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

李声涛　17214643　宋日辉　17214675
黄莉　17214619　管卓群　17214612
周晓梅　17214729

School of Data and Computer Science, SYSU

December 19, 2017

# Content

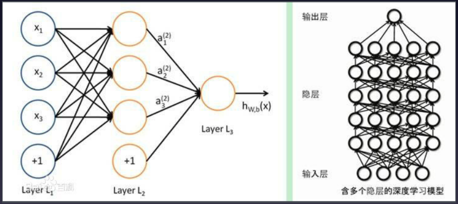
- Introduction
- Problem Statement
- Deep Learning on Point Sets
- Experiment
- Conclusion

# Introduction

- *Title:*
  - PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation
- *Keywords:*
  - Deep Learning
  - Point Sets (3D)
  - Classification and Segmentation
  - PointNet

# Deep Learning

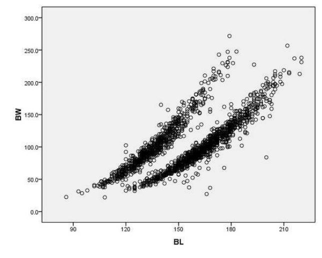
- Part of a broader family of machine learning methods
- Based on learning data representations
- Deep learning architectures: DNN, DBN, RNN, CNN...



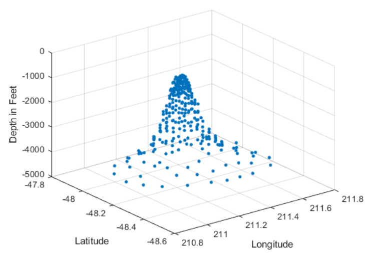
(a) Deep Learning Structure

# Point Sets

- A set of points.
- 2D point: $(x, y)$
- 3D point: $(x, y, z)$



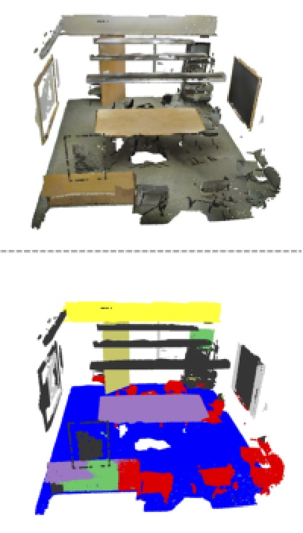
(b) 2D Point Set



(c) 3D Point Set (Point Cloud)

# Classification and Segmentation

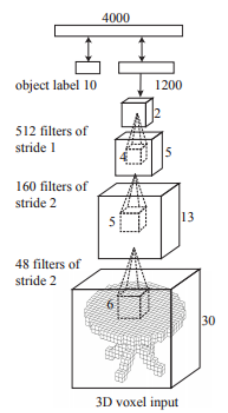
- Classification: distinguish different classes
- Segmentation: figure out different parts



(d) Room

# Neural Network Based Methods

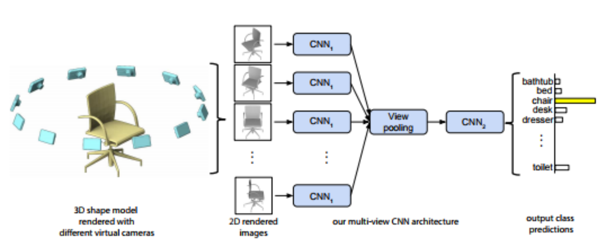
- Volumetric CNNs: 3D voxel grids
  - Constrained by resolution



(e) 3D ShapeNet

# Neural Network Based Methods

- Multi-view CNNs: collections of images
    - Nontrivial to extend them to scene understanding or other 3D tasks.



(f) MVCNN

# PointNet

- A novel deep net architecture
- Input: points
- Output: points and scores for each label
- Tasks: 3D shape classification, shape part segmentation, and scene semantic parsing
- Simple, effective and robust

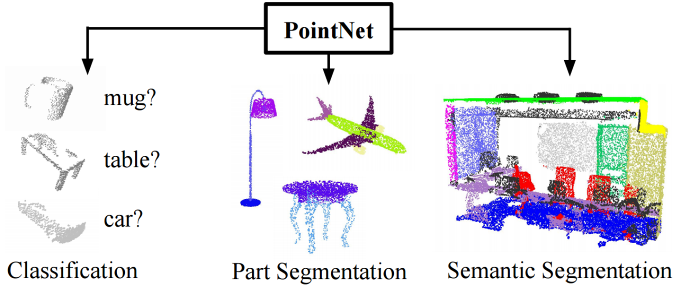


Figure: Task of PointNet

# Problem Statement

- A point cloud is represented as a set of 3D points:
    - $\{P_i | i = 1, ..., n\}$
    - $P_i = (x, y, z)$
    - Extra feature channels: color, normal, etc.

