Loop Closure Detection via Locality Preserving Matching with Global Consensus

Jiayi Ma, *Senior Member, IEEE*, Kaining Zhang, and Junjun Jiang, *Senior Member, IEEE*

***Abstract*—A critical component of visual simultaneous localization and mapping is loop closure detection (LCD), an operation judging whether a robot has come to a pre-visited area. Concretely, given a query image (i.e., the latest view observed by the robot), it proceeds by first exploring images with similar semantic information, followed by solving the relative relationship between candidate pairs in the 3D space. In this work, a novel appearance-based LCD system is proposed. Specifically, candidate frame selection is conducted via the combination of Super-features and aggregated selective match kernel (ASMK). We incorporate an incremental strategy into the vanilla ASMK to make it applied in the LCD task. It is demonstrated that this setting is memory-wise efficient and can achieve remarkable performance. To dig up consistent geometry between image pairs during loop closure verification, we propose a simple yet surprisingly effective feature matching algorithm, termed locality preserving matching with global consensus (LPM-GC). The major objective of LPM-GC is to retain the local neighborhood information of true feature correspondences between candidate pairs, where a global constraint is further designed to effectively remove false correspondences in challenging sceneries, *e.g.*, containing numerous repetitive structures. Meanwhile, we derive a closed-form solution that enables our approach to provide reliable correspondences within only a few milliseconds. The performance of the proposed approach has been experimentally evaluated on ten publicly available and challenging datasets. Results show that our method can achieve better performance over the state-of-the-art in both feature matching and LCD tasks. We have released our code of LPM-GC at https://github.com/jiayi-ma/LPM-GC.**

***Index Terms*—Feature matching, locality preserving matching, loop closure detection, SLAM.**

I. Introduction

SIMULTANEOUS localization and mapping (SLAM) system is an enabling technology for mobile devices to accomplish self-localization and map construction [1], [2]. Generally speaking, the trajectory calculated directly via the frame-to-frame visual odometry is prone to generate drift by virtue of errors in each pose estimation. To resolve the problem, the loop closure detection (LCD) module is introduced. It identifies re-observations during navigation, followed by i) recovering an Sim(3) pose from a closed-loop event ii) optimizing the pose graph by incorporating the recovered Sim(3) pose as a constraint. As such, effective LCD methods can help with providing accurate pose estimation, as well as mapping the environment with global consensus. The paper in hand mainly discusses how to search closed-loop events, and focuses on appearance information captured by exteroceptive cameras.

Appearance-based LCD is achieved via visual place recognition [3], and is typically composed of two components: *candidate frame selection* and *loop closure identification*. The former operation, to a large degree, greatly coincides with the image retrieval task due to its trying to search the most similar visual entry into an already captured database, *viz.*, pre-visited camera observations. This step aims to find as many valid candidate frames as possible to acquire a high recall rate. Most state-of-the-art approaches devote themselves to searching for a more reasonable representation for images. This has transformed from aggregating local descriptors [4]–[6], to directly train a network with ranking-based or classification losses [7]–[10]. The former approaches are usually constructed on the bag-of-words (BoW) scheme [11]. It is achieved by first quantizing the descriptor space via k-means clustering [12] to produce visual words, then coupling with the widely used term-frequency inverse-document-frequency (TF-IDF) technique to generate visual word histograms that represent the incoming sensory measurements. Due to the neglect of spatial information of scenes, the performance of the initial BoW strategy would degenerate when repetitive or similar structures appear. This drawback impels the research community to adopt more sophisticated solutions [13]–[16] which enhance spatial information by altering the encoding pattern.

With regard to loop closure verification, it is also known as geometric consistency checking module, aiming at mining geometric relationship between candidate pairs and enhancing the precision of the LCD system. It is achieved by first searching for reliable correspondences between image pairs and then recovering the camera pose through the calculation of fundamental or essential matrix [17]. In the process of reliable correspondence construction, a putative set is first established as per local descriptor similarity. Later, false matches of the putative set are filtered via a feature matching approach. The most broadly utilized algorithm in this process is RANSAC [18]. It optimizes the maximum consensus loss by alternating between hypothesis and verification, which ends

This work was supported by the Key Research and Development Program of Hubei Province (2020BAB113). Recommended by Associate Editor Yang Xiang. *(Corresponding author: Jiayi Ma.)*

Citation: J. Y. Ma, K. N. Zhang, and J. J. Jiang, "Loop closure detection via locality preserving matching with global consensus," *IEEE/CAA J. Autom. Sinica*, 2022, DOI: 10.1109/JAS.2022.105926

J. Y. Ma and K. N. Zhang are with the Electronic Information School, Wuhan University, Wuhan 430072, China (e-mail: jyma2010@gmail.com; zkn19961212@whu.edu.cn).

J. J. Jiang is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China (e-mail: jiangjunjun@hit.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2022.105926

with describing the mapping relationship of putative correspondences via a parametric model. This technique is effective when the geometric model between an image pair is rigid. Yet in LCD tasks, the existence of camera distortion would lead to *non-rigid deformation* whilst dynamic occlusions and similar structures such as trees, cars and buildings would result in *a high outlier ratio* [19]. These two factors make the performance of RANSAC degenerate severely. Thus using RANSAC would cut down both accuracy and efficiency when challenging scenes appear in the LCD task.

The contributions of this paper are three-fold:

1) A state-of-the-art deep local descriptors (i.e., Super-features [20]) proposed for image retrieval is applied in the LCD task. However, in image retrieval, its coupling with aggregated selective match kernel (ASMK) [16] needs to extract descriptors of all database images in advance, which cannot be directly applied in LCD where the database scales up with the navigation of a robot/vehicle. So we incorporate an incremental strategy within Super-features+ASMK, which is demonstrated both computationally and memory-wise efficient and can yield great LCD performance.

2) To tackle the drawback of RANSAC, a novel feature matching approach called locality preserving matching with global consensus (LPM-GC) is proposed to efficiently establish reliable correspondences in complex LCD scenes, e.g., image pairs with many repetitive patterns and large mismatch proportions. Specifically, LPM-GC aims to retain the local neighborhood information among putative matches and make the motion vectors across the whole image consistent with each other. A solution with closed form is further deduced which enables our model to realize feature matching in only a few milliseconds.

3) Extensive experiments are performed to measure the performance of the proposed approach on both feature matching and LCD tasks. We show that our approach can acquire impressive results in long-term traversal.

The remainder of the paper is organized as follows. Section II introduces related work in the context of LCD in brief. Section III presents our proposed LCD system in detail, including candidate frame selection using Super-features + ASMK, and loop closure identification using LPM-GC. Section IV gives comprehensive experimental results and comparative evaluation against existing advanced LCD approaches. The whole paper ends with conclusion remarks outlined in Section V.

II. Related Work

The state-of-the-art LCD methods typically put attention on how to select candidate frames more effectively, whereas the verification process is scarcely studied. RANSAC is typically regarded as the default setting to remove mismatches. To give the readers a deeper understanding of the points of interest in our work, we first review prevalent LCD approaches and then give an introduction to feature matching methods [21].

A. Candidate Frame Selection

There are two classification criteria for the candidate frame selection techniques. One is judging whether the algorithm adopts on-line or off-line strategy, and the other is determining whether the algorithm is image-based or sequence-based. Often, off-line methods train visual dictionaries in advance, according to which BoW vectors are calculated directly to represent images [4], [22]–[25]. For example, a probabilistic model over the BoW representation is employed in FAB-MAP [22], [23], where a Chow Liu tree is additionally involved to determine the probabilities of co-occurrence of visual words. Later, to make LCD more compatible with the realtime requirement, [4] proposes to construct the visual dictionary based on binary descriptors. On-line strategies refer to building visual dictionaries on the fly, and they become increasingly popular on account of their scene-agnostic abilities [26]–[28]. For example, Angeli *et al.* [26] constructed on-line dictionaries by performing comparison between each new descriptor and its nearest visual word. Distances greater than a pre-defined threshold would lead to new word generation. iBoW-LCD [28] proposes to save a binary dictionary in the form of a hierarchical tree. Then the update of the dictionary is performed by adding or deleting visual words based on the similarity of binary descriptors. Tsintotas *et al.* [27] proposed to generate on-line visual words via a tracking approach, followed by a voting scheme to determine candidate frames.

Image-based and sequence-based strategies are distinguished by the smallest unit in a system, respectively. The former takes each image as an individual and judges whether a loop-closing pair appears in it, while the latter takes the image continuing in time as a group and determines whether a loop-closing pair arises in it [5], [29]–[31]. For example, [30] utilizes an on-line dictionary. The input image stream is partitioned into places based on a simple feature matching scheme and visual words of each place are yielded via a clustering technique. As what is done in [27], candidate frames are finally identified based on a voting scheme. In PRE-VieW [31], an off-line dictionary is adopted and a new sequence is instigated if there are enough new visual words observed. Then candidate frames are searched from sequence-to-sequence, to image-to-image matching, in L2-norm scoring.

Recently, the convolutional neural networks (CNNs) have been adopted to tackle the LCD task [6], [8]–[10], [32], [33]. Existing approaches mainly focus on how to boost the performance of image representation via CNN. They may generate global descriptors via CNNs directly, or via deep local features coupled with BoW-related strategies. However, regardless of the way in which the candidate frame selection session is conducted, the subsequent processes of feature matching as well as pose estimation are still based on local features. Therefore, designing networks to extract local features suitable for LCD-related tasks is another worthwhile research topic.

B. Loop Closure Identification & Feature Matching

Feature matching is a fundamental problem in many computer vision problems [21], [34]. In particular, RANSAC [18] is the most commonly utilized approaches for feature matching in LCD tasks. But its parametric-model-based and resampling-based nature limit its performance in scenes with non-rigid deformation or gross outliers. Although considerable

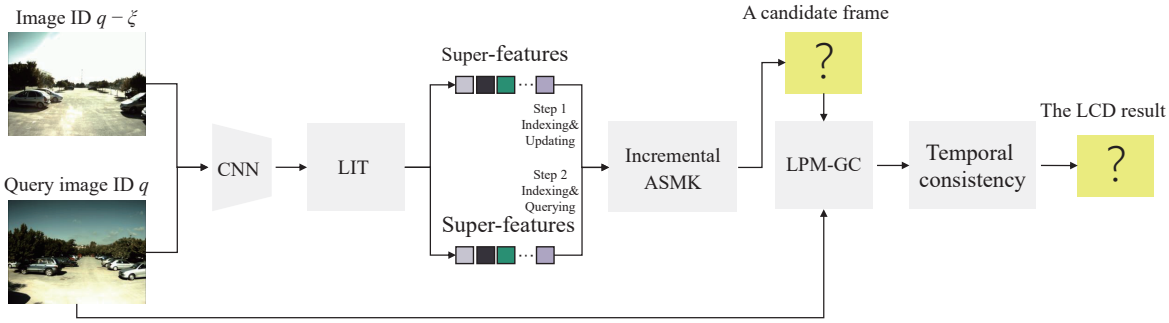


Fig. 1. An overview of the proposed LCD system and detailed introduction can be found in Section III-A. CNN is ResNet50 excluding the last convolutional block, while LIT stands for an iterative module exploited to extract Super-features. q is the ID of the query image, and ξ equals to $f \cdot \mathcal{T}$ with f and $\mathcal{T}$ standing for the frame rate (Hz) and non-search time (s), respectively. Due to the sequential nature of LCD data, images captured consecutively would have high visual commonality. So to prevent false-positive detection from happening among them, we only add images with IDs smaller or equal to $q-\xi$ to the database.

variants [35]–[38] have been proposed to improve its performance, the drawback mentioned above still exists. To address this issue and model in high dimension representation, numerous non-parametric techniques are investigated and have shown promising performance in both rigid and non-rigid matching [39]–[42], but the optimal solution is challenging to determine due to the large search space particularly when massive outliers and/or dynamic objects exist in the image pairs. In addition, another strategy, termed as graph matching, is widely studied as well, which achieves feature matching by optimizing a quadratic programming problem [43], [44]. Nevertheless, its low efficiency cannot guarantee the realtime requirement of LCD tasks. For the sake of fast matching, some methods are lately published based on local information consistency [45]–[48]. In particular, Zhang *et al.* [49] proposed the local motion and structure consensus for robust feature matching during loop closure verification.

