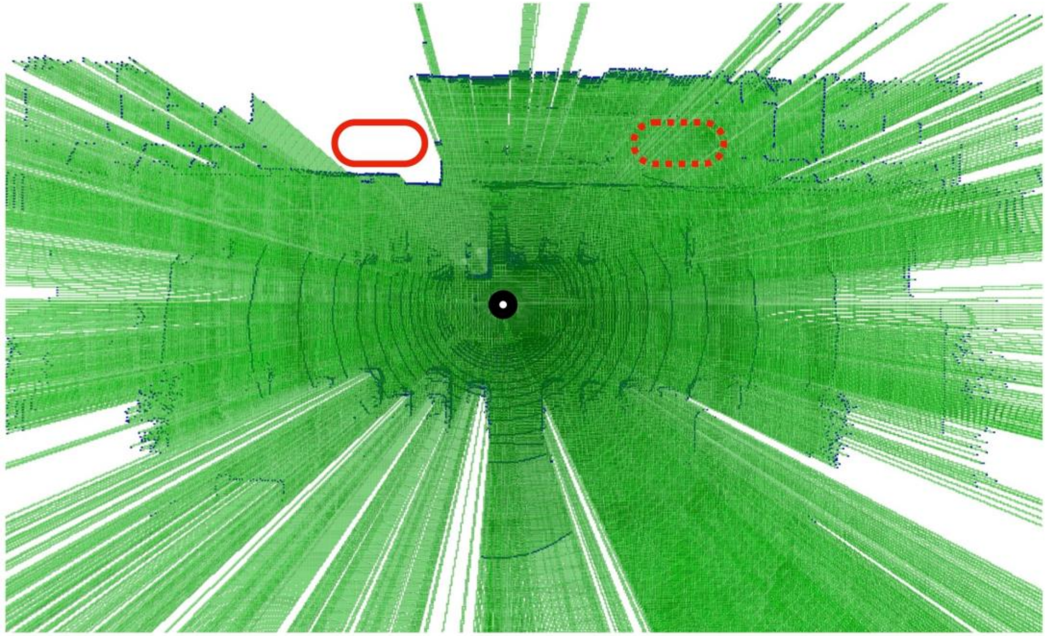
THE ROBOTICS INSTITUTE

ARGO AI

CVPR 2020

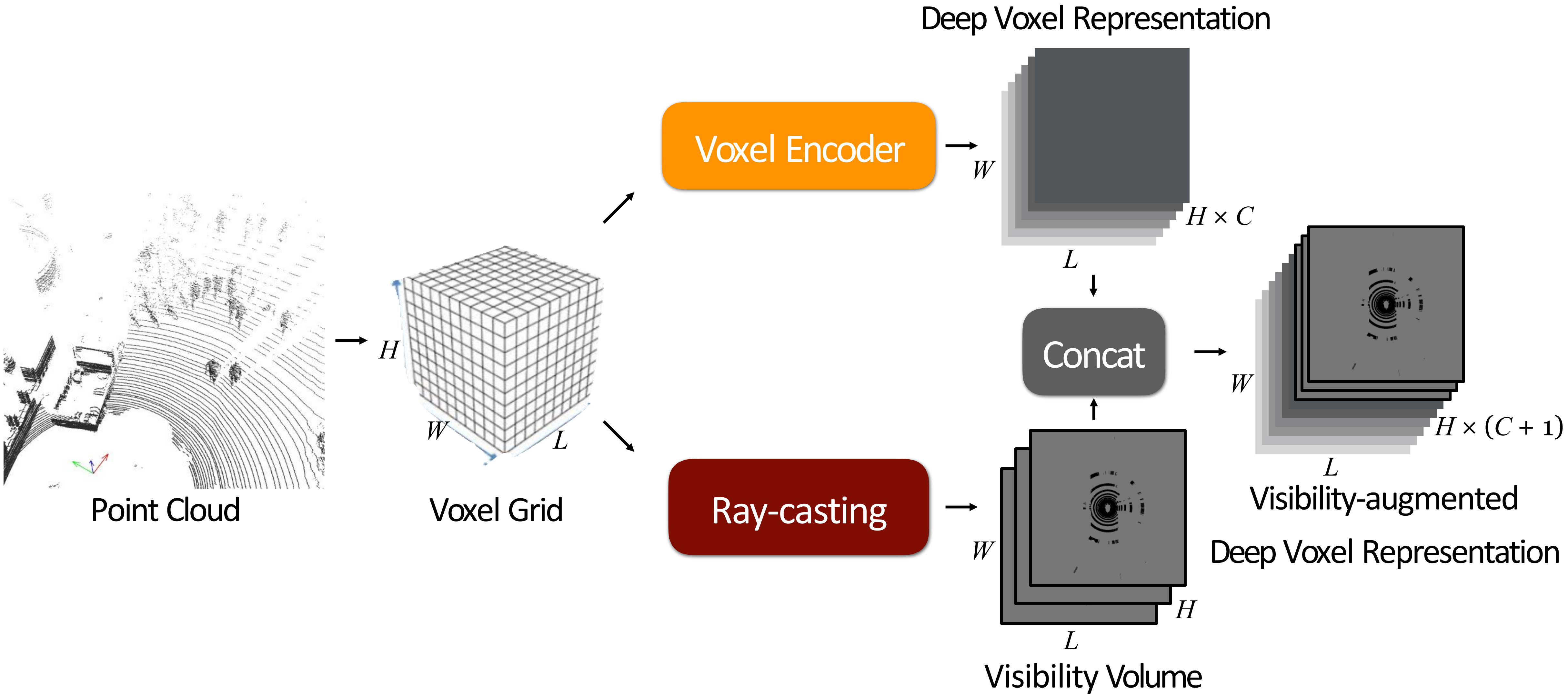
What is a good representation for LiDAR data?



