Loop-Closure Detection Using Local Relative Orientation Matching

Jiayi Ma, *Member, IEEE*, Xinyu Ye, Huabing Zhou, *Member, IEEE*, Xiaoguang Mei, *Member, IEEE*, and Fan Fan

***Abstract*— Loop-closure detection (LCD), which aims to recognize a previously visited location, is a crucial component of the simultaneous localization and mapping system. In this paper, a novel appearance-based LCD method is presented. In particular, we propose a simple yet surprisingly useful feature matching algorithm for real-time geometrical verification of candidate loop-closures, termed as *local relative orientation* matching (LRO). It aims to efficiently establish reliable feature correspondences based on preserving local topological structures between the query image and candidate frame. To effectively retrieve candidate loop closures, we introduce the aggregated selective match kernel framework into the LCD task, which can effectively represent images and reduce the quantization noise of the traditional bag-of-words framework. In addition, the SuperPoint neural network is employed to extract reliable interest points and feature descriptors. Extensive experimental results demonstrate that our LRO can significantly improve the LCD performance, and the proposed overall LCD method can achieve much better performance over the current state-of-the-art on six publicly available datasets.**

***Index Terms*— Loop-closure detection, SLAM, feature matching, place recognition, ASMK.**

I. Introduction

RECENT developments in the field of autonomous mobile robotics and intelligent transportation systems have led to an increasing interest in simultaneous localization and mapping (SLAM) [1]–[5], which is used to construct a reliable map of the unknown environment at the same time accurately estimate the location of the robot. However, trajectories estimated by the robot would inevitably appear some accumulated drift over time due to imprecise sensor measures or environmental conditions. As one of the most important modules of the SLAM system, loop-closure detection (LCD), also known as place recognition [6], aims to amend the accumulated error by identifying if the robot has returned to a previously visited place. In general, a robust LCD method can provide precise pose estimation and improve the accuracy of the whole SLAM system [7]. In the last decades, motivated by the rich visual information, camera sensors have become the primary device on mobile robots [8]. Therefore, the LCD is achieved by distinguishing whether the current captured image has been taken from a pre-visited place, which is also termed as appearance-based LCD.

The appearance-based LCD problem can be turned into an on-line image retrieval task. In this case, the current input image is regarded as the query image, while pre-visited images are considered as database images. Therefore, the key factors influencing the performance of an appearance-based LCD method include three aspects: image feature extraction, candidate frame selection, and loop closing pair verification. For feature extraction and candidate frame selection, most LCD methods [9], [10] employ the bag-of-words (BoW) framework to quantize the descriptor space into visual words, after generating descriptors of local features like scale-invariant feature transform (SIFT) [11], speeded-up robust features (SURF) [12], oriented FAST and rotated BRIEF (ORB) [13], or local difference binary (LDB) [14]. By applying the term frequency-inverse document frequency technique, the BoW model can use a compact vector to represent an image. Then BoW combines the inverted index method to quickly calculate the similarity between the current query image and previous images, thus finds the potential loop closing pairs, *i.e.*, candidate frames. For loop closing pair verification, both time consistency verification and spatial consistency verification can be adopted. In particular, random sample consensus (RANSAC) [15] is a feature matching algorithm following a hypothesized-and-verify strategy, which has been widely utilized in LCD methods as geometrical consistency checking.

Although the existing methods are usually effective, several challenges for appearance-based LCD methods still exist. Firstly, environmental change, such as illumination change, viewpoint change, or dynamic object occlusion, makes it difficult to identify pre-visited places. In this case, the repeatability and discriminability of features are essential. Secondly, because the conventional BoW framework discards the spatial information between visual words and suffers from the quantization error problem [16], it is sensitive to the perceptual

Manuscript received 29 June 2020; revised 31 December 2020 and 20 February 2021; accepted 17 April 2021. Date of publication 4 May 2021; date of current version 8 July 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62003247, Grant 61773295, and Grant 61903279; in part by the Key Research and Development Program of Hubei Province under Grant 2020BAB113; and in part by the Natural Science Fund of Hubei Province under Grant 2019CFA037. The Associate Editor for this article was H. Jula. *(Corresponding author: Fan Fan.)*

Jiayi Ma, Xiaoguang Mei, and Fan Fan are with the Electronic Information School, Wuhan University, Wuhan 430072, China (e-mail: jyma2010@gmail.com; meixiaoguang@gmail.com; fanfan@whu.edu.cn).

Xinyu Ye is with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: xinyu_ye@sjtu.edu.cn).

Huabing Zhou is with the Hubei Key Laboratory of Intelligent Robot, Wuhan Institute of Technology, Wuhan 430073, China (e-mail: zhouhuabing@gmail.com).

Digital Object Identifier 10.1109/TITS.2021.3074520

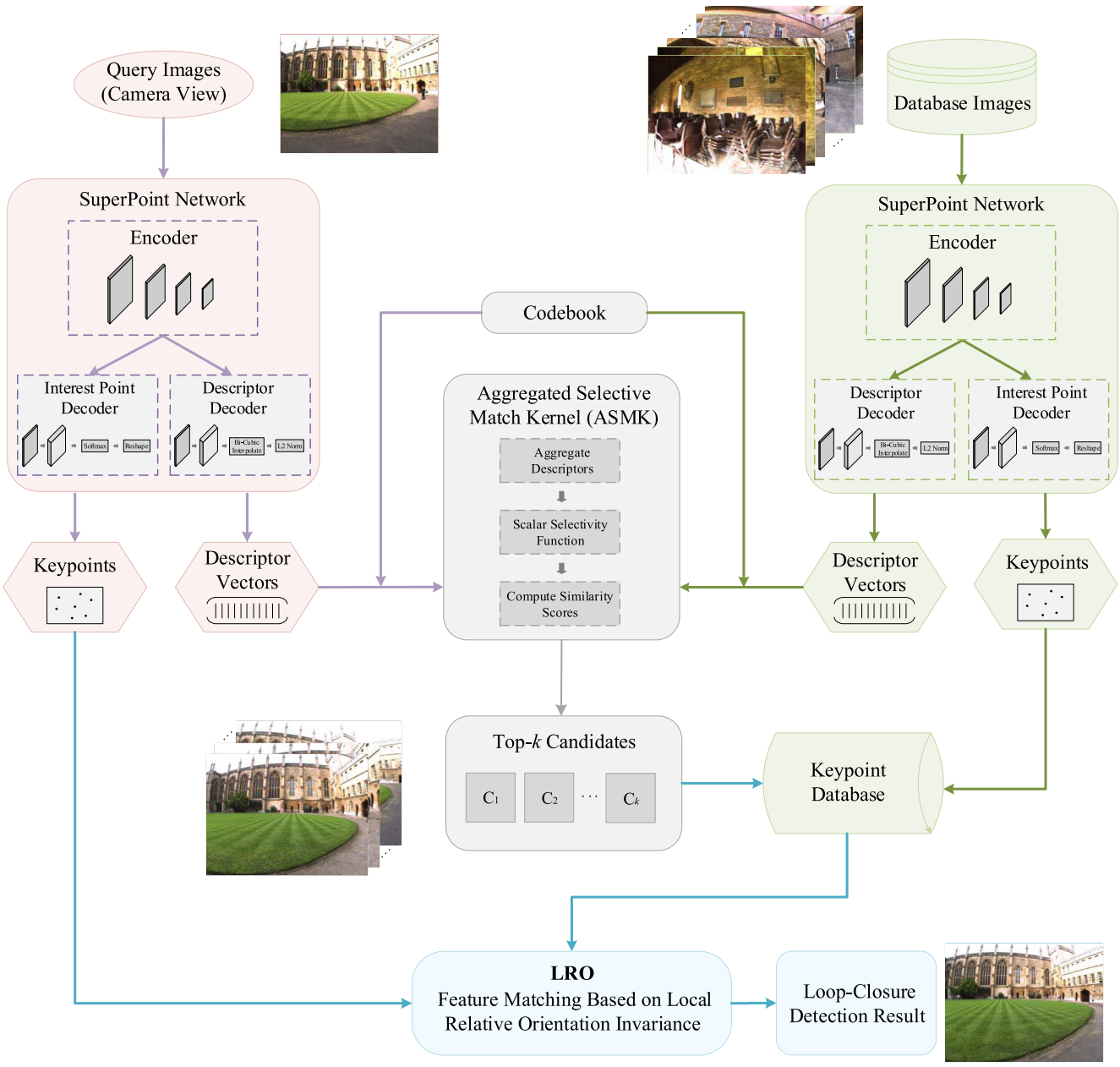

Fig. 1. An overview of the proposed LCD framework.

aliasing phenomenon, which is a kind of scene with plenty of similar objects and common features. Thirdly, the candidate loop closing pairs may involve complex scenarios, such as scenes with many repetitive patterns and high outlier rates, which would lead to false matches between image pairs and fail to verify loop closing pairs. In particular, the frequently-used RANSAC in LCD relies on predefined parametric models thus cannot handle the non-rigid transformation and other complex transformations between image pairs. Moreover, its runtime grows exponentially with the increase of outlier rate.