Also, during recent years, learning-based methods are increasingly investigated to address feature matching problem and achieve encouraging performance. In addition to implement feature detection and description, many researchers try to identify reliable matches and/or estimate the camera pose information by means of deep convolutional networks. Yi *et al.* [50] performed a first try to simultaneously conduct inliers/outliers recognition, as well as pose estimation. Ma *et al.* [51] designed a high-dimensional feature based on local information for each putative correspondence, followed by feeding the features into a network to identify inliers/outliers. Sarlin *et al.* [52] introduced a neural network that is able to achieve feature matching by constructing a flexible context aggregation mechanism based on attention, and it can learn priors over geometric transformations and regularities of the 3D world through end-to-end training from image pairs. In our work, a novel algorithm for matching complex scenes in the LCD task is designed, and our experiments demonstrate its significant improvements comparing with the competitors.

III. Methodology

This section illustrates the proposed LCD system. As the schematic diagram shown in Fig. 1, given a query image, two procedures involving updating and querying the database are performed one after another. Then the identified candidate frame would constitute an image pair with the query one, followed by being transferred into the stage of geometric confirmation which is achieved via our LPM-GC. A loop-closing event is confirmed only when an adequate number of matches are retained as well as the temporal consistency test is passed. **Note that notations in different subsections of this section are mutually independent except for special illustrations.**

A. Candidate Frame Selection via Super-Features & ASMK

Super-Features: Often, ASMK is coupled with handcrafted local descriptors to generate global embeddings of images. Recently, some literatures show that learned local descriptors can highly outperform the handcrafted ones [20], [53]. Thus in this paper, we exploit a learned one dubbed Super-features proposed in *feature integration-based retrieval* (FIRe) [20] to achieve candidate frame selection. It is obtained from the middle level of a network and trained directly on the features themselves. In the following, we would first give a brief introduction to its background.

Given an image, an ordered set of Super-features of it can be extracted via a module called local feature integration transformer (LIT). To be specific, ResNet50 without the last convolutional block serves as the backbone to firstly extract local features $\mathcal{U} = \{\mathbf{u}_l \in \mathbb{R}^D, l = 1,2,\ldots,L\}$ of images. Then LIT can be formulated as a function $\Phi(\mathcal{U}): \mathbb{R}^{L\times D} \longrightarrow \mathbb{R}^{N\times d}$ that takes as input the set of local features and outputs N Super-features. It is defined as an iterative module with $t \in [1,T]$

$$\Phi(\mathcal{U}) = \mathcal{Q}^T, \qquad \mathcal{Q}^t = \phi(\mathcal{U}, \mathcal{Q}^{t-1}) \tag{1}$$

where $\phi(\cdot)$ is the core function applied T times and $\mathcal{Q}^0 = \{\mathbf{q}_n^0 \in \mathbb{R}^d, n = 1,2,\ldots,N\}$ is initialized with a set of learnable templates. Equation (1) indicates that Super-features are progressively formed by iterative refinement of the templates, conditioned on the local features $\mathcal{U}$ from the backbone.

Specifically for $\phi(\cdot)$, it is inspired by the Transformer architecture [54] which takes *key*, *query*[1] and *value* as inputs. In our case, *key* and *value* are respectively calculated via layer normalization and the corresponding linear projection functions $K(\cdot)$ and $V(\cdot)$ on the local features $\mathbf{u}_l \in \mathcal{U}$ in each itera-

[1] In the introduction of "Super-features", the term "query" refers to the input of the Transformer architecture, which has different meaning with that in image retrieval and LCD.

tion, while *query* is calculated via layer normalization together with another linear projection $Q(\cdot)$ on each $\mathbf{q}_n^{t-1}$ in the t-th iteration. The three linear projection functions $K(\cdot)$, $V(\cdot)$ and $Q(\cdot)$ respectively project the input to dimensions d_k, d_v and d_q, and we set $d_k = d_v = d_q = d = 1024$ as per FIRe. Formally, in the t-th iteration, the attention scores of each component of *key* over all ones of *query* are given by

$$\boldsymbol{\alpha}_l = \frac{\hat{\boldsymbol{\alpha}}_l}{\sum_{i=1}^{L}\hat{\boldsymbol{\alpha}}_i}, \hat{\boldsymbol{\alpha}}_l = \frac{e^{\boldsymbol{M}_l}}{\sum_{n=1}^{N} e^{M_{ln}}}, M_{ln} = \frac{K(\boldsymbol{u}_l)\cdot Q\left(\boldsymbol{q}_n^{t-1}\right)}{\sqrt{d}} \tag{2}$$

where all attention scores can be denoted as $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_L] \in \mathbb{R}^{N\times L}$ with $\boldsymbol{\alpha}_l \in \mathbb{R}^{N\times 1}$. Afterwards, $\mathcal{Q}^t$ can be updated by

$$\begin{cases} \mathcal{Q}^t = \phi\left(\mathcal{U}, \mathcal{Q}^{t-1}\right) \\ \phi(\mathcal{U}, \mathcal{Q}^{t-1}) = \mathrm{MLP}(\psi(\mathcal{U}, \mathcal{Q}^{t-1})) + \psi(\mathcal{U}, \mathcal{Q}^{t-1}) \\ \psi(\mathcal{U}, \mathcal{Q}^{t-1}) = \boldsymbol{\alpha}\cdot V(\mathcal{U}) + \mathcal{Q}^{t-1}. \end{cases} \tag{3}$$

Finally, the Super-features are generated by performing whitening and L2-normalization on $\mathcal{Q}^T$. Intuitively, each Super-feature is a function over all local features in $\mathcal{U}$, invariant to permutation of its elements. Thus the set of Super-features is ordered, based on which the whole network can be trained by comparing Super-features with the same ID among images with the same landmark.

In most deep image retrieval approaches, local features boil down to the localized map activations of a neural network, where the optimization of the network is performed on the global feature aggregated by these local ones. In this case, training actually happens on the global feature while testing acts on local feature matching. This makes a discrepancy between training and testing, meanwhile leading to the redundancy of the local features. FIRe addresses the abovementioned two issues by training on the local features themselves, which eventually leads to the compactness of Super-features.

Incremental ASMK: After obtaining local descriptors, ASMK is exploited for indexing and retrieval. Note that we use the binarized version of it to improve computational and memory-wise efficiency for the LCD task. Suppose we have obtained a set $\mathcal{X} = \{\mathbf{x} | \mathbf{x} \in \mathbb{R}^d\}$ of $n = |\mathcal{X}|$ d-dimensional local descriptors of an image, where $|\cdot|$ denotes the cardinality of a set. Based on a k-means quantizer, a visual codebook $C = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k\}$ comprising $|C| = k$ visual words is acquired in advance. We denote by $q(\mathbf{x})$ the nearest neighbor word for each local descriptor $\mathbf{x}$, and denote by $\mathcal{X}_\mathbf{c} = \{\mathbf{x} | \mathbf{x} \in \mathcal{X} : q(\mathbf{x}) = \mathbf{c}\}$ the subset of descriptors in $\mathcal{X}$ that are assigned to the visual word $\mathbf{c}$. Denoting by $\mathcal{X}$ and $\mathcal{Y}$ the local descriptors of images X and Y, while $C_\mathcal{X}$ and $C_\mathcal{Y}$ the words occurring in the two images respectively, then their similarity can be calculated by

$$K(X, Y) = \gamma(\mathcal{X})\gamma(\mathcal{Y}) \sum_{\mathbf{c}\in C_\mathcal{X}\cap C_\mathcal{Y}} \sigma(\Phi(\mathcal{X}_\mathbf{c})^T \Phi(\mathcal{Y}_\mathbf{c})). \tag{4}$$

Taking items *w.r.t.* $\mathcal{X}$ for illustration, $\Phi(\mathcal{X}_\mathbf{c})$ is an aggregated vector representation of local descriptors in $\mathcal{X}$ assigned to the same visual word $\mathbf{c}$. $\sigma(\cdot)$ is a scalar selectivity function, and $\gamma(\mathcal{X})\gamma(\mathcal{Y}) = \sum_{\mathbf{c}\in C_\mathcal{X}\cap C_\mathcal{Y}} \sigma(\Phi(\mathcal{X}_\mathbf{c})^T \Phi(\mathcal{Y}_\mathbf{c}))$ is a normalization factor. Denoting an aggregated residual for a visual word $\mathbf{c}$ by $V(\mathcal{X}_\mathbf{c}) = \sum_{\mathbf{x}\in\mathcal{X}_\mathbf{c}} \mathrm{sign}(\mathbf{x} - q(\mathbf{x}))$, then $\Phi(\mathcal{X}_\mathbf{c})$ can be calculated by a normalized aggregated residual $\hat{V}(\mathcal{X}_\mathbf{c}) = V(\mathcal{X}_\mathbf{c}) / \|V(\mathcal{X}_\mathbf{c})\|$. Regarding $\sigma(u)$, it is defined as a threshold polynomial selectivity function

$$\sigma_\phi(u) = \begin{cases} \mathrm{sign}(u)|u|^\phi, & u > \gamma \\ 0, & u \le \gamma. \end{cases} \tag{5}$$

ASMK enjoys saving memory requirements and searching efficiently by exploiting an inverted file indexing structure. Since the scale of the database increases with the navigation of a robot/vehicle, we revise the vanilla inverted file[2] to an incremental one to make it applied in LCD, and show an example of it in Fig. 2. Specifically, we denote I_{new} the image to be inserted into the inverted file. Then the update of the inverted file starts by searching the nearest neighbor of each descriptor of I_{new} from the codebook. As the gray block shown in Fig. 2, each word in the codebook is attached with a list and if a word occurs in an image, the ID of the image would be recorded in the corresponding list. Regarding the blue one, each image also corresponds to a list, recording the IDs of words occurring in each image and their accumulated binary residuals. Thus when I_{new} occurs, the inverted file is updated by renewing the gray and the blue blocks, respectively. Similarly, the procedure of querying starts by first extracting the descriptors of the query image I_q and then searching the nearest neighbor of each descriptor from the codebook. Based on the gray block shown in Fig. 2, database images which have the same words with I_q are first identified, then similarity calculated by (4) would be performed based on the information recorded in the blue block. Intuitively, the indexing latency and the memory footprint of our incremental ASMK strategy are proportionate to the number of non-empty ASMK clusters of the codebook, i.e., the number of words with at least one assignment. This property is related to the compactness of the used descriptor. Thus exploiting Super-feature can improve computational and memory-wise efficiency of the candidate frame selection stage, which is demonstrated in the experimental part.

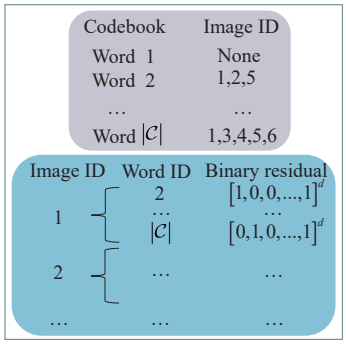


Fig. 2. An example illustrating the structure of the inverted file exploited in our incremental ASMK. The gray block records which images each word occurs in, while the blue one records the words appearing in each image and their corresponding binary residuals.

Overall, there are three elements that guarantee high execution speed and low memory consumption of our candidate

[2] https://github.com/jenicek/asmk

frame selection strategy: the compact Super-features, the binarilized version of ASMK, and the incremental inverted file. In this paper, we extract Super-features by the official model and set $\phi = 3$, $\gamma = 0$ to conduct the calculation in ASMK.

How to Perform Candidate Frame Selection? Not all images captured before the query one (I_q) would be added to the database. This is because they share more or less co-visibility areas with I_q due to the sequential nature of the incoming instances in LCD. If detection is triggered among them, false-positive events would occur unavoidably. To address this issue, we first denote q the image ID of I_q, and r the image ID of any image captured before the query one. Since no LCD events would occur at the start of a traversal, we only consider the case of $q > \xi$, where ξ equals to $f \cdot \mathcal{T}$ with f and $\mathcal{T}$ standing for the frame rate and non-search time, respectively. Then only when $r = q - \xi$ is satisfied would I_r be applied to update the inverted file. In other words, updating starts when $q = \xi + 1$. When the procedure of updating is finished, similarity measurement in (4) would be performed between I_q and the database. Finally, the most similar database image would constitute a candidate pair with I_q and be sent to the stage of geometric verification.

