PPF-FoldNet: Unsupervised Learning of Rotation Invariant 3D Local Descriptors

(Haowen Deng, Tolga Birdal, Slobodan Ilic, TUM Chair for Computer Aided Medical Procedures & Augmented Reality)

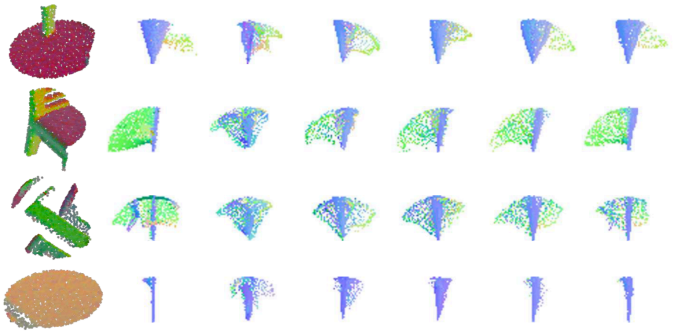
Simon Boche

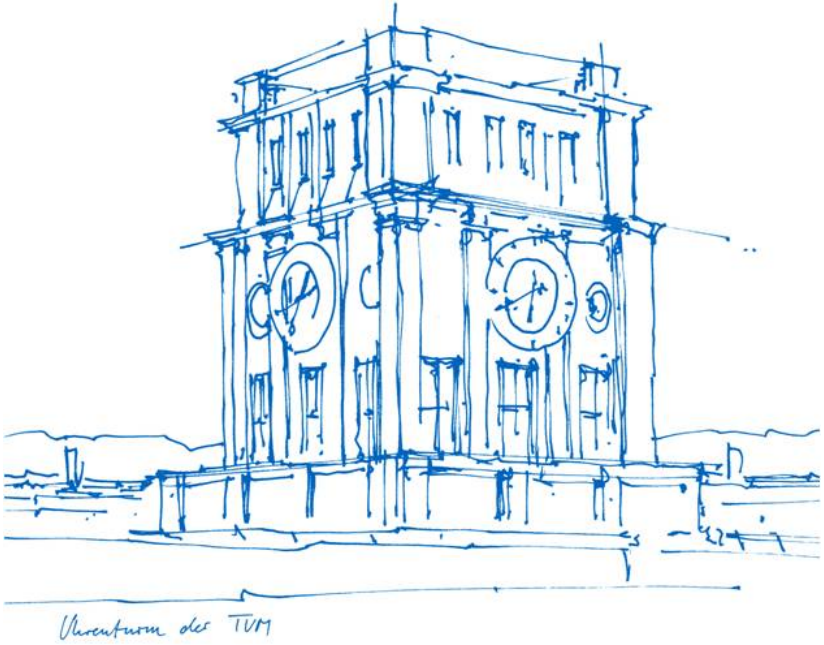
Technische Universität München

Seminar: Recent Advances in 3D Computer Vision

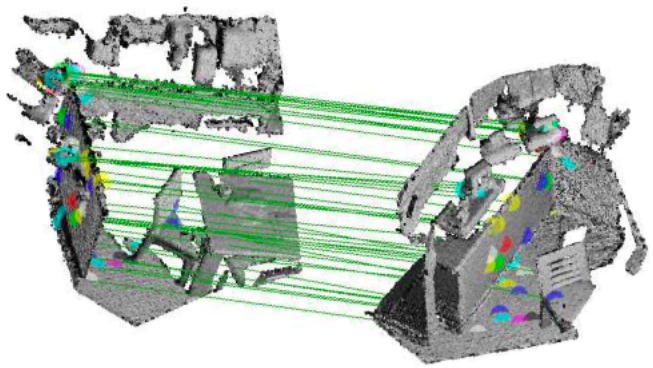
Computer Vision Group

02.07.2019



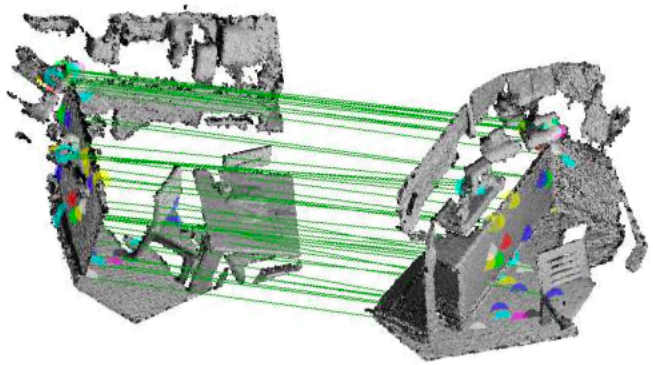


Motivation



- **Local descriptors** are essential tools in many computer vision tasks:
  - Object detection
  - Pose Estimation
  - SLAM
  - …
- Methods **well established in 2D**
- **3D descriptors** still lack good discriminative power and repeatability
- **Recent research** aims to apply **deep learning approaches** for 3D feature extraction

Motivation



- **BUT:** So far, learning 3D descriptor suffers from one of the following:

a) **Supervised learning requires enormous amount of data**

b) **Sensitivity to 6DoF rotations**

c) **Expensive pre-processing of input data**

d) **Unsatisfactory performance**

Outline

I. Background / Related work

a. Point clouds and point pairs
b. Related work (PointNet, FoldingNet, PPFNet)

