POINTNET: DEEP LEARNING ON POINT SETS FOR 3D CLASSIFICATION AND SEGMENTATION

AUTHOR: CHARLES R. QI HAO SU KAICHUN MO LEONIDAS J. GUIBAS

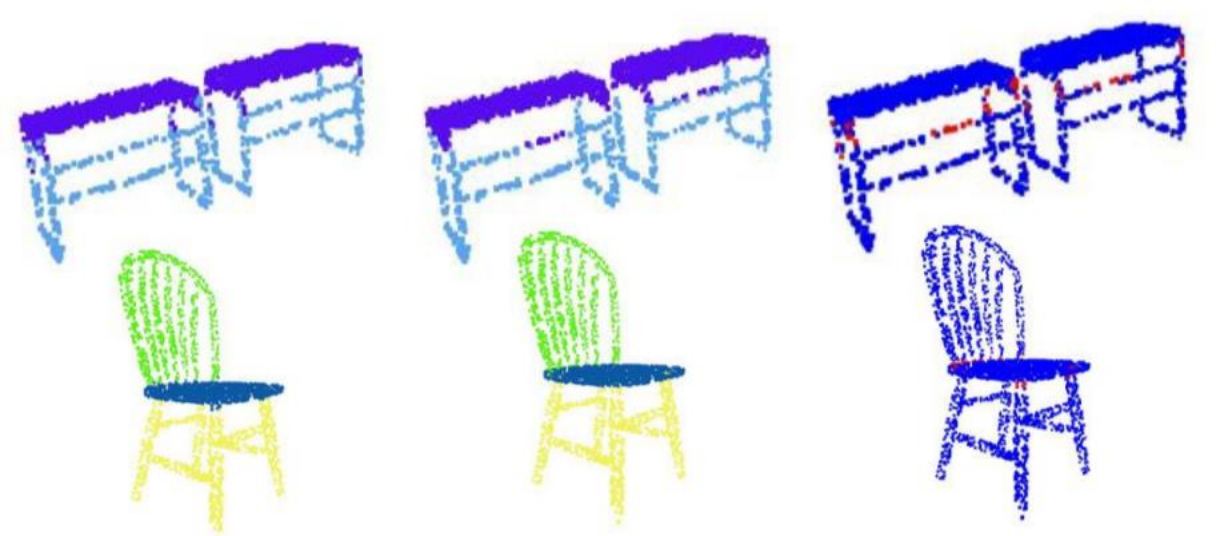
PRESENTER: LIHAO QIU

2 CONTENTS

3 ABSTRACT

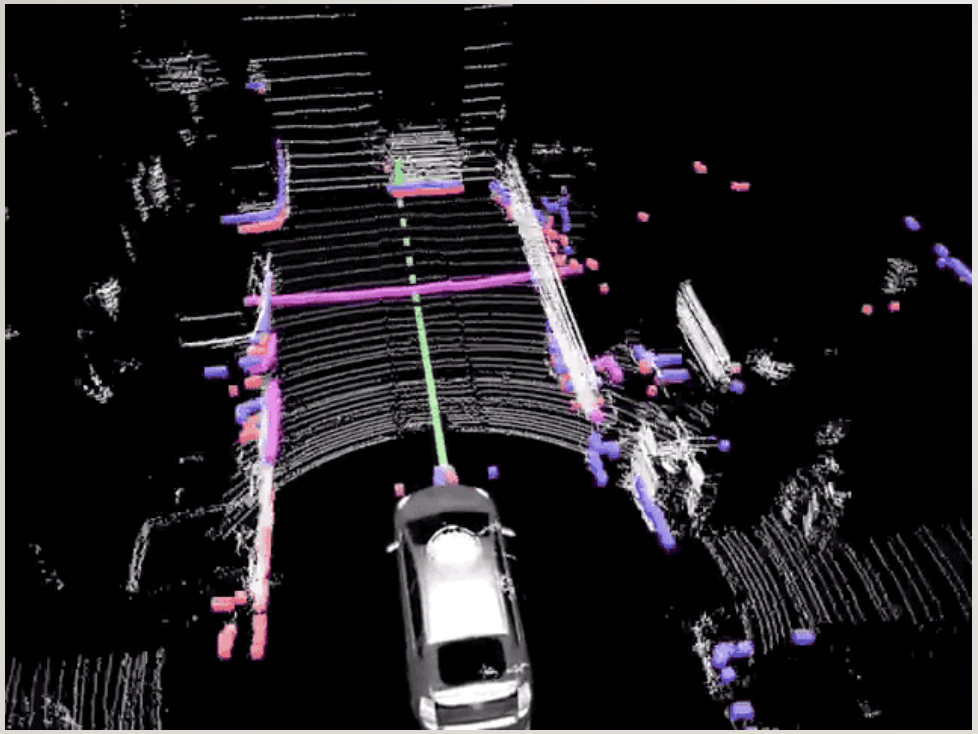
- Point cloud is a type of geometric data structure. Due to its irregular format, most researchers transform such data to regular 3D voxel grids or collections or images. However, this renders unnecessary volumes and causes issues. In this paper, a novel type of neural network(PointNet) is designed, which directly consumes point clouds, which well respects the permutation invariance of points in the input. This network provides a unified architecture for applications ranging from object classification, part segmentation, to scene semantic parsing.