To cope with the above three challenges, in this paper, we propose a novel and effective appearance-based LCD approach, which follows a two-stage strategy: first selects candidate frames for query place and then verifies the candidate places via robust feature matching, as illustrated in Fig. 1. In order to obtain reliable candidate frames, we employ a deep neural network named SuperPoint [17] to extract robust local features, which can capture higher-level semantic information and is more robust against appearance change than hand-crafted descriptors. Subsequently, a retrieval framework named aggregated selective match kernel (ASMK) [18] is used to replace the traditional BoW model for candidate frame selection. ASMK can alleviate the quantization noise problem meanwhile economize memory space by binarizing the visual word. After the above first stage, the places similar in appearance to the query place will compete as candidates due to ignoring the spatial geometric relationship between local features. To this end, in the second stage we

present an efficient and effective feature matching algorithm to check the geometric relationship between the candidate and the query image, especially for challenging scenes in LCD tasks. Extensive experiments on six publicly available datasets demonstrate that our method can significantly improve the LCD performance and outperforms the current state-of-the-art.

The primary contributions of this paper include the following aspects. On the one hand, we present a new feature matching algorithm termed as *local relative orientation* matching (LRO) for real-time geometrical verification. LRO sufficiently explores the geometric relationship by preserving the neighborhood topological structure, which converts neighborhood topology invariance to relative orientation invariance and utilizes the property of affine-invariant to enhance robustness, so that it is robust against complex scenarios, such as repetitive patterns or high outlier rate, which often occur in the LCD task. On the other hand, to reduce the quantization noise of the BoW framework, for the first time we introduce the ASMK framework [18] into the LCD task. ASMK combines an aggregation procedure with a selectivity function to generate image representations, which can obtain more precise similarity scores between images while enjoying savings in memory and computation requirements. In addition, we employ the SuperPoint neural network [17] to extract feature points and generate local descriptors, and demonstrate its superiority over the traditional corner detectors and descriptors. It is capable of promoting the discriminability and repeatability of features against different illumination conditions and viewpoints.

The rest of the paper is structured as follows. In Section II, we briefly survey relevant work in the field of LCD. Section III introduces our proposed method and explicitly describes the feature extraction method and retrieval framework. Subsequently, in Section IV, we describe how to verify loop closing pairs using our proposed feature matching algorithm in detail. Section V presents thorough experiments compared with the state-of-the-art, followed by conclusion remarks in Section VI.

II. Related Work

A robust LCD algorithm is typically required to be able to identify the same place undergoing viewpoint change, illumination change, dynamic object, perceptual aliasing, or involving many complex structures. In order to achieve this goal, current related works mainly improve the LCD performance from the following three aspects.

A. Feature Extraction

Reliable and robust image representation is vital for LCD methods to address complex and changeable environments. Generally speaking, features can be divided into two categories: global features and local features. The global descriptor treats the entire image as a single descriptor. It is computationally efficient but sensitive to environmental changes such as viewpoint changes and partial occlusion. Gist [19], histogram of oriented gradient [20], and color histograms [21] have been widely used in LCD methods based on global features. Compared with global descriptors, local descriptors are less sensitive to scale and viewpoint changes but computationally expensive. The local feature extraction is regularly divided into two steps: detecting keypoints and generating descriptors. In particular, SURF [22], SIFT [23], and ORB [24] are the most widely used local features in LCD. Recently, convolution neural networks (CNNs) have been used to extract image features and improve the LCD performance. In particular, An *et al.* [25] used the lightweight neural network MobileNet to achieve real-time performance significantly. Khaliq *et al.* [26] developed a novel CNN-based regional feature for achieving better performance of their place recognition approach. DeTone *et al.* [17] presented the SuperPoint network to simultaneously compute pixel-level interest point locations and associate local feature descriptors. Sarlin *et al.* [27] proposed a monolithic CNN named HF-Net to simultaneously extract global descriptors and local features for accurate 6-DoF localization. In this paper, our proposed LCD approach also focuses on exploiting learned local features as robust image representation.

B. Candidate Frame Selection

There are two categories of LCD methods to select candidate frames, *i.e.*, image sequence matching-based method and single image matching-based method. Bampis *et al.* [28] dynamically partitioned the image stream into image sequences and provided individual image associations so as to accurately detect loop closing pairs. In this paper, we principally focus on LCD approaches on the basis of single image matching, where the LCD task can be regarded as an image retrieval problem. Specifically, Angeli *et al.* [29] extended the BoW framework to incremental conditions and used Bayesian filtering to estimate the probability of loop closure. Recently, instead of using hand-crafted features, Hou *et al.* [30] combined CNN features with the BoW framework to achieve faster performance. Yue *et al.* [16] also applied SuperPoint neural network to extract features and combined with an incremental BoW scheme. In addition to the BoW framework, there are several other retrieval methods which show more precise results, such as vector of locally aggregated descriptors (VLAD) [31] and Fisher vector [32]. Arandjelovic *et al.* [33] proposed a new generalized VLAD layer, called NetVLAD, which can insert into any CNN architecture. Khaliq *et al.* [26] combined VLAD with CNN-based regional features to achieve promising results with a smaller visual word dictionary.

C. Loop Closing Verification

In general, LCD can utilize the information that comes from the following three aspects to verify candidate frames: topological map, metric map, and geometrical information [34]. This paper focuses on using geometrical information of image feature points to enhance the robustness of LCD systems [35]. RANSAC [15] is widely utilized in LCD approaches. It can robustly establish reliable correspondences and estimate transformation between image pairs, but it relies on a predefined parameter model. Therefore, it fails to work in the case of complex transformations, *e.g.*, non-rigid. In addition, RANSAC shows inferior results when the outlier ratio of an image pair is large. These motivate researches to develop new geometrical verification methods for LCD. Yue *et al.* [16]

proposed a graph verification method, which builds an undirected triangular graph for candidate images to graph match with the current image. Sarlin *et al.* [36] introduced an attention-based graph neural network termed SuperGlue for local feature matching, which is readily integrated into SLAM systems since it can run in real-time. Moreover, several locality consistency assumption-based methods [37], [38] are proposed in recent researches of image matching and achieved encouraging performance in both accuracy and efficiency. In this paper, we aim to design a tailor-made feature matching algorithm based on locality consistency to handle common and challenging scenarios in the LCD tasks.

III. Method

Figure 1 illustrates an overview of the proposed LCD method, which consists of three parts: using SuperPoint to extract features, employing ASMK to select candidate frames, and adopting LRO to verify loop closing pairs. On the off-line stage, we train a visual vocabulary used to quantize the feature descriptor space into the visual words. On the on-line stage, we employ ASMK to aggregate descriptors extracted by SuperPoint and compute the similarity between the query image and the database images, then select the top k most similar images from database images as candidate frames. In order to accurately reject false candidate frames, we provide a geometric verification method based on LRO to match the query image and its candidate frames. Specifically, when their inlier ratios are high enough, they are identified as correct loop closing pairs. Next, we will discuss the three parts in detail.

A. Feature Extraction Based on SuperPoint Network

Since there exist many disturbances in the images faced by LCD, such as viewpoint change, illuminate change and dynamic object occlusion, it is crucial for an appearance-based LCD approach to extract repeatable and distinctive image features. In this paper, we employ a pre-trained SuperPoint network [17] by a self-supervised framework to extract robust interest points and feature descriptors. The main advantages of the SuperPoint network consist of the following three aspects. First, SuperPoint can simultaneously detect the precise locations of interest feature points and generate feature descriptors, rather than first performing interest point detection and then coupling with feature descriptors in traditional way, which cannot share computation and representation across these two tasks. Second, as opposed to patch-based neural networks, SuperPoint is a fully-convolutional model operated on full-sized images. Finally, the SuperPoint network can achieve real-time performance by using the GPU technique.