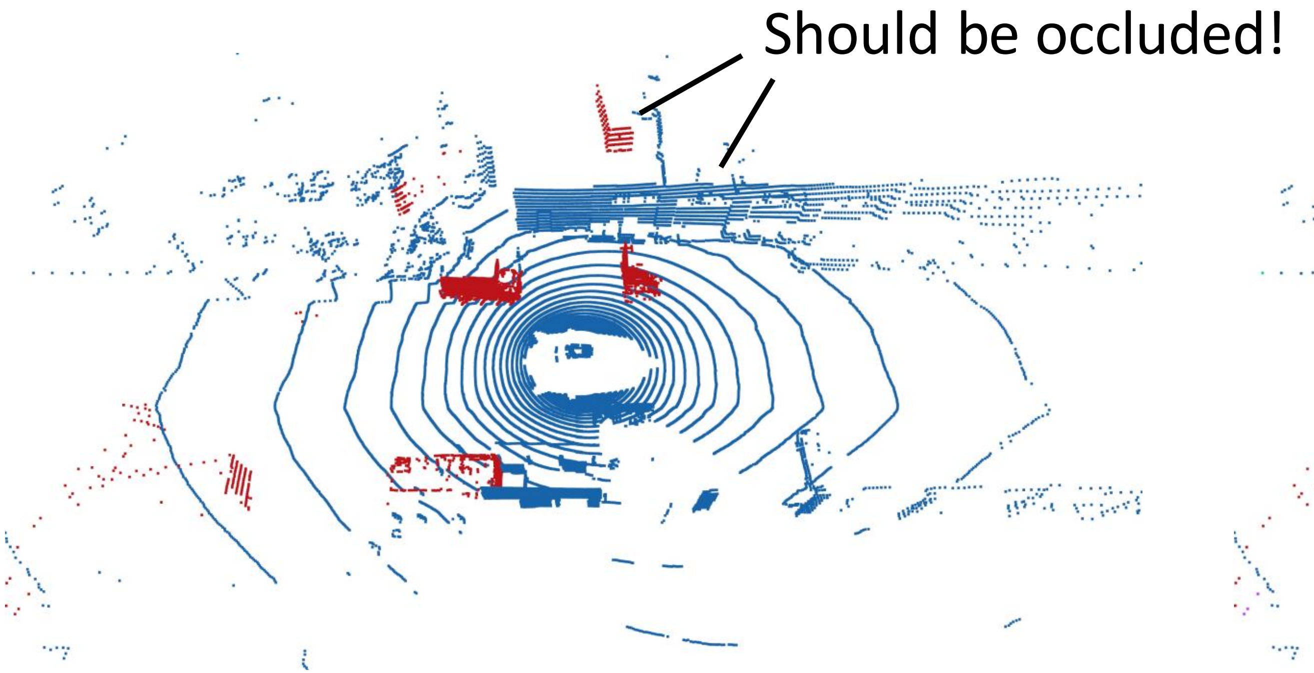


- LiDAR data provides more than just point measurements
- Rays emanating from the sensor to each 3D point must pass through free space
- Representing LiDAR data as (x, y, z)s fundamentally destroys such freespace information

