LiDAR-Based Global Localization Using Histogram of Orientations of Principal Normals

Lun Luo, Si-Yuan Cao, Zehua Sheng, and Hui-Liang Shen

***Abstract*—Global localization on LiDAR point cloud maps is a challenging task because of the sparse nature of point clouds and the large size difference between LiDAR scans and the maps. In this paper, we solve the LiDAR-based global localization problem based upon the plane-motion assumption. We first project the clouds into Bird's-eye View (BV) images and transform the problem into a BV image matching problem. We then introduce a novel local descriptor, *i.e.*, Histogram of Orientations of Principal Normals (HOPN), to perform matching. The HOPN descriptor encodes the point normals of the clouds, and is more effective in matching BV images than the common image descriptors. In addition, we present the consensus set maximization algorithm to robustly estimate a rigid pose from the HOPN matches in the case of the low inlier ratio. The experimental results on three large-scale datasets show that our method achieves state-of-the-art global localization performance when using either single LiDAR scans or local maps.**

***Index Terms*—Global localization, pose estimation, LiDAR, point cloud, descriptor, image matching, optimization, intelligent vehicles.**

I. Introduction

GLOBAL localization is one of the key components for intelligent vehicles to achieve full autonomy. In practice, it can initialize the vehicle pose tracking algorithms without any prior information of the pose, and can also re-localize the vehicle when the pose tracking fails due to wrong data association [1]. Although the global navigation satellite system (GNSS) can provide high-precision positioning results in open areas, it is not always reliable in urban environments since the satellite signals can easily be affected by buildings [2] or bad weather conditions. To cope with the GNSS-denied environment, many localization methods using onboard sensor data such as images [3], [4] and the scans of light detection and ranging (LiDAR) sensors [5]–[7] have been proposed in the literature. Among these methods, LiDAR-based ones are shown to be more accurate than the image-based counterparts due to the availability of precise depth information [5]. In addition, they are more robust because of the invariance to illumination and view changes [6].

Manuscript received 16 January 2022; revised 13 March 2022 and 3 April 2022; accepted 14 April 2022. Date of publication 21 April 2022; date of current version 24 October 2022. *(Corresponding author: Hui-Liang Shen.)*

The authors are with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: 11831030@zju.edu.cn; karlcao@hotmail.com; shengzehua@zju.edu.cn; shenhl@zju.edu.cn).

Data is available on-line at https://github.com/zjuluolun/HOPN and https://www.youtube.com/watch?v=xXm3txQalAo.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIV.2022.3169153.

Digital Object Identifier 10.1109/TIV.2022.3169153

LiDAR-based global localization can be cast as a problem of estimating global poses of the local LiDAR scans with respect to a pre-built map. Since no prior information of the pose is available, this problem cannot be solved by Simultaneously Localization and Mapping (SLAM) algorithms [8]. One straightforward solution is to detect keypoints and assign each keypoint a local descriptor [9], [10]. Global localization can then be achieved through descriptor matching. However, the local point cloud descriptors [9], [10] may not provide sufficient discrimination in large-scale outdoor environments since they are sensitive to point density and noise. Some methods [11], [12] extend the visual place recognition methods to the LiDAR. They build a database storing the global descriptors of keyframes and perform localization through descriptor retrieval. These keyframe-based methods work well when the keyframes of the map are provided, but cannot be employed to match LiDAR scans with the raw point cloud map. The segment-based methods [13], [14] suppose that the segments belonging to partial or full objects can incorporate the strengths of the local and global descriptors while reducing their individual drawbacks. To this end, they extract segments from the point clouds and match them for localization. Similarly, there are also some methods [15], [16] that perform localization using high-level object features on road scenes such as curbs. Since the segments and objects cannot be perfectly extracted from sparse LiDAR scans, these methods need to travel a long distance to build a compact point cloud before a successful localization.

In road scenes, the vehicle usually moves on the ground plane and thus its pose is restricted to the 2D space. Based on this observation, some methods [12], [17], [18] project point clouds into 2D images before extracting descriptors and achieve competitive performance compared to the methods utilizing raw point clouds [5], [7], [19]. However, these methods have two inherent limitations. First, due to the sparse nature of the point clouds, the projected images are usually textureless and suffer from intensity distortion. Thus, the descriptors extracted from the images are not robust to noise and view changes. Second, the pixel intensities of the image formed by point densities [12] or heights [17], [18] lose much structure information of the point clouds. As a result, the descriptors are less distinctive in the localization problem.

In this work, we project the point clouds into bird's-eye view (BV) images [12] by discretizing the ground plane into regular grids. Based on the plane-motion assumption, the transform of the BV image pair is rigid and is similar to the 2D transform of the corresponding LiDAR scan pair. This allows us to achieve

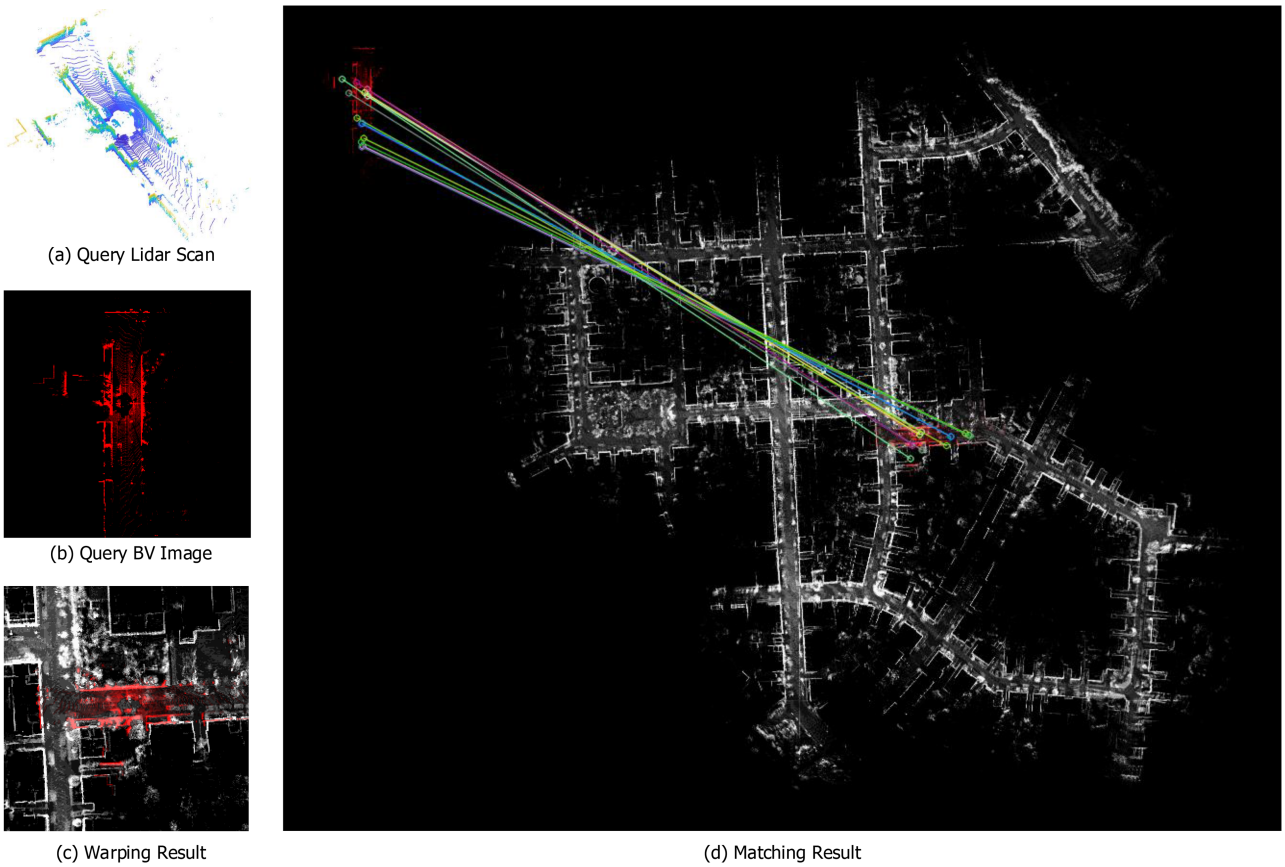

Fig. 1. Illustration of a LiDAR scan matching with the global map on the sequence "*00*" of the KITTI dataset. (a) The query LiDAR scan. (b) The query BV image colored red for visualization. (c) The overlapping of the warped query BV image (red) and the global BV image (grayscale). (d) The keypoint correspondences of the local query BV image (top left corner) and the global BV image. Our method can produce excellent matching despite of the large difference between the LiDAR scan and global map. It is validated that the average success rate of our method is 88.7% on the KITTI dataset.