Figure 2 illustrates the basic architecture of SuperPoint. It has a shared encoder to reduce the dimensionality of the input image, which includes convolutional layers, spatial pooling layers, and non-linear activation functions. After the encoder, the output is fed into an interest point decoder and a descriptor decoder. Both decoders utilize the non-learned upsampling method to reduce training difficulty and runtime.

B. Aggregated Selective Match Kernel

To efficiently select candidate frames of the current query image, compact image representations are required to retrieve database images quickly. Clearly, directly using feature descriptors to represent images is time- and memory-consuming. To address this issue, we introduce the aggregated selective match kernel framework [18] to compact image representation and effectively compute the similarity between images. Compared with the BoW model, ASMK can overcome the problem of quantization noise and achieve higher precision.

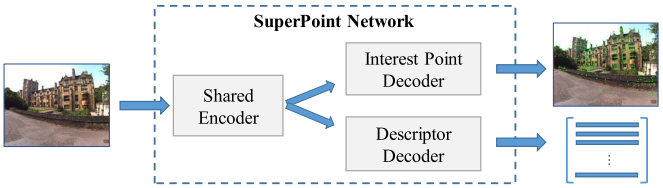

Fig. 2. The basic architecture of the SuperPoint network.

Assume that an image X is described by a set $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$, which consists of n D-dimensional local descriptors extracted by the SuperPoint network. A vocabulary $\mathcal{U}$ can be learned by a clustering algorithm g (*e.g.*, K-means) to quantize the descriptors from a large number of images, where $\mathcal{U} = \{u_1, u_2, \ldots, u_t\}$ contains t visual words. By assigning n descriptors in $\mathcal{X}$ to the t visual words in $\mathcal{U}$, we can obtain a collection $\{\mathcal{X}_{u_1}, \mathcal{X}_{u_2}, \ldots, \mathcal{X}_{u_t}\}$, where $\mathcal{X}_u = \{x : x \in \mathcal{X}, g(x) = u\}$ represents a set of descriptors assigned to the same visual word u. For image Y, we can also obtain descriptor sets $\mathcal{Y}$ and $\mathcal{Y}_u$.

Similar to VLAD [31], the aggregate vector $V(\mathcal{X}_u)$ for a visual word is defined as the sum of the difference between descriptors assigned to the same visual word and the visual word, *i.e.*,

$$V(\mathcal{X}_u) = \sum_{x \in \mathcal{X}_u} (x - g(x)). \tag{1}$$

Then, $\hat{V}(\mathcal{X}_u)$ is defined as the l_2-normalization of $V(\mathcal{X}_u)$:

$$\hat{V}(\mathcal{X}_u) = V(\mathcal{X}_u) / \| V(\mathcal{X}_u) \|_2. \tag{2}$$

In the same manner, $V(\mathcal{Y}_u)$ and $\hat{V}(\mathcal{Y}_u)$can be obtained. Next, ASMK uses the inner product to compute the similarity

$$r = \hat{V}(\mathcal{X}_u)^\top \hat{V}(\mathcal{Y}_u). \tag{3}$$

In order to reduce weight for false correspondences, ASMK defines a scalar selectivity function

$$\sigma_\alpha(r) = \begin{cases} \text{sign}(r) \cdot |r|^\alpha, & \text{if } r > \tau, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Therefore, the similarity between two images X and Y is defined as

$$\mathcal{K}(\mathcal{X}, \mathcal{Y}) = \gamma(\mathcal{X}) \gamma(\mathcal{Y}) \sum_{u \in \mathcal{U}} \sigma_\alpha \left(\hat{V}(\mathcal{X}_u)^\top \hat{V}(\mathcal{Y}_u) \right), \tag{5}$$

where the normalization factor $\gamma(\cdot)$ is calculated as:

$$\gamma(\mathcal{X}) = \left(\sum_{u \in \mathcal{U}} \sigma_\alpha \left(\hat{V}(\mathcal{X}_u)^\top \hat{V}(\mathcal{X}_u) \right) \right)^{-1/2}. \tag{6}$$

By binarizing $\mathcal{X}_u$ before applying the selectivity function, the binarization of similarity function $\mathcal{K}$ can be determined as:

$$\mathcal{K}^{\star}(\mathcal{X},\mathcal{Y}) = \gamma(\mathcal{X})\gamma(\mathcal{Y}) \sum_{u\in\mathcal{U}} \sigma_{\alpha}\left\{\hat{b}\left(\hat{V}(\mathcal{X}_u)\right)^{\top} \hat{b}\left(\hat{V}(\mathcal{Y}_u)\right)\right\}, \tag{7}$$

where $\hat{b}(\cdot)$ is a normalized function of $b(\cdot)$ (see Eq. (2) for example), and $b(\cdot)$ is an element-wise binarization function, *i.e.*, $b_i(x) = 1$ if $x_i \geq 0$, otherwise, $b_i(x) = -1$.

C. Temporal Constraint

During LCD tasks, the image currently captured by the camera is considered as a query image, and the previously captured images are regarded as reference images. We aim to find candidate frames of the query image from reference images. However, due to the frame rate of camera acquisition and the velocity of camera movement, images that are close in time are likely to be similar in appearance. In order to avoid such images becoming candidate frames, we specify a temporal constraint to require N_{tem} neighborhood images of query image not to participate in the similarity calculation, and N_{tem} is determined as:

$$N_{tem} = f \cdot T, \tag{8}$$

where f is the frame rate, and T is a pre-defined parameter.

By the ASMK framework, we obtain the similarity scores between the query image and its reference images. The top-k reference images with the greatest scores and satisfied with the temporal constraint are considered as candidate frames.

IV. Loop Closing Verification Based on LRO

In this section, we describe a novel feature matching method for verifying loop closing pairs, which follows a two-step strategy. The first step is to establish a set of putative matches of feature point pairs by considering all possible correspondences between the current image and its candidate frame meanwhile filtering out matches whose feature descriptors are sufficiently different. Specifically, we combine SuperPoint features with the distance ratio method [11]. For each feature point extracted by the SuperPoint network in the current image, the method searches the two nearest neighbors in descriptor space in the candidate frame, then computes their distance ratio. Only when passed distance ratio test, the putative match can be kept. The second step is to remove mismatches in the putative set using our proposed local relative orientation matching method, which aims to preserve local neighborhood topological structures of feature correspondences. The LRO method is designed based on our previous proposed robust feature matching method, termed as locality preserving matching (LPM) [37]. In the following, we first give a brief introduction to the LPM approach, and then point out its limitations on the LCD task and propose our new model accordingly.

A. Locality Preserving Matching

Assume that there has a set of N putative feature correspondences $\mathcal{S} = (\mathbf{x}_i, \mathbf{y}_i)_{i=1}^{N}$, where $\mathbf{x}_i$ and $\mathbf{y}_i$ respectively denote the feature point coordinates extracted from two given images using SuperPoint, and $(\mathbf{x}_i, \mathbf{y}_i)$ denotes the i-th putative correspondences. The goal of LPM is to filter out the outliers in $\mathcal{S}$ and construct accurate correspondences accordingly.

If the image pair undergoes a rigid transformation, then the distance between any feature correspondence will be preserved. Denoting the unknown inlier set as $\mathcal{I}$, the optimal solution of LPM is:

$$\mathcal{I}^{*} = \arg\min_{\mathcal{I}} C(\mathcal{I}; \mathcal{S}, \lambda), \tag{9}$$

where C denotes the cost function, and it is defined as:

$$C(\mathcal{I}; S, \lambda) = \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{I}} (d(\mathbf{x}_i, \mathbf{x}_j) - d(\mathbf{y}_i, \mathbf{y}_j))^2 + \lambda(N - |\mathcal{I}|). \tag{10}$$

In Eq. (10), d denotes a measure of distance, and $|\cdot|$ is the cardinality of a set. The first term penalizes any match that does not preserve the distance of a point pair, and the second term discourages the outliers. Parameter λ is used for trading off the two terms.

