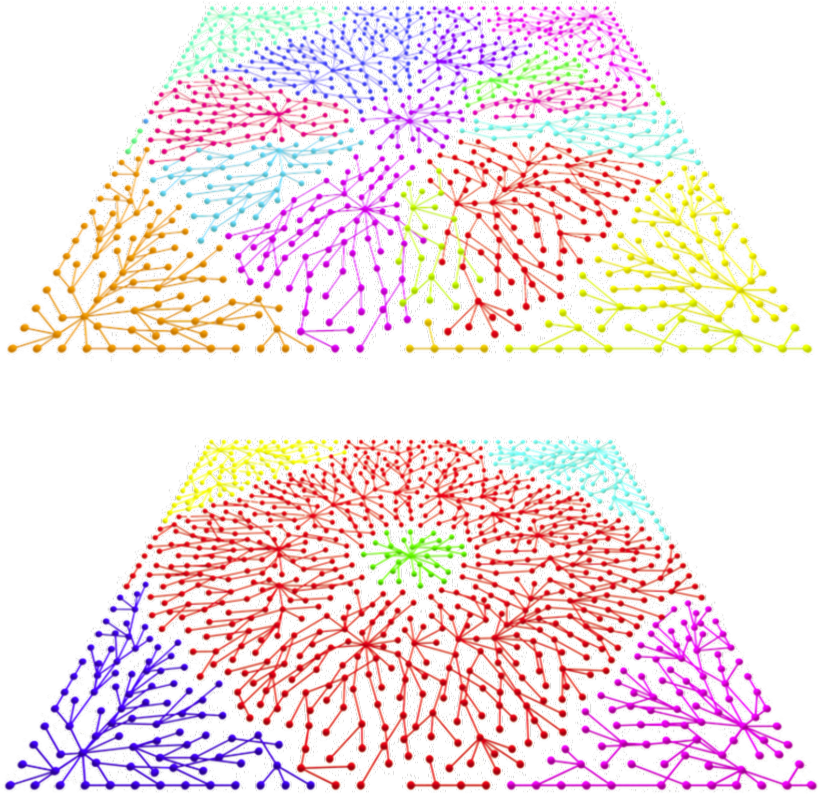
CS233, CME251: Geometric and Topological Data Analysis

Leonidas Guibas
Computer Science Department
Stanford University



Lecture 14
27 May 2020

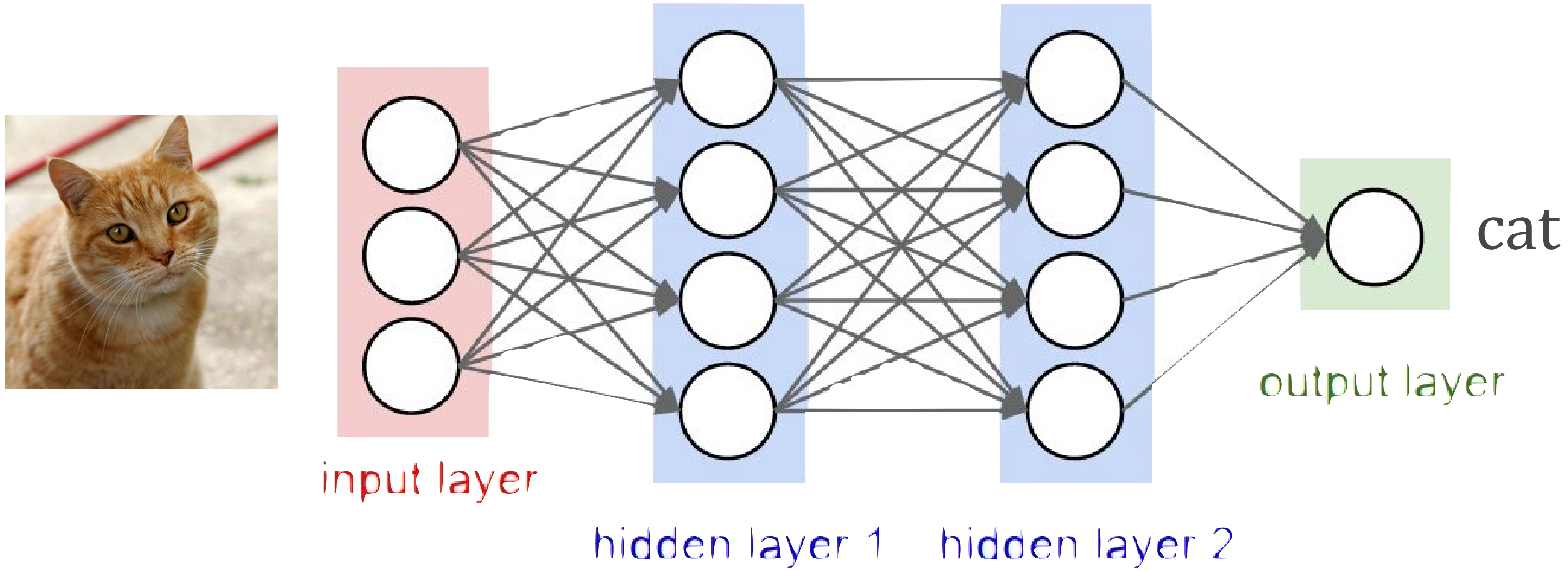
Leonidas Guibas Laboratory
Geometric Computing

Last Time: Deep Nets, Multi-View and Volumetric Approaches to 3D

Deep Learning

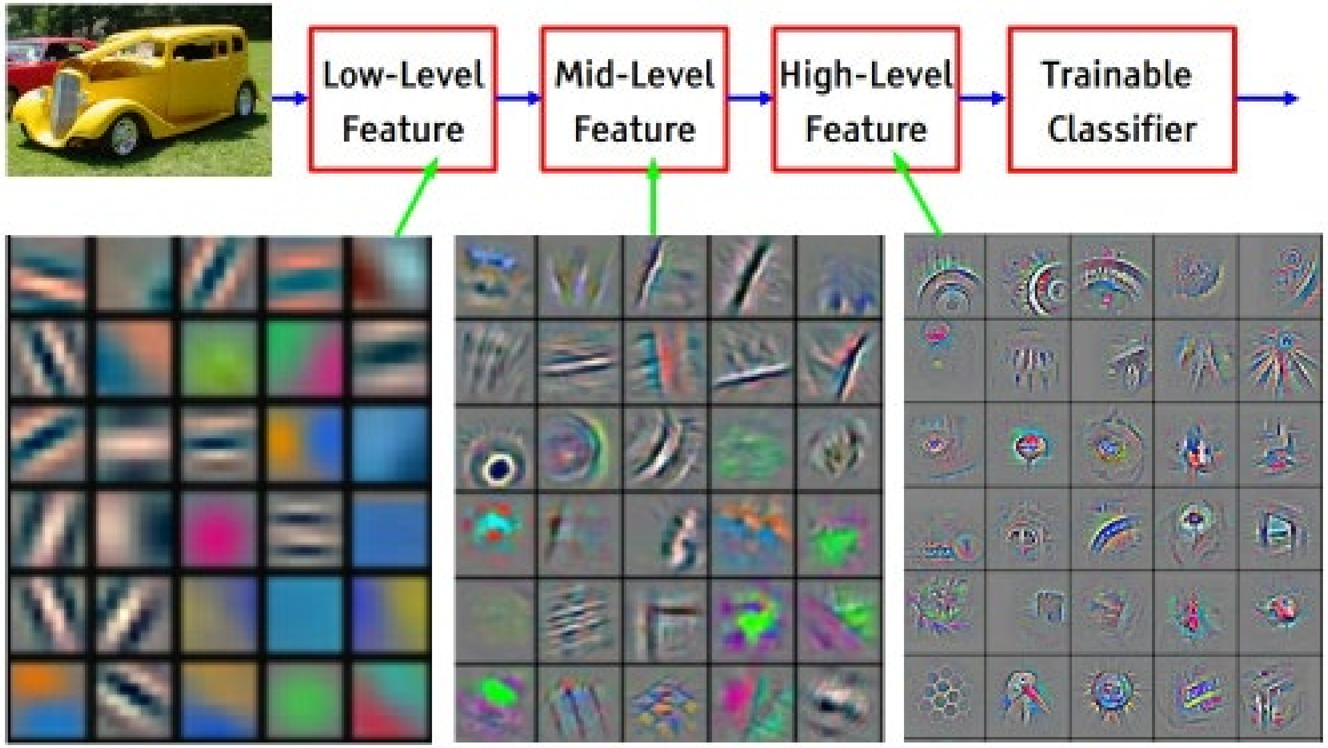
- Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction.

Deep Learning by Y. LeCun et al. Nature 2015



Neural Networks as Feature Extractors

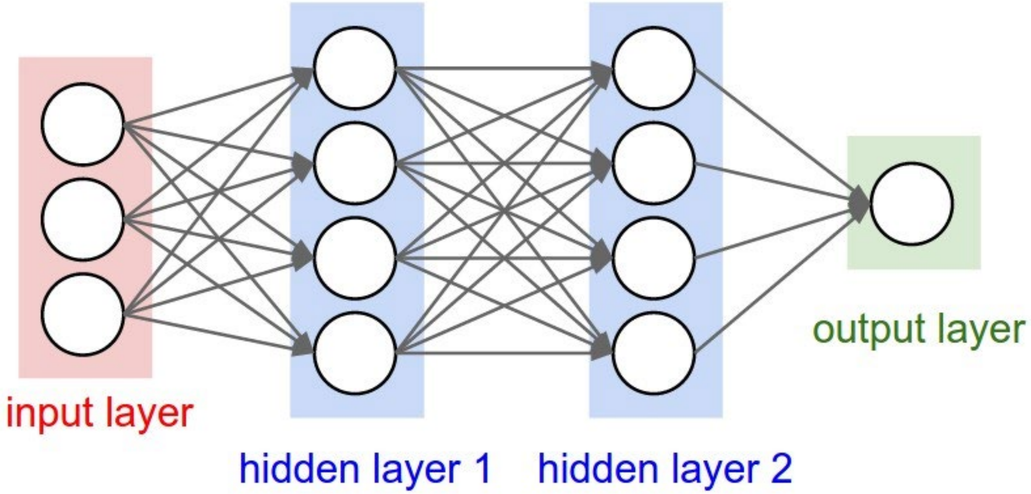
- Neural networks extract powerful features from data



Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

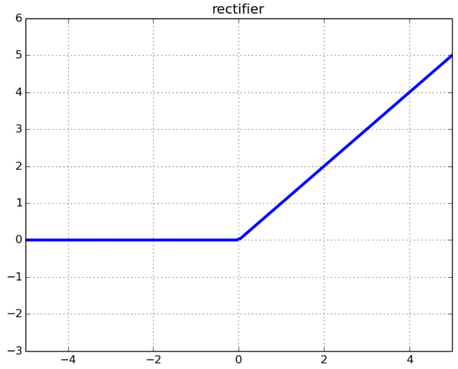
Image Credits: Yan LeCun

Neural Networks Non-Linearities



f: non-linear activation function

$$\mathrm{ReLU}(x) = \begin{cases} 0 \text{ if } x < 0 \\ x \text{ if } x >= 0 \end{cases}$$

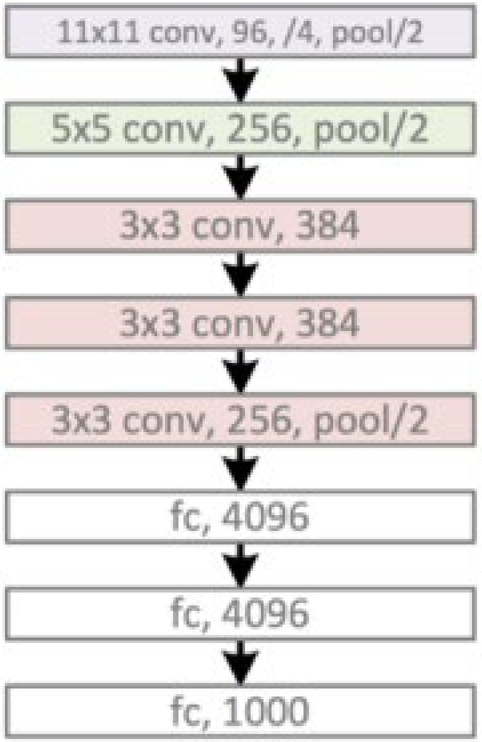


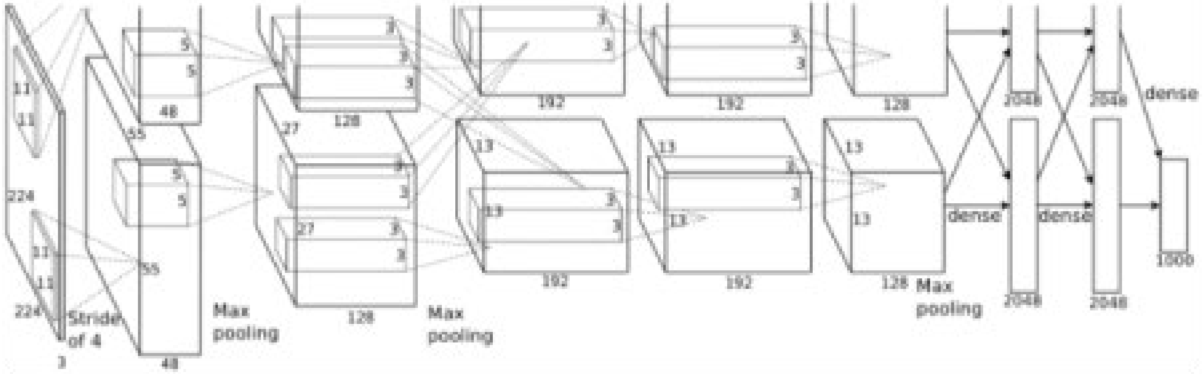
Model: Multi-Layer Perceptron (MLP)

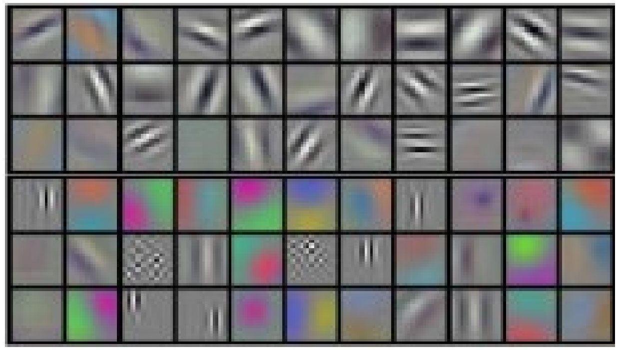
$$y' = W_3 f(W_2 f(W_1 x + b_1) + b_2) + b_3)$$

Convolutional Neural Networks

AlexNet







The first work that popularized Convolutional Networks in Computer Vision

Convolutional Neural Networks

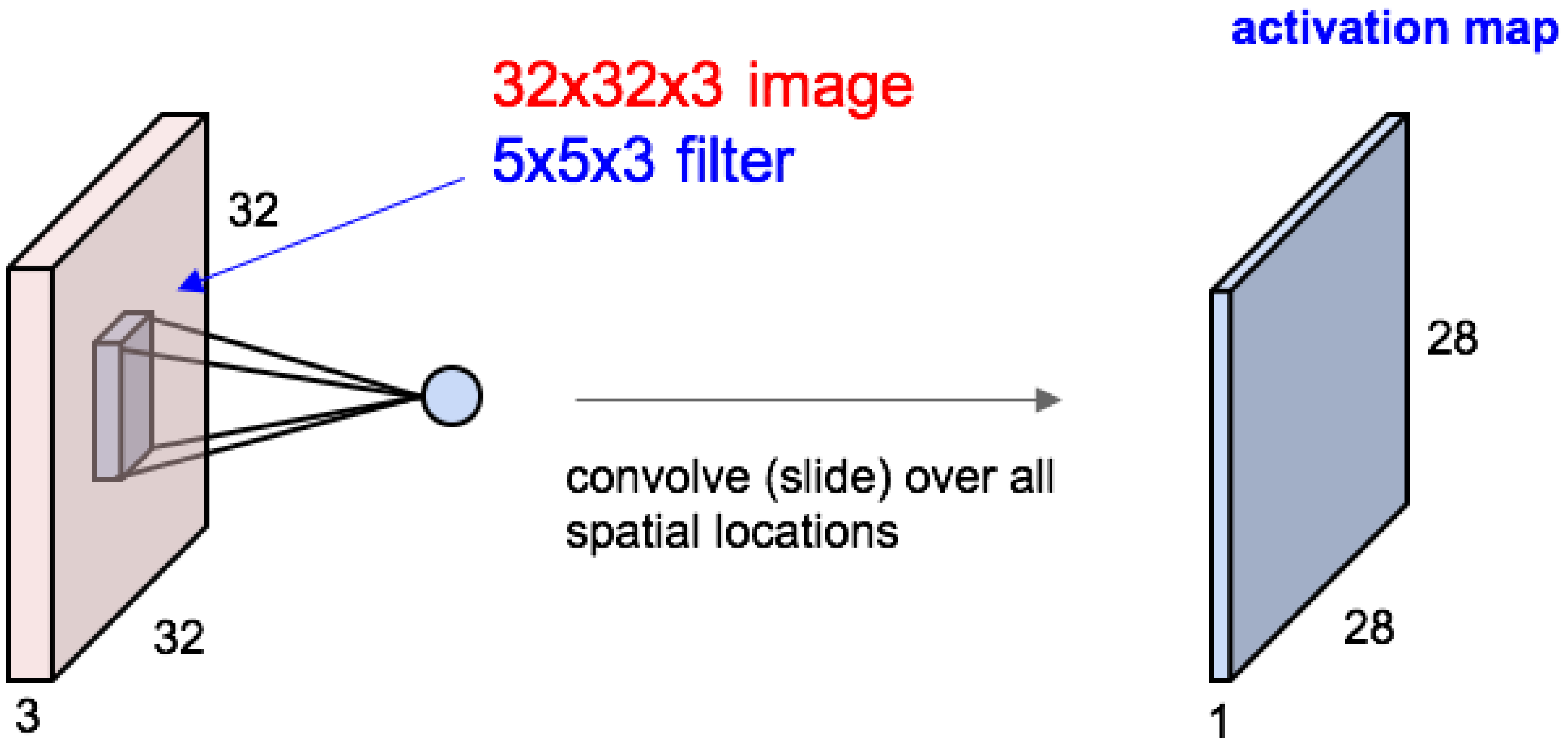


Image Credits: Andrej Karpathy

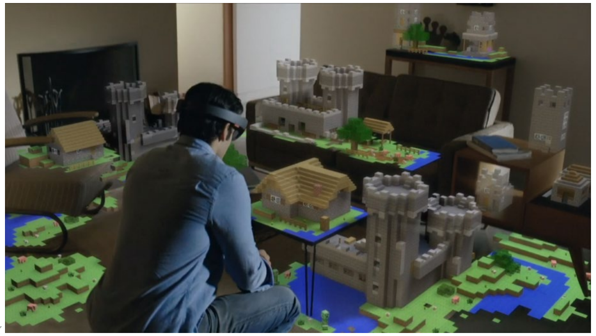
Convolutional Neural Networks

- Filters are doing pattern matching

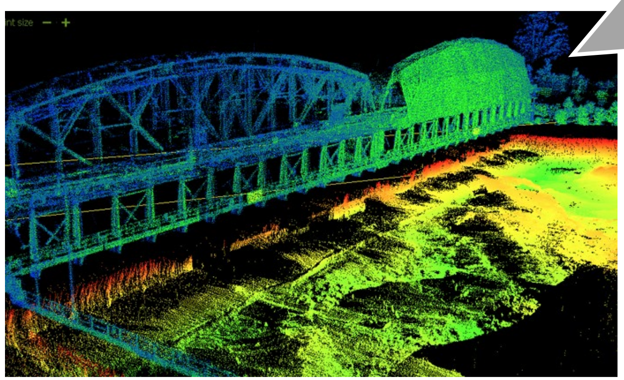
Image Credits: Yan LeCun

3D Applications

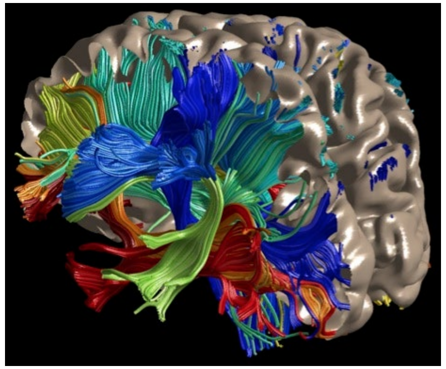
Robotics



Augmented Reality



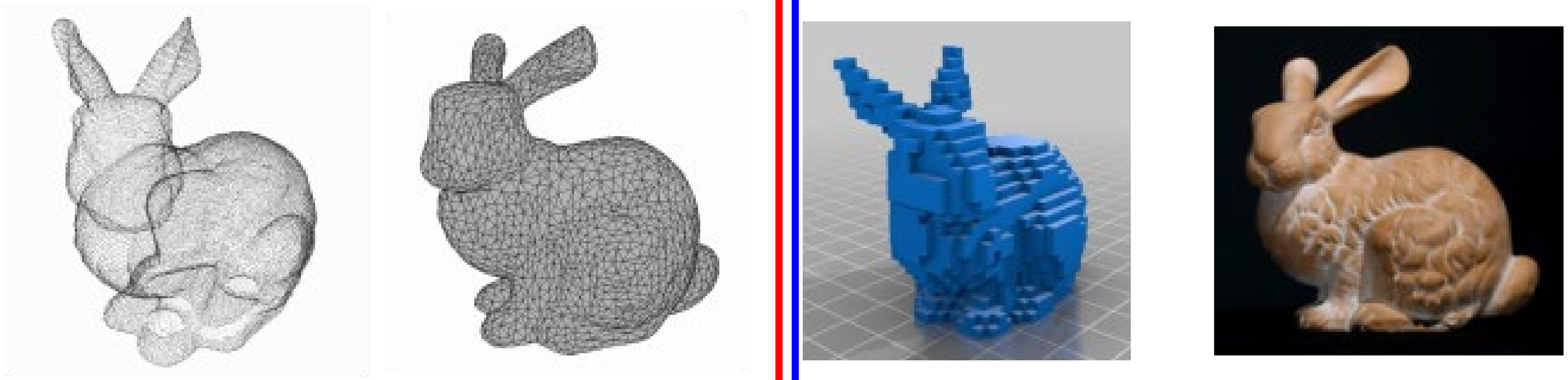
Autonomous driving



Medical Image Processing

3D Representations

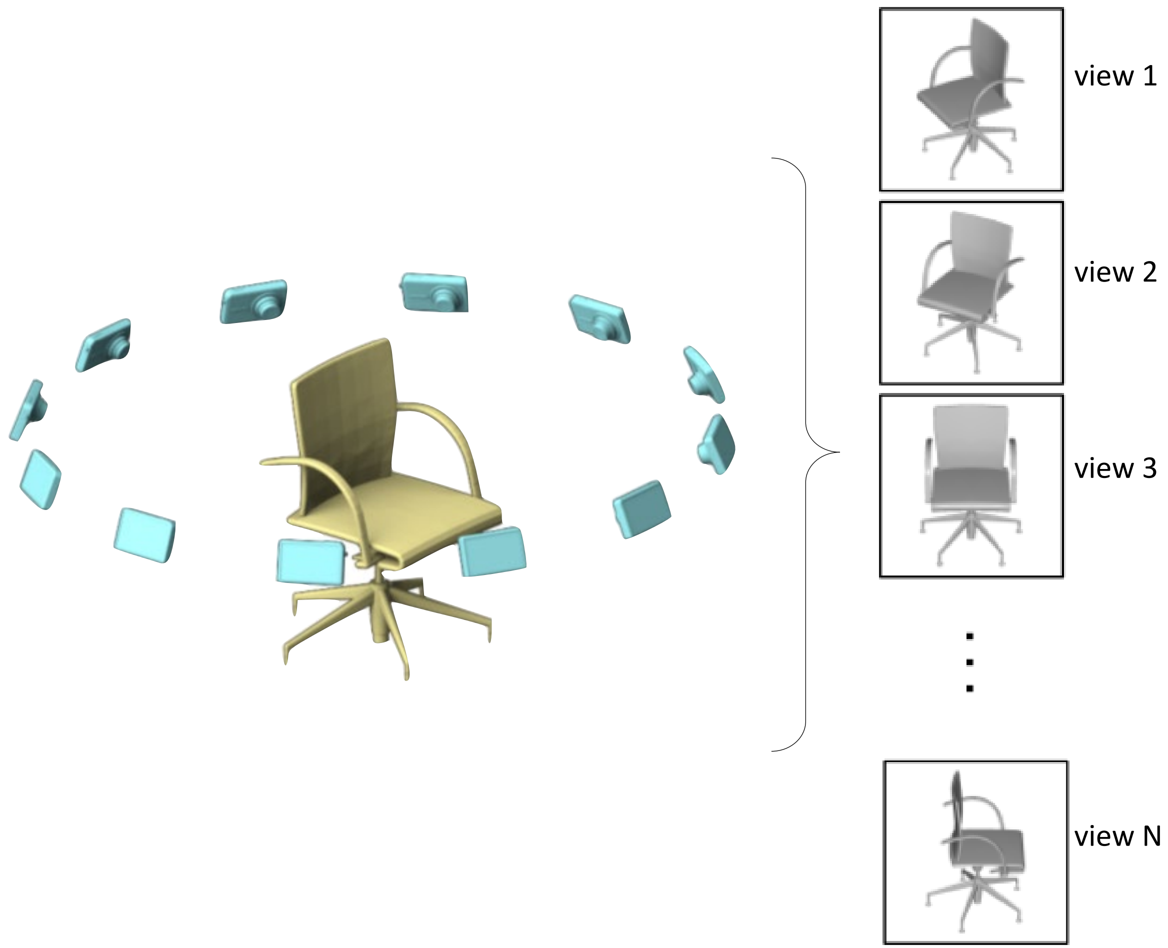
- Design good 3D representation for NN to consume



Point Cloud | Mesh | Volumetric | Image

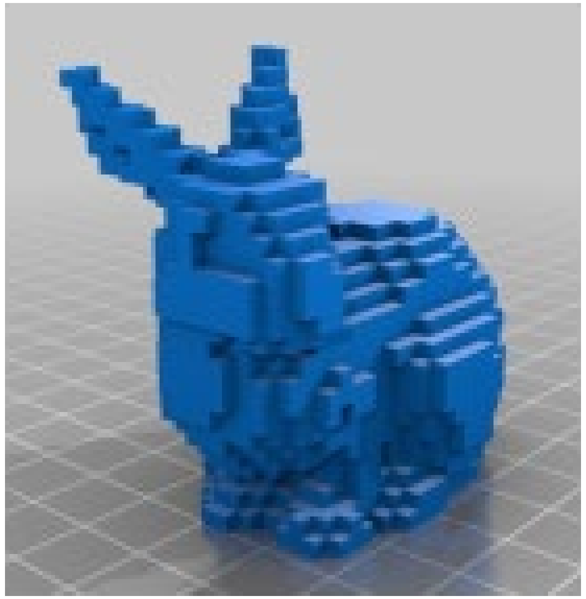
Irregular | Regular

Multi-view Representation



\+ Powerful 2D CNNs

Volumetric Representation



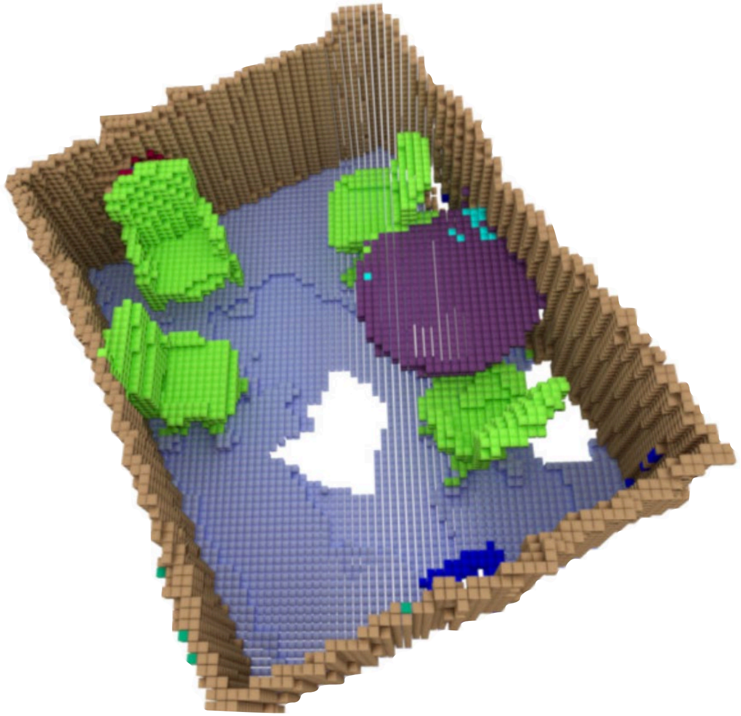
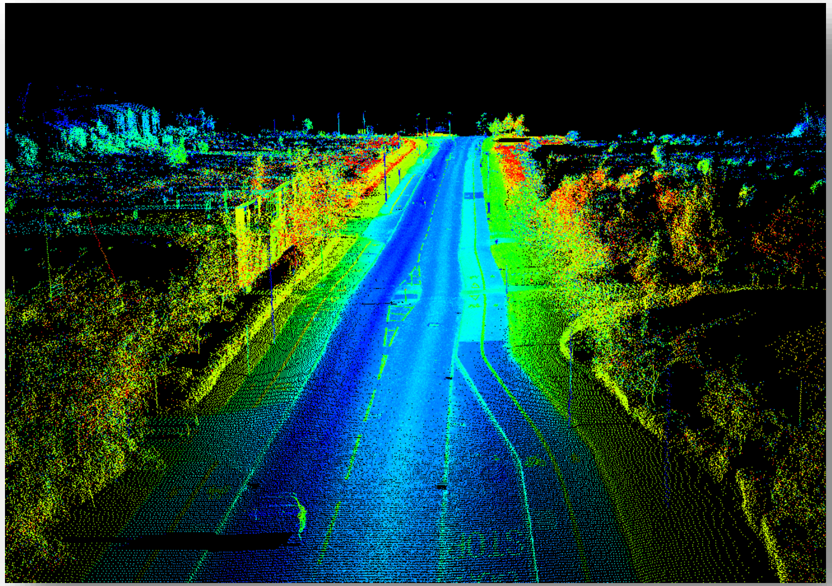


Image Credits: Scannet

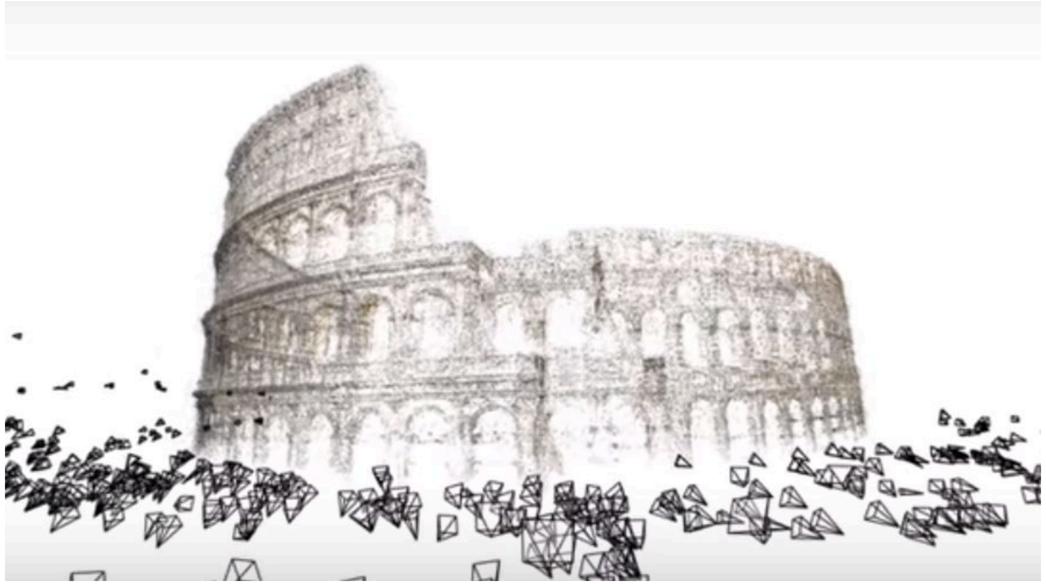
Today: Deep Learning on Point Cloud Data

Non-regular 3D data

Point Clouds are Commonplace



Lidar point clouds (LizardTech)



Structure from motion (Microsoft)

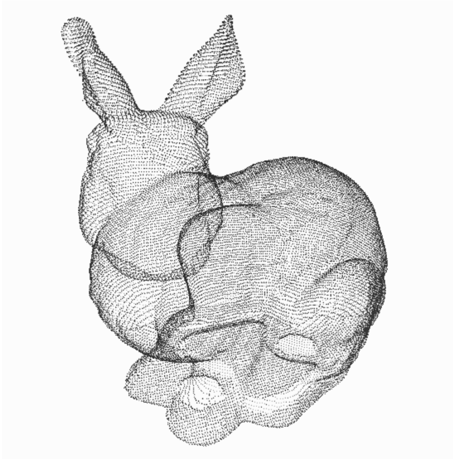
Depth camera (Intel)

A Common 3D Representation: Point Cloud

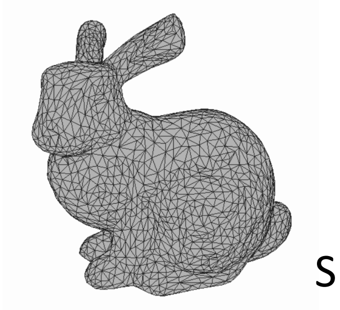
- **Point clouds are close to raw sensor data**
- **Point clouds are representationally simple**

LiDAR

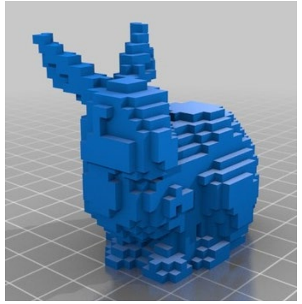
Depth Sensor



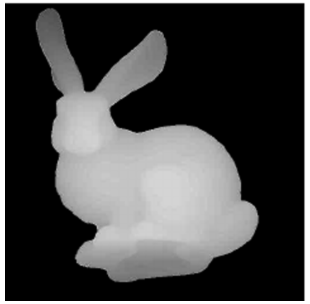
Point Cloud



Surface Mesh

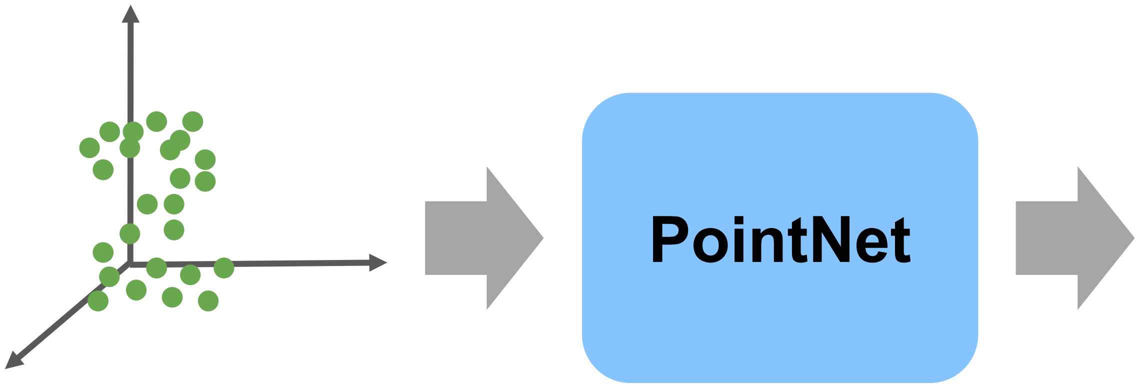


Volumetric



Depth Map

Deep Nets for PCs: PointNet and PointNet++



Object Classification

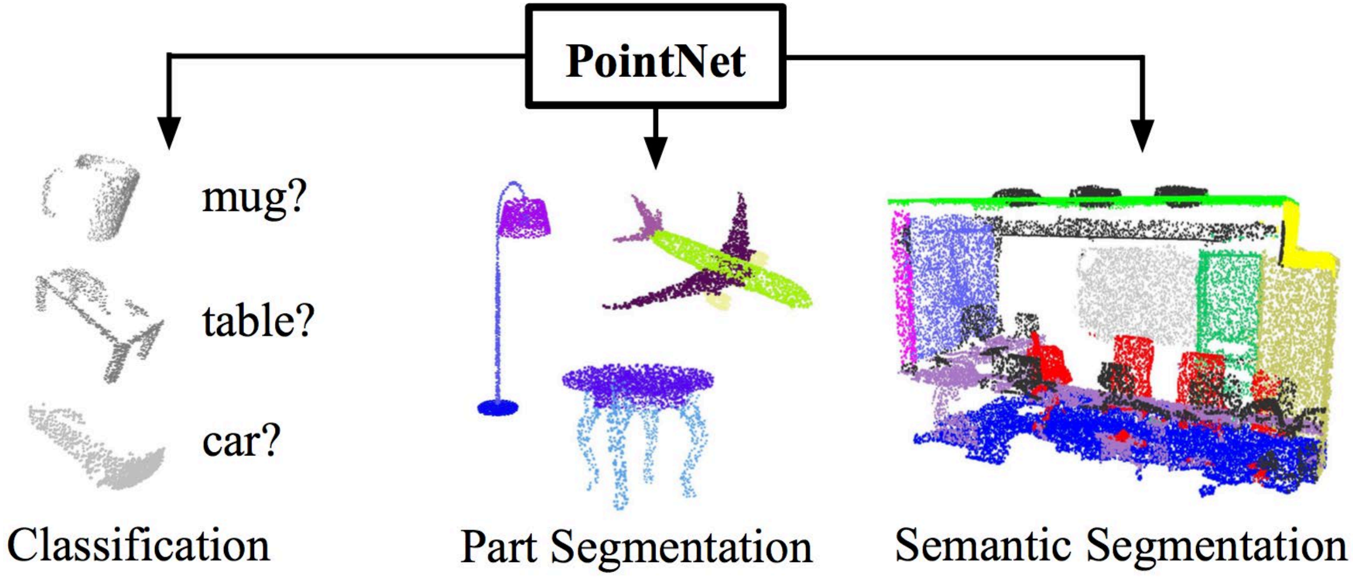
Object Part Segmentation

Semantic Scene Parsing

...

End-to-end learning for irregular point data

Unified framework for various tasks



Charles R. Qi, Hao Su, Kaichun Mo, Leonidas J. Guibas. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. (CVPR'17)

Invariances

The model has to respect key desiderata for point clouds:

Point Permutation Invariance

Point cloud is a set of unordered points

Spatial Transformation Invariance

Point cloud rigid motions should not alter classification results

Sampling Invariance

Output a function of the underlying geometry and not the sampling

Permutation Invariance: Symmetric Functions

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

Examples:

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

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

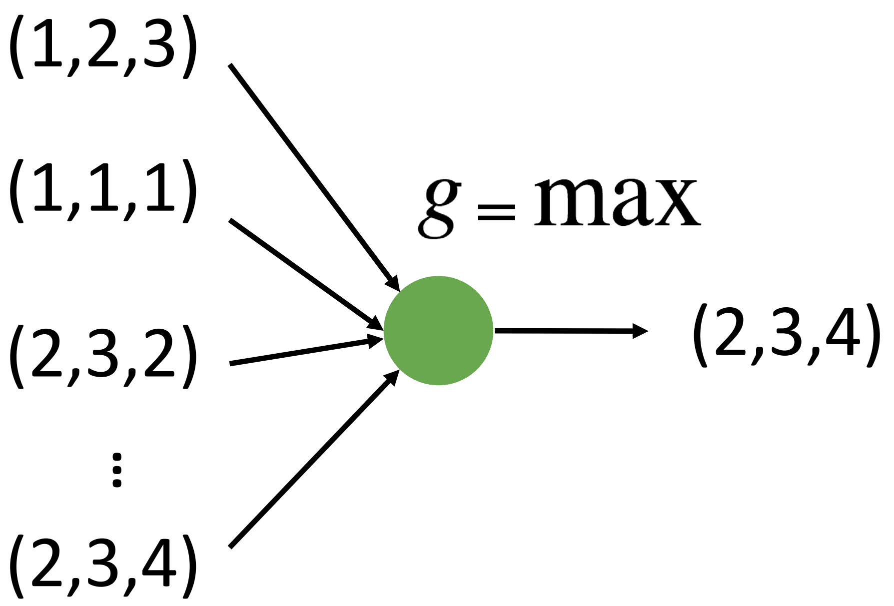
...

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

Construct Symmetric Functions by Neural Networks

Simplest form: directly aggregate all points with a symmetric operator g

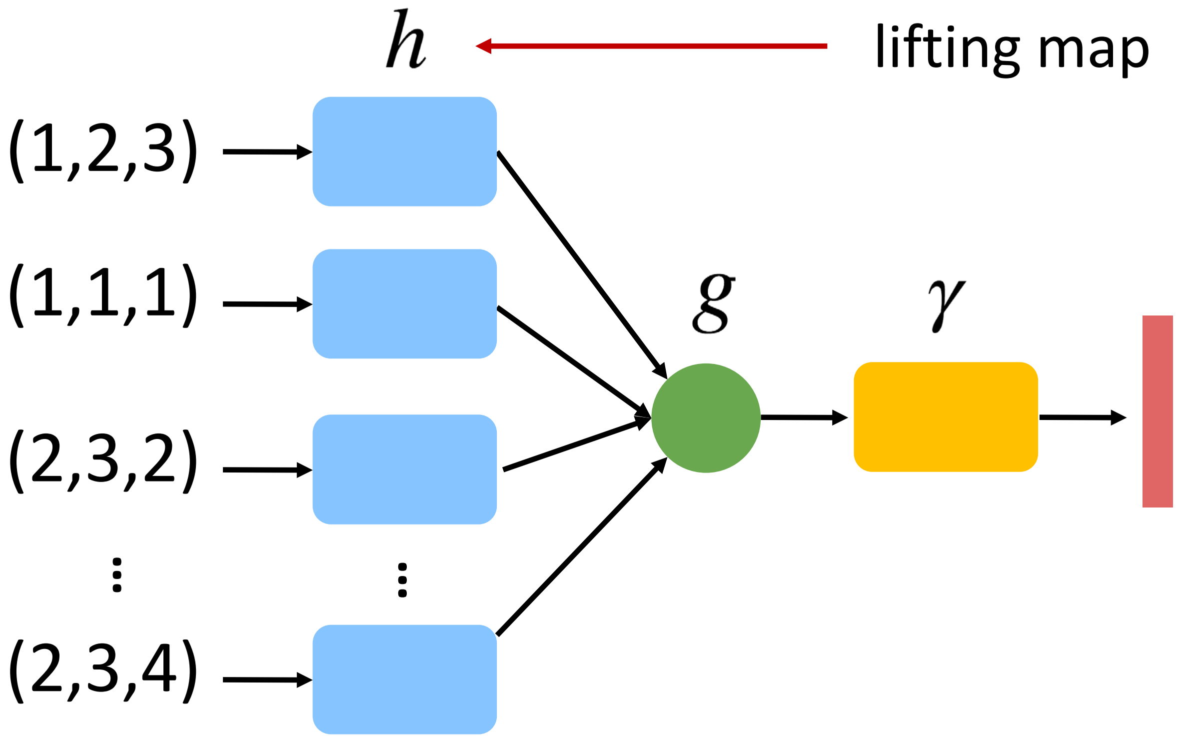
Just discovers simple extreme/aggregate properties of the geometry.



