DH3D:
Deep Hierarchical 3D Descriptors for Robust Large-Scale 6DoF Relocalization

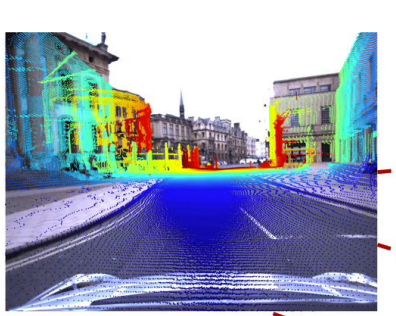
Juan Du, Rui Wang, Daniel Cremers
ECCV 2020

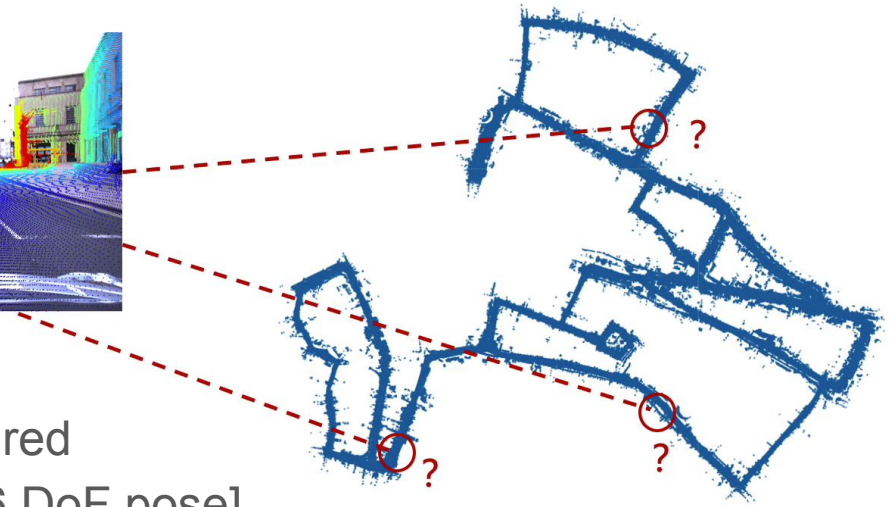
presented by Vojtěch Pánek
RMP AAG, CIIRC, CTU

Large-scale point cloud relocalization task

Inputs:

- map [point cloud]
- query [point cloud]



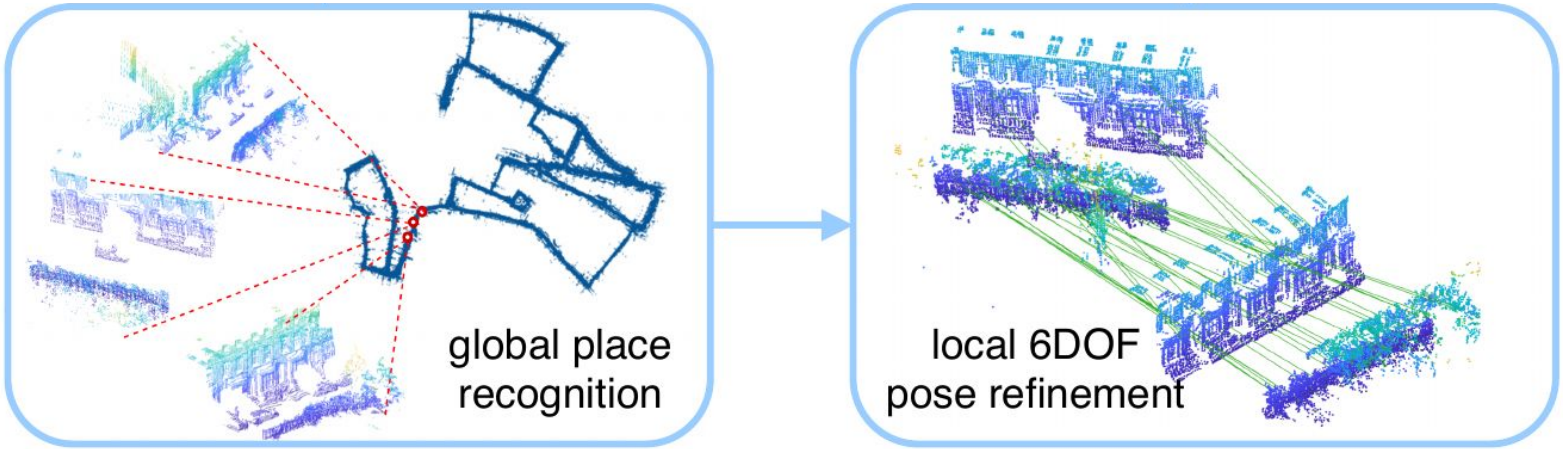


Output:

- pose from which the query was captured (within the map coordinate system) [6 DoF pose]

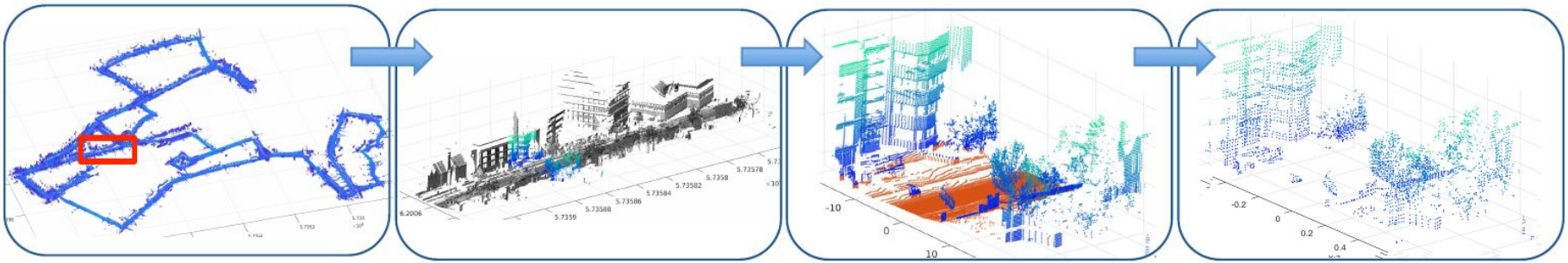
Splitting the problem

- global place recognition - get coarse pose estimate
  - point cloud retrieval
- local pose refinement - use the coarse estimate and produce a precise pose
  - point cloud registration



