RSKDD-Net: Random Sample-based Keypoint Detector and Descriptor

Fan Lu, Guang Chen, Yinlong Liu, Zhongnan Qu, Alois Knoll

Institute of Intelligent Vehicles, Tongji University

Motivation

3D Keypoint detector and descriptor are two main components in point cloud registration. However, compared to numerous keypoint detector and descriptor in 2D images, 3D keypoint detector and descriptor have not been deeply explored.

3D keypoint detector and descriptor can be categorized into two groups:

- **Handcrafted keypoint detector and descriptor**
  - Detector: SIFT3D, Harris3D, ISS
  - Descriptor: FPFH, USC, SHOT
- **Learning-based detector and descriptor**
  - 3DFeatNet (ECCV 2018)
  - USIP (ICCV 2019)

Motivation

Limitations of previous learning based works: **high time complexity**

- 3DFeatNet: Estimate saliency for **each point in the point cloud** and select keypoints based on the predicted saliency. The per-point saliency estimation is time-consuming.

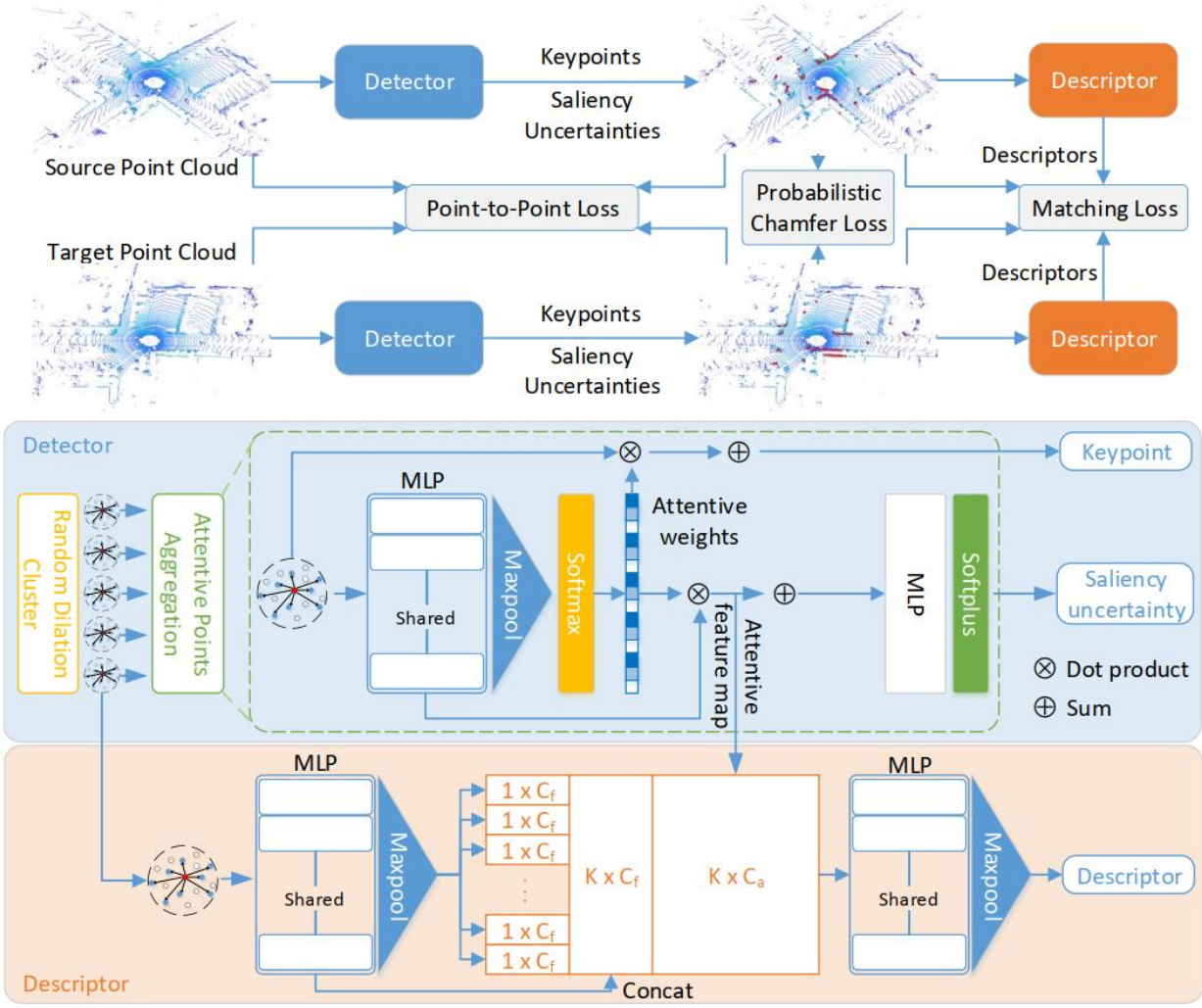
- USIP: Utilize **furthest point sampling (FPS)** to generate keypoint candidates and select keypoints based on saliency estimation. However, FPS is an inefficient sampling method so that is time consuming.