Construct Symmetric Functions by Neural Networks

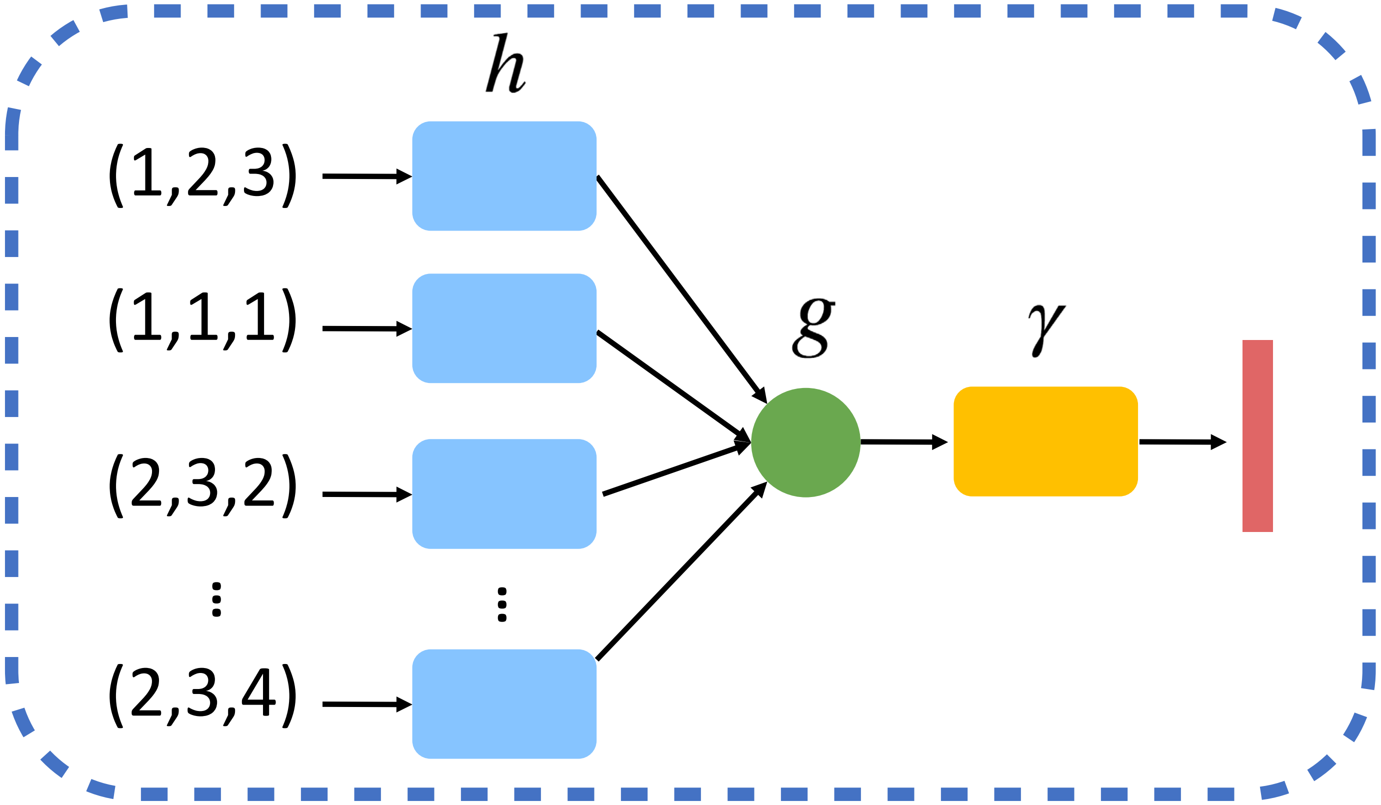
Embed points in a high-dim space before aggregation.

Aggregation in the (redundant) high-dim space encodes more interesting properties of the geometry.



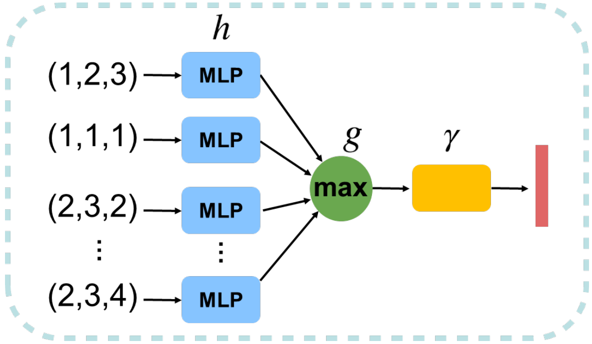
Construct Symmetric Functions by Neural Networks

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



PointNet (vanilla)

Symmetric Functions: Polynomials



$$2\sum_{i\neq j} x_i x_j = \left(\sum_i x_i\right)^2 - \sum_i x_i^{\,2} \qquad \sum_{i\neq j}(x_i - x_j)^2 = 3\sum_i x_i^{\,2} - \left(\sum_i x_i\right)^2$$

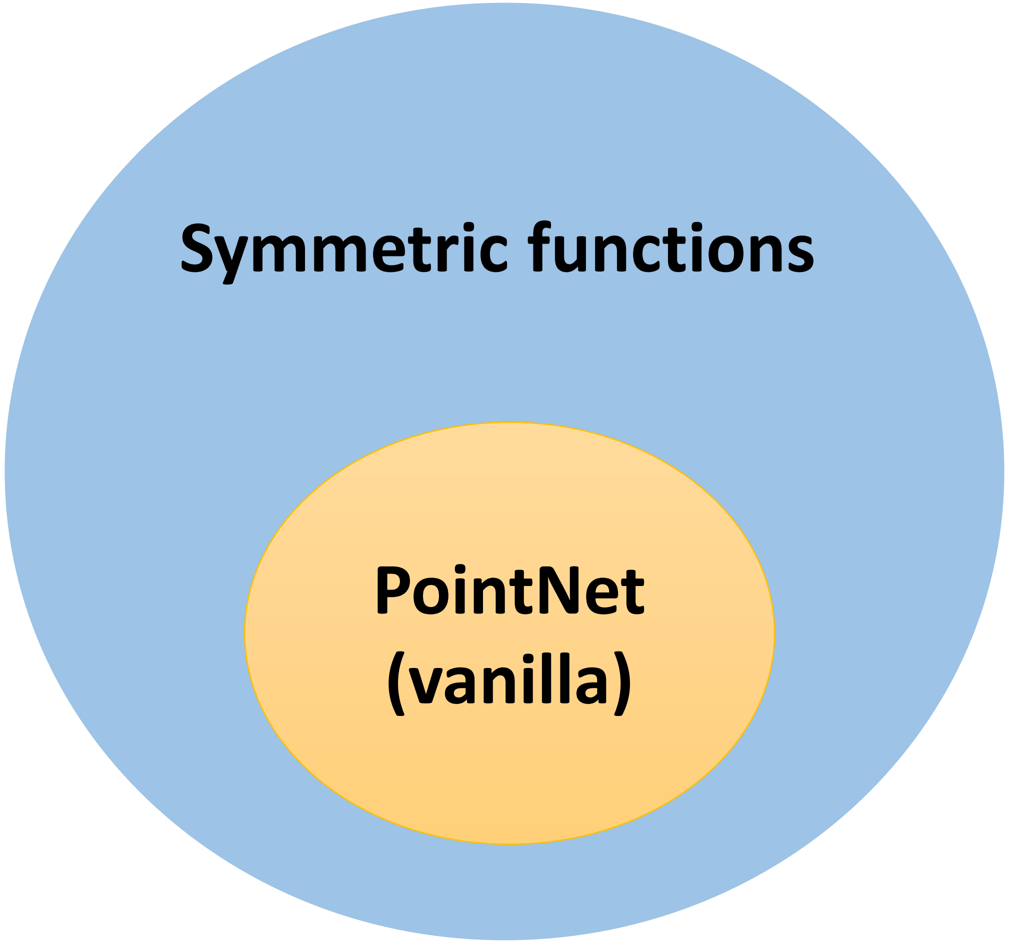
- In fact, any symmetric polynomial in the x_i can be expressed as a polynomial in sums of the form

$$\sum_i x_i^{\,k}$$

and can be computed by

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

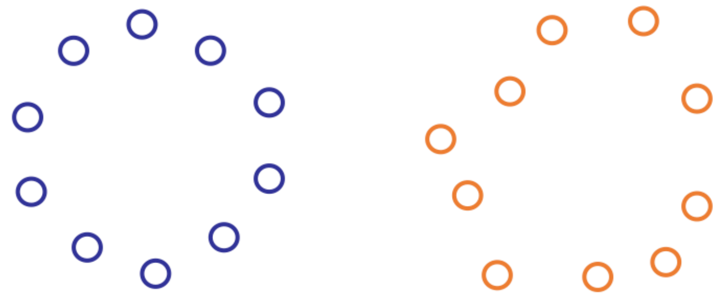
What Symmetric Functions Can Be Constructed By PointNet?



PointNet as a Universal Approximation to Set Functions

Hausdorff continuous:

$f: 2^{\mathcal{X}} \rightarrow \mathbb{R}$ is a continuous set function w.r.t Hausdorff distance



$\mathcal{S}$ $\mathcal{S}'$ (perturbed)

if $d_{Hausdorff}(S, S') \approx 0$, then $f(S) \approx f(S')$

Theorem

A Hausdorff continuous set function $f: 2^{\mathcal{X}} \rightarrow \mathbb{R}$ can be arbitrarily approximated by PointNet.

$$\left| f(S) - \boxed{\gamma\left(\underset{x_i \in S}{\mathrm{MAX}}\{h(x_i)\}\right)} \right| < \epsilon$$

$S \subseteq R^d$

PointNet (vanilla)

Voxel occupancy maps

Invariances

The model has to respect key properties of point clouds:

Point Permutation Invariance

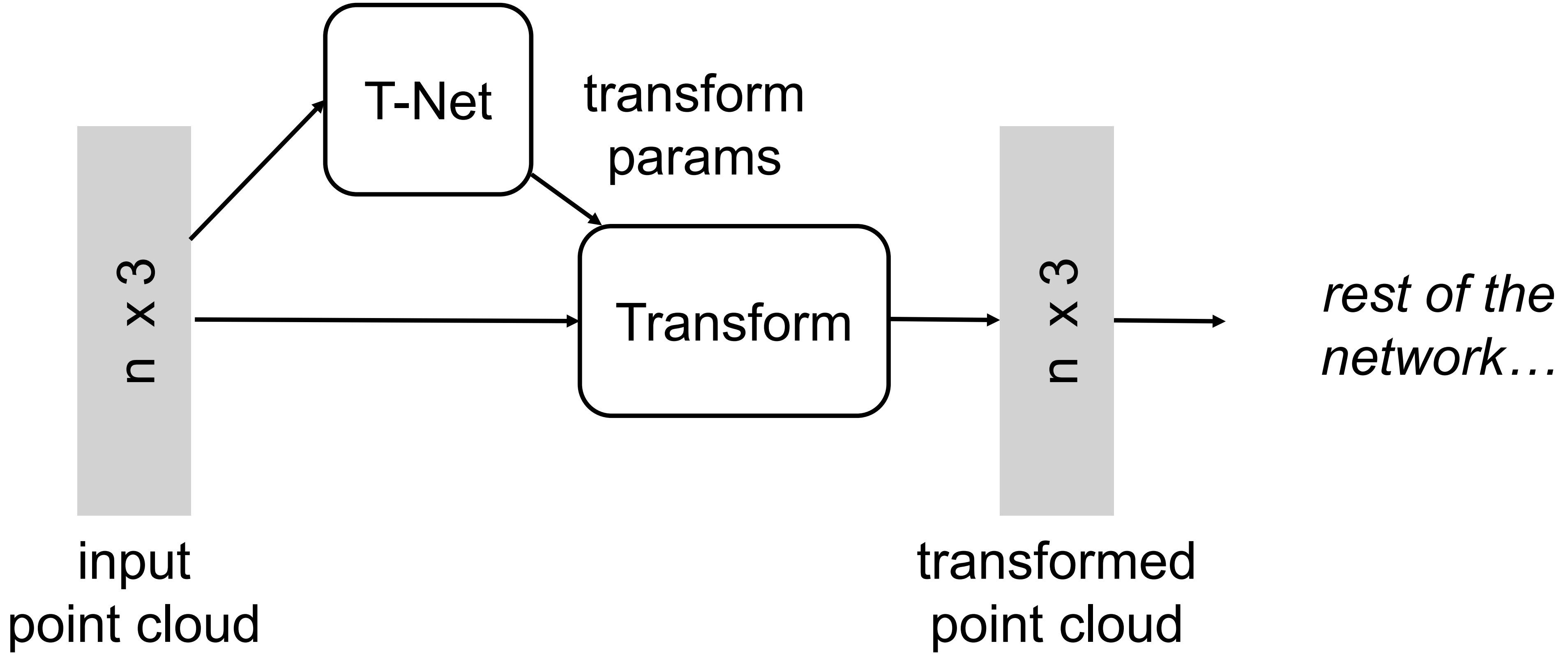
Point cloud is a set of unordered points

Spatial Transformation Invariance

Point cloud rigid motions should not alter classification results

Input Alignment by Transformer Network

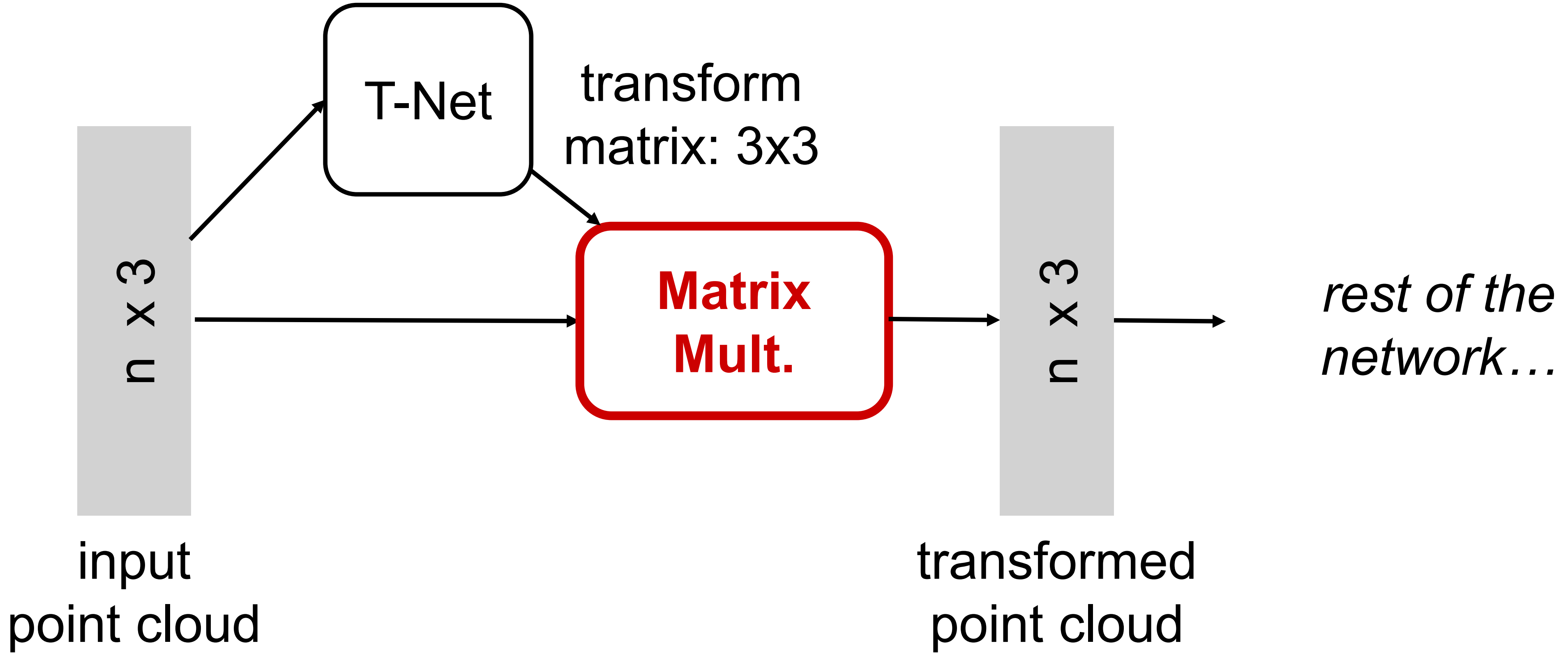
Idea: Data dependent transformation for automatic alignment



Input Alignment by Transformer Network

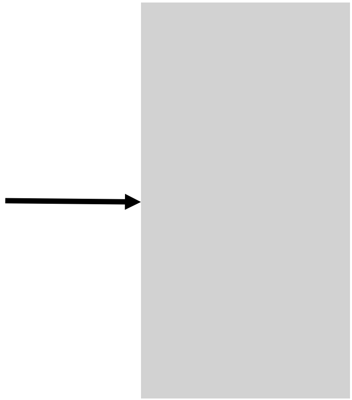
Idea: Data dependent transformation for automatic alignment

The transformation is just matrix multiplication!



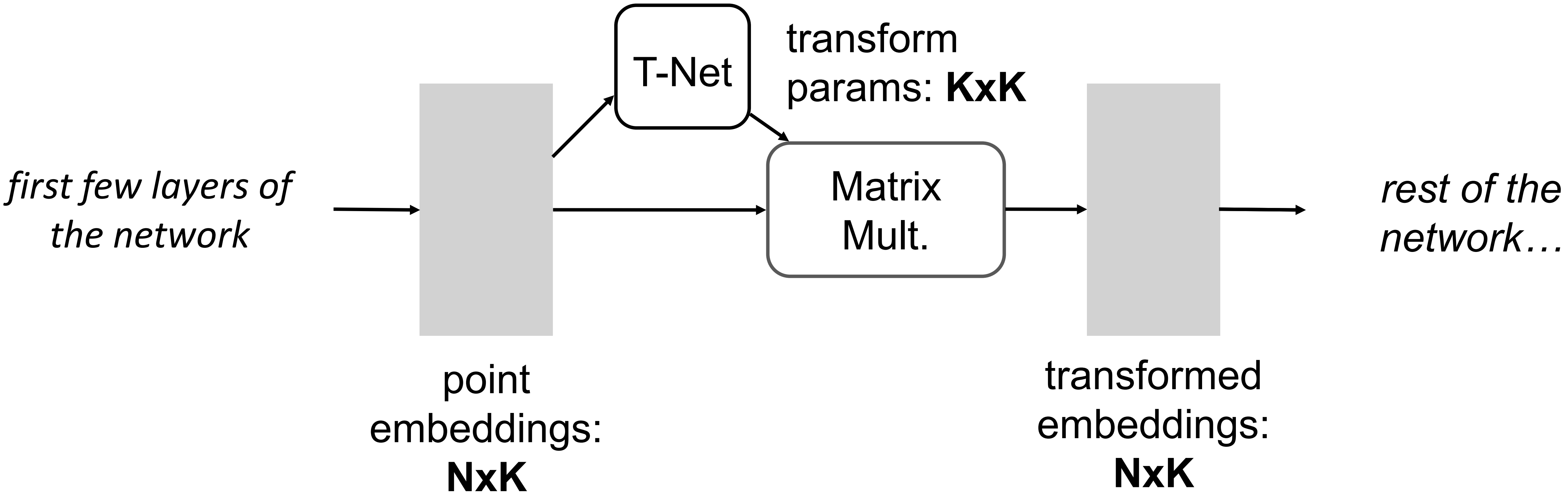
Embedding Space Alignment

first few layers of the network

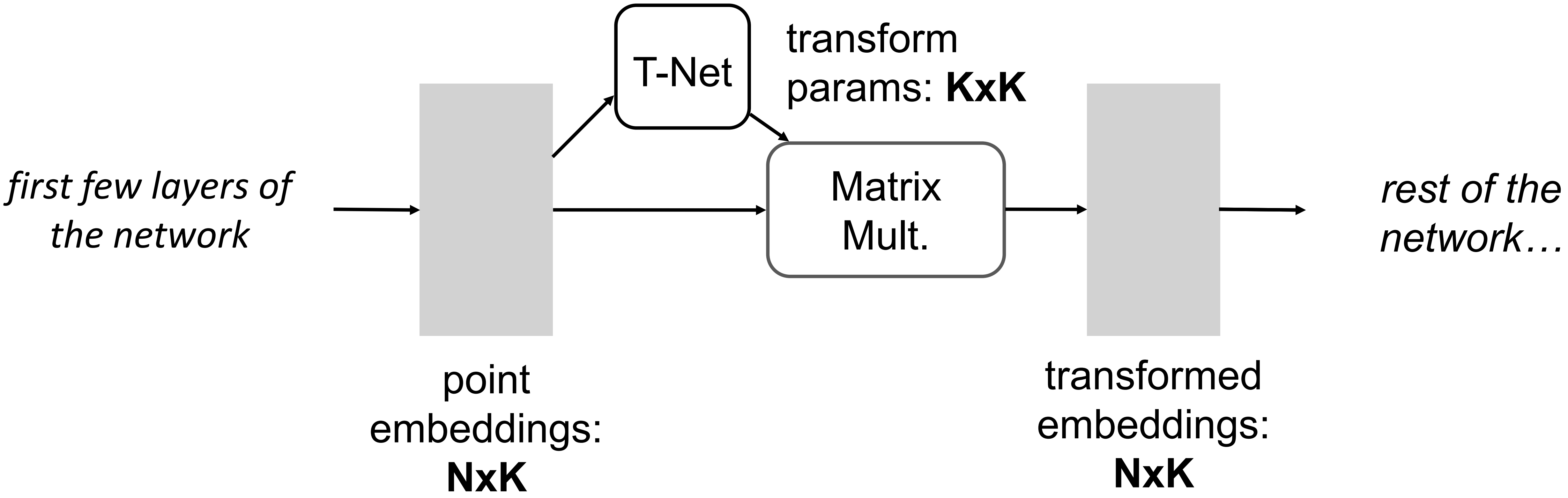


point embeddings: **NxK**

Embedding Space Alignment



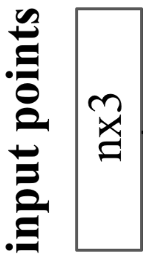
Embedding Space Alignment



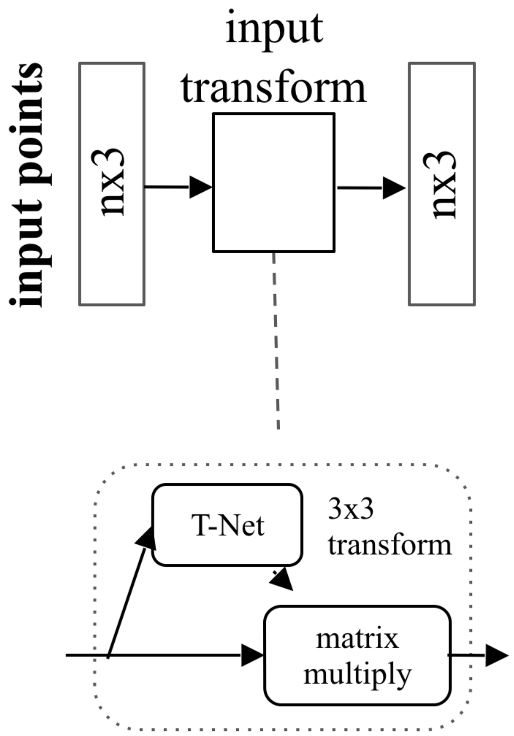
Regularization loss:

Transform matrix close to orthogonal: $L_{reg} = \|I - AA^T\|_F^2$

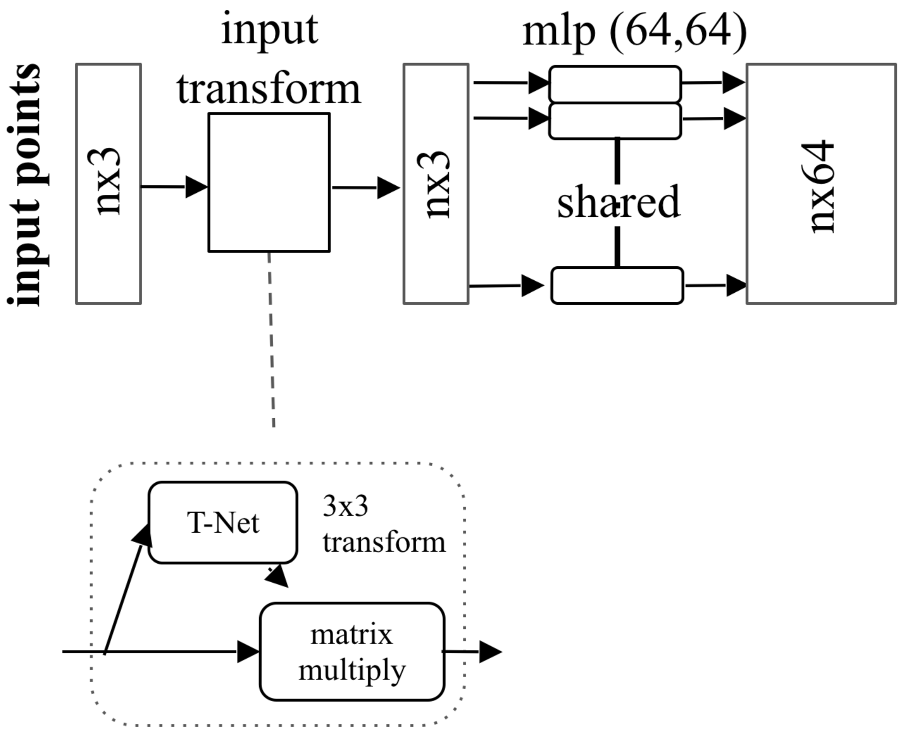
PointNet Classification Network



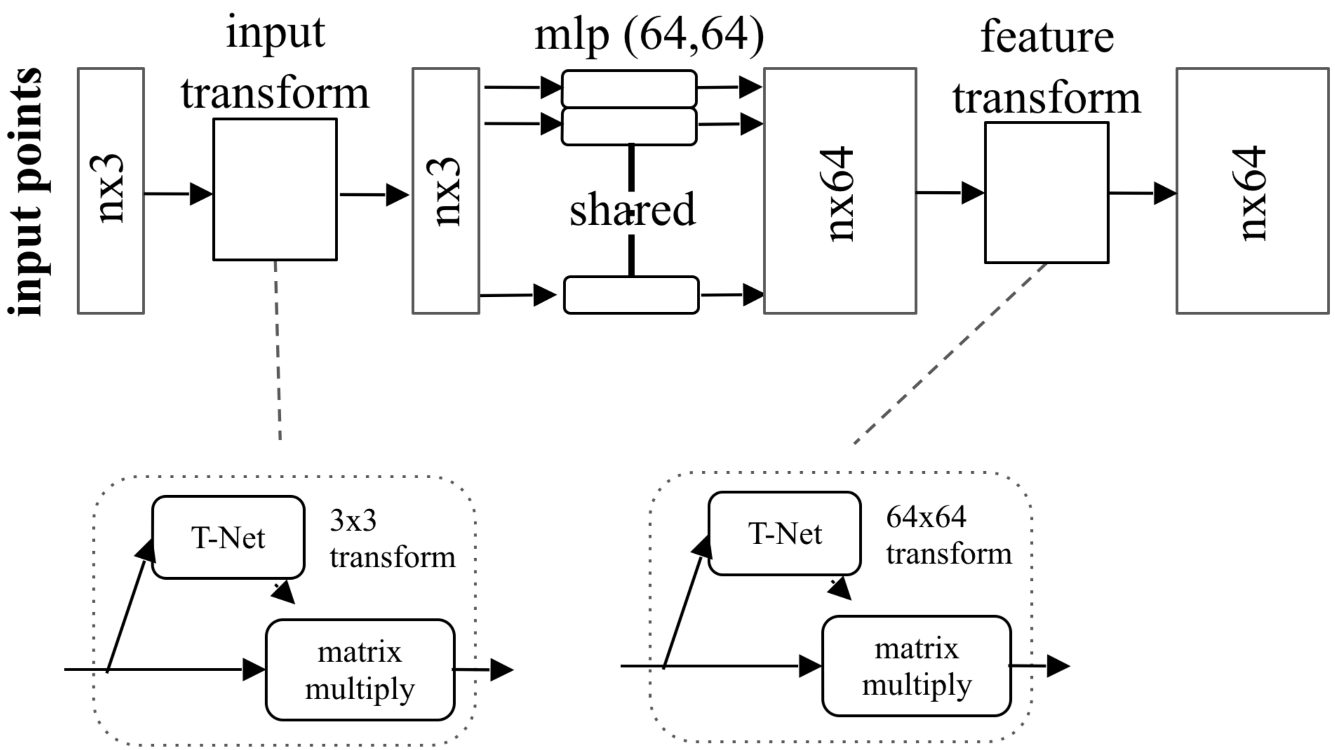
PointNet Classification Network



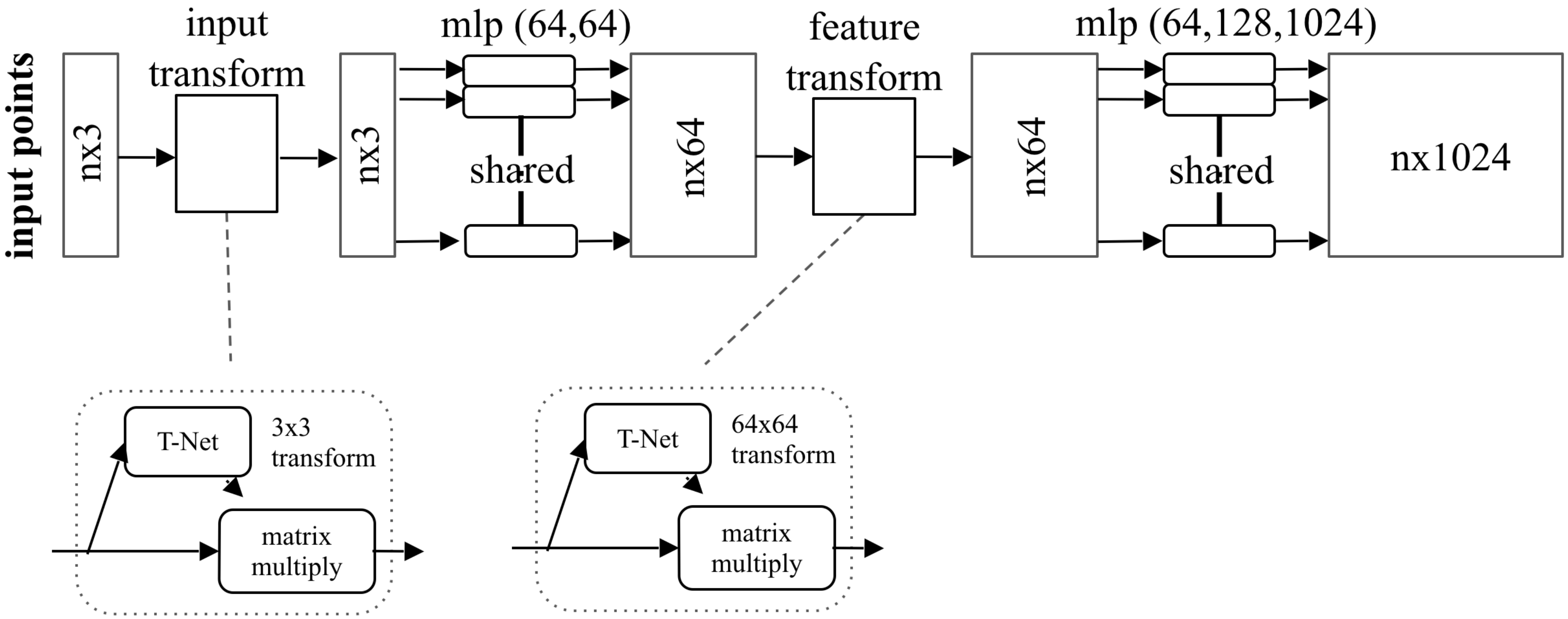
PointNet Classification Network



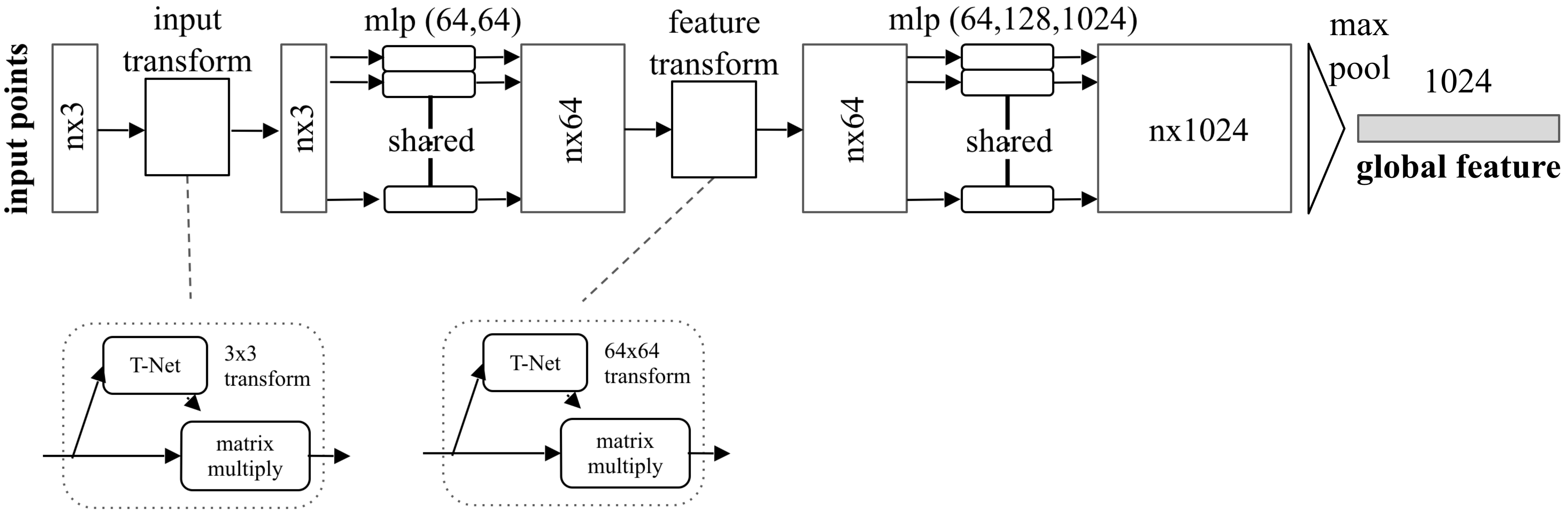
PointNet Classification Network



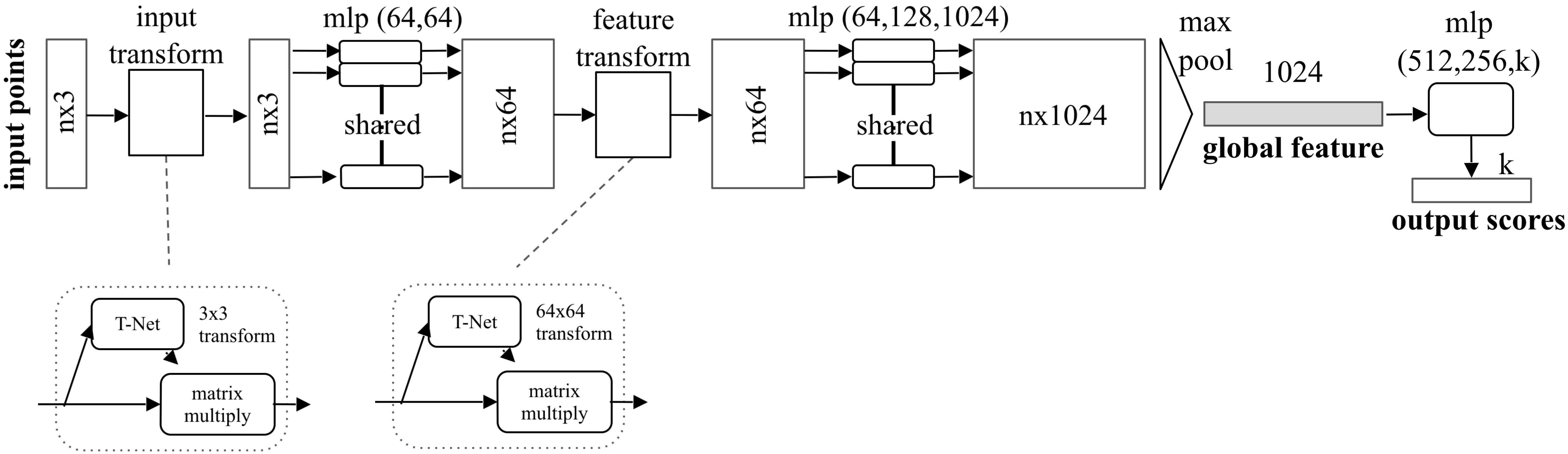
PointNet Classification Network



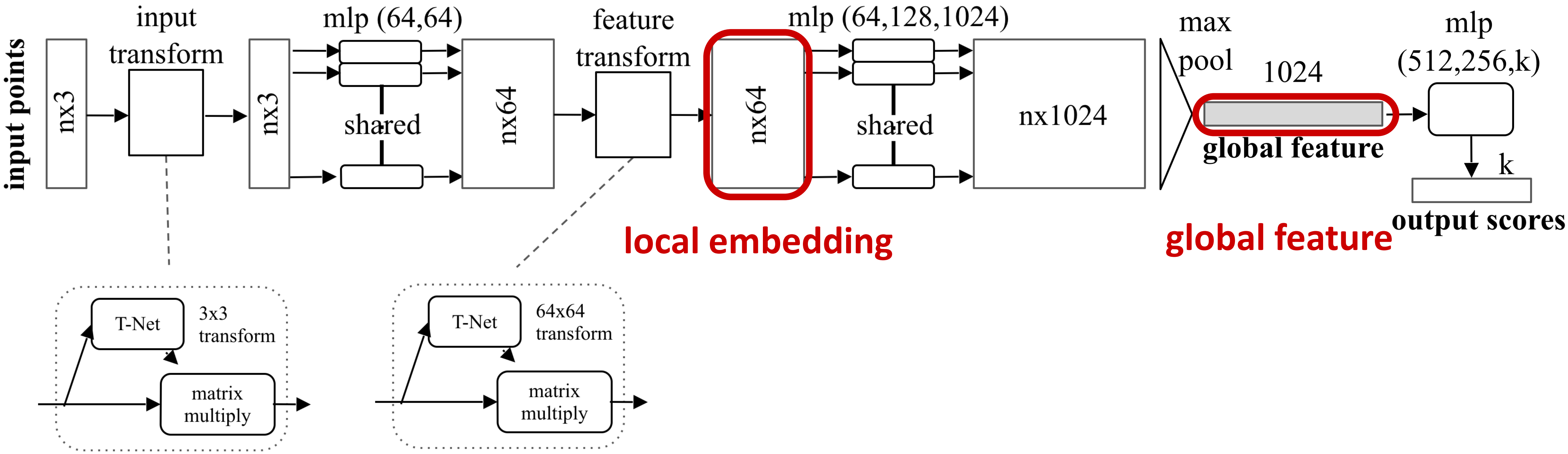
PointNet Classification Network



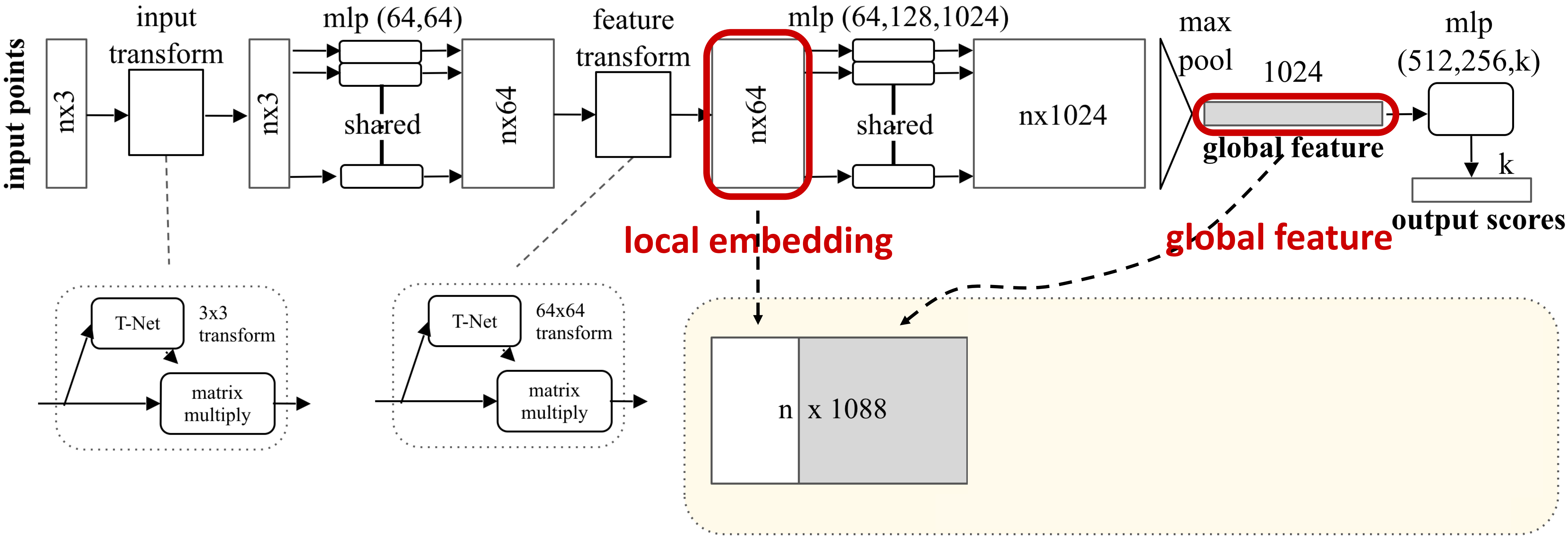
PointNet Classification Network



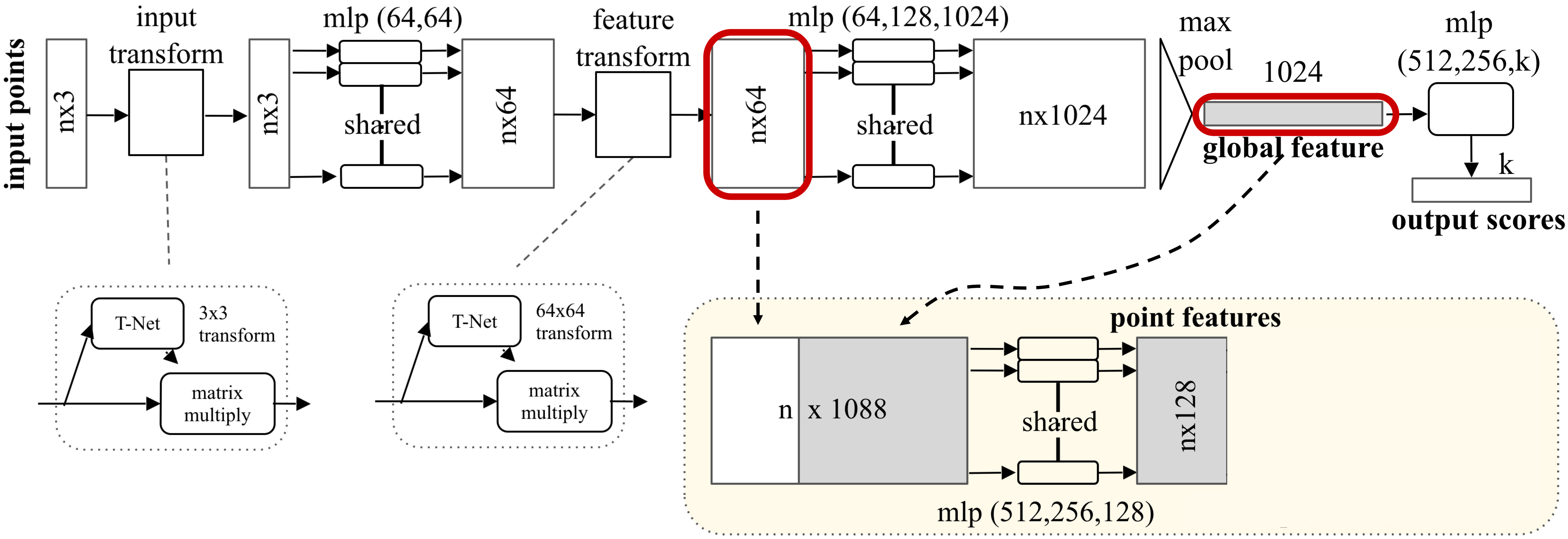
Extension to PointNet Segmentation Network