A Simple Approach to Augment Visibility



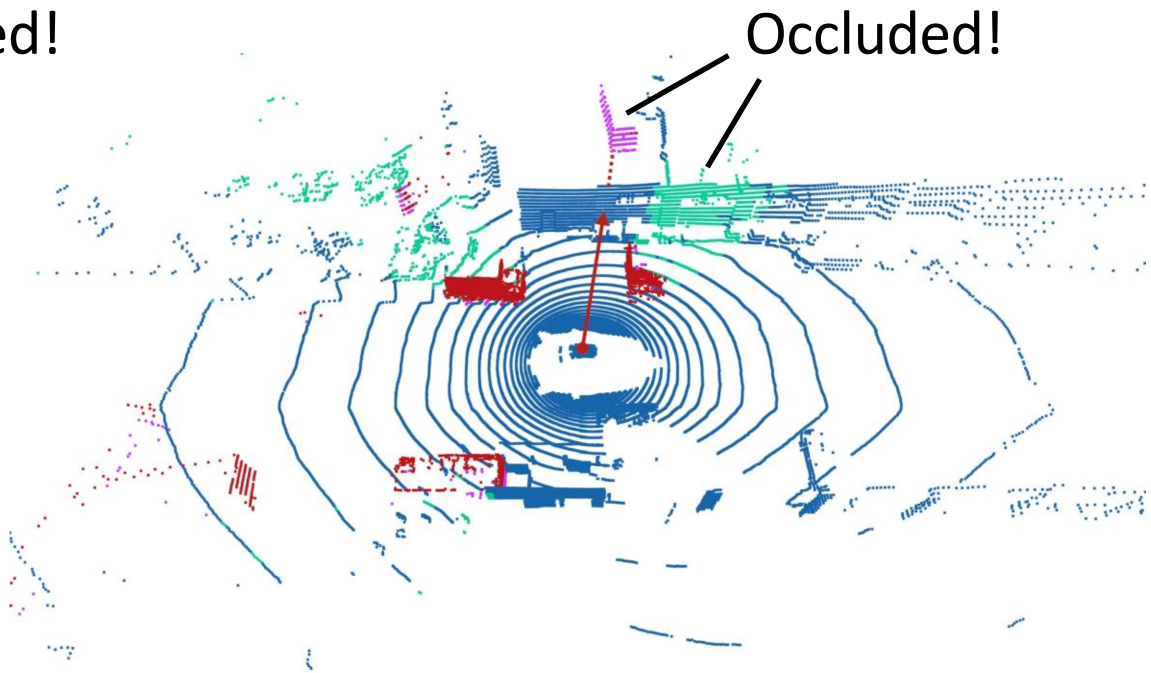
Visibility-aware LiDAR Synthesis



Naive Object Augmentation

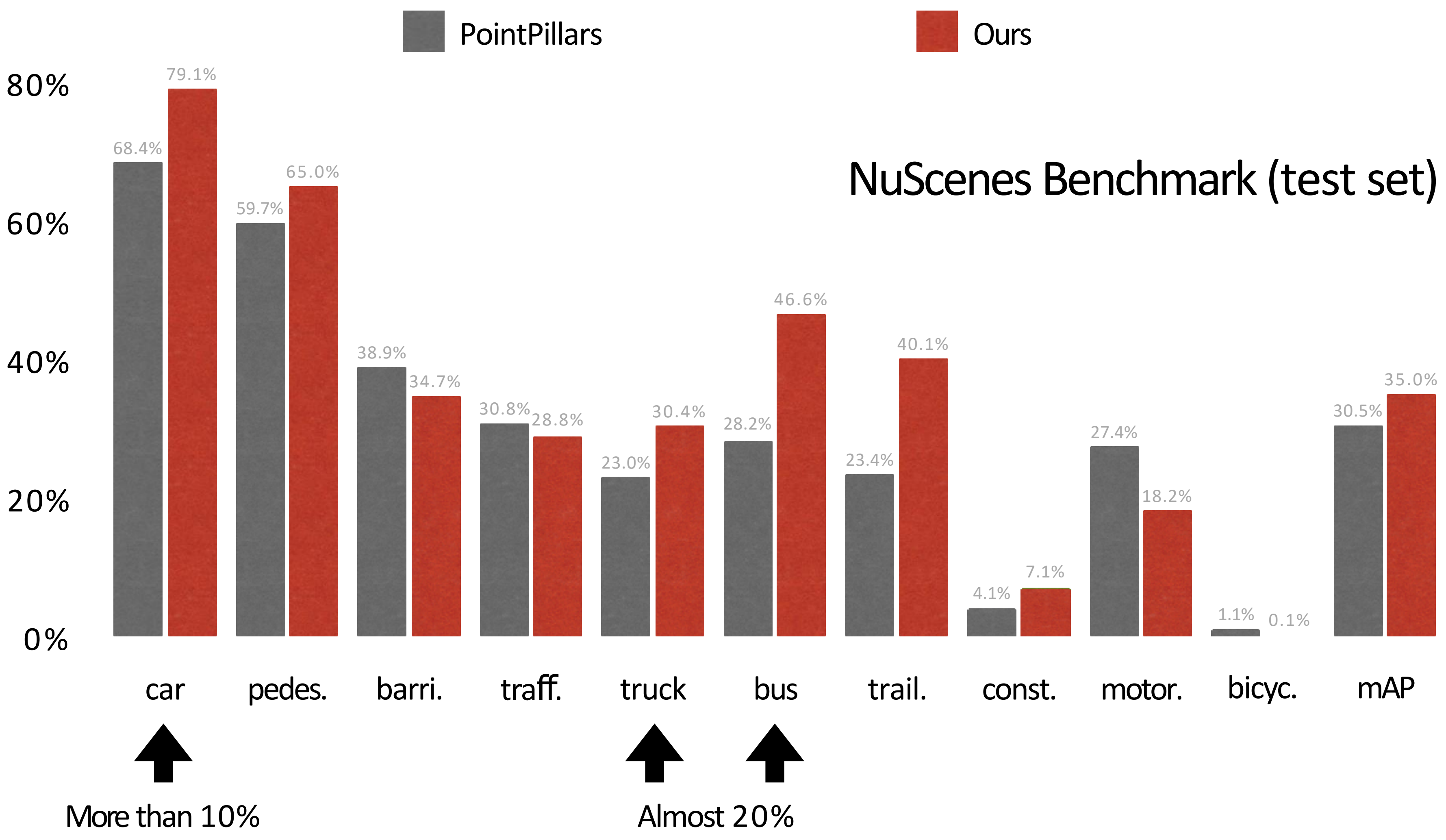
PointPillars, Lang et al., CVPR'19

SECOND, Yan et al., Sensors'18



Visibility-aware Object Augmentation

Improve PointPillars by 4.5% in overall mAP