localization by matching BV images. However, the common image-based descriptors [12], [17], [18], [20] cannot handle the matching problem well since BV images suffer from intensity distortion and lose much structure information of the point clouds. To solve the problem, we introduce the Histogram of Orientations of Principal Normals (HOPN) computed from cloud normals. Compared to image-based descriptors, HOPN is more distinctive in the localization problem by using 3D information. To estimate global poses from the HOPN matches robustly, we use the consensus set maximization [21] for outlier rejection and solve the problem using a fast voting-based algorithm. With the distinctive HOPN descriptor and the efficient pose estimation technique, our method shows the powerful capability of matching local LiDAR scans with large global maps as shown in Fig. 1. Furthermore, our method does not rely on the special map representations such as keyframe databases [11], [12] or segment maps [13], [14] that need extra treatments to the LiDAR scans when constructing the map. Instead, it performs global localization on the raw point cloud map. To summarize, the contributions of this work are as follows:

- We propose a novel local descriptor called HOPN for the LiDAR-based global localization problem. The descriptor encodes the principal normals and can match LiDAR scans with the large global map effectively.
- We introduce a voting-based algorithm for the consensus set maximization problem constrained by a 2D rigid transform. With the objective of maximizing the number of inliers, the algorithm improves the robustness of global pose estimation.
- We compare our method with the state-of-the-arts and validate the superiority of our method in terms of localization capability and long-term robustness on three large-scale datasets.

The remainder of this paper is organized as follows. Section II overviews the related work in the fields of LiDAR-based global localization. Section III introduces the framework of our global localization method. Section IV elaborates the design of the HOPN descriptor, and Section V presents the solution to the consensus set maximization problem. Section VI shows the experimental results and Section VII finally concludes the paper.

II. Related Work

The global localization methods can be approximately divided into three categories according to the features they employed, *i.e.*, local features, global features, and objects.

The local feature-based methods design local descriptors for point clouds and match them for localization. Some methods

exploit the local characteristics of point cloud with geometric measures. For example, the fast point feature histograms (FPFH) method [9] uses the relationship between the neighbors of interest points and encodes them into a histogram. The Signature of Histograms of OrienTations (SHOT) [10] builds a local reference frame and leverages the orientation distribution of the normals in local regions. Although these methods can align local LiDAR scans efficiently, they usually lack descriptive capability in outdoor environments due to their sensitivity to point cloud density and noise. To leverage the local descriptors more efficiently, the keypoints voting method [22] performs localization using the 3D Gestalt descriptors. The method depends on the keypoints extraction procedure, but actually, the repeatable 3D keypoint detection is still an open problem in the literature. Instead of directly processing 3D point clouds, some methods [23], [24] adopt 2D occupancy grid maps for LiDAR global localization and build local descriptors based on the statistics of the occupancy probability. Some studies [25], [26] directly extract conventional image descriptors such as ORB [27] and SIFT [20] from LiDAR images for place recognition. However, the 2D descriptors in these methods are less distinctive since they do not consider the 3D structure information of point clouds. In comparison, our method employs the keypoints from 2D BV images and builds the HOPN descriptors using the normals of 3D points.

The global feature-based methods generate global descriptors for LiDAR scans and perform place recognition through descriptor retrieval. A main trend of these methods is to extract local features [11], [12] from a set of training point clouds and train a bag-of-words [3] model. Then the global descriptors are generated using the model and global localization is achieved through descriptor retrieval and local registration. Theoretically, any descriptors including our HOPN descriptor can be employed in such framework. The drawback is that it cannot be employed to match LiDAR scans with point cloud maps due to the place recognition mechanism. What's more, the bag-of-words model usually has poor generalization performance in unseen environments. There are some works [5], [7], [19], [28]–[30] that leverage neural network to generate global descriptors. Similar to the bag-of-words approach, these learning based methods may show poor performance across scenes. Instead of training models, some methods [6], [17], [31] exploit the projection statistics of the raw point cloud for global descriptor generation. However, these methods cannot perform metric localization since they do not extract local features.

The object-based methods aim to represent the environment using objects rather than using local or global descriptors. SegMatch [13] performs segmentation on LiDAR scans and builds a segment map. On the online global localization, it extracts segments from the query LiDAR scan and matches them with the map. SegMap [14] further assigns the segments descriptors learned by neural network to improve the matching performance. The main drawback of these methods is that they need to travel a long distance to extract distinctive segments. What's more, they cannot be employed on the point cloud map directly as segmentation to LiDAR scans is required during map constructions. Following SegMap, OneShot [32] extracts distinct segments from single LiDAR scans. However, it relies on the fusion of the image information and LiDAR point cloud. Some methods [15], [16] extract salient curbs on roads to perform localization. However, the curbs are simple geometry features and lack descriptive power in large-scale environment.

III. Global Localization Framework

In this section, we first introduce our global localization framework, and then present the details of BV image formation and global 3D pose estimation.

A. Framework

Under the plane-motion assumption, our method achieves global localization by matching 2D BV images as illustrated in Fig. 2. For each query, we first align the consecutive LiDAR scans using LOAM [8] to form a local map since a single scan may fail to capture enough scene structures. Note that the local map is cropped using a cubic window with a side length of $2C$ meters, with $C = 50$, to reduce computation. We then uniformly discretize the ground plane into grids. In each grid, the normalized point density forms the pixel value of the BV image. We detect keypoints on the image with the FAST [33] detector and construct the HOPN descriptors using the principal normals of point clouds (see Section IV). We perform feature matching with the global BV image via the nearest neighbor searching. From the resulting matches, we estimate the 2D pose of the local BV image by solving a consensus set maximization problem (see Section V). We recover the coarse 3D pose of the local map using a similar transform. If more accurate localization is desired, we can optionally use the Iterative Closest Point (ICP) algorithm [34] to refine the pose.

B. BV Image Formation

As illustrated in Fig. 3(a), we use the right-handed coordinate system, with the x-axis pointing forward, the y-axis pointing left, and the z-axis pointing upward. We regard the x-y plane as the motion plane. Given a point cloud $\mathcal{P}$, we distribute the points evenly using a voxel grid filter with leaf size of G meters. We discretize the x-y plane into grids with the resolution of G meters. For the cropped local map with the side length of $2C$ meters, the BV image I is a matrix of size $\lceil \frac{2C}{g} \rceil \times \lceil \frac{2C}{g} \rceil$. The intensity of I at position (u, v) is computed as [12]

$$I(u, v) = \frac{\min(N_g, N_m)}{N_m}, \tag{1}$$

where N_g denotes the number of points at (u, v). N_m denotes the normalization factor, and is set to the 99th percentile of the numbers of points in the grids. As illustrated in Fig. 1, the objects with vertical structures in the environment (*e.g.*, poles, facades, guideposts) form clear edges in BV images, which can facilitate the keypoint detection process.

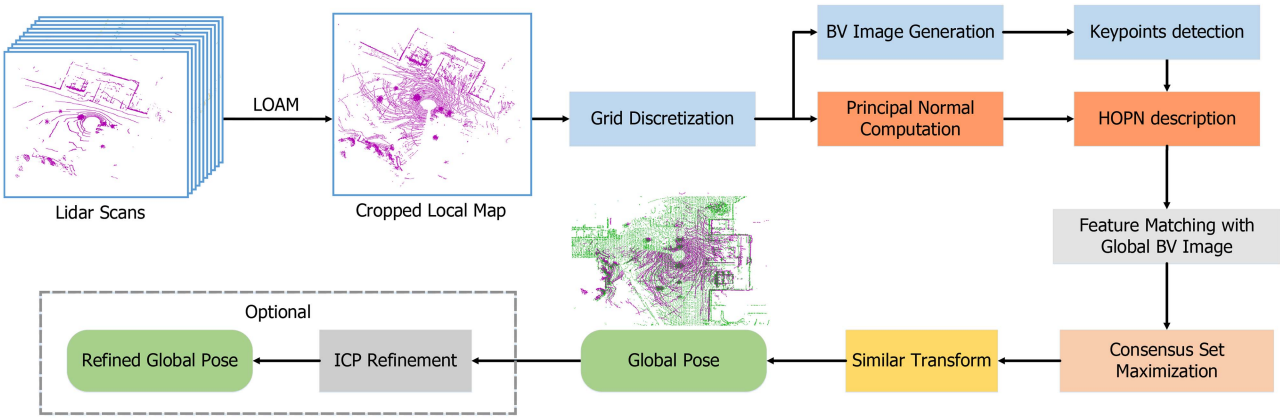

Fig. 2. Global localization framework. We first form the query local map by aligning the consecutive LiDAR scans, and then discretize the ground plane into regular grids. We detect the keypoints from the BV image and build the HOPN descriptor (see Section IV) from principal normals. We solve the feature matching problem using consensus set maximization (see Section V) and recover the global pose with the similar transform. If necessary, the pose can be further refined with ICP. The red and black boxes show the details of registration before and after the ICP refinement, respectively.