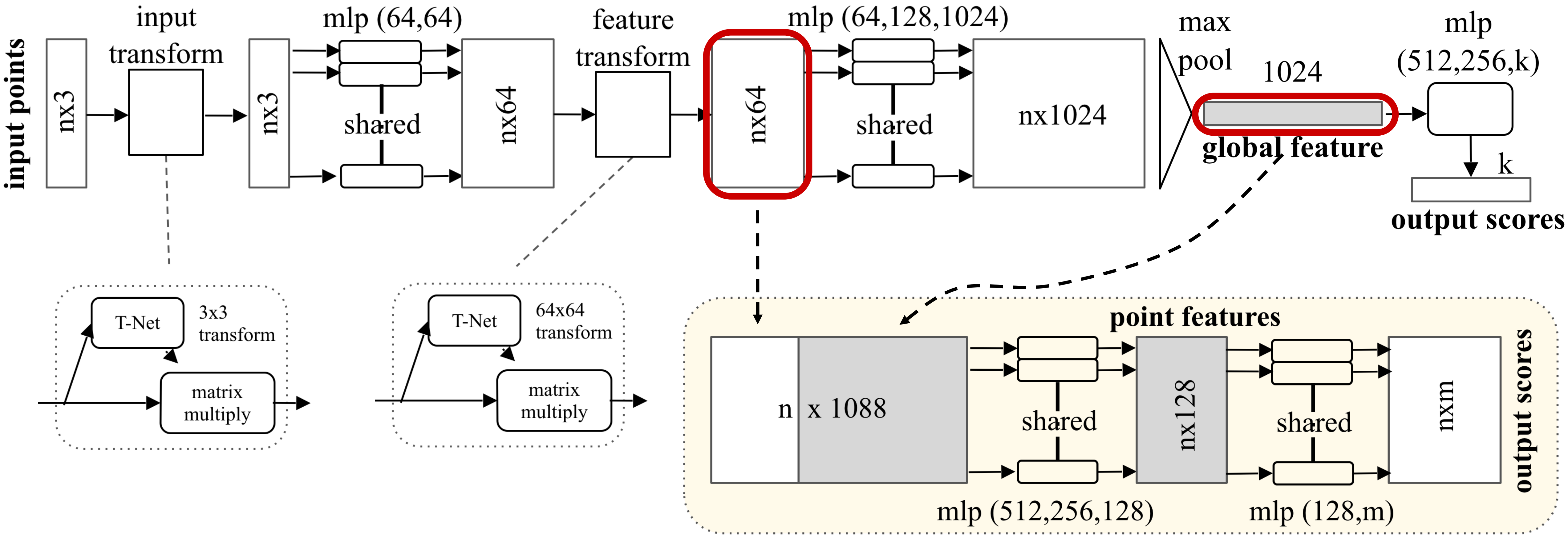
Extension to PointNet Segmentation Network



Extension to PointNet Segmentation Network



Extension to PointNet Segmentation Network



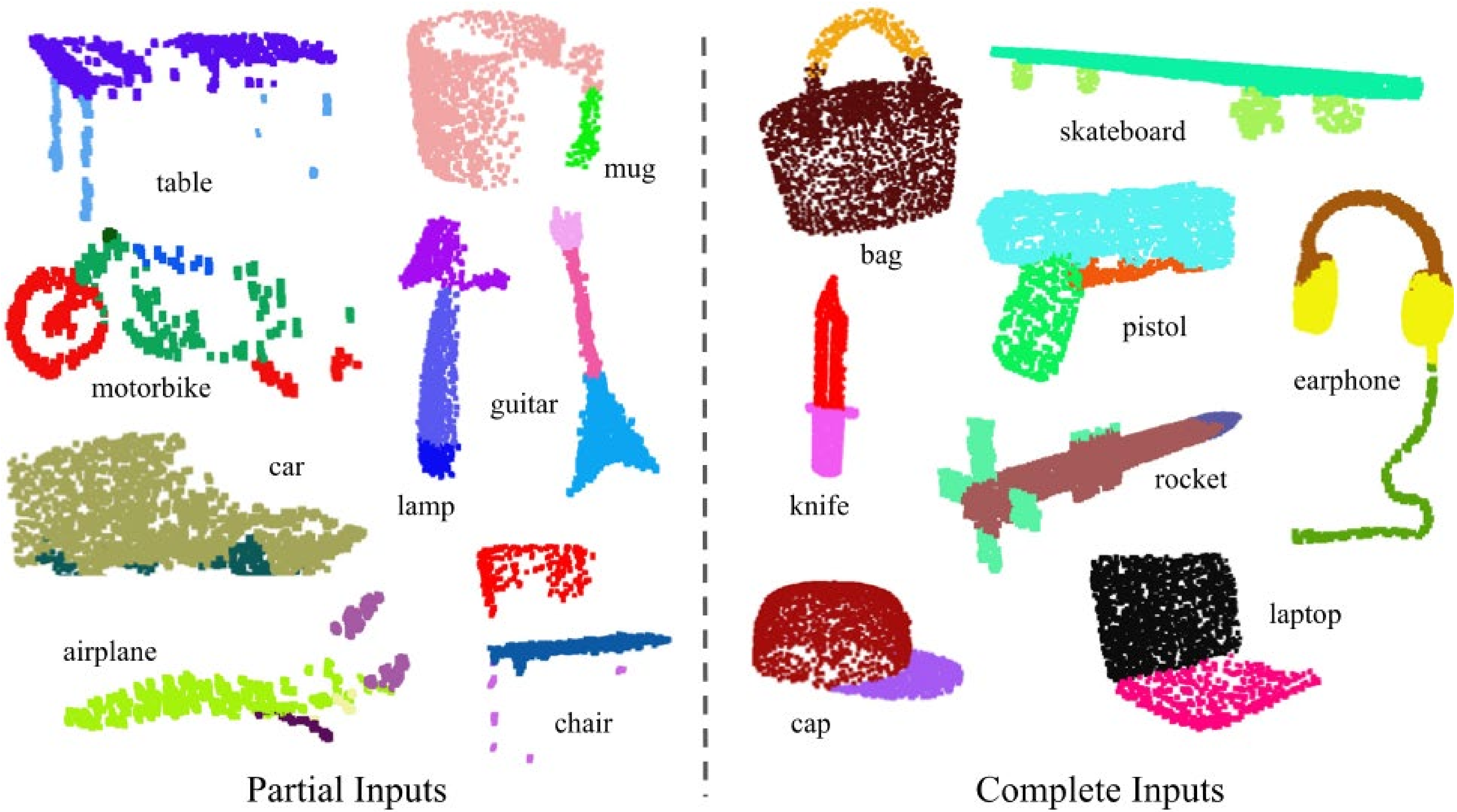
Results on Object Classification

3D CNNs

	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



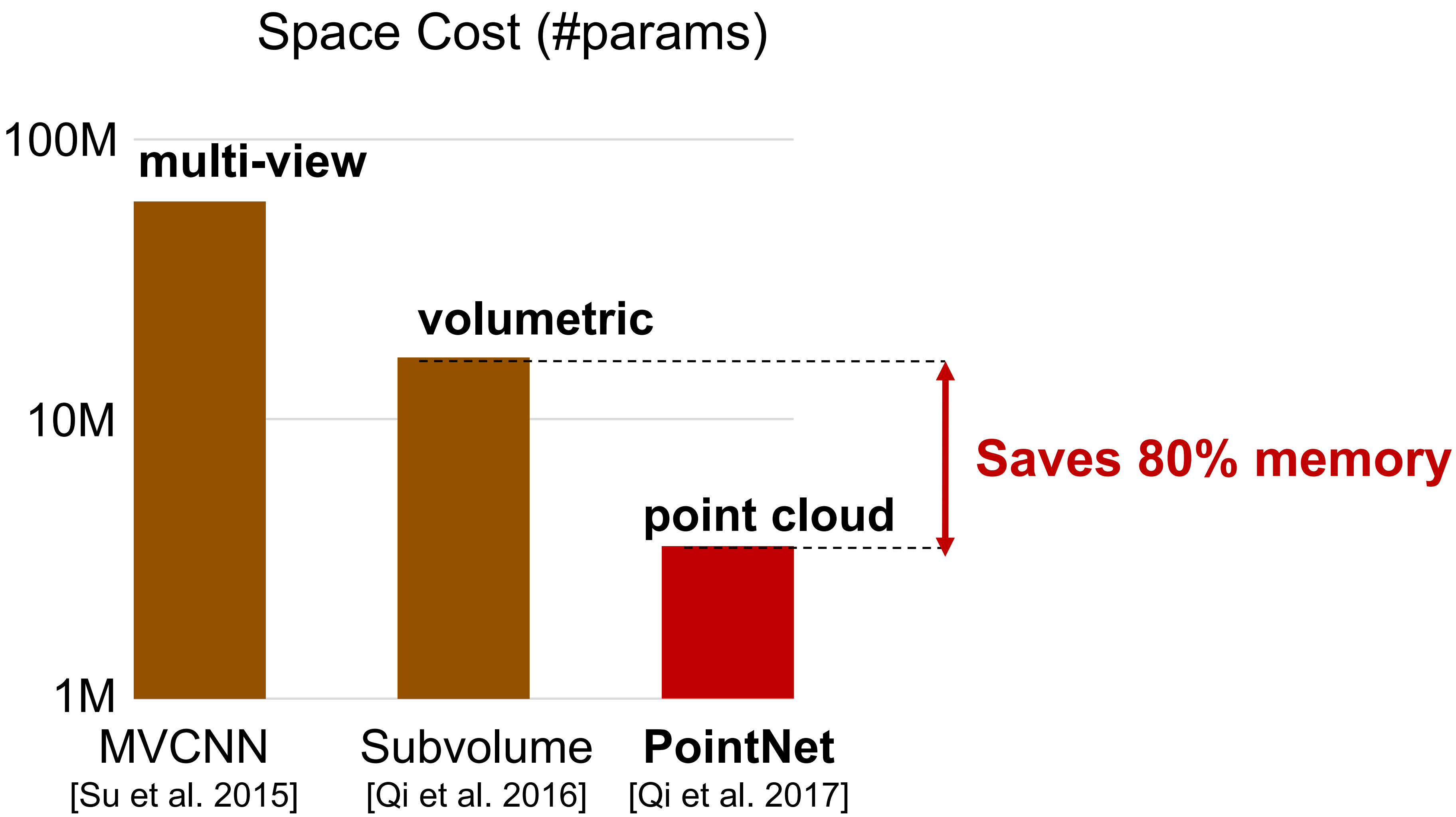
Results on Semantic Scene Parsing

Input

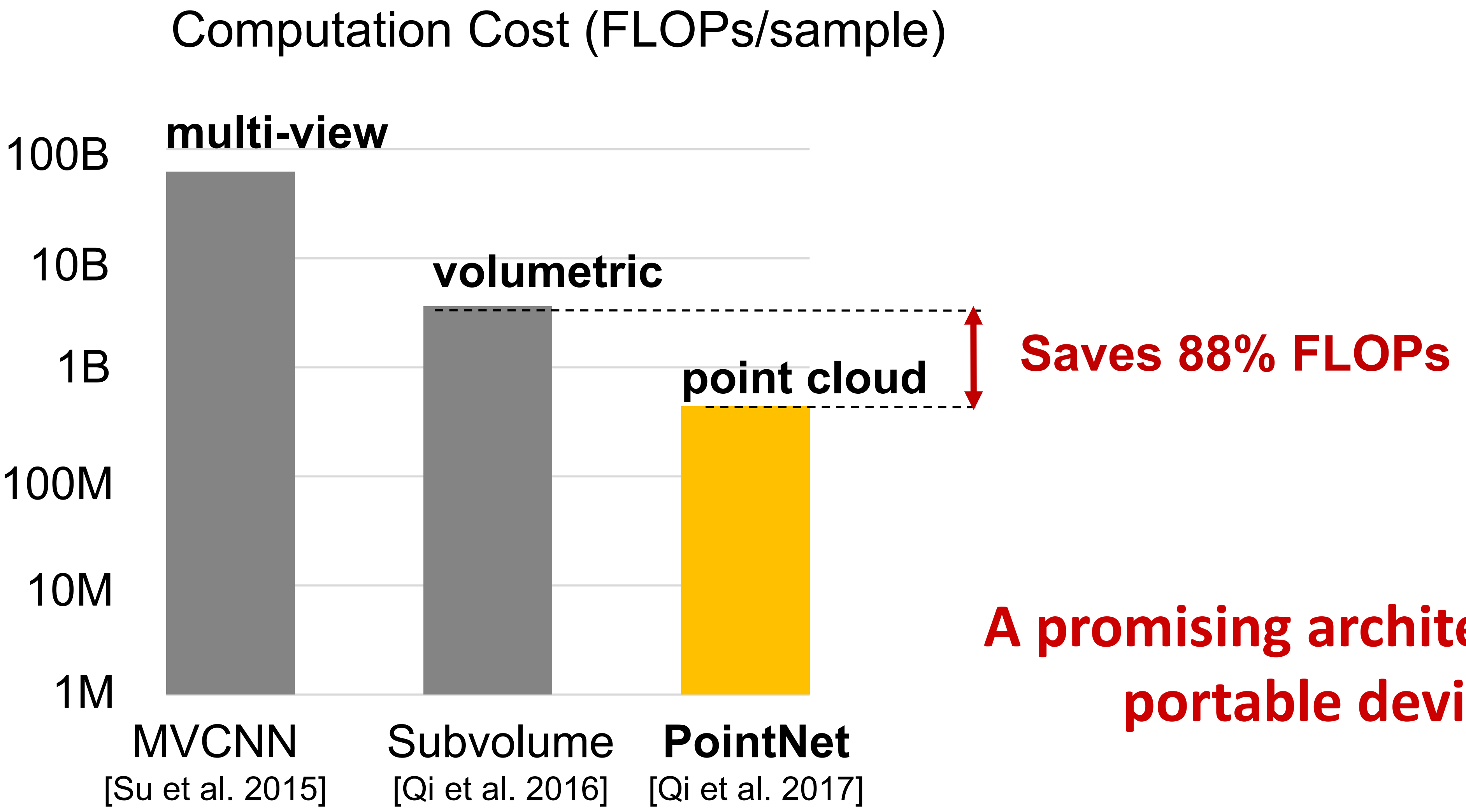
Output

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

PointNet is Light-Weight and Fast

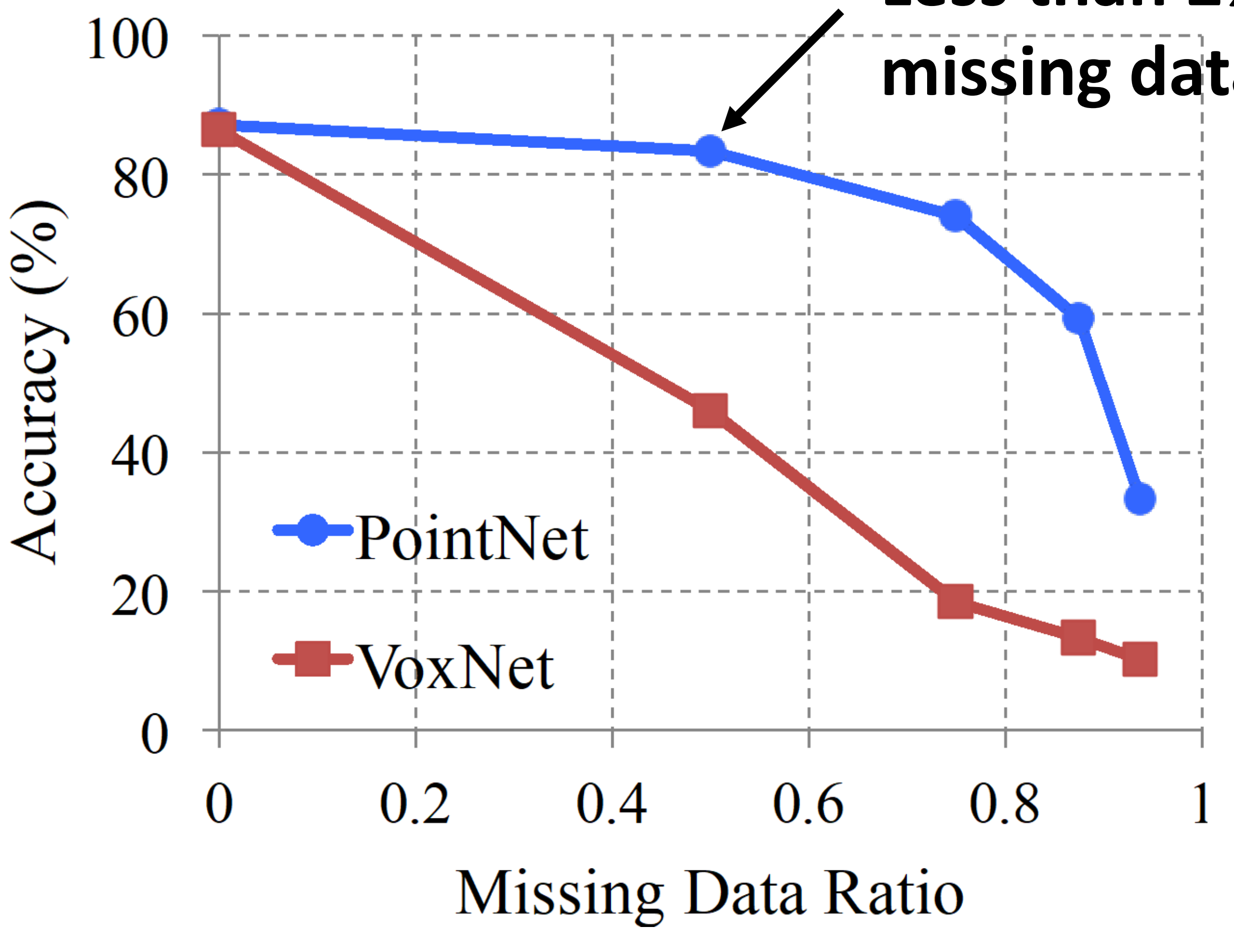


PointNet is Light-Weight and Fast



A promising architecture for portable devices!

PointNet is Robust to Data Corruption



Less than 2% accuracy drop with 50% missing data

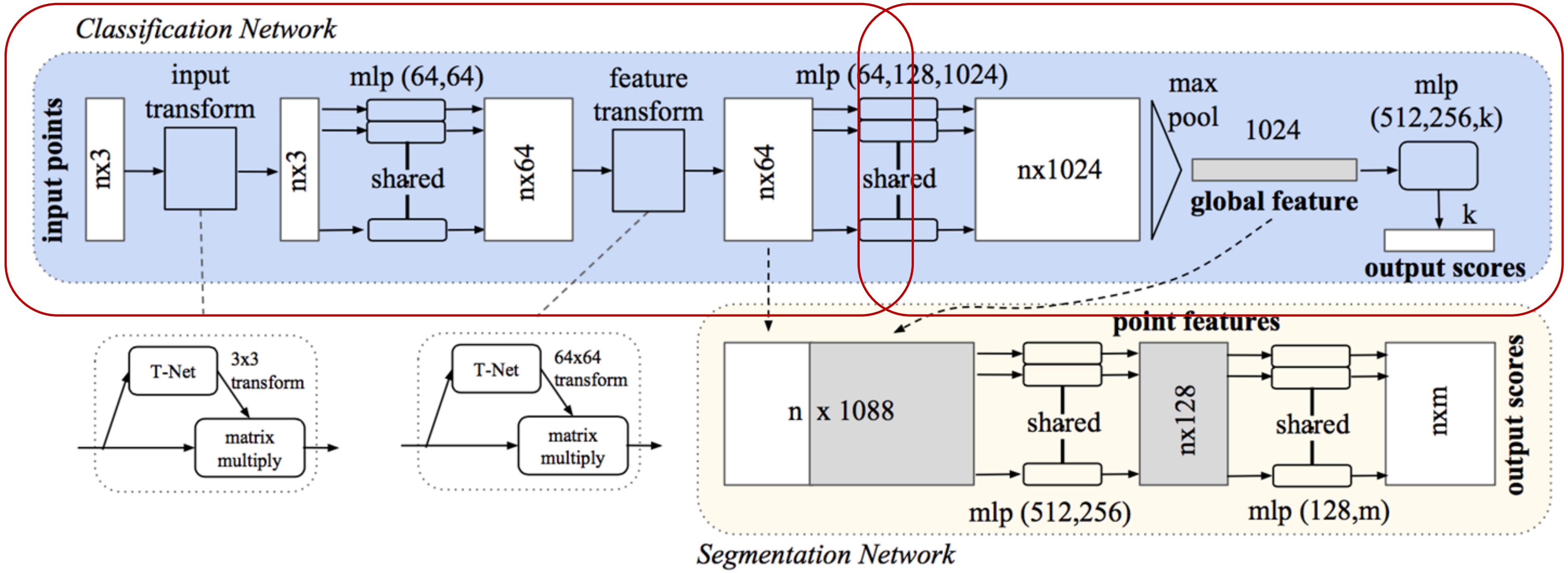
Why is PointNet so robust to missing data?

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

Visualizing Global Point Cloud Features

Original Shape

Learning Interesting Points

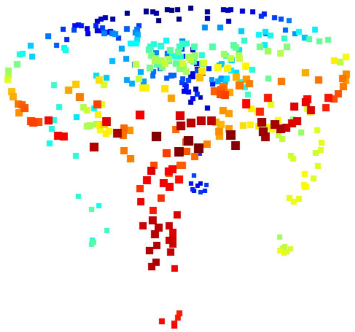


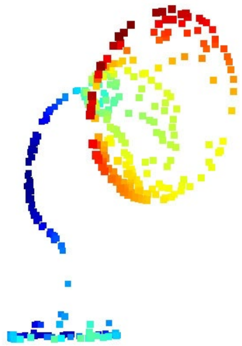
Pointnet learns optimization criteria, which in turn pick interesting points

Visualizing Global Point Cloud Features

Original Shape

Critical Points



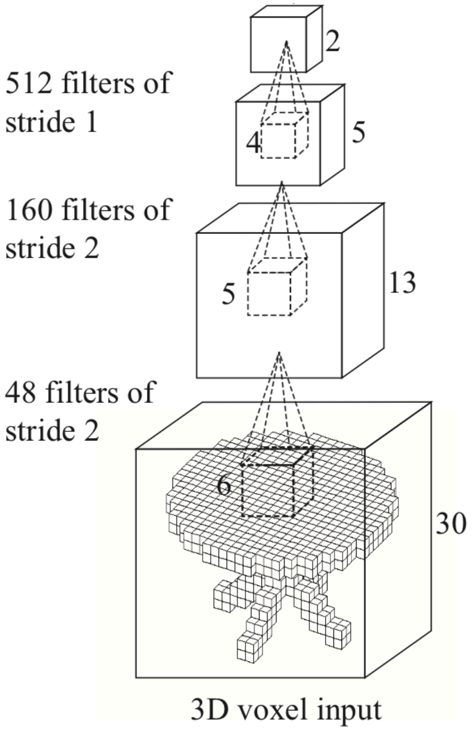


PointNet learns to pick perceptually interesting points

From PointNet to PointNet++

Limitations of PointNet

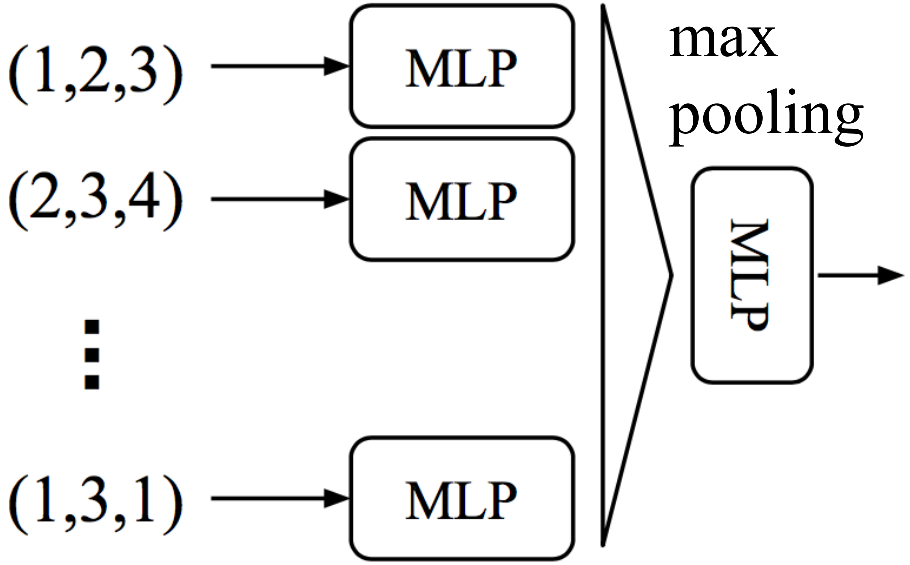
Hierarchical feature learning
multiple levels of abstraction



3D CNN [Wu et al.2015]

v.s.

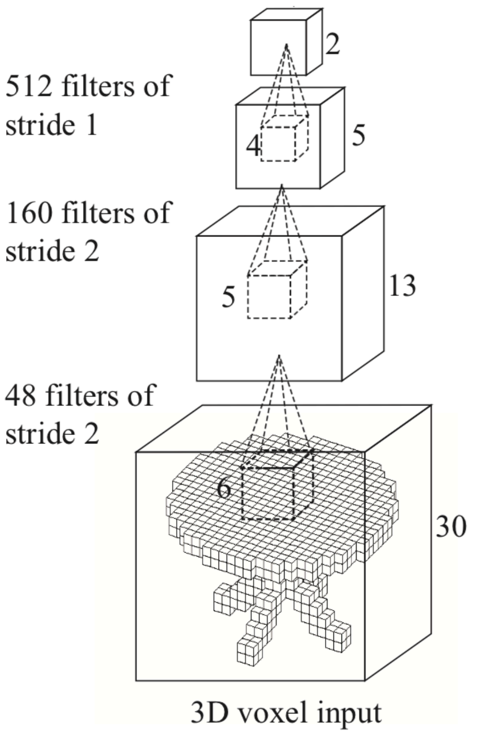
Global feature learning
either **one** point, or **all** points



PointNet (vanilla) [Qi et al.2017]

Limitations of PointNet

Hierarchical feature learning
multiple levels of abstraction



3D CNN [Wu et al.2015]

v.s.

Global feature learning
either **one** point or **all** points

No local context

Limited local invariance

PointNet (vanilla) [Qi et al.2017]

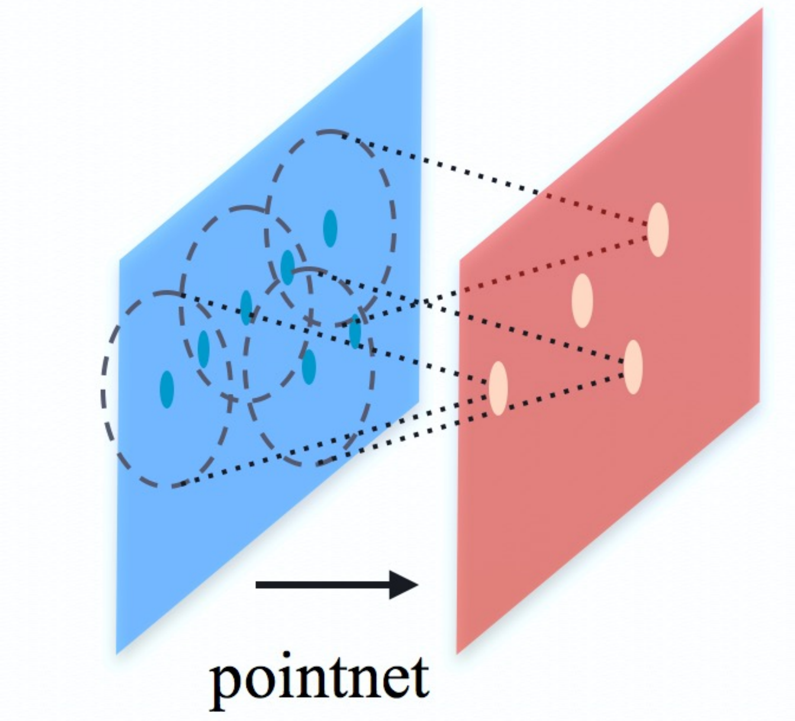
PointNet++

Basic idea: Recursively apply pointnet at local regions.

✓ Hierarchical feature learning

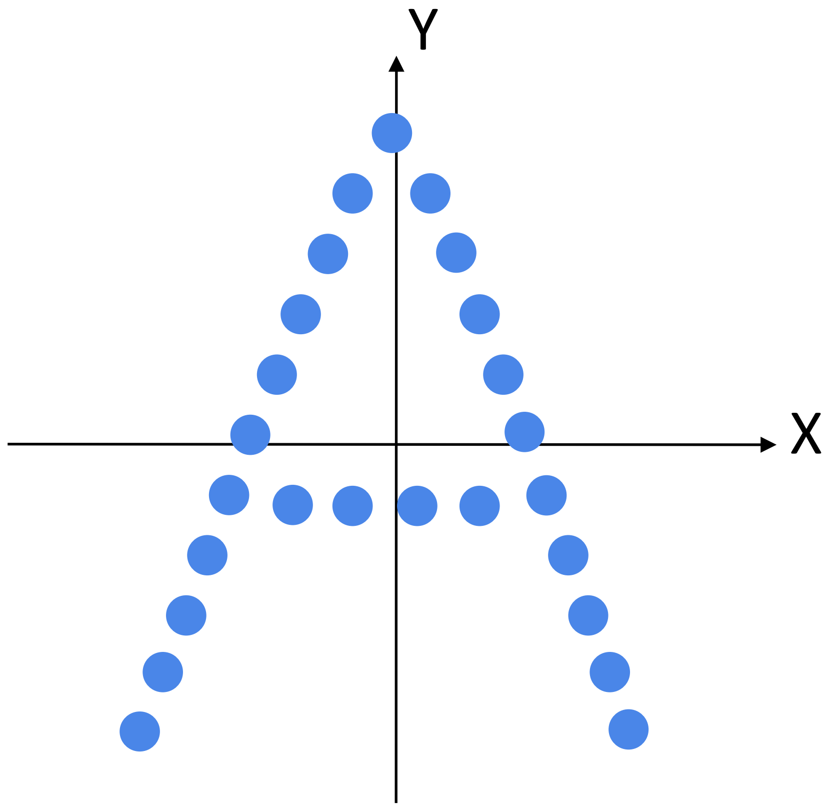
✓ Local translation invariance

✓ Permutation invariance



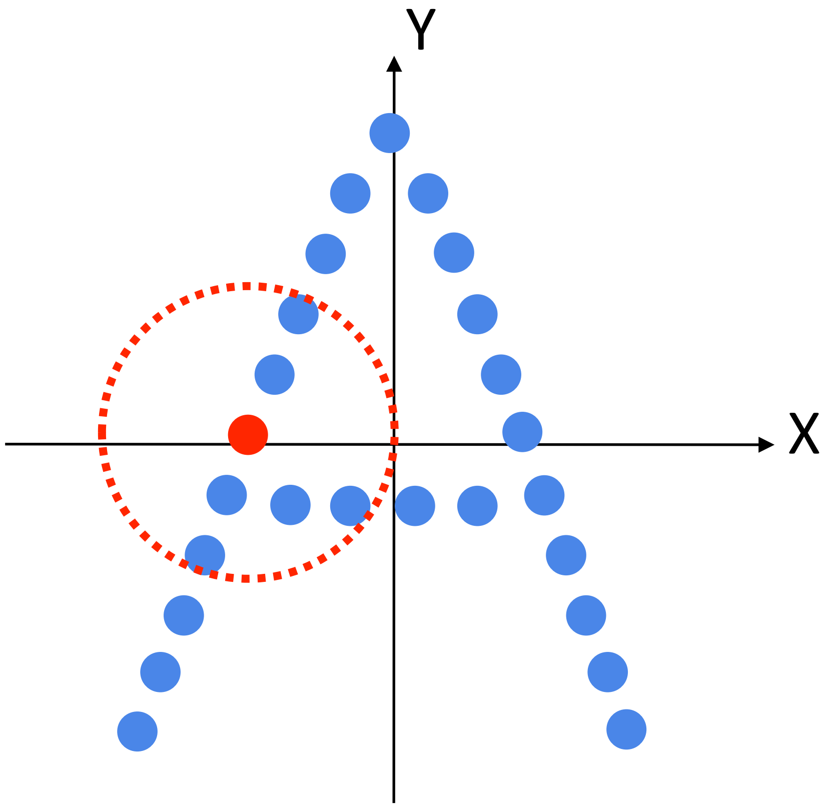
Charles R. Qi, Li Yi, Hao Su, Leonidas Guibas. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space (NIPS’17)

Hierarchical Point Feature Learning



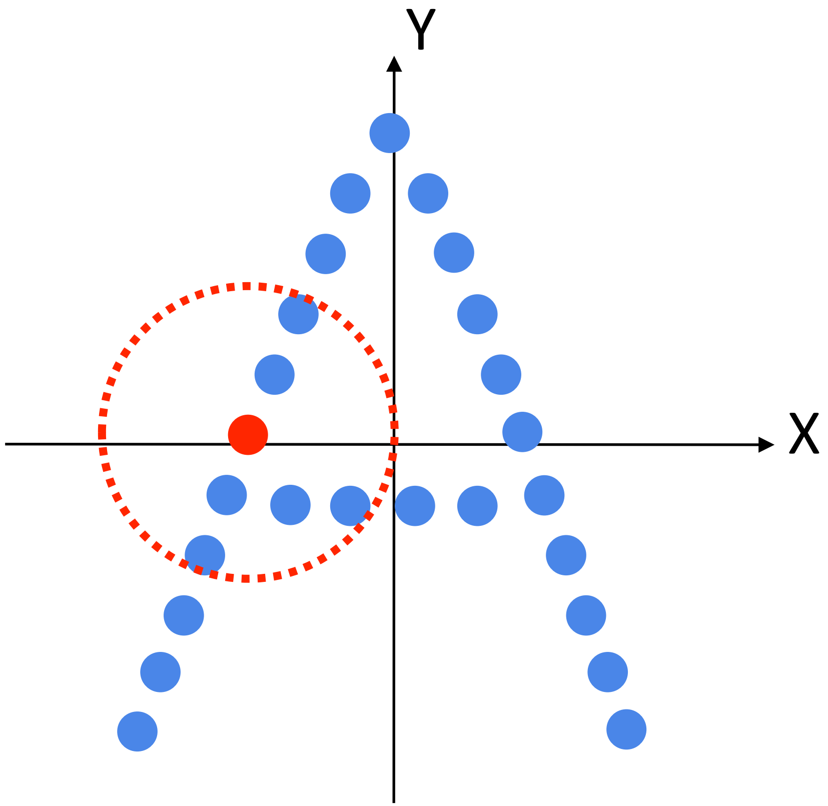
N points in (X,Y)

Hierarchical Point Feature Learning

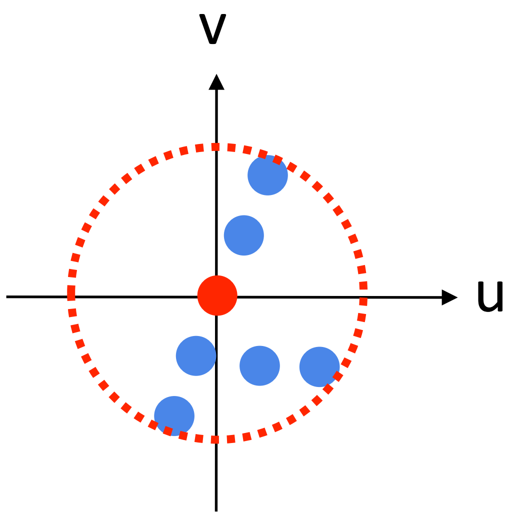


N points in (X,Y)

Hierarchical Point Feature Learning

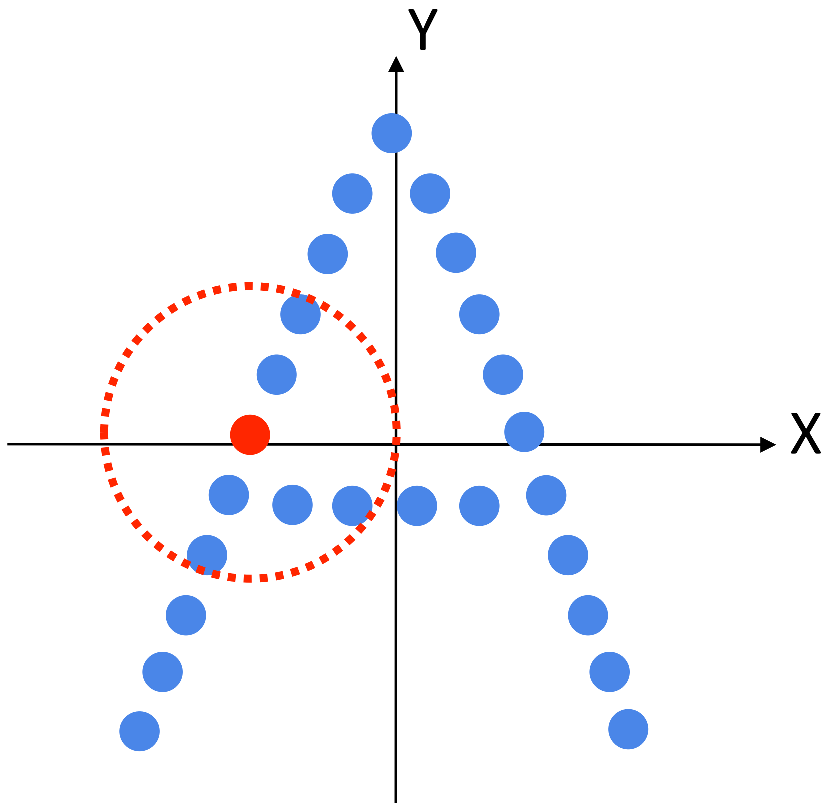


N points in (X,Y)



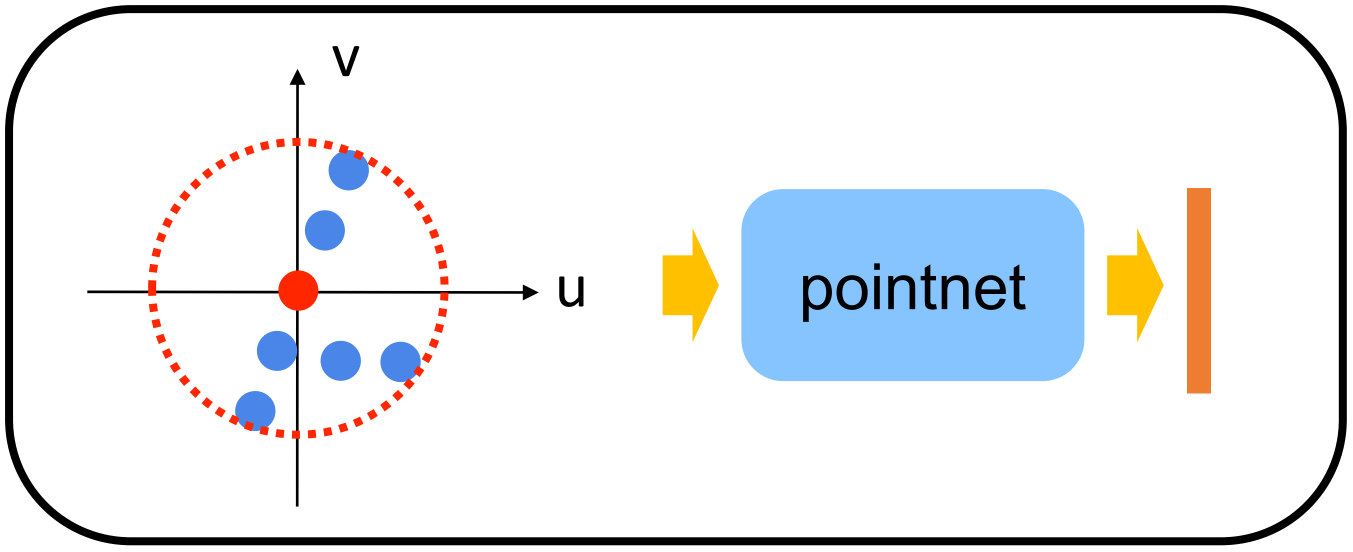
k points in local coordinates (u,v)