Methodology

We introduce **random sampling** concept for 3D keypoint detector and descriptor. Random sampling is efficient however can cause information loss.

Proposed novel strategies:

- **Random dilation cluster** to enlarge the receptive field.
- **Attentive point aggregation** to generate keypoints and predict saliency uncertainty.
- **Matching loss** to train the descriptor.

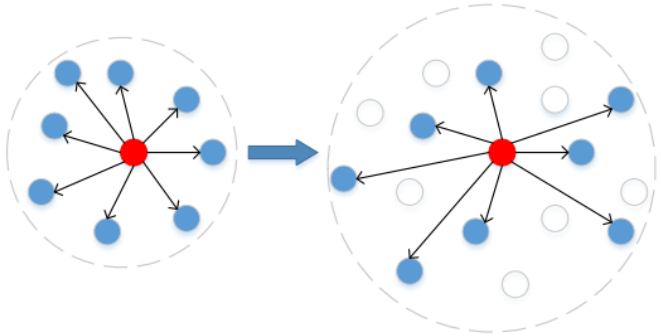


Methodology

Random dilation cluster

Enlarging the receptive field of a random sampled cluster center can weaken the negative impact of random sampling.

- **Generally method**: k-nearest neighbor (knn)/ball query.
- **Our strategy**: query $\alpha_d \times K$ neighbor points and random select K points from them.
- **Effect**: enlarging receptive field from K to $\alpha_d \times K$ without hardly any increase of time complexity.



Methodology

Attentive points aggregation

Previous method for predicting new keypoint:

- Predict a offset for each sampled cluster center point (USIP)
- Just use the raw point (3DFeatNet)

Our method for predicting new keypoint:

- Predict attentive weights for each neighbor points
- The predicted keypoint can be represented as the weighted sum of neighbor points

The advantage of our method:

- The predicted keypoint is within the convex hull of the input cluster
- The attention mechanism tends to give higher weights for informative points

Methodology

Matching loss

Soft assignment strategy: Explicitly estimate the **correspondences between keypoints** in two point clouds according to the Euclidean distance of the descriptors.

The corresponding keypoint are represented as **weighted sum** of all keypoints in the other point cloud

$$s_{ij}^{\mathcal{S}} = \frac{\exp(\frac{1/d_{ij}^{\mathcal{S}}}{t})}{\sum_{j=1}^{M} \exp(\frac{1/d_{ij}^{\mathcal{S}}}{t})} \qquad \hat{x}_i^{\mathcal{S}} = \sum_{j=1}^{M} s_{ij}^{\mathcal{S}} \cdot \tilde{x}_j^{\mathcal{T}}$$

The proposed matching loss can be seen as the distance between corresponding keypoints.

$$\mathcal{L}_{matching} = \sum_{i=1}^{M} \tilde{w}_i^{\mathcal{S}} \left\| \mathbf{R}\tilde{x}_i^{\mathcal{S}} + \mathbf{t} - \hat{x}_i^{\mathcal{S}} \right\|_2^2 + \sum_{i=1}^{M} \tilde{w}_i^{\mathcal{T}} \left\| \mathbf{R}\hat{x}_i^{\mathcal{T}} + \mathbf{t} - \tilde{x}_i^{\mathcal{T}} \right\|_2^2$$