However, when two images undergo some complex transformations, such as non-rigid transformation, the absolute distance between two points cannot be maintained well. Nevertheless, due to physical constraints, the distribution of neighboring point pairs after transformation should be well preserved, although two images captured in the same place may have a significant viewpoint change or non-rigid deformation. Therefore, by preserving only local structures, the cost function in Eq. (10) becomes:

$$\begin{aligned} C(\mathcal{I}; \mathcal{S}, \lambda) = \sum_{i\in\mathcal{I}} \frac{1}{2K} \Bigg(& \sum_{j|\mathbf{x}_j\in\mathcal{N}_{\mathbf{x}_i}} \left(d(\mathbf{x}_i, \mathbf{x}_j) - d(\mathbf{y}_i, \mathbf{y}_j)\right)^2 \\ & + \sum_{j|\mathbf{y}_j\in\mathcal{N}_{\mathbf{y}_i}} \left(d(\mathbf{x}_i, \mathbf{x}_j) - d(\mathbf{y}_i, \mathbf{y}_j)\right)^2 \Bigg) \\ & + \lambda(N - |\mathcal{I}|), \end{aligned} \tag{11}$$

where the K nearest neighbors of feature point $\mathbf{x}_i$ are regarded as its neighborhood $\mathcal{N}_{\mathbf{x}_i}$, $1/2K$ is a normalizing factor, and LPM quantizes the distance metric d into two levels as:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} 0, & \mathbf{x}_j \in \mathcal{N}_{\mathbf{x}_i} \\ 1, & \mathbf{x}_j \notin \mathcal{N}_{\mathbf{x}_i}, \end{cases} \tag{12}$$

$$d(\mathbf{y}_i, \mathbf{y}_j) = \begin{cases} 0, & \mathbf{y}_j \in \mathcal{N}_{\mathbf{y}_i} \\ 1, & \mathbf{y}_j \notin \mathcal{N}_{\mathbf{y}_i}. \end{cases} \tag{13}$$

B. Relative Orientation of Topological Structures

On account of the fact that LPM primarily concentrates on the composition of the neighborhood, its performance of mismatch removal would be degraded in complex scenes, such as involving many repetitive patterns (*e.g.*, Fig. 3) or a high outlier rate. To address this problem, we aim to strengthen the topological constrains for the neighborhood structure of correspondences. In particular, we propose a novel feature matching method based on local relative orientation invariance and affine-invariant, termed as LRO.

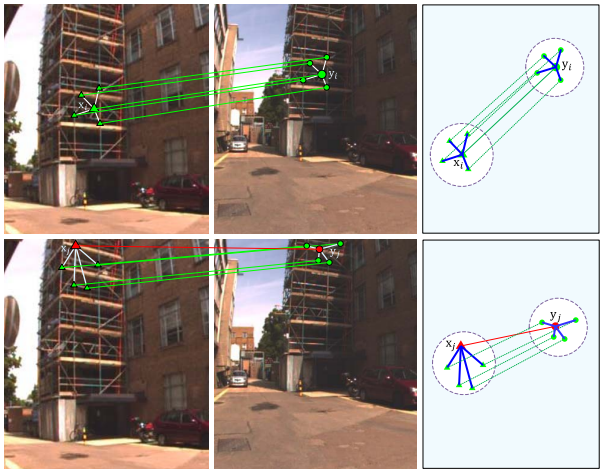

Fig. 3. Different topological construction similarities shown by inlier (top) and outlier (bottom). Green: true match; red: false match.

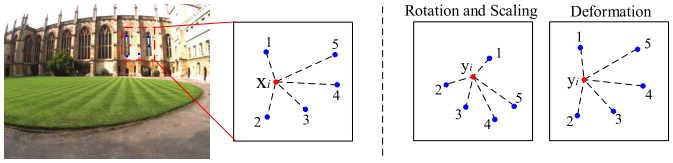

Fig. 4. The different neighborhood topological structures under rotation, scaling transformations and deformation.

As shown in the dashed circle in Fig. 3, we search the four nearest neighbors under the Euclidean distance as the neighborhood of the feature point. On the top of Fig. 3, $\mathbf{x}_i$ and $\mathbf{y}_i$ respectively denote the feature point in the query image and one of its candidate frames, and $(\mathbf{x}_i, \mathbf{y}_i)$ is a true match, also termed as inlier. The green lines connect the correspondences in the putative set $\mathcal{S} = (\mathbf{x}_i, \mathbf{y}_i)_{i=1}^{N}$, which co-occur in these two neighborhoods. The blue lines represent the topological neighborhood structure of feature points. As we can see, the topological structures of the two feature points are extremely consistent. However, on the bottom of Fig. 3, $\mathbf{x}_j$ and $\mathbf{y}_j$ are a false match, also termed as outlier. We can see that the topological structures are almost completely inconsistent, especially reflected in that neighbors are distributed in the different relative orientations of the feature point. Although the outlier has the same number of common neighbors with the inlier in the top of Fig. 3, our algorithm can still identify outliers and inliers by analyzing their topological structures of neighborhoods.

To illustrate the core idea of LRO more clearly, Fig. 4 shows changes in neighborhood topological structures under different image transformations. When an image is rotated or scaled, feature point $\mathbf{y}_i$ and its neighborhood only change the absolute coordinates in the image, and the orientation of each neighbor relative to the center point $\mathbf{y}_i$ is not changed. To formulate this relative orientation invariance, we consider the measurement of the angle formed by the center feature point and its neighbors. More specifically, each angle formed between the center feature point and its neighbors should be kept consistent after image transformation. In order to cope with complex cases, such as deformation shown in the right of Fig. 4, we set a pre-defined threshold to describe the degree of angle consistent. With a reasonable threshold, the algorithm would be robust and able to cope with image distortion.

Relative Orientation

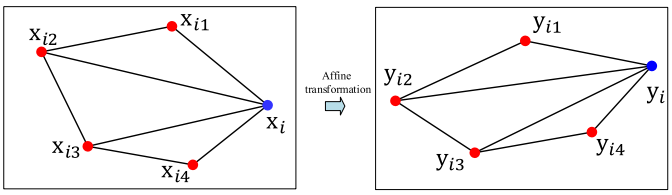

Fig. 5. The topological structures under an affine transformation, where $(\mathbf{x}_i, \mathbf{y}_i)$ is a putative correspondence, $\mathbf{x}_{i1}$, $\mathbf{x}_{i2}$, $\mathbf{x}_{i3}$, $\mathbf{x}_{i4}$ are the four nearest neighbors of $\mathbf{x}_i$ and respectively correspond to $\mathbf{y}_{i1}$, $\mathbf{y}_{i2}$, $\mathbf{y}_{i3}$, $\mathbf{y}_{i4}$ after the affine transformation.

C. The Affine-Invariant of Topological Structures

However, when viewpoint changes and serious geometric distortions occur in the image, the above property of local relative orientation cannot be preserved well. The transformations cannot be modeled in global, but in general an affine or homography transformation can well model the topological structures of local regions inside the image. Therefore, we introduce affine-invariant into the LRO algorithm. According to the theory of multiple view geometry in computer vision, the affine transformation has three important invariants: parallel lines, ratios of lengths of parallel line segments, and ratios of areas. In this paper, we use the ratio of areas to further verify correspondences that do not conform to the invariant of local relative orientation. As shown in Fig. 5, given a feature point $\mathbf{x}_i$ and its neighbors $\mathbf{x}_{i1}$, $\mathbf{x}_{i2}$, $\mathbf{x}_{i3}$, $\mathbf{x}_{i4}$, they can form three triangles: $\Delta\mathbf{x}_i\mathbf{x}_{i1}\mathbf{x}_{i2}$, $\Delta\mathbf{x}_i\mathbf{x}_{i2}\mathbf{x}_{i3}$, and $\Delta\mathbf{x}_i\mathbf{x}_{i3}\mathbf{x}_{i4}$. After affine transformation, points $\mathbf{x}_i$, $\mathbf{x}_{i1}$, $\mathbf{x}_{i2}$, $\mathbf{x}_{i3}$, $\mathbf{x}_{i4}$ are converted to $\mathbf{y}_i$, $\mathbf{y}_{i1}$, $\mathbf{y}_{i2}$, $\mathbf{y}_{i3}$, $\mathbf{y}_{i4}$, and $\Delta\mathbf{x}_i\mathbf{x}_{i1}\mathbf{x}_{i2}$, $\Delta\mathbf{x}_i\mathbf{x}_{i2}\mathbf{x}_{i3}$, $\Delta\mathbf{x}_i\mathbf{x}_{i3}\mathbf{x}_{i4}$ are transformed to $\Delta\mathbf{y}_i\mathbf{y}_{i1}\mathbf{y}_{i2}$, $\Delta\mathbf{y}_i\mathbf{y}_{i2}\mathbf{y}_{i3}$, $\Delta\mathbf{y}_i\mathbf{y}_{i3}\mathbf{y}_{i4}$, respectively. Moreover, the areas of these triangles satisfy the following equation based on the affine invariant:

Affine-Invariant

$$S_{\Delta\mathbf{x}_i\mathbf{x}_{i1}\mathbf{x}_{i2}} : S_{\Delta\mathbf{x}_i\mathbf{x}_{i2}\mathbf{x}_{i3}} : S_{\Delta\mathbf{x}_i\mathbf{x}_{i3}\mathbf{x}_{i4}} = S_{\Delta\mathbf{y}_i\mathbf{y}_{i1}\mathbf{y}_{i2}} : S_{\Delta\mathbf{y}_i\mathbf{y}_{i2}\mathbf{y}_{i3}} : S_{\Delta\mathbf{y}_i\mathbf{y}_{i3}\mathbf{y}_{i4}}. \quad (14)$$

By using a simple transform, we have:

$$\frac{S_{\Delta\mathbf{x}_i\mathbf{x}_{i1}\mathbf{x}_{i2}}}{S_{\Delta\mathbf{x}_i\mathbf{x}_{i2}\mathbf{x}_{i3}}} = \frac{S_{\Delta\mathbf{y}_i\mathbf{y}_{i1}\mathbf{y}_{i2}}}{S_{\Delta\mathbf{y}_i\mathbf{y}_{i2}\mathbf{y}_{i3}}}, \quad \frac{S_{\Delta\mathbf{x}_i\mathbf{x}_{i2}\mathbf{x}_{i3}}}{S_{\Delta\mathbf{x}_i\mathbf{x}_{i3}\mathbf{x}_{i4}}} = \frac{S_{\Delta\mathbf{y}_i\mathbf{y}_{i2}\mathbf{y}_{i3}}}{S_{\Delta\mathbf{y}_i\mathbf{y}_{i3}\mathbf{y}_{i4}}}. \quad (15)$$

Eq. (15) indicates that the ratios of any two triangles in the neighborhood are invariant to affine transformations.

D. Problem Formulation

To exploit the relative orientation invariance of topological structures and affine invariant for feature matching, here we redesign the cost function in Eq. (9). We first define a set of common feature correspondences $Q_i = \{(\mathbf{x}_{iq}, \mathbf{y}_{iq})\}_{q=1}^{L}$ for the putative match $(\mathbf{x}_i, \mathbf{y}_i)$, where $(\mathbf{x}_{iq}, \mathbf{y}_{iq})$ is a putative match with $\mathbf{x}_{iq} \in \mathcal{N}_{\mathbf{x}_i}$ and $\mathbf{y}_{iq} \in \mathcal{N}_{\mathbf{y}_i}$, and L is the total number of common correspondences for $(\mathbf{x}_i, \mathbf{y}_i)$.

Figure 6 shows the i-th putative correspondence together with its neighborhood elements in the putative set $\mathcal{S} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$, where $\{(\mathbf{x}_{iq}, \mathbf{y}_{iq})\}_{q=1}^{4}$ are common elements

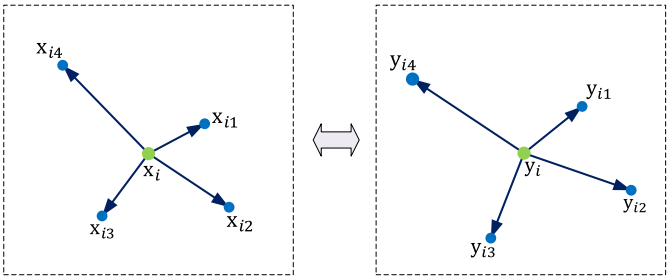

Fig. 6. The neighborhood topological structure diagram of the i-th feature point pairs $(\mathbf{x}_i, \mathbf{y}_i)$, and $(\mathbf{x}_{i1}, \mathbf{y}_{i1})$, $(\mathbf{x}_{i2}, \mathbf{y}_{i2})$, $(\mathbf{x}_{i3}, \mathbf{y}_{i3})$, $(\mathbf{x}_{i4}, \mathbf{y}_{i4})$ are common elements in the two neighborhoods.

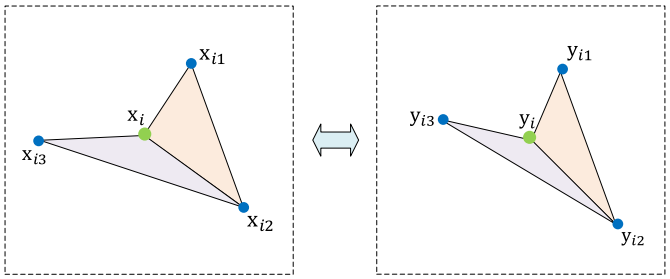

Fig. 7. The affine-invariant of topological structures. Points $\mathbf{x}_i$, $\mathbf{x}_{i1}$, $\mathbf{x}_{i2}$, $\mathbf{x}_{i3}$ form two triangles $\Delta\mathbf{x}_i\mathbf{x}_{i1}\mathbf{x}_{i2}$ and $\Delta\mathbf{x}_i\mathbf{x}_{i2}\mathbf{x}_{i3}$, which correspond to $\Delta\mathbf{y}_i\mathbf{y}_{i1}\mathbf{y}_{i2}$ and $\Delta\mathbf{y}_i\mathbf{y}_{i2}\mathbf{y}_{i3}$ after transformation.

obtained from the two neighborhoods $\mathcal{N}_{\mathbf{x}_i}$ and $\mathcal{N}_{\mathbf{y}_i}$. Suppose that feature points $\mathbf{x}_i$ and $\mathbf{x}_{iq}$ form vector $\mathbf{vx}_{iq}$. In order to represent the relative orientation invariance of topological structures, we compare the similarity of angles formed by two neighborhoods. Specifically, the angle value formed between $\mathbf{vx}_{iq}$ and $\mathbf{vx}_{i(q+1)}$ is defined as:

$$\theta x\,(i,q) = \arccos \frac{\langle \mathbf{vx}_{iq}, \mathbf{vx}_{i(q+1)} \rangle}{|\mathbf{vx}_{iq}| \cdot |\mathbf{vx}_{i(q+1)}|}, \tag{16}$$

where $\arccos(\cdot)$ is the arccosine function, and $\langle \cdot, \cdot \rangle$ denotes the inner product. Similarly, we can obtain the angle value formed between $\mathbf{vy}_{iq}$ and $\mathbf{vy}_{i(q+1)}$:

$$\theta y\,(i,q) = \arccos \frac{\langle \mathbf{vy}_{iq}, \mathbf{vy}_{i(q+1)} \rangle}{|\mathbf{vy}_{iq}| \cdot |\mathbf{vy}_{i(q+1)}|}. \tag{17}$$

The similarity of these two angles is defined as:

$$s\,(i,q) = 1 - \frac{|\theta x\,(i,q) - \theta y\,(i,q)|}{\max\{\theta x\,(i,q)\,, \theta y\,(i,q)\}}, \tag{18}$$

where $s\,(i,q) \in [0,1]$ and a larger value indicates a higher similarity. We construct a loss function term $s_{LRO}(i,q)$ of the neighborhood topological structure by setting a predefined threshold τ_1:

$$s_{LRO}(i,q) = \begin{cases} f(i,q), & s\,(i,q) \le \tau_1, \\ 0, & s\,(i,q) > \tau_1. \end{cases} \tag{19}$$

If the similarity $s_{LRO}(i,q)$ is higher than the threshold τ_1, there has no punishment for the i-th feature correspondence. Otherwise, we further verify whether the topological structure conforms to an affine model.