B. Loop Closure Identification via LPM-GC

The second stage of our framework involves a geometric confirmation step. It is achieved by feature matching, serving to test whether the image pair has consistent relative planar geometry. For this, we first build a putative set by coupling the descriptor similarity with a ratio test [55]. Then our interest lies in how to filter the false correspondences in the set. Mathematically, we denote by $S_p = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$ the putative set with $\mathbf{x}_i$ and $\mathbf{y}_i$ being the coordinates of the keypoints. If $(\mathbf{x}_i, \mathbf{y}_i)$ constitutes a true correspondence, we call $\mathbf{x}_i/\mathbf{y}_i$ an inlier, otherwise an outlier. Our LPM-GC functions as a filter, which guarantees as many inliers are recognized as possible.

1) Locality Preserving Matching: Locality preserving matching (LPM) [45] is proposed for effective mismatch removal. The main contribution of it is that it does not need any advance estimation for the transformation between two images and can realize accurate correspondences in real time. In the following, we would revisit it in brief.

Generally speaking, the local neighboring information of a keypoint would not change freely in spite of local non-rigid deformation. As such, we can distinguish inliers and outliers in accordance with the consistency of local neighboring information. This leads to a cost function of

$$C(\mathcal{I}; S_p, \lambda) = \sum_{i \in \mathcal{I}} \frac{1}{2K} \left(\sum_{j|\mathbf{x}_j \in \mathcal{N}_{\mathbf{x}_i}} \left(d\left(\mathbf{x}_i, \mathbf{x}_j\right) - d\left(\mathbf{y}_i, \mathbf{y}_j\right) \right)^2 + \sum_{j|\mathbf{y}_j \in \mathcal{N}_{\mathbf{y}_i}} \left(d\left(\mathbf{x}_i, \mathbf{x}_j\right) - d\left(\mathbf{y}_i, \mathbf{y}_j\right) \right)^2 \right) + \lambda (N - |\mathcal{I}|) \tag{6}$$

where $\mathcal{I}$ is the inlier set, the first term means that the pairwise distance of a true correspondence should be retained, the second term penalizes the outliers, λ controls the trade-off between the two terms, $\mathcal{N}_{\mathbf{x}}$ is the K nearest neighbors of $\mathbf{x}$, $d(\cdot,\cdot)$ is a distance measuring function, and $|\cdot|$ denotes the cardinality of a set. With the cost function defined in (6), the optimization problem can be given by

$$\mathcal{I}^* = \arg\min_{\mathcal{I}} C(\mathcal{I}; S_p, \lambda). \tag{7}$$

Thus our goal is to find an optimal solution $\mathcal{I}^*$ which can minimize the cost function in (6).

Generally, $d(\cdot,\cdot)$ can be defined in the form of L1-norm or L2-norm. But this kind of distance between a point pair can not be retained well due to the existence of non-rigid transformation (e.g., scale change) between an image pair. To ease this concern, the output of $d(\cdot,\cdot)$ is defined in a discrete domain

$$d(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} 0, & \mathbf{x}_j \in N_{\mathbf{x}_i} \\ 1, & \mathbf{x}_j \notin N_{\mathbf{x}_i} \end{cases} \tag{8}$$

and $d(\mathbf{y}_i, \mathbf{y}_j)$ has a similar definition.

2) LPM With Global Consensus: As can be seen from the cost function defined in (6), LPM only concentrates on preserving the local neighborhood structures of putative matches. Therefore, its ability of distinguishing outliers and inliers would decrease when the scenes are complex, for example, image pairs involving many repetitive patterns or a large number of mismatches, leading to wrong judgements on loop closure verification. To address this issue, we design a global consensus based on LPM, termed as LPM-GC, to promote the matching accuracy. In the following, we first introduce two strategies to improve the neighborhood structure characterization [45], and then detail the consensus of global motion structure.

Multi-Scale Strategy: As the locations of keypoints in an image are typically not subject to uniform distribution and different putative sets have various proportion of outliers, the optimal value of neighbor number K could vary, and using a fixed K will reduce the robustness. Therefore, by considering a multi-scale neighborhood strategy, i.e., $\mathbf{K} = \{K_m\}_{m=1}^{M}$, the cost function C in (6) can be rewritten as

$$C = \sum_{i \in \mathcal{I}} \sum_{m=1}^{M} \frac{1}{2MK_m} \left(\sum_{j|\mathbf{x}_j \in \mathcal{N}_{\mathbf{x}_i}^{K_m}} \left(d\left(\mathbf{x}_i, \mathbf{x}_j\right) - d\left(\mathbf{y}_i, \mathbf{y}_j\right) \right)^2 + \sum_{j|\mathbf{y}_j \in \mathcal{N}_{\mathbf{y}_i}^{K_m}} \left(d\left(\mathbf{x}_i, \mathbf{x}_j\right) - d\left(\mathbf{y}_i, \mathbf{y}_j\right) \right)^2 \right) + \lambda (N - |\mathcal{I}|) \tag{9}$$

where $\mathcal{N}_{\mathbf{x}}^{K_m}$ denotes the K_m nearest neighbors of $\mathbf{x}$.

By using the definition of the distance in (8), we can simplify the cost function in (9) as follows:

$$C = \sum_{i \in \mathcal{I}} \sum_{m=1}^{M} \frac{1}{MK_m} \sum_{j|\mathbf{x}_j \in \mathcal{N}_{\mathbf{x}_i}^{K_m}} d\left(\mathbf{y}_i, \mathbf{y}_j\right) + \lambda (N - |\mathcal{I}|). \tag{10}$$

We provide the derivation of (10) in the Appendix.

Consensus of Neighborhood Topological Structure: As shown in Fig. 3, $(\mathbf{x}_i, \mathbf{y}_i)$ is a randomly selected putative match of this image pair. The points within the pink circle are their $K = 5$ nearest neighbors and pink points $(\mathbf{x}_j, \mathbf{y}_j)_{j=1}^{n_i^{se}}$ are the cor-

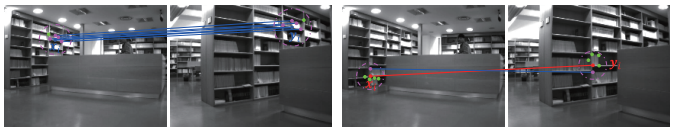


Fig. 3. Illustration of the consensus of neighborhood topological structure. The putative match $(\mathbf{x}_i, \mathbf{y}_i)$ (connected by bold lines) is an inlier in the left (blue) and an outlier in the right (red). Centring on $\mathbf{x}_i$ and $\mathbf{y}_i$, the points within the pink circle are their K nearest neighbors, where the pink ones represent their local support, e.g., common elements in $\mathcal{N}_{\mathbf{x}_i}$ and $\mathcal{N}_{\mathbf{y}_i}$, while green for otherwise.

responding n_i^{se} local support elements, e.g., intersection of the two neighborhoods $\mathcal{N}_{\mathbf{x}_i}$ and $\mathcal{N}_{\mathbf{y}_i}$, where $n_i^{se}=4$ and $n_i^{se}=1$ in the two examples, respectively. Here we define the motion vector of a correspondence $(\mathbf{x}_i, \mathbf{y}_i)$ as $\mathbf{m}_i = \mathbf{y}_i - \mathbf{x}_i$, while that of its local support as $\mathbf{m}_j = \mathbf{y}_j - \mathbf{x}_j$. Intuitively, there are two implicit rules existing in the local information: i) n_i^{se} of a true correspondence would be larger than that of a false correspondence ii) the motion vector of an inlier tends to be more consistent with those of its local supports. Both of the rules suggest that inliers generally have a more similar neighborhood topology than outliers. This conclusion can further help with inlier recognition. For this, the neighborhood topology consensus can be mathematically defined as

$$s(\mathbf{m}_i, \mathbf{m}_j) = \frac{\min\left\{|\mathbf{m}_i|, |\mathbf{m}_j|\right\}}{\max\left\{|\mathbf{m}_i|, |\mathbf{m}_j|\right\}} \cdot \cos\left\langle \mathbf{m}_i, \mathbf{m}_j \right\rangle \tag{11}$$

where the first and second terms reflect the length and orientation consistency between $\mathbf{m}_i$ and $\mathbf{m}_j$, respectively. This definition results in $s(\mathbf{m}_i, \mathbf{m}_j) \in [-1, 1]$, and a larger $s(\mathbf{m}_i, \mathbf{m}_j)$ indicates higher consensus. As per (11), a two-level local topological cost $d(\mathbf{m}_i, \mathbf{m}_j)$ can be defined as

$$d(\mathbf{m}_i, \mathbf{m}_j) = \begin{cases} 0, & s(\mathbf{m}_i, \mathbf{m}_j) \geq \tau \\ 1, & s(\mathbf{m}_i, \mathbf{m}_j) < \tau. \end{cases} \tag{12}$$

Consensus of Global Motion Structure: In terms of the fact that correct loop-closing pairs are always the image pairs captured within a few meters, the length of motion vectors between putative matches should maintain a high degree of consensus. As Fig. 4 shows, the first row successively depicts the loop closure events where the robot is in the ideal case, goes ahead, turns right, and turns left. To characterize the consensus of $(\mathbf{x}_i, \mathbf{y}_i)$ with respect to the whole putative set S_p, we calculate the lengths of the motion vectors $\{\mathbf{m}_i\}_{i=1}^N$ and normalize them into $[0, 1]$ to deal with scale changes, denoted as $\{l_i\}_{i=1}^N$. The bottom of Fig. 4 illustrates the distribution of $\{l_i\}_{i=1}^N$ from 0 to 1. Clearly, the lengths of motion vectors associated by inliers in these four scenes all have high consensus, while those of outliers always scatter randomly.

Based on this observation, we develop a global consensus factor $d_{gc}(\mathbf{x}_i, \mathbf{y}_i)$ to refine the original cost function. We first adopt a clustering strategy such as mean shift [56] to cluster $\{l_i\}_{i=1}^N$, and subsequently obtain several clusters $\{C_j\}_{j=1}^J$ automatically without requiring to set the cluster number J in advance. Note that in the mean shift clustering, there is a parameter r used to assign the points around a clustering center with radius r to that cluster. The value of r can influence the consensus degree of the motion vectors. A large r means the inclusiveness of each cluster is high, leading to a low consistency, and vice versa.

After obtaining the clusters $\{C_j\}_{j=1}^J$, we calculate their cardinalities denoted as $\{n_j\}_{j=1}^J$. For a putative match $(\mathbf{x}_i, \mathbf{y}_i)$, if its corresponding motion vector length l_i belongs to cluster C_j, then the support value α_i can be defined as

$$\alpha_i = \frac{n_j}{N}. \tag{13}$$

As we can see, when l_i belongs to a large cluster, α_i will be large, which means that the putative match $(\mathbf{x}_i, \mathbf{y}_i)$ has a large degree of consensus with respect to the whole putative set S_p.

Based on the support value α_i, we can finally define the global consensus distance d_{gc} as follows:

$$d_{gc}(\mathbf{x}_i, \mathbf{y}_i) = 1 - e^{-\frac{l_i^2}{\alpha_i}} \tag{14}$$

where the numerator of the exponent in the Gaussian function, i.e., l_i^2, characterizes the prior of $(\mathbf{x}_i, \mathbf{y}_i)$. Ideally, a loop-closing pair should have two same images, then we have $l_i = 0$ for a true match $(\mathbf{x}_i, \mathbf{y}_i)$, and the corresponding global consensus distance will be minimized, i.e., $d_{gc}(\mathbf{x}_i, \mathbf{y}_i) = 0$.