II. PPF-FoldNet: Method description

a. Basic concept
b. Network architecture

III. Experimental validation

a. Data preparation and definition of accuracy
b. Results

IV. Summary

From Point Cloud to Point Pairs

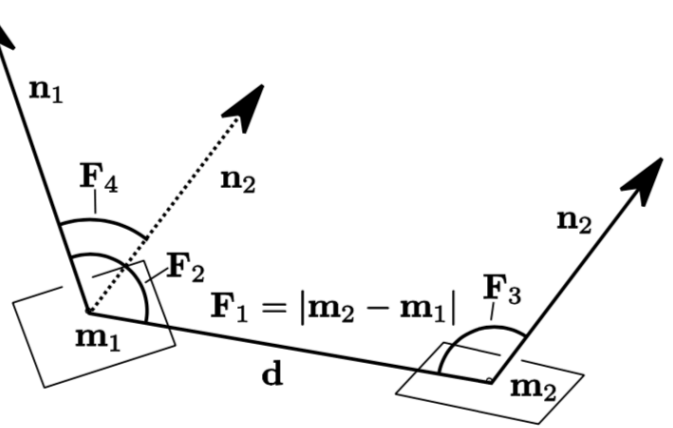
Input Point Cloud

- Set of oriented points
  $$\boldsymbol{X} = \{\boldsymbol{x_i} \in \mathbb{R}^6\}$$
- Every point $\boldsymbol{p}$ is assigned a normal vector $\boldsymbol{n} \in \mathbb{R}^3$ such that:
  $$\boldsymbol{x_i} = \{\boldsymbol{p}, \boldsymbol{n}\} \in \mathbb{R}^6$$

- **Local Patch Representation:**
  Local patch is a subset $\boldsymbol{\Omega_r} \subset \boldsymbol{X}$ around a reference point $\boldsymbol{x_r}$

Point Pairs

- Every patch can be encoded as a collection of point pair features
- Every point pair $(\boldsymbol{m_1}, \boldsymbol{m_2})$ can be defined by four parameters:
  1. **Distance** between points
  2. **Angle** between a **normal** and the **distance vector**
  3. **… and vice versa**
  4. **Angle** between the **two normal vectors**



Related Work

- **PointNet (Qi et al., Stanford, 2017)**
  - Pure point cloud Input
  - Supervised MLP

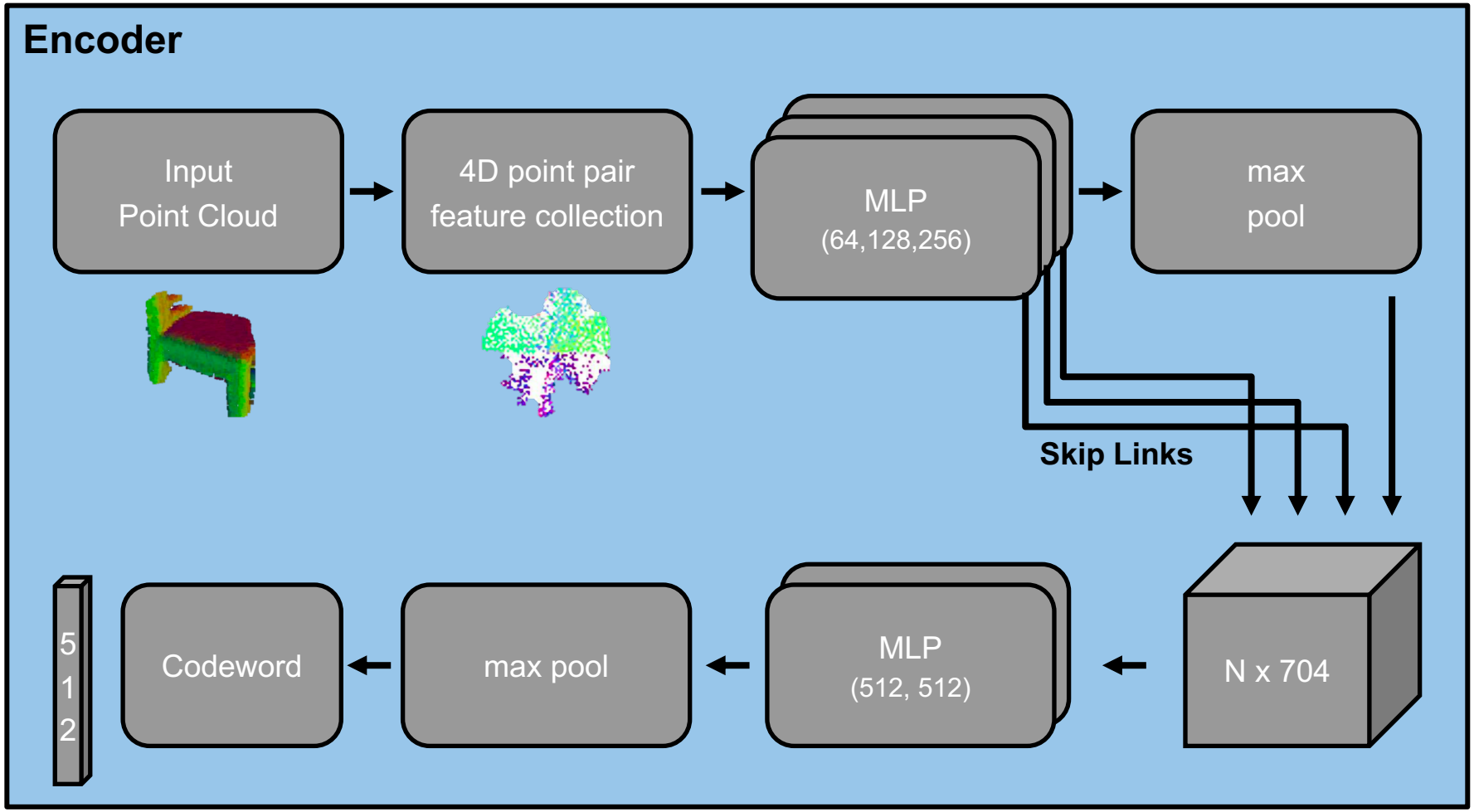
- **FoldingNet (Yang et al., Carnegie Mellon University, 2018)**
  - Autoencoder structure
  - Introduces folding operations as a strong decoder alternative