# Object Classification

- The input:
  Directly sampled from a shape
  Pre-segmented from a scene point cloud.
- The output:
  This deep network outputs k scores for all the k candidate classes.

# Semantic Segmentation

- The input:
  A single object for part region segmentation
  A sub-volume from a 3D scene for object region segmentation.
- The output:
  This model will output $n \times m$ scores for each of the n points and each of the m semantic subcategories.

# Deep Learning on Point Sets

- Properties of Point Sets
- PointNet Architecture

# Properties of Point Sets

- Unordered
- Interaction among points
- Invariance under transformations

# PointNet Architecture

- Symmetry Function for Unordered Input
- Local and Global Information Aggregation
- Joint Alignment Network

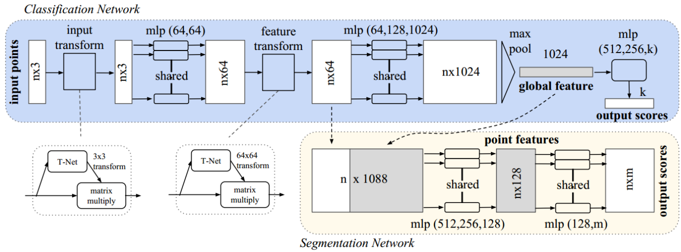


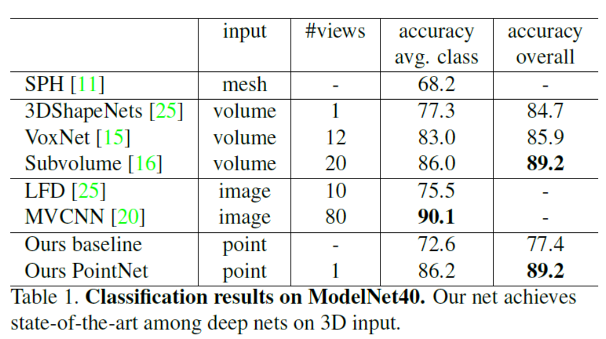
Figure: Architecture

# Experiments

- Applications
- Architecture Design Analysis
- Visualizing PointNet
- Time and Space Complexity Analysis

# Applications-3D Object Classification

- 12,311 CAD models
- from 40 man-made object categories,
- split into 9,843 for training and 2,468 for testing

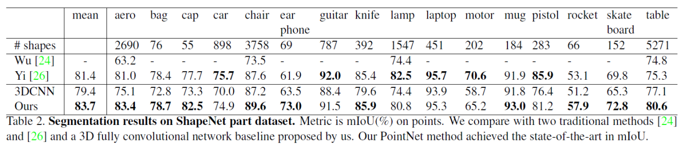


| | input | #views | accuracy<br>avg. class | accuracy<br>overall |
|---|---|---|---|---|
| SPH [11] | mesh | - | 68.2 | - |
| 3DShapeNets [25] | volume | 1 | 77.3 | 84.7 |
| VoxNet [15] | volume | 12 | 83.0 | 85.9 |
| Subvolume [16] | volume | 20 | 86.0 | **89.2** |
| LFD [25] | image | 10 | 75.5 | - |
| MVCNN [20] | image | 80 | **90.1** | - |
| Ours baseline | point | - | 72.6 | 77.4 |
| Ours PointNet | point | 1 | 86.2 | **89.2** |

Table 1. **Classification results on ModelNet40.** Our net achieves state-of-the-art among deep nets on 3D input.

# Applications-3D Object Part Segmentation

- ShapeNet part data set
- 16,881 shapes from 16 categories, annotated with 50 parts in total
- mIoU?