As illustrated in Fig. 7, it is possible to exist an affine transformation between these two neighborhood structures. The ratios of the area of triangles are defined as:

$$\mu_q^{\mathbf{x}_i} = \frac{S_{\Delta\mathbf{x}_i\mathbf{x}_{iq}\mathbf{x}_{i(q+1)}}}{S_{\Delta\mathbf{x}_i\mathbf{x}_{i(q+1)}\mathbf{x}_{i(q+2)}}}, \quad \mu_q^{\mathbf{y}_i} = \frac{S_{\Delta\mathbf{y}_i\mathbf{y}_{iq}\mathbf{y}_{i(q+1)}}}{S_{\Delta\mathbf{y}_i\mathbf{y}_{i(q+1)}\mathbf{y}_{i(q+2)}}}, \tag{20}$$

where $q \in \{0, 1, \ldots, L-1\}$, in particular, when $q = L-1$, we set $q+2 = 1$. According to affine-invariant previously mentioned and Eq. (15), we can infer that $\mu_q^{\mathbf{x}_i}$ and $\mu_q^{\mathbf{y}_i}$ should remain consistent when neighborhood structures fit an affine transformation. Considering various interference in real images, we set a predefined threshold τ_2 to establish a measure function $f(i,q)$:

$$f(i,q) = \begin{cases} 0, & |\mu_q^{\mathbf{x}_i} - \mu_q^{\mathbf{y}_i}| \le \tau_2, \\ 1, & |\mu_q^{\mathbf{x}_i} - \mu_q^{\mathbf{y}_i}| > \tau_2. \end{cases} \tag{21}$$

If the difference between $\mu_q^{\mathbf{x}_i}$ and $\mu_q^{\mathbf{y}_i}$ is below the threshold τ_2, it shows that this part of neighborhood topological structures conform to the affine transformation, and hence the i-th feature correspondence should not be penalized. Otherwise, they should be penalized.

Finally, we design a loss function d_{LRO} of the i-th putative correspondence by summing all of the loss function terms of its neighborhood, which is defined as:

$$d_{LRO}\,(\mathbf{x}_i, \mathbf{y}_i) = \sum_{q=1}^{L-1} s_{LRO}\,(i,q)\,. \tag{22}$$

Ideally, for a true match $(\mathbf{x}_i, \mathbf{y}_i)$, the function $d_{LRO}(\mathbf{x}_i, \mathbf{y}_i)$ should be equal to 0.

By adding the loss function d_{LRO} into Eq. (11), we obtain the cost function of our LRO algorithm:

$$\begin{aligned} C(\mathcal{I}; S, \lambda) = \sum_{i\in\mathcal{I}} \frac{1}{2K} \Bigg(& \sum_{j|\mathbf{x}_j\in\mathcal{N}_{\mathbf{x}_i}} \big(d(\mathbf{x}_i, \mathbf{x}_j) - d(\mathbf{y}_i, \mathbf{y}_j)\big)^2 \\ & + \sum_{j|\mathbf{y}_j\in\mathcal{N}_{\mathbf{y}_i}} \big(d(\mathbf{x}_i, \mathbf{x}_j) - d(\mathbf{y}_i, \mathbf{y}_j)\big)^2 \\ & + d_{LRO}(\mathbf{x}_i, \mathbf{y}_i) \Bigg) + \lambda(N - |\mathcal{I}|). \end{aligned} \tag{23}$$

E. A Closed-Form Solution

We set an $N \times 1$ binary vector $\mathbf{p}$ to associate the putative set $\mathcal{S}$, where $p_i = 1$ represents that the i-th correspondence $(\mathbf{x}_i, \mathbf{y}_i)$ is a true match and $p_i = 0$ for a false match. With the distance in Eq. (12) and the binary vector $\mathbf{p}$, the cost function in Eq. (23) is simplified as:

$$\begin{aligned} C(\mathbf{p}; S, \lambda, \tau_1, \tau_2) = \sum_{i=1}^{N} \frac{p_i}{2K} \Bigg(& \sum_{j|\mathbf{x}_j\in\mathcal{N}_{\mathbf{x}_i}} d(\mathbf{y}_i, \mathbf{y}_j) \\ & + \sum_{j|\mathbf{y}_j\in\mathcal{N}_{\mathbf{y}_i}} d(\mathbf{x}_i, \mathbf{x}_j) + d_{LRO}(\mathbf{x}_i, \mathbf{y}_i) \Bigg) \\ & + \lambda(N - \sum_{i=1}^{N} p_i). \end{aligned} \tag{24}$$

The item $\sum_{j|\mathbf{x}_j\in\mathcal{N}_{\mathbf{x}_i}} d(\mathbf{y}_i,\mathbf{y}_j)$ can be rewritten as:

$$\begin{aligned}\sum_{j|\mathbf{x}_j\in\mathcal{N}_{\mathbf{x}_i}} d(\mathbf{y}_i,\mathbf{y}_j) &= \sum_{j|\mathbf{x}_j\in\mathcal{N}_{\mathbf{x}_i},\mathbf{y}_j\in\mathcal{N}_{\mathbf{y}_i}} d(\mathbf{y}_i,\mathbf{y}_j) \\ &\quad + \sum_{j|\mathbf{x}_j\in\mathcal{N}_{\mathbf{x}_i},\mathbf{y}_j\notin\mathcal{N}_{\mathbf{y}_i}} d(\mathbf{y}_i,\mathbf{y}_j) \\ &= 0 + count(j \mid \mathbf{x}_j\in\mathcal{N}_{\mathbf{x}_i},\mathbf{y}_j\notin\mathcal{N}_{\mathbf{y}_i}) \\ &= K - count(j \mid \mathbf{x}_j\in\mathcal{N}_{\mathbf{x}_i},\mathbf{y}_j\in\mathcal{N}_{\mathbf{y}_i}) \\ &= K - n_i, \end{aligned} \tag{25}$$

where $count(\cdot)$ counts the number of elements in the set, and n_i is the number of the common elements in the two neighborhoods $\mathcal{N}_{\mathbf{x}_i}$ and $\mathcal{N}_{\mathbf{y}_i}$. Similarly,

$$\sum_{j|\mathbf{y}_j\in\mathcal{N}_{\mathbf{y}_i}} d(\mathbf{x}_i,\mathbf{x}_j) = K - n_i = \sum_{j|\mathbf{x}_j\in\mathcal{N}_{\mathbf{x}_i}} d(\mathbf{y}_i,\mathbf{y}_j). \tag{26}$$

Substituting Eq. (26) into Eq. (24), we can obtain the cost function

$$\begin{aligned}C(\mathbf{p};S,\lambda,\tau_1,\tau_2) = \sum_{i=1}^{N}\frac{p_i}{K}\Bigg(&\sum_{j|\mathbf{x}_j\in\mathcal{N}_{\mathbf{x}_i}} d(\mathbf{y}_i,\mathbf{y}_j) \\ &+ d_{LRO}(\mathbf{x}_i,\mathbf{y}_i)\Bigg) + \lambda(N-\sum_{i=1}^{N} p_i). \end{aligned} \tag{27}$$

Considering that the putative matches are usually not uniformly distributed across image pairs, we employ a multi-scale neighborhood representation of LPM, which searches a set of neighborhoods with sizes $\mathbf{K}=\{K_m\}_{m=1}^{M}$, *e.g.*, $\{\mathcal{N}_{\mathbf{x}_i}^{K_m}\}_{m=1}^{M}$ and $\{\mathcal{N}_{\mathbf{y}_i}^{K_m}\}_{m=1}^{M}$, then Eq. (27) becomes:

$$\begin{aligned}\mathrm{C}(\mathbf{p};\mathrm{S},\lambda,\tau_1,\tau_2) = \sum_{i=1}^{N} p_i\Bigg(\sum_{m=1}^{M}\frac{1}{MK_m}\Bigg(&\sum_{j|\mathbf{x}_j\in\mathcal{N}_{\mathbf{x}_i}} d(\mathbf{y}_i,\mathbf{y}_j) \\ &+ d_{LRO}(\mathbf{x}_i,\mathbf{y}_j)\Bigg)\Bigg) + \lambda(N-\sum_{i=1}^{N} p_i). \end{aligned} \tag{28}$$

To optimize the objective function in Eq. (28), we merge the items related to p_i and Eq. (28) is reorganized as:

$$C(\mathbf{p};S,\lambda,\tau_1,\tau_2) = \sum_{i=1}^{N} p_i(c_i-\lambda) + \lambda N, \tag{29}$$

where

$$c_i = \sum_{m=1}^{M}\frac{1}{MK_m}\Bigg(\sum_{j|\mathbf{x}_j\in N_{\mathbf{x}_i}} d(\mathbf{y}_i,\mathbf{y}_j) + d_{LRO}(\mathbf{x}_i,\mathbf{y}_i)\Bigg). \tag{30}$$