Hierarchical Point Feature Learning



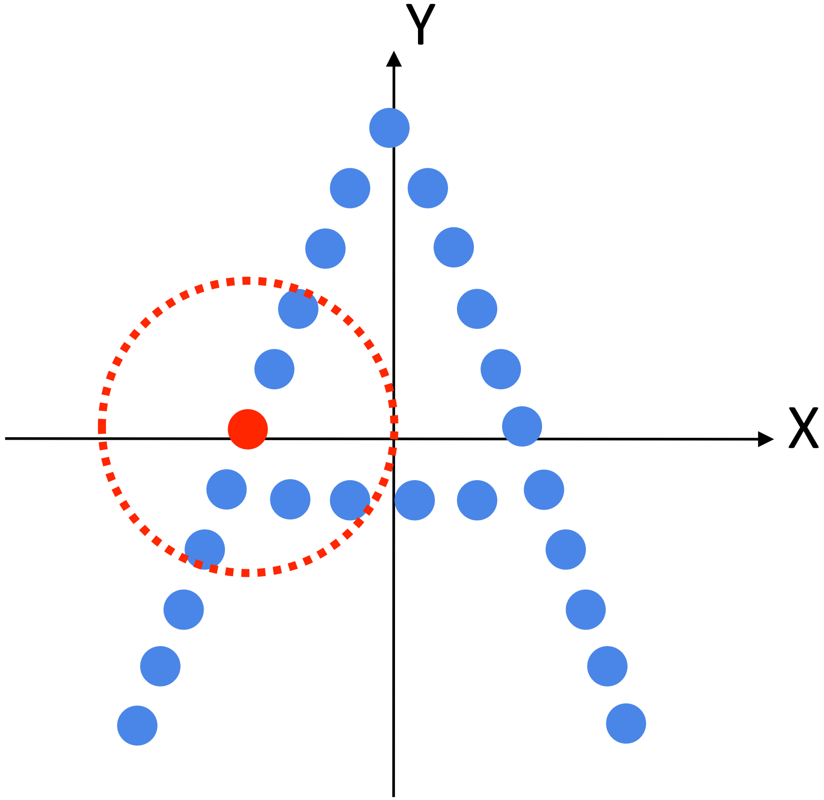
N points in (X,Y)

Apply pointnet at a local region

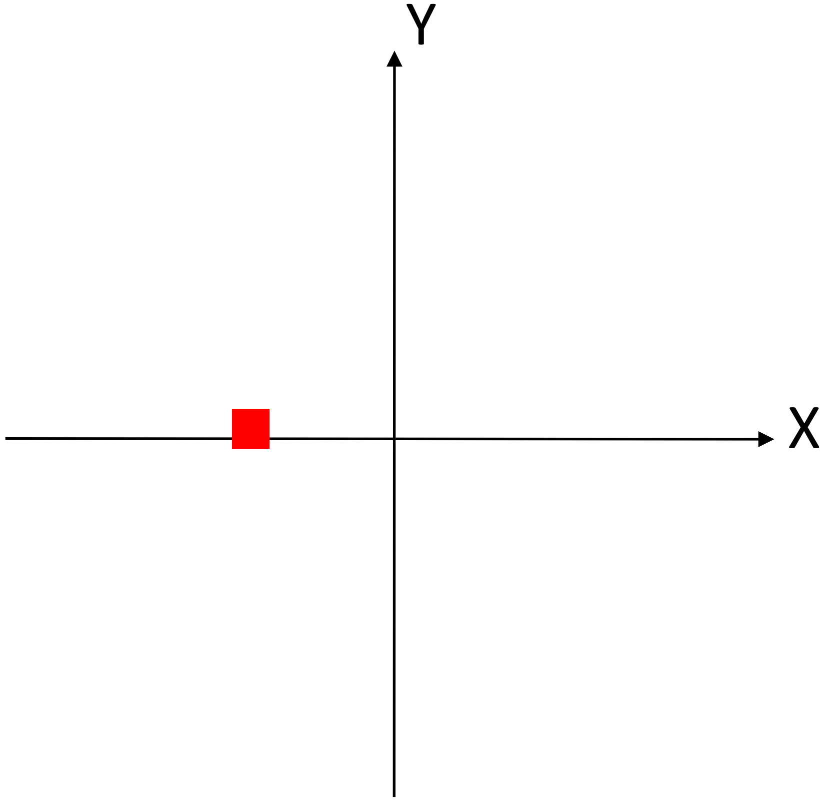


k points in local coordinates (u,v)

Hierarchical Point Feature Learning



N points in (X,Y)

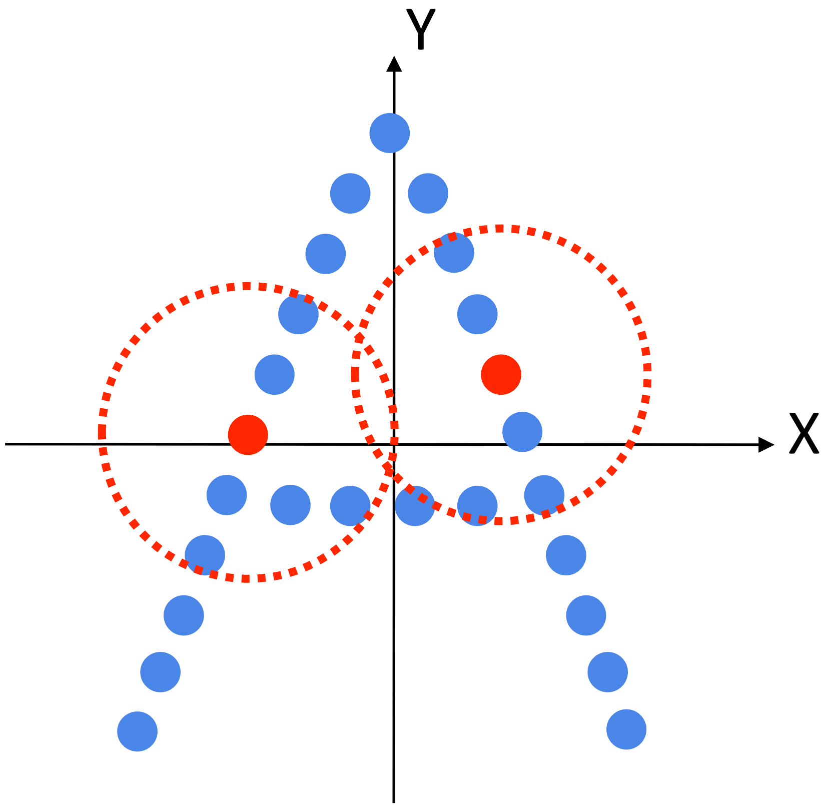


points in (X,Y, **F**)

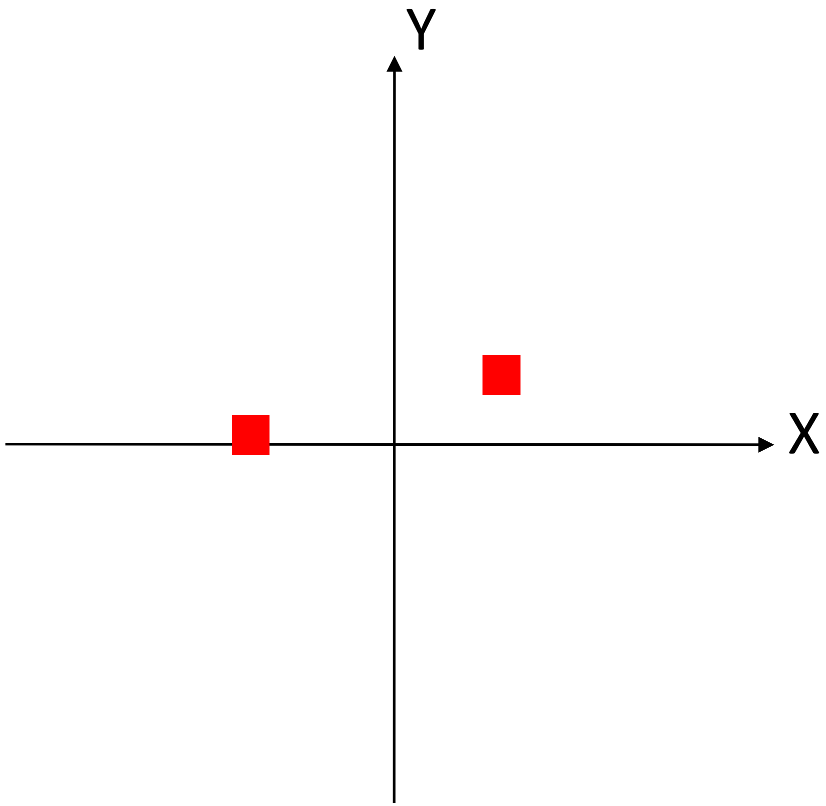
Euclidean space

high-dim feature space

Hierarchical Point Feature Learning

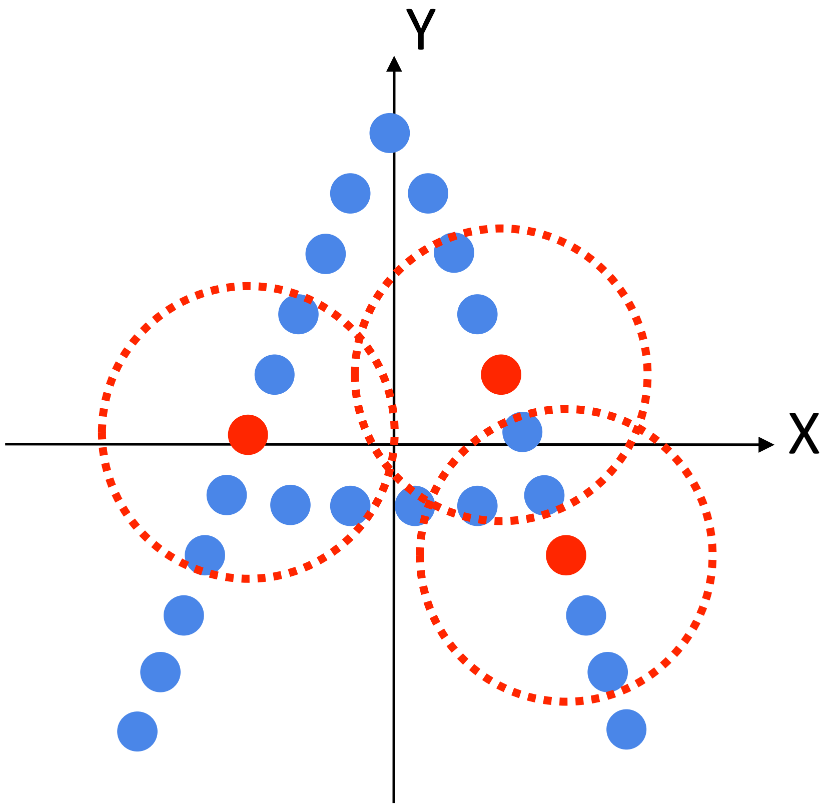


N points in (X,Y)

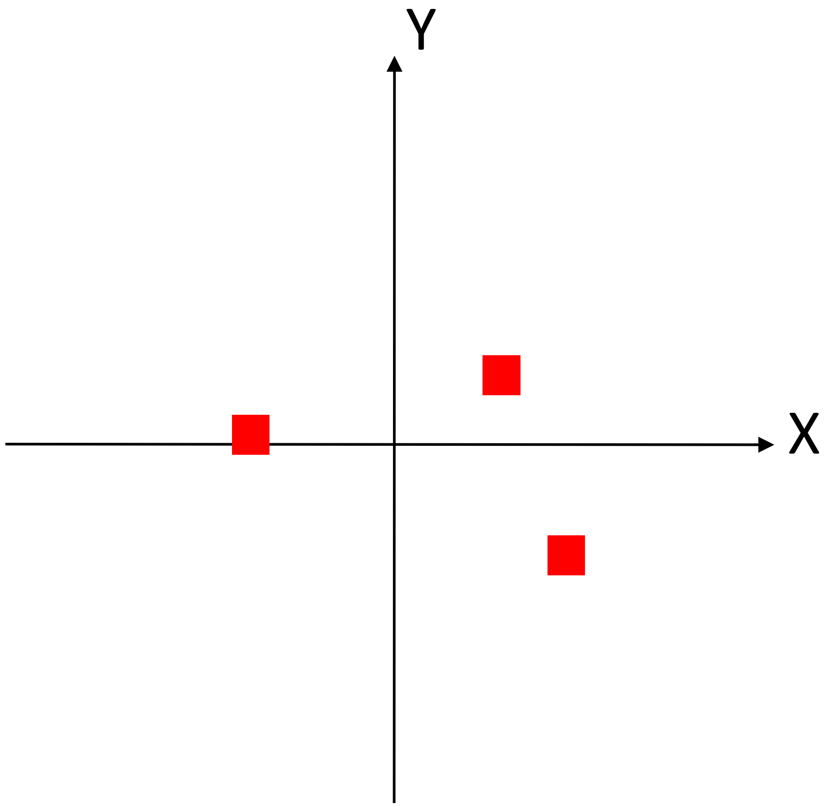


points in (X,Y, **F**)

Hierarchical Point Feature Learning

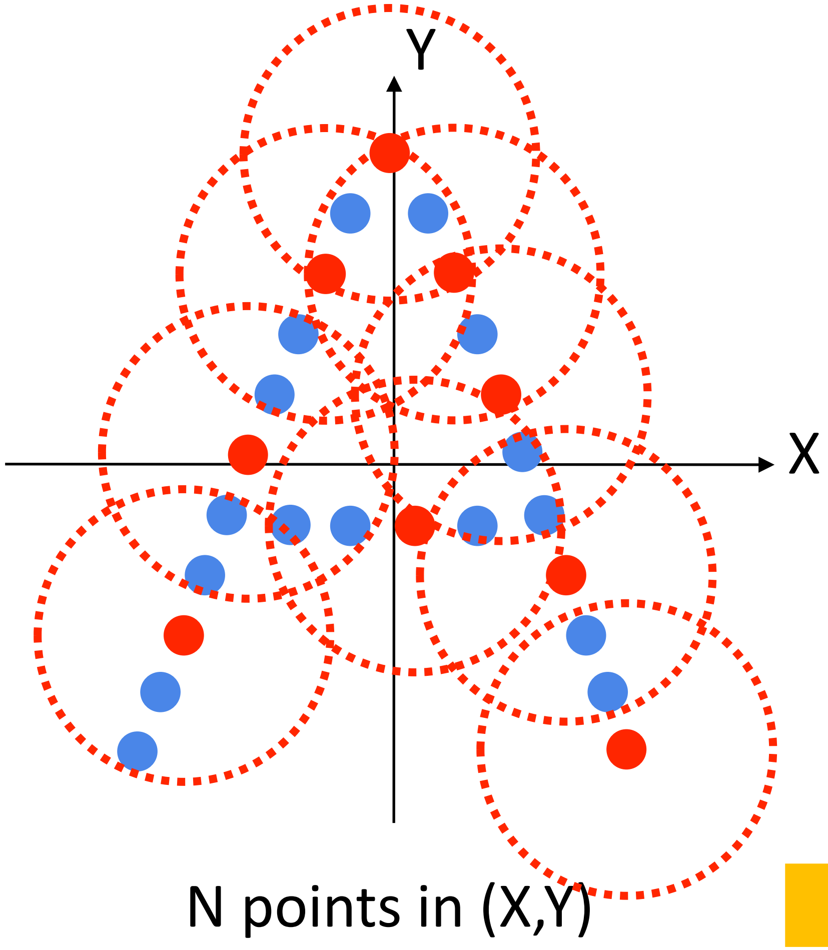


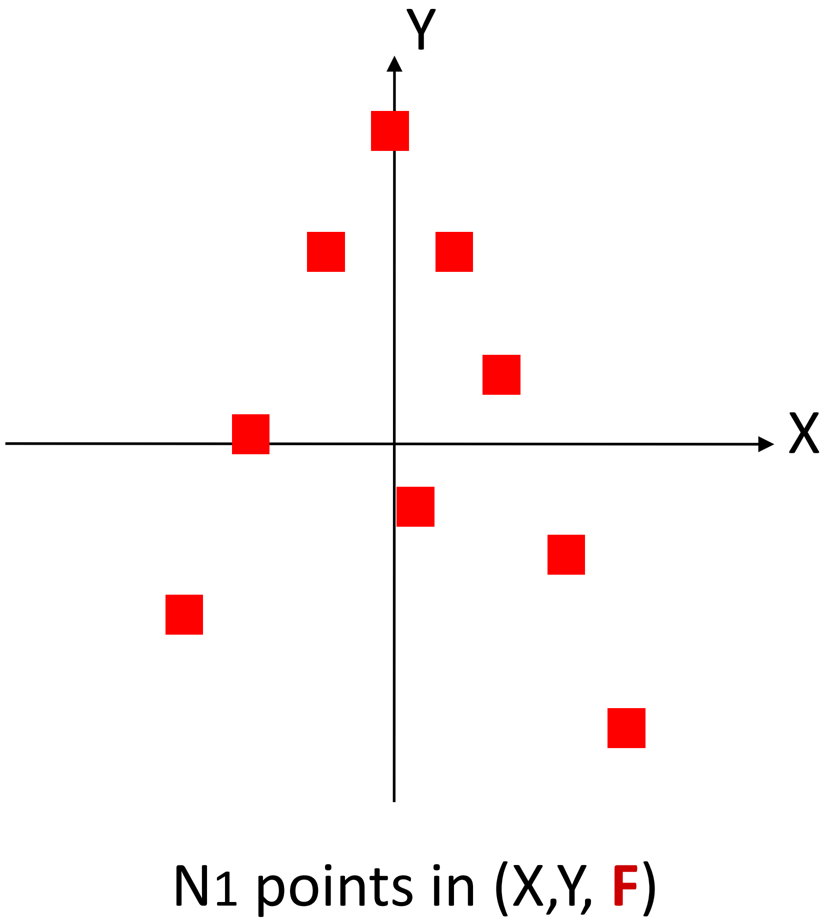
N points in (X,Y)



points in (X,Y, **F**)

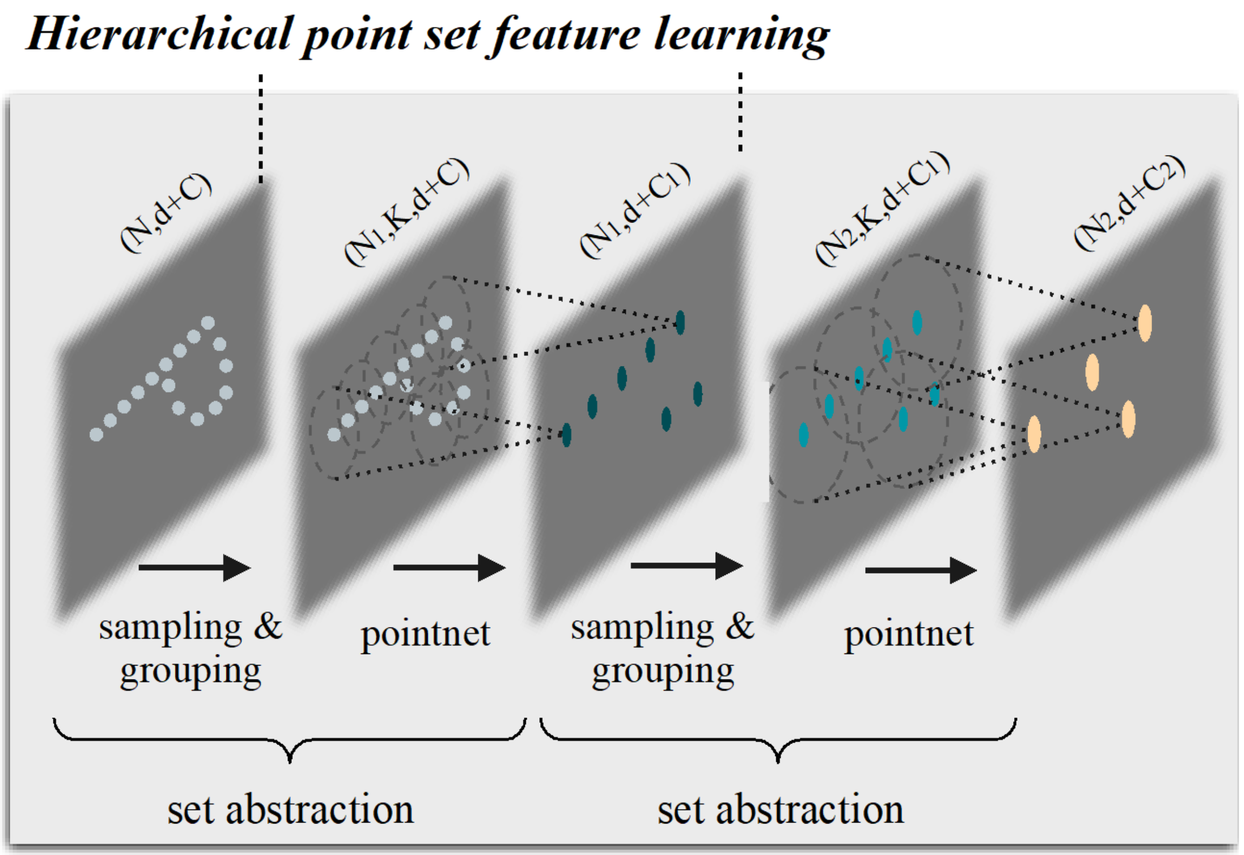
Hierarchical Point Feature Learning





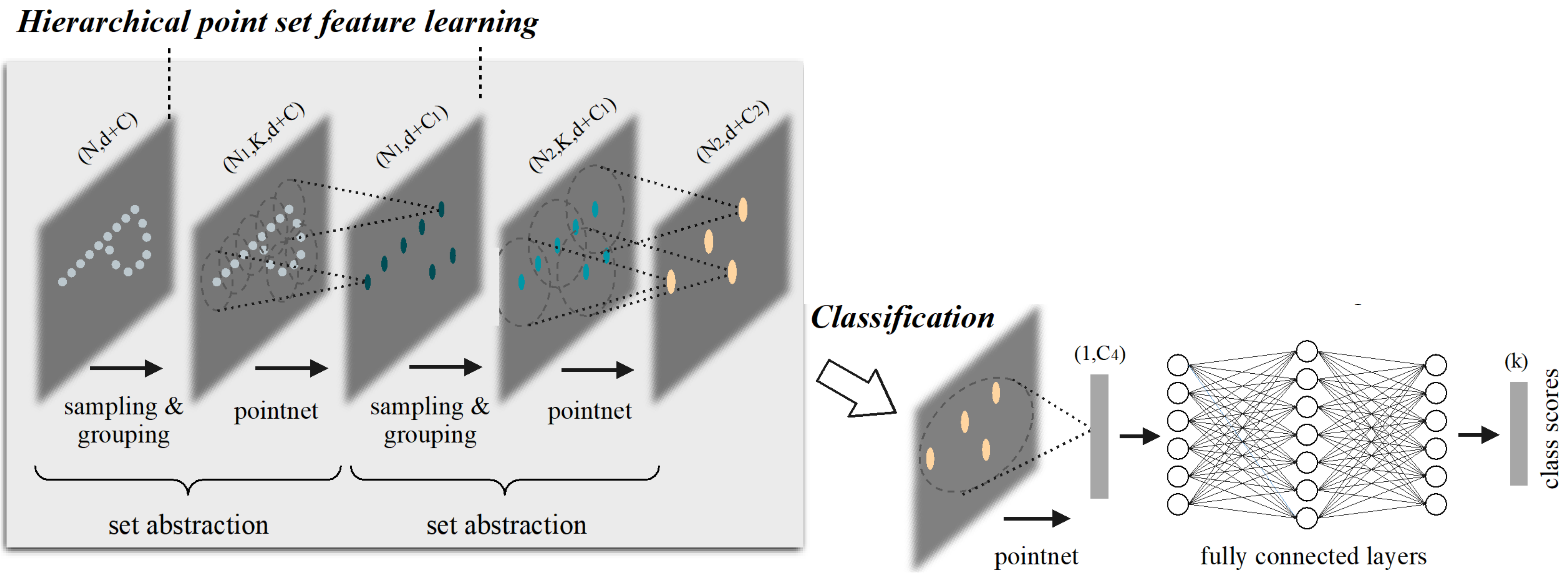
Set Abstraction: farthest point sampling + grouping + pointnet

PointNet++ for Classification and Segmentation

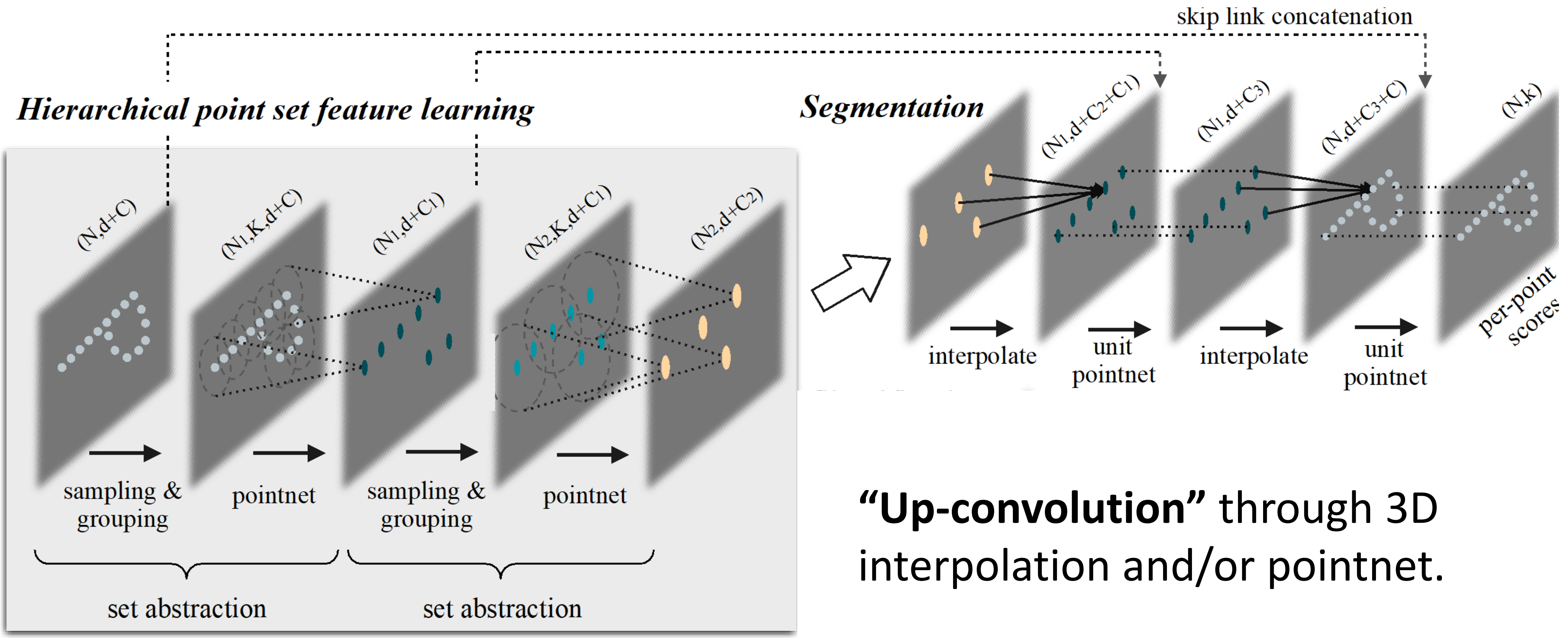


Caveat: Shouldn't feature dimensions from the lower layers affect connectivity at the higher layers?

PointNet++ for Classification and Segmentation



PointNet++ for Classification and Segmentation

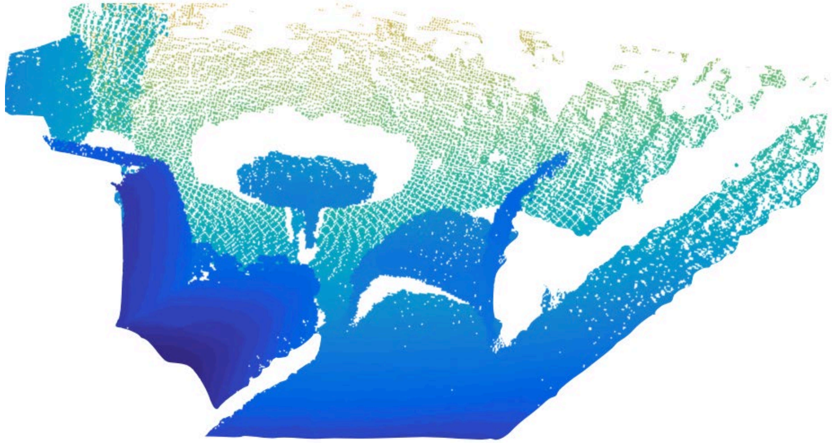


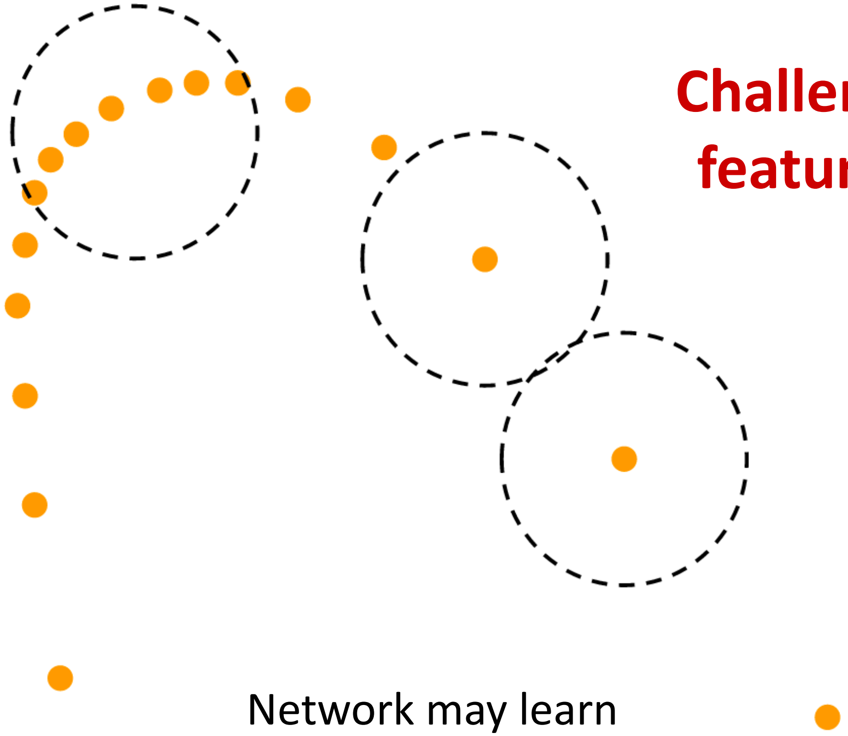
"Up-convolution" through 3D interpolation and/or pointnet.

Non-uniform Sampling Density in Point Clouds

Density variation is a common issue in 3D point cloud processing

- perspective effect, radial density variation, motion etc.

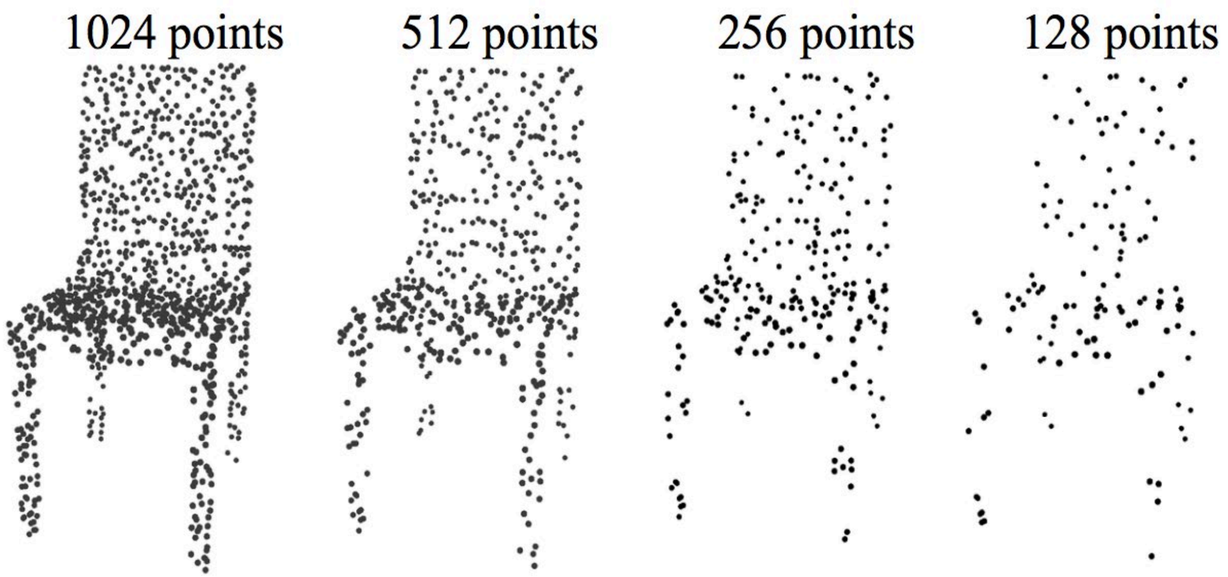


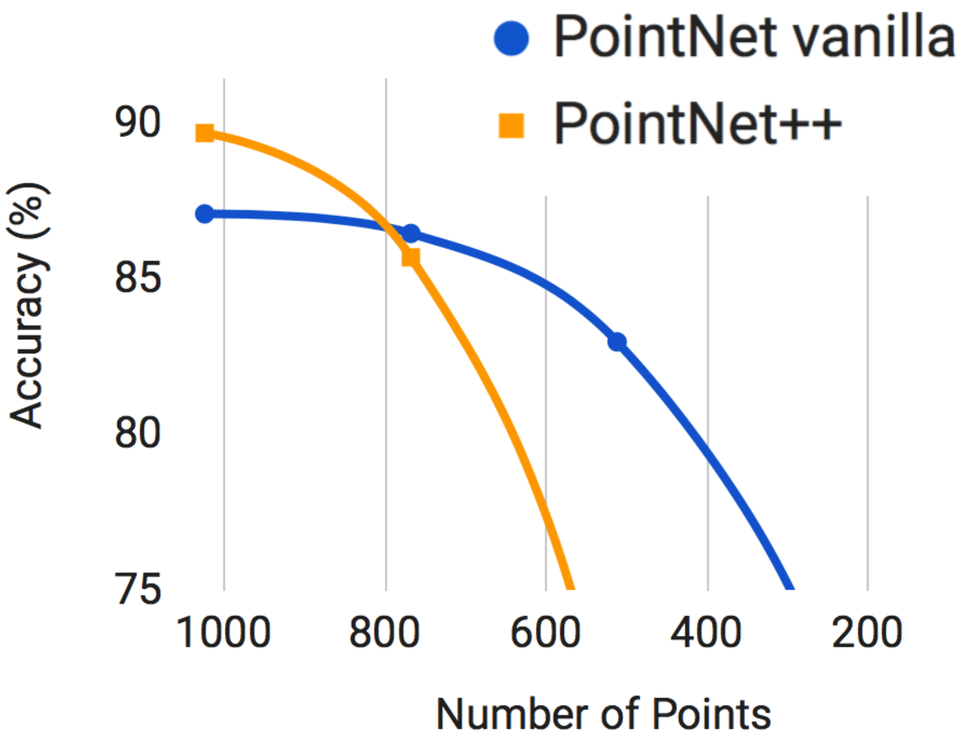


Challenge for local feature learning!

Network may learn sampling pattern – not underlying geometry

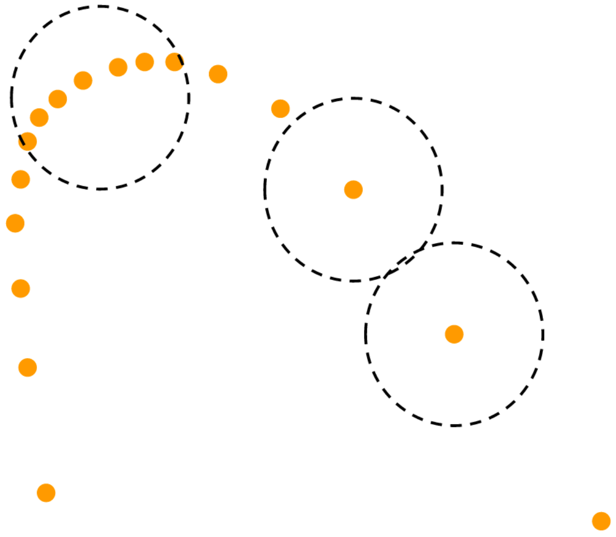
Density Variation Affects Hierarchy

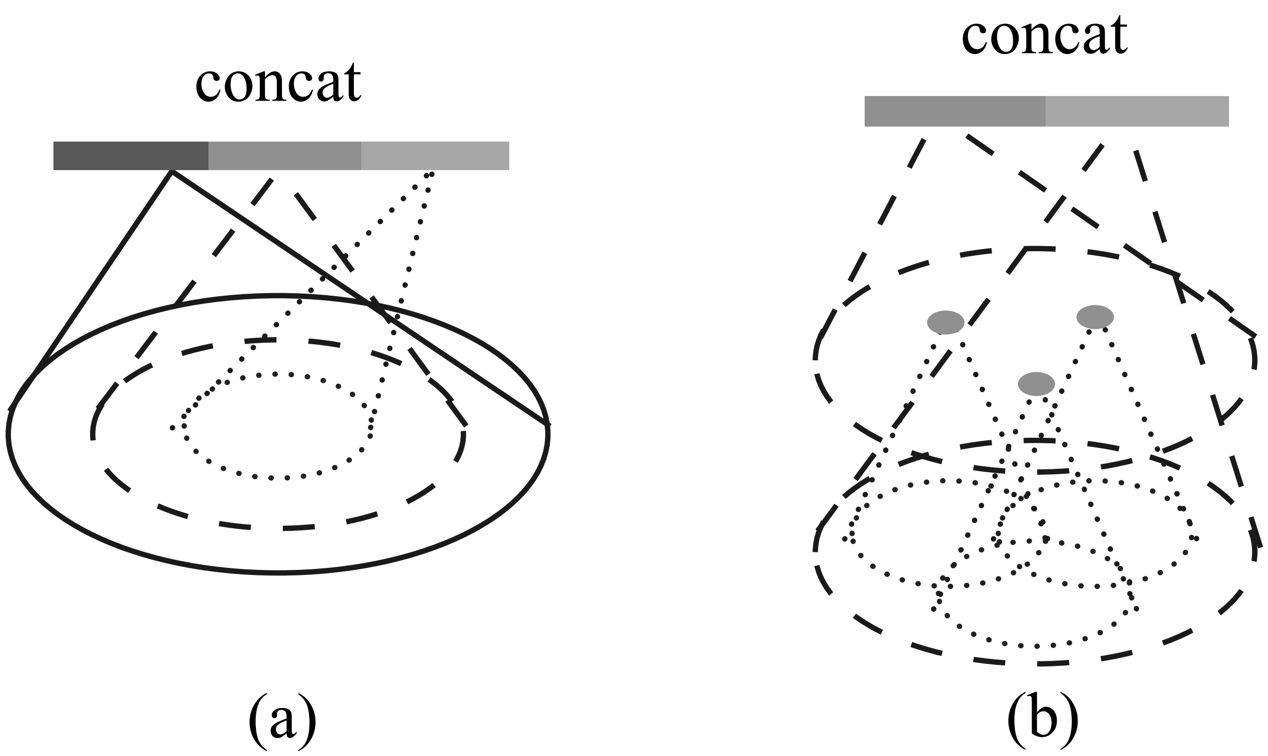




Small kernels suffer from varying densities!