Advantages of proposed matching loss:

- Weakly supervised
- Explicitly estimate the correspondence

Experiments

Dataset

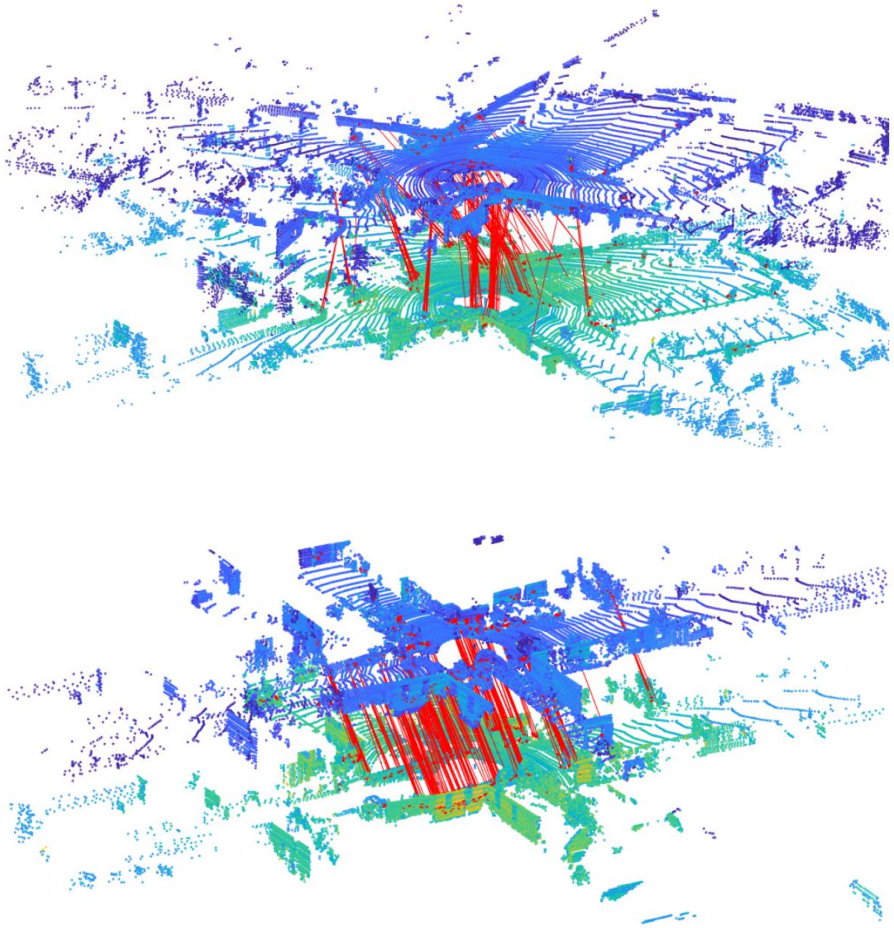
- KITTI Odoemtry dataset
- Ford Campus Vision and LiDAR dataset

Baselines

- Handcrafted detector and descriptor
  - Harris3D + FPFH
  - SIFT3D + FPFH
  - ISS + FPFH
- Learning-based detector and descriptor
  - 3DFeatNet
  - USIP