Global place recognition

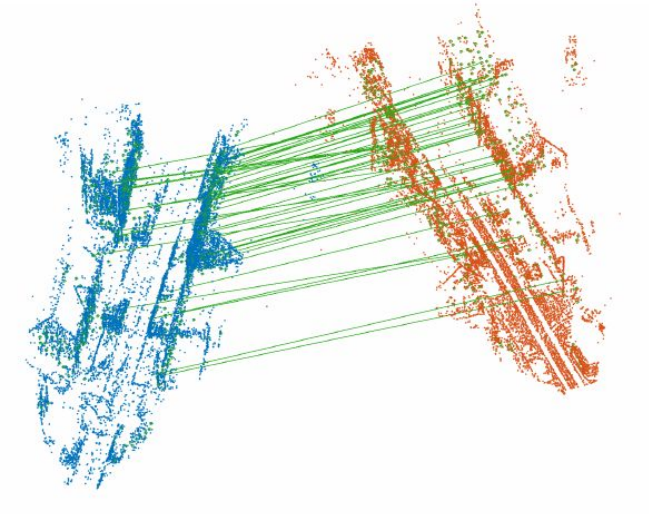
- point cloud retrieval
  - get the most similar point cloud to the query one and use its pose within the map as the coarse pose estimate of the query
- cutouts (submaps) for retrieval database
  - analogy with images in image retrieval database
  - 20 m cutout length, 10 m stride, downsample to 0.2 m voxel grid

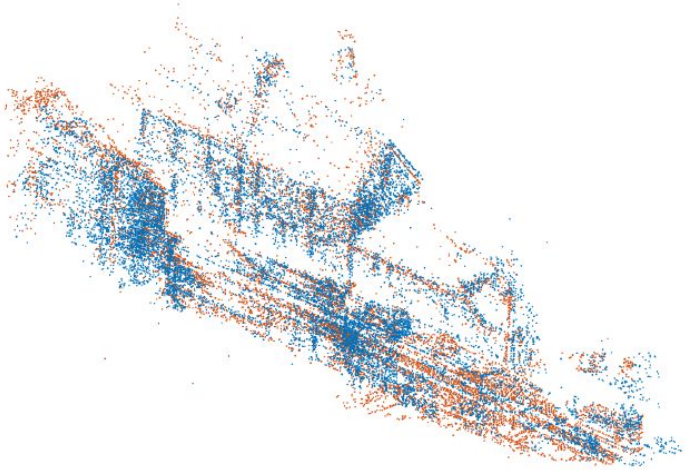


(a) Full route (b) Zoomed-in region (c) Submap (d) Downsampled [2]

Local pose refinement

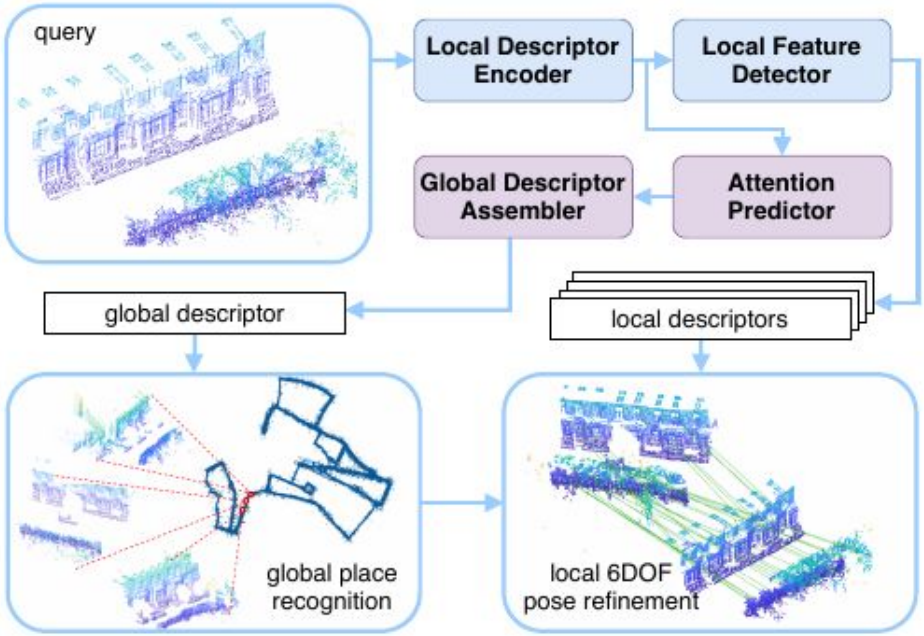
- point cloud registration
  - improve the coarse pose estimate by searching for the transformation between query and submap point clouds





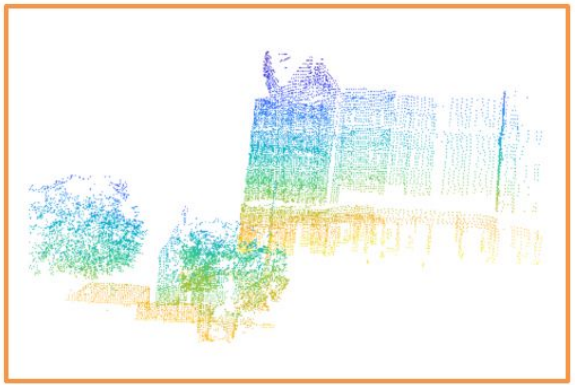
Keypoint detection and description

- local descriptors
  - used for point cloud registration
  - describe-and-detect pipeline
    - describe all points
    - detect the important ones
- global descriptor
  - describes the whole point cloud
  - generated by aggregation of local descriptors
- local and global descriptors generated in single forward pass



Parallels with image domain

- detect and describe salient points / patches
  - filter out non-salient samples
  - create description useful for sample matching
  - problem: unstructured samples
- instance retrieval
  - quickly obtain coarse pose estimate
  - problem: instance definition
- global description by aggregation
  - reuse the local features
  - image domain: BoW [20, 21], VLAD [22], NetVLAD [23]
- joint description and detection of local features
  - detector has more information
  - image domain: e.g. D2-Net [16]



State of the Art - local descriptors

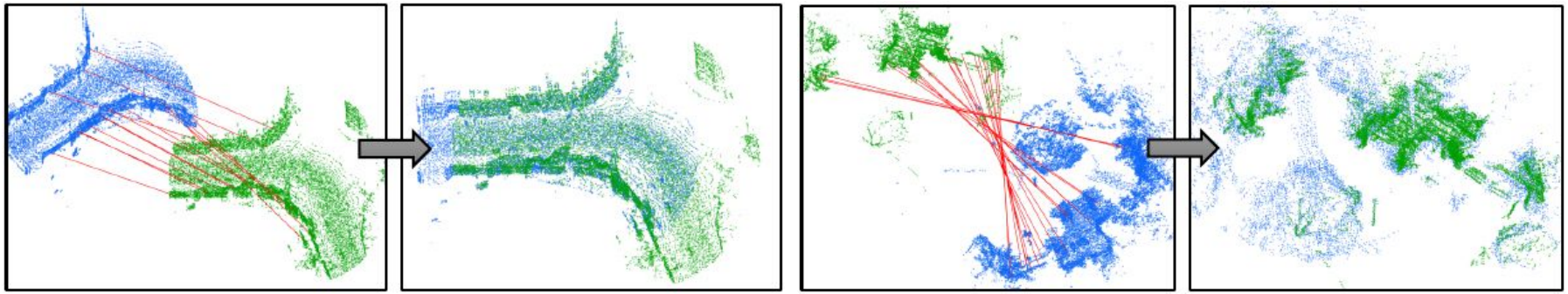
Unique Shape Context (USC) [5]