Robust Learning Under Varying Sampling Density



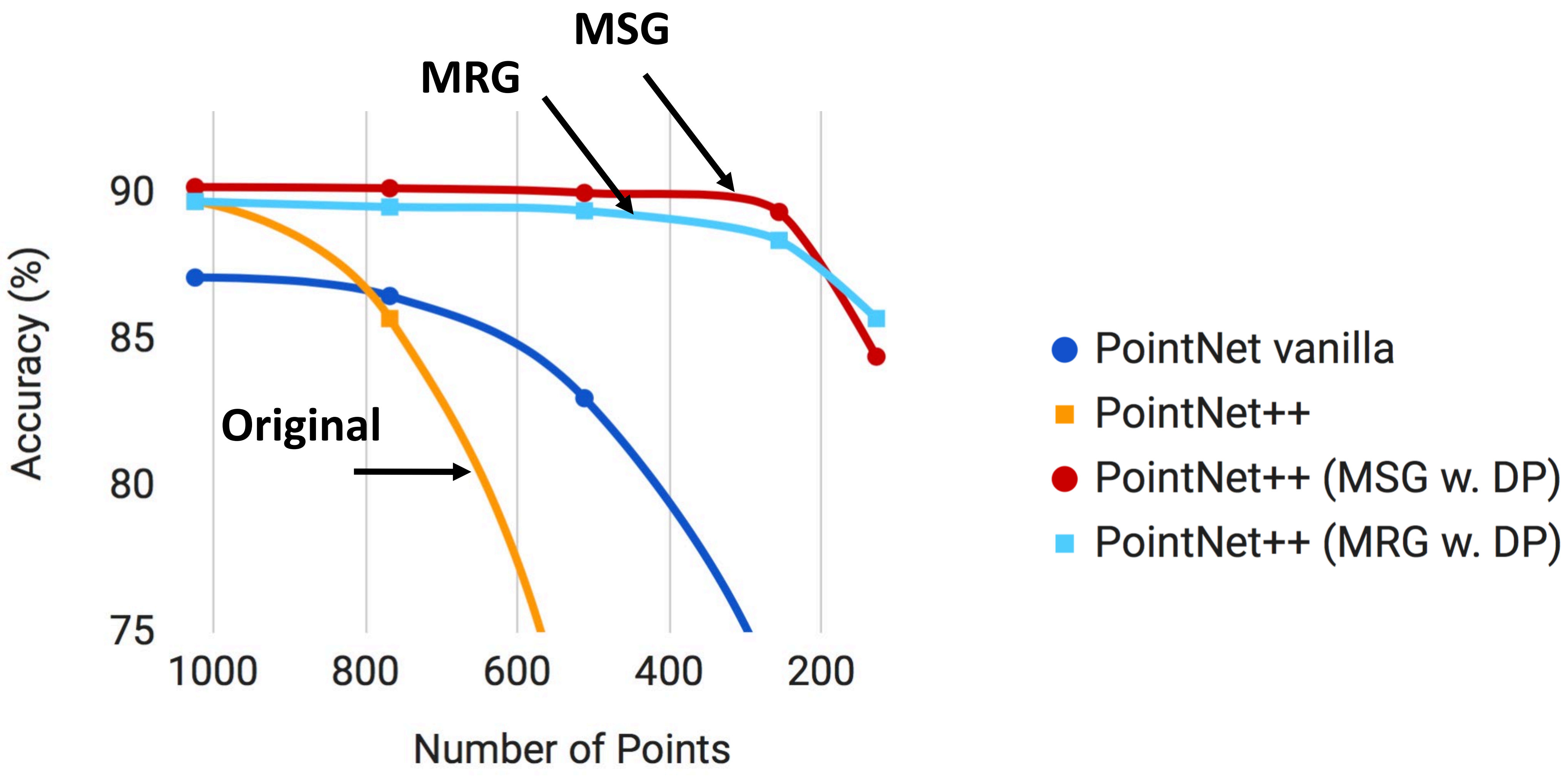


Multi-scale grouping (MSG)

Multi-res grouping (MRG)

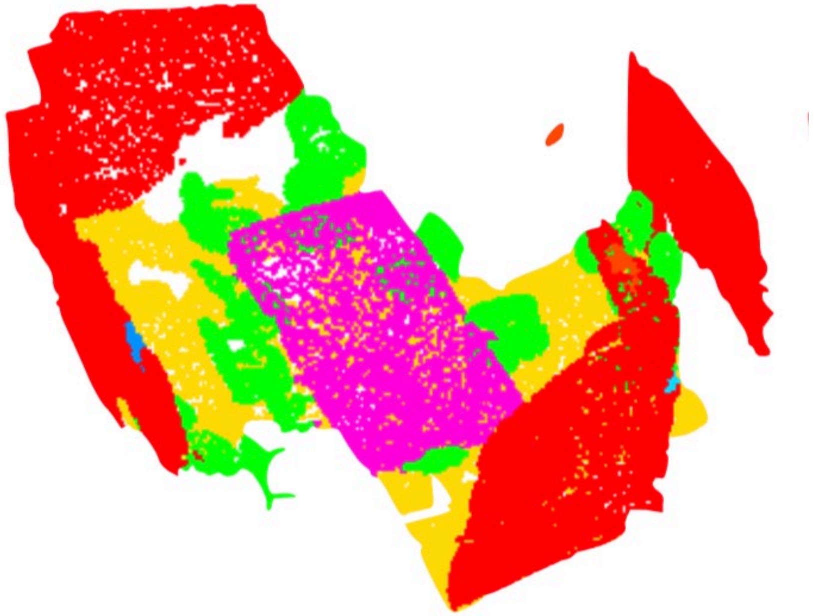
During Training: input point dropout with random dropout ratio

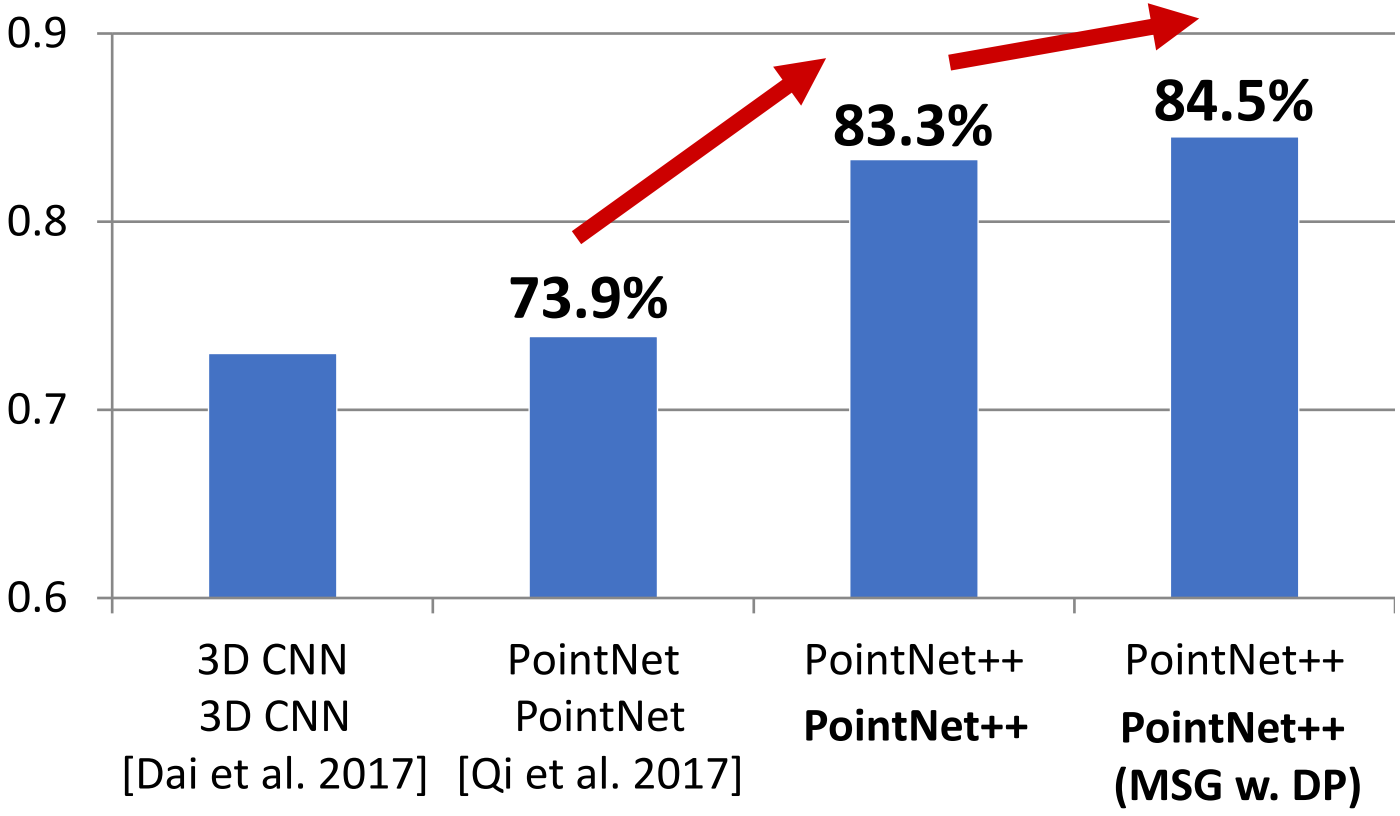
Robust Learning Under Varying Sampling Density



PointNet++ Results: Scene Parsing

Better accuracy with hierarchical learning.

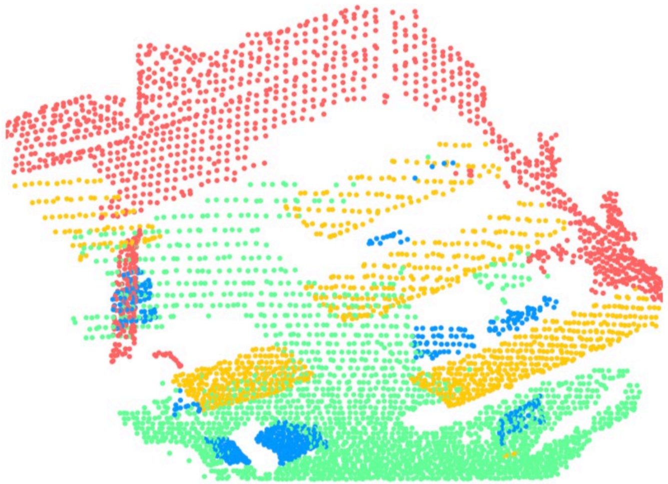


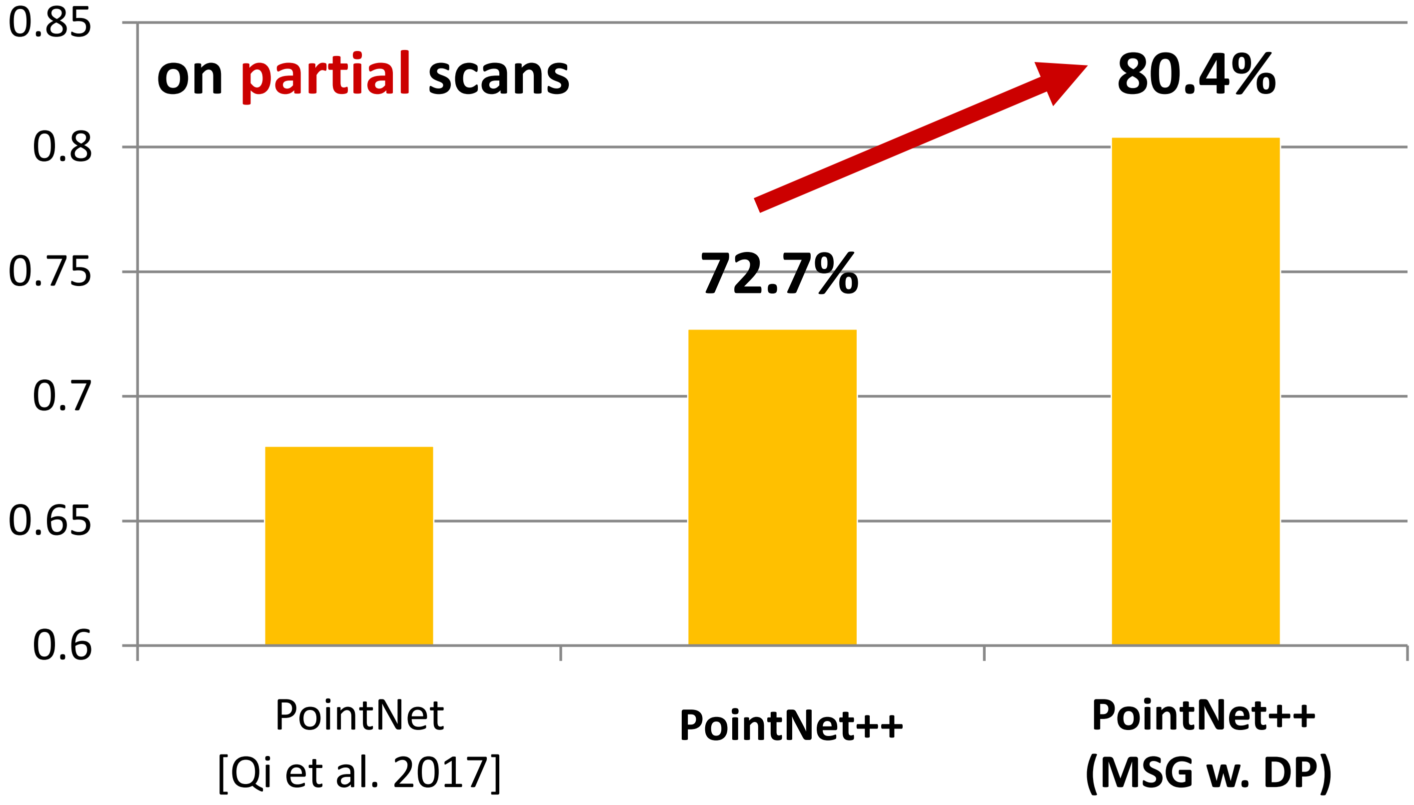


dataset: ScanNet; metric: per-point semantic classification accuracy (%)

PointNet++ Results: Scene Parsing

Robust layers for non-uniform densities (MSG) help a lot.

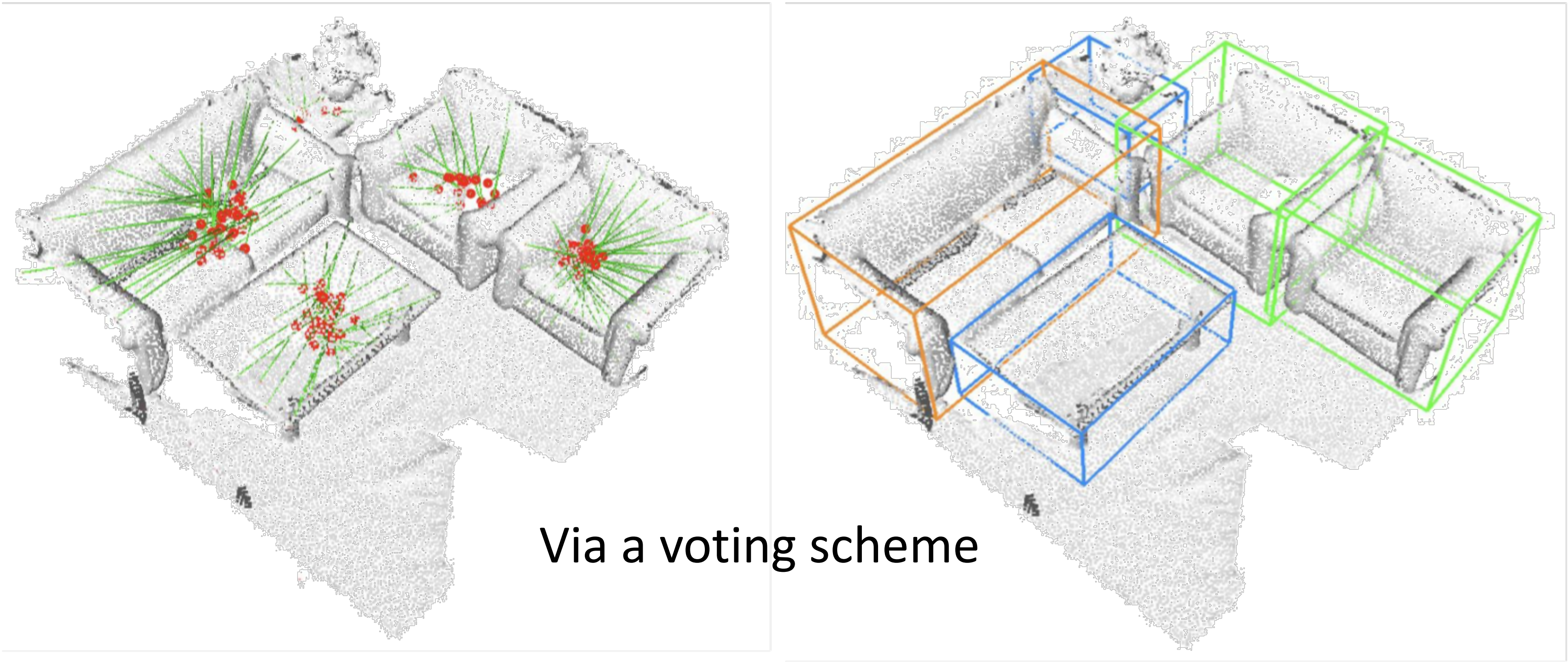




dataset: ScanNet; metric: per-point semantic classification accuracy (%)

Object Detection in Point Clouds

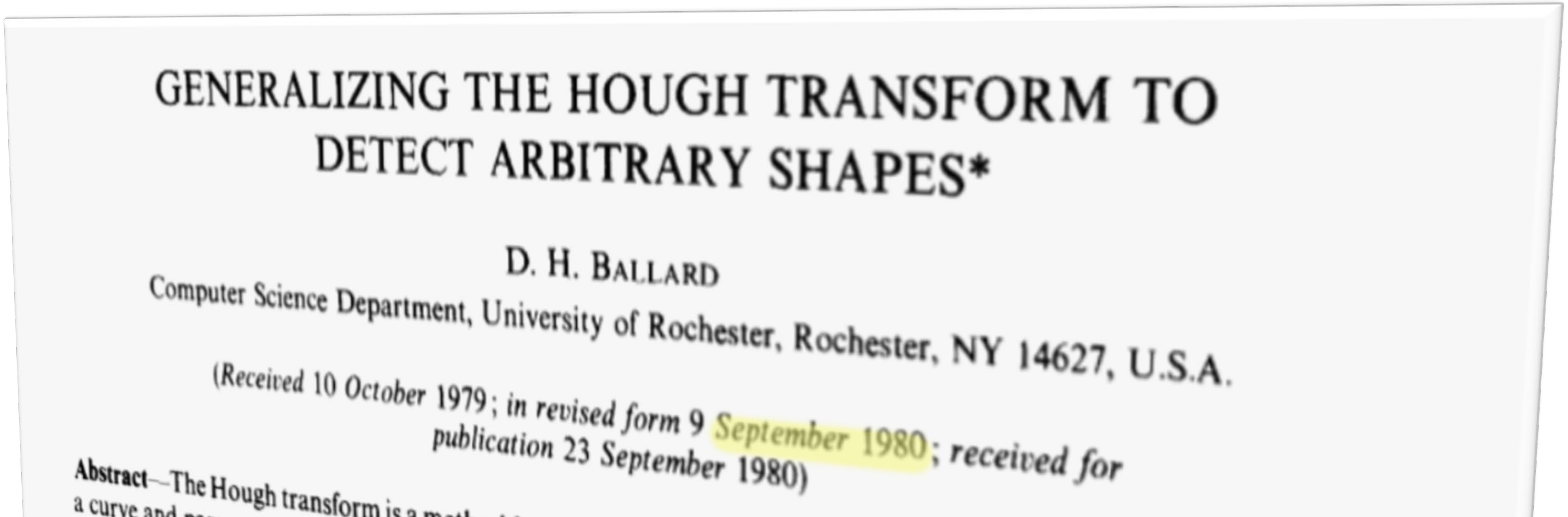
Point Cloud Object Amodal Bounding Box Detection



Via a voting scheme

[Charles R. Qi, Or Litany, Kaiming He, Leonidas J. Guibas.
Deep Hough Voting for 3D Object Detection in Point Clouds. ICCV '19]

Generalized Hough Transform

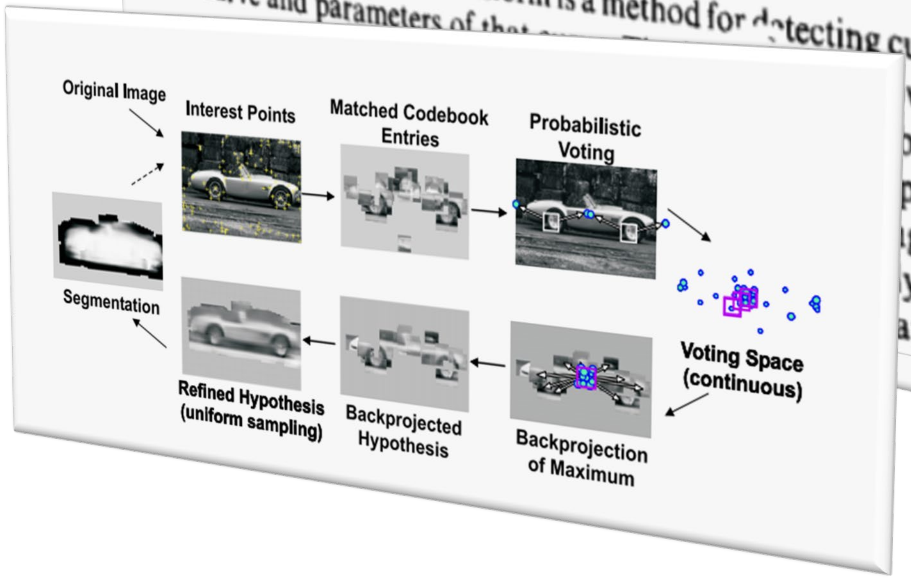


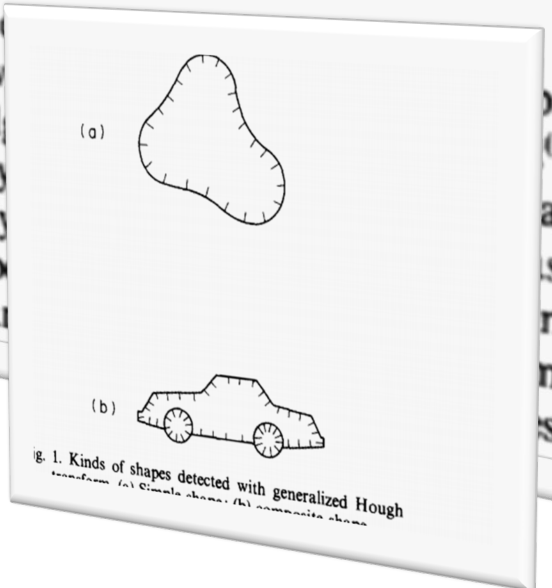
GENERALIZING THE HOUGH TRANSFORM TO DETECT ARBITRARY SHAPES*

D. H. BALLARD
Computer Science Department, University of Rochester, Rochester, NY 14627, U.S.A.

(Received 10 October 1979; in revised form 9 September 1980; received for publication 23 September 1980)

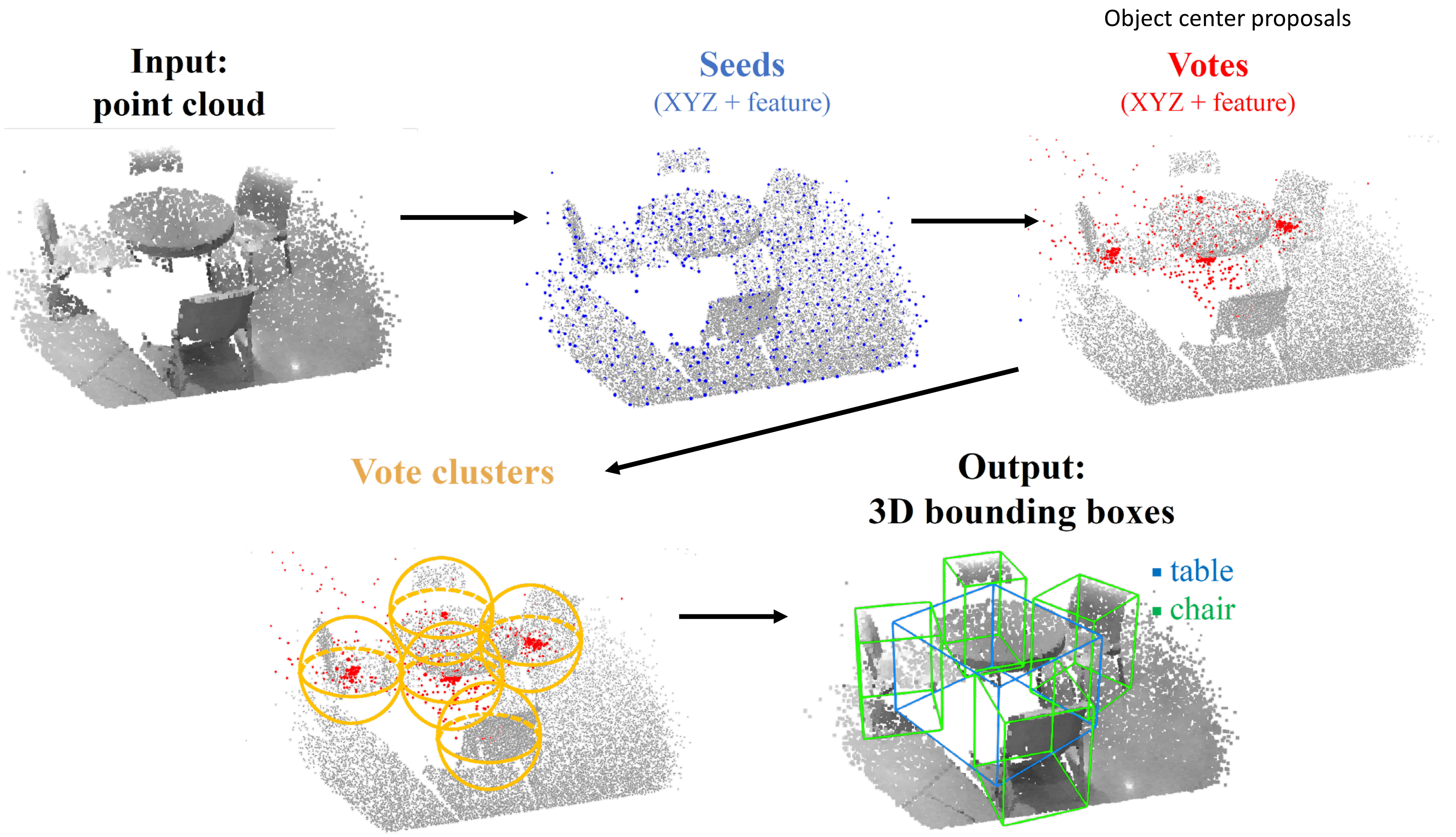
Abstract—The Hough transform is a method for detecting curves by
a curve and parameters of that





ig. 1. Kinds of shapes detected with generalized Hough

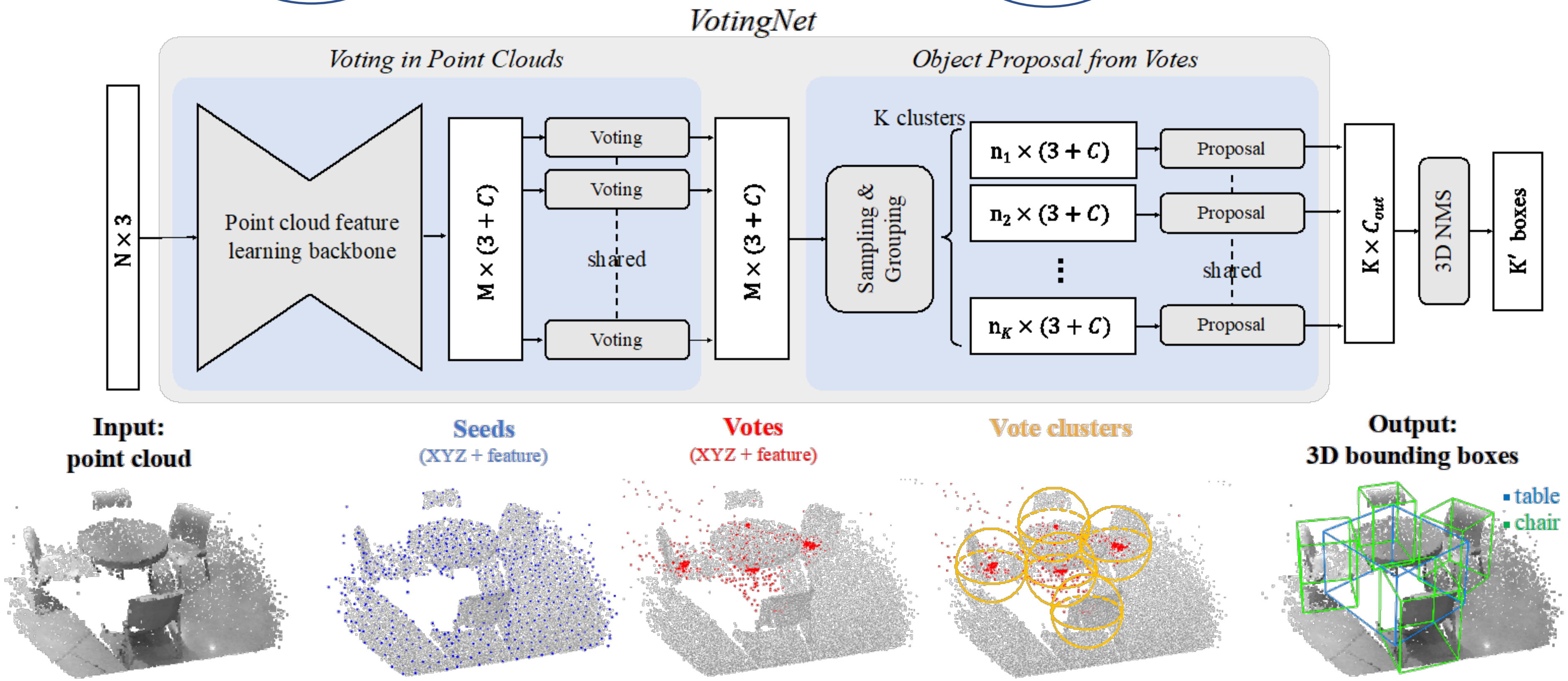
Deep Hough Voting – A Two-Stage Approach



VoteNet – A Two-Stage Approach

Pointnet++

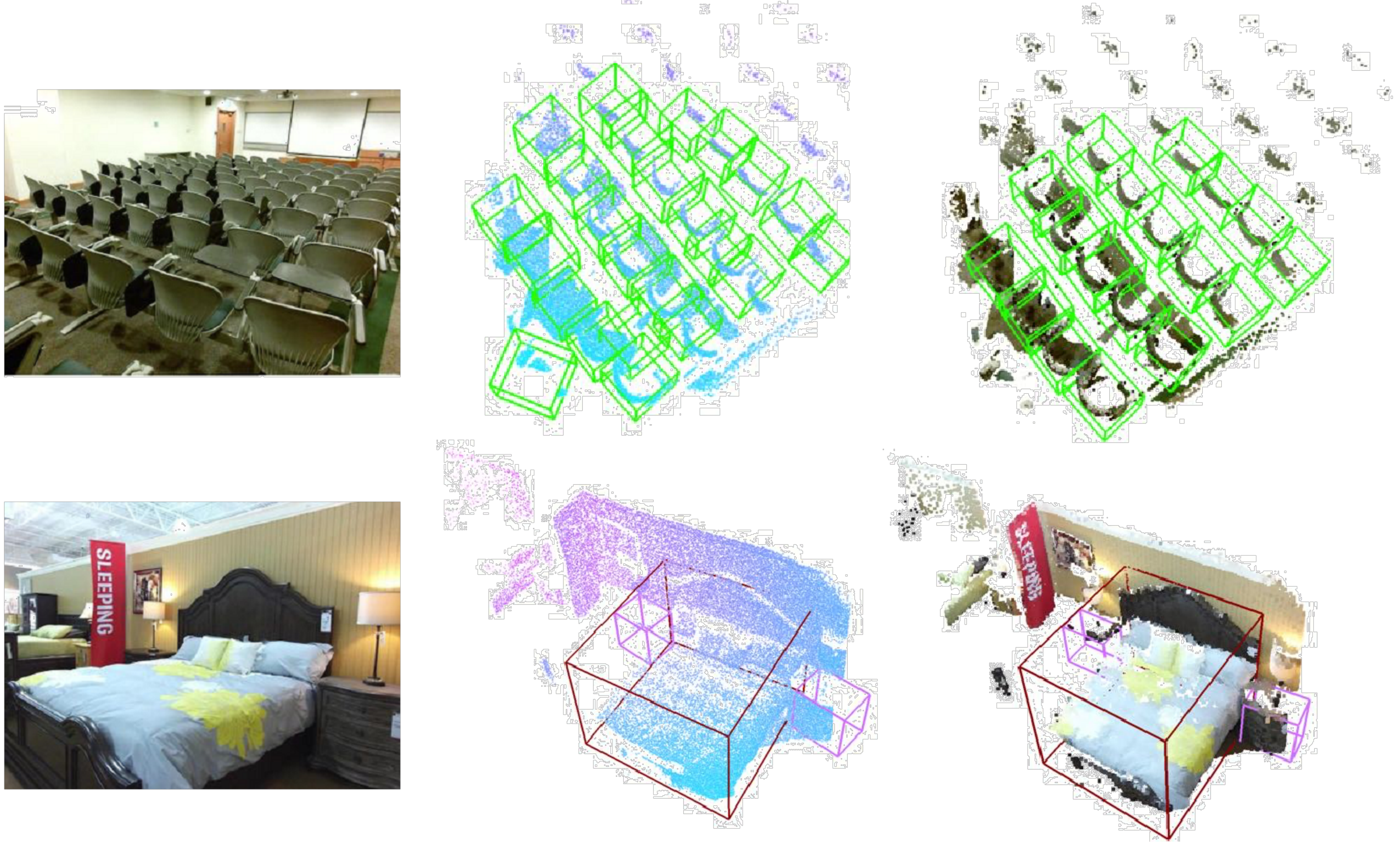
Pointnet++



A capsule network in disguise …

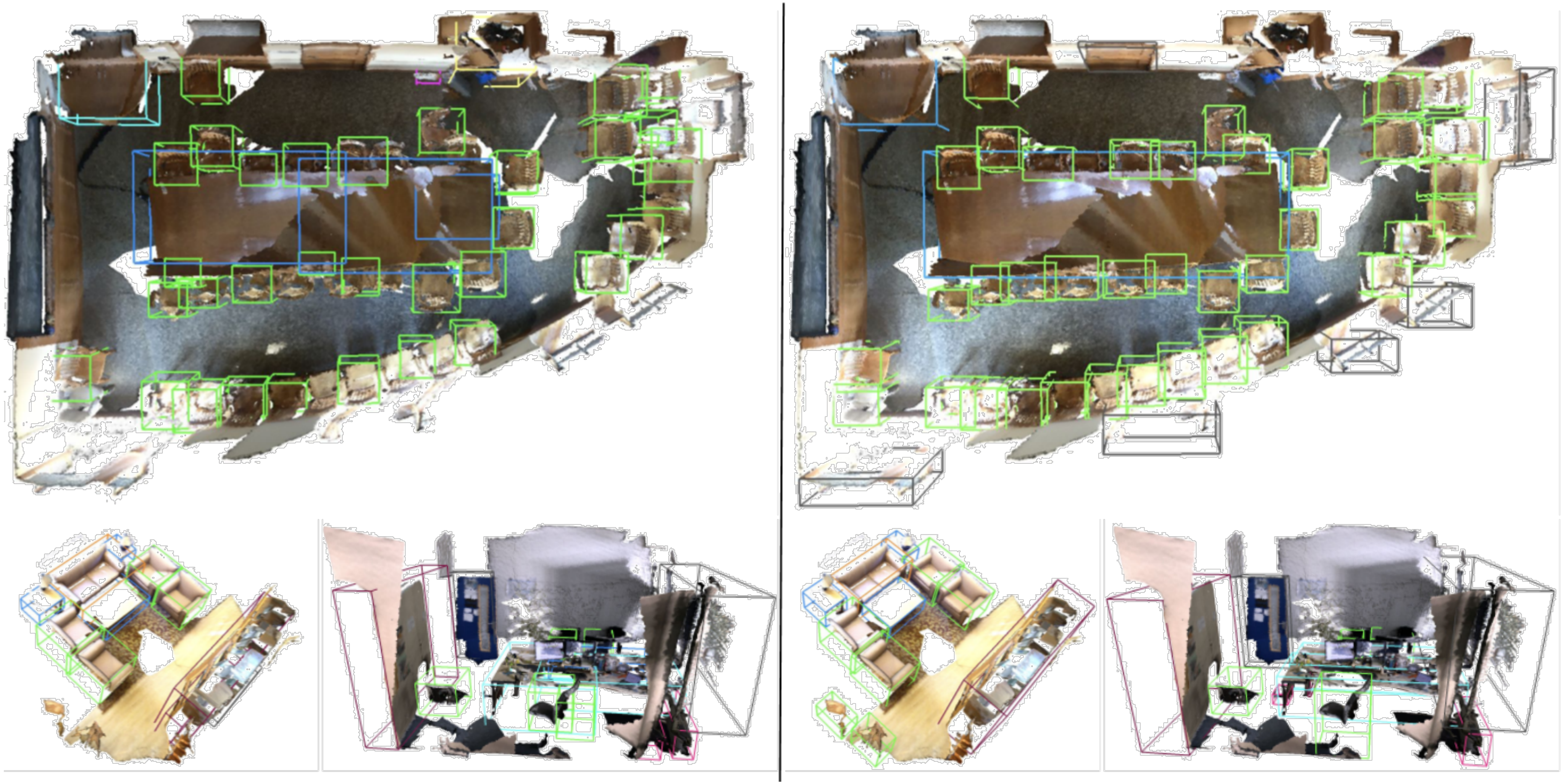
Results on SUN RGB-D

Image of the scene | VotingNet prediction | Ground truth

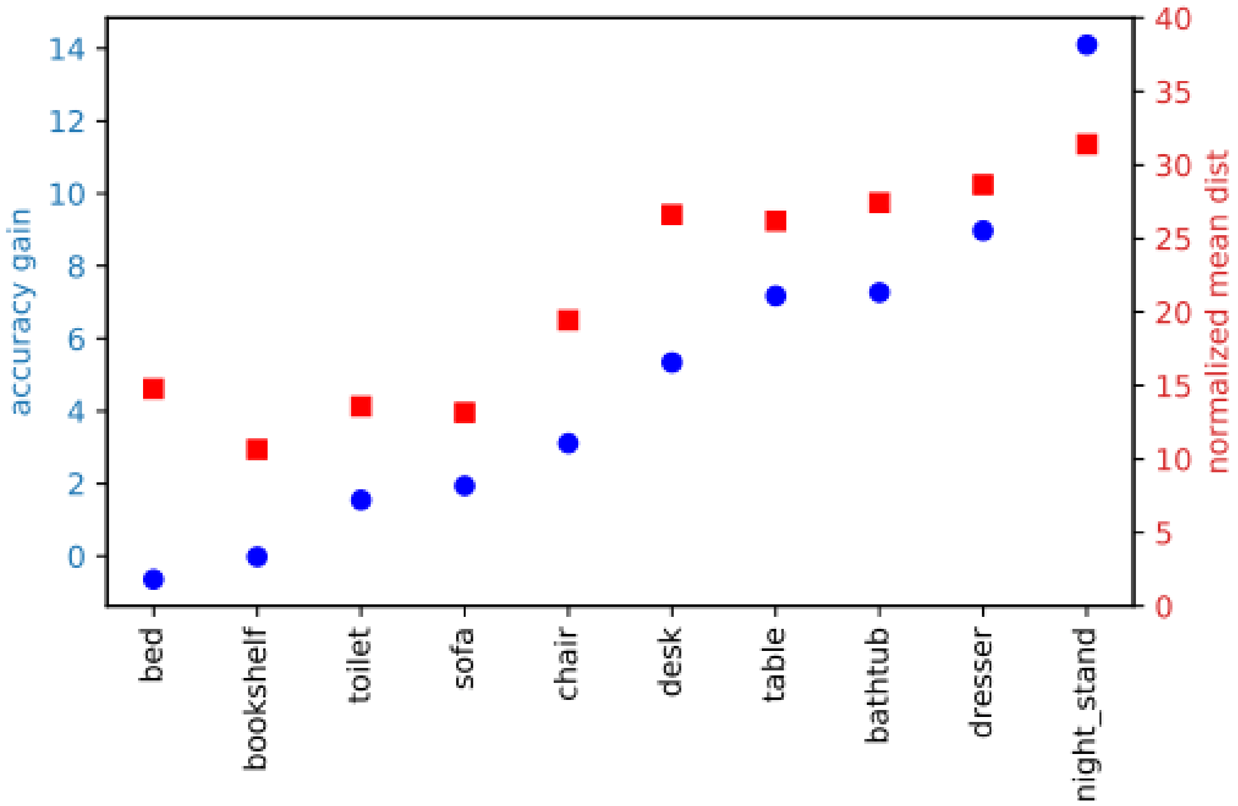


Results on ScanNet

VotingNet prediction | Ground truth



Performance Gain Vs. Centroid Distance



Quantitative Results

SUN RGB-D

Deep sliding shapes
Clouds of oriented gradients

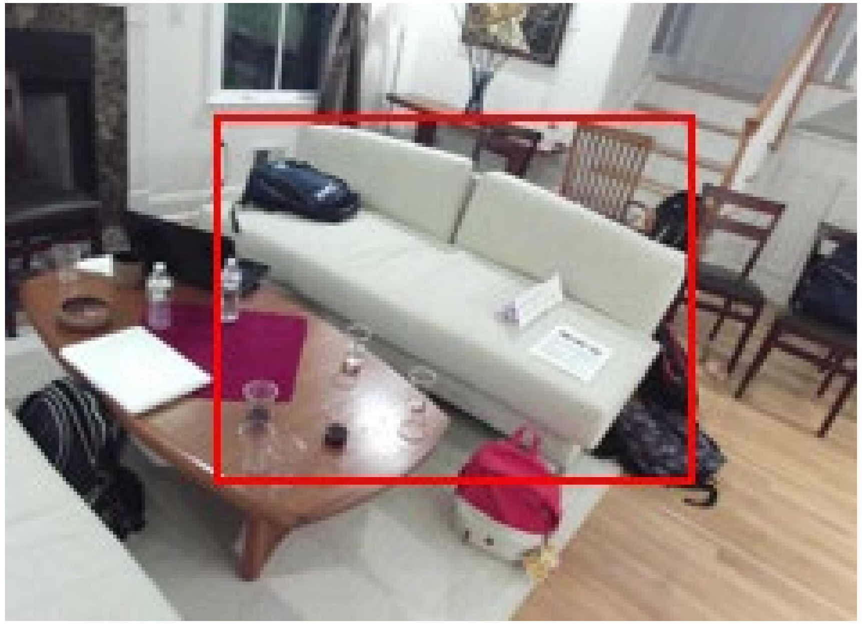
Frustum pointnet

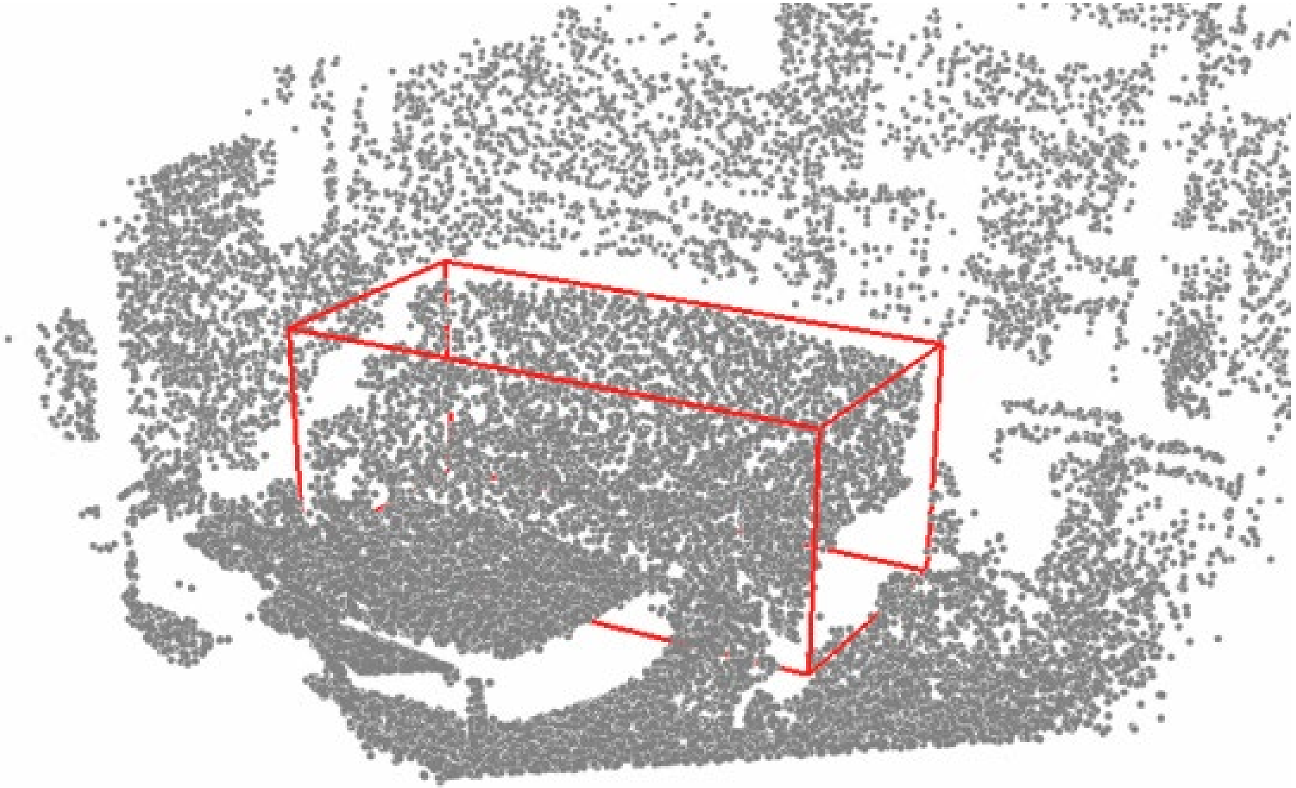
	Input	bathtub	bed	bookshelf	chair	desk	dresser	nightstand	sofa	table	toilet	mAP
DSS [37]	Geo + RGB	44.2	78.8	11.9	61.2	20.5	6.4	15.4	53.5	50.3	78.9	42.1
COG [33]	Geo + RGB	58.3	63.7	31.8	62.2	**45.2**	15.5	27.4	51.0	**51.3**	70.1	47.6
2D-driven [17]	Geo + RGB	43.5	64.5	31.4	48.3	27.9	25.9	41.9	50.4	37.0	80.4	45.1
F-PointNet [30]	Geo + RGB	43.3	81.1	**33.3**	64.2	24.7	**32.0**	58.1	61.1	51.1	**90.9**	54.0
VotingNet (ours)	Geo only	**74.4**	**83.0**	28.8	**75.3**	22.0	29.8	**62.2**	**64.0**	47.3	90.1	**57.7**