Considering the multi-scale strategy in (10), the local topological cost factor in (12), and the global consensus factor in (14), we obtain our final cost function as follows:

$$\begin{aligned} C(\mathcal{I}; S_p, \lambda) = \sum_{i\in\mathcal{I}} \Bigg(\sum_{m=1}^{M} \frac{1}{MK_m} \Bigg(& \sum_{j|\mathbf{x}_j \in \mathcal{N}_{\mathbf{x}_i}^{K_m}} d\left(\mathbf{y}_i, \mathbf{y}_j\right) \\ & + \sum_{j|\mathbf{x}_j \in \mathcal{N}_{\mathbf{x}_i}^{K_m}, \mathbf{y}_j \in \mathcal{N}_{\mathbf{y}_i}^{K_m}} d\left(\mathbf{m}_i, \mathbf{m}_j\right) \Bigg) + \mu \cdot d_{gc}(\mathbf{x}_i, \mathbf{y}_i) \Bigg) \\ & + \lambda(N - |\mathcal{I}|). \end{aligned} \tag{15}$$

3) A Solution with Closed Form: We associate each putative correspondence $(\mathbf{x}_i, \mathbf{y}_i)$ with a binary code p_i to indicate its correctness. Specifically, $p_i = 1$ suggests $(\mathbf{x}_i, \mathbf{y}_i)$ constitutes a pair of inliers while $p_i = 0$ means a pair of outliers. In this case, the objective function in (15) can be simplified with $|\mathcal{I}| = \sum_{i=1}^N p_i$ to

$$C(\mathbf{p}; S_p, \lambda) = \sum_{i=1}^{N} p_i \left(c_i + d_{gc}^i - \lambda \right) + \lambda N \tag{16}$$

where

$$\begin{aligned} c_i = \sum_{m=1}^{M} \frac{1}{MK_m} \Bigg(& \sum_{j|\mathbf{x}_j \in \mathcal{N}_{\mathbf{x}_i}^{K_m}} d\left(\mathbf{y}_i, \mathbf{y}_j\right) \\ & + \sum_{j|\mathbf{x}_j \in \mathcal{N}_{\mathbf{x}_i}^{K_m}, \mathbf{y}_j \in \mathcal{N}_{\mathbf{y}_i}^{K_m}} d\left(\mathbf{m}_i, \mathbf{m}_j\right) \Bigg) \end{aligned} \tag{17}$$

$$d_{gc}^i = \mu \cdot d_{gc}(\mathbf{x}_i, \mathbf{y}_i). \tag{18}$$

The final inlier set is determined by the optimal solution of $\mathbf{p}$ in (16). Intuitively, $\{c_i\}_{i=1}^N$ reflect the neighborhood structures of each correspondence, which can be calculated in

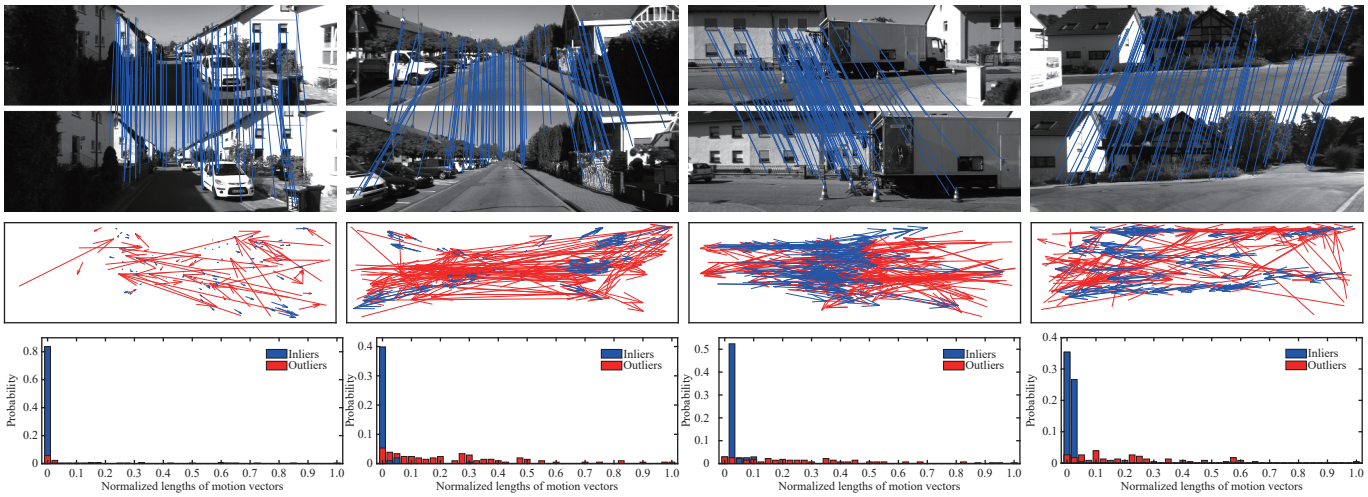


Fig. 4. Illustration of the global consensus in loop-closing pairs. Top: feature matching results of loop-closing pairs (only inliers provided). Middle: motion fields of four different scenes in the top row, where the head and tail of each vector correspond to the spatial positions of two corresponding feature points in the two images (blue: inliers; red: outliers). Bottom: distribution of the lengths of motion vectors. For each bin, we overlap the inlier and outlier probabilities, where the one with smaller probability is shown in the outer layer.

advance with regard to the known information of local neighborhood. As such, the only unknown variable in (16) is p_i. To optimize the cost function and make it minimum, p_i should be set to 1 when $c_i + d_{gc}^i > \lambda$ while 0 when $c_i + d_{gc}^i \leq \lambda$. This is equivalently a simple linear assignment problem, i.e., optimizing a truncated loss without any regularization term [57]. This kind of solver results in a closed-form solution for p_i, and we formally summarize it as

$$p_i = \begin{cases} 0, & c_i + d_{gc}^i > \lambda, \\ 1, & c_i + d_{gc}^i \leq \lambda, \end{cases} \quad i = 1, \ldots, N \tag{19}$$

which subsequently leads to the optimal inlier set of

$$\mathcal{I}^* = \{i | p_i = 1,\ i = 1, \ldots, N\}. \tag{20}$$

The pipeline of our LPM-GC is summarized in Algorithm 1.

Algorithm 1: The LPM-GC Algorithm

Input: putative set $S_p = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, parameters $\mathbf{K}$, r, μ, λ, τ
Output: inlier set $\mathcal{I}^*$
1 Normalize lengths of motion vectors to get $\{l_i\}_{i=1}^N$;
2 Conduct mean shift clustering to obtain clusters;
3 Calculate support value α_i using (13);
4 Calculate global consensus distance using (14);
5 Construct neighbors $\{\mathcal{N}_{\mathbf{x}_i}^{K_m}, \mathcal{N}_{\mathbf{y}_i}^{K_m}\}_{m=1,i=1}^{M,N}$ using S_p;
6 Calculate local topological cost using (12);
7 Calculate $\{c_i\}_{i=1}^N$ using (17);
8 Calculate $\{d_{gc}^i\}_{i=1}^N$ using (18);
9 Determine $\mathcal{I}^*$ using (19) and (20).

4) Computational Complexity Analysis: For N points, the time complexity of searching K_m nearest neighbors for each of them is about $O((K_m + N)\log N)$. Because the corresponding $K_m (m < M)$ neighborhood $\mathcal{N}_{\mathbf{x}_i}^{K_m}$ of $\mathbf{x}_i$ can be acquired directly after obtaining the K_M neighborhood, its time complexity can be noted as $O((K_M + N)\log N)$.

If ignoring d_{gc}^i, there are merely some addition operations in the calculation of the major cost c_i in (17), and the time complexity of calculating c_i is less than $O(MK_MN)$. When considering d_{gc}^i, the time requirement of the mean shift operation in it scales as $O(cN^2)$ with c denoting the number of iterations [56]. Therefore, the total time complexity of our LPM-GC is close to $O((K_M + N)\log N + MK_MN + cN^2)$, which can be naturally simplified as $O(N^2)$ on account of $MK_M \ll N$. Since we at most extract top-500 features of each image (i.e., the maximum value of N is less than 500), the time complexity of our feature matching method has a stable upper bound. Regarding the space complexity of LPM-GC, it is $O(N)$ (simplified from $O(MK_MN)$) due to the memory requirement for storing $\mathcal{N}_{\mathbf{x}_i}^{K_M}$ and $\mathcal{N}_{\mathbf{y}_i}^{K_M}$. Overall, LPM-GC has quadratic time complexity and linear space complexity in regard to the number of putative correspondences.

C. Temporal Consistency

If the cardinality of the inlier set $|\mathcal{I}^*|$ preserved by LPM-GC is sufficient enough (i.e., $|\mathcal{I}^*| > \sigma_1$), the image pair is most likely to represent the same scene. Nevertheless, it does not mean a loop-closing pair could be verified due to the existence of strong perceptual aliasing examples (i.e., several distinct places look similar). For this, only loop-closing events that persist over time are retained. This operation is termed as temporal consistency check and we follow the way in [6] to achieve it. Specifically, we use the vector $\mathbf{e}$ to record how many times each database image is marked as a loop-closing event. As such, the temporal consistency factor θ_r of the reference image I_r can be denoted as

$$\theta_r = \begin{cases} \dfrac{\mathbf{e}_r}{\sum_{j=1}^{r+f} \mathbf{e}_j}, & 0 < r \leq f \\ \dfrac{\mathbf{e}_r}{\sum_{j=r-f}^{r+f} \mathbf{e}_j}, & f < r < N_T - f \\ \dfrac{\mathbf{e}_r}{\sum_{j=r-f}^{N_T} \mathbf{e}_j}, & N_T - f \leq r \leq N_T \end{cases} \tag{21}$$

where $\mathbf{e}_r$, $\mathbf{e}_j$ respectively indicate the r-th, j-th elements of the

TABLE I
INFORMATION OF THE USED DATASETS

Dataset		Description	Image resolution	# Images	Frame rate (Hz)	Distance (km)
KITTI [58]	Seq# 00 (K00)	Outdoor, dynamic	1241×376	4541	10	3.7
	Seq# 02 (K02)			4661		5.0
	Seq# 05 (K05)		1226×370	2761		2.2
	Seq# 06 (K06)			1101		1.2
St. Lucia [59]	100909 12:10 (St1210)	Outdoor, dynamic	640×480	19251	15	~17.6
	100909 14:10 (St1410)			20894		
	180809 15:45 (St1545)			21434		
	190809 08:45 (St0845)			21815		
Malaga 2009 Parking 6L (Malaga) [60]		Outdoor, slightly dynamic	1024×768	3474	7	1.2
City Centre (CC) [23]		Outdoor, dynamic	640×480	1237	10	1.9

vector **e** (i.e., the image ID of a database image), and N_T represents the number of images acquired up to time T. A larger θ_r indicates a higher probability of the occurrence of a false-positive event. So far, the trigger of a loop-closing event is confined to two conditions: i) $|\mathcal{I}^*| > \sigma_1$ ii) $\theta_r < \sigma_2$.

IV. EXPERIMENTAL RESULTS

In this section, the performance of our method is evaluated with the state-of-the-art competitors in both feature matching and LCD tasks. Next, we first give a brief introduction of the experimental setup in our work, and then illustrate our results. We run our codes on an Intel(R) Core(TM) i9-9920X CPU @ 3.50GHZ machine with three TITAN RTX GPUs. All operations are conducted on CPU excluding the extraction of Super-features.

A. Datasets and Settings

Datasets: Ten challenging sequences are selected for evaluation, and their detailed information is shown in Table I. Due to the rich visual information including high resolution and frame rate provided by the recorded data, the scenes in KITTI sequences [58] are relatively easy to allow successful loop closure identification. In contrast, K02 is the most challenging one among the four due to the existence of dynamic cars and illumination changes. St. Lucia [59] involves long-term traversals, suffering from large illumination changes and plenty of dynamic occlusions (cars and pedestrians). The frame rates of the four sequences in it are downsampled to 3 Hz to avoid high content overlap among consecutive images. Regarding Malaga [60], images in it are collected on repetitive and very close roads, which easily leads to strong perceptual aliasing examples. It is collected by a stereo camera and only the right camera measurements are chosen for evaluation. All sequences introduced before are car-mounted trajectories, while CC [23] is a ground one collected by a mobile robot. Loop-closing pairs in it are equipped with wide baseline or large depth difference due to the variable speed of the robot. It also comprises stereo images and we adopt the streams recorded by the left camera to evaluate.

Ground Truth: In the case of LCD, ground truth (GT) is composed of a binary matrix to suggest the relationship between an image pair. Specifically, if a true loop-closing event happens between images I_i and I_j, then $\mathbf{GT}_{ij}$ is equal to 1 otherwise to 0. It is initially produced in accordance with GPS information, where image pairs captured shorter than a distance metric range of 40 meters are labelled as true loop-closing events. Nevertheless, this would lead to false labels for some image pairs, typically when two images look similar, yet the captured distance surpasses the GPS's threshold. To tackle this problem, the combination of proximity in space and equality in appearance is considered in [10] to generate refined GT. Concretely, the GT files are refined by manually checking each labeled pair and we exploit human-labeled GT matrices provided in [10] for evaluation.