Point Feature Histogram (PFH) [6]

Fully Convolutional Geometric Features (FCGF) [7]

3DFeatNet [8]

[8]



State of the Art - local detectors

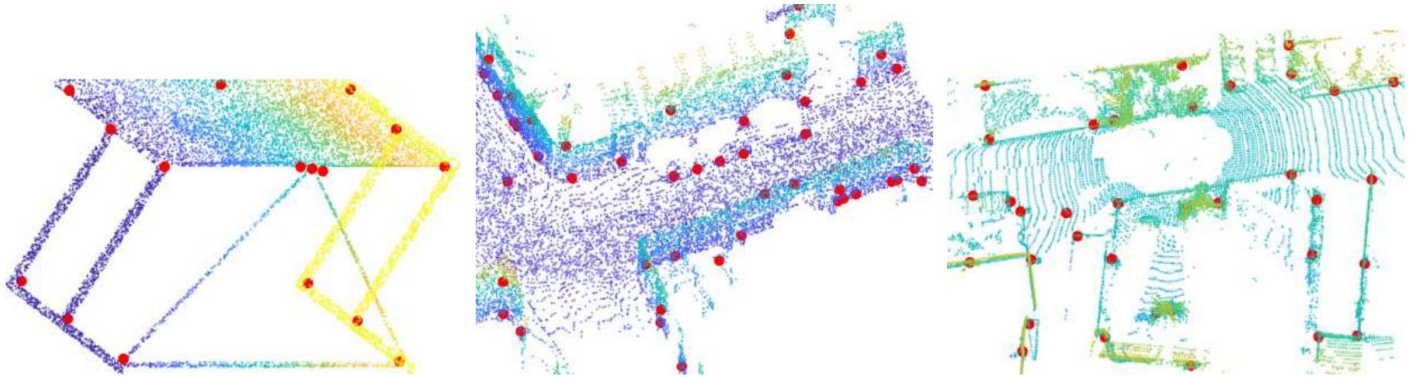
Intrinsic Shape Signatures (ISS) [9]

SIFT-3D [10]

Harris-3D [11]

3DFeatNet [8]

USIP [12]



[12]

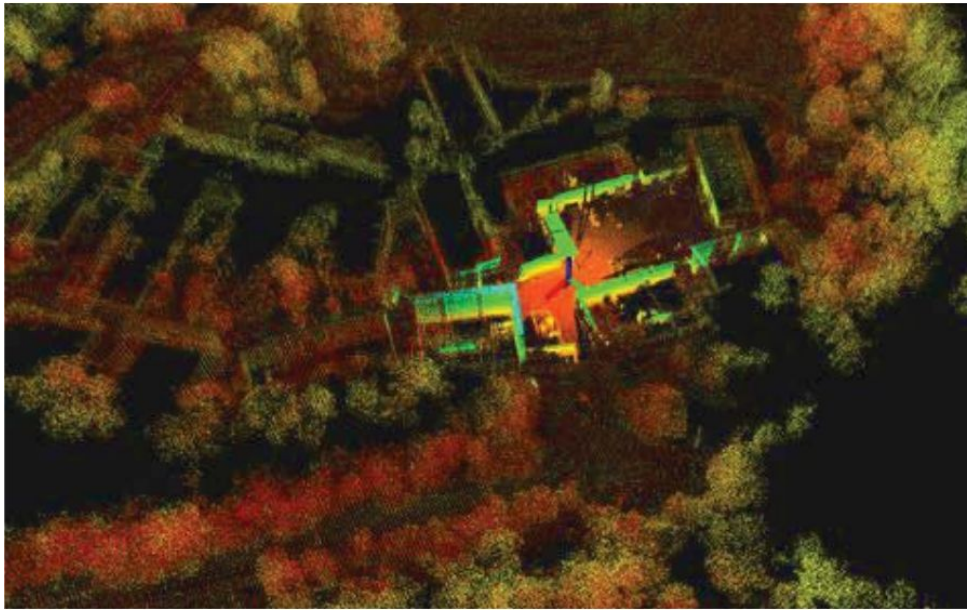
State of the Art - global descriptors

histograms of points elevation [13]

DELIGHT [14]

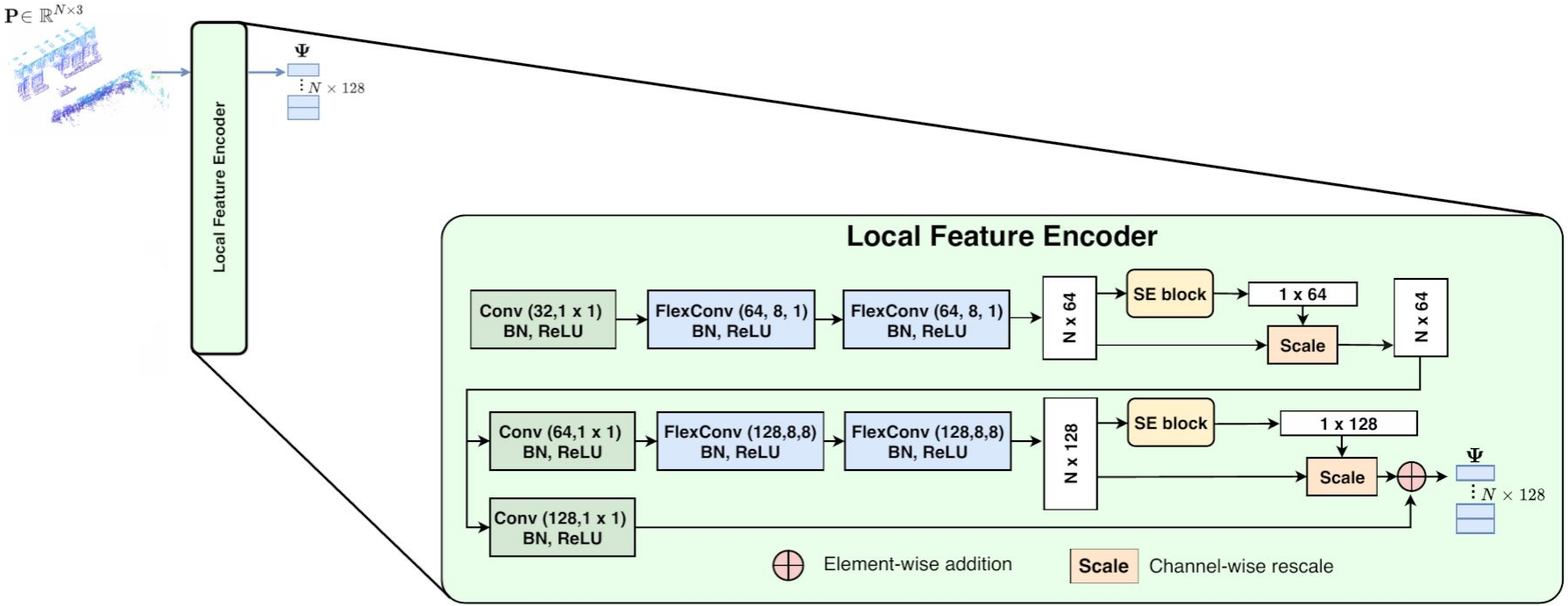
Point-NetVLAD [2]