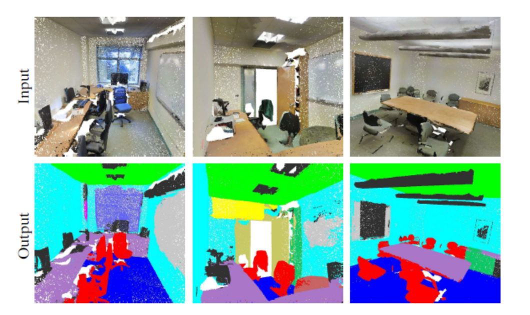


| | mean | aero | bag | cap | car | chair | ear phone | guitar | knife | lamp | laptop | motor | mug | pistol | rocket | skate board | table |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # shapes | | 2690 | 76 | 55 | 898 | 3758 | 69 | 787 | 392 | 1547 | 451 | 202 | 184 | 283 | 66 | 152 | 5271 |
| Wu [24] | - | 63.2 | - | - | - | 73.5 | - | - | - | 74.4 | - | - | - | - | - | - | 74.8 |
| Yi [26] | 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** |

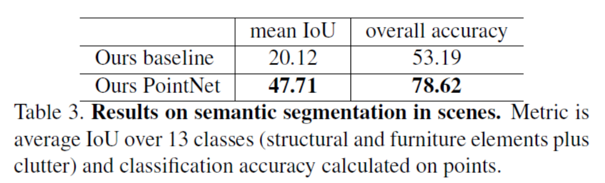
Table 2. **Segmentation results on ShapeNet part dataset.** Metric is mIoU(%) on points. We compare with two traditional methods [24] and [26] and a 3D fully convolutional network baseline proposed by us. Our PointNet method achieved the state-of-the-art in mIoU.

# Applications-Semantic Segmentation in Scenes

- Stanford 3D semantic parsing data set
- The dataset contains 3D scans from Matterport scanners in 6 areas including 271 rooms. Each point in the scan is annotated with one of the semantic labels from 13 categories (chair, table, floor, wall etc)

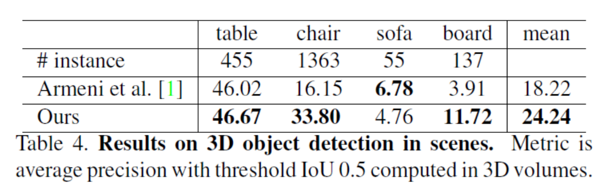


# Applications-Semantic Segmentation in Scenes



| | mean IoU | overall accuracy |
|---|---|---|
| Ours baseline | 20.12 | 53.19 |
| Ours PointNet | **47.71** | **78.62** |

Table 3. **Results on semantic segmentation in scenes.** Metric is average IoU over 13 classes (structural and furniture elements plus clutter) and classification accuracy calculated on points.

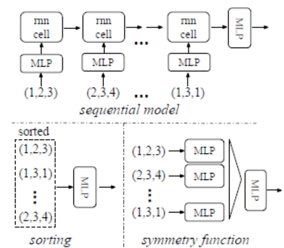


| | table | chair | sofa | board | mean |
|---|---|---|---|---|---|
| # instance | 455 | 1363 | 55 | 137 | |
| Armeni et al. [1] | 46.02 | 16.15 | **6.78** | 3.91 | 18.22 |
| Ours | **46.67** | **33.80** | 4.76 | **11.72** | **24.24** |

Table 4. **Results on 3D object detection in scenes.** Metric is average precision with threshold IoU 0.5 computed in 3D volumes.

# Architecture Design Analysis

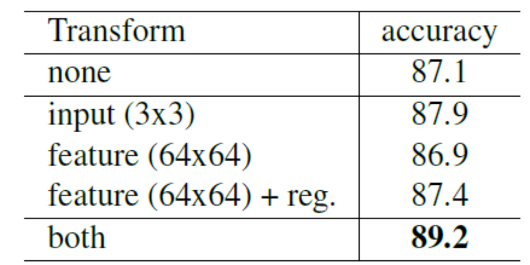
- Three approaches to achieve order invariance.
- ModelNet40 shape classification problem



| | accuracy |
|---|---|
| MLP (unsorted input) | 24.2 |
| MLP (sorted input) | 45.0 |
| LSTM | 78.5 |
| Attention sum | 83.0 |
| Average pooling | 83.8 |
| Max pooling | **87.1** |