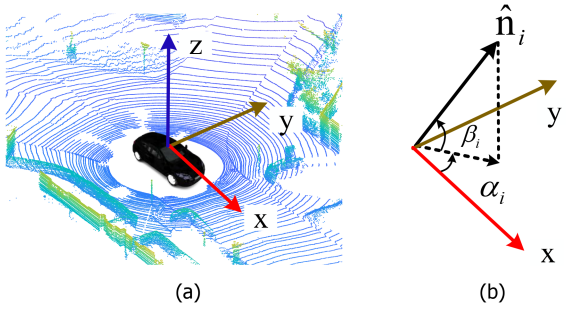

Fig. 3. Illustration of the coordinate system and principal normal. (a) The vehicle moves on the x-y plane. (b) The azimuthal angle α_i and elevation angle β_i associated with the principal norm $\hat{\mathbf{n}}_i$.

C. 3D Pose Estimation

In this work, we first estimate the 2D pose from BV images and then compute the 3D pose of point maps according to the principle of the similar transform.

Let I_l and I_g be the BV images of the local point map $\mathcal{M}_l$ and the global point map $\mathcal{M}_g$, respectively. The aim of pose estimation is to find a 2D transform with rotation θ and translation (t_u, t_v) so that the transformed local BV image $I_l(u, v)$ matches the global BV image $I_g(u,' v')$. With the 2D transform, the coordinates of the two BV images are related as

$$\begin{pmatrix} u' \\ v' \\ 1 \end{pmatrix} = \begin{pmatrix} \cos(\theta) & \sin(\theta) & t_u \\ -\sin(\theta) & \cos(\theta) & t_v \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} u \\ v \\ 1 \end{pmatrix}. \tag{2}$$

As the ground plane is discretized uniformly, the computed transform of BV images is similar to the 2D transform of the corresponding point maps in the x-y plane. The translations along the x- and y-axes are, respectively, $t_x = G \cdot t_u$ and $t_y = G \cdot t_v$. For the z-axis translation t_z, we crop the matched sub-map $\mathcal{M}'_g$ on $\mathcal{M}_g$ using a cubic window centering at (t_x, t_y) with a side length of $2C$ meters. We set t_z to be the difference about the z-axis of the centroids of $\mathcal{M}_l$ and $\mathcal{M}'_g$. Since the vehicle is assumed to move on the plane, we set the roll and pitch angles to zero. Consequently, the 3D transform between the point maps $\mathcal{M}_l$ and $\mathcal{M}_g$ becomes

$$\mathbf{T} = \begin{pmatrix} \cos(\theta) & \sin(\theta) & 0 & t_x \\ -\sin(\theta) & \cos(\theta) & 0 & t_y \\ 0 & 0 & 1 & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{3}$$

IV. Our Descriptor

The common image descriptors, such as ORB [27] and SIFT [20], are usually computed from image intensities or gradients. The intensity of the BV image, which represents the normalized point density of the point cloud, is not informative enough to represent local structures. Furthermore, the point density can be greatly affected by the distance between the vehicle and objects. For example, the point cloud is dense for nearby scenes while sparse for far scenes. Hence, the common intensity-based descriptors are not suitable for the BV image matching problem.

Compared to the point density, the point normal is more suitable for representing local structures since it indicates the orientation of the surface. However, current normal-based descriptors [9], [10] are sensitive to noise and view changes, because the normals are usually nonuniformly distributed along with the points in the cloud. We note that, in road scenes, the normals of local structures usually point to the same direction (see Fig. 4). From the bird's-eye view, we can find a principal normal that represents the overall orientation of the structure regardless of the point distribution along the z-axis. Based on this observation, we compute the principal normals in the uniform grids of the ground plane and construct the Histogram of Orientations of Principal Normals (HOPN) descriptor for BV image matching.

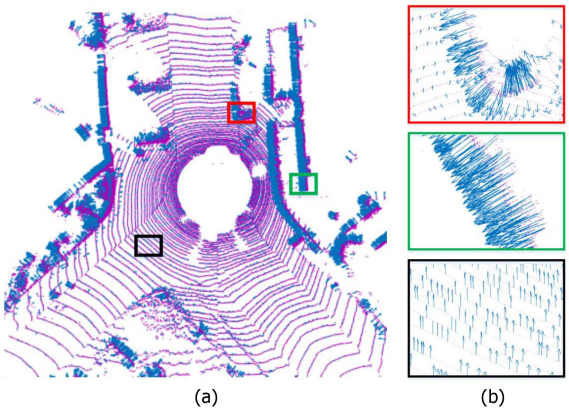

Fig. 4. Illustration of the distribution of the point normals in the road scene. (a) A LiDAR scan with normals from the bird's-eye view. (b) The normals of three typical structures on the road which are a car (red), the wall (green), and the ground (black). It can be seen that the normals in the local area usually point to the same direction.

A. Computing Principal Normals

For a point cloud $\mathcal{P}$ with N_p points, we first compute its normal cloud $\mathcal{N} = \{\mathbf{n}_i | i = 1, .., N_p\}$ where $\mathbf{n}_i = (n_{ix}, n_{iy}, n_{iz})^\mathsf{T}$. We then discretize the ground space into uniform grids of side size G meters, which is identical to the leaf size of voxel filter in the BV image formation. Let $\mathcal{N}_K = \{\mathbf{n}_{ik} | k = 1, .., K\}$ be the set of the normals in the i-th grid and its 8-neighbor grids. The principal normal $\hat{\mathbf{n}}_i$ is computed as the unit vector to which the mean projection length from the normals in set $\mathcal{N}_K$ is the largest, *i.e.*,

$$\begin{aligned} \hat{\mathbf{n}}_i = & \arg\max_{\mathbf{n}} \sum_k \frac{w_{ik}}{K} ||\mathbf{n}^\mathsf{T} \mathbf{n}_{ik}||_2^2 \\ & \text{s.t.} \quad \mathbf{n}^\mathsf{T}\mathbf{n} = 1, \end{aligned} \tag{4}$$

where w_{ik} is the weight. Since the far normals should make less contribution to the result, we set $w_{ik} = e^{-d_{ik}}$, where d_{ik} is the distance of $\mathbf{n}_{ik}$ to the center of the i-th grid. Problem (4) can be solved by constrained convex optimization. We write its Lagrange function as

$$\begin{aligned} f(\mathbf{n}) &= \sum_k \frac{w_{ik}}{K} ||\mathbf{n}^\mathsf{T} \mathbf{n}_{ik}||_2^2 - \lambda(\mathbf{n}^\mathsf{T}\mathbf{n} - 1) \\ &= \sum_k \frac{w_{ik}}{K} \mathbf{n}^\mathsf{T} \mathbf{n}_{ik} \mathbf{n}_{ik}^\mathsf{T} \mathbf{n} - \lambda(\mathbf{n}^\mathsf{T}\mathbf{n} - 1), \end{aligned} \tag{5}$$

where λ is the Lagrange multiplier. By computing its partial derivative with respect to $\mathbf{n}$ and set it to zero, we have

$$\sum_k \frac{w_{ik}}{K} \mathbf{n}_{ik} \mathbf{n}_{ik}^\mathsf{T} \mathbf{n} = \lambda \mathbf{n}. \tag{6}$$

This indicates that the optimal normal $\mathbf{n}$ is the eigenvector of the covariance matrix $\sum_k \frac{w_{ik}}{K} \mathbf{n}_{ik} \mathbf{n}_{ik}^\mathsf{T}$. Substituting (6) into (5) yields $f(\mathbf{n}) = \lambda$. Since the largest length is expected, $\hat{\mathbf{n}}_i$ is then the eigenvector corresponding to the maximal eigenvalue. Fig. 5 illustrates the distributions of the point normals and the principal normals of a local surface.

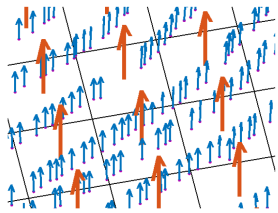

Fig. 5. Illustration of the point normals (blue arrows) and the principal normals (orange arrows) of a local surface. The point normal is computed for each point (purple points) of the point cloud. The principal normal is computed for each regular grid (indicated by black lines) of the motion plane using the point normals.