LocNet [16]



[14]

Local features - descriptor



Flex Convolution (FlexConv) [17]

- encodes local spatial relationship between points
- grid neighborhood in convolution replaced by spatial distance of the points

$$f_{FlexConv}(\boldsymbol{p}_l) = \sum_{\boldsymbol{p}_{l_i} \in N_k(\boldsymbol{p}_l)} \omega(\boldsymbol{p}_{l_i}, \boldsymbol{p}_l) \cdot h(\boldsymbol{p}_{l_i})$$

$$w(\boldsymbol{p}_l, \boldsymbol{p}_{l_i} \mid \theta, \theta_b) = \langle \theta, \boldsymbol{p}_l - \boldsymbol{p}_{l_i} \rangle + \theta_b$$

$\boldsymbol{p}_l$ = point

N_k = k-NN of $\boldsymbol{p}_l$

$\omega(\boldsymbol{p}_{l_i}, \boldsymbol{p}_l)$ = kernel function

$h(\boldsymbol{p}_{l_i})$ = point-wise encoding function

$\langle \theta, \boldsymbol{p}_l - \boldsymbol{p}_{l_i} \rangle$ = matrix to vector multiplication

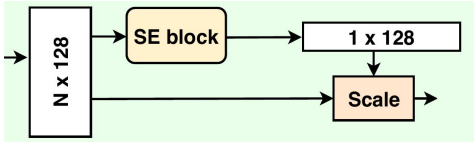
$h(\boldsymbol{p}_{l_i}) \in \mathbb{R}^C$

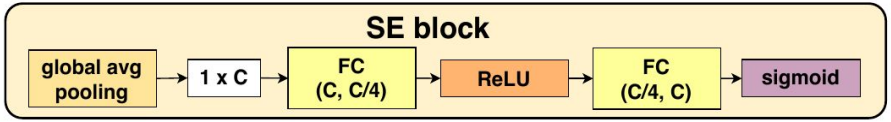
$\theta \in \mathbb{R}^{C \times 3}$

$\theta_b \in \mathbb{R}^C$

Squeeze-and-Excitation (SE) [18]

- models dependencies between channels coming from Flex Convolution





- squeeze operation - global average pooling over points

$$z = f_{sq}(\boldsymbol{U}) \qquad f_{sq} : \mathbb{R}^{N \times C} \rightarrow \mathbb{R}^{C}$$

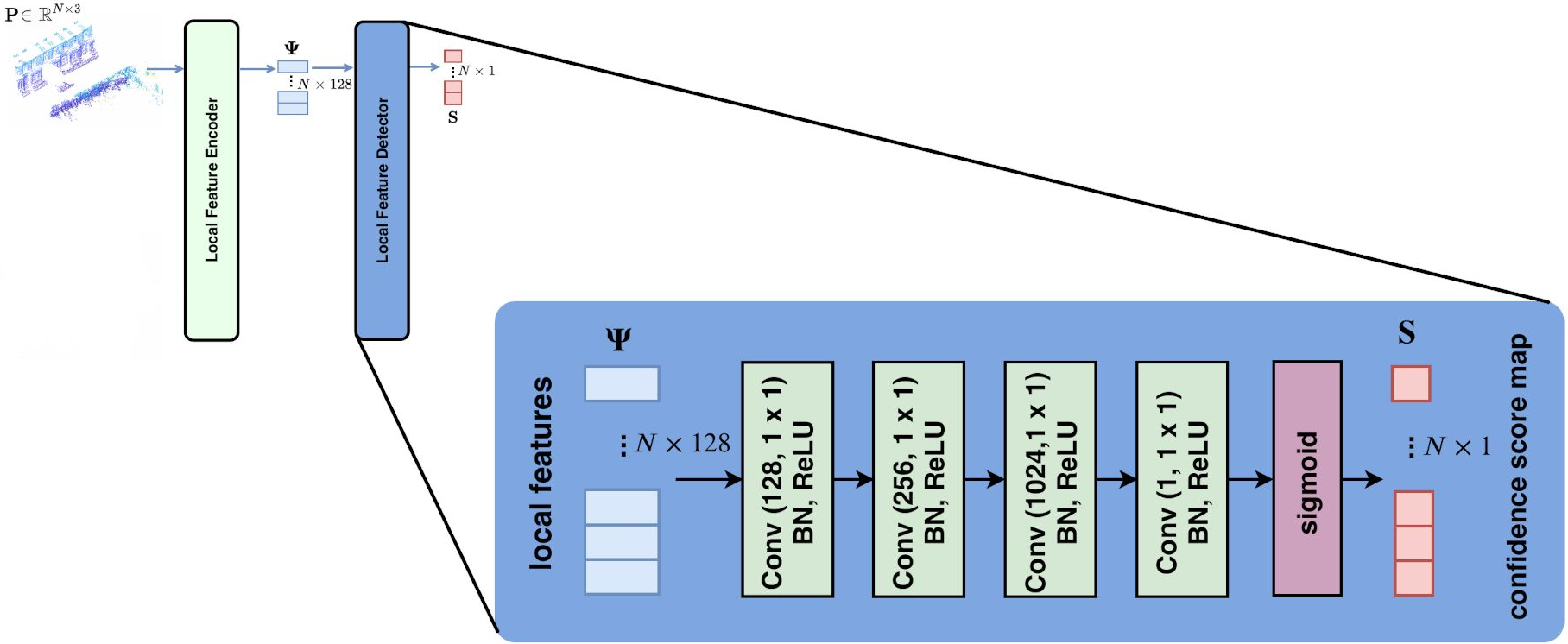
- excitation operation - capture channel-wise dependencies

$$s = f_{ex}(z) \qquad f_{ex} : \mathbb{R}^{C} \rightarrow \mathbb{R}^{C}$$

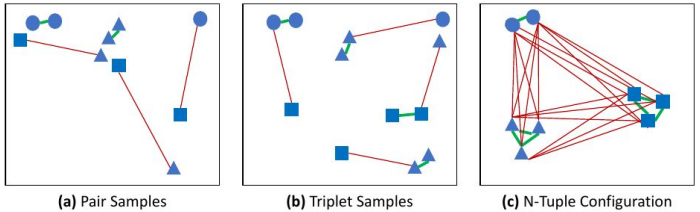
- scale operation - attention selection of different channel-wise features