- **PPFNet (Deng et al, TUM, 2018)**
  - Combined point pair feature (PPF) and point input
  - Supervised architecture
  - Learning local features informed by context of the scene
  - Memory Bottleneck

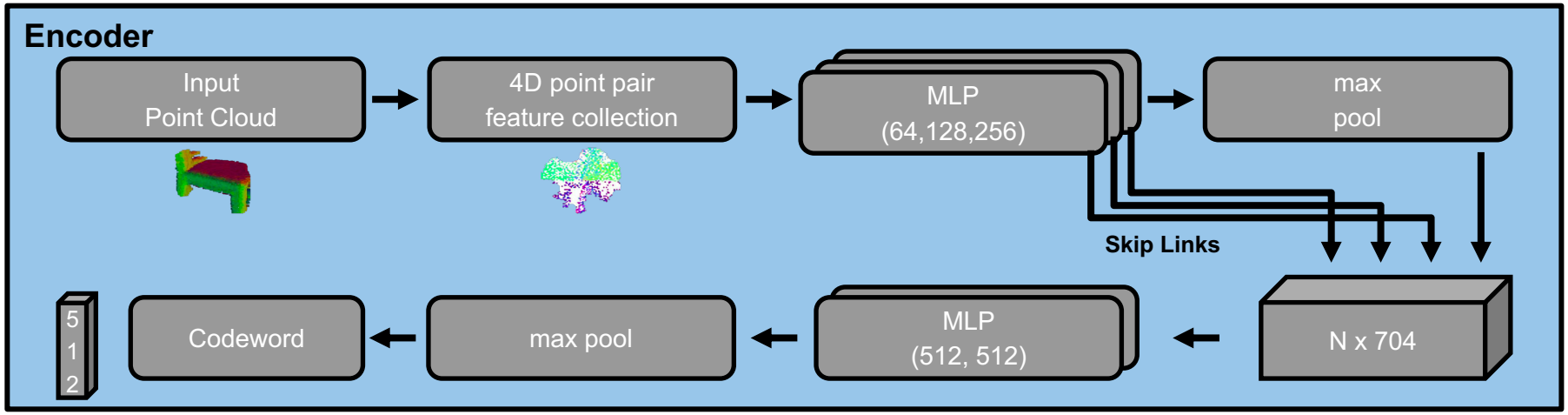
PPF-FoldNet: Key Concept

- PPF-FoldNet builds up an **autoencoder** structure
  - Neural Network that **reconstructs its input**
- PPF-FoldNet **combines** elements of the previously introduced works
  - **PointNet Encoder Structure**
  - **FoldingNet Decoder Structure**
- PPF-FoldNet makes use of **4D point pair features** instead of pure points to learn **rotation invariant (local) features**