Since we compute the principal normal in the same grids of the BV image, each pixel in the image has an associated principal normal. However, it is not convenient to leverage $\hat{\mathbf{n}}_i$ since $\hat{\mathbf{n}}_i$ is a unit vector and its elements are not independent. Note that $\hat{\mathbf{n}}_i$ has two degrees of freedom and it can be uniquely represented by its azimuthal angle α_i and elevation angle β_i, as shown in Fig. 3. Thus, we instead use these two angles in HOPN descriptor construction. We compute α_i and β_i as

$$\begin{aligned} \alpha_i &= \tan^{-1}\left(\frac{\hat{n}_{iy}}{\hat{n}_{ix}}\right), \\ \beta_i &= \tan^{-1}\left(\frac{|\hat{n}_{iz}|}{\sqrt{\hat{n}_{ix}^2 + \hat{n}_{iy}^2}}\right). \end{aligned} \tag{7}$$

In road scenes, α_i indicates the orientation of the local structure on the x-y plane and β_i is the vertical degree. Fig. 6 illustrates the principal normals, as well as the associated azimuthal angles and elevation angles of a local map in the NCLT dataset [35] from the bird's-eye view.

B. Constructing HOPN Descriptor

We use a histogram distribution technique similar to SIFT [20] to construct the HOPN descriptor. While SIFT is based on pixel gradients, our HOPN descriptor is built using principal normals.

For each keypoint, we use a local patch of $J \times J$ pixels centering at the keypoint for feature description. To obtain a rotation-invariant descriptor, we need to align the patch with respect to its dominant orientation, which is determined from the structures of the scene. To this end, we construct a histogram $\text{hist}(b)$ with N_b bins for the azimuthal angles of the principal normals. Here we set $N_b = 12$ and $b \in \{0, 1, \ldots, N_b - 1\}$. As vertical structures are more informative in determining the dominant direction when compared with the horizontal ones, we construct the histogram in a weighted manner,

$$\text{hist}(b) = \sum_i \cos(\beta_i) \mathbb{I}_b(\alpha_i), \quad b = 0, 1, \ldots, N_b - 1, \tag{8}$$

so that the histogram can pay more attention to vertical structures. The indicator function $\mathbb{I}_b(\alpha_i)$ is defined as

$$\mathbb{I}_b(\alpha_i) = \begin{cases} 1, & \text{if } \alpha_i \in [\frac{b}{N_b}\pi - \frac{\pi}{2}, \frac{b+1}{N_b}\pi - \frac{\pi}{2}) \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

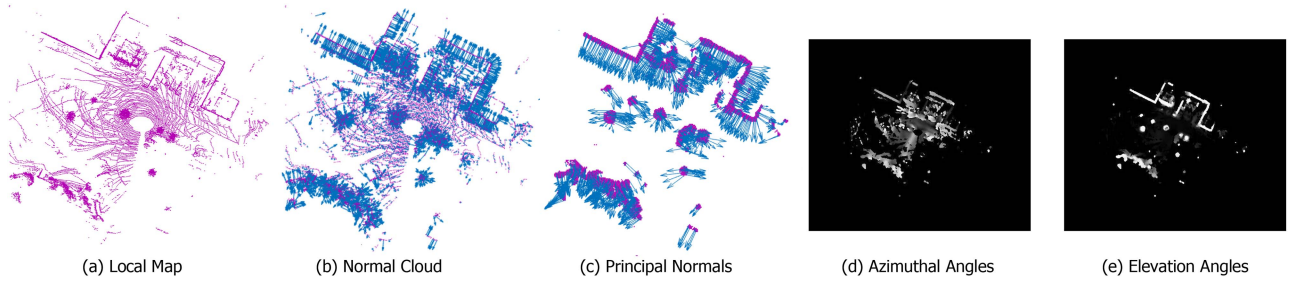

Fig. 6. The principal normals and its associated azimuthal and elevation angles of a normal cloud. (a) A local map from the NCLT dataset. (b) The normal cloud of the local map. (c) The principal normals of the local map. For concision, only the principal normals with their elevation angle larger than $\frac{\pi}{6}$ are displayed. (d) The azimuthal angles of the principal normals. (e) The elevation angles of the principal normals.

We detect the peak position b_0 of the histogram and rotate the local patch with respect to the dominant orientation angle $\theta_0 = \frac{b_0}{N_B}\pi - \frac{\pi}{2}$. Accordingly, the principal normal in the patch rotates as

$$\hat{\mathbf{n}}_i' = \begin{pmatrix} \cos(\theta_0) & \sin(\theta_0) & 0 \\ -\sin(\theta_0) & \cos(\theta_0) & 0 \\ 0 & 0 & 1 \end{pmatrix} \hat{\mathbf{n}}_i. \tag{10}$$

Thus, the elevation angle of $\hat{\mathbf{n}}_i'$ keeps unchanged while the azimuthal angle becomes

$$\alpha_i' = \tan^{-1}\left(\frac{-\sin(\theta_0)\hat{n}_{ix} + \cos(\theta_0)\hat{n}_{iy}}{\cos(\theta_0)\hat{n}_{ix} + \sin(\theta_0)\hat{n}_{iy}}\right). \tag{11}$$

To construct the HOPN descriptor, we divide the rotated patch into 6×6 sub-grids. For each sub-grid, we build a local distribution histogram of the azimuthal angles with B bins. Similar to the dominant orientation computation, we assign a weight to the magnitude of each sample pixel added to the histogram using the cosine value of the elevation angle, considering that vertical structures are more informative. Additionally, we weight the magnitude of the sample pixels in the patch using a Gaussian function with standard deviation $\frac{J}{2}$ to alleviate abrupt changes in the feature description when the keypoint position shifts. Finally, the HOPN descriptor is built by cascading all the local histograms of the sub-grids.

It is worth noting that the sign of the dominant orientation is ambiguous since we do not specify the direction of the principal normals. For a single LiDAR scan, we can eliminate this ambiguity by forcing the normals to point to the coordinate origin. For the 3D point cloud map, however, the directions of the normals cannot be determined in this way since a point may come from multiple LiDAR scans because of the mapping and filtering procedure. In our global localization framework, we solve this problem by assigning each keypoint in the local map with two descriptors where their dominant orientations have opposite directions. Note that this operation is not needed in the global map description. As a consequence, the number of the nearest neighbor searching doubles in descriptor matching.

V. Our Consensus Set Maximization Algorithm

Since we perform the HOPN descriptor matching between the local map and the global one, the inliers usually form a small part in the resulting match set. In this case, the conventional robust estimation method such as RANSAC [36] may fail to produce reliable solutions due to its random sampling mechanism. To estimate the optimal rigid transform having the maximal number of inliers, we formulate the pose estimation as the consensus set maximization problem [21].

The consensus set maximization problem can be solved using exhaustive searching algorithms such as branch-and-bounding [21], [37] and tree search [38]. However, these algorithms are computationally expensive and thus not suitable for the global localization task. In our scene, the transform constraint of the maximization problem is rigid. We note that, given a rotation angle, we can compute the optimal translation through fast parameter voting. Based on this consideration, we introduce a voting-based algorithm to solve the consensus set maximization problem efficiently and accurately.

A. Problem Formulation

We denote $\{(\mathbf{x}_i^l, \mathbf{x}_i^g) | i = 1, \ldots, N_F\}$ as the keypoint match set, where $\mathbf{x}_i^l \in \mathbb{R}^2$ and $\mathbf{x}_i^g \in \mathbb{R}^2$ are, respectively, the coordinates of the keypoint extracted from the local and global maps, and N_F is the number of keypoints. We regard the i-th match as an inlier with respect to the rigid transform $(\theta, \mathbf{t})$ when the residual $\|\mathbf{R}_\theta \mathbf{x}_i^l - \mathbf{x}_i^g + \mathbf{t}\|$ is less than a predefined threshold ϵ. We aim to maximize the number of inliers, mathematically

$$\begin{aligned} &\text{maximize}_{\theta, \mathbf{t}, \mathcal{I} \subseteq \Omega} \quad |\mathcal{I}| \\ &\text{s.t.} \quad \|\mathbf{R}_\theta \mathbf{x}_i^l - \mathbf{x}_i^g + \mathbf{t}\|_2 \leq \epsilon, \quad \forall i \in \mathcal{I}, \end{aligned} \tag{12}$$

where $\Omega = \{1, 2, \ldots, N_F\}$ denotes the index set, $\mathcal{I}$ the consensus set, and $|\mathcal{I}|$ the size of $\mathcal{I}$. $\mathbf{R}_\theta$ represents the rotation matrix of θ and $\mathbf{t}$ represents the 2D translation (t_u, t_v).