4 INTRODUCTION OF POINT CLOUD



Object detection, classification, part segmentation

5 INTRODUCTION OF POINT CLOUD



Autonomous driving

6 INTRODUCTION OF POINT CLOUD

- Three main properties:

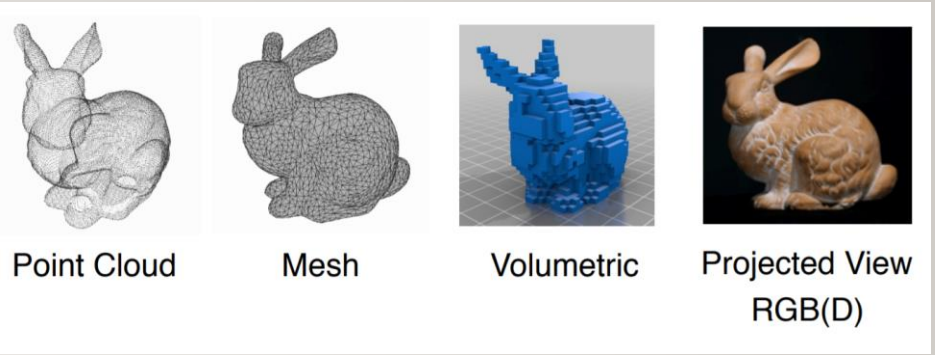
1. Unordered. Unlike pixel arrays in images or voxel arrays in volumetric grids, point cloud is a set of points without specific order. A network that consumes N 3D point sets needs to be invariant to N factorial permutations of the input set in data feeding order.
2. Interaction among points. Points are from a space with a distance metric, which means that points are not isolated, and neighboring points form a meaningful subset. Therefore, the model must be able to capture local structures from nearby points, and the combinatorial interactions among local structures.
3. Invariance under transformation. As a geometric object, the learned representation of the point set should be invariant to certain transformations. For example, rotating and translating points all together should not modify the global point cloud category nor the segmentation of the points.

7 RELATED WORK

- Point Cloud Features

Most existing features for point cloud are handcrafted towards specific tasks. Features are typically intrinsic or extrinsic. Hence cannot be generalized.

8 RELATED WORK



- Deep Learning on 3D Data

1. Applying 3D CNN to voxelized shapes. (Specific data formats, in order to perform weight sharing and other kernel optimizations). This leads to low resolution and high computation cost. (3D convolution)
2. Render 3D point cloud into 2D images and apply 2D conv nets to classify them. This method cannot be extended to scene understanding or other tasks such as point classification and shape completion.
3. Spectral CNNs on meshes. This method is constrained on manifold meshes.

9 RELATED WORK

- Deep Learning on Unordered Sets

From a data structure point of view, a point cloud is an unordered set of vectors. Most works in deep learning focus on regular input representations like sequences, images and volumes. Not much work has been done in deep learning on point sets.

PROBLEM STATEMENT

- Design a deep learning framework that can directly consumes unordered point sets(raw) as inputs. Each point is represented by its coordinates: $P_i = (x_i, y_i, z_i)$. This network should be able to perform object classification, part segmentation and semantic segmentation.