Evaluation Metrics: Both feature matching and LCD tasks adopt *Precision* and *Recall* for evaluation. The former is the ratio between true positives and all system identifications, whereas the latter is that between true positives and ground truth. Additionally, $F-socre = (2 \cdot Precision \cdot Recall)/(Precision + Recall)$ is applied for comprehensive evaluation in feature matching, and the *the maximum recall rate at 100% precision* (MR) is used in LCD. In the evaluation of LCD, tuples of precision and recall are generated by the varied threshold σ_1. Meanwhile, an identification would be regarded as a true positive when it locates within 10 neighboring frames of the true loop-closing examples indicated by GT.

Implementation and Parameter Illustration: In the procedure of geometric verification, we extract top-500 SURF [61] features for each image. A thorough parameter illustration of our approach is presented in Table II. We set those of ASMK, LPM-GC (excluding r, λ) and LCD (excluding $\mathcal{T}, \sigma_1$) according to [16], [45] and [6], respectively, while that of $\mathcal{T}$ in LCD is set empirically. Later, we would vary the value of σ_1 to generate different tuples of (*Precision*,*Recall*) in the LCD task, thus we denote its value by "\" in the table.

To determine the optimal values of r and λ in LPM-GC, we create a small dataset (*LCD*) by collecting 32 true loop-closing pairs from the LCD datasets shown in Table I. Then putative sets of these image pairs are created by SIFT and ratio test (RT) with a threshold of 1.5, and true correspondences are identified by manually checking each of the putative match. Experimental results are shown in Fig. 5. When r varies from 0 to 0.04, the left plot illustrates the average clusters formed in

TABLE II
PARAMETER ILLUSTRATION

Parameter	Description	Value
ASMK	ϕ, associated with the selectivity function in (5)	3
	γ, the threshold in the selectivity functionin in (5)	0
LPM-GC	K_M, size of the neighborhood	[4, 6, 8]
	τ, threshold of local topological cost in (12)	10
	μ, weight of the global consensus distance in (15)	0.3
	r, radius in mean shift clustering	0.02
	λ, inlier threshold in (19)	0.8
LCD	$\mathcal{T}$, consuming time in non-search area	40
	σ_1, inlier threshold of loop closure events	\
	σ_2, threshold of temporal consistency	0.3

the 32 different scenes according to the length of motion vectors. We conduct a group of experiments and found $r = 0.02$ can achieve stable performance on loop-closing pairs. Thus we set $r = 0.02$ in the subsequent experiments. In this case, the number of clusters is approximately equal to 17. Afterwards, considering the objective function of our LPM-GC defined in (16), we report the distribution of c_i (i.e., the cost of neighborhood topological structure) and $c_i + d_{gc}^i$ (i.e., the cost of neighborhood topological structure with global consensus constraint) in the middle and right plots of Fig. 5. Clearly, the separability between inliers and outliers has been enhanced by introducing the global consensus constraint. According to the results in the right plot of Fig. 5, we set $\lambda = 0.8$ to discriminate inliers and outliers.

B. Results on Feature Matching

To evaluate the performance of LPM-GC on feature matching, six feature matching methods involving RT in SIFT [55], RANSAC [18], MAGSAC++ [38], LGSC [47], GMS [46] and LPM [45] are selected for comparison. For this, another dataset (*VGG*) is created in the same way where *LCD* is created. The detailed information of the two datasets is as follow:

-*VGG:* This dataset is made up of 40 image pairs collected in [62], mainly composed of situations with wide baseline or image pairs with homography transformation. We select 30 image pairs for evaluation, where the average size and inlier ratio of the putative sets attain 744 and 80.14%, respectively.

-*LCD:* This dataset is composed of 32 image pairs collected from the LCD datasets shown in Table I. Scenes in this dataset are challenging for feature matching due to the existence of similar structures and dynamic occlusions. The average size and inlier ratio of the putative sets attain 351 and 52.59%, respectively.

Intuitive Results: Fig. 6 presents intuitive feature matching results of seven methods in three common LCD situations extracted from *LCD*. The first one describes an outdoor situation with strong illumination changes and similar structures (trees), the second one includes camera distortions, repetitive patterns (windows) and a dynamic occlusion and the third one shows a simple scene occurring commonly between a loop-closing pair but also with some similar structures. The corresponding precisions, recalls and F-scores are labelled in the image pairs. From the results, we can see that since the existence of illumination changes and similar structures mentioned above, the discrimination capability of SIFT descriptors is poor, and that's why the precisions of RT are unsatisfactory. Regarding the parametric methods RANSAC and MAGSAC++, their F-scores are extremely poor in the second scene. It is because the low inlier ratio and non-rigid deformation caused by the camera distortion make them fail to estimate a good solution. It seems that the graph-based method LGSC is not good at recalls in the three scenes, which is possibly caused by its strict condition of identifying an inlier. GMS achieves feature matching based on a statistic framework. Thus to get a satisfying result, the size of the putative set should be large. In the challenging scenes, some false matches also have consistent local information, which would lead vanilla LPM to preserve them incorrectly. By introducing the global constraint in our LPM-GC, the weakness of LPM could be improved and that's why LPM-GC performs marginally better than LPM in the three scenes.

Quantitative Results: A more comprehensive quantitative evaluation result is illustrated in Fig. 7. It can be seen that the performance of RANSAC and MAGSAC++ drops dramatically in *LCD*. Compared with recall, both LGSC and GMS are better at precision. Conversely, our LPM-GC is better at recall, but it yields a notable boost in both precision and recall compared with LPM. Overall, LPM-GC has high F-scores in the two feature matching datasets, which indicates its capability of generating satisfying precision-recall trade-off.

The runtime statistics of different methods are presented in the fourth column of Fig. 7. It can be seen that RANSAC is fast when dealing with *VGG*, but becomes slow in *LCD*. MACSAC++ and LGSC are slow in both datasets. GMS spends very little time but its results on precision and recall metrics are not satisfying. Due to the requirement of mean-shift clustering, LPM-GC generally costs several milliseconds more than LPM, but it is still fast and only needs about 10 ms in average to establish reliable matches, which can satisfy the realtime requirement of LCD.

C. Analysis on Super-Features and ASMK

The response score of a Super-feature can be reflected by its L2-norm. To perform ASMK, a codebook of 65536 clusters is first learned by extracting top-1000 Super-features from each training image of SfM-120k [63] (involving totally 20000 images) at a single image scale. The testing configuration of FIRe involves restricting the input image to a maximum resolution of 1024 pixels and extracting Super-features from the image at 7 resolutions/scales $\{2.0, 1.414, 1.0, 0.707, 0.5, 0.353, 0.25\}$. This costs about 101.1 ms for a K00 instance with the resolution of 1024×311 (downsampled from 1241×376), which cannot satisfy the realtime requirement for the LCD task. So in order to guarantee both performance and efficiency, we experiment with different image pyramids on K00, St1210 and Malaga. The results are shown in Table III. It can be seen that the image pyramid of $[1.0, 0.707]$ can achieve a trade-off between MR and latency, thus we choose it to conduct the subsequent experiments.

We also want to show how Super-features achieve memory-

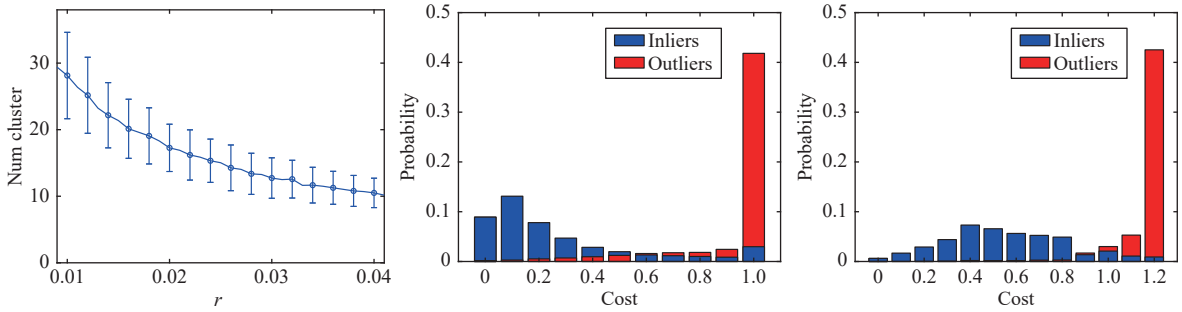


Fig. 5. Parameter analysis. Left: the change of cluster number with respect to the radius parameter r. Middle: the distribution of c_i in (16) (i.e., the cost of neighborhood topological structure). Right: the distribution of $c_i + d_{gc}^i$ in (16) (i.e., the cost of neighborhood topological structure with global consensus constraint). For each bin, we overlap the inlier and outlier probabilities, where the one with smaller probability is shown in the outer layer.

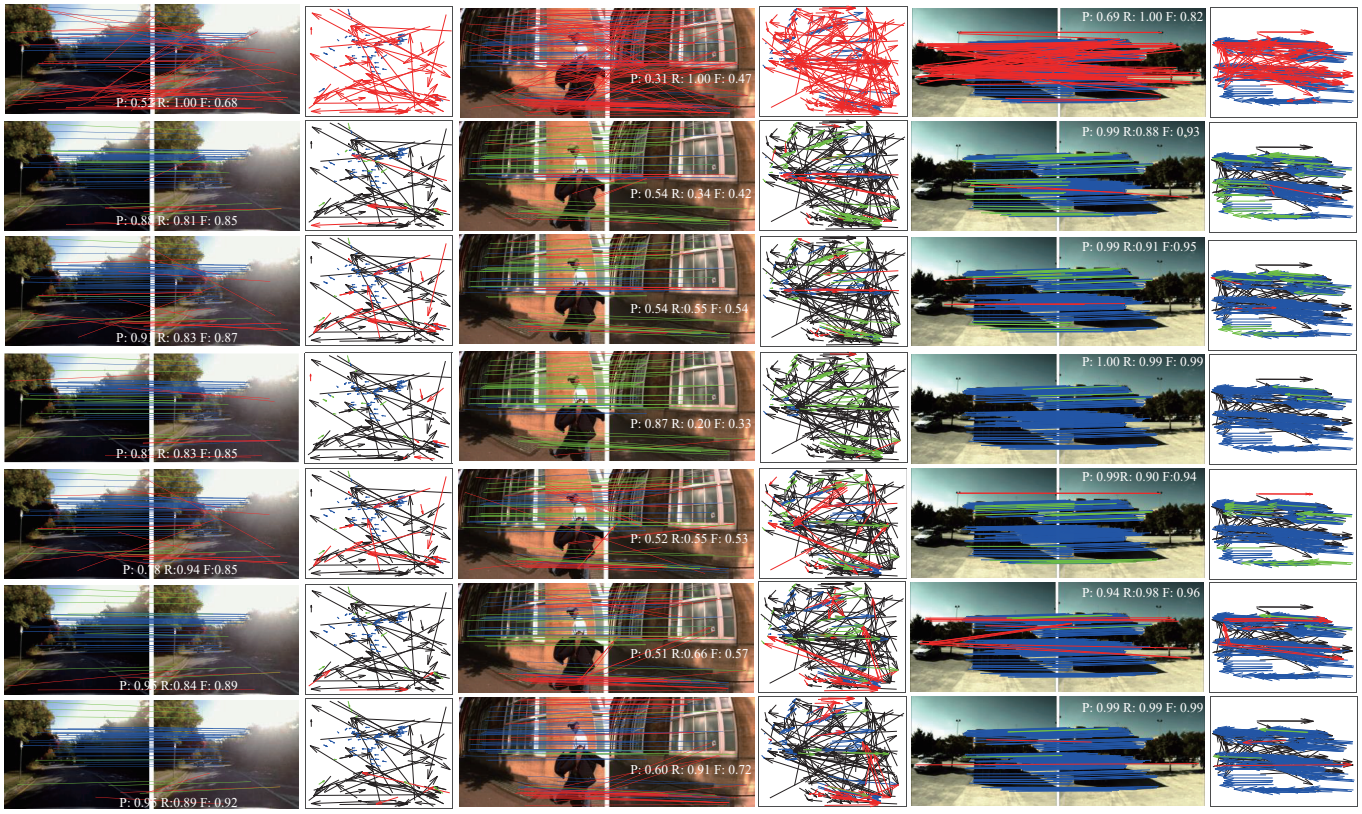


Fig. 6. From top to bottom, intuitive feature matching results of seven approaches including RT [55], RANSAC [18], MAGSAC++ [38], LGSC [43], GMS [46], LPM [45] and our LPM-GC. The inlier percentages are 52.03%, 30.77% and 69.18% in the three scenes, respectively. In each group of results, the head and tail of each arrow in the right motion field correspond to the positions of two corresponding feature points in the left image pair (black = true negative, blue = true positive, red = false positive, green = false negative). For visibility, the true negatives are not shown in the left image pairs. The corresponding precisions, recalls and F-scores are labelled in the image pairs.