B. Solution

To solve the consensus set matrix problem (12) efficiently, our idea is to first discretize the rotation space $[0, 2\pi)$ with the resolution $\Delta\theta$ and then solve the sub-problems of (12) under all the discretized angles. As the sub-problems can be solved via translation voting, our solution can be much more efficient than the exhaustive searching algorithms [21], [37], [38].

We introduce our solution in the following. Given a discretized angle θ, the sub-problem of (12) under this angle is

$$\begin{aligned} &\text{maximize}_{\mathbf{t},\mathcal{I}\subseteq\Omega} \ \ |\mathcal{I}| \\ &\text{s.t.} \quad \|\mathbf{t}_i+\mathbf{t}\|_2 \le \epsilon, \quad \forall i \in \mathcal{I}, \end{aligned} \tag{13}$$

where $\mathbf{t}_i \triangleq \mathbf{R}_\theta \mathbf{x}_i^l - \mathbf{x}_i^g$. The constraint of problem (13) implies that if the i-th match is an inlier, $\mathbf{t}_i$ must be a point falling into the circle centering at $\mathbf{t}$ with the radius ϵ. Hence, the objective of problem (13) can be regarded as maximizing the number of points $\mathbf{t}_i$ that can be enclosed in the circle of radius ϵ.

We solve the sub-problem via translation voting. We discretize the 2D translation space into a grid of bins with the resolution Δt. For each point $\mathbf{t}_i$, we cast votes into its corresponding bin and the neighbor bins in a circle of radius $\frac{\epsilon}{\Delta t}$. The translation with the most votes is the optimal translation and the corresponding votes is the maximal consensus set size. By traversing all the quantized angles and solving the corresponding sub-problems, we obtain the optimal quantized transform and the consensus set $\hat{\mathcal{I}}$.

To reduce the quantization error introduced by transform discretization, we use the inliers to compute a refined transform $(\hat{\mathbf{R}}_\theta, \hat{\mathbf{t}})$ via singular value decompsion (SVD) [39]. Specifically, we compute the covariance matrix $\mathbf{C}$ of the inliers as

$$\mathbf{C} = \sum_{i\in\hat{\mathcal{I}}} \left(\mathbf{x}_i^l - \bar{\mathbf{x}}^l\right)\left(\mathbf{x}_i^g - \bar{\mathbf{x}}^g\right)^\mathsf{T}, \tag{14}$$

where

$$\bar{\mathbf{x}}^l = \frac{1}{|\hat{\mathcal{I}}|}\sum_{i\in\hat{\mathcal{I}}}\left(\mathbf{x}_i^l\right), \quad \bar{\mathbf{x}}^g = \frac{1}{|\hat{\mathcal{I}}|}\sum_{i\in\hat{\mathcal{I}}}(\mathbf{x}_i^g), \tag{15}$$

and obtain its decomposition

$$\mathbf{C} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}, \tag{16}$$

where $\mathbf{\Sigma}$ is a diagonal matrix of singular values, and $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices. Then we compute the rotation matrix $\hat{\mathbf{R}}_\theta$ as

$$\hat{\mathbf{R}}_\theta = \mathbf{U}\mathbf{V}^\mathsf{T}, \tag{17}$$

and compute the translation vector $\hat{\mathbf{t}}$ as

$$\hat{\mathbf{t}} = \bar{\mathbf{x}}^g - \hat{\mathbf{R}}_\theta \bar{\mathbf{x}}^l. \tag{18}$$

Our algorithm is globally optimal under the specified angle resolution $\Delta\theta$ and translation resolution Δt. Theoretically, we need to set the resolutions as small as possible to ensure accuracy. In practice, we found that our algorithm can achieve good performance with relatively large resolutions, and our experiments will show that our algorithm is more effective in the global localization problem than the state-of-the-art RANSAC-based methods.

VI. Experiments

In this section, we present the experimental results of our method. We compare the global localization performance with SegMatch [13], SegMap [14], SIFT [20], and BVFT [12]. For SegMatch and SegMap, we build segment maps and perform localization by segment matching. While SegMap is a deep learning method, we adopt the pre-trained model on the KITTI dataset for evaluation. For SIFT and BVFT, we extract descriptors from BV images and perform localization using our consensus set maximization algorithm. For our method, we do not use the ICP refinement when comparing with the competitors for the sake of fairness. To investigate how ICP influences the performance, we compare the localization accuracy before and after ICP refinement in Section VI-G. The source codes of SegMatch, SegMap, and BVFT are publicly available on the websites.[1] For SIFT, we use the implementation of OpenCV.[2]

A. Datasets

We conduct the experiments on three large-scale datasets: KITTI dataset [40], NCLT dataset [35] and Oxford Radar RobotCar dataset [41].

The KITTI dataset provides 3D LiDAR scans generated by a Velodyne 64 beam LiDAR sensor mounted on a car. The dataset provides accurate ground truth based on RTK-GPS. In this work, we choose the sequence "*00*" as the benchmark since it lasts 3.7 km and contains a large loop. We use the scans collected at the loop (the scans of seconds 340 to 385) as queries and use the scans before the loop (the scans of seconds 0 to 300) for map building. In particular, the scans of seconds 300-340 are excluded for separating the map and queries.

The NCLT dataset was created at the University of Michigan North Campus using a Velodyne32-HDL LiDAR sensor. The dataset provides ground-truth poses based on a large SLAM solution using LiDAR scan matching and high-accuracy RTK-GPS. We note that the sequences of the dataset are collected on varying routes and cover different parts of the campus. Therefore, we choose the sequence "*2012-01-15*" which can cover most areas of the campus for map building. We use the scans of the sequence "*2012-02-04*" as the queries. For long-term performance evaluation, we also use the sequences collected in different seasons, including "*2012-03-17*," "*2012-06-15*," "*2012-09-28*," "*2012-11-16*," and "*2013-02-23*".

The Oxford Radar RobotCar dataset was created at Oxford and contains sparse 3D LiDAR scans generated by two Velodyne32-VLP LiDAR scanners. We note that the fields of view of the scanners are partially obscured by each other. To form a complete LiDAR scan, we align the data at the same timestamp from the two scanners using the calibration information of the dataset. Since the sequences of the dataset were created on similar routes in 7 days, we only use two sequences in the experiment for convenience. Specifically, we use the sequence "*2019-01-10-11-46-21*" to build a map and the sequence "*2019-01-11-12-26-55*" for evaluation.

B. Metrics

We deploy the localization success rate to evaluate the localization performance when using single scans. We compute the translation error e_t between the estimated translation $\hat{\mathbf{t}}$ and the

[1] https://github.com/ethz-asl/segmap, https://github.com/zjuluolun/BVMatch
[2] https://github.com/opencv/opencv

ground truth $\mathbf{t}$ as

$$e_t = \|\hat{\mathbf{t}} - \mathbf{t}\|_2, \tag{19}$$

and compute the rotation error e_r using the estimated rotation matrix $\hat{\mathbf{R}}$ and the ground truth $\mathbf{R}$ as [42]

$$e_r = \arccos\left(\frac{1}{2}\left(\text{trace}\left(\hat{\mathbf{R}}^{-1}\mathbf{R}\right) - 1\right)\right). \tag{20}$$

We set a threshold $\boldsymbol{\tau} = (\tau_t, \tau_r)$, and regard the localization as successful when both of the conditions $e_t < \tau_t$ and $e_r < \tau_r$ satisfy. The success rate is defined as

$$\text{success rate} = \frac{\text{number of successful localizations}}{\text{total number}}. \tag{21}$$

We use a tight threshold $\boldsymbol{\tau} = (2\,\text{m}, 5^\circ)$ and a loose threshold $\boldsymbol{\tau} = (5\,\text{m}, 10^\circ)$ in the evaluation. The tight one is commonly used in local registration applications [7], [12]. In the experiments we focus more on the loose one, considering that global localization aims to estimate an coarse pose for further pose tracking rather than directly provides precise poses.

Ideally, a vehicle should recover localization without moving when it loses the track of the map. However, this is usually hard to achieve since single LiDAR scans may lack structure information. In this case, the vehicle could travel a dead-reckoning distance to accumulate scene structures. It is crucial to minimize the distance required for successful localization since traveling blindly is unsafe. Following the experimental setup in SegMatch [13], we use the probability of traveling a given distance without successful localization [13] to evaluate the localization performance. We build local maps using LOAM [8] and run the methods on the datasets. We record the travel distance between the successful localizations and computed the probability as

$$P(x) = \frac{S(x)}{S_{\text{totoal}}}, \tag{22}$$

where $S(x)$ is the sum of distance traveled without localization for greater than or equal to x meters, and S_{totoal} is the total traveling distance. The coarse ground truth threshold is used for this metric. It is obvious that the lower $P(x)$, the better the localization performance.