# Architecture Design Analysis

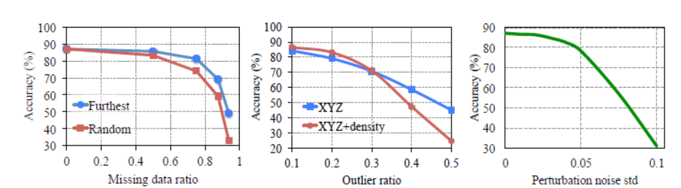
- Effectiveness of Input and Feature Transformations
- ModelNet40 shape classification problem



| Transform | accuracy |
|---|---|
| none | 87.1 |
| input (3x3) | 87.9 |
| feature (64x64) | 86.9 |
| feature (64x64) + reg. | 87.4 |
| both | **89.2** |

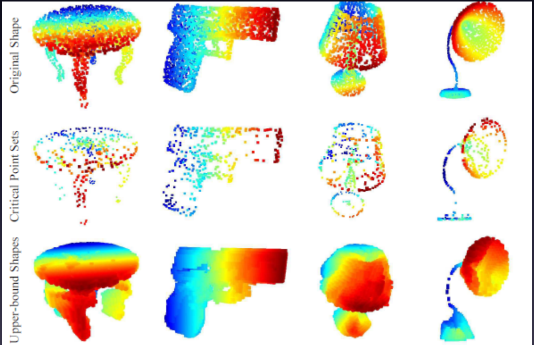
# Architecture Design Analysis

- Robustness Test
- ModelNet40 shape classification problem



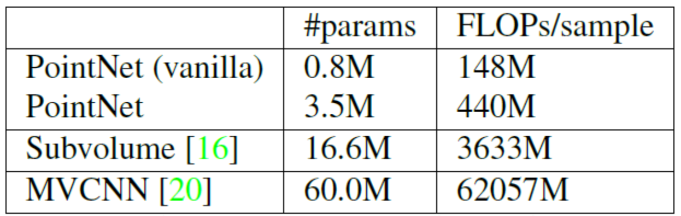
# Visualizing PointNet

- critical point sets $C_S$ & the upper-bound shapes $N_S$



# Time and Space Complexity Analysis

- PointNet’s space and time, complexity is O(N)
- point cloud classification: 1K objects/second
- semantic segmentation: 2 rooms/second
- 1080X GPU on TensorFlow



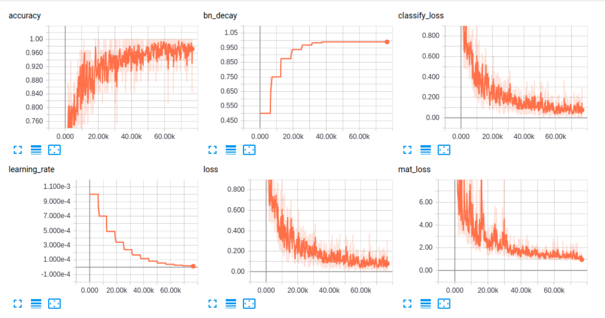
| | #params | FLOPs/sample |
|---|---|---|
| PointNet (vanilla) | 0.8M | 148M |
| PointNet | 3.5M | 440M |
| Subvolume [16] | 16.6M | 3633M |
| MVCNN [20] | 60.0M | 62057M |

# Conclusion

- A brief introduction
- PointNet architecture
- Experiment result

# Repeat the experiment

- Tensorflow
- CPU: i7 - 5700
- GPU: Geforce 1070
- Training time: 2h31min(classification) and 16h28min(part segmentation)

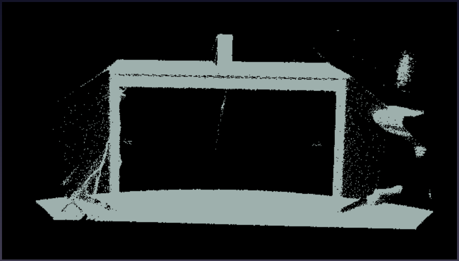


# Heuristics

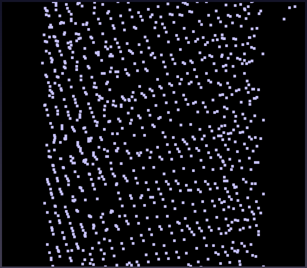
- Use it on robots
- Learn to create a novel structure
- Figure out a research idea

# Traditional Point Cloud Processing

- Edge-based methods
- Model-based methods
- Region-based methods
- Attributes-based methods
- Graph-based methods



(a) A cup on the desk



(b) A part of leg