Outline

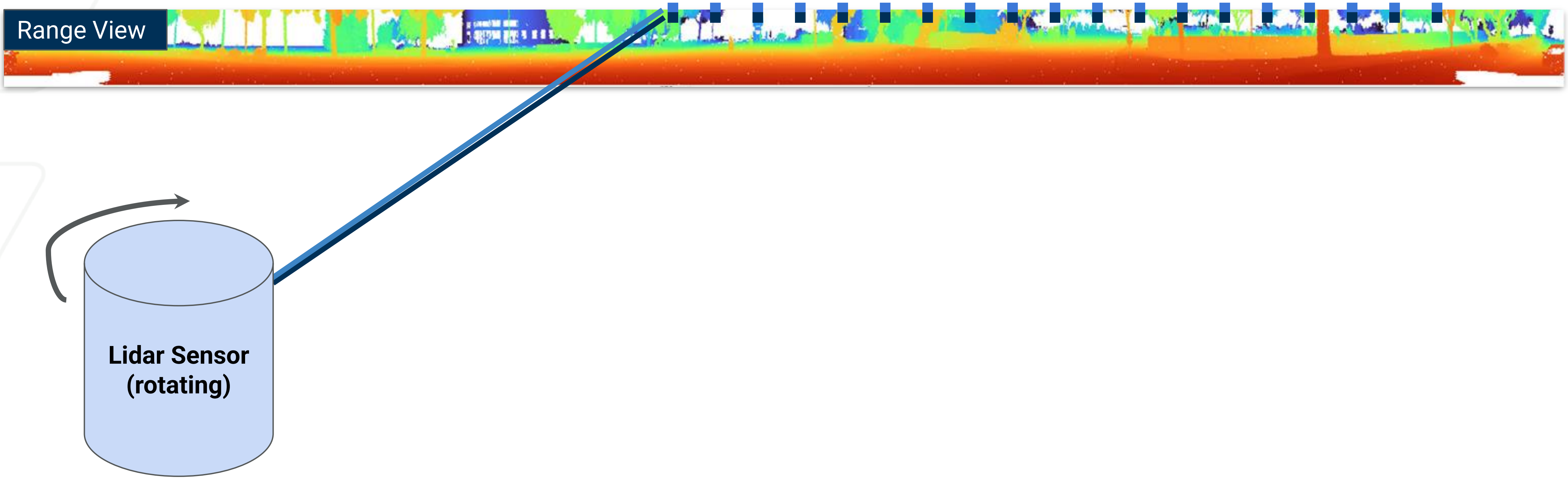
- What is lidar?
- How do we make decisions about point clouds?
  - PointNet – orderless point processing
  - VoxelNet – voxel-based point processing
  - PointPillars – bird's eye view point processing
    - Exploiting Visibility for 3D Object Detection
  - Range view object detection