wise efficiency in the LCD task. As analyzed earlier, the memory footprint of a database image I can be measured by the number of non-empty ASMK clusters in it, i.e., clusters with at least one assignment which can be denoted as $|C(I)|$. We choose the other two features, involving a handcrafted one (SURF) and a learning-based one (Key.Net [64]+Hynet [65])[3] for comparison. The codebooks in the three settings (SURF, Key.Net+Hynet, Super-features) are learned accordingly in the way described before. We vary the number of features extracted before aggregation in $\{200, 400, 600, 800, 1000\}$ and show the results in Fig. 8. Three conclusions can be drawn from the results. First, when extracting more than 600 features, Super-features at most use 228.11, 167.20 and 189.86 clusters of the codebook on K00, St1210 and Malaga, while the number of clusters used by SURF and Key.Net+Hynet scales up with that of extracted features. This can reflect the compactness of Super-features, which is beneficial to memory saving in the LCD task. Second, the updating/querying latency in ASMK is closely related to the number of used clusters. Thus using Super-features can satisfy the realtime requirement of LCD task. Third, the advantage of Super-features can be shown in more challenging scenes where

[3] https://github.com/axelBarroso/Key.Net-Pytorch

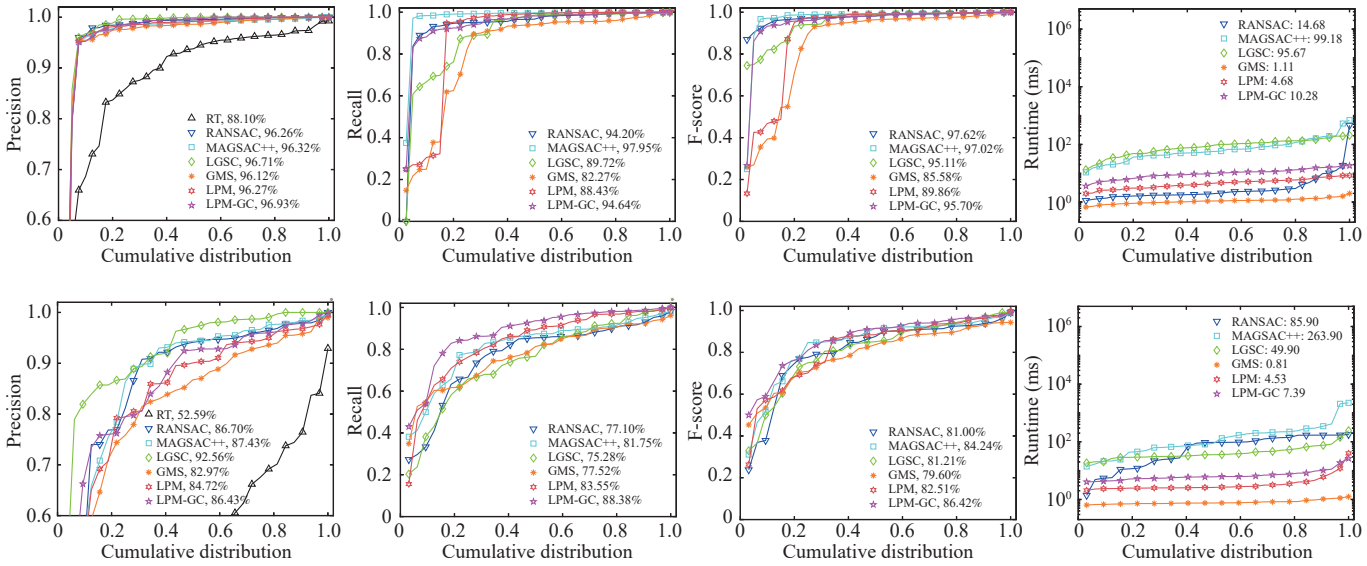


Fig. 7. Quantitative feature matching performance of LPM-GC and five competitors on *VGG* (top) and *LCD* (bottom). As the putative set is constructed on SIFT and RT, we show the curve of RT in the precision plot, which also reflects the inlier rate. The coordinate (x, y) on the curves means that there are $100 * x$ percents of image pairs which have precisions, recalls, F-scores or runtime no more than y.

TABLE III
LCD RESULTS ON DIFFERENT IMAGE PYRAMIDS

1.414	1.0	0.707	K00		St1210		Malaga	
			MR (%)	Latency (ms)	MR (%)	Latency (ms)	MR (%)	Latency (ms)
	√		94.67	13.9	89.65	12.9	65.93	23.6
	√	√	95.94	22.2	90.02	21.0	67.45	32.7
√	√		96.70	30.5	89.70	27.9	61.77	59.1
√		√	96.07	29.5	90.12	25.9	65.51	50.9
√	√	√	95.94	38.5	89.05	34.9	66.62	71.1

dynamic occlusions (St1210) or similar structures (Malaga) appear.

D. Results on Loop Closure Detection

Evaluation with Varied Feature Matching Methods: Different feature matching methods including RANSAC, LPM and LPM-GC are exploited to conduct geometric confirmation to demonstrate the superiority of LPM-GC in the LCD task. All evaluations are carried out with 600 Super-features extracted, and the results are shown in Table IV. It can be seen that RANSAC performs consistently worse than the other two approaches on all datasets excluding K06, meanwhile with more running time cost. As for LPM, it performs marginally worse than LPM-GC on all datasets excluding K00 and St1545. The performance gap attains largest on CC. Overall, through using LPM-GC, we can significantly improve MR with an extremely low cost of efficiency.

Efficiency Analysis of The LCD Pipeline: Before, only one candidate frame ($k = 1$) is searched for each query image to perform geometric verification. This procedure only costs about 9.0 ms via LPM-GC, thus we want to continue to explore how the performance would change when we increase the number of candidate frames. Table V shows the LCD results of $k = 1, 3, 5$. It can be concluded that MR generally tends to increase as the number of selected candidate frames grows. But a larger k also leads to lower efficiency. We observe that executing LPM-GC three times only costs 26.5 ms, meanwhile leading to satisfying performance on MR. So afterwards we would only discuss under the case of $k = 3$ the efficiency of our LCD system together with comparative results with other state-of-the-art LCD approaches.

The runtime plots of each component of our LCD system on K00 and Malaga datasets are presented in Fig. 9. Specifically, "Extracting" involves extracting Super-features of the query image and the image used for updating the inverted file, thus it is actually performed twice. Similarly, "Matching" involves extracting SURF features for two images as well as performing LPM-GC three times. In addition to the fact that the querying time grows as the database expands, the executive time of other parts remains stable. Overall, totally 141.8 ms and 160.4 ms would be cost when performing LCD on K00 and Malaga, respectively.

Failure Cases: The navigation paths of four sequences involving K00, St1410, Malaga and CC, together with the respective identified loop closure events of them, are depicted in Fig. 10. It can be observed that false-negative events would unavoidably occur when the appearance changes significantly. For example, instances captured at the same crossroad with

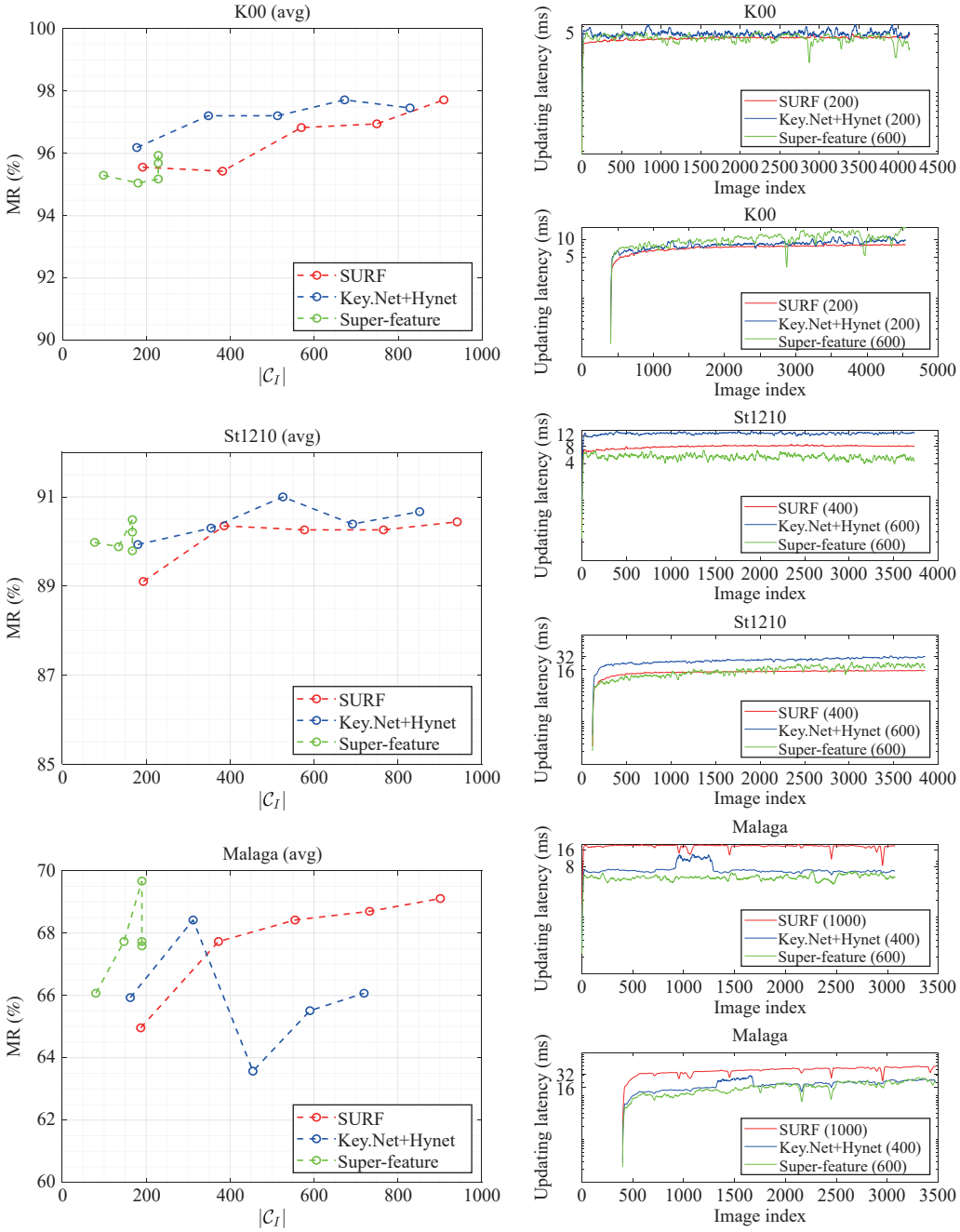


Fig. 8. Performance versus memory footprint and executive latency for SURF, Key.Net+Hynet and Super-feature. Left: $|\mathcal{C}_I|$ shows the average number of clusters per image used in ASMK (proportionate to memory usage). Right: For Super-feature, we show its executive time of extracting 600 features. For the other two, we show their executive time of the cases where i) better performance than Super-feature-600 appears with relatively few features extracted, or ii) the best performance of itself is achieved if no result is better than Super-feature-600. Relevant information has been labelled in the legend.