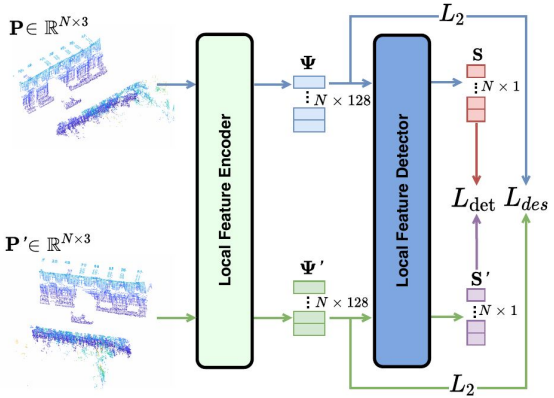
$$f_{scale}(u_c, s_c) = s_c \circ u_c$$

Local features - detector



Local feature learning





- loss used for training

$$L = L_{desc} + \lambda L_{det}$$

- description loss - N-tuple loss [19]

$$L_{desc} = \sum{}^{*} \left(\frac{\boldsymbol{M} \circ \boldsymbol{D}}{\|\boldsymbol{M}\|_F^2} + \eta \frac{max(\mu - (1 - \boldsymbol{M}) \circ \boldsymbol{D}, 0)}{N^2 - \|\boldsymbol{M}\|_F^2} \right)$$

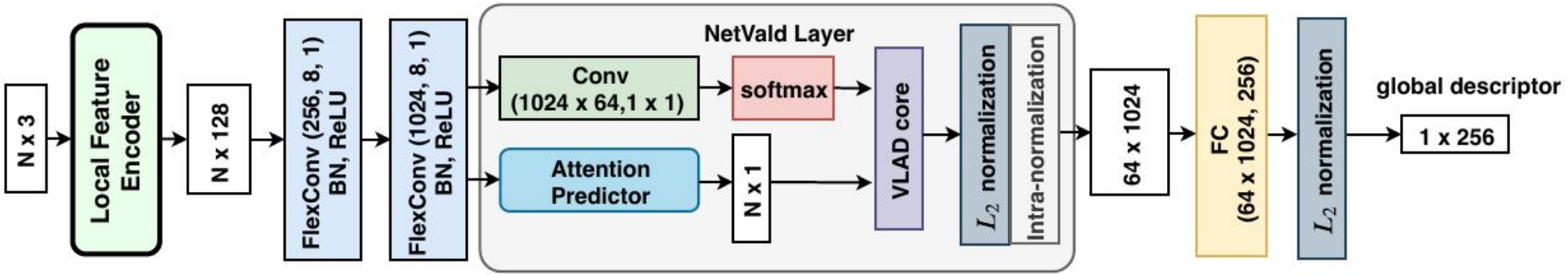
$$\boldsymbol{M} \in \mathbb{R}^{N\times N}, \boldsymbol{M}_{i,j} \in \{0,1\}, \boldsymbol{D} \in \mathbb{R}^{N\times N}, \boldsymbol{D}(i,j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|$$

- detection loss
  - unsupervised learning wrt detection score
  - if point has high saliency score, then its NN (in feature space) should be the correct match
  - single NN unstable → use average successful rate AR_i
    - find k-NN of i-th point, j-th best is the correct match

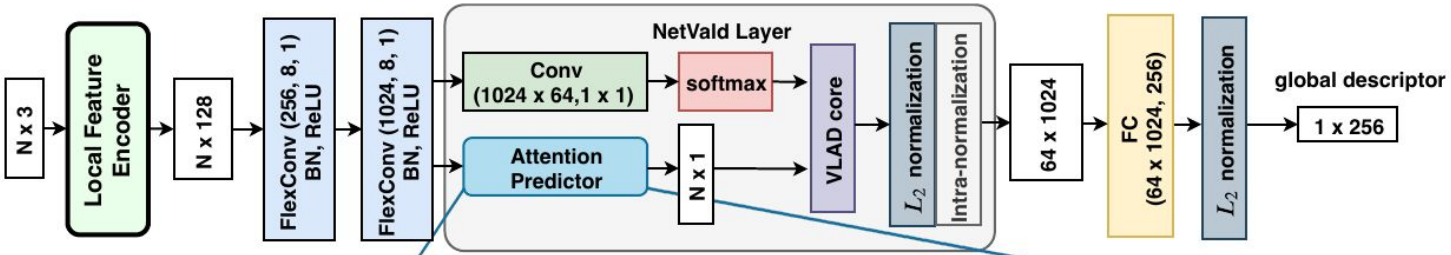
$$L_{det} = \frac{1}{N} \sum_{i=1}^{N} 1 - [\kappa(1 - s_i) + s_i \cdot \text{AR}_i] \quad \text{AR}_i = \frac{k - (j-1)}{k}$$

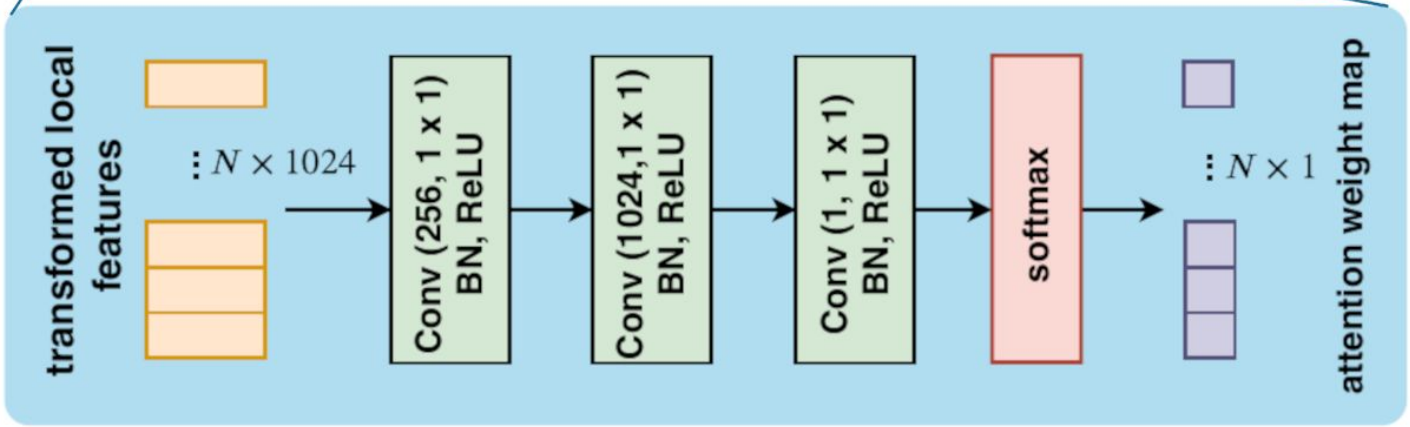
Global point cloud description

- PCAN (Point Contextual Attention Network) [24]
  - Point-NetVLAD [2] extended by attention prediction