What Matters in Range View 3D Object Detection

https://github.com/benjaminrwilson/range-view-3d-detection

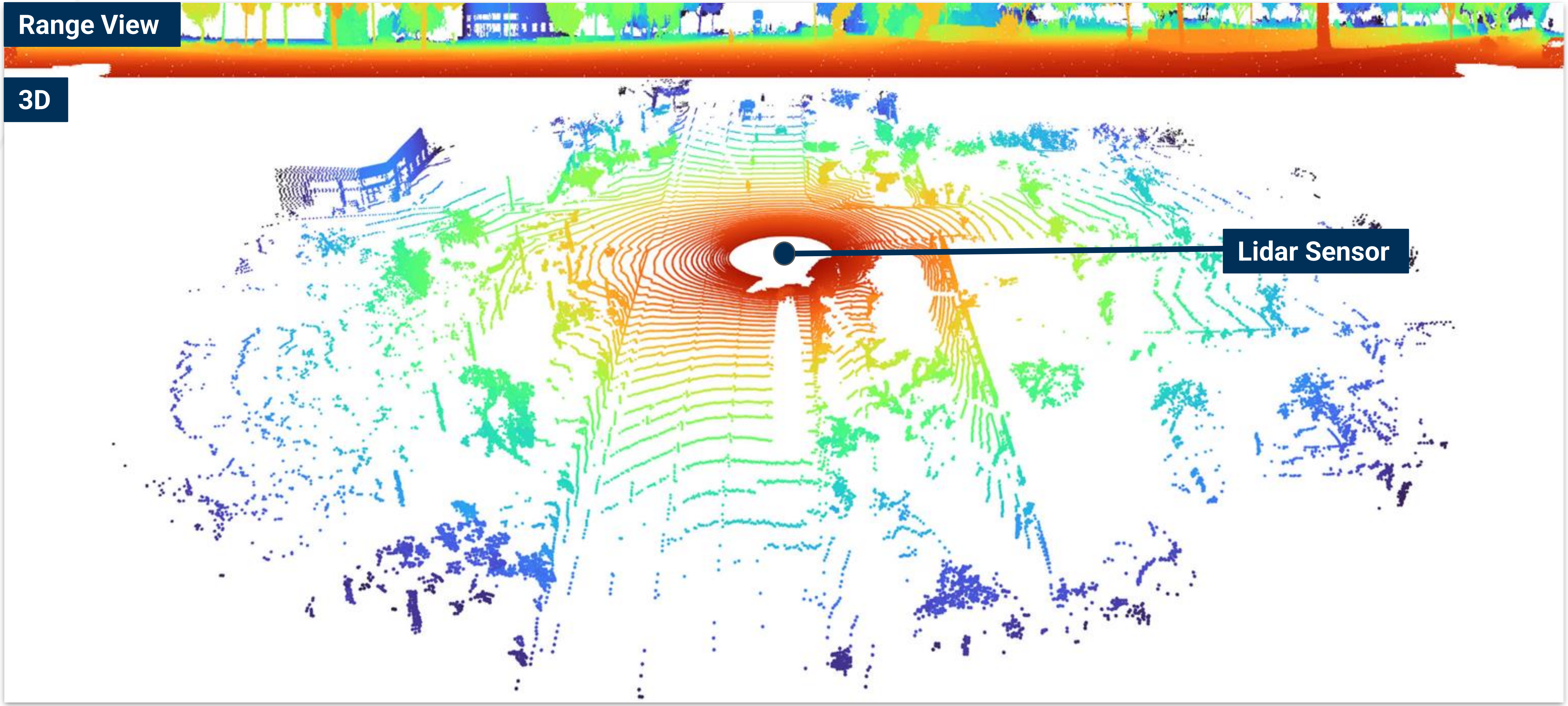
Benjamin Wilson, Nicholas Autio Mitchell, Jhony Kaesemodel Pontes, James Hays.
"What Matters in Range View 3D Object Detection." 8th Annual Conference on Robot Learning. 2024.