Object classification: output class labels for the entire input

Part/Semantic segmentation: output per point part/segment labels for each point of the input

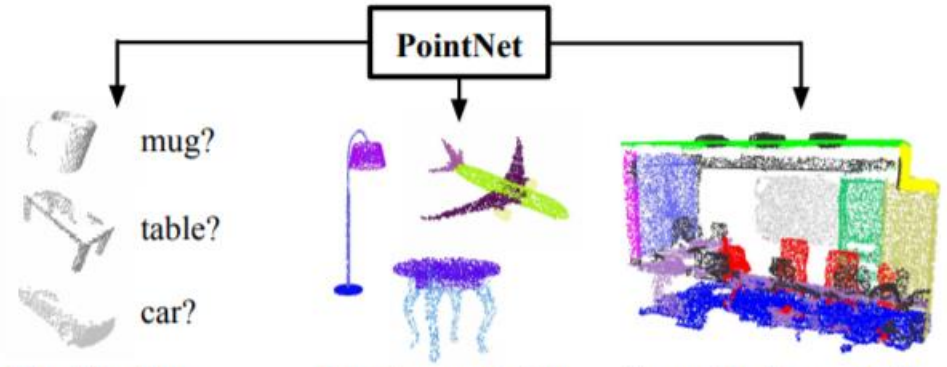


Figure 1. **Applications of PointNet.** We propose a novel deep net architecture that consumes raw point cloud (set of points) without voxelization or rendering. It is a unified architecture that learns both global and local point features, providing a simple, efficient and effective approach for a number of 3D recognition tasks.