ScanNetV2

	Input	mAP@0.25	mAP@0.5
DSS [42, 12]	Geo + RGB	15.2	6.8
MRCNN 2D-3D [11, 12]	Geo + RGB	17.3	10.5
F-PointNet [34, 12]	Geo + RGB	19.8	10.8
GSPN [54]	Geo + RGB	30.6	17.7
3D-SIS [12]	Geo + 1 view	35.1	18.7
3D-SIS [12]	Geo + 3 views	36.6	19.0
3D-SIS [12]	Geo + 5 views	40.2	22.5
3D-SIS [12]	Geo only	25.4	14.6
VoteNet (ours)	Geo only	**58.6**	**33.5**

Images and Point Clouds are Complementary



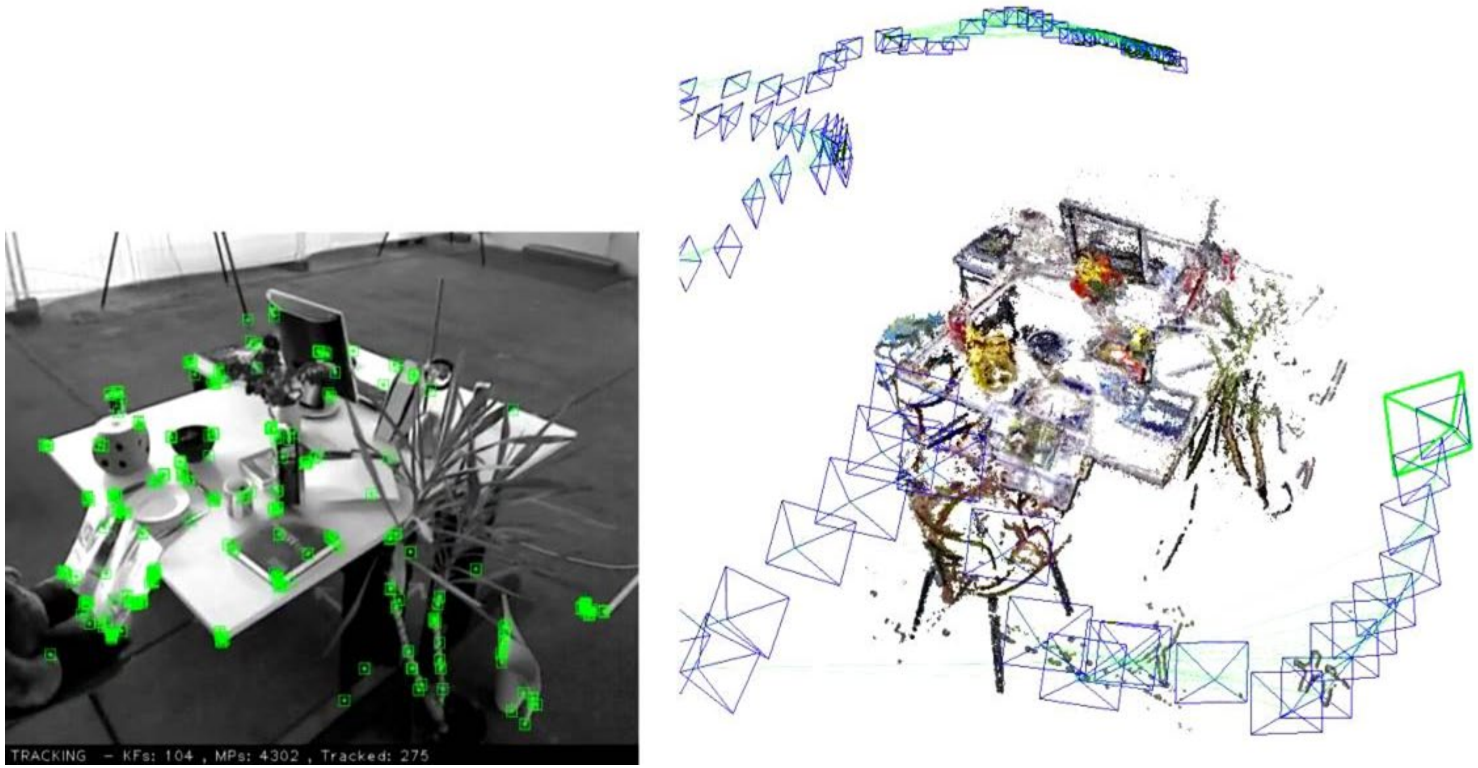


- High resolution
- See "blind regions"

- Absolute depth and scale

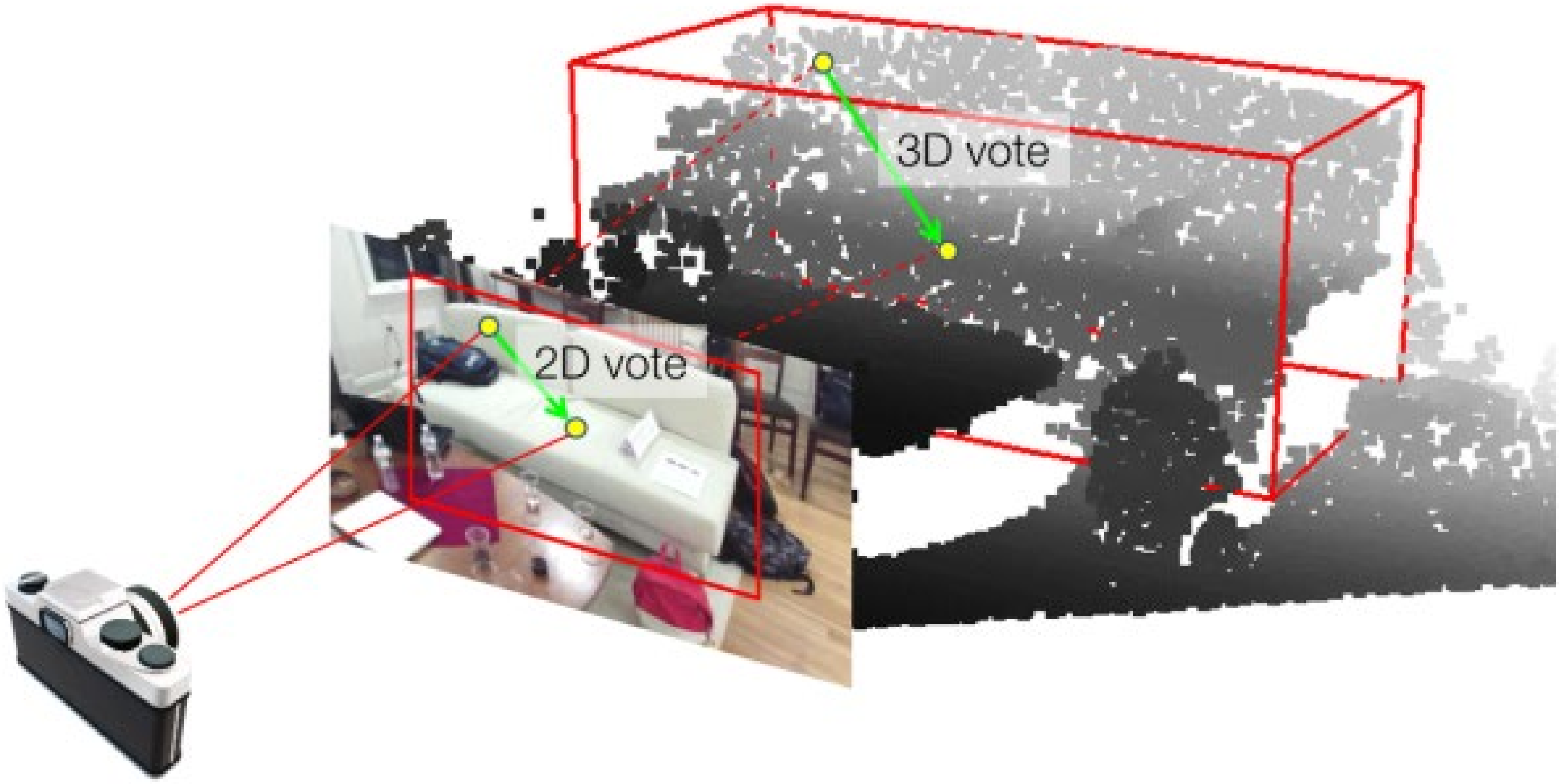
3D Detection with Sparse Points

Application: 3D detection from monocular video, using sparse SLAM keypoints.



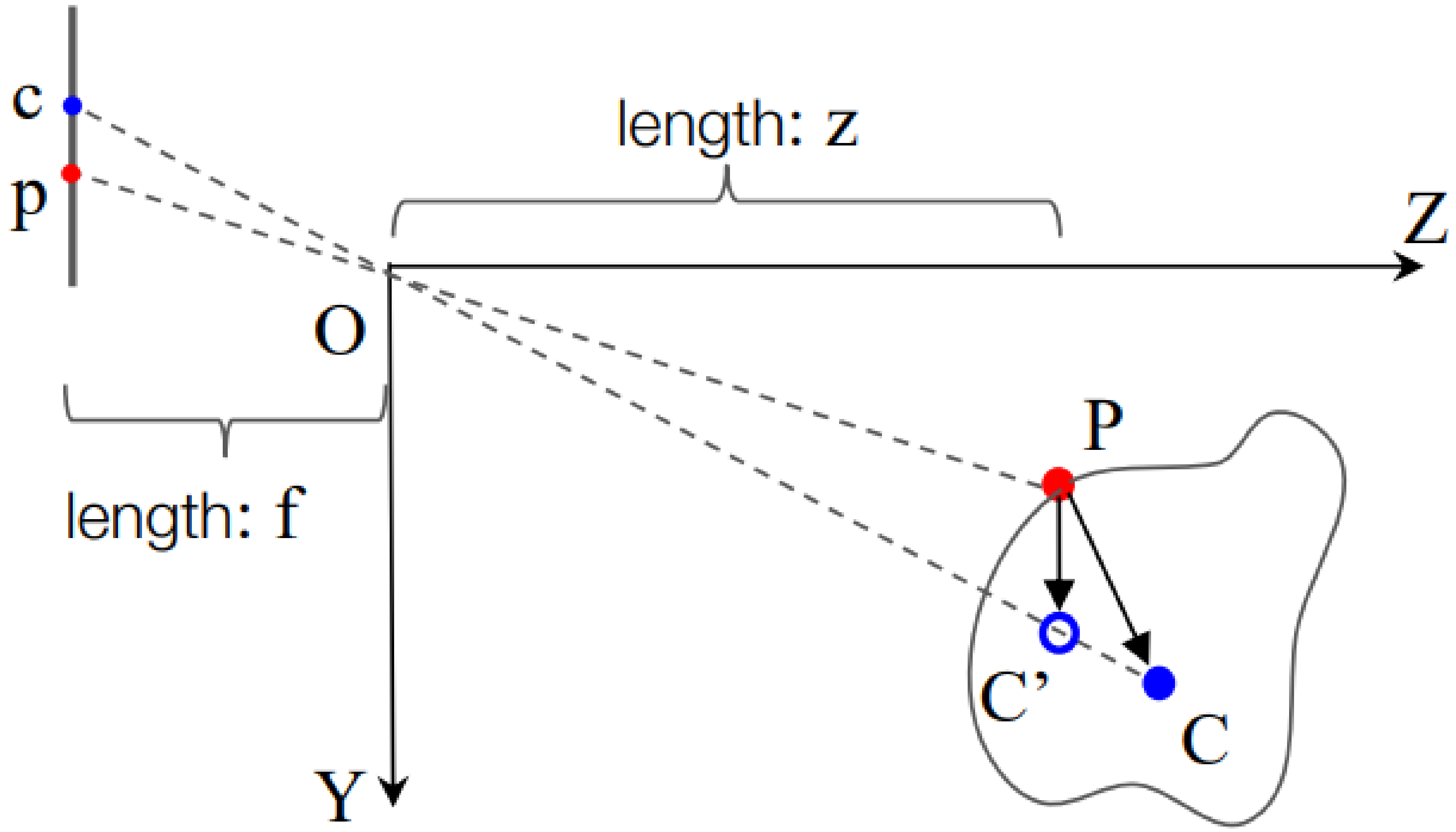
Picture: ORB-SLAM results

Basic idea: *ImVoteNet*



C. Qi, X. Chen, O. Litany, L.J. Guibas CVPR 2020

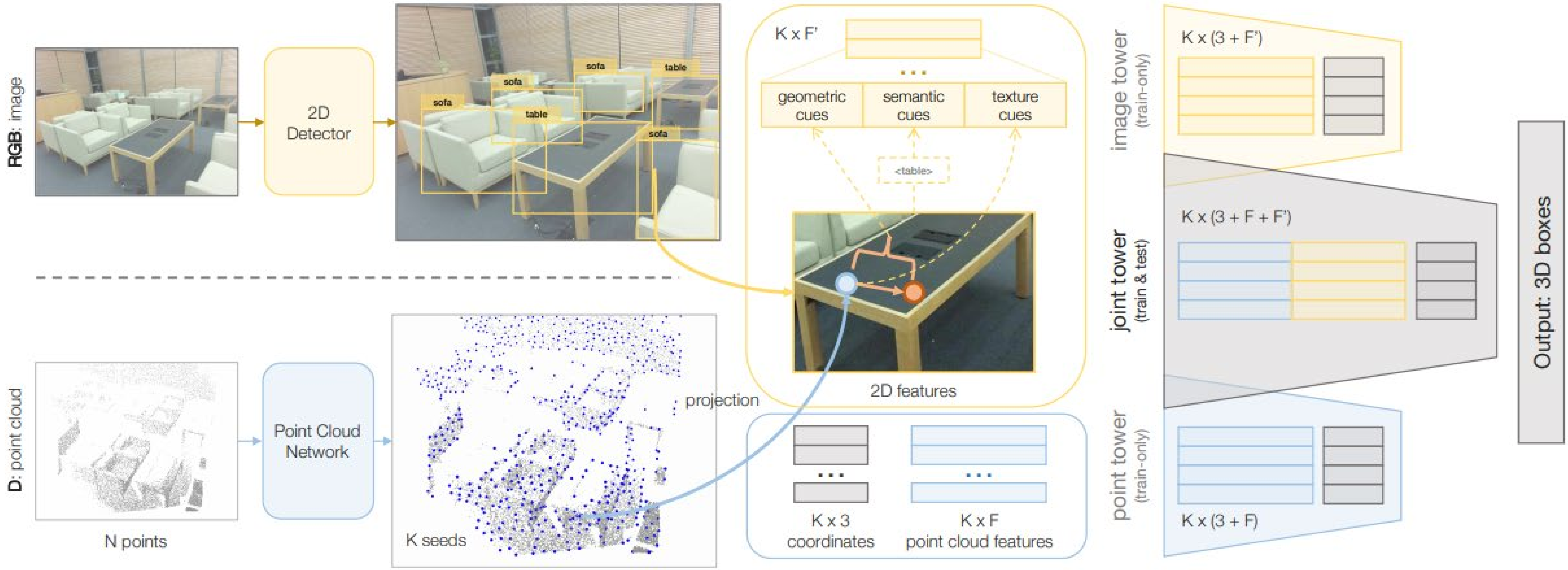
Lifting 2D features



displacement and ray direction:

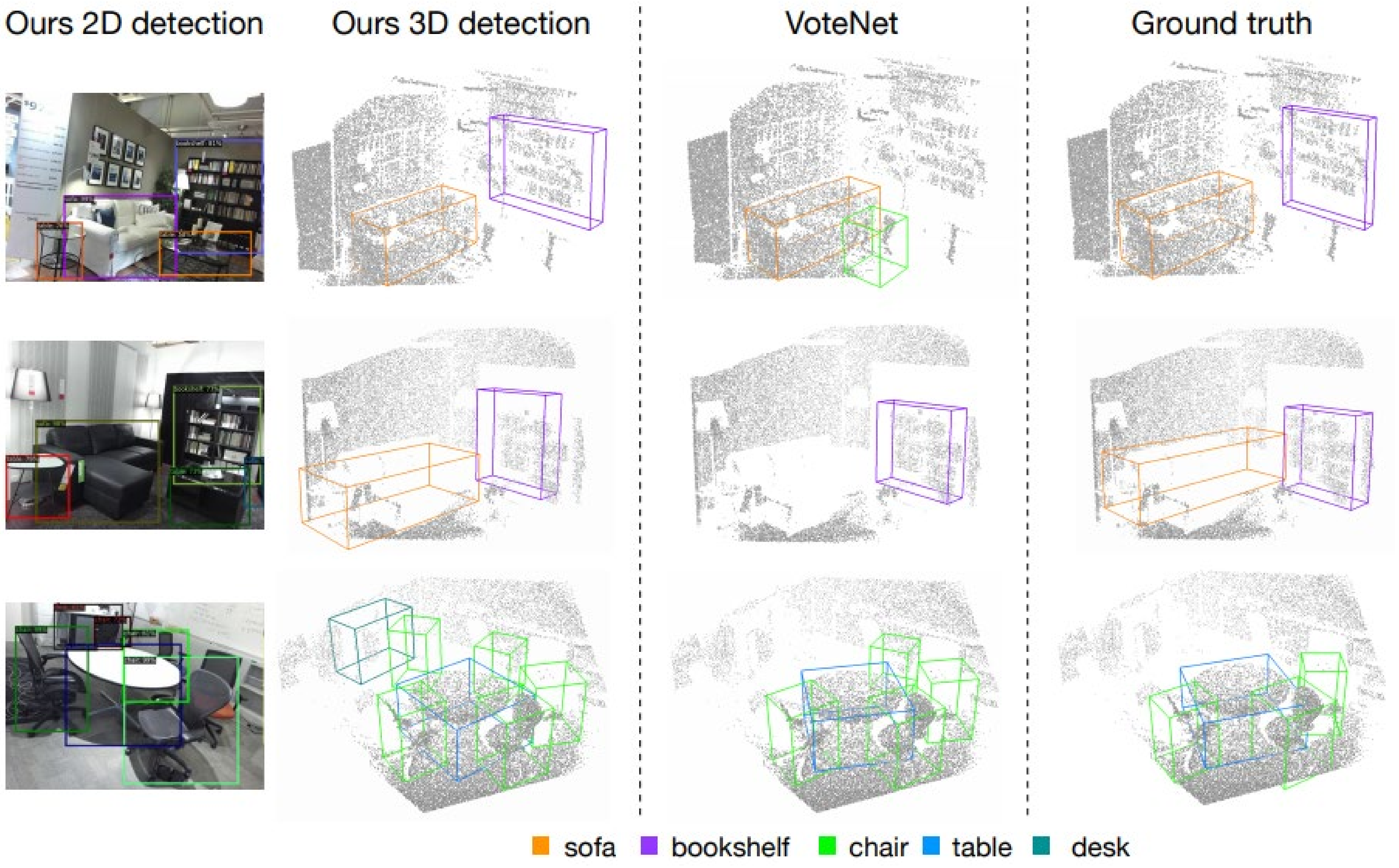
$$(\frac{\Delta u}{f} z_1, \frac{\Delta v}{f} z_1, \frac{\overrightarrow{OC'}}{\|\overrightarrow{OC'}\|}).$$

ImVoteNet Architecture



Results on SUN RGB-D

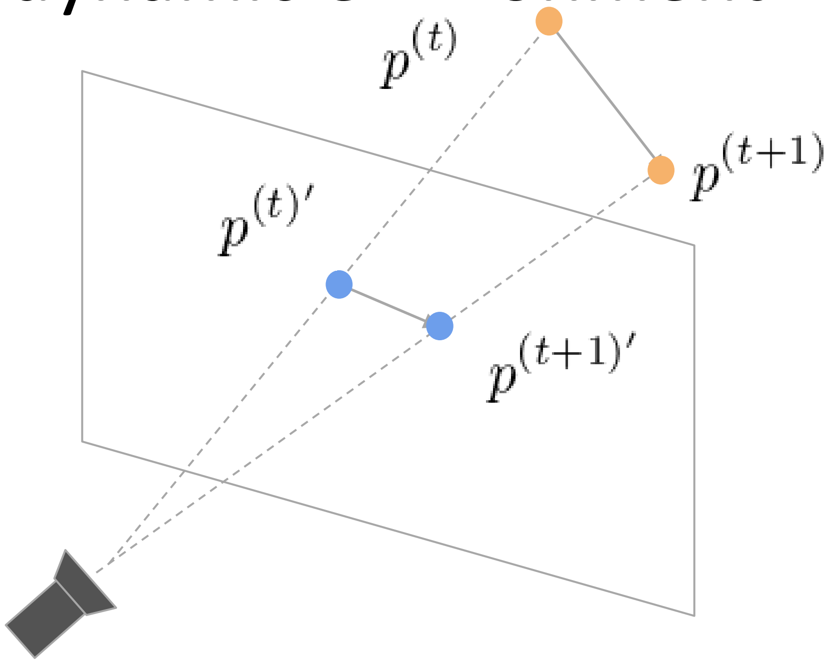
methods	RGB	bathtub	bed	bookshelf	chair	desk	dresser	nightstand	sofa	table	toilet	mAP
DSS [39]	✓	44.2	78.8	11.9	61.2	20.5	6.4	15.4	53.5	50.3	78.9	42.1
COG [34]	✓	58.3	63.7	31.8	62.2	**45.2**	15.5	27.4	51.0	**51.3**	70.1	47.6
2D-driven [15]	✓	43.5	64.5	31.4	48.3	27.9	25.9	41.9	50.4	37.0	80.4	45.1
PointFusion [43]	✓	37.3	68.6	37.7	55.1	17.2	23.9	32.3	53.8	31.0	83.8	45.4
F-PointNet [29]	✓	43.3	81.1	33.3	64.2	24.7	32.0	58.1	61.1	51.1	**90.9**	54.0
VOTENET [28]	✗	74.4	83.0	28.8	75.3	22.0	29.8	62.2	64.0	47.3	90.1	57.7
+RGB	✓	70.0	82.8	27.6	73.1	23.2	27.2	60.7	63.7	48.0	86.9	56.3
+region feature	✓	71.7	86.1	34.0	74.7	26.0	34.2	64.3	66.5	49.7	88.4	59.6
IMVOTENET	✓	**75.9**	**87.6**	**41.3**	**76.7**	28.7	**41.4**	**69.9**	**70.7**	51.1	90.5	**63.4**


Ours 2D detection
Ours 3D detection
VoteNet
Ground truth
sofa
bookshelf
chair
table
desk

3D Motion in Point Clouds

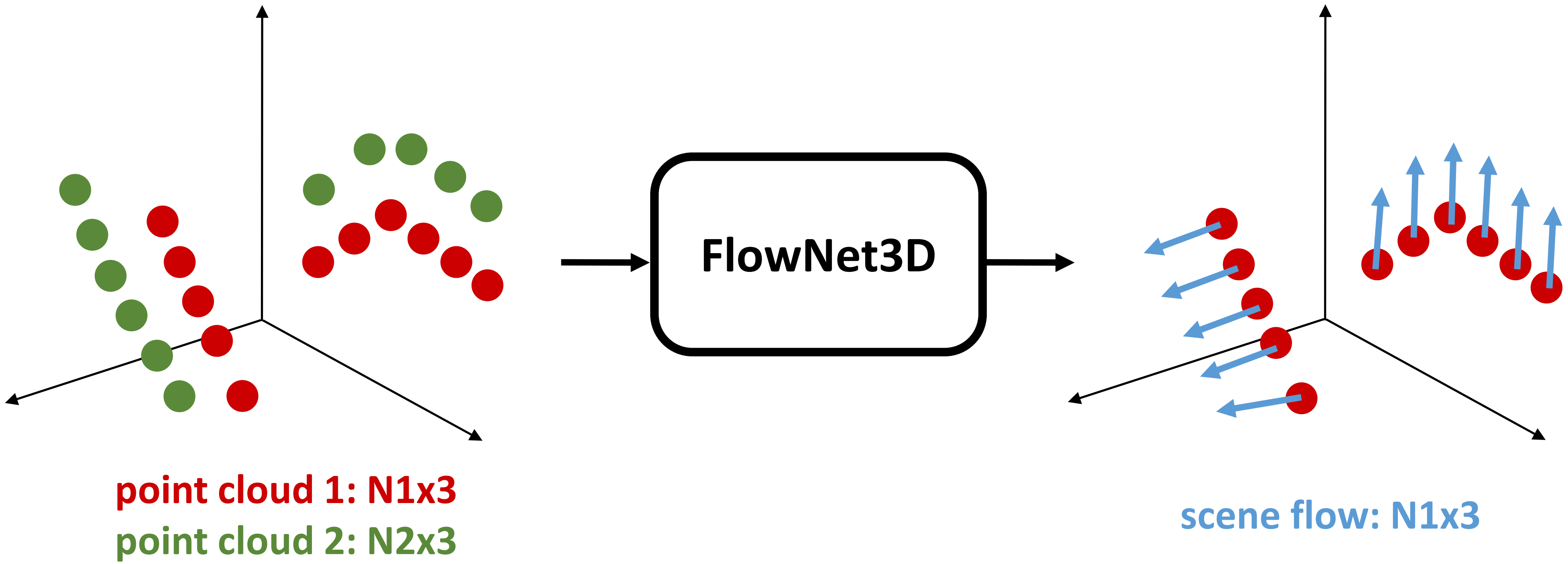
Scene Flow [Vedula et al. 1999]

- Scene flow: 3D motion field of points
- Optical flow is its projection to 2D image plane.
- Low-level understanding of a dynamic environment



Our Approach: FlowNet3D

- Directly learning scene flow in 3D point clouds, with 3D deep learning architectures.



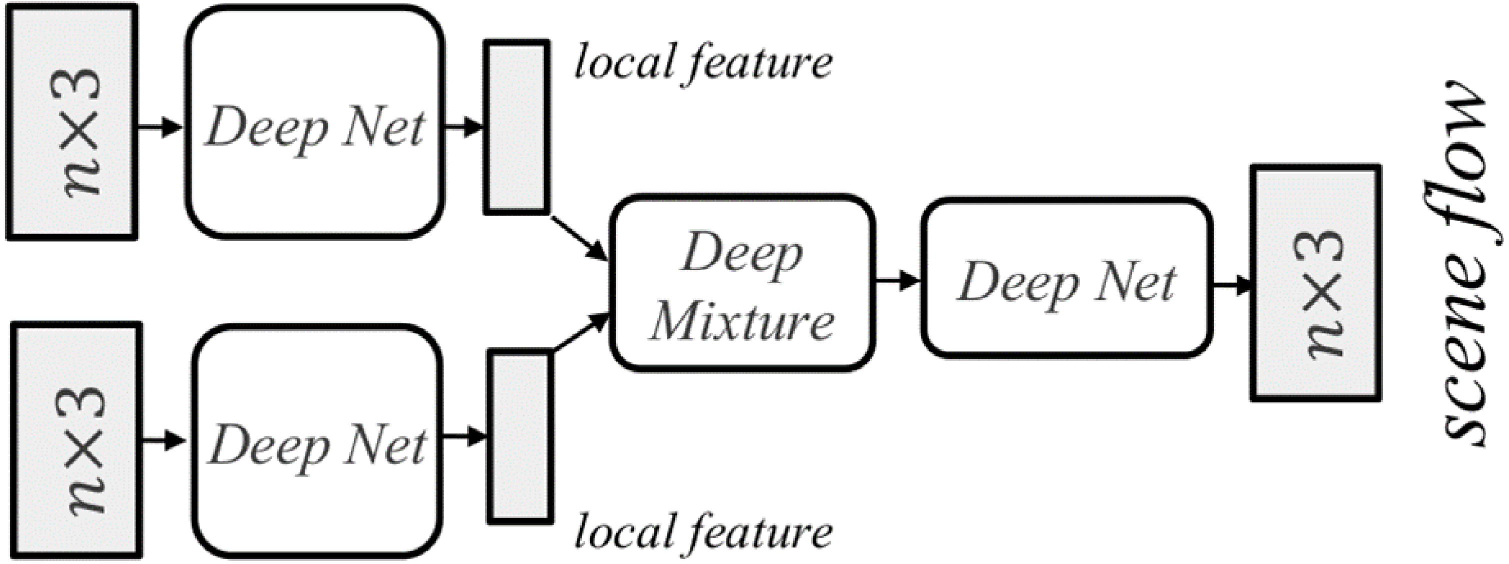
Xingyu Liu, Charles R. Qi, Leonidas Guibas. Learning Scene Flow in 3D Point Clouds, (CVPR 2019).

Deep Net Architecture

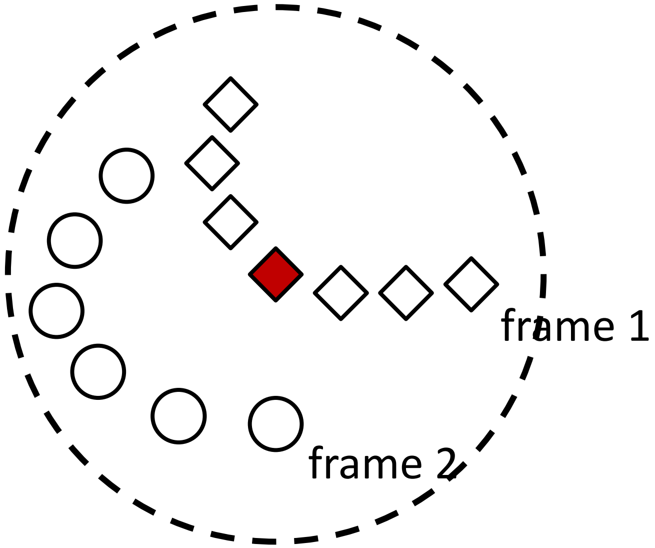
- How to learn point cloud features?
- Where in the network architecture to mix point features from consecutive frames?
- How to mix them?

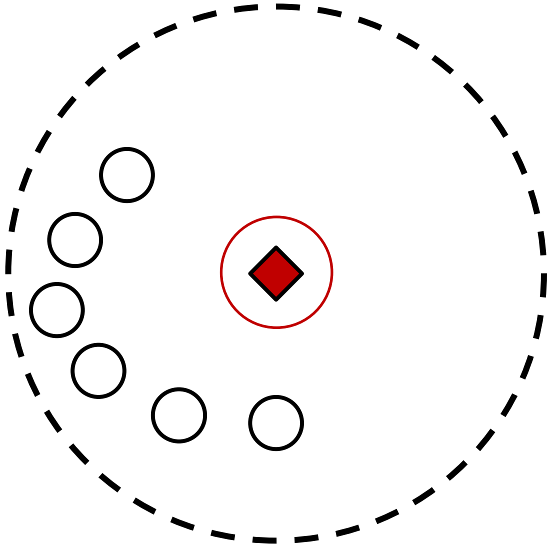
Intermediate level

Middle-level
deep mixture

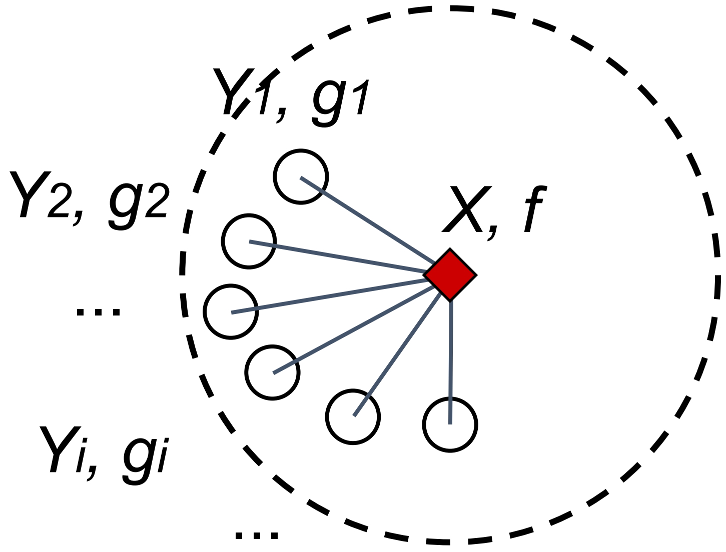


Middle-Level Mixing





Point Attributes



dist(g_1, f), Y_1-X

dist(g_2, f), Y_2-X

⋮

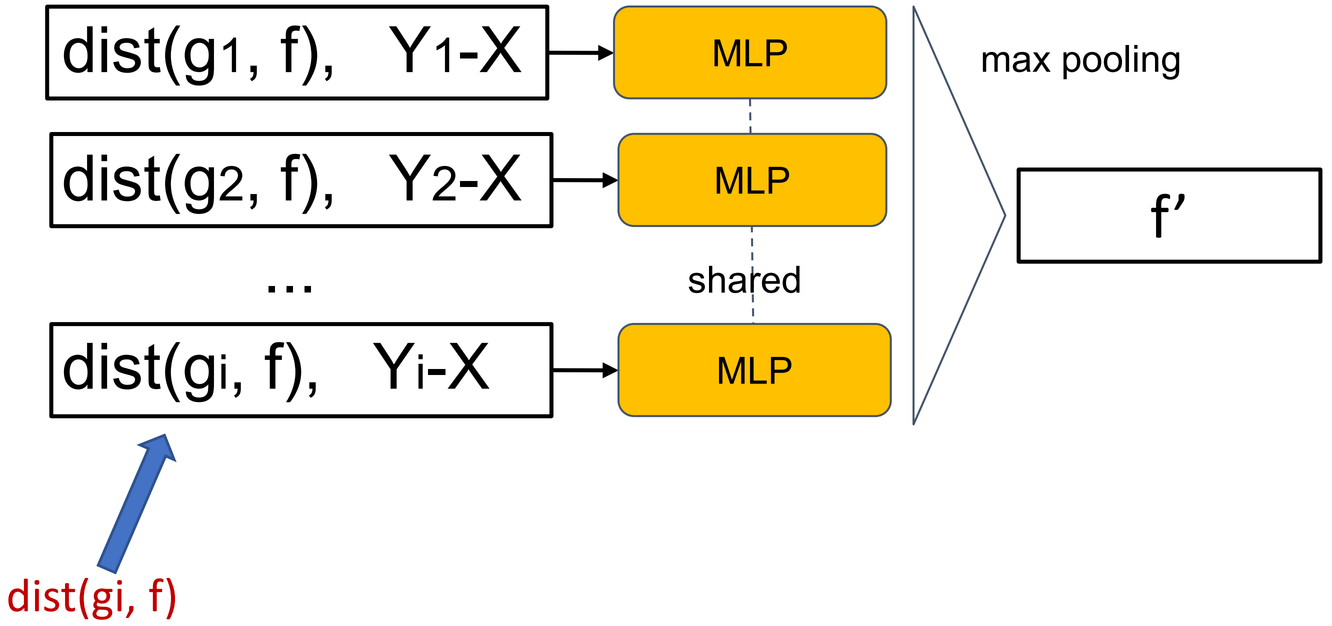
dist(g_i, f), Y_i-X

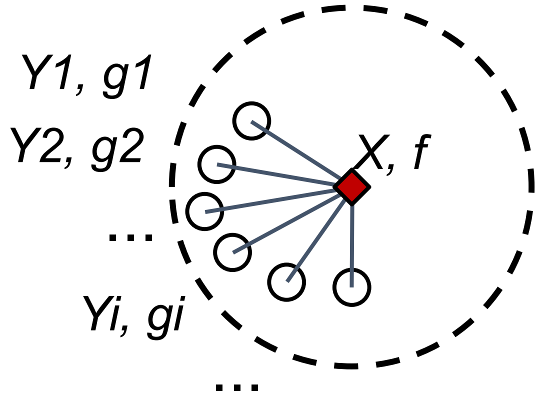
⋮

Naive approach: concatenation

dist(g_1, f), Y_1-X	dist(g_2, f), Y_2-X	...