Due to the neighborhood relationship among the feature points is fixed, $\{c_i\}_{i=1}^{N}$ can be estimated in advance. Hence, the only unknown variable in Eq. (29) is p_i. Any putative match with c_i smaller than λ will lead to a negative term of Eq. (29), and vice versa. Consequently, the optimal solution of $\mathbf{p}$ is determined as:

$$p_i = \begin{cases} 1, & c_i \le \lambda \\ 0, & c_i > \lambda, \quad i = 1,\ldots,N. \end{cases} \tag{31}$$

Algorithm 1 The LRO Algorithm

Input: putative set $\mathcal{S}=\{(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^{N}$, parameters $\mathbf{K}$, λ, τ_1, τ_2
Output: inlier set $\mathcal{I}^*$
1 Construct neighborhood $\{\mathcal{N}_{\mathbf{x}_i}^{K_m},\mathcal{N}_{\mathbf{y}_i}^{K_m}\}_{m=1,i=1}^{M,N}$ on $\mathcal{S}$;
2 Compute distance $\{d_{LRO}(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^{N}$ using Eq. (22);
3 Calculate cost $\{c_i\}_{i=1}^{N}$ using Eq. (30);
4 Determine $\mathcal{I}^*$ using Eqs. (32) and (31).

Finally, the optimal inlier set $\mathcal{I}^*$ is represented as:

$$\mathcal{I}^* = \{i \mid p_i = 1, i = 1,\ldots,N\}. \tag{32}$$

The procedure of our LRO is summarized in Alg. 1. In this paper, we use it to find an inlier set $\mathcal{I}^*$ and calculate the proportion of inliers in image pairs that consist of the current query image and its candidate frames. If the inlier proportion value is higher than the pre-defined threshold, the image pair is regarded as a true loop closing pair; otherwise, we reject this candidate frame.

F. Computational Complexity

Considering that the putative set $\mathcal{S}$ has N feature point correspondences, the time complexity of searching K_m nearest neighbors for each feature point is about $O((K_m+N)\log N)$ by using K-D tree [39]. When searching multi-scale neighborhoods with $\mathbf{K}=\{K_m\}_{m=1}^{M}$, the time complexity of Line 1 in Alg. 1 becomes $O((\sum_{m=1}^{M} K_m + MN)\log N)$. In Eq. (29), the calculation of cost c_i includes computing the angle and area of neighborhood structures, and the time complexity is less than $O(2N\sum_{m=1}^{M} K_m)$. Therefore, the total time complexity of our LRO is close to $O((\sum_{m=1}^{M} K_m + MN)\log N + 2N\sum_{m=1}^{M} K_m)$. As both M and $\sum_{m=1}^{M} K_m$ are constants and far less than N, the time complexity of LRO can be simply written as $O(N\log N)$. That is to say, LRO has linearithmic complexity about the element number of the given putative set, which is significant for the LCD tasks.

V. Experimental Results

In this section, we evaluate the performance of the proposed LCD approach on publicly available datasets and compare it with the state-of-the-art. In the following, we first describe the experimental datasets, and then present implementation details of our methods. Finally, we report the results on the feature matching and LCD tasks in detail.

A. Datasets

To more comprehensively evaluate the LCD performance, we select several publicly available datasets based on different surroundings and camera settings, such as image resolution, frame rate, and movement velocity, as shown in Table I.

The KITTI [40] dataset denotes the KITTI odometry benchmark, which consists of 22 outdoor sequences. There has 12 sequences containing loop closures. We use KITTI 00, 05, 06 in our experiments and adopt the ground truth for LCD provided by [41].

TABLE I
DETAILS ABOUT THE DATASETS USED IN OUR EVALUATION

Dataset	Description	Image Resolution	#Images	Frame Rate (Hz)
KITTI 00	Outdoor, dynamic	1241×376	4544	10
KITTI 05	Outdoor, dynamic	1241×376	2762	10
KITTI 06	Outdoor, dynamic	1226×370	1104	10
City Centre	Urban, dynamic	640×480	2474	<7
Lip6 Outdoor	Urban, highly dynamic	240×192	1063	1
Malaga 2009 Parking 6L	Outdoor, slightly dynamic	1024×768	3474	7.5

TABLE II
PARAMETER SETTINGS

Parameter	Description	Value
$\mathbf{K}$	Size of the neighborhood	[6, 8, 10]
τ_1	Threshold of locality relative orientation	0.55
τ_2	Threshold of affine invariance	0.2
λ	Threshold of distinguishing inliers and outliers	0.55
T	Time constraint	10
k	Number of candidate frames	5

Malaga 2009 Parking 6L [42] dataset was gathered at a parking lot of Malaga, which contains a wealth of visual information and many similar objects, such as trees and roads. These appearances would degrade the LCD performance.

City Centre [43] dataset was collected by cameras mounted on a wheeled robotic platform. Every 1.5m triggered an image collection to capture images from the left and right of the robot. This dataset contains illumination change, viewpoint change, and dynamic objects occlusion, making it difficult to detect loop closing pairs.

Lip6 Outdoor [29] dataset was recorded by a hand-held camera at an average walking speed. There are plenty of dynamic objects in the dataset, and the image solution is low, which is challenging for LCD.

B. Implementation Details

Table II summarizes the parameter settings of our approach. For feature extraction, we employ a pre-trained SuperPoint model to detect feature points and generate 256-dimensional local descriptors. The descriptors are utilized by the ASMK* framework to select candidate frames, and the feature points are mainly used to verify candidate frames by the LRO algorithm. The experiments are performed on a PC with Intel(R) Xeon(R) CPU E5-2673 v3 @ 2.40GHz and 64 GB memory. We run the SuperPoint network using a GeForce GTX 1080 Ti GPU under the PyTorch framework. In addition, ASMK and LRO are implemented by MATLAB code. The visual vocabulary of the ASMK framework is created offline with the SuperPoint descriptors extracted from Oxford Buildings [44] and Paris [45] datasets. To further reduce the LCD runtime, we select the top 300 reliable features points for LRO to check candidate loop closing pairs.

C. Results on Feature Matching

Before evaluating the LCD performance, we first verify the effectiveness and efficiency of our LRO algorithm, compared with six state-of-the-art feature matching methods, including RANSAC [15], ICF [46], GMS [38], RFM-SCAN [47], LMR [48] and LPM [37]. We implement these algorithms based on publicly available codes and tune parameters to achieve their best performance as far as possible we can.

To present more intuitive results on the feature matching performance, we select four representative images of loop closing pairs and manually build ground truth by checking each putative match in each image pair. As shown in Fig. 8, the first two image pairs represent two challenging situations for LCD: repetitive structures and non-rigid transformation, the third image pair represents a more general case undergoing some slight viewpoint and illumination changes, while the last image pair undergoes significant illumination and viewpoint variation, moderate scale variation and dynamic object occlusion. In the results, we see that RANSAC does not achieve good performance, especially when non-rigid transformation occurs such as the second and the last image pairs, as it only relies on a parametric model. The non-parametric model-based ICF fails in these four situations since it cannot cope with the scenes involving large depth discontinuity or motion inconsistency. GMS has no satisfying result either, especially on the first and the last image pairs, because it requires large-scale putative matches to achieve better performance. In general, RFM-SCAN, LMR and LPM can obtain better performance as they can deal with scenes involving non-rigid transformation and motion inconsistency, and have excellent performance on the third image pair. However, their performance would be degraded in case of repetitive patterns in the first image pair and complex transformation in the last image pair. Moreover, RANSAC and GMS also show worse performance in the last image pair than the performance in the third image pair, namely, they can handle moderate appearance change but do not cope well with significant appearance variation in the LCD tasks. In contrast, our LRO clearly has the best matching results among these methods when handling both challenging scenes and common scenes.