POINT CLOUD REVIEW

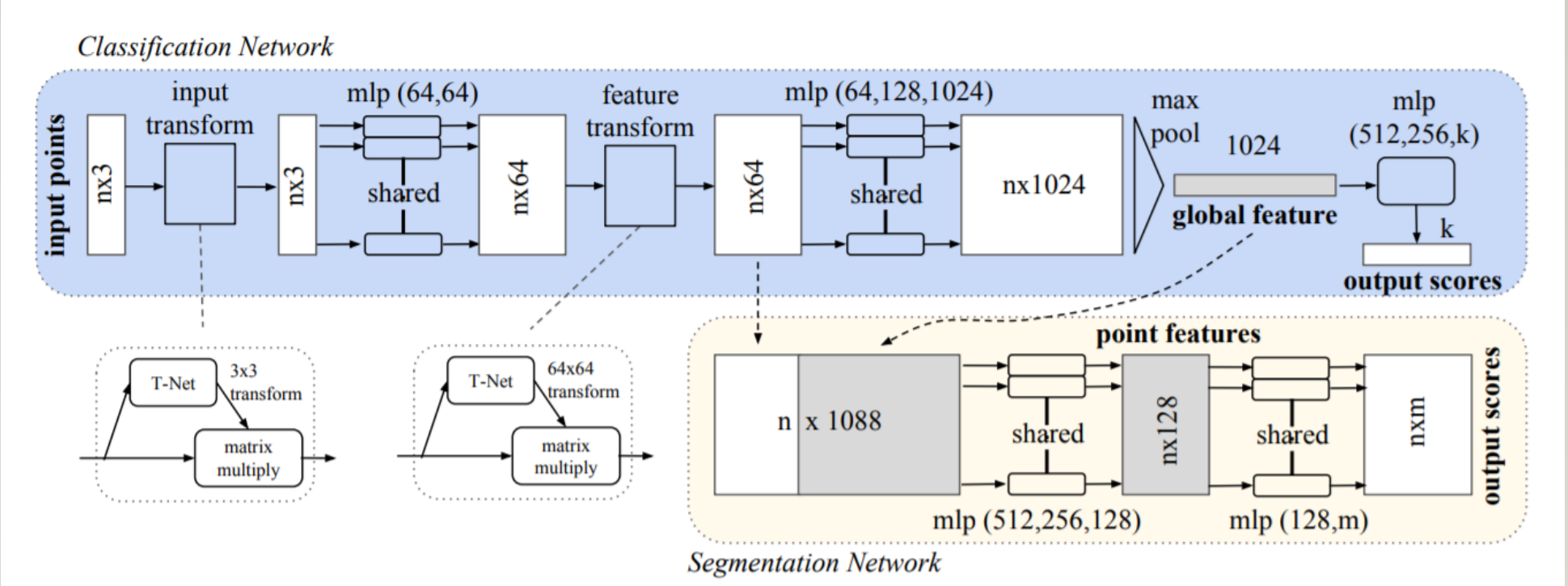
- Three properties

1. Unordered,
2. Interaction among points
3. Invariance under transformations.

12 METHOD COMPARISON

1. Sort input into a canonical order: in high dimension space there in fact does not exist an ordering that is stable w.r.t. point perturbations in the general sense.
2. Treat the input as a sequence to train an RNN, but augment the training data by all kinds of permutations: hoping that by training the RNN with randomly permuted sequences, the RNN will become invariant to input order. However, the order does matter and cannot be totally omitted.
3. Use a simple symmetric function to aggregate the information from each point

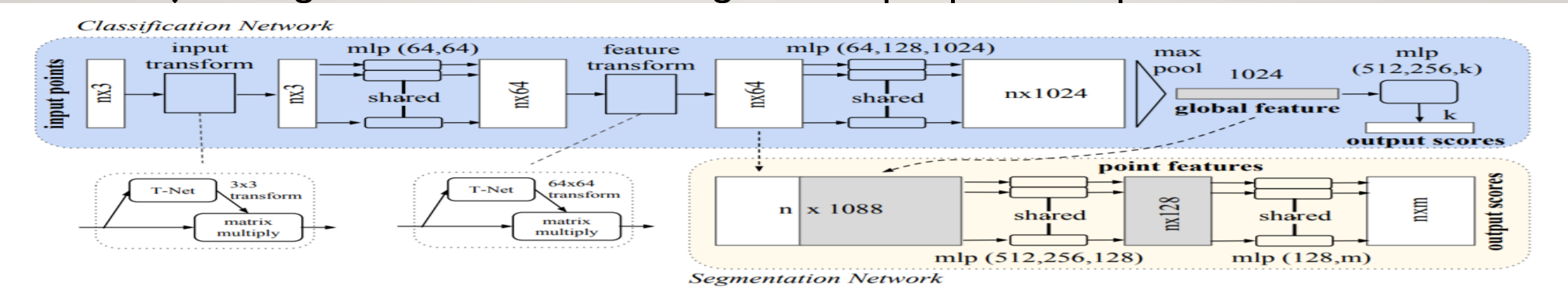
13 INTRODUCTION OF POINTNET



14 INTRODUCTION OF POINTNET

- Three key modules:

1. A max pooling layer as a symmetric function to aggregate information from all points
2. A local and global information combination structure
3. Two joint alignment networks that align both input parts and point features



15 INTRODUCTION OF POINTNET

- Symmetric function for unordered input(max pooling)

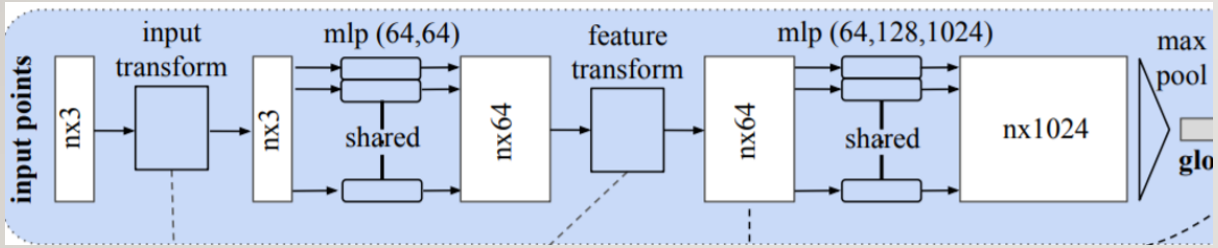
The idea is to approximate a general function defined on a point set by applying a symmetric function on transformed elements in the set:

$$f(\{x_1, \ldots, x_n\}) \approx g(h(x_1), \ldots, h(x_n)), \qquad (1)$$

where $f : 2^{\mathbb{R}^N} \rightarrow \mathbb{R}$, $h : \mathbb{R}^N \rightarrow \mathbb{R}^K$ and $g : \underbrace{\mathbb{R}^K \times \cdots \times \mathbb{R}^K}_{n} \rightarrow \mathbb{R}$ is a symmetric function.