Attention predictor





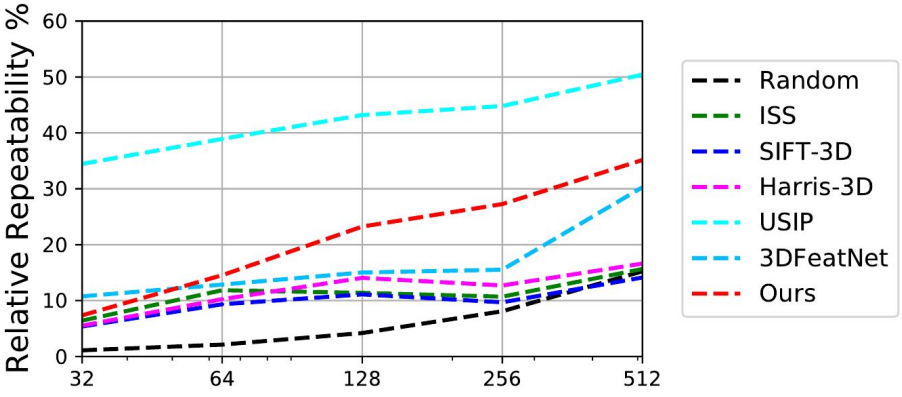
Experiments - ablation study

- local feature training without $L_{det} \rightarrow L = L_{desc}$
- weakly supervised local feature training [8]
- local feature encoder without Squeeze-and-Excitation blocks

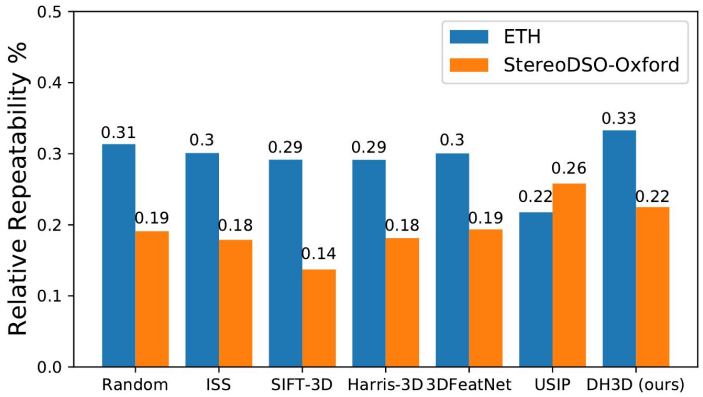
Method	RTE(m)	RRE(°)	Succ.	Iter.
w/o L_{det}	0.43	1.52	93.72	3713
Weak Sup.	0.48	1.78	90.82	3922
w/o SE	0.39	1.24	95.18	3628
Default	**0.23**	**0.95**	**98.49**	**1972**

Experiments - keypoint repeatability

- Oxford RobotCar



- ETH (1024), StereoDSO (512)



Experiments - point cloud registration

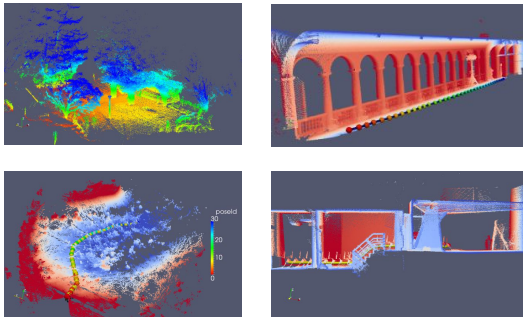
- Oxford RobotCar dataset

	RTE (m) / RRE(°) / Succ. (%) / Iter.				
	FPFH	3DSmoothNet	3DFeatNet	FCGF	DH3D
Random	0.44/1.84/89.8/7135	0.34/1.39/96.2/7274	0.43/1.62/90.5/9898	0.61/ 2.01/39.87/7737	0.33/1.31/92.1/6873
ISS	0.39/1.60/92.3/7171	0.32/1.21/96.8/6301	0.31/1.08/97.7/7127	0.56/1.89/43.99/7799	**0.30**/1.04/97.9/4986
Harris-3D	0.54/2.31/47.5/9997	0.31/1.19/97.4/5236	0.35/1.33/95.0/9214	0.57/1.99/46.82/7636	0.34/1.20/96.4/5985
3DFeatNet	0.43/2.01/73.7/9603	0.34/1.34/95.1/7280	**0.30**/1.07/98.1/2940	0.55/1.89/43.35/5958	0.32/1.24/95.4/2489
USIP	0.36/1.55/84.3/5663	**0.28**/**0.93**/98.0/**584**	**0.28**/**0.81**/**99.1**/**523**	0.41/1.73/53.42/3678	**0.30**/1.21/96.5/**1537**
DH3D	0.75/1.85/55.6/8697	0.32/1.22/96.0/3904	**0.28**/1.04/**98.2**/2908	0.38/1.48/49.47/4069	**0.23**/**0.95**/**98.5**/1972

[3]

- ETH dataset

	RTE (m) / RRE(°) / Succ. (%) / Iter.				
	SI	3DSmoothNet	3DFeatNet	FCGF	DH3D
Random	0.36/4.36/95.2/7535	**0.18**/2.73/**100**/**986**	0.30/4.06/95.2/6898	0.69/52.87/17.46/10000	0.25/3.47/**100**/5685
ISS	0.37/5.07/93.7/7706	**0.15**/**2.40**/**100**/**986**	0.31/3.86/90.5/6518	0.65/24.78/6.35/10000	0.19/2.80/93.8/3635
Harris-3D	0.35/4.83/90.5/8122	**0.15**/2.41/**100**/**788**	0.27/3.96/88.9/6472	0.43/55.70/6.35/10000	0.22/3.47/93.4/4524
3DFeatNet	0.35/5.77/87.3/7424	**0.17**/2.73/**100**/1795	0.33/4.50/95.2/6058	0.52/47.02/3.17/10000	0.27/3.58/93.7/6462
USIP	0.32/4.06/92.1/6900/	**0.18**/2.61/**100**/**1604**	0.31/3.49/82.5/7060	0.54/27.62/15.87/10000	0.29/3.29/95.2/4312
DH3D	0.42/4.65/81.3/7922	0.38/3.49/**100**/5108	0.36/**2.38**/**95.5**/3421	0.56/48.01/15.87/10000	0.3/**2.02**/**95.7**/3107