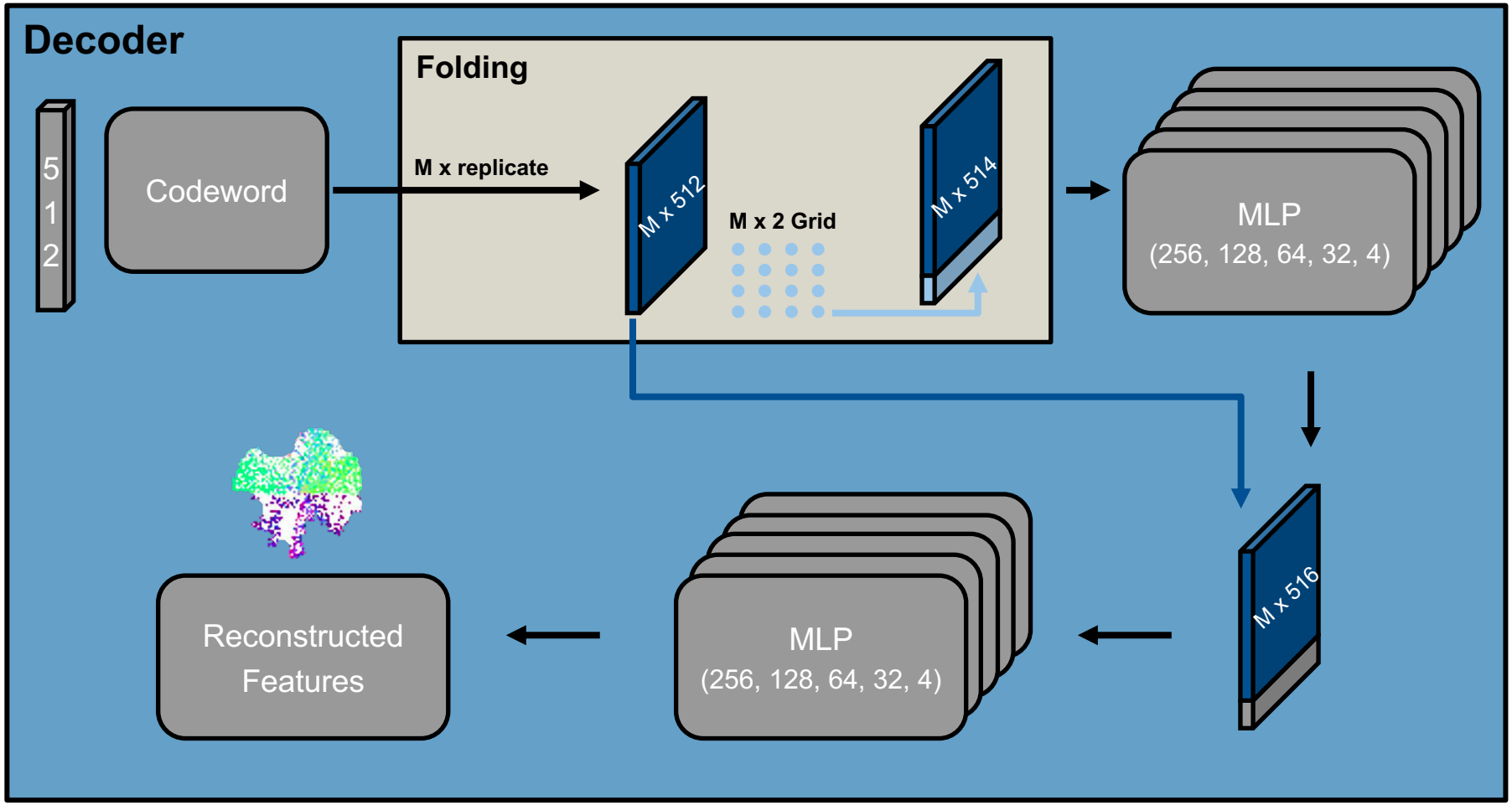
PPF-FoldNet: Network Architecture



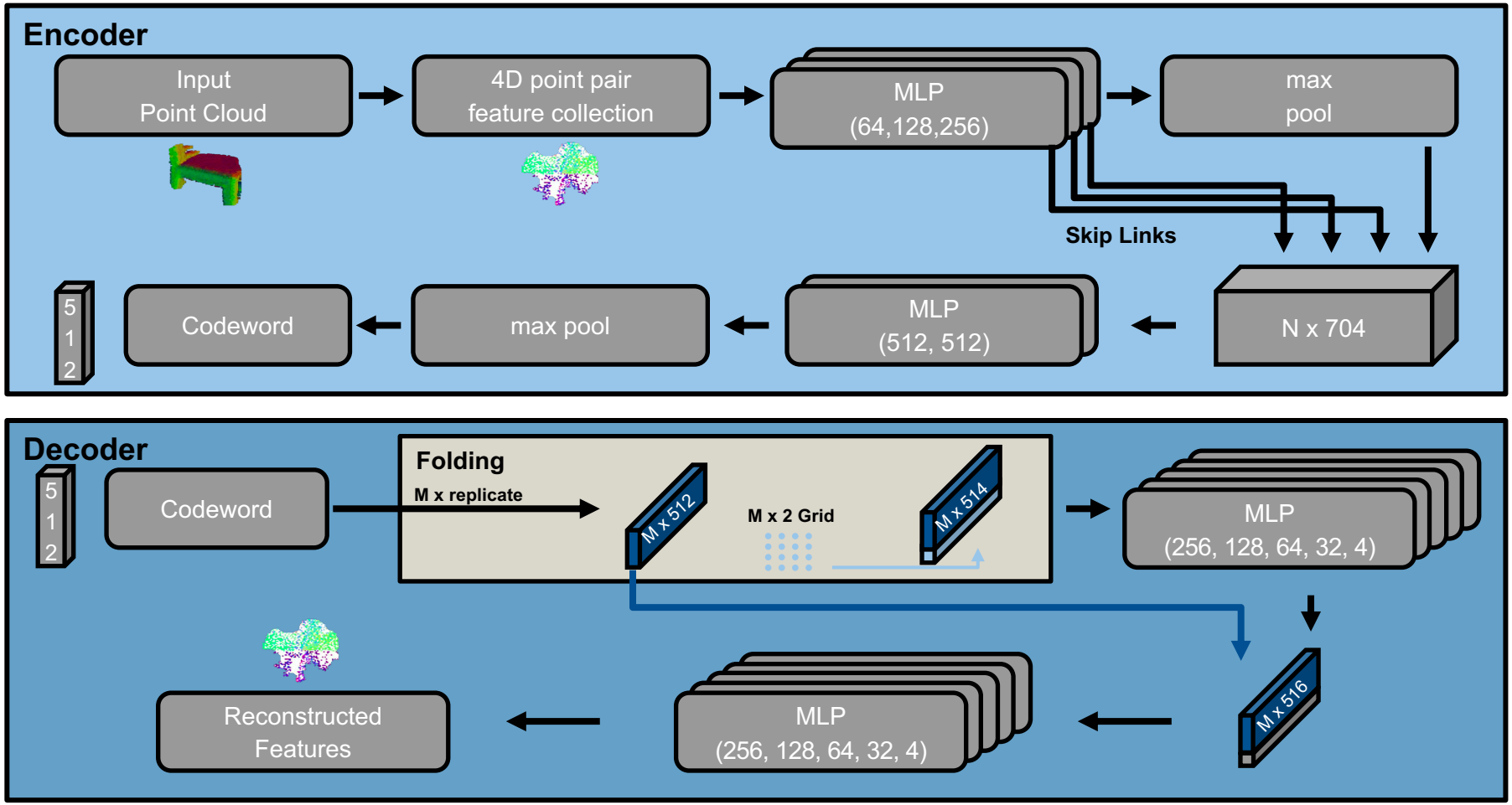
PPF-FoldNet: Network Architecture



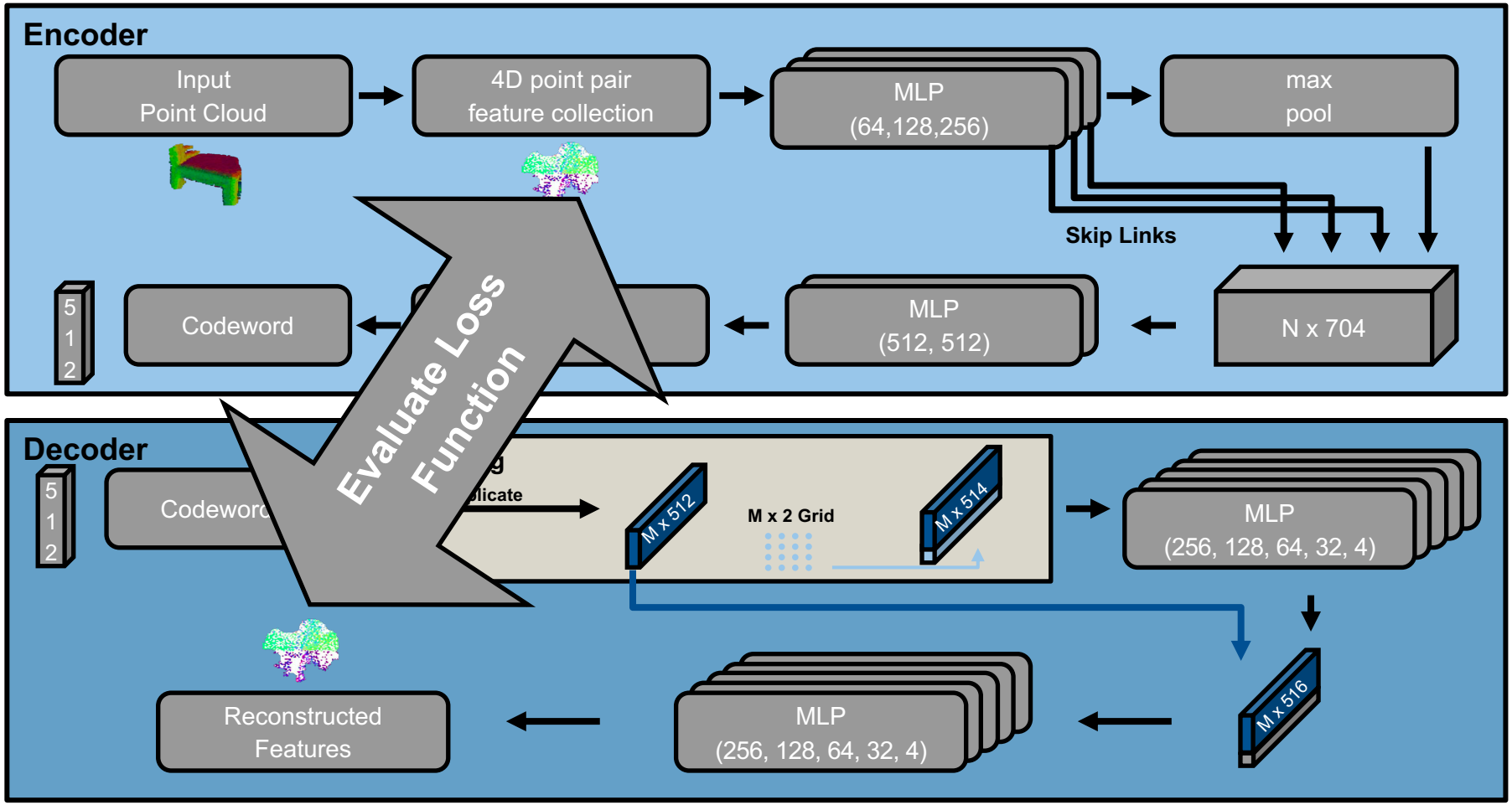
PPF-FoldNet: Network Architecture



PPF-FoldNet: Network Architecture



PPF-FoldNet: Network Architecture



PPF-FoldNet: Optimization

- **Input:** N x 4
- **Output:** M x 4

- **How to evaluate the loss function** for two sets of unequal cardinality?
  - **Chamfer Loss** for two sets $\boldsymbol{F}, \widehat{\boldsymbol{F}}$:

$$d(\boldsymbol{F}, \widehat{\boldsymbol{F}}) = \max\left\{ \frac{1}{|\boldsymbol{F}|} \sum_{f \in \boldsymbol{F}} \min_{\hat{f} \in \widehat{\boldsymbol{F}}} \left\| f - \hat{f} \right\|_2 \quad , \quad \frac{1}{|\widehat{\boldsymbol{F}}|} \sum_{\hat{f} \in \widehat{\boldsymbol{F}}} \min_{f \in \boldsymbol{F}} \left\| f - \hat{f} \right\|_2 \right\}$$

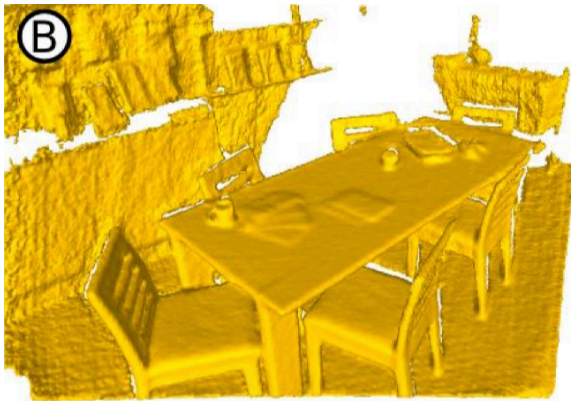
Experimental Validation: Data Preparation

- Use of **3DMatch** Benchmark Dataset (Princeton University):
  - 62 scenes (54 for training & evaluation, 8 for testing)
  - Provides fragments fused from 50 consecutive depth frames
  - Only 3D shape is used

- Fragments are downsampled with **spatial uniformity**

- Points within 30 cm vicinity of each reference point form a **local patch**

- For comparison: patches downsampled to **2048 points** each (also extended version with **5000 points**)



Experimental Validation: Accuracy Evaluation

- **How to evaluate the accuracy** of the network?