16 INTRODUCTION OF POINTNET

- h is approximated by a multi-layer perceptron network and g is approximated by a composition of a single variable function and a max pooling function. Through a collection of h, we can learn a number of f's to capture different properties of the set.



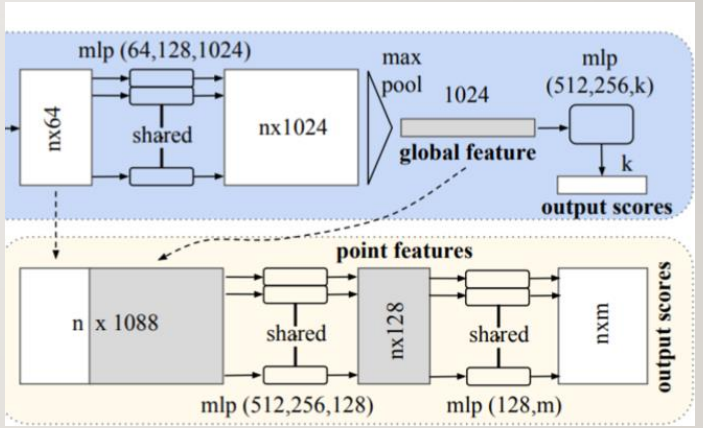
$$f(\{x_1, \ldots, x_n\}) \approx g(h(x_1), \ldots, h(x_n)), \qquad (1)$$

where $f : 2^{\mathbb{R}^N} \to \mathbb{R}$, $h : \mathbb{R}^N \to \mathbb{R}^K$ and $g : \underbrace{\mathbb{R}^K \times \cdots \times \mathbb{R}^K}_{n} \to \mathbb{R}$ is a symmetric function.

17 INTRODUCTION OF POINTNET

- Local and global information aggregation

Point segmentation requires a combination of local and global knowledge. The solution is to feed back the global point cloud feature vector to per point features by concatenating the global feature with each of the point features. With this modification the network is able to predict per point quantities that rely on both local geometry and global semantics.



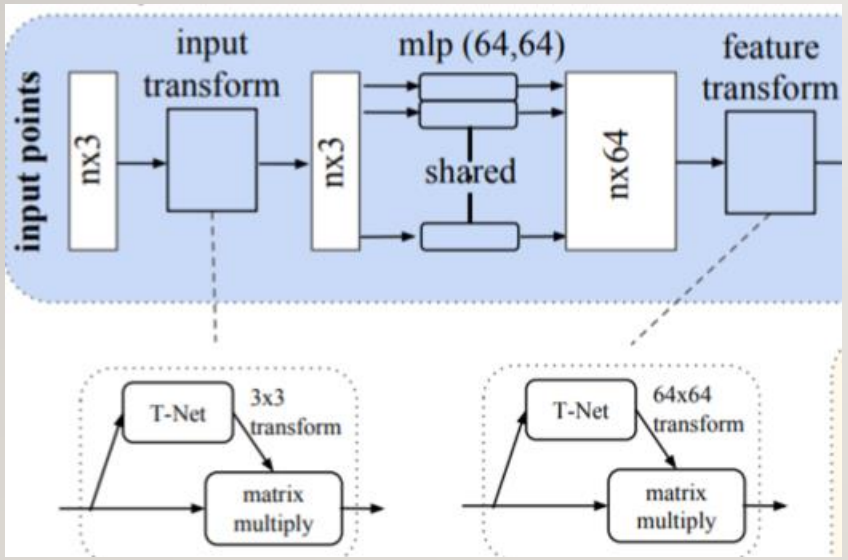
INTRODUCTION OF POINTNET

- Joint alignment network

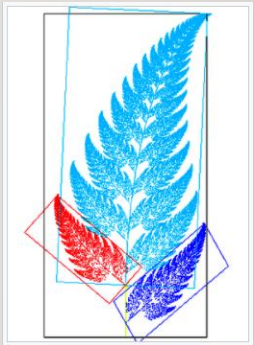
The semantic labeling of a point cloud has to be invariant if the point cloud undergoes certain geometric transformations. A natural solution is to align all input set to a canonical space before feature extraction. An affine transformation matrix by a mini-network(T-net) is directly applied to input. The T-net itself resembles the big network and is composed by basic modules of point independent feature extraction, max pooling and fully connected layers. This idea can be further extended to the alignment of feature space. A second T-net is inserted on point features and predict a feature transformation matrix to align features from different input point clouds.