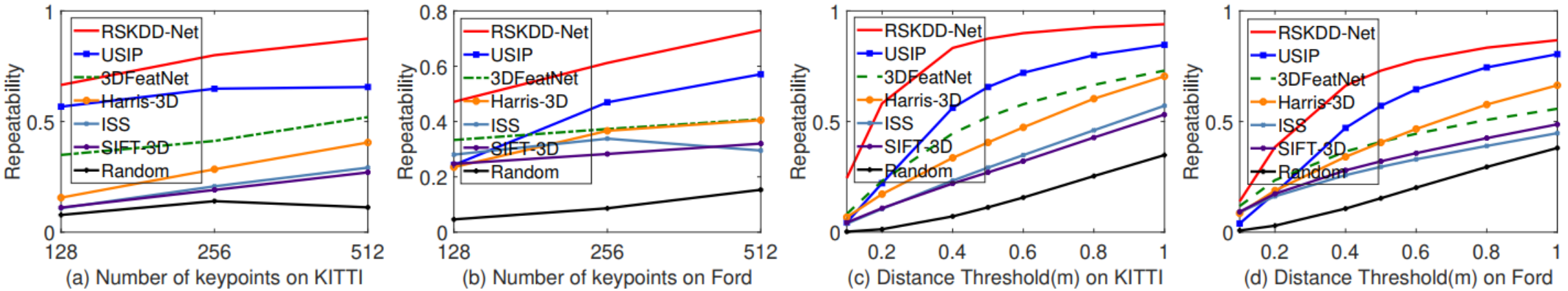
Evaluation metrics

- Repeatability
- Precision
- Registration performance

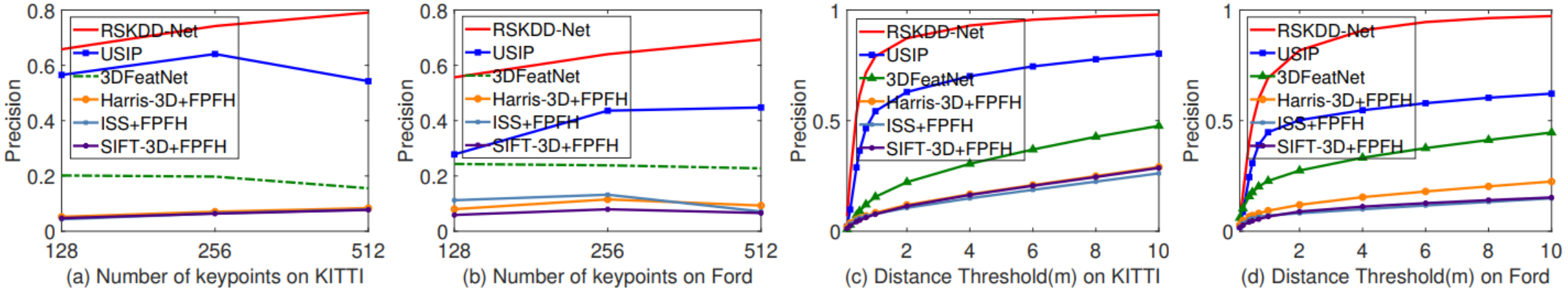


Experiments

Repeatability



Precision



Experiments

Registration performance

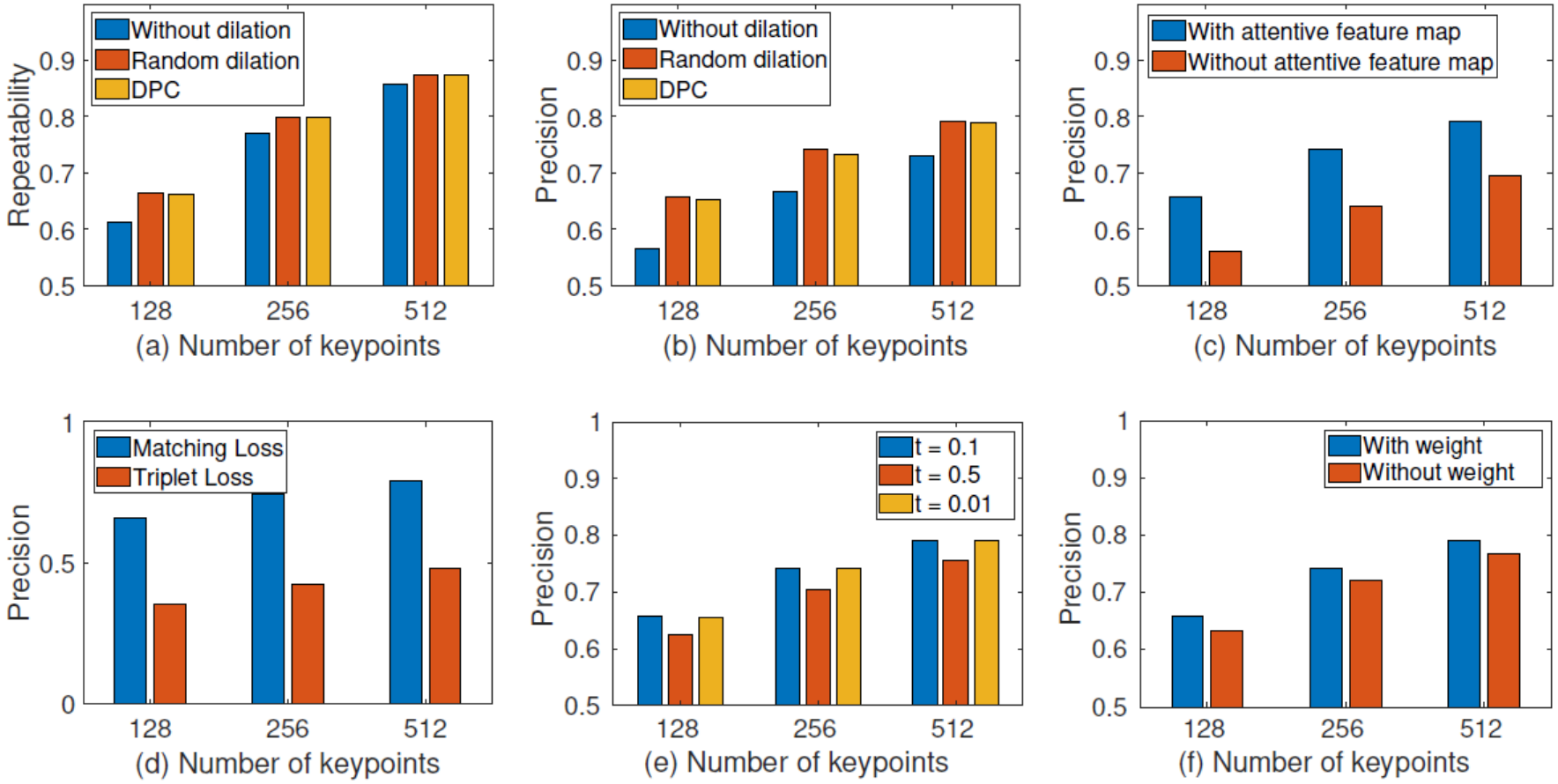
Methods	KITTI dataset					Ford dataset				
	RTE (m)	RRE (deg)	Inlier	Iter	Success	RTE (m)	RRE (deg)	Inlier	Iter	Success
Harris+FPFH	0.38 ± 0.33	1.79 ± 1.24	0.018	10000	82.9%	0.51 ± 0.59	0.48 ± 0.90	0.187	935	74.0%
ISS+FPFH	0.59 ± 0.39	1.24 ± 0.98	0.024	10000	92.3%	0.54 ± 0.56	0.70 ± 1.16	0.160	1364	74.4%
SIFT+FPFH	0.57 ± 0.39	1.39 ± 0.96	0.040	9973	92.5%	0.56 ± 0.56	0.68 ± 1.11	0.243	376	75.3%
3DFeatNet	0.31 ± 0.26	0.73 ± 0.64	0.093	6591	97.9%	0.37 ± 0.42	0.61 ± 0.73	0.100	5642	91.3%
USIP	0.10 ± 0.05	$\mathbf{0.35 \pm 0.21}$	0.243	468	**100%**	0.12 ± 0.06	$\mathbf{0.38 \pm 0.39}$	0.195	870	**100%**
RSKDD-Net	$\mathbf{0.09 \pm 0.06}$	0.50 ± 0.28	**0.586**	**32**	99.9%	$\mathbf{0.11 \pm 0.08}$	0.58 ± 0.41	**0.505**	**41**	99.5%

