HRegNet: A Hierarchical Network for Large-scale Outdoor LiDAR Point Cloud Registration – Supplementary material

Fan Lu[1], Guang Chen[1], Yinlong Liu[2], Lijun Zhang[1], Sanqing Qu[1], Shu Liu[3], Rongqi Gu[4]
[1]Tongji University, [2]Technische Universität München, [3]ETH Zurich, [4]Westwell lab
{lufan,guangchen,tjedu_zhanglijun,2011444}@tongji.edu.cn
Yinlong.Liu@tum.de, liush@ethz.ch, rongqi.gu@westwell-lab.com

1. Feature extraction

1.1. Network architecture

The inputs of the feature extraction module in layer $l+1$ are the keypoints $X_l \in \mathbb{R}^{M_l \times 3}$, descriptors $D_l \in \mathbb{R}^{M_l \times C_l^D}$, saliency uncertainties $\Sigma_l \in \mathbb{R}^{M_l}$ and features $F_l \in \mathbb{R}^{M_l \times C_l^F}$ in last layer l. For the first layer, the input keypoints are the original point cloud, the saliency uncertainties are initialized to 1 and the features and descriptors are initialized to None. We start by adopting a Weighted Farthest Point Sampling (WFPS) [4] to sample a set of candidate keypoints $\hat{X}_{l+1} = \{\hat{x}_1, \cdots, \hat{x}_{M_{l+1}}\} \in \mathbb{R}^{M_{l+1} \times 3}$. Given the saliency uncertainties $\Sigma_l = \{\sigma_1, \cdots, \sigma_{M_l}\}$, the weight of each point can be calculated as

$$w_i = M_l \frac{1/\sigma_i}{\sum_{j=1}^{M_l} 1/\sigma_j} \tag{1}$$

Compared to standard Farthest Point Sampling (FPS) [3], WFPS incorporates the weights of points into the sampling process, thus the algorithm can concentrate more on the points with larger weights and reject unreliable points. The detailed description of WFPS can be seen in [4].

Detector network: After performing WFPS, we obtain a set of candidate keypoints. We use kNN search to generate clusters centered on the candidate keypoints and each cluster consists of K neighboring points. The features of a cluster consist of the relative distances and coordinates of the neighboring points to the center point and also the features of neighboring points. The cluster features are inputted into a 3-layer Shared-MLP to generate a feature map $\tilde{F} \in \mathbb{R}^{M_{l+1} \times K \times C_{l+1}^F}$. After that, a max-pool layer and a Softmax function are followed to predict attentive weights for the neighboring points. The generated keypoints X_{l+1} can be represented as the weighted sum of neighboring points. Besides, we also produce an attentive feature map $\hat{F}$ by weighting the feature map $\tilde{F}$ using the predicted attentive weights. The features $F_{l+1} \in \mathbb{R}^{M_{l+1} \times C_{l+1}^F}$ of the generated keypoints can be represented as the summation of the attentive feature map of neighboring points in clusters. We further input F_{l+1} into a 3-layer MLP with a Softplus function to predict the saliency uncertainties $\Sigma_{l+1} \in \mathbb{R}^{M_{l+1}}$.

Descriptor network: The features of clusters are inputted into another 3-layer Shared-MLP with a max-pool layer to generate global features, which are duplicated and concatenated with the attentive feature map $\hat{F}$ from the detector network and also individual features of neighboring points. The concatenated feature map is further passed into a 2-layer Shared-MLP with a max-pool layer to generate the final descriptors $D_{l+1} \in \mathbb{R}^{M_{l+1} \times C_{l+1}^D}$ of the keypoints.

1.2. Loss functions and training details

The training of the proposed method can be divided into 3 stages. We first train the detector network of the feature extraction module using the probabilistic chamfer loss in USIP [1], which aims to minimize the distances between the keypoints in source and target point clouds and meanwhile optimize saliency uncertainties. Then, We utilize the matching loss in RSKDD-Net [2] to train the descriptor network based on the pre-trained detector network, which requires only the relative transformation between two point clouds. Finally, the HRegNet can be trained based on the pre-trained feature extraction module and the weights of the detector network is fixed during this training for stability.

1.3. Details

The detailed network structure of the feature extraction module is shown in Table 1. M denotes the number of keypoints and K is the number of searched neighboring points in a cluster in the given layer. The first row in Detector Convs represents the channel numbers of the first Shared-MLP in the detector network and the second row is the channel numbers of the second MLP. Similarly, Descriptor Convs also denotes the channel numbers of two Shared-MLPs in the descriptor network.