19 INTRODUCTION OF POINTNET

Joint alignment networks



Affine transformation



An image of a fern-like fractal (Barnsley's fern) that exhibits affine self-similarity. Each of the leaves of the fern is related to each other leaf by an affine transformation. For instance, the red leaf can be transformed into both the dark blue leaf and any of the light blue leaves by a combination of reflection, rotation, scaling, and translation.

THEORETICAL ANALYSIS OF POINTNET ★

- Suppose we have a continuous set function: $X = \{S: S \in [0,1]^m \; and \; |S| = n\}$, $f : X \to R$ is a continuous set function on X w.r.t. Hausdorff distance $d_H(.,.)$: for all $\varepsilon > 0$, exists $\delta > 0$, for any $S, S' \in X$, if $d_H(S, S') < \delta$, then $|f(S) - f(S')| < \varepsilon$

This indicates f can be arbitrarily approximated by PointNet given enough neurons at the max pooling layerR^K x R^K x R^K ... R^K. The more the neurons are, the larger the K is.

THEORETICAL ANALYSIS OF POINTNET

- Hausdorff distance: the longest distance you can be forced to travel by an adversary who chooses a point in one of the two sets, from where you then must travel to the other set. It is a criteria which measures the similarity between two point sets.

Given point set A and B, if $d_H(\mathrm{A}, B) = 0$, then point set A and B are the same.

THEORETICAL ANALYSIS OF POINTNET

- Theorem 1. Suppose $f : X \to R$ is a continuous set function w.r.t. Hausdorff distance $d_H(.,.)$, for all $\varepsilon > 0$, exists a continuous function h and a symmetric function $g(x_1, \ldots x_n) = \gamma o MAX$, such that for any $S \in X$,

$$|f(S) - \gamma\,(\underset{x_i \in S}{\mathrm{MAX}}\{h(x_i)\})| < \varepsilon$$

where xi is the full list of elements in S ordered arbitrarily(input point cloud), γ is a continuous function(structure before max pooling layer), and MAX is a vector max operator that takes n vectors as input and returns a new vector of the element-wise maximum(max pooling layer).

Theorem 1 is basically saying that PointNet is able to interpret raw point cloud input.

THEORETICAL ANALYSIS OF POINTNET

- Bottleneck dimension and stability

Theoretically and experimentally the authors found that the expressiveness of PointNet is strongly affected by the dimension of the max pooling layer, i.e., K.

We define $\mathrm{u} = (\underset{x_i \in S}{\mathrm{MAX}}\{h(x_i)\}$ to be the sub-network of f which maps a point set in $[0,1]^m$ to a K-dimensional vector.

24 THEORETICAL ANALYSIS OF POINTNET

- Theorem 2. Suppose u: $X \to R^K$ such that u = $(\underset{x_i \in S}{\mathrm{MAX}}\{h(x_i)\}$ and $f = \gamma\ o\ u$, Then,

(a) For all S, exists C_S, $N_S \in X, f(T) = f(S)$ if $C_S \in T \in N_S$;

(b) $|C_S| \leq K$

Theorem2(a) says that $f(S)$ is unchanged to the input corruption if all points in C_S are preserved; it is also unchanged with extra noise points up to N_S

Theorem2(b) says that C_S only contains a bounded number of points, determined by K(number of neurons in max pooling layer)

THEORETICAL ANALYSIS OF POINTNET

- C_S : critical point set of S,
- K: bottleneck dimension of f

Combined with the continuity of h, this explains the robustness of PointNet w.r.t point perturbation, corruption and extra noise points. The robustness is gained in analogy to the sparsity principle in machine learning models. Intuitively, PointNet learns to summarize a shape by a sparse set of key points.

26 EXPERIMENTAL RESULTS OF POINTNET

- 3D Object Classification: evaluate PointNet on ModelNet40 shape classification benchmark. While previous methods focus on volumetric and multi-view image representations, PointNet is the first one to directly take in raw point cloud.

	input	#views	accuracy avg. class	accuracy overall
SPH [11]	mesh	-	68.2	-
3DShapeNets [28]	volume	1	77.3	84.7
VoxNet [17]	volume	12	83.0	85.9
Subvolume [18]	volume	20	86.0	**89.2**
LFD [28]	image	10	75.5	-
MVCNN [23]	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.