What is the Range View?



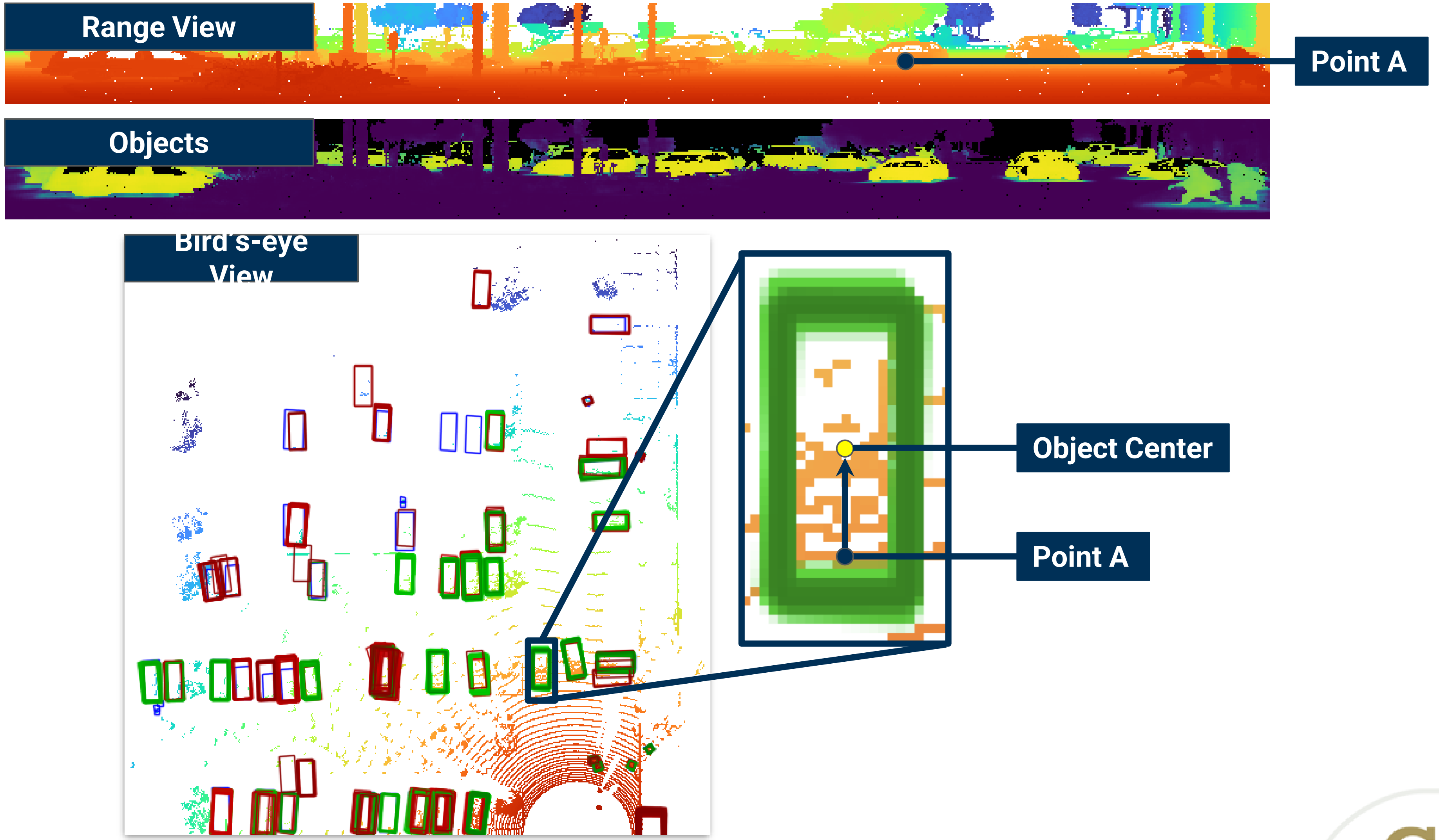