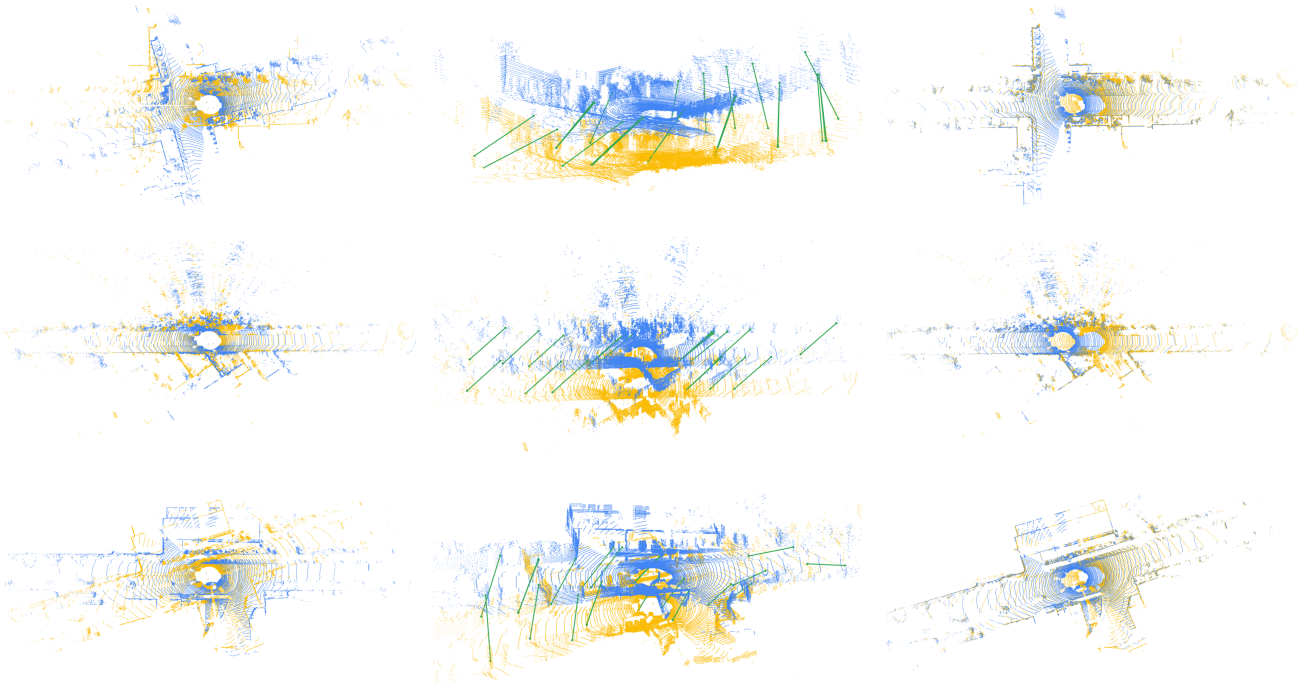

Figure 1. Qualitative results on KITTI dataset. Left: Two point clouds to be aligned; Middle: The green lines represent the correspondences of keypoints; Right: Two aligned point clouds.