- For a given pair of fragments $\boldsymbol{P}$ and $\boldsymbol{Q}$ which can be aligned by a rigid-body transformation $\boldsymbol{T}$, we define:

  - Set of matches from network ($g(\cdot)$ mapping from input to feature space):

$$\boldsymbol{M} = \left\{\{\boldsymbol{p_i}, \boldsymbol{q_i}\} : g(\boldsymbol{p_i}) = NN\left(g(\boldsymbol{q_i}), g(\boldsymbol{P})\right), g(\boldsymbol{q_i}) = NN\left(g(\boldsymbol{p_i}), g(\boldsymbol{Q})\right)\right\}$$

  - Set of Ground-Truth matches $\boldsymbol{M_{GND}}$:

$$\boldsymbol{M_{GND}} = \left\{\{\boldsymbol{p_i}, \boldsymbol{q_i}\} : (\boldsymbol{p_i}, \boldsymbol{q_i}) \in \boldsymbol{M}, \|\boldsymbol{p_i} - \boldsymbol{T}\boldsymbol{q_i}\|_2 < \tau_1\right\}$$

Experimental Validation: Accuracy Evaluation

- **How to evaluate the accuracy** of the network?

- With previous definitions, we can define an **inlier ratio r_{in}**:

$$r_{in} = \frac{|\boldsymbol{M_{GND}}|}{|\boldsymbol{M}|} > \tau_2$$

- As feature quality measure we define the **recall $\boldsymbol{R}$** for fragment pairs **S** considered to match under ground-truth alignment:

$$R = \frac{1}{|\boldsymbol{S}|} \sum_{i=1}^{|\boldsymbol{S}|} \mathbb{I}\left(r_{in}\left(\boldsymbol{S_i} = (\boldsymbol{P_i}, \boldsymbol{Q_i})\right) > \tau_2\right)$$

Experimental Validation: Results

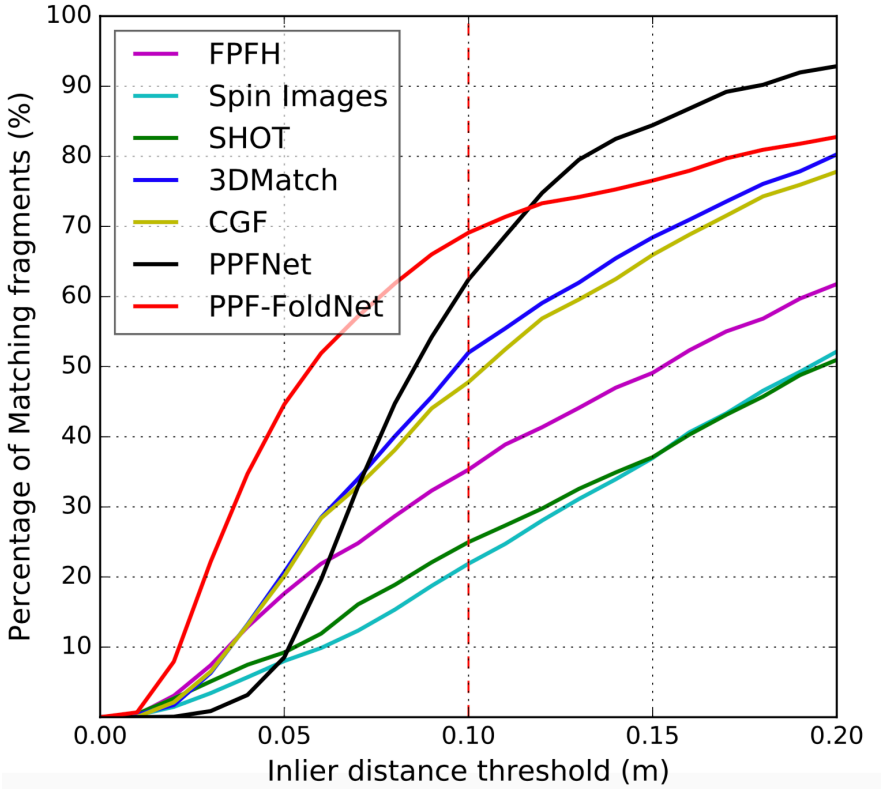
- For comparing PPF-FoldNet against other works on 3DMatch benchmark, the parameters are set to $\boldsymbol{\tau_1 = 10cm}$ and $\boldsymbol{\tau_2 = 5\%}$:

	Spin Image	FPFH	3DMatch	PPFNet	FoldNet	PPF-FoldNet	PPF-FoldNet 5K
Kitchen	0.1937	0.3063	0.5751	**0.8972**	0.5949	0.7352	0.7866
Home 1	0.3974	0.5833	0.7372	0.5577	0.7179	0.7564	**0.7628**
Home 2	0.3654	0.4663	**0.7067**	0.5913	0.6058	0.625	0.6154
Hotel 1	0.1814	0.2611	0.5708	0.5796	0.6549	0.6593	**0.6814**
Hotel 2	0.2019	0.3269	0.4423	0.5769	0.4231	0.6058	**0.7115**
Hotel 3	0.3148	0.5000	0.6296	0.6111	0.6111	0.8889	**0.9444**
Study	0.0548	0.1541	0.5616	0.5342	**0.7123**	0.5753	0.6199
MIT Lab	0.1039	0.2727	0.5455	**0.6364**	0.5844	0.5974	0.6234
Average	0.2267	0.3589	0.5961	0.6231	0.6130	0.6804	**0.7128**

Experimental Validation: Results

- Influence of **inlier distance threshold τ_1** and **inlier ratio threshold τ_2** is investigated:

  ➔ PPF-FoldNet outperforms other methods especially for **strict distance requirements**

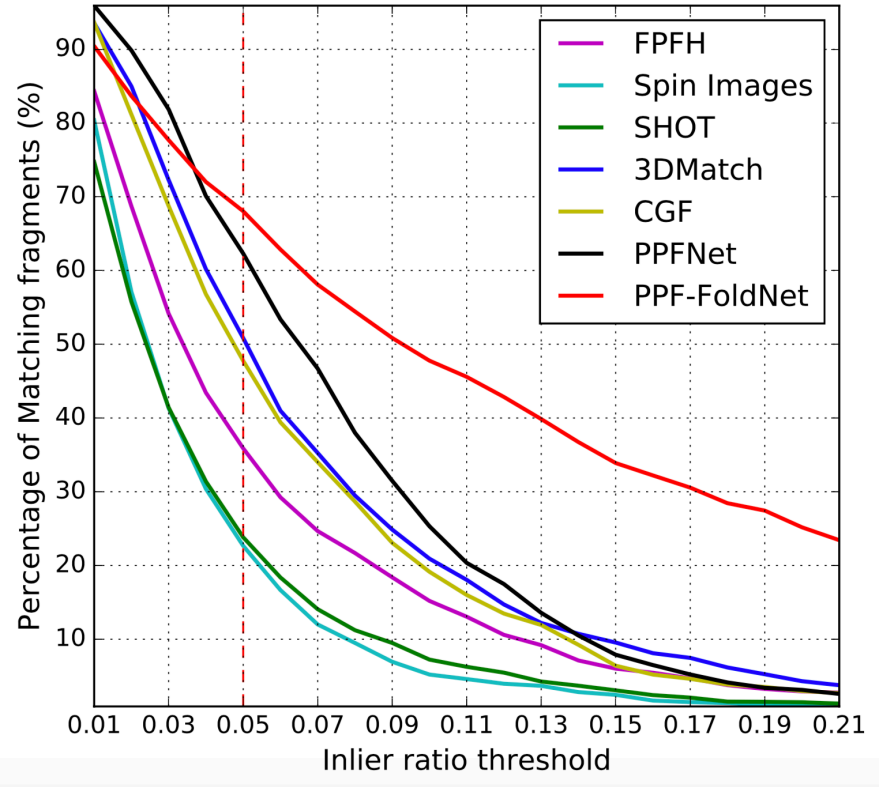


Experimental Validation: Results

- Influence of **inlier distance threshold τ_1** and **inlier ratio threshold τ_2** is investigated:

  ➔ PPF-FoldNet outperforms other methods especially for **strict distance requirements**

  ➔ PPF-FoldNet outperforms other methods especially when demanding **high inlier ratios**

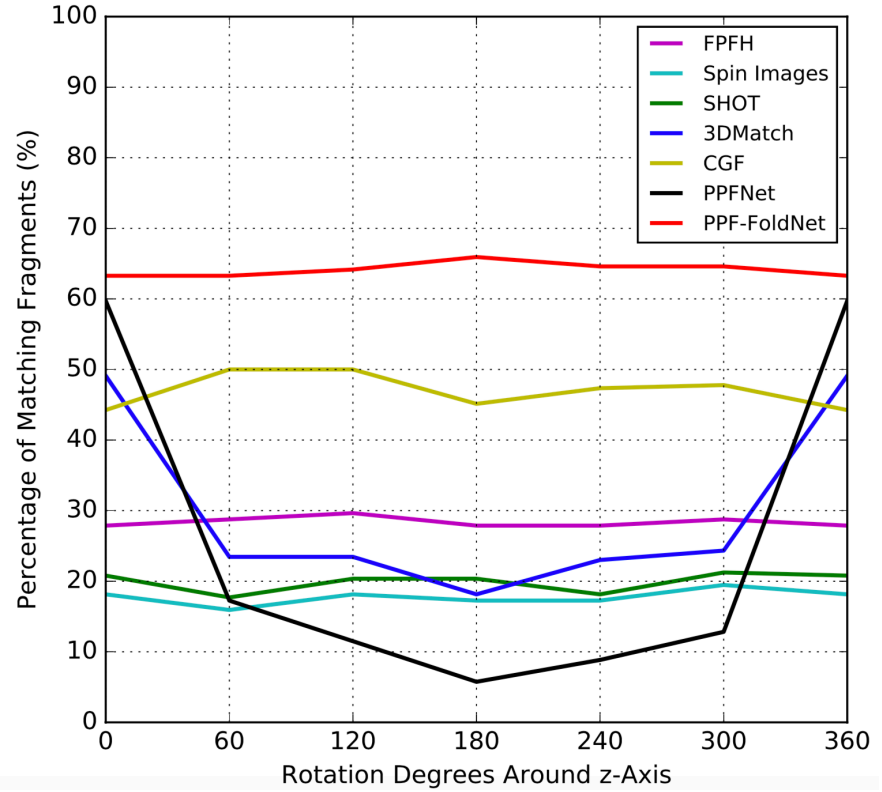


Experimental Validation: Rotation Invariance

- To test the presented method on rotation invariance, **random fragments** are **gradually rotated** around the z-axis from 60° to 360° in steps of 60°

  ➔ Recall does not depend on rotation angle

  ➔ PPF-FoldNet outperforms every other method



Experimental Validation: Rotation Invariance

- Introducing a new benchmark by **rotating all fragments** randomly (*Rotated 3DMatch Benchmark*) and comparing the performance again shows robustness of PPF-FoldNet to rotations:

	Spin Image	FPFH	3DMatch	PPFNet	FoldNet	PPF-FoldNet	PPF-FoldNet 5K
Kitchen	0.1779	0.2905	0.004	0.002	0.0178	0.7352	**0.7885**
Home 1	0.4487	0.5897	0.0128	0.0000	0.0321	0.7692	**0.7821**
Home 2	0.3413	0.4712	0.0337	0.0144	0.0337	0.6202	**0.6442**
Hotel 1	0.1814	0.3009	0.0044	0.0044	0.0133	0.6637	**0.6770**
Hotel 2	0.1731	0.2981	0.0000	0.0000	0.0096	0.6058	**0.6923**
Hotel 3	0.3148	0.5185	0.0096	0.0000	0.0370	0.9259	**0.963**
Study	0.0582	0.1575	0.0000	0.0000	0.0171	0.5616	**0.6267**
MIT Lab	0.1169	0.2857	0.026	0.0000	0.0260	0.6104	**0.6753**
Average	0.226**5**	0.364	0.0113	0.0026	0.0233	0.6865	**0.7311**

Summary

- ✓ **Unsupervised autoencoder** structure to learn **local features**
- ✓ Operates on **point pair features** (PPF)
- ✓ **Combines best attributes** of its ancestors (PointNet, FoldingNet, PPFNet)
- ✓ Outperforms most state-of-the-art approaches under varying conditions (**rotation invariance**, **point cloud density**)
- ✓ PPF-FoldNet can be shown to **generalize** well on **unseen input**.

References

[1] Deng H., Birdal,T., Ilic, S.: „PPF-FoldNet: Unsupervised Learning of Rotation Invariant 3D Local Descriptors“. European Conference on Computer Vision (ECCV) 2018.

[2] Qi, C.R., Su, H., Mo, K., Guibas, L.J.: „Pointnet: Deep learning on point sets for 3d classi- fication and segmentation“. Proc. Computer Vision and Pattern Recognition (CVPR) 2017.

[3] Yang, Y., Feng, C., Shen, Y., Tian, D.: „Foldingnet: Point cloud auto-encoder via deep grid deformation“. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2018.

[4] Deng, H., Birdal, T., Ilic, S.: „Ppfnet: Global context aware local features for robust 3d point matching“. Computer Vision and Pattern Recognition (CVPR) 2018.

[5] Zeng, A., Song, S., Nießner, M., Fisher, M., Xiao, J., Funkhouser, T.: “3DMatch: Learning Local Geometric Descriptors from RGB-D Reconstructions”. Computer Vision and Pattern Recognition (CVPR) 2017.

Thanks for your attention.

Backup

Implementation Details & Hardware Configuration

- Implementation Details:
  - **TensorFlow**
  - Variables initialized by **Xavier‘s algorithm**
  - **ADAM** optimizer
  - **Exponentially decaying learning rate** (starting at 1e-3; truncated at 1e-4)
  - **Batch size**: 32

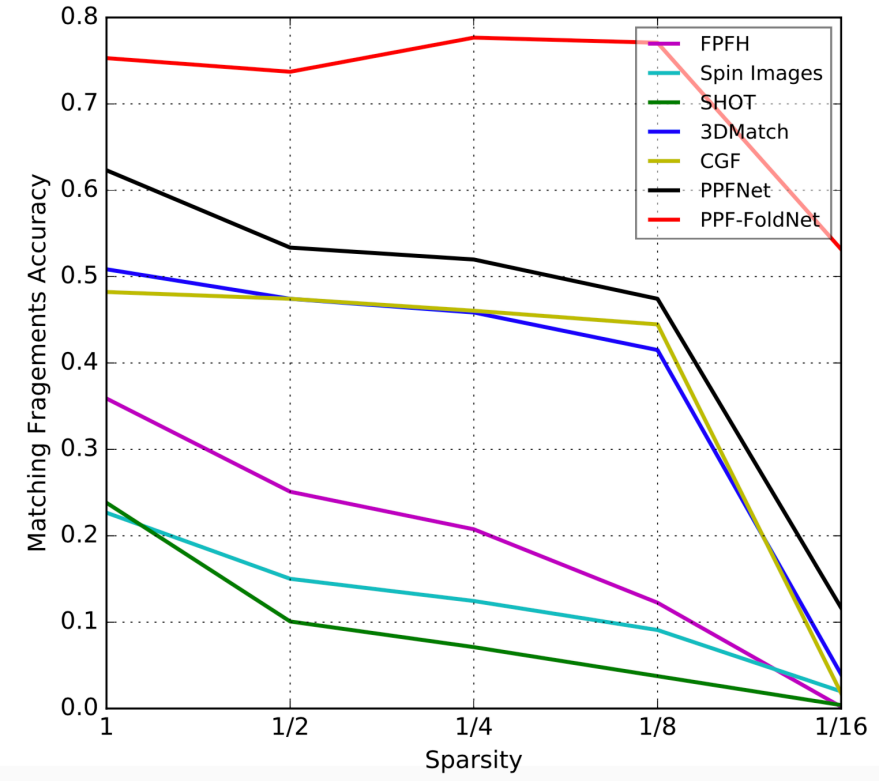
- Hardware Configuration:
  - **NVIDIA TitanX Pascal GPU**
  - **Intel Core i7 3.2 GHz CPU**

Experimental Validation: Results

- PPFNet is shown to be robust against changes in point cloud density.

➔ PPF-FoldNet least affected by decrease in point cloud density.

➔ Common Problem in many point cloud representations



Experimental Validation: Generalizability

- PPF-FoldNet only trained on small portion of scenes
- Chess scene divided into training and test splits
- Loss values measured on all other scenes as well (not part of training)

➔ Loss decreases for all 7 scenes with similar trend

➔ PPF-FoldNet can generalize on unseen input