Georgia Tech

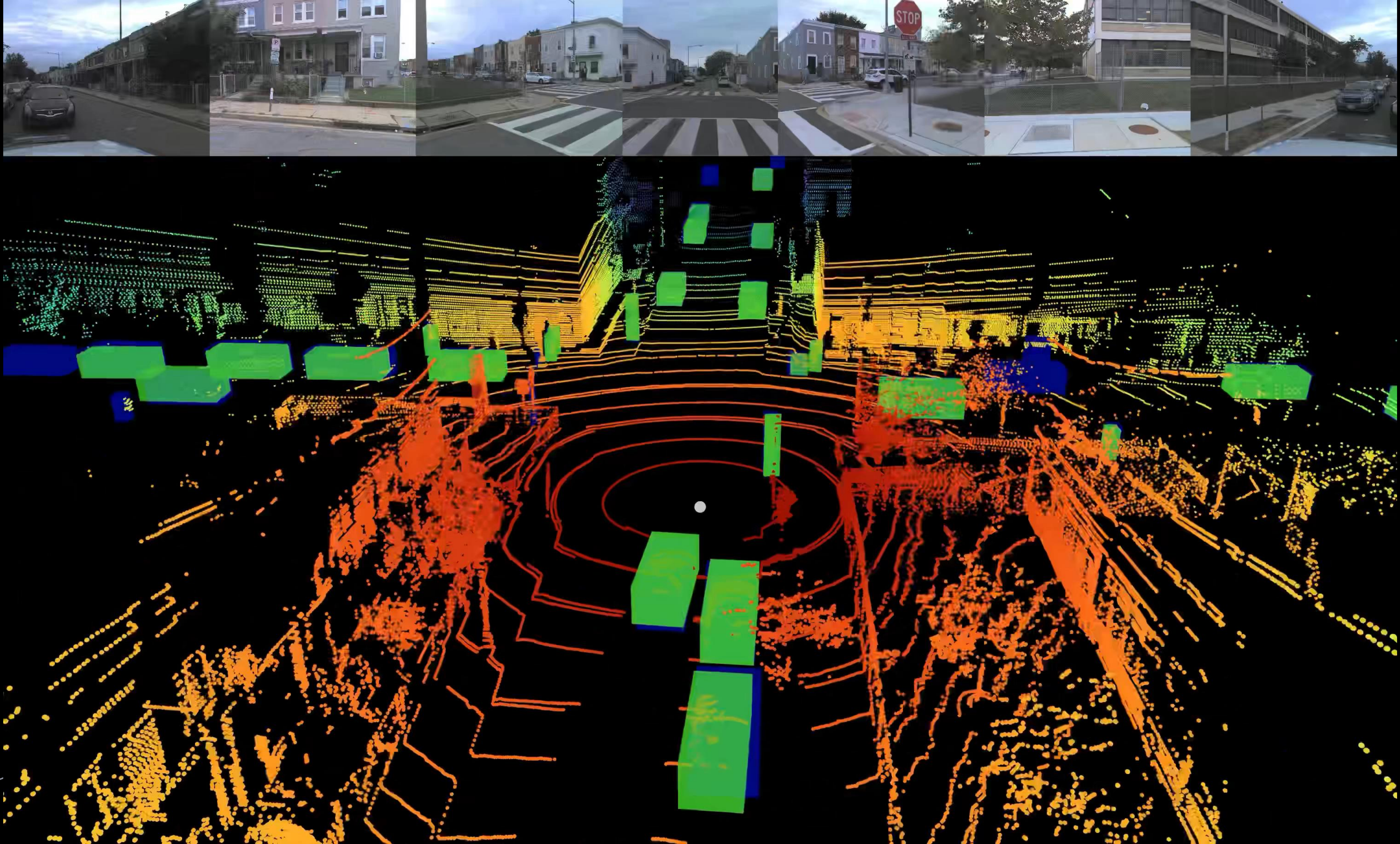
What is the Range View?