C. Parameter Effects

We conduct experiments on the KITTI dataset to find the optimal parameters of the HOPN descriptor and the transform resolutions of the consensus set maximization algorithm. We adopt the localization success rate using single scans under the loose threshold as the evaluation metric.

There are three main parameters relating to the HOPN descriptor, *i.e.,* the grid size G, the number of orientation intervals B, and the local patch size J. We design three independent experiments, with each experiment having only one variable parameter, to determine the suitable values of these parameters. The experiment setup is detailed in Table I. Note that we fix the transform resolutions of the consensus set maximization algorithm to small values ($\Delta\theta = 1$ degree and $\Delta t = 1$) to guarantee the pose estimation accuracy.

TABLE I
THE PARAMETER SETTINGS OF HOPN DESCRIPTOR

Experiments	Variable	Fixed Parameters
G	$G = \{0.2, 0.3, 0.4, 0.5, 0.6\}$	$B = 8,\ J = 64$
B	$B = \{2, 4, 6, 8, 10\}$	$G = 0.4,\ J = 64$
J	$J = \{32, 48, 64, 80, 96\}$	$G = 0.4,\ B = 6$

TABLE II
THE SUCCESS RATES OF GLOBAL LOCALIZATION UNDER THE DIFFERENT PARAMETER SETTINGS OF HOPN DESCRIPTOR

The results of parameter G					
Fixed parameter	$B = 8,\ J = 64$				
G	0.2	0.3	0.4	0.5	0.6
Success rate (%)	54.3	83.5	87.5	72.4	65.0
The results of parameter B					
Fixed parameter	$G = 0.4,\ J = 64$				
B	2	4	6	8	10
Success rate (%)	74.4	84.4	87.2	87.5	87.6
The results of parameter J					
Fixed parameter	$G = 0.4,\ B = 6$				
J	32	48	64	80	96
Success rate (%)	70.3	88.7	87.2	76.1	66.8

TABLE III
SUCCESS RATES OF GLOBAL LOCALIZATION ON DIFFERENT DATASETS WHEN USING SINGLE LIDAR SCANS

	$\boldsymbol{\tau} = (5\,\text{m}, 10^\circ)$			$\boldsymbol{\tau} = (2\,\text{m}, 5^\circ)$		
	KITTI	NCLT	Oxford	KITTI	NCLT	Oxford
SIFT[20]	28.7	17.3	0.5	19.1	13.7	0.0
BVFT[12]	32.0	17.8	6.7	30.1	15.2	4.2
Ours	**88.7**	**53.7**	**45.3**	**81.2**	**45.9**	**36.3**

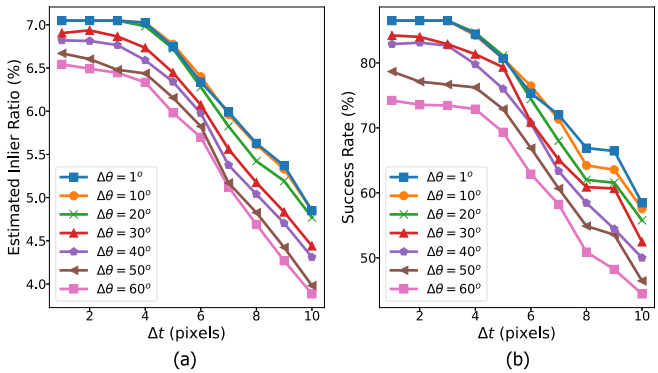

Fig. 7. Performance of the consensus set maximization algorithm under different transform resolution combinations. (a) The estimated inlier ratio. (b) The success rate of pose estimation.

We have three main observations from Table II. **First**, the success rate increases when the grid size G varies from 0.2 to 0.4 but decreases when G becomes further larger. This is because the principal normal computation can be easily affected by noise when G is too small while the locality of the principal normal cannot be well preserved when G is too large. Hence we set $G = 0.4$ meters. **Second**, the success rate increases with the number of orientation intervals B. This is reasonable since the

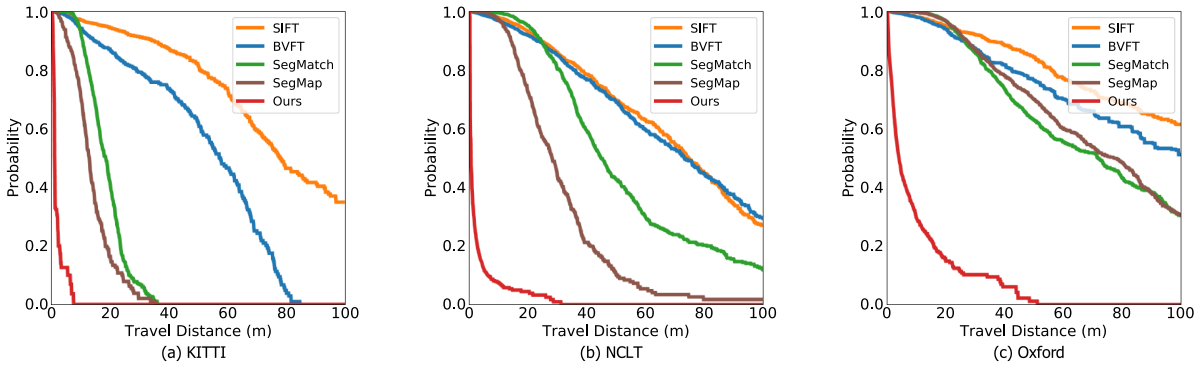

Fig. 8. Probability of traveling a given distance without successful localization on different datasets.

descriptor is more informative when more discrete orientations are encoded. It is noticed that the performance improvement is negligible when $B > 6$. Taking both the computation complexity and success rate of localization into account, we set $B = 6$. **Third**, we achieve the highest success rate when the patch size J is 48. When J is too small, the information encoded into the descriptor will be insufficient. On the contrary, when J is too large, the descriptor will be less distinctive since the description patch of adjacent keypoints will overlap. Based on the above observations, we fixed the parameters of the HOPN descriptor as $G = 0.4$, $B = 6$, $J = 48$ in the following experiments.

We conduct a parameter searching experiment to find the proper transform resolutions of the consensus maximization algorithm. We perform HOPN descriptor matching between the single LiDAR scans of the KITTI dataset and the global map, and obtain 450 match sets. By applying the ground truth transform, we find that the average inlier ratio of these match sets is 7.04%. We estimate poses from the sets using our algorithm and compute the average inlier ratio and success rate under different resolution combinations. From the results illustrated in Fig. 7, we observe that when $\Delta\theta \leq 20^\circ$ and $\Delta t \leq 3$ pixels, the estimated inlier ratio is approximately constant at 7.04%, which is identical to the inlier ratio computed using ground truth. Correspondingly, the success rate 88.7% is the highest. When the resolutions further increase, the estimated number of inliers and the success rate both decrease. Hence, we set $\Delta\theta = 20^\circ$ and $\Delta t = 3$ in our global localization method.

D. Global Localization Performance

We compare the localization performance using single LiDAR scans with SIFT [20] and BVFT [12]. We do not compare with SegMatch [13] and SegMap [14] since they have to accumulate local maps to extract sufficient segments, and cannot be employed to localize single scans. As the amount of scans in the NCLT and the Oxford datasets is very large, we sample the scans for the convenience of evaluation. On the NCLT dataset, we sample scans every 8 meters since the dataset provides accurate ground truth every 8 meters. After the sampling, we obtain 990 scans. For the Oxford dataset, we opt for the same sample strategy as the NCLT dataset, and obtain 1238 scans. Fig. 9 shows the cumulative distribution of the localization error on the three datasets. Our method demonstrates a superior performance by localizing much more scans with higher accuracy. Table III shows the localization success rate under different thresholds. It can be seen that our method performs better than SIFT and BVFT. We note that the performance on the KITTI dataset of all the methods is much better than the performance on the other two datasets. According to our observations, this is mainly due to two reasons. First, the evaluation data and the map data of the NCLT and Oxford datasets are collected on different dates, which leads to structural changes between the data. Second, these two datasets contain challenging places such as long corridors. At these places, the field of view of the LiDAR sensor is small due to occlusion, and accordingly, the LiDAR scans do not have much information.

Fig. 8 shows the probability of given a travel distance without successful localization of the global localization methods. On the KITTI dataset, our method can successfully localize the vehicle within 7.5 meters. While SegMatch [13], SegMap [14], and BVFT [12] can also localize the vehicle as the travel distance accumulates, they have to travel much longer distances than our method. On the NCLT and the Oxford datasets, our method can achieve localization within 31.2 and 53.5 meters, respectively. However, the compared methods fail to perform localization at some places even when the travel distance is larger than 100 meters.