Runtime

Input points	4096			8192			16384		
Keypoints	128	256	512	128	256	512	128	256	512
3DFeatNet	66.8	70.8	81.4	136.2	156.2	169.9	367.6	413.7	420.7
USIP	76.6	163.6	296.7	99.6	171.0	310.9	115.2	203.4	378.5
RSKDD-Net	**3.8**	**4.1**	**4.7**	**4.3**	**5.2**	**6.5**	**5.7**	**8.5**	**10.1**

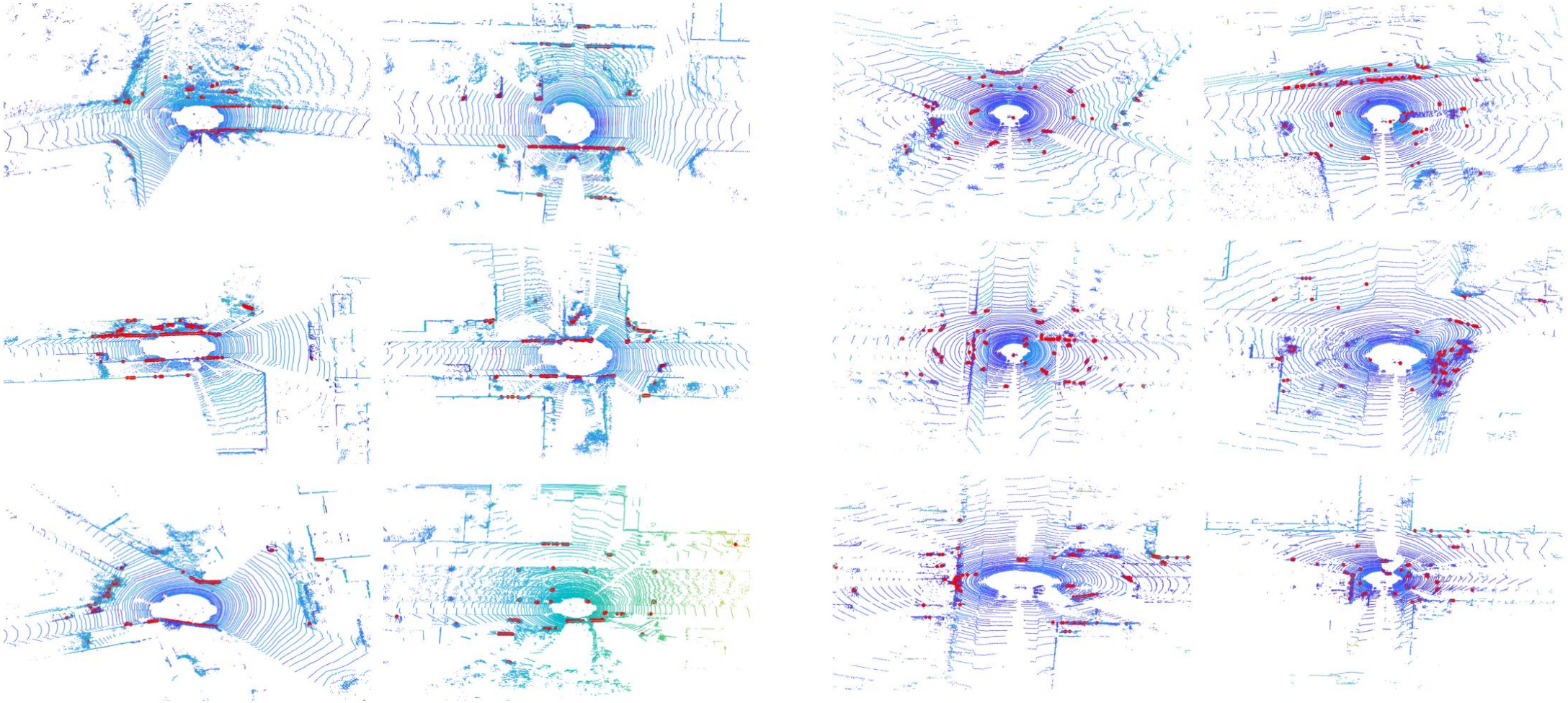
Experiments

Ablations



Experiments

- More qualitative results



Thank you !