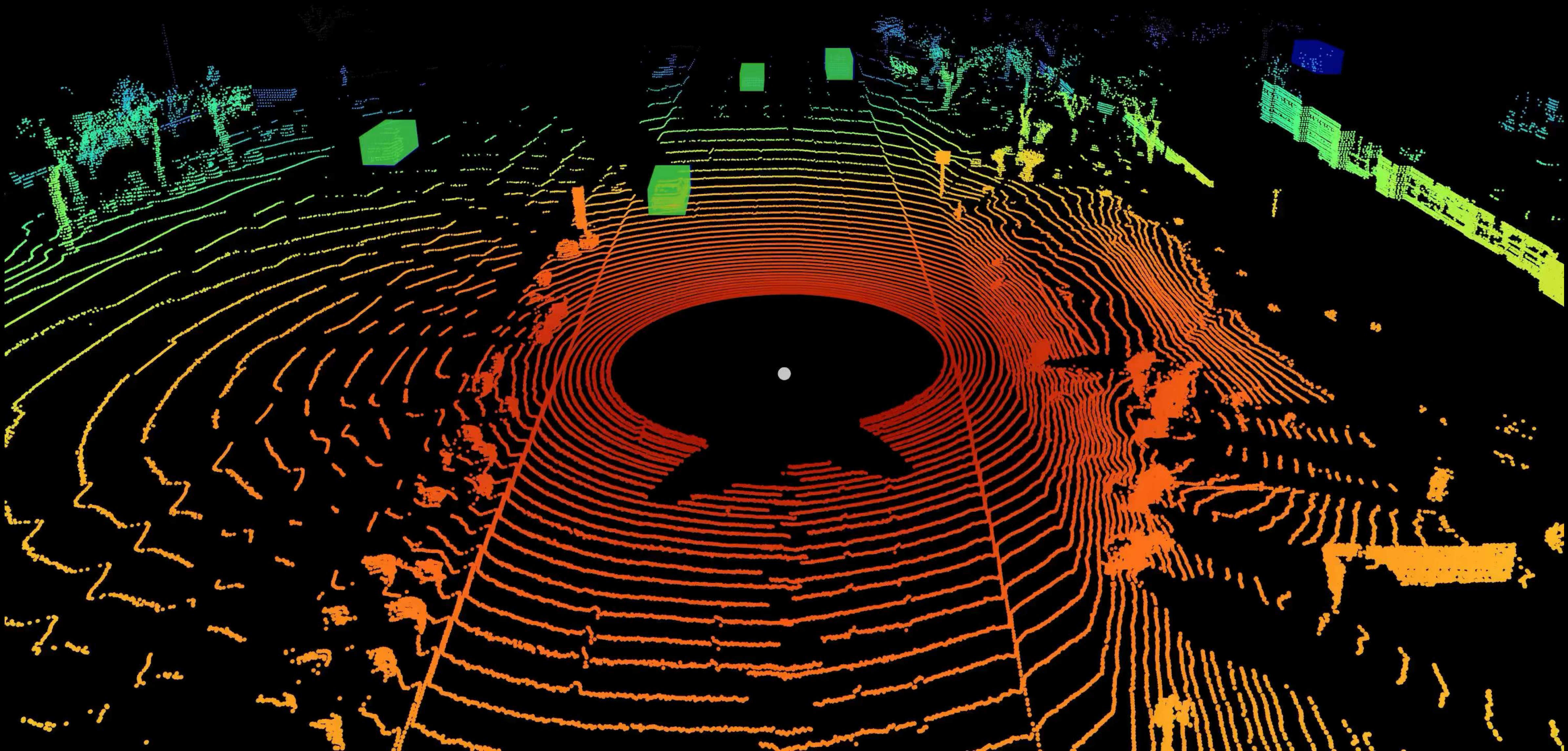


Georgia Tech

3D Object Detection in the Range View




STOP



Outline

- What is lidar?
- How do we make decisions about point clouds?
  - PointNet – orderless point processing
  - VoxelNet – voxel-based point processing
  - PointPillars – bird’s eye view point processing
    - Exploiting Visibility for 3D Object Detection
  - Range view object detection

Summary

- Popular CNN backbones aren’t a direct fit for 3D point processing tasks.
- It’s not clear how to best use deep learning on 3D data
  - Use a truly permutation invariant representation (PointNet)
  - Render multiple 2D views of the 3D data
  - Use a voxel representation (VoxelNet)
  - Use a bird’s a view representation (PointPillars)
  - Use a range image