A More Structured Approach



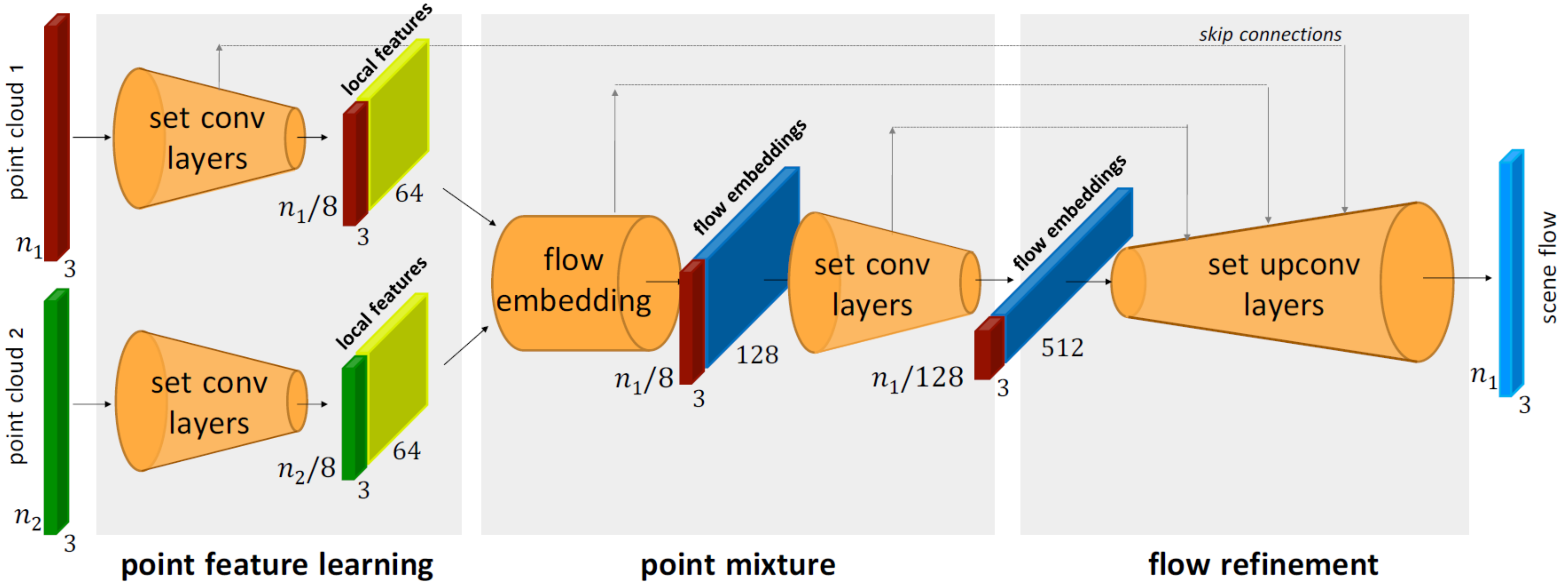


“Distance” functions:

- Euclidean distance (scalar)
- Cosine distance (scalar)
- Element-wise product (vector)

Let the network learn the distance function …

FlowNet3D



Composed of many many mini-pointnet++ modules …

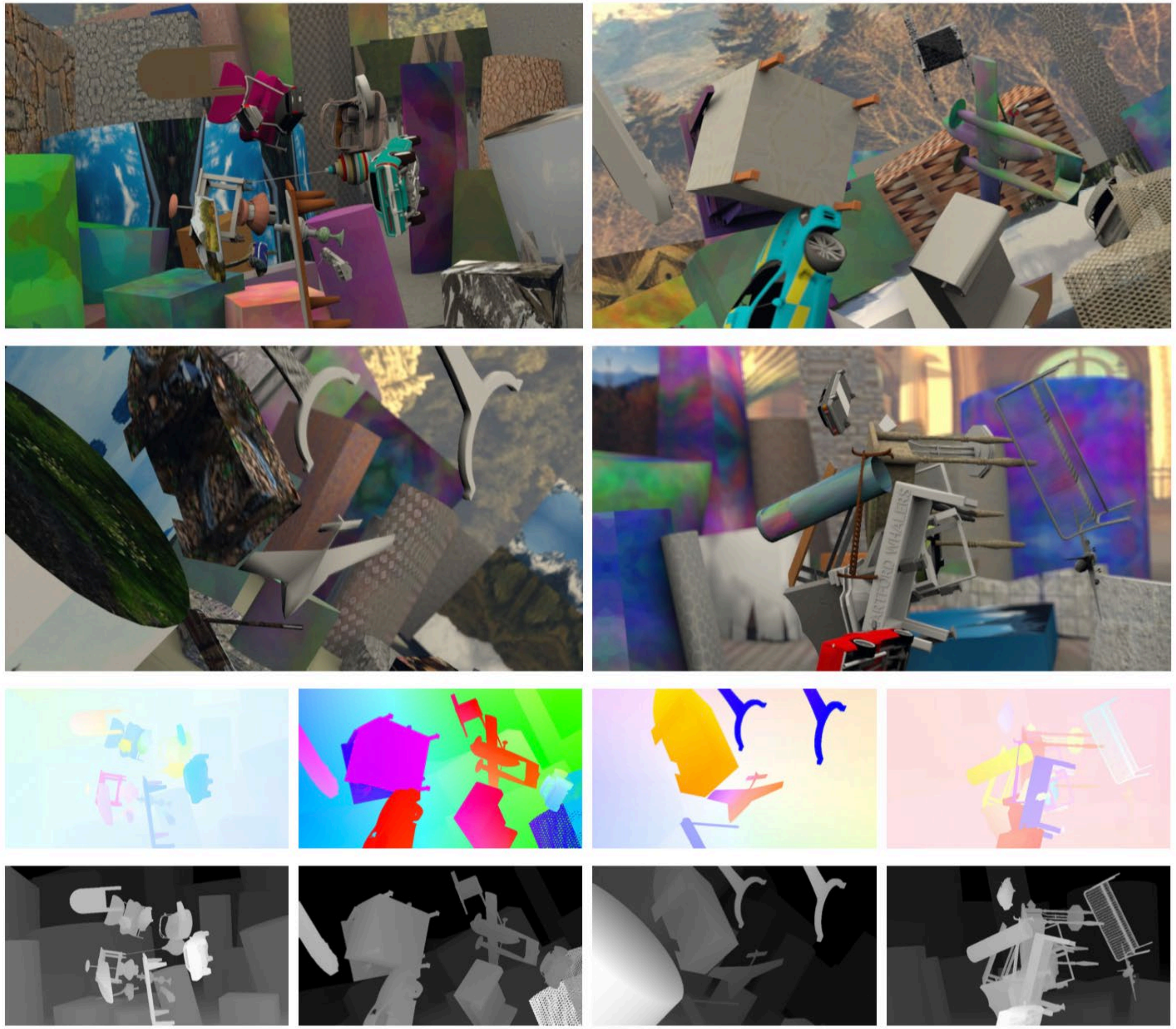
Pointnet++

Training on Synthetic Data

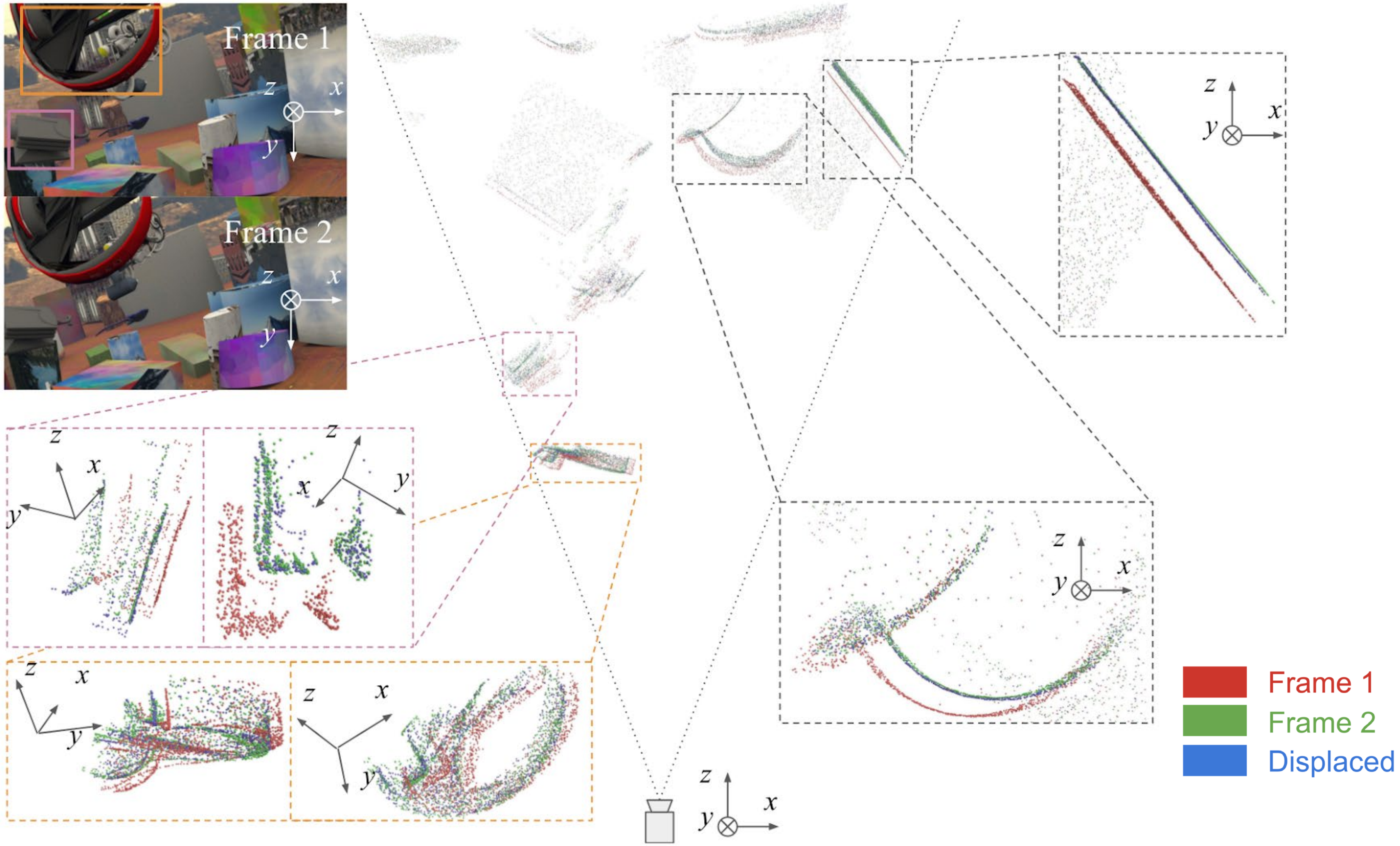
FlyingThings3D [Mayer et al. 2016]
dataset from MPI

Random ShapeNet objects

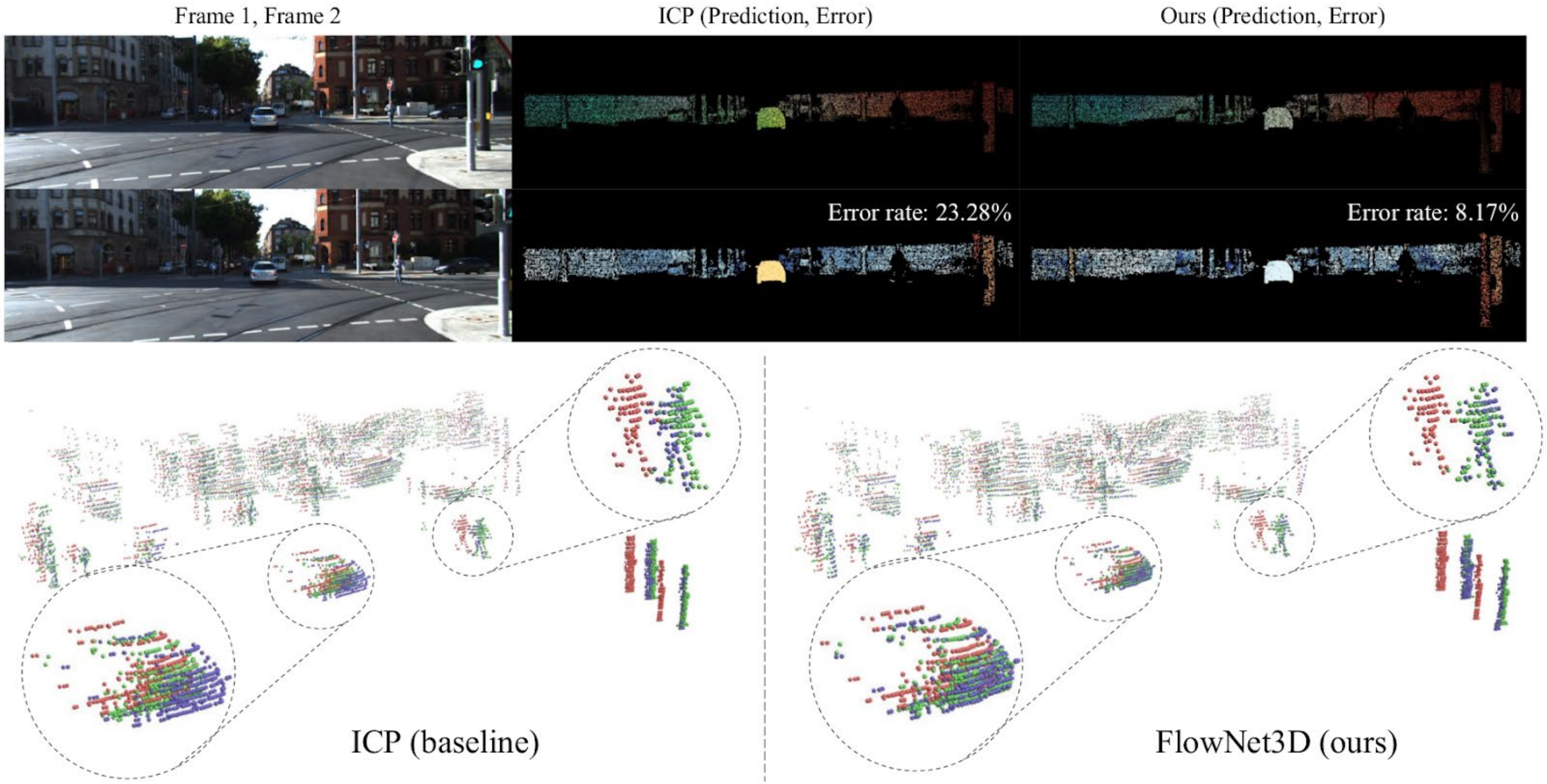
Very challenging dataset with
strong occlusions and large motions.



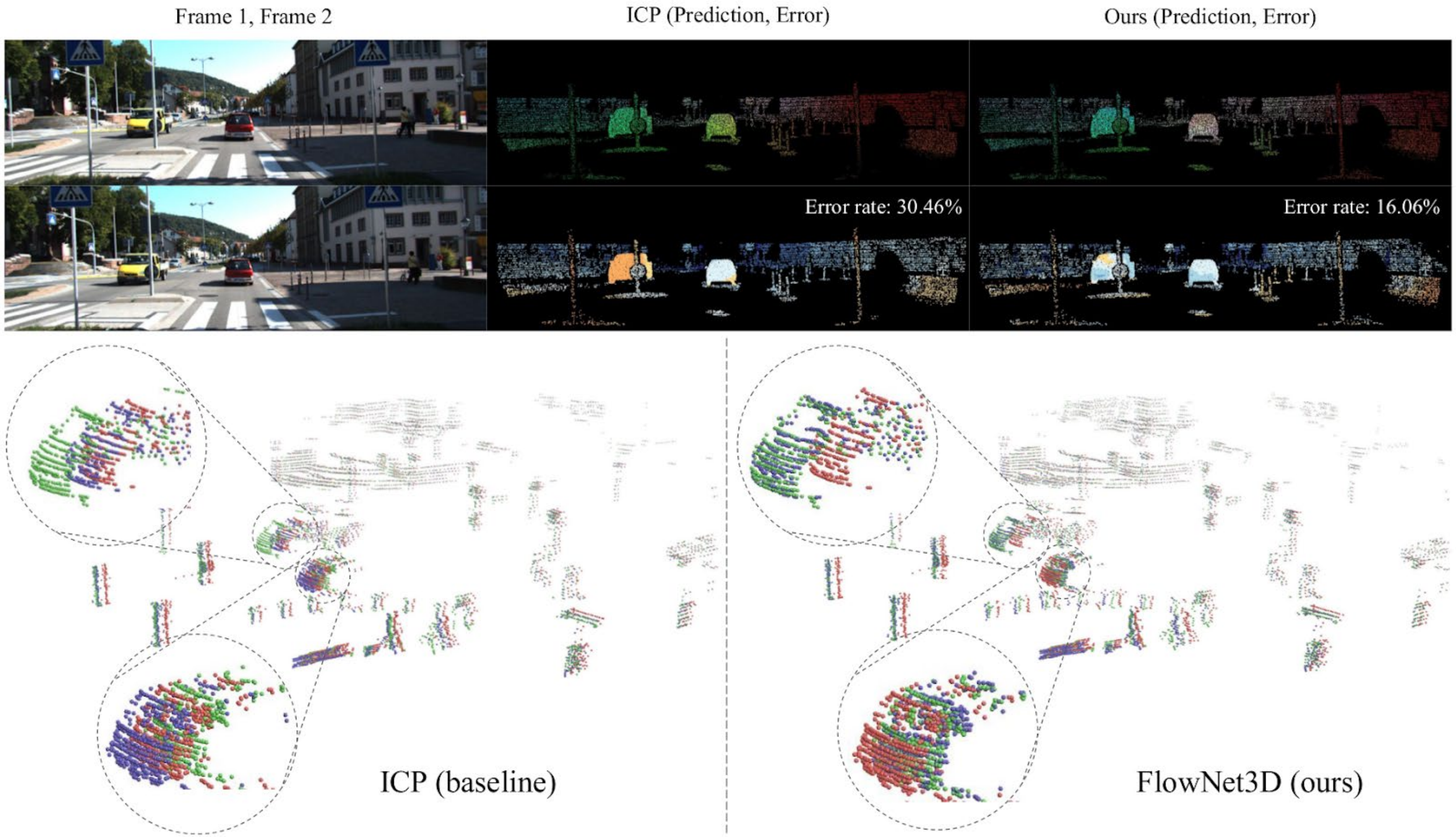
FlyingThings3D Results



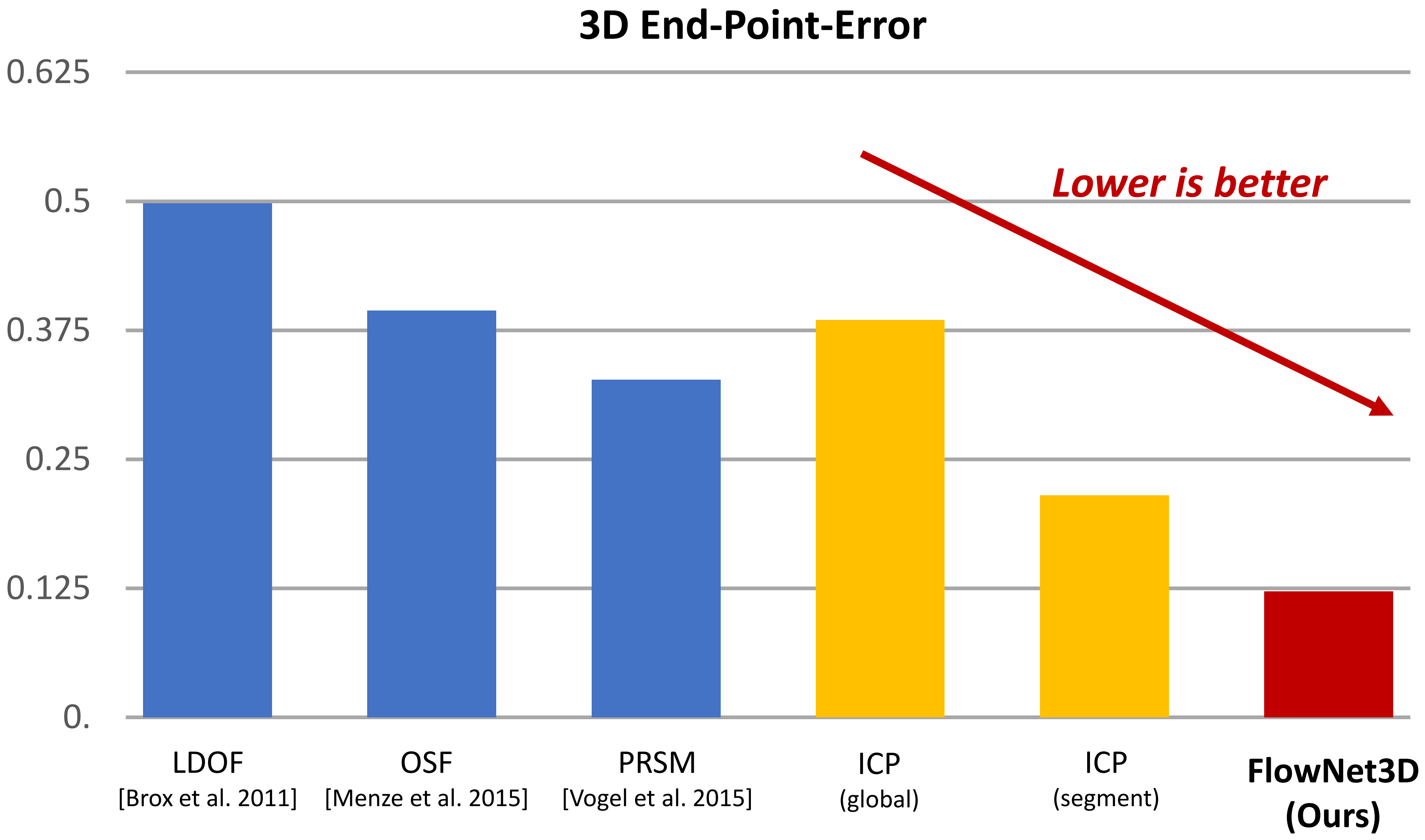
KITTI Results



KITTI Results

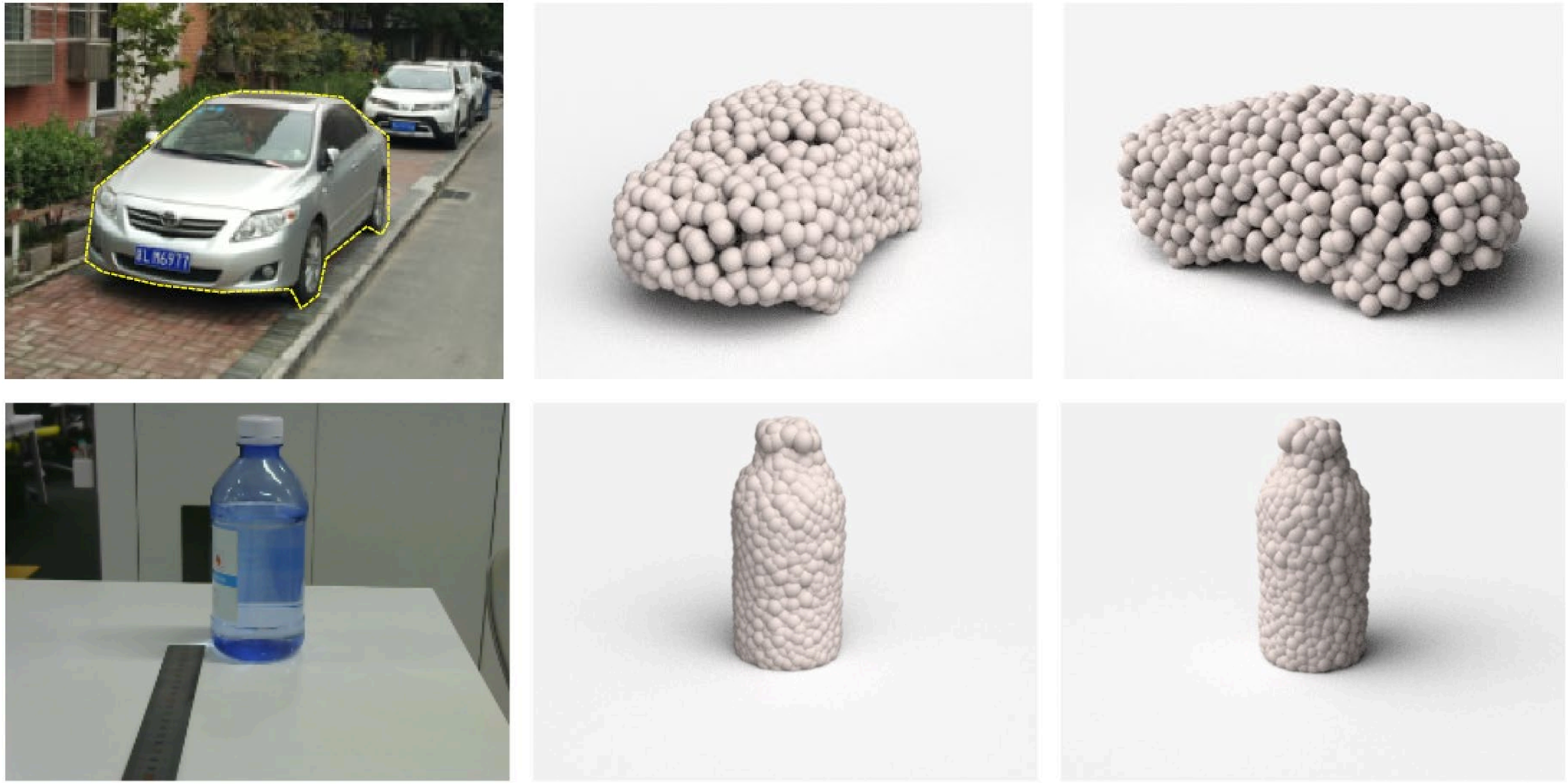


Generalizing to KITTI: Quantitative



Point-Set Generation

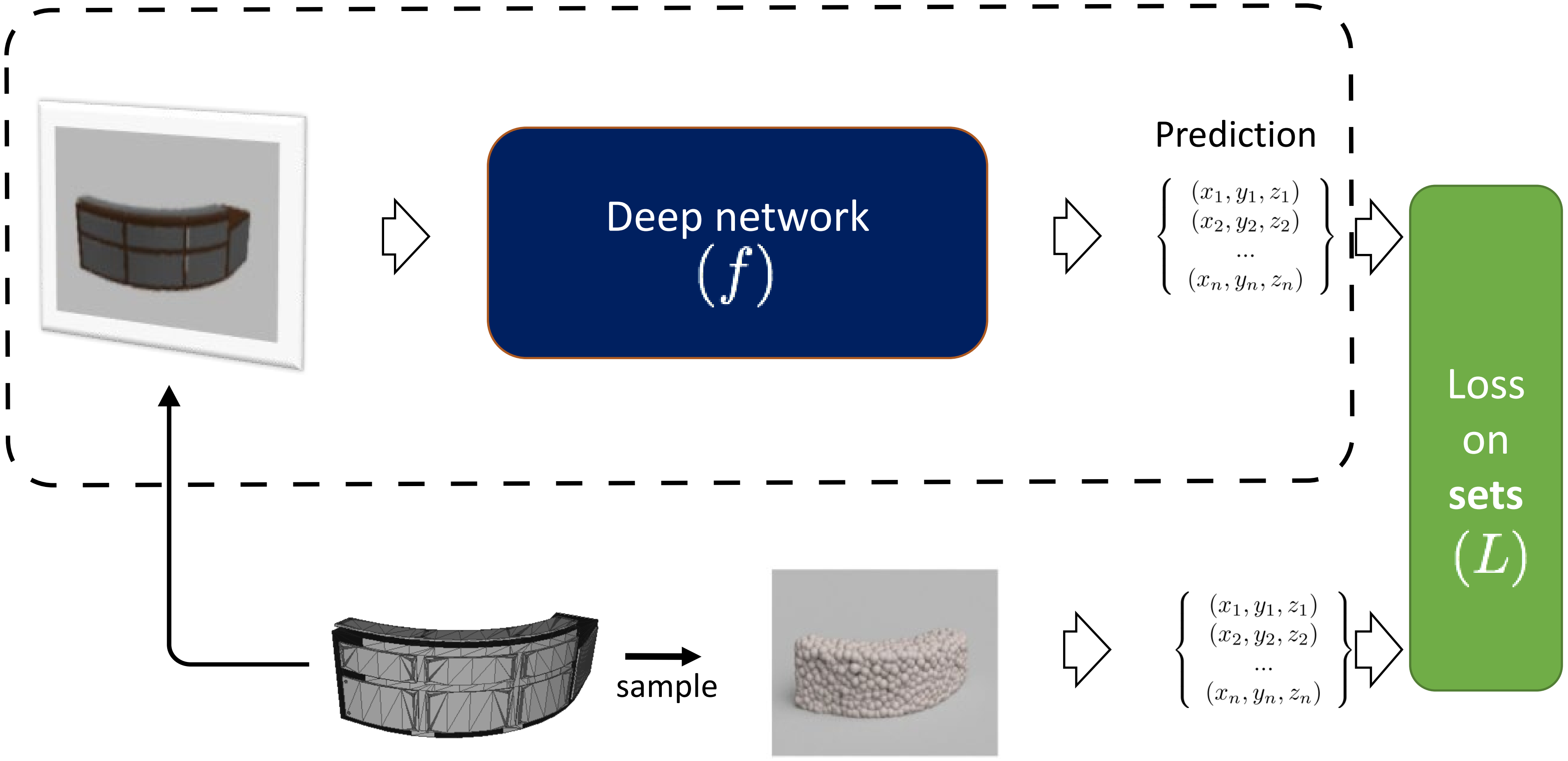
Point Cloud Synthesis from a Single Image



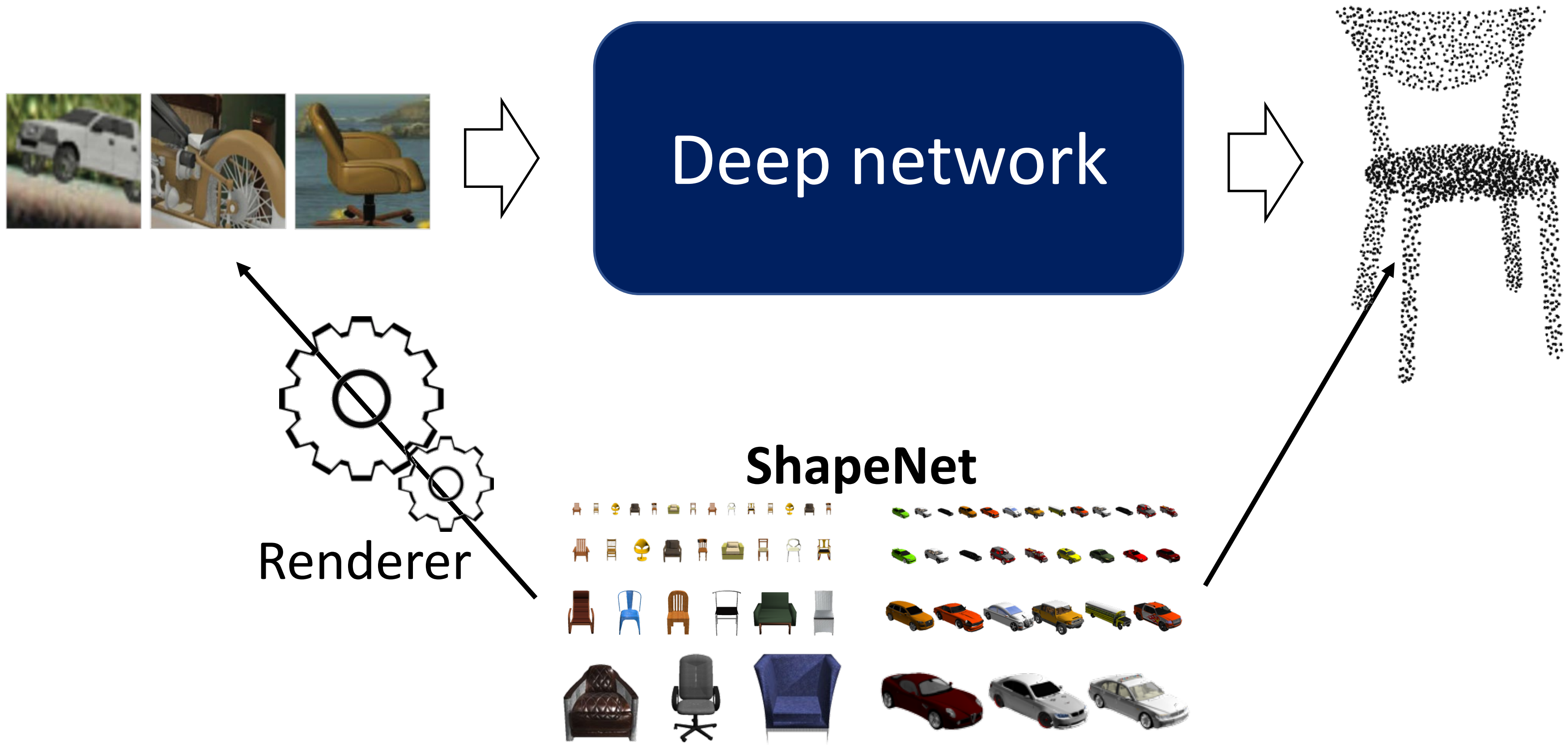
Input Reconstructed 3D point cloud

[H. Su, H. Fan, LG, 2017]

End-to-End Learning



Synthesize for Learning



Point Cloud Distance Metrics

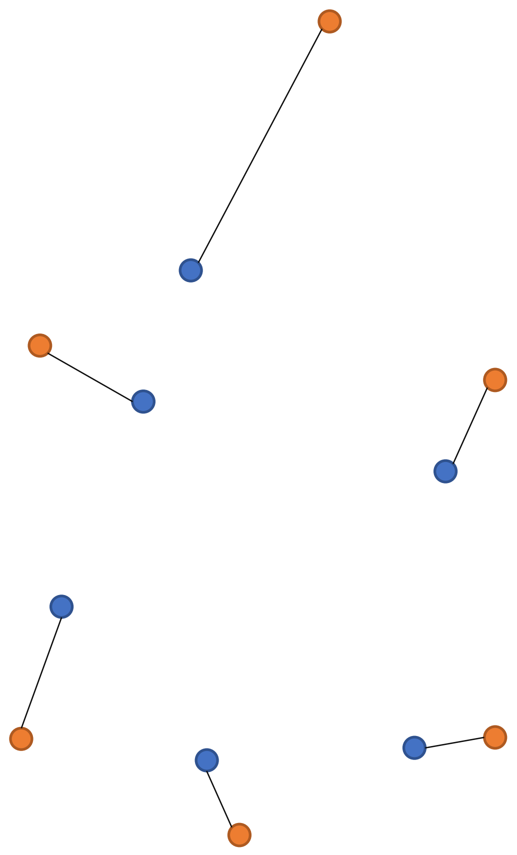
Worst case: Hausdorff distance (HD)

Average case: Chamfer distance (CD)

Optimal case: Earth Mover's distance (EMD)

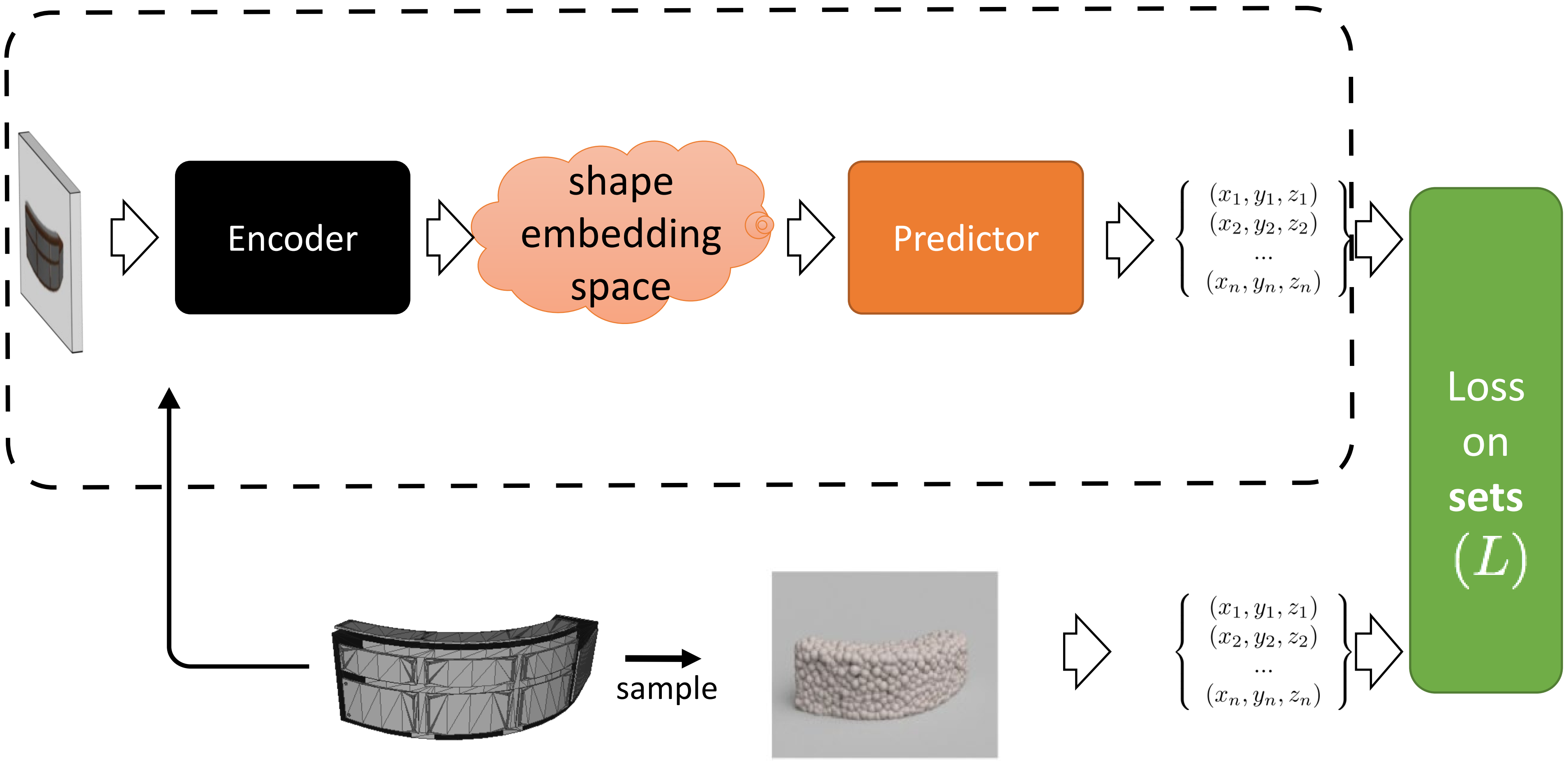
$$d_{EMD}(S_1, S_2) = \min_{\phi: S_1 \to S_2} \sum_{x \in S_1} \|x - \phi(x)\|_2$$

where $\phi : S_1 \to S_2$ is a bijection.

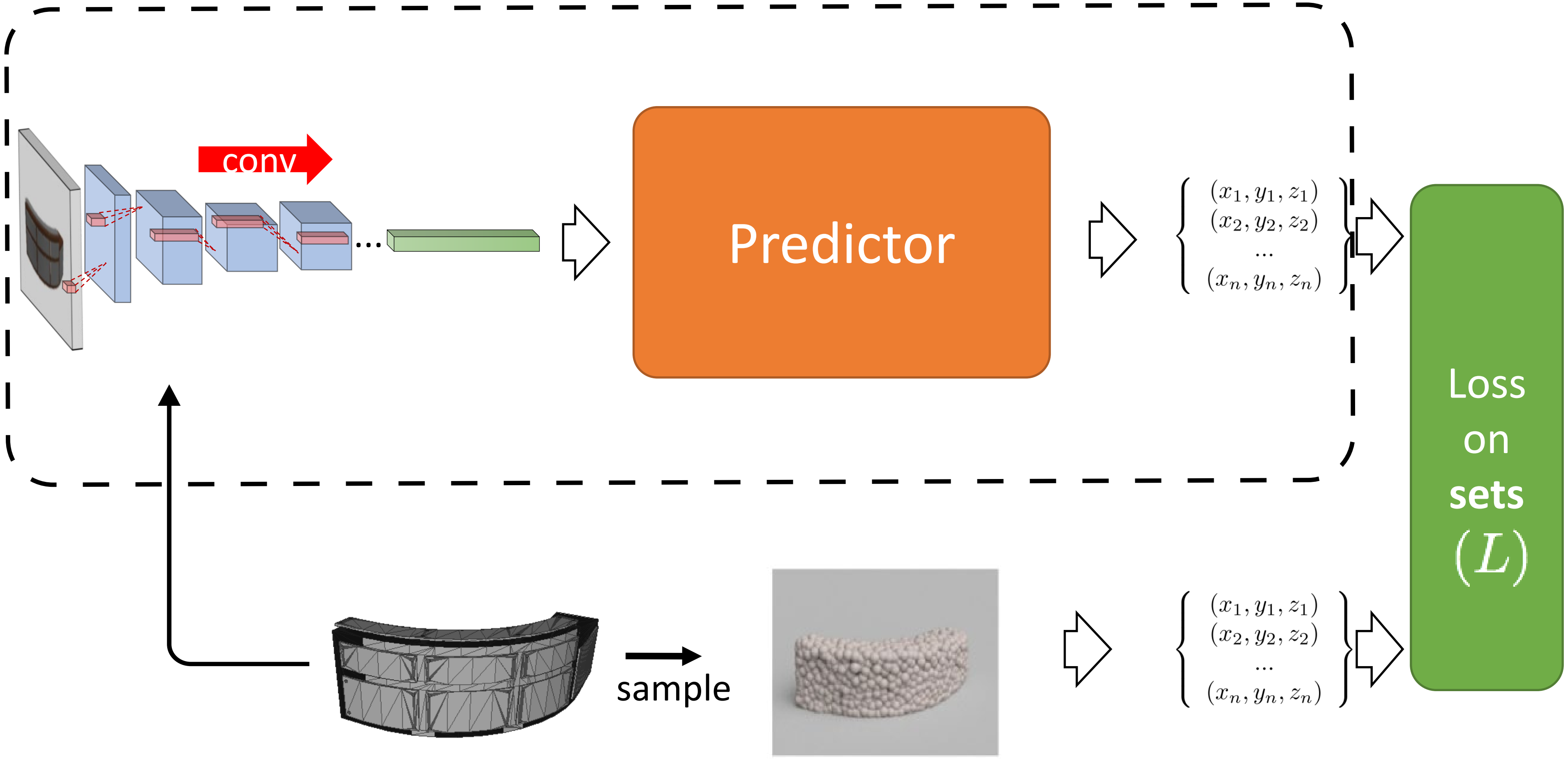


Solves the optimal transportation (bipartite matching) problem!