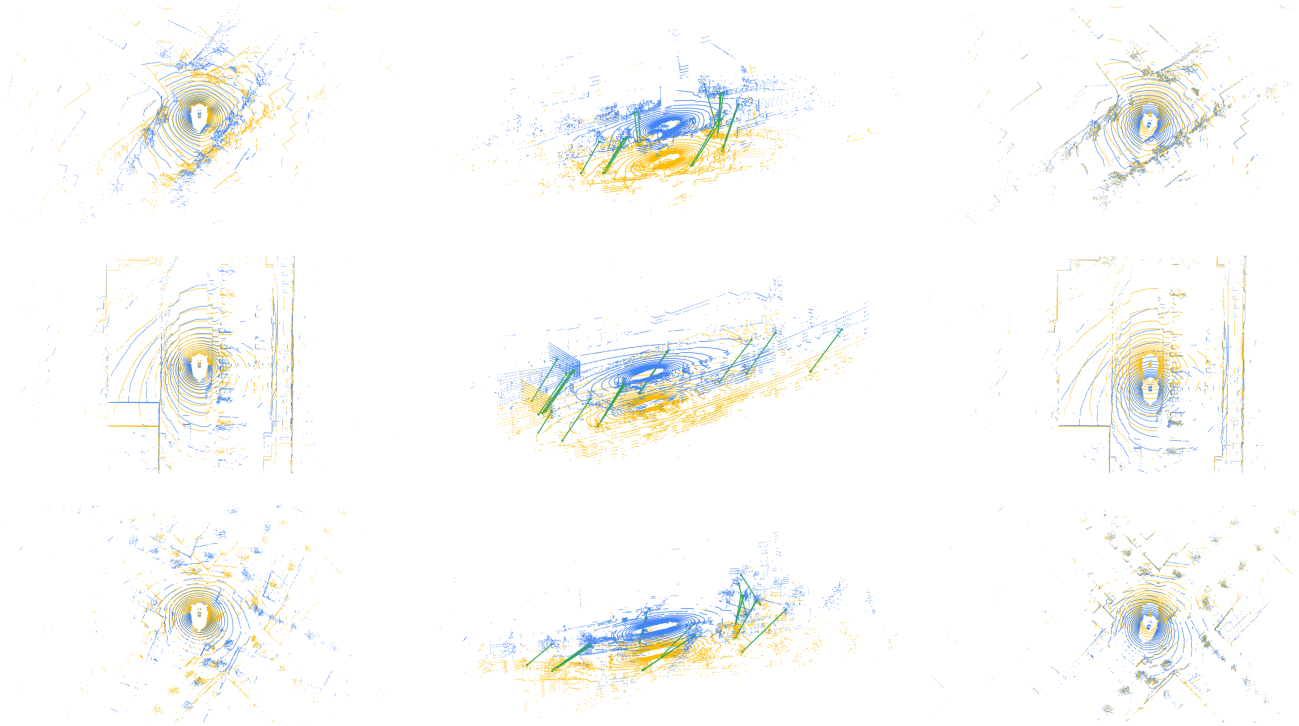

Figure 2. Qualitative results on NuScenes dataset. Left: Two point clouds to be aligned; Middle: The green lines represent the correspondences of keypoints; Right: Two aligned point clouds.

Table 1. Detailed structure of the feature extraction module.

Layer	M	K	Detector Convs	Descriptor Convs
1	1024	64	[32,32,64] [64,64,1]	[32,32,64] [32,64]
2	512	32	[64,64,128] [128,128,1]	[64,64,128] [64,128]
3	256	16	[128,128,256] [256,256,1]	[128,128,256] [128,256]

Table 2. Detailed structure of the proposed HRegNet.

Layer	K	Convs1	Convs2	Neighbor Convs
1	8	[128,128,128]	[128,128,1]	/
2	8	[256,256,256]	[256,256,1]	/
3	8	[512,512,512]	[512,512,1]	[256,256,256]

2. Details of HRegNet

The detailed network architecture of the proposed HRegNet is shown in Table 2. K is the number of candidate keypoints. Convs1 and Convs2 represent the channel numbers of two Shared-MLPs in the correspondence network, respectively. Neighbor Convs denotes the channel numbers of Shared-MLP in the neighbor encoding module.

Qualitative results: We provide several qualitative results on KITTI dataset and NuScenes dataset in Fig. 1 and Fig. 2. The left column displays two point clouds to be aligned, the middle column displays the predicted corresponding keypoints in coarse registration and the right column shows the aligned two point clouds based on the estimated transformation. According to the results, the proposed HRegNet can precisely predict relative transformation between two point clouds.

References

[1] Jiaxin Li and Gim Hee Lee. Usip: Unsupervised stable interest point detection from 3d point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 361–370, 2019. 1

[2] Fan Lu, Guang Chen, Yinlong Liu, Zhongnan Qu, and Alois Knoll. Rskdd-net: Random sample-based keypoint detector and descriptor. *Advances in Neural Information Processing Systems*, 33, 2020. 1

[3] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, volume 30, 2017. 1

[4] Yao Zhou, Guowei Wan, Shenhua Hou, Li Yu, Gang Wang, Xiaofei Rui, and Shiyu Song. Da4ad: End-to-end deep attention-based visual localization for autonomous driving. In *European Conference on Computer Vision*, pages 271–289. Springer, 2020. 1