TABLE IV

MR (%) WHEN USING DIFFERENT FEATURE MATCHING METHODS FOR LOOP CLOSURE IDENTIFICATION. "MEAN LATENCY" INDICATES THE AVERAGE RUNNING TIME OF ONE MISMATCH REMOVAL. THE BEST RESULTS ARE LABELLED IN BOLD

	K00	K02	K05	K06	St1210	St1410	St1545	St0845	Malaga	CC	Mean latency (ms)
RANSAC	94.54	70.72	92.53	**98.53**	87.33	67.38	80.85	85.67	55.96	68.70	165.6
LPM	96.07	77.26	94.94	**98.53**	89.93	76.66	**87.81**	89.97	66.76	84.58	**4.6**
LPM-GC	**96.19**	**78.19**	**95.18**	**98.53**	**90.49**	**78.34**	86.50	**91.00**	**67.73**	**87.21**	9.0

TABLE V

MR (%) WHEN DIFFERENT NUMBER OF CANDIDATE FRAMES IS EXPLOITED FOR LOOP CLOSURE IDENTIFICATION. "MEAN LATENCY" SHOWS THE AVERAGE RUNNING TIME FOR EXECUTING LPM-GC ONCE, THREE TIMES AND FIVE TIMES, RESPECTIVELY

	K00	K02	K05	K06	St1210	St1410	St1545	St0845	Malaga	CC	Mean latency (ms)
k = 1	96.19	78.19	95.18	98.53	90.49	78.34	86.50	91.00	67.73	87.21	9.0
k = 3	97.46	80.01	95.57	98.90	90.21	80.63	88.90	93.21	68.42	88.45	26.5
k = 5	97.21	81.29	95.89	98.90	90.44	81.98	89.47	93.28	68.70	88.21	45.3

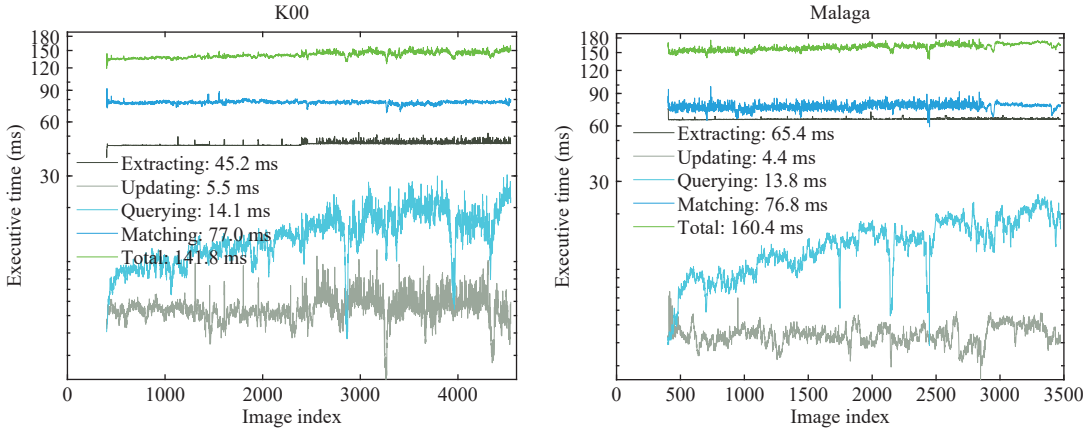


Fig. 9. Executive time per image for each main processing step of LCD system, measured on datasets with relatively large resolutions.

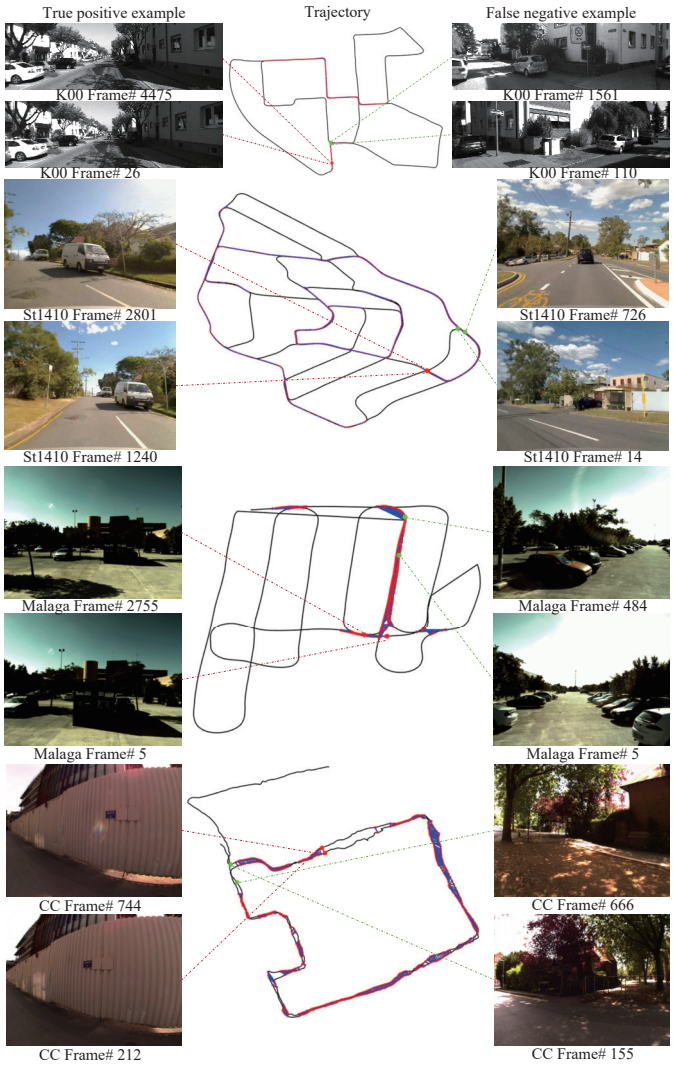


Fig. 10. Illustration of the LCD task. From left to right: true positive, robots' trajectory, and false positive. From top to bottom: K00, St1410, Malaga, and CC. The black trajectory is acquired from GPS logs, and the results of loop closure identification are obtained under the maximum recall rate at 100% precision. We label the loop closure pair as red hollow points, along with connecting them via a blue line. The red and green solid points are the specific illustrations of true-positive and false-negative detections, respectively.

different orientations, or scenes dominated by dynamic occlusions, may both lead to failure of loop closure identification.

Comparative Results: Table VI shows the comparative results of our approach against the state-of-the-art LCD techniques. Specifically, the reproducing results of [10], [28], [31], [66] are cited from [10]; those of [8] are cited from [8]; those of [27] are reproduced by ourselves on the basis of available codes. It can be seen that our approach can yield the best results on all datasets expect for CC. The performance gaps among different methods on KITTI sequences are relatively smaller than those on St. Lucia sequences. It is due to the fact that instances in KITTI datasets comprise abundant prominent features, which makes it relatively easy for approaches to identify true loop-closing pairs. However, when scenes turn more challenging in St. Lucia datasets, the superiority of our feature-matching-based LCD approach comes into being. The remarkable performance on St. Lucia datasets also hints that our method has the potential to be put to long-term driving tasks. As for Malaga, nearly all approaches yield poor results on it by virtue of extremely high similarity of scenes. Regarding CC, we perform marginally worse than FILD++ and ESA-VLAD, but still achieve a satisfying result.

V. Conclusion

In this paper, an effective method for the LCD task is introduced. We first perform candidate frame selection based on Super-features as well as an incremental ASMK strategy, and then achieve geometrical verification via a proposed feature matching technique termed as LPM-GC. From the experimental results we found that the combination of Super-features and ASMK can achieve both computational and memory-wise efficiency in the LCD task. As for LPM-GC, it focuses on recognizing reliable matches between image pairs on the basis of retaining local neighborhood information and global spatial consensus among the feature points, and hence it can deal with complex scenes in LCD such as image pairs with many repetitive patterns or large mismatch proportions. More importantly, our LPM-GC has a closed-form solution which can fulfill the verification of one image pair in only a few milliseconds, enabling the whole system to operate more efficiently. Through experiments, it is concluded that our feature-matching-based LCD pipeline has great potential in the long-

TABLE VI
COMPARATIVE RESULTS OF THE MAXIMUM RECALL RATE AT 100% PRECISION. ROLD INDICATES THE BEST

	K00	K02	K05	K06	St1210	St1410	St1545	St0845	Malaga	CC
PREVIeW [31]	89.47	71.96	87.71	80.15	60.93	23.06	49.79	56.69	33.93	49.63
iBoW-LCD [28]	76.50	72.22	53.07	95.53	70.02	68.06	87.50	59.36	57.48	82.03
Kazmi and Mertsching [66]	90.39	79.49	81.41	97.39	80.06	58.10	72.55	80.13	50.98	75.58
Tsintotas *et al.* [27]	93.78	78.50	94.46	96.69	76.52	51.79	65.11	78.13	52.22	36.30
FILD++ [10]	94.92	73.52	95.42	98.16	83.39	66.41	81.36	87.86	62.74	90.01
ESA-VLAD [8]	**97.46**	79.94	\	**98.90**	89.14	79.59	88.47	92.94	\	**90.31**
Ours	**97.46**	**80.01**	**95.57**	**98.90**	**90.21**	**80.63**	**88.90**	**93.21**	**68.42**	88.45

term SLAM system.

APPENDIX
DERIVATION OF EQUATION (10)

According to (8), the square items in (9) can be simplified and the cost function can be rewritten as

$$C=\sum_{i\in\mathcal{I}}\sum_{m=1}^{M}\frac{1}{2MK_m}\left(\sum_{j|\mathbf{x}_j\in\mathcal{N}_{\mathbf{x}_i}^{Km}} d\left(\mathbf{y}_i,\mathbf{y}_j\right)+\sum_{j|\mathbf{y}_j\in\mathcal{N}_{\mathbf{y}_i}^{Km}} d\left(\mathbf{x}_i,\mathbf{x}_j\right)\right)+\lambda(N-|\mathcal{I}|). \tag{22}$$

The first item $\sum_{j|\mathbf{x}_j\in\mathcal{N}_{\mathbf{x}_i}^{Km}} d\left(\mathbf{y}_i,\mathbf{y}_j\right)$ can be calculated by

$$\begin{aligned}\sum_{j|\mathbf{x}_j\in\mathcal{N}_{\mathbf{x}_i}^{Km}} d\left(\mathbf{y}_i,\mathbf{y}_j\right)&=\sum_{j|\mathbf{x}_j\in\mathcal{N}_{\mathbf{x}_i}^{Km},\mathbf{y}_j\in\mathcal{N}_{\mathbf{y}_i}^{Km}} d\left(\mathbf{y}_i,\mathbf{y}_j\right)+\sum_{j|\mathbf{x}_j\in\mathcal{N}_{\mathbf{x}_i}^{Km},\mathbf{y}_j\notin\mathcal{N}_{\mathbf{y}_i}^{Km}} d\left(\mathbf{y}_i,\mathbf{y}_j\right)\\&=0+count\left(j|\mathbf{x}_j\in\mathcal{N}_{\mathbf{x}_i}^{Km},\mathbf{y}_j\notin\mathcal{N}_{\mathbf{y}_i}^{Km}\right)\\&=K_m-count\left(j|\mathbf{x}_j\in\mathcal{N}_{\mathbf{x}_i}^{Km},\mathbf{y}_j\in\mathcal{N}_{\mathbf{y}_i}^{Km}\right)\\&=K_m-n_i^{se}\end{aligned} \tag{23}$$

where $count(\cdot)$ functions as the calculation of the set cardinality, while n_i^{se} denotes the number of common members in $\mathcal{N}_{\mathbf{x}_i}^{Km}$ and $\mathcal{N}_{\mathbf{y}_i}^{Km}$. Likewise, it can be proved that $\sum_{j|\mathbf{y}_j\in\mathcal{N}_{\mathbf{y}_i}^{Km}} d(\mathbf{x}_i,\mathbf{x}_j)=K_m-n_i^{se}=\sum_{j|\mathbf{x}_j\in\mathcal{N}_{\mathbf{x}_i}^{Km}} d\left(\mathbf{y}_i,\mathbf{y}_j\right)$. According to the above two equations, the cost function in (9) can be converted into (10) straightforwardly.

REFERENCES

[1] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Trans. Robot.*, vol. 32, no. 6, pp. 1309–1332, 2016.

[2] W. Huang, G. Zhang, and X. Han, "Dense mapping from an accurate tracking SLAM," *IEEE/CAA J. Autom. Sinica*, vol. 7, no. 6, pp. 1565–1574, 2020.