E. Long-Term Robustness

As the appearance of an environment may change over time, the long-term robustness is an essential aspect to evaluate a global localization method. To this end, we compare the performance of our method with SegMatch [13], SegMap [14], SIFT [20], and BVFT [12] on the NCLT dataset. We localize the scans of the evaluation sequences collected across a year on the map. Table IV shows the localization success rate using single LiDAR scans. It can be seen that our method outperforms SIFT and BVFT throughout the year. Fig. 10 shows the probability of given a travel distance without successful localization. Our method achieves localization with much shorter travel distances than the compared methods. It can localize the vehicle within 40 meters regardless of season changes.

TABLE IV
SUCCESS RATES OF DIFFERENT DATE SEQUENCES IN THE NCLT DATASET WHEN USING SINGLE LIDAR SCANS

	$\tau = (5\text{ m}, 10^\circ)$					$\tau = (2\text{ m}, 5^\circ)$				
	2012-03-17	2012-06-15	2012-09-28	2012-11-16	2013-02-23	2012-03-17	2012-06-15	2012-09-28	2012-11-16	2013-02-23
SIFT [36]	14.0	11.3	11.2	8.0	12.2	12.4	8.6	9.4	6.5	9.7
BVFT [12]	14.3	12.1	11.8	14.1	13.5	13.6	11.5	11.4	14.1	13.0
Ours	**65.0**	**54.9**	**42.0**	**45.4**	**59.8**	**49.7**	**41.5**	**32.4**	**33.3**	**45.1**

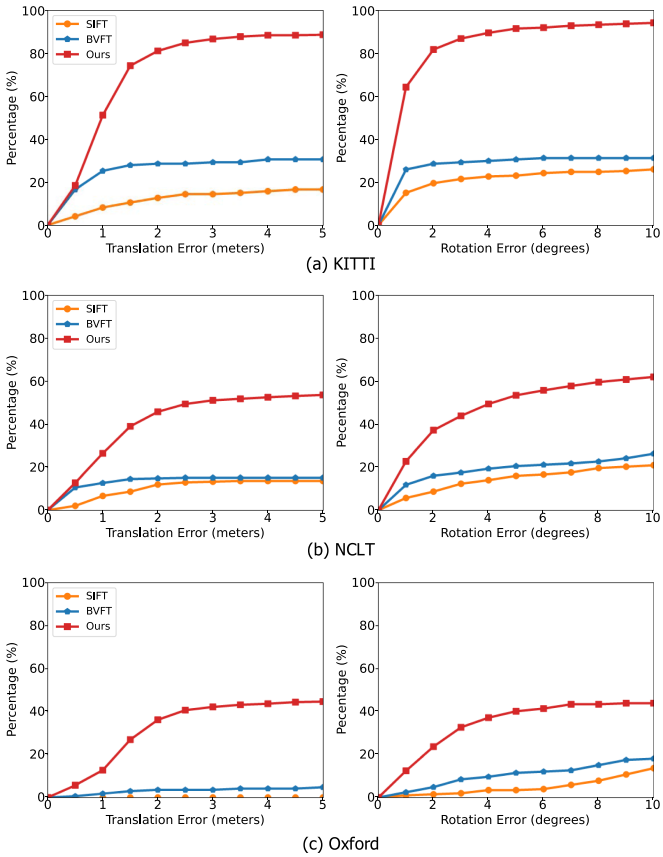

Fig. 9. Cumulative distribution of the localization errors on different datasets when using single LiDAR scans. The distribution of the translation and rotation errors are computed independently.

F. Performance of Consensus Set Maximization

We compare our consensus set maximization algorithm with RANSAC [36] and its two variants GC-RANSAC [44] and R1P-RANSAC [43]. For RANSAC and GC-RANSAC, we use the confidence of 0.99 as the stopping criterion. For R1P-RANSAC, we fixed the scale factor to 1 since there is no scale difference between BV images. We test the algorithms on the sequence "*00*" of the KITTI dataset. We match the LiDAR scans with the map using our HOPN descriptor and adopt these algorithms for pose estimation. Table V shows the localization success rate using single LiDAR scans. Our algorithm achieves the highest success rate while the RANSAC-based methods fail to estimate the poses of some scans due to their random sample mechanism. Fig. 11 illustrates the probability of given a travel distance without successful localization. It can be seen that with our pose estimation algorithm, our localization method can achieve localization with shorter travel distances.

TABLE V
SUCCESS RATES OF DIFFERENT POSE ESTIMATION ALGORITHMS ON THE KITTI DATASET WHEN USING SINGLE LIDAR SCANS

	$\tau = (5\text{ m}, 10^\circ)$	$\tau = (2\text{ m}, 5^\circ)$
RANSAC [20]	79.6	70.3
R1P-RANSAC [43]	69.8	61.3
GC-RANSAC [44]	80.0	72.0
Ours	**88.7**	**81.2**

TABLE VI
LOCALIZATION ACCURACY BEFORE AND AFTER ICP REFINEMENT ON THE NCLT DATASET

	e_t (m)		e_r (deg)	
	Mean	Std.	Mean	Std.
Before	0.61	0.46	1.28	1.55
After	0.09	0.08	0.41	0.47

TABLE VII
AVERAGE TIME COST FOR EACH QUERY OF THE COMPARED METHODS

Method	SIFT	BVFT	SegMatch	SegMap	Ours
Time (in seconds)	0.28	0.43	0.08	0.09	0.39

G. Pose Refinement With ICP

When more accurate localization is desired, we can use ICP to refine the coarse global poses. We compare the localization accuracy of our method before and after the ICP refinement on the NCLT dataset. We compute the translation error e_t and the rotation error e_r for each query and show the results in Table VI. It is observed that the localization error can be significantly reduced after ICP refinement.

H. Running Time

We implemented our method using C++ and run all the experiments on a desktop computer equipped with an Intel Quad-Core 3.40 GHz i5-7500 CPU and 16 GB RAM. The average time cost of our method is 0.25 seconds to extract HOPN descriptors, 0.13 seconds to match the descriptors using the nearest neighbor searching, and 0.01 seconds to perform consensus set maximization. The total time cost for each query is 0.39 seconds. We show the average time cost of each compared method in Table VII. As can be seen, our method runs faster than BVFT [12] but slower than SIFT [20], SegMatch [13], and SegMap [14]. This is a limitation of our method.

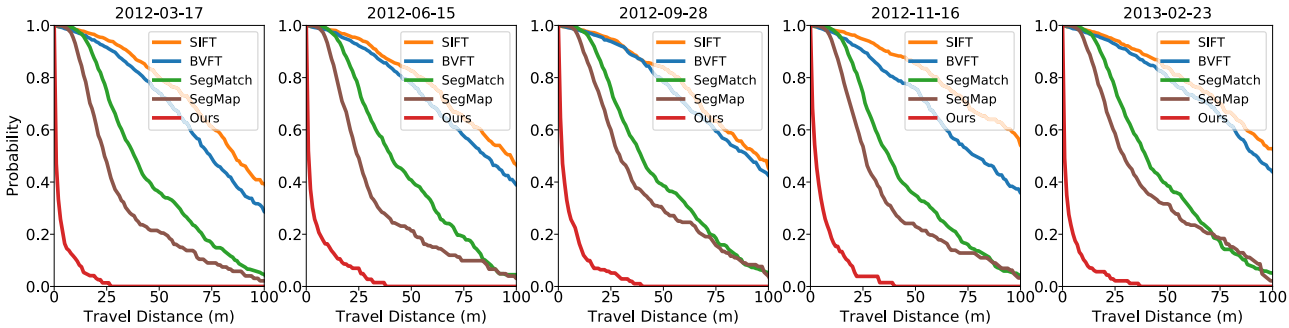

Fig. 10. Long term performance evaluation. Each plot shows the probability of traveling a given distance without successful localization on the different dates on the NCLT dataset.

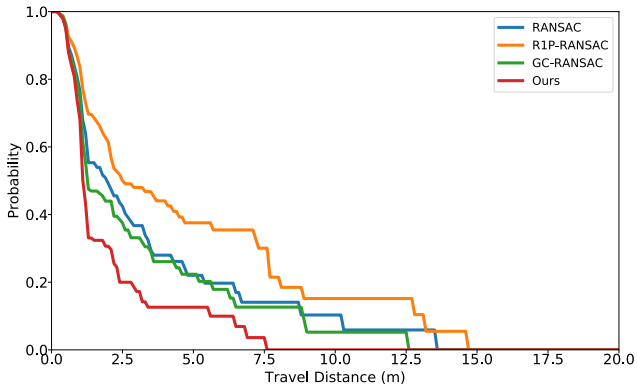

Fig. 11. Probability of traveling a given distance without successful localization using different pose estimation algorithms on the KITTI dataset.