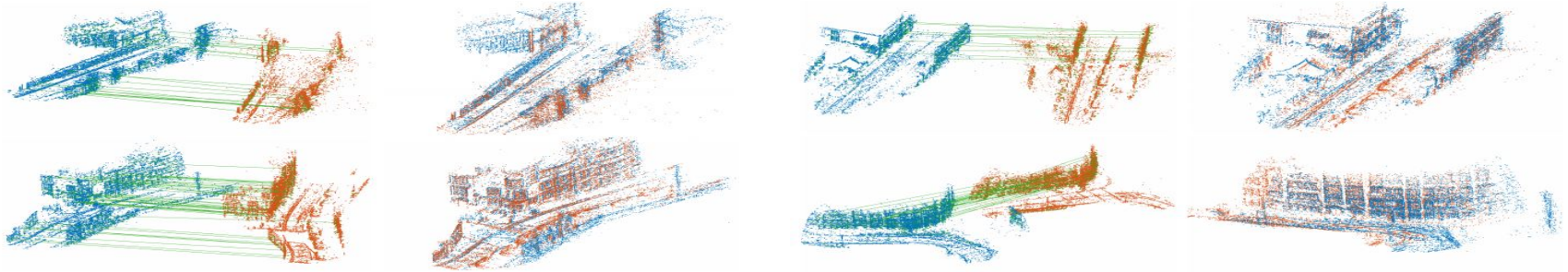


[4]

Experiments - visual SLAM - registration

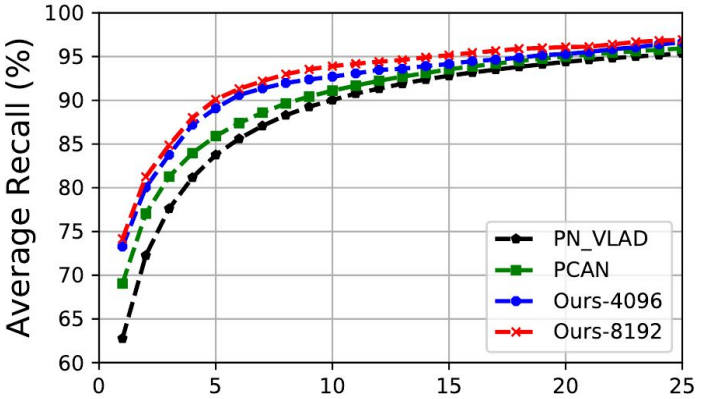
- Oxford RobotCar dataset

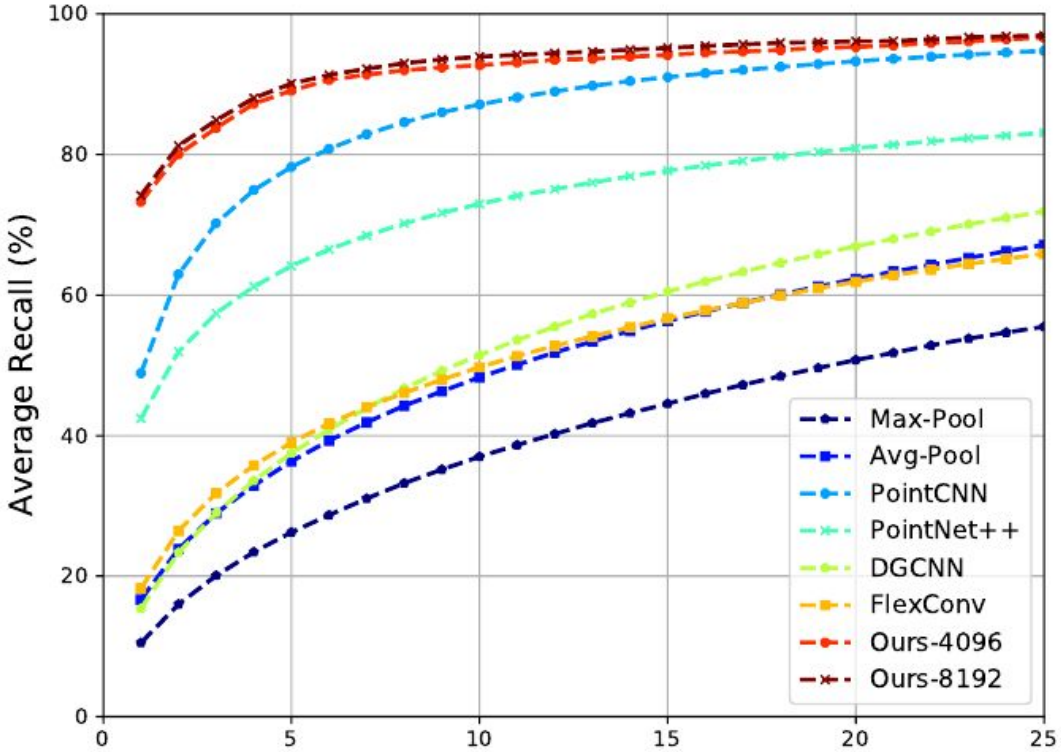
	RTE (m) / RRE(°) / Succ. (%) / Iter.				
	FPFH	3DSmoothNet	3DFeatNet	FCGF	DH3D
Random	0.56/2.82/53.13/9030	0.70/2.19/73.1/6109	0.72/2.37/69.0/9661	0.51/2.65/74.93/5613	0.70/2.23/71.9/7565
ISS	0.56/3.03/43.58/9210	0.67/2.15/79.1/6446	0.58/2.41/71.9/9776	0.51/2.57/71.94/6015	0.48/**1.72**/**90.2**/6312
Harris-3D	0.49/2.67/45.67/9130	0.48/2.07/74.9/6251	0.66/2.26/64.5/9528	0.48/2.63/74.03/5482	0.39/2.27/68.1/7860
3DFeatNet	0.62/3.05/35.52/7704	**0.38**/2.22/66.6/5235	0.92/1.97/84.1/8071	0.54/2.64/60.90/**4409**	0.74/2.38/80.9/7124
USIP	0.54/2.98/48.96/7248	0.39/2.27/77.3/5593	0.85/2.24/69.9/8389	0.51/2.65/67.46/**3846**	0.65/2.45/68.1/6824
DH3D	0.60/2.92/48.96/8914	**0.35**/2.01/77.9/5764	0.41/**1.84**/**89.3**/7818	0.48/2.43/69.55/**5002**	**0.36**/**1.58**/**90.6**/7071