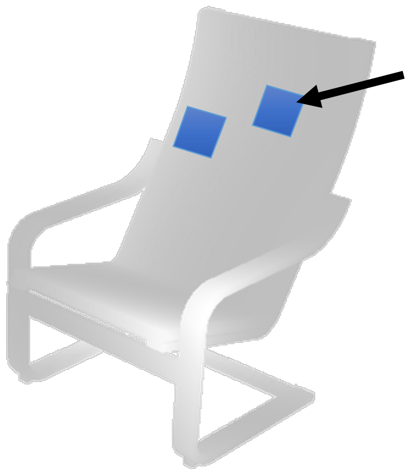
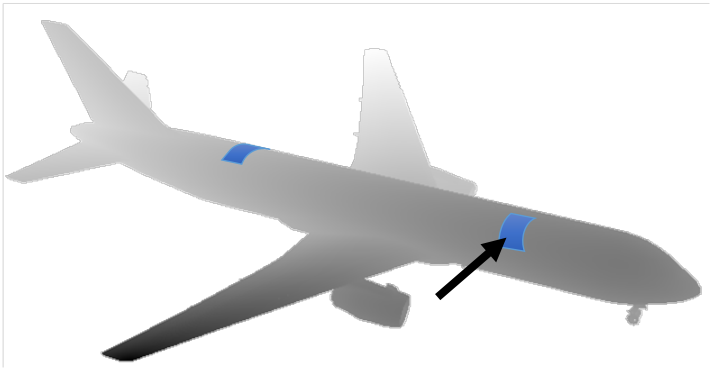
End-to-End Learning



End-to-End Learning



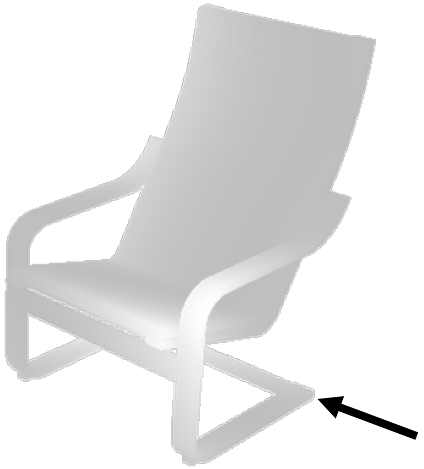
Natural Statistics of Object Geometry

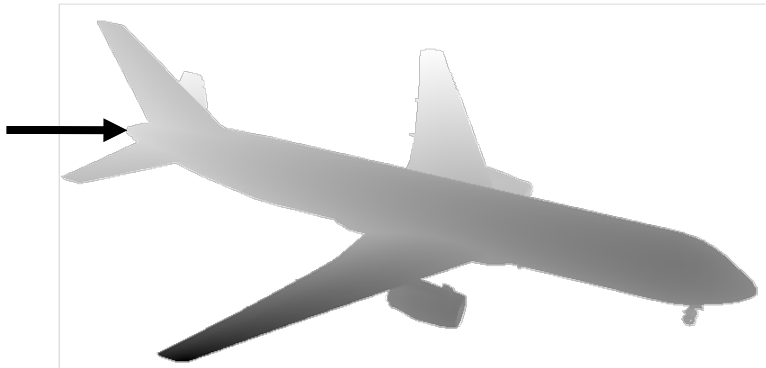
 

- Many local structures are common
  - e.g., planar patches, cylindrical patches
  - **strong local correlation** among point coordinates

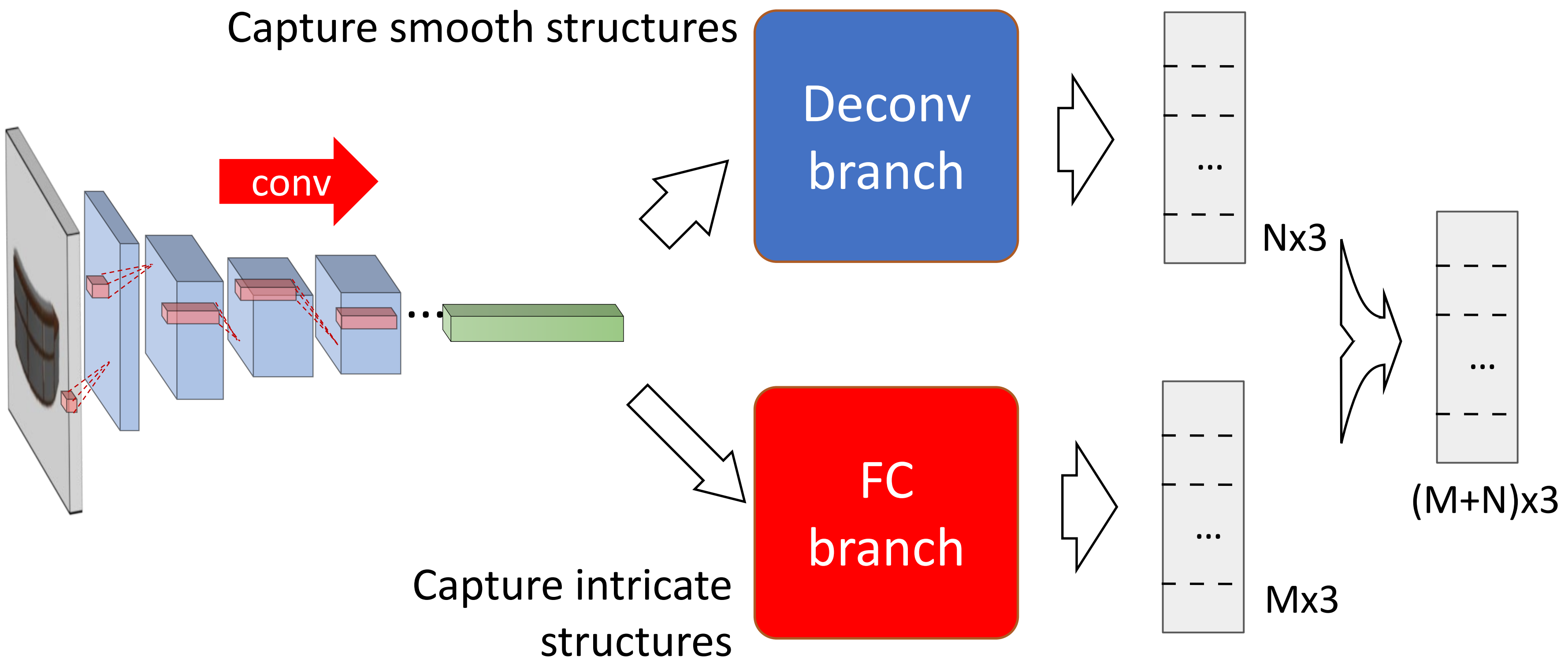
Natural Statistics of Object Geometry

- Many local structures are common/shared
    - e.g., planar patches, cylindrical patches
    - **strong local correlation** among point coordinates
- But also some intricate local structures
    - **some points have high variability neighborhoods**





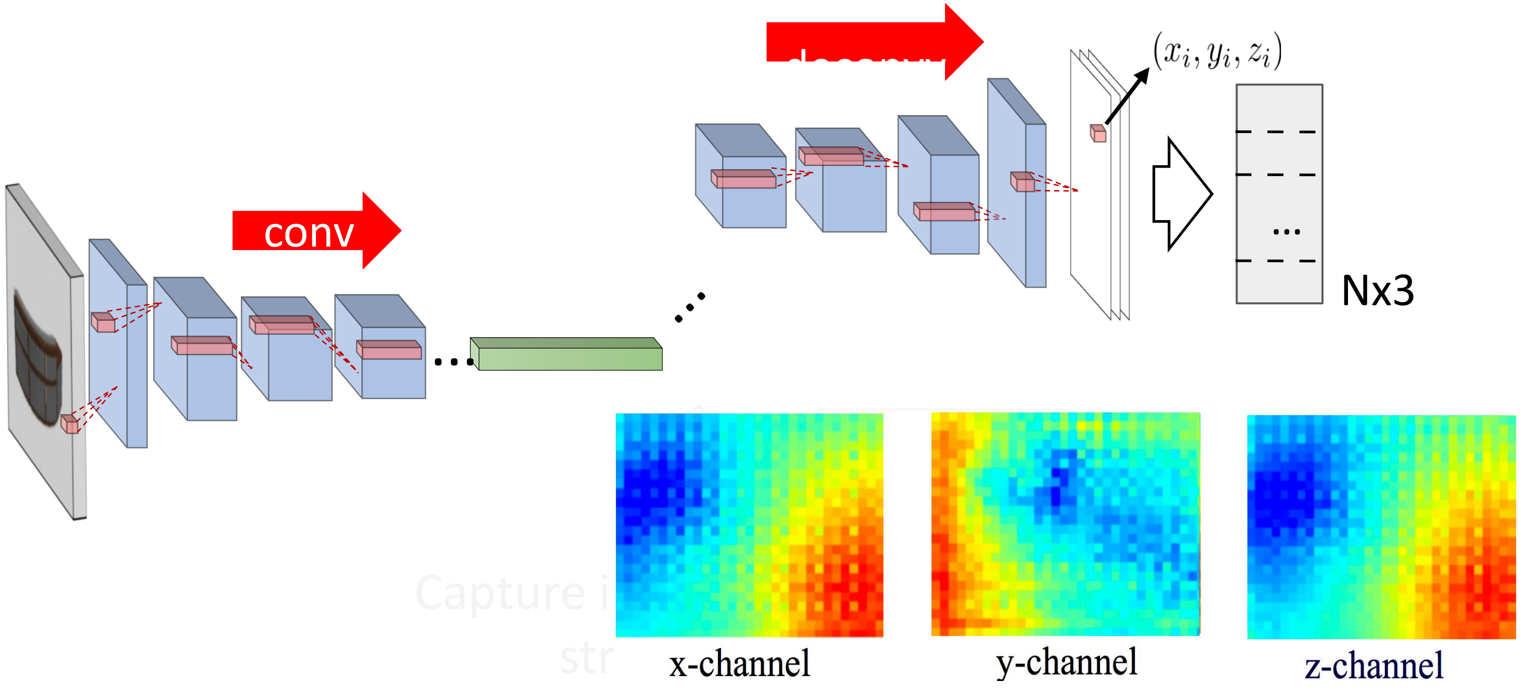
Two-Branch Architecture



Set union by array concatenation

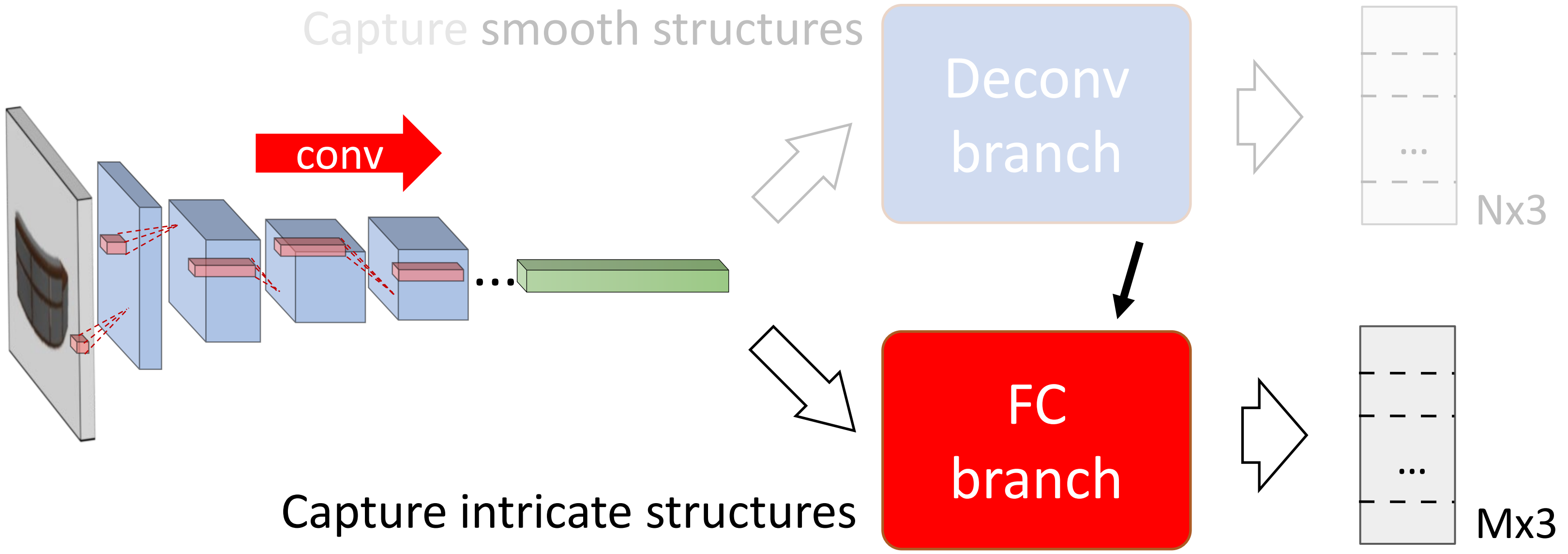
Deconvolution Branch

Parametrization / coordinate map



- Deconvolution induces a smooth coordinate map
- Geometrically, learns a smooth parameterization

Fully Connected Branch



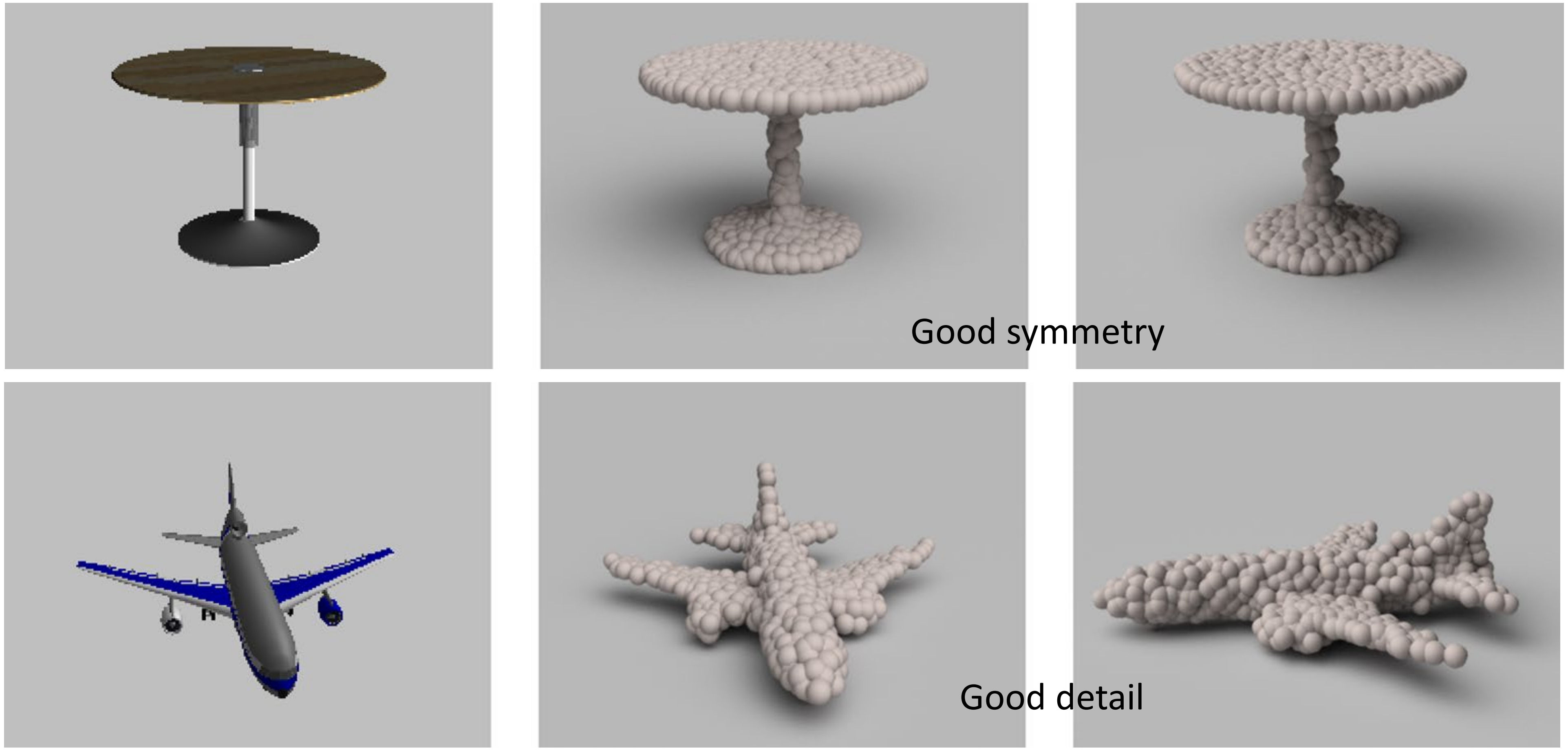
The Two Branches

blue: deconv branch – **large, consistent, smooth** structures

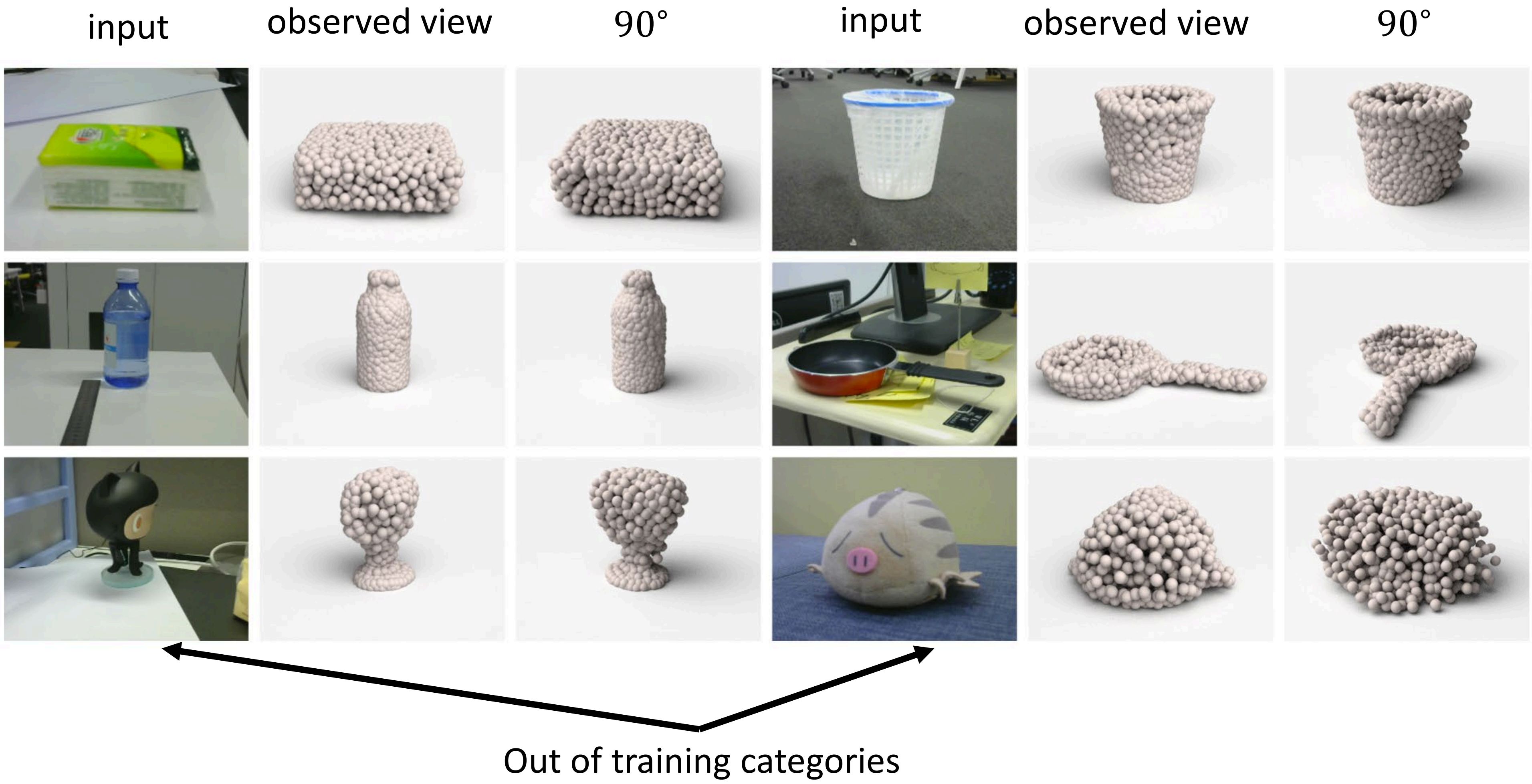
red: fully-connected branch – **more intricate** structures

Example Results

Same view New view

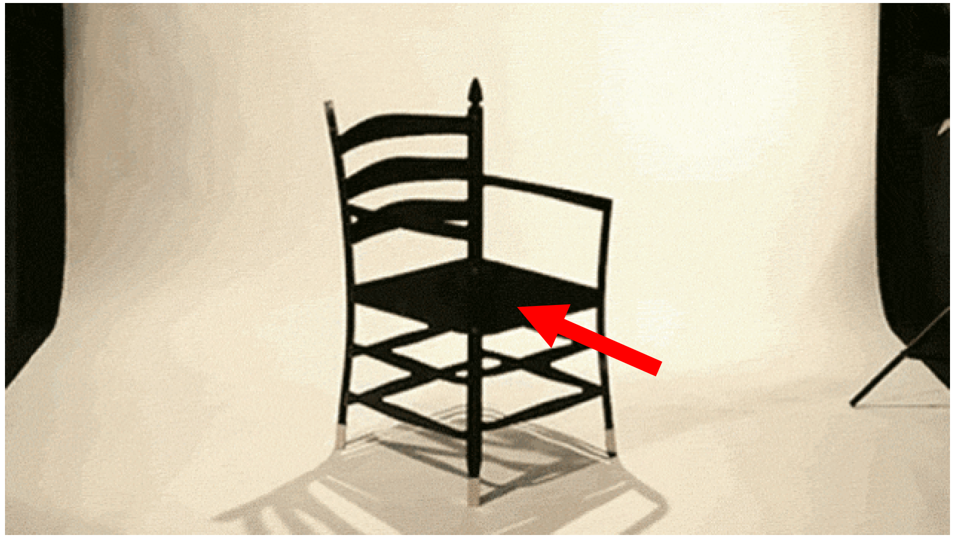


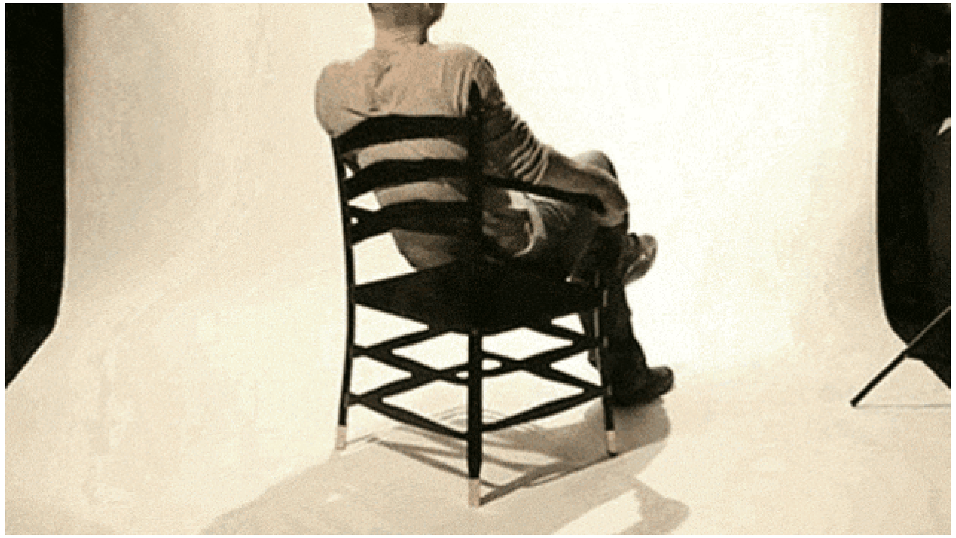
From Real Images



Influence of Distance Metrics

- A fundamental issue: inherent ambiguity in prediction



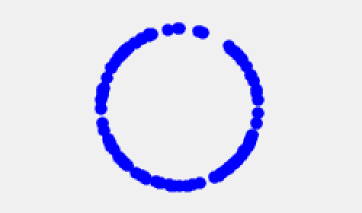


- By loss minimization, the network tends to predict a "**mean shape**" that **averages out** uncertainty

Distance Metrics Affect Mean Shapes

$$\bar{x} = \underset{x}{\operatorname{argmin}}\ \mathbb{E}_{s\sim\mathbb{S}}[d(x, s)]$$

continuous hidden variable (radius)



Input

EMD mean

Chamfer mean

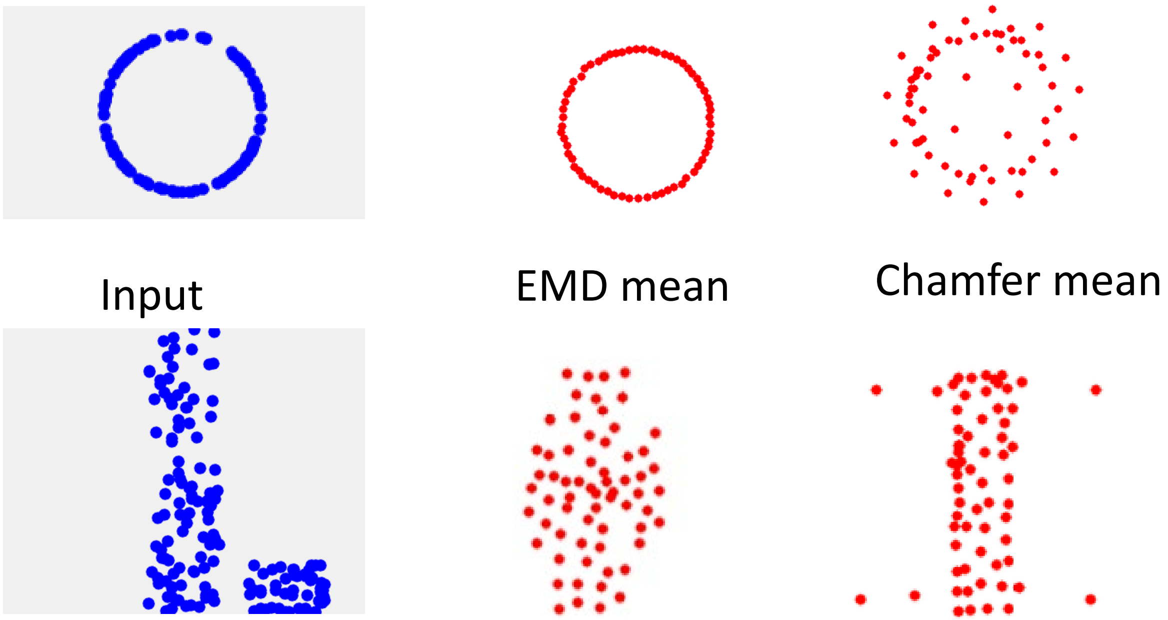
The mean shape carries characteristics of the distance metric

Distance Metrics Affect Mean Shapes

The mean shape carries characteristics of the distance metric

continuous
hidden variable
(radius)

discrete
hidden variable
(add-on location)

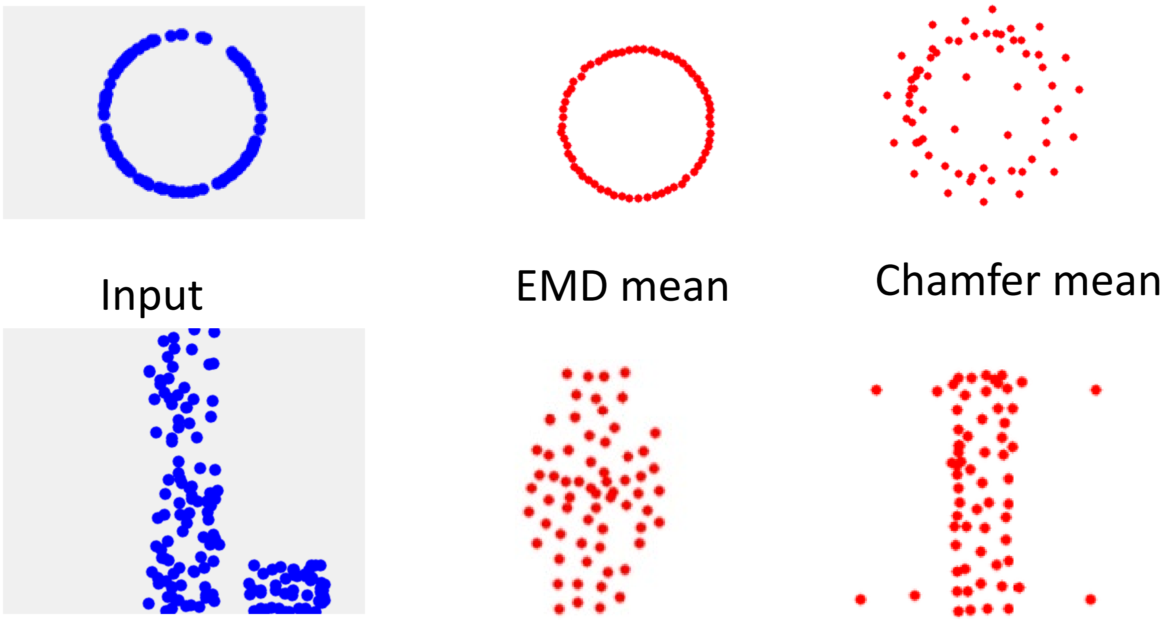


Distance Metrics Affect Mean Shapes

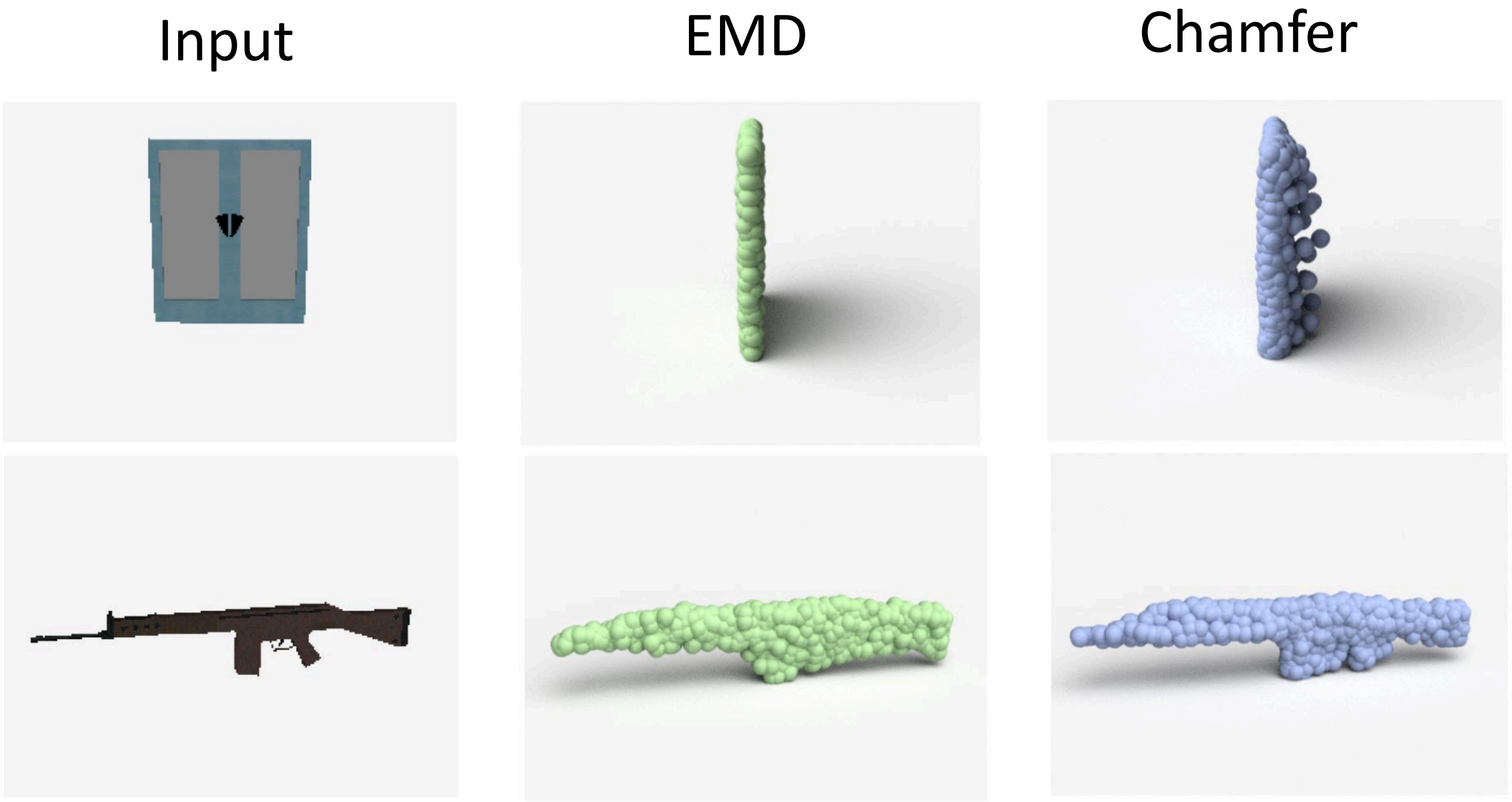
The mean shape carries characteristics of the distance metric

continuous
hidden variable
(radius)

discrete
hidden variable
(add-on location)



EMD vs CD Predictions



Future Directions for Point Cloud Deep Learning

Future Directions

- **Scalability**

How to scale up from processing 100k points to 1M or even 10M points?

(1024 x 1024 image ~ 1M pixels)

Trade-offs in neighborhood sampling

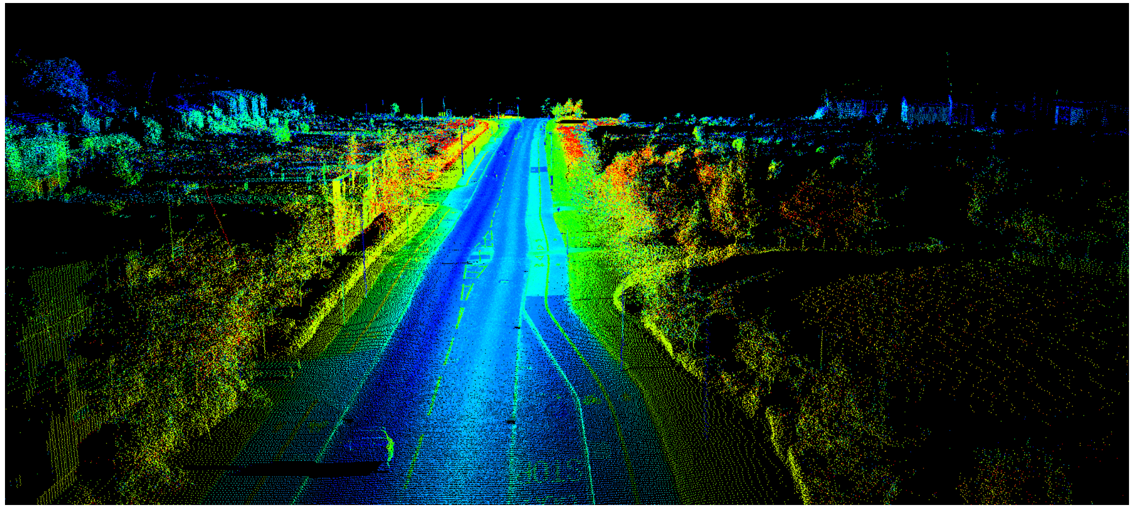
More memory efficient operators

Future Directions

- Scalability
- **Multi-modality**

RGB images

High resolution
Rich textures

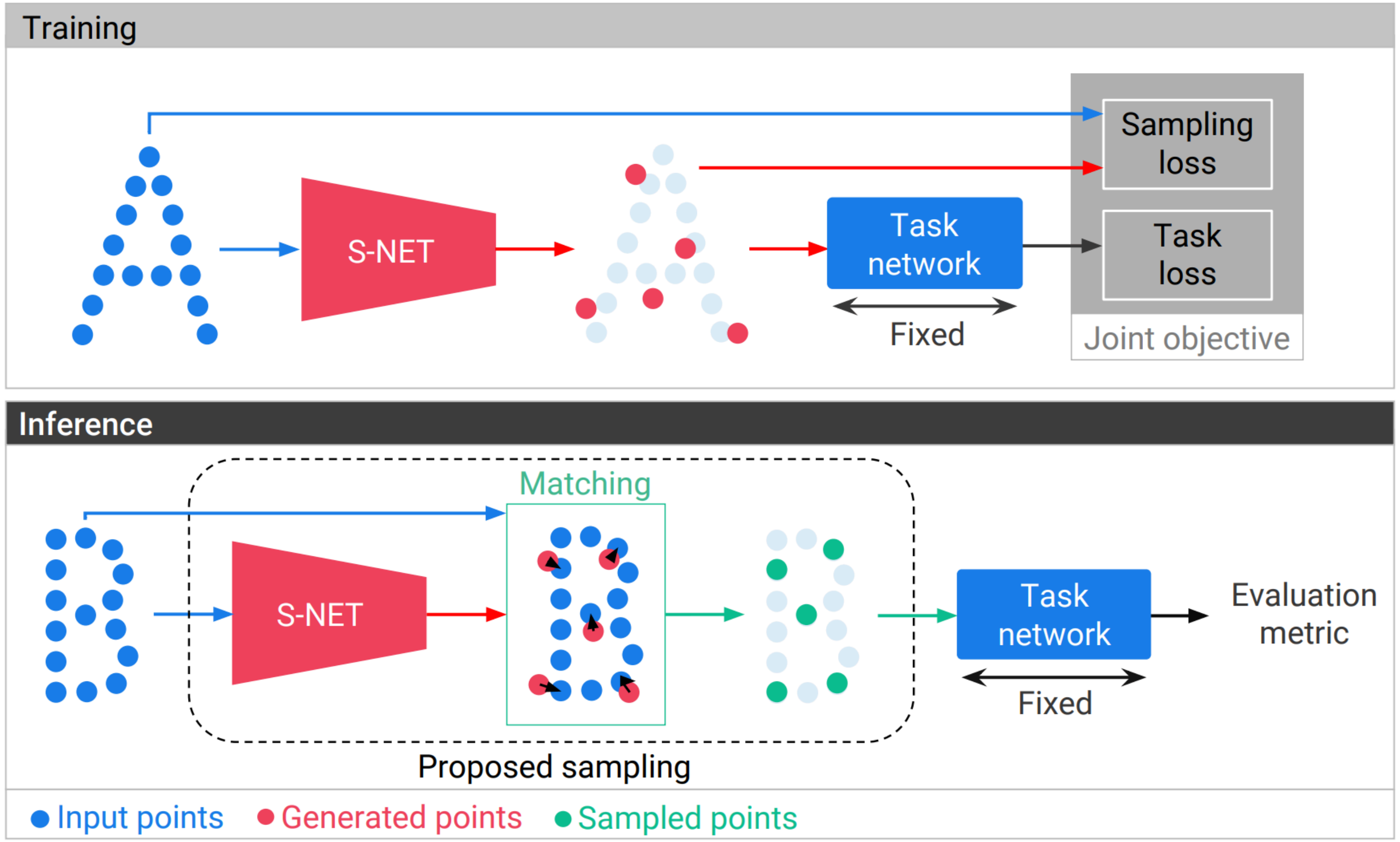


Lidar point clouds

Accurate depth
Accurate 3D geometry

Future Directions

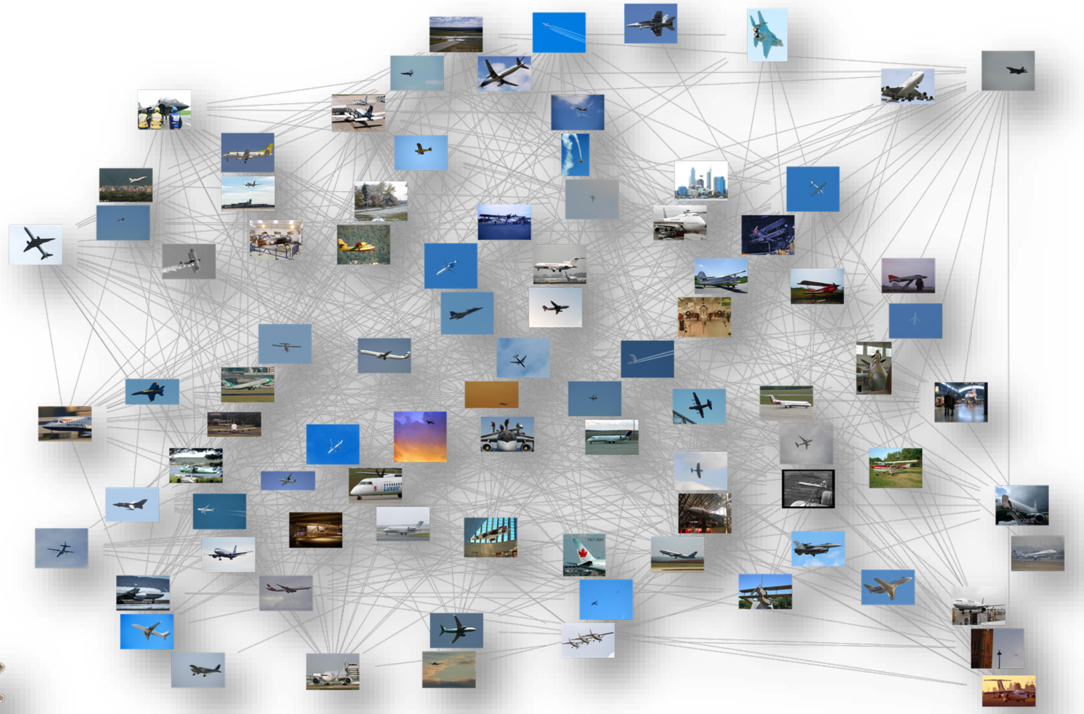
- Scalability
- Multi-modality
- **Sampling**



Learning to sample [Dovrat et al.]

General Set / Graph Processors

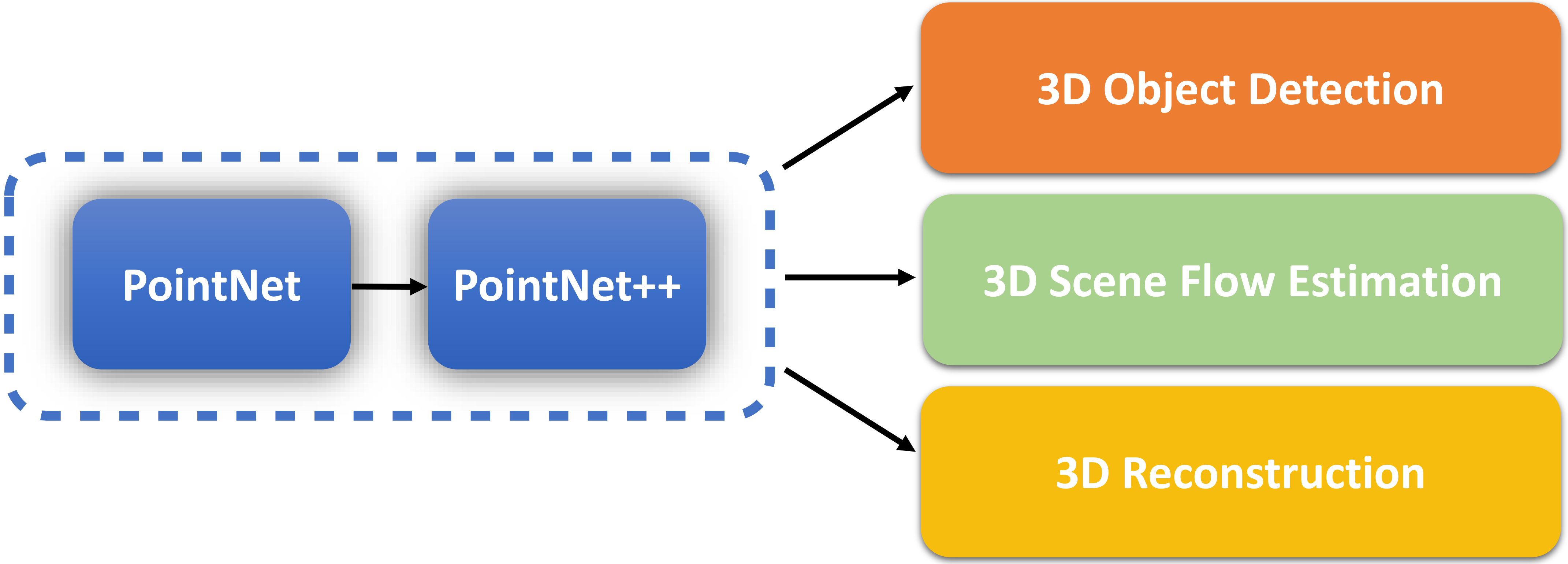
- Scalability
- Multi-modality
- Sampling
- **Set processing**



Conclusions: Real-World 3D Understanding

- **Novel architectures for deep learning on point clouds** – PointNet and PointNet++, respecting invariances, light-weight and robust to data corruption, a unified framework for various tasks.

- **Successful applications in 3D scene understanding**.



The End