27 EXPERIMENTAL RESULTS OF POINTNET

- 3D object part segmentation: given a 3D scan or a mesh model, the task is to assign part category label (e.g. chair leg, cup handle) to each point or face. The data set selected to evaluate this is ShapeNet part data set. The author formulate part segmentation as a per-point classification problem. Evaluation metric is mIoU on points.

	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 [27]	-	63.2	-	-	-	73.5	-	-	-	74.4	-	-	-	-	-	-	74.8
Yi [29]	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 [27] and [29] and a 3D fully convolutional network baseline proposed by us. Our PointNet method achieved the state-of-the-art in mIoU.

28 EXPERIMENTAL RESULTS OF POINTNET

- Semantic segmentation in scenes: extend part segmentation to semantic scene segmentation, where point labels become semantic object classes instead of object part labels. The authors experiment on Stanford 3D semantic parsing data set. Each point in the scan is annotated with one of the semantic labels.

	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.

EXPERIMENTAL RESULTS OF POINTNET

- 3D object detection system: built based on semantic segmentation output. Comparing with state-of-the-art method, which is based on a sliding shape method (with conditional random fields post processing) with SVMs trained on local geometric features and global room context feature in voxel grids.

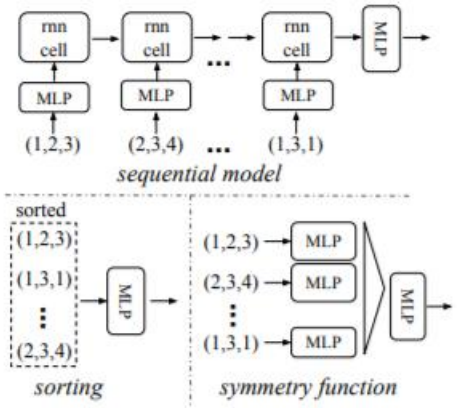
	table	chair	sofa	board	mean
# instance	455	1363	55	137	
Armeni et al. [1]	46.02	16.15	**6.78**	3.91	18.22
Ours	**46.67**	**33.80**	4.76	**11.72**	**24.24**

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

31 ARCHITECTURE DESIGN ANALYSIS

- Comparison with alternative order-invariant methods

Comparison among: MLP on unsorted and sorted points, RNN model that considers input point as a sequence and a model based on symmetry functions. The symmetry operation experimented include max pooling, average pooling and an attention based weighted sum.



	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**

Figure 5. **Three approaches to achieve order invariance.** Multi-layer perceptron (MLP) applied on points consists of 5 hidden layers with neuron sizes 64,64,64,128,1024, all points share a single copy of MLP. The MLP close to the output consists of two layers with sizes 512,256.

32 ARCHITECTURE DESIGN ANALYSIS

- Effectiveness of input and feature transformation(joint alignment network)

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

Table 5. **Effects of input feature transforms.** Metric is overall classification accuracy on ModelNet40 test set.

33 ARCHITECTURE DESIGN ANALYSIS

- Robustness test: to show that PointNet is robust to various kinds of input corruptions. (Theorem 2)

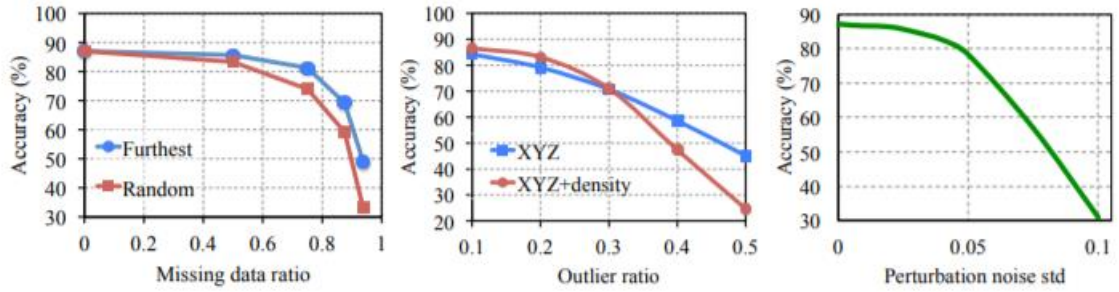


Figure 6. **PointNet robustness test.** The metric is overall classification accuracy on ModelNet40 test set. Left: Delete points. Furthest means the original 1024 points are sampled with furthest sampling. Middle: Insertion. Outliers uniformly scattered in the unit sphere. Right: Perturbation. Add Gaussian noise to each point independently.