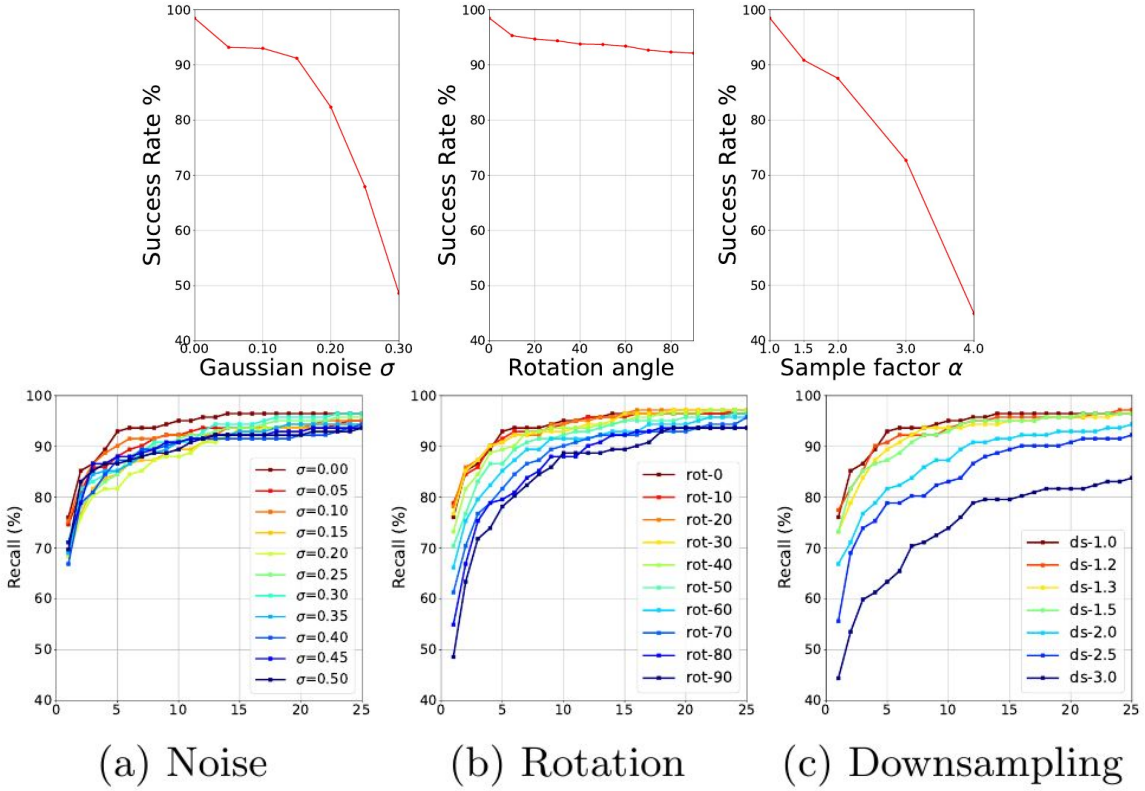
Experiments - point cloud retrieval

Method	@1%	@1
PN_MAX	73.44	58.46
PN_VLAD	81.01	62.18
PCAN	83.81	69.76
Ours-4096	**84.26**	**73.28**
Ours-8192	**85.30**	**74.16**





Experiments - robustness



Thank you!

References

All non-referenced image materials are from [1]

[1] Du, Juan, Rui Wang, and Daniel Cremers. "DH3D: Deep Hierarchical 3D Descriptors for Robust Large-Scale 6DoF Relocalization." ECCV 2020. + supplementary material, conference slides and videos
[2] Angelina Uy, Mikaela, and Gim Hee Lee. "Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition." CVPR 2018.
[3] Oxford RobotCar: https://robotcar-dataset.robots.ox.ac.uk/documentation/
[4] ETH ASL dataset: https://projects.asl.ethz.ch/datasets/doku.php?id=home
[5] Tombari, Federico, Samuele Salti, and Luigi Di Stefano. "Unique shape context for 3D data description." ACM workshop on 3D object retrieval, 2010.
[6] Rusu, Radu Bogdan, Zoltan Csaba Marton, Nico Blodow, and Michael Beetz. "Persistent point feature histograms for 3D point clouds." IAS-10, 2008.
[7] Choy, Christopher, Jaesik Park, and Vladlen Koltun. "Fully convolutional geometric features." ICCV 2019.
[8] Yew, Zi Jian, and Gim Hee Lee. "3DFeat-Net: Weakly supervised local 3D features for point cloud registration." ECCV 2018.
[9] Zhong, Yu. "Intrinsic shape signatures: A shape descriptor for 3D object recognition." ICCV 2009.
[10] Rister, Blaine, Mark A. Horowitz, and Daniel L. Rubin. "Volumetric image registration from invariant keypoints." IEEE Transactions on Image Processing 26, no. 10 (2017): 4900-4910.
[11] Sipiran, Ivan, and Benjamin Bustos. "Harris 3D: a robust extension of the Harris operator for interest point detection on 3D meshes." The Visual Computer 27, no. 11 (2011): 963.
[12] Li, Jiaxin, and Gim Hee Lee. "USIP: Unsupervised stable interest point detection from 3D point clouds." ICCV 2019.

References

[13] Röhling, Timo, Jennifer Mack, and Dirk Schulz. "A fast histogram-based similarity measure for detecting loop closures in 3-d lidar data." IROS 2015.
[14] Cop, Konrad P., Paulo VK Borges, and Renaud Dubé. "Delight: An efficient descriptor for global localisation using lidar intensities." ICRA 2018.
[15] Yin, Huan, Li Tang, Xiaqing Ding, Yue Wang, and Rong Xiong. "LocNet: Global localization in 3D point clouds for mobile vehicles." IV 2018.
[16] Dusmanu, Mihai, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. "D2-net: A trainable cnn for joint description and detection of local features." CVPR 2019.
[17] Groh, Fabian, Patrick Wieschollek, and Hendrik PA Lensch. "Flex-convolution." ACCV 2018.
[18] Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." CVPR 2018.
[19] Deng, Haowen, Tolga Birdal, and Slobodan Ilic. "PPFNet: Global context aware local features for robust 3D point matching." CVPR 2018.
[20] Csurka, Gabriella, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. "Visual categorization with bags of keypoints." ECCV, 2004.
[21] Sivic, Josef, Bryan C. Russell, Alexei A. Efros, Andrew Zisserman, and William T. Freeman. "Discovering objects and their location in images." ICCV 2005.
[22] Jégou, Hervé, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. "Aggregating local descriptors into a compact image representation." CVPR 2010.
[23] Arandjelovic, Relja, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. "NetVLAD: CNN architecture for weakly supervised place recognition." CVPR 2016.
[24] Zhang, Wenxiao, and Chunxia Xiao. "PCAN: 3D attention map learning using contextual information for point cloud based retrieval." CVPR 2019.