[3] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Trans. Robot.*, vol. 32, no. 1, pp. 1–19, 2015.

[4] D. Gálvez-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, 2012.

[5] L. Bampis, A. Amanatiadis, and A. Gasteratos, "Encoding the description of image sequences: A two-layered pipeline for loop closure detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2016, pp. 4530–4536.

[6] K. Zhang, X. Jiang, and J. Ma, "Appearance-based loop closure detection via locality-driven accurate motion field learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 2350–2365, 2022.

[7] 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.

[8] Y. Xu, J. Huang, J. Wang, Y. Wang, H. Qin, and K. Nan, "ESA-VLAD: a lightweight network based on second-order attention and netvlad for loop closure detection," *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 6545–6552, 2021.

[9] S. An, G. Che, F. Zhou, X. Liu, X. Ma, and Y. Chen, "Fast and incremental loop closure detection using proximity graphs," arXiv preprint arXiv: 1911.10752, 2019.

[10] S. An, H. Zhu, D. Wei, K. A. Tsintotas, and A. Gasteratos, "Fast and incremental loop closure detection with deep features and proximity graphs," *J. Field Robot.*, vol. 39, no. 4, pp. 473–493, 2022.

[11] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 1–8.

[12] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, 1967, pp. 281–297.

[13] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[14] H. Jégou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *Int. J. Comput. Vis.*, vol. 87, no. 3, pp. 316–336, 2010.

[15] R. Ji, L.-Y. Duan, J. Chen, H. Yao, J. Yuan, Y. Rui, and W. Gao, "Location discriminative vocabulary coding for mobile landmark search," *Int. J. Comput. Vis.*, vol. 96, no. 3, pp. 290–314, 2012.

[16] G. Tolias, Y. Avrithis, and H. Jégou, "Image search with selective match kernels: Aggregation across single and multiple images," *Int. J. Comput. Vis.*, vol. 116, no. 3, pp. 247–261, 2016.

[17] A. M. Andrew, "Multiple view geometry in computer vision," *Kybernetes*, 2001.

[18] 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.

[19] J. Ma, J. Wu, J. Zhao, J. Jiang, H. Zhou, and Q. Z. Sheng, "Nonrigid point set registration with robust transformation learning under manifold

regularization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 12, pp. 3584–3597, 2019.

[20] P. Weinzaepfel, T. Lucas, D. Larlus, and Y. Kalantidis, "Learning super-features for image retrieval," in *Proc. Int. Conf. Learn Representations*, 2022.

[21] J. Ma, X. Jiang, A. Fan, J. Jiang, and J. Yan, "Image matching from handcrafted to deep features: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 1, pp. 23–79, 2021.

[22] M. Cummins and P. Newman, "Appearance-only SLAM at large scale with fab-map 2.0," *Int. J. Rob. Res.*, vol. 30, no. 9, pp. 1100–1123, 2011.

[23] M. Cummins and P. Newman, "Fab-map: Probabilistic localization and mapping in the space of appearance," *Int. J. Rob. Res.*, vol. 27, no. 6, pp. 647–665, 2008.

[24] R. Mur-Artal and J. D. Tardós, "Fast relocalisation and loop closing in keyframe-based SLAM," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2014, pp. 846–853.

[25] E. S. Stumm, C. Mei, and S. Lacroix, "Building location models for visual place recognition," *Int. J. Rob. Res.*, vol. 35, no. 4, pp. 334–356, 2016.

[26] A. Angeli, D. Filliat, S. Doncieux, and J.-A. Meyer, "A fast and incremental method for loop-closure detection using bags of visual words," *IEEE Trans. Robot.*, vol. 24, no. 5, pp. 1027–1037, 2008.

[27] K. A. Tsintotas, L. Bampis, and A. Gasteratos, "Modest-vocabulary loop-closure detection with incremental bag of tracked words," *Robot. Auto. Syst.*, vol. 141, p. 103782, 2021. DOI: 10.1016/j.robot.2021.103782.

[28] E. Garcia-Fidalgo and A. Ortiz, "iBoW-LCD: An appearance-based loop-closure detection approach using incremental bags of binary words," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 3051–3057, 2018.

[29] K. A. Tsintotas, L. Bampis, and A. Gasteratos, "Probabilistic appearance-based place recognition through bag of tracked words," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 1737–1744, 2019.

[30] K. A. Tsintotas, L. Bampis, and A. Gasteratos, "Assigning visual words to places for loop closure detection," in *Proc. IEEE Int. Conf. Robot Automat.*, 2018, pp. 1–7.

[31] L. Bampis, A. Amanatiadis, and A. Gasteratos, "Fast loop-closure detection using visual-word-vectors from image sequences," *Int. J. Rob. Res.*, vol. 37, no. 1, pp. 62–82, 2018.

[32] Y. Xia, J. Li, L. Qi, and H. Fan, "Loop closure detection for visual SLAM using pcanet features," in *Proc. Int. Joint Conf. Neural Netw.*, 2016, pp. 2274–2281.

[33] J. Ma, S. Wang, K. Zhang, Z. He, J. Huang, and X. Mei, "Fast and robust loop-closure detection via convolutional auto-encoder and motion consensus," *IEEE Trans. Ind. Informat.*, vol. 18, no. 6, pp. 3681–3691, 2022.

[34] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, "SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 7, pp. 1200–1217, 2022.

[35] O. Chum, J. Matas, and J. Kittler, "Locally optimized RANSAC," in *Proc. Joint Pattern Recognit. Symp.*, 2003, pp. 236–243.

[36] R. Raguram, O. Chum, M. Pollefeys, J. Matas, and J.-M. Frahm, "USAC: A universal framework for random sample consensus," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 2022–2038, 2012.

[37] D. Barath, J. Matas, and J. Noskova, "MAGSAC: Marginalizing sample consensus," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10197–10205.

[38] D. Barath, J. Noskova, M. Ivashechkin, and J. Matas, "Magsac++, a fast, reliable and accurate robust estimator," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1304–1312.

[39] X. Li and Z. Hu, "Rejecting mismatches by correspondence function," *Int. J. Comput. Vis.*, vol. 89, no. 1, pp. 1–17, 2010.

[40] J. Ma, J. Zhao, J. Tian, A. L. Yuille, and Z. Tu, "Robust point matching via vector field consensus," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1706–1721, 2014.

[41] H. Chen, X. Zhang, S. Du, Z. Wu, and N. Zheng, "A correntropy-based affine iterative closest point algorithm for robust point set registration," *IEEE/CAA J. Autom. Sinica*, vol. 6, no. 4, pp. 981–991, 2019.

[42] C. Leng, H. Zhang, G. Cai, Z. Chen, and A. Basu, "Total variation constrained non-negative matrix factorization for medical image registration," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 5, pp. 1025–1037, 2021.

[43] X. Jiang, Y. Xia, X.-P. Zhang, and J. Ma, "Robust image matching via local graph structure consensus," *Pattern Recognit.*, p. 108588, 2022.

[44] J. Ma, A. Fan, X. Jiang, and G. Xiao, "Feature matching via motion-consistency driven probabilistic graphical model," *Int. J. Comput. Vis.*, 2022.

[45] J. Ma, J. Zhao, J. Jiang, H. Zhou, and X. Guo, "Locality preserving matching," *Int. J. Comput. Vis.*, vol. 127, no. 5, pp. 512–531, 2019.

[46] J. Bian, W.-Y. Lin, Y. Matsushita, S.-K. Yeung, T.-D. Nguyen, and M.-M. Cheng, "GMS: Grid-based motion statistics for fast, ultra-robust feature correspondence," *Int. J. Comput. Vis.*, vol. 128, no. 6, pp. 1580–1593, 2020.

[47] X. Jiang, J. Ma, J. Jiang, and X. Guo, "Robust feature matching using spatial clustering with heavy outliers," *IEEE Trans. Image Process.*, vol. 29, pp. 736–746, 2020.

[48] J. Ma, Z. Li, K. Zhang, Z. Shao, and G. Xiao, "Robust feature matching via neighborhood manifold representation consensus," *ISPRS J. Photogramm. Remote Sens.*, vol. 183, pp. 196–209, 2022.

[49] K. Zhang, J. Ma, and J. Jiang, "Loop closure detection with reweighting netvlad and local motion and structure consensus," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 6, pp. 1087–1090, 2022.

[50] K. M. Yi, E. Trulls, Y. Ono, V. Lepetit, M. Salzmann, and P. Fua, "Learning to find good correspondences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2666–2674.

[51] J. Ma, X. Jiang, J. Jiang, J. Zhao, and X. Guo, "LMR: Learning a two-class classifier for mismatch removal," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 4045–4059, 2019.

[52] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4938–4947.

[53] G. Tolias, T. Jenicek, and O. Chum, "Learning and aggregating deep local descriptors for instance-level recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 460–477.

[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–11.

[55] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[56] Y. Cheng, "Mean shift, mode seeking, and clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 8, pp. 790–799, 1995.

[57] A. Fan, J. Ma, X. Jiang, and H. Ling, "Efficient deterministic search with robust loss functions for geometric model fitting," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[58] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *Int. J. Rob. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.

[59] A. J. Glover, W. P. Maddern, M. J. Milford, and G. F. Wyeth, "FAB-MAP + RatSLAM: Appearance-based slam for multiple times of day," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2010, pp. 3507–3512.

[60] J.-L. Blanco, F.-A. Moreno, and J. Gonzalez, "A collection of outdoor robotic datasets with centimeter-accuracy ground truth," *Auton. Robots*, vol. 27, no. 4, p. 327, 2009. DOI: 10.1007/s10514-009-9138-7.

[61] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features, " in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 404–417.

[62] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F.

Schaffalitzky, T. Kadir, and L. Van Gool, "A comparison of affine region detectors," *Int. J. Comput. Vis.*, vol. 65, no. 1-2, pp. 43–72, 2005.

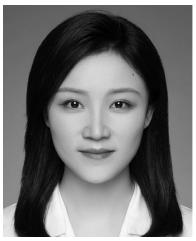
[63] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1655–1668, 2018.

[64] A. B. Laguna and K. Mikolajczyk, "Key.Net: Keypoint detection by handcrafted and learned cnn filters revisited," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022.

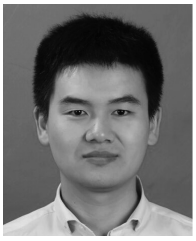
[65] Y. Tian, A. Barroso Laguna, T. Ng, V. Balntas, and K. Mikolajczyk, "HyNet: Learning local descriptor with hybrid similarity measure and triplet loss," in *Adv. Neural Inf. Process. Syst.*, 2020, pp. 7401–7412.

[66] S. A. M. Kazmi and B. Mertsching, "Detecting the expectancy of a place using nearby context for appearance-based mapping," *IEEE Trans. Robot.*, vol. 35, no. 6, pp. 1352–1366, 2019.

Jiayi Ma (Senior Member, IEEE) received the B.S. degree in information and computing science and the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology, in 2008 and 2014, respectively. He is currently a Professor with the Electronic Information School, Wuhan University. He has authored or co-authored more than 200 refereed journal and conference papers, including IEEE TPAMI/TIP, IJCV, CVPR, ICCV, ECCV, etc. His research interests include computer vision, machine learning, and pattern recognition. Dr. Ma has been identified in the 2019–2021 Highly Cited Researcher lists from the Web of Science Group. He is an Area Editor of *Information Fusion*, and an Editorial Board Member of *Neurocomputing*.



Kaining Zhang received the B.S. degree from the Electronic Information School, Wuhan University, in 2019. She is currently pursuing the Ph.D. degree at the Multi-Spectral Vision Processing Lab of Wuhan University. Her current research interests include robotics, machine learning and image processing.



Junjun Jiang (Senior Member, IEEE) received the B.S. degree from the Department of Mathematics, Huaqiao University, in 2009, and the Ph.D. degree from the School of Computer, Wuhan University, in 2014. He is currently a Professor with the School of Computer Science and Technology, Harbin Institute of Technology. He won the Finalist of the World's FIRST 10K Best Paper Award at ICME 2017, and the Best Student Paper Runner-up Award at MMM 2015. He received the 2016 China Computer Federation (CCF) Outstanding Doctoral Dissertation Award and 2015 ACM Wuhan Doctoral Dissertation Award. His research interests include image processing and computer vision.