34 VISUALIZING POINTNET

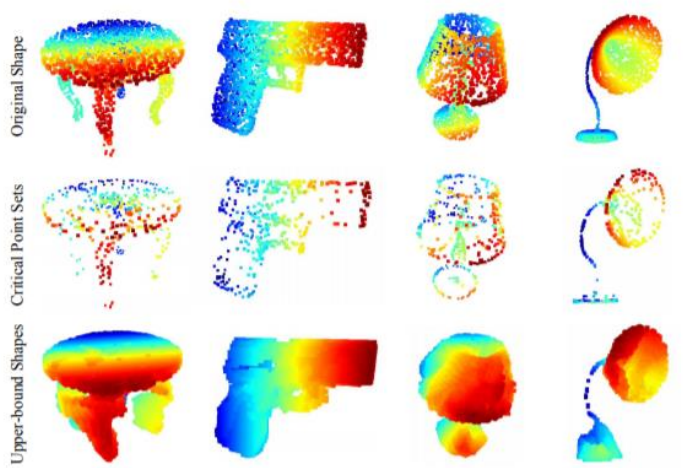


Figure 7. **Critical points and upper bound shape.** While critical points jointly determine the global shape feature for a given shape, any point cloud that falls between the critical points set and the upper bound shape gives exactly the same feature. We color-code all figures to show the depth information.

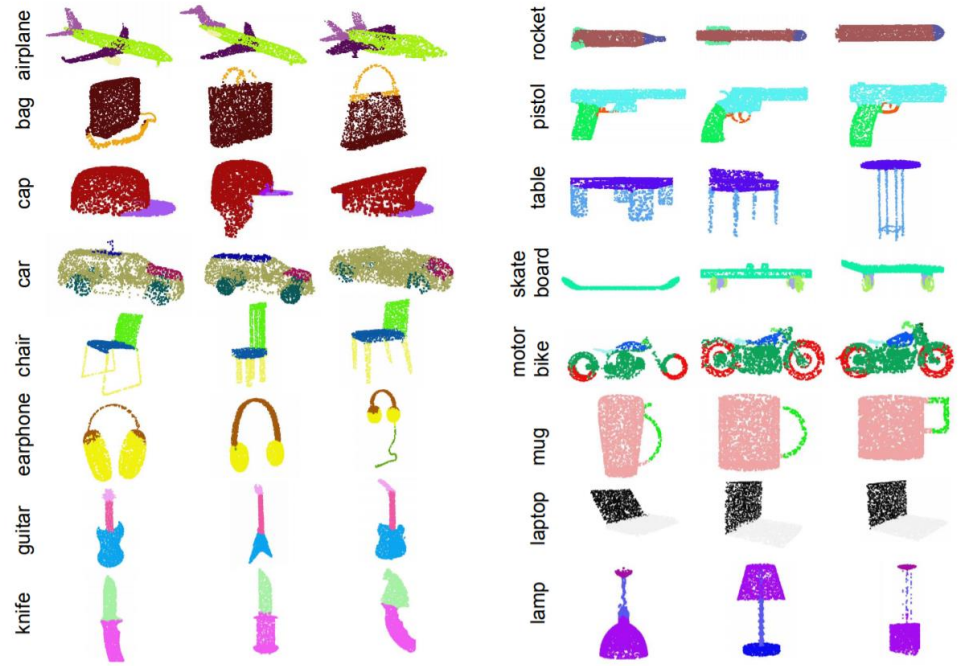


Figure 21. **PointNet segmentation results on complete CAD models.**

CONCLUSION

- PointNet is a novel deep neural network that directly consumes point cloud. It provides a unified approach to a number of 3D recognition tasks including object classification, part segmentation and semantic segmentation, while obtaining on par or better results than state-of-the-art on standard benchmarks.
- Cons: Semantic segmentation is not that accurate, with only 47% mIoU, where the ones nowadays can achieve over 80%.

36 THANK YOU

- Question?