VII. Conclusion

This paper presents a global localization method that can localize local LiDAR scans on the large global map. The method introduces a novel local descriptor called HOPN based on the principal normals to perform matching. Additionally, it presents a fast algorithm to estimate poses from the HOPN match set of the low inlier ratio. Compared to the state-of-the-arts, it can be employed on the point cloud map directly and is more effective for global localization. In our future work, we will focus on improving its performance over short travel distances and reducing its time complexity.

References

[1] C. Cadena *et al.*, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Trans. Robot.*, vol. 32, no. 6, pp. 1309–1332, Dec. 2016.

[2] T. Kos, I. Markezic, and J. Pokrajcic, "Effects of multipath reception on GPS positioning performance," in *Proc. Int. Symp. Electron.*, Mar. 2010, pp. 399–402.

[3] D. Galvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Trans. Robot.*, vol. 28, no. 5, pp. 1188–1197, Oct. 2012.

[4] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5297–5307.

[5] M. A. Uy and G. Hee Lee, "PointNetVLAD: Deep point cloud based retrieval for large-scale place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4470–4479.

[6] F. Cao, F. Yan, S. Wang, Y. Zhuang, and W. Wang, "Season-invariant and viewpoint-tolerant LiDAR place recognition in GPS denied environments," *IEEE Trans. Ind. Electron.*, vol. 68, no. 1, pp. 563–574, Jan. 2021.

[7] J. Du, R. Wang, and D. Cremers, "DH3D: Deep hierarchical 3D descriptors for robust large-scale 6DoF relocalization," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 744–762.

[8] Z. Ji and S. Sanjiv, "LOAM: LiDAR odometry and mapping in real-time," in *Proc. Robot., Sci. Syst.*, 2014, pp. 109–111.

[9] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (FPFH) for 3D registration," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2009, pp. 3212–3217.

[10] F. Tombari, S. Salti, and L. Di Stefano, "Unique signatures of histograms for local surface description," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 356–369.

[11] B. Steder, M. Ruhnke, S. Grzonka, and W. Burgard, "Place recognition in 3D scans using a combination of bag of words and point feature based relative pose estimation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2011, pp. 1249–1255.

[12] L. Luo, S.-Y. Cao, B. Han, H.-L. Shen, and J. Li, "BVMatch: Lidar-based place recognition using bird's-eye view images," *IEEE Robot. Automat. Lett.*, vol. 6, no. 3, pp. 6076–6083, Jul. 2021.

[13] R. Dubé, D. Dugas, E. Stumm, J. Nieto, R. Siegwart, and C. Cadena, "SegMatch: Segment based place recognition in 3D point clouds," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2017, pp. 5266–5272.

[14] R. Dubé, A. Cramariuc, D. Dugas, J. I. Nieto, R. Siegwart, and C. Cadena, "Segmap: 3D segment mapping using data-driven descriptors," in *Proc. Robot., Sci. Syst.*, Jun. 2018.

[15] A. Y. Hata and D. F. Wolf, "Feature detection for vehicle localization in urban environments using a multilayer LiDAR," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 2, pp. 420–429, Feb. 2016.

[16] Z. Wang, J. Fang, X. Dai, H. Zhang, and L. Vlacic, "Intelligent vehicle self-localization based on double-layer features and multilayer LiDAR," *IEEE Trans. Intell. Veh.*, vol. 5, no. 4, pp. 616–625, Dec. 2020.

[17] G. Kim and A. Kim, "Scan context: Egocentric spatial descriptor for place recognition within 3D point cloud map," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 4802–4809.

[18] G. Kim, B. Park, and A. Kim, "1-day learning, 1-year localization: Long-term LIDAR localization using scan context image," *IEEE Robot. Automat. Lett.*, vol. 4, no. 2, pp. 1948–1955, Apr. 2019.

[19] W. Zhang and C. Xiao, "PCAN: 3D attention map learning using contextual information for point cloud based retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12428–12437.

[20] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91–110, Nov. 2004.

[21] H. Li, "Consensus set maximization with guaranteed global optimality for robust geometry estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 1074–1080.

[22] M. Bosse and R. Zlot, "Place recognition using keypoint voting in large 3D lidar datasets," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2013, pp. 2677–2684.

[23] J.-L. Blanco, J. Gonzalez, and J.-A. Fernandez-Madrigal, "A new method for robust and efficient occupancy grid-map matching," in *Proc. Iberian Conf. Pattern Recognit. Image Anal.*, 2007, pp. 194–201.

[24] M. Rapp, T. Giese, M. Hahn, J. Dickmann, and K. Dietmeyer, "A feature-based approach for group-wise grid map registration," in *Proc. IEEE Int. Conf. Intell. Transp. Syst.*, 2015, pp. 511–516.

[25] K. Wang, S. Jia, Y. Li, X. Li, and B. Guo, "Research on map merging for multi-robotic system based on RTM," in *Proc. IEEE Int. Conf. Inf. Automat.*, 2012, pp. 156–161.
[26] T. Shan, B. Englot, F. Duarte, C. Ratti, and R. Daniela, "Robust place recognition using an imaging LiDAR," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 5469–5475.
[27] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 2564–2571.
[28] Z. Liu *et al.*, "LPD-Net: 3D point cloud learning for large-scale place recognition and environment analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2831–2840.
[29] X. Chen *et al.*, "OverlapNet: Loop closing for LiDAR-based SLAM," in *Proc. Robot., Sci. Syst.*, Jun. 2020.
[30] H. Yin, Y. Wang, X. Ding, L. Tang, S. Huang, and R. Xiong, "3D LiDAR-based global localization using siamese neural network," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 4, pp. 1380–1392, Apr. 2020.
[31] L. He, X. Wang, and H. Zhang, "M2DP: A novel 3D point cloud descriptor and its application in loop closure detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2016, pp. 231–237.
[32] S. Ratz, M. Dymczyk, R. Siegwart, and R. Dubé, "OneShot global localization: Instant LiDAR-visual pose estimation," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 5415–5421.
[33] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 430–443.
[34] P. Besl and N. D. McKay, "A method for registration of 3-D shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 239–256, Feb. 1992.
[35] N. Carlevaris-Bianco, A. K. Ushani, and R. M. Eustice, "University of michigan north campus long-term vision and LiDAR dataset," *Int. J. Robot. Res.*, vol. 35, no. 9, pp. 1023–1035, 2016.
[36] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.
[37] Y. Zheng, S. Sugimoto, and M. Okutomi, "Deterministically maximizing feasible subsystem for robust model fitting with unit norm constraint," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1825–1832.
[38] T.-J. Chin, P. Purkait, A. Eriksson, and D. Suter, "Efficient globally optimal consensus maximization with tree search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 758–772, Apr. 2017.
[39] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares fitting of two 3-D point sets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 9, no. 5, pp. 698–700, May 1987.
[40] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3354–3361.
[41] D. Barnes, M. Gadd, P. Murcutt, P. Newman, and I. Posner, "The oxford radar RobotCar dataset: A radar extension to the oxford RobotCar dataset," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 6433–6438.
[42] T. Sattler *et al.*, "Benchmarking 6DOF outdoor visual localization in changing conditions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8601–8610.
[43] H. Zhou, T. Zhang, and J. Jagadeesan, "Re-weighting and 1-point RANSAC-based PnP solution to handle outliers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 12, pp. 3022–3033, Dec. 2019.
[44] D. Barath and J. Matas, "Graph-cut RANSAC," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6733–6741.

Lun Luo received the B.E. degree in 2018 from Zhejiang University, Hangzhou, China, where he is currently working toward the Ph.D. degree with the College of Information Science and Electronic Engineering. His research interests include SLAM and LiDAR-based localization.

Si-Yuan Cao received the B.E. degree from Tianjin University, Tianjin, China, in 2016. He is currently working toward the Ph.D. degree with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China. His research interests include multispectral/multimodal image alignment and image processing.

Zehua Sheng received the B.E. degree in 2017 from Zhejiang University, Hangzhou, China, where he is currently working toward the Ph.D. degree. His research focuses on image denoising.

Hui-Liang Shen received the B.E. and Ph.D. degrees in electronic engineering from Zhejiang University, Hangzhou, China, in 1996 and 2002, respectively. From 2001 to 2005, he was a Research Associate and Research Fellow with The Hong Kong Polytechnic University, Hong Kong. He is currently a Full Professor with the College of Information Science and Electronic Engineering, Zhejiang University. His research interests include multispectral imaging, image processing, computer vision, and machine learning.