We further conduct quantitative experiments on two feature matching datasets: *LCD* and *VGG* [49]. The statistical results are reported in Fig. 9. The *LCD* dataset consists of 25 loop closing image pairs from the above mentioned six datasets, where the ground-truth correspondences are established by manually checking the correctness of each putative correspondence in each image pair. *VGG* is a publicly available dataset for general feature matching tasks, which contains 40 image pairs that obey homography. Their average inlier

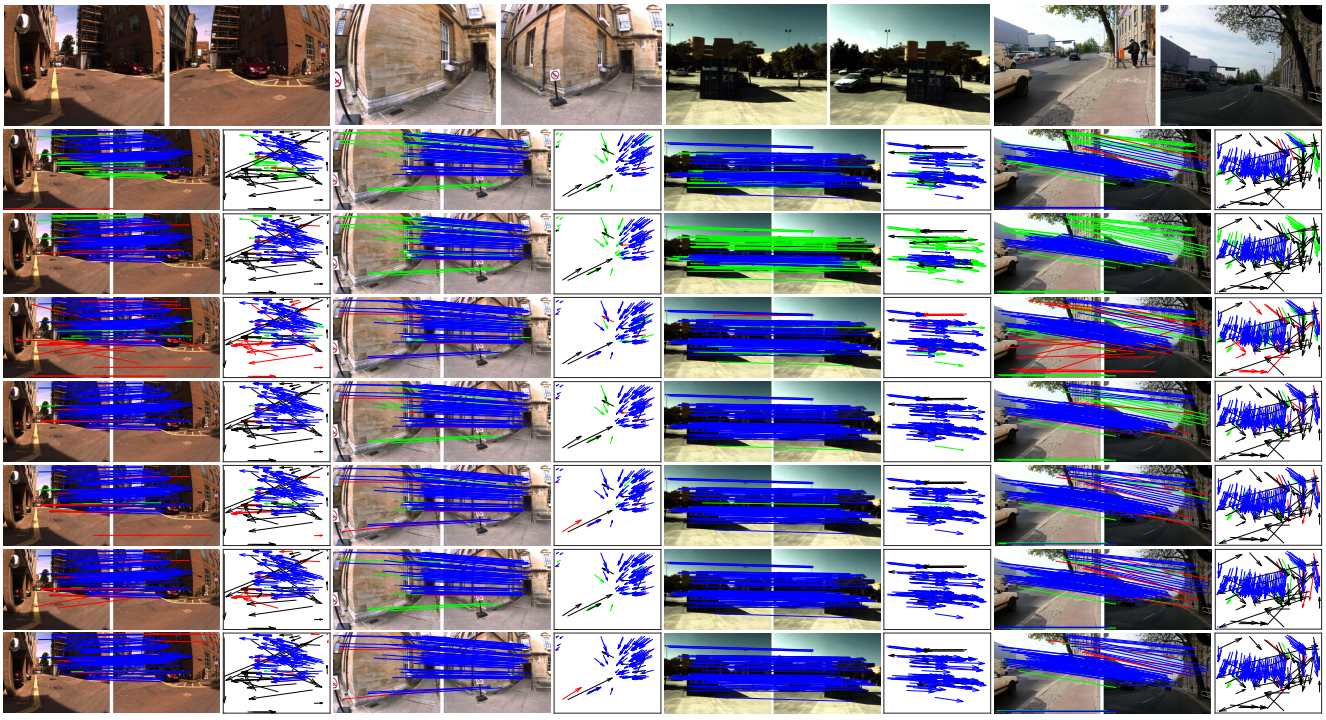

Fig. 8. Feature matching results of our LRO and other six methods on four typical image pairs (as shown in the first row) for the LCD task. From the second row to the last row: results of RANSAC [15], ICF [46], GMS [38], RFM-SCAN [47], LMR [48], LPM [37] and LRO. In each group, the head and tail of each arrow in the right motion field correspond to feature points in the left two images. Various color lines indicate different types of matching results (black lines denote true negative, blue lines denote true positive, red lines denote false positive, and green lines denote false negative).

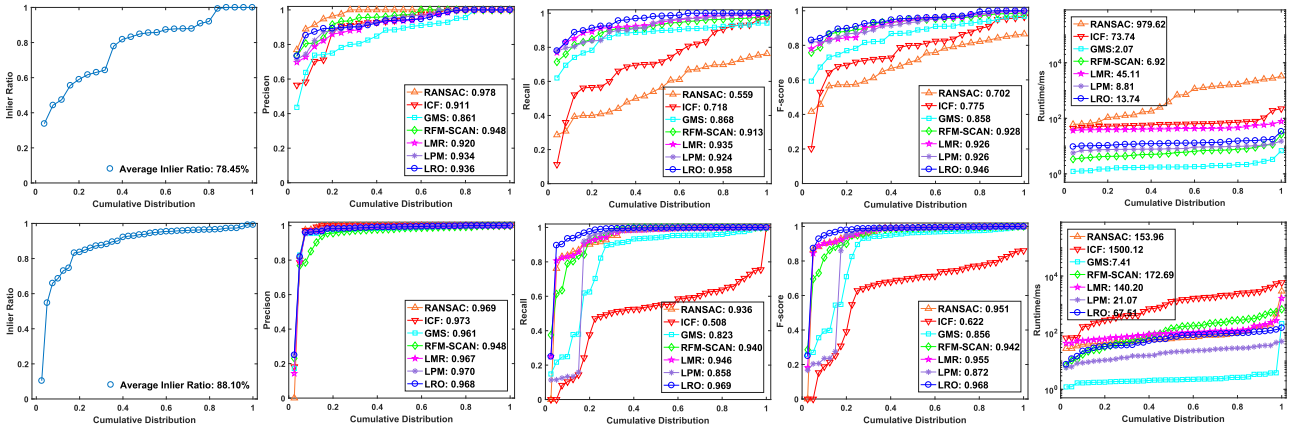

Fig. 9. Quantitative comparisons of feature matching using RANSAC [15], ICF [46], GMS [38], RFM-SCAN [47], LMR [48], LPM [37] and LRO on two datasets, such as *LCD* (top) and *VGG* [49] (bottom). Left to right: Initial inlier ratio, precision, recall, and runtime with respect to the cumulative distribution. A point on the curve with coordinate (x, y) denotes that there are $100 \times x$ percents of image pairs which have initial inlier ratio, precision, recall, or runtime no more than y.

ratios are 78.45% and 88.10%, respectively. We characterize the matching performance by precision, recall and F-score. The precision is defined as the percentage of true inliers among preserved matches by a matching algorithm, and the recall is the ratio of the preserved true inliers among the whole ground-truth inlier set. F-score is determined as the harmonic mean of precision and recall, which is equal to $2 \cdot Precision \times Recall/(Precision + Recall)$. From the results, we see that the precision of our LRO only attains a medium level among the seven methods, but it can achieve satisfying recall and the best F-score, that is to say, LRO has the best precision-recall trade-off. This is because our method can effectively deal with motion inconsistency or non-rigid deformation, unlike RANSAC and other existing methods preserving only a part of true matches that obey some specific motion models. As we can see, the F-score index of LRO is higher than LPM on all the datasets. It shows that compared with LPM, LRO has stronger constraints for common scenes and can also be adaptable to complex scenes. For the runtime statistics, in most cases, the average runtime of our LRO is less than the other state-of-the-art competitors only except for GMS and LPM, and occasionally over RFM-SCAN. In fact, LRO can further reduce runtime by filtering more unstable feature correspondences when used for real-time LCD tasks.

TABLE III
COMPARATIVE RESULTS ON VARIOUS LCD METHODS WITH DIFFERENT FEATURES AND IMAGE REPRESENTATION FRAMEWORKS

	DBoW3		**VLAD**				**ASMK**			
	SuperPoint		SURF		SuperPoint		SURF		SuperPoint	
Datasets	MR (%)	AUC	MR (%)	AUC	MR (%)	AUC	MR (%)	AUC	MR (%)	AUC
City Centre	80.11	0.9385	67.29	0.9396	83.91	0.9807	70.78	0.9647	**88.38**	**0.9826**
Lip6 Outdoor	57.66	0.8910	75.83	0.9880	81.83	0.9884	71.00	0.9896	**87.33**	**0.9904**
Malaga 6L	43.31	0.8139	61.20	0.8984	64.38	0.9459	75.25	0.9554	**76.29**	**0.9768**
KITTI 00	73.43	0.9446	92.51	0.9743	85.53	0.9522	**96.70**	**0.9983**	96.33	0.9972
KITTI 05	87.55	0.9492	84.93	0.9649	90.61	0.9930	85.15	0.9907	**95.71**	**0.9987**
KITTI 06	96.79	0.9954	85.00	0.9761	91.79	0.9850	96.43	0.9961	**97.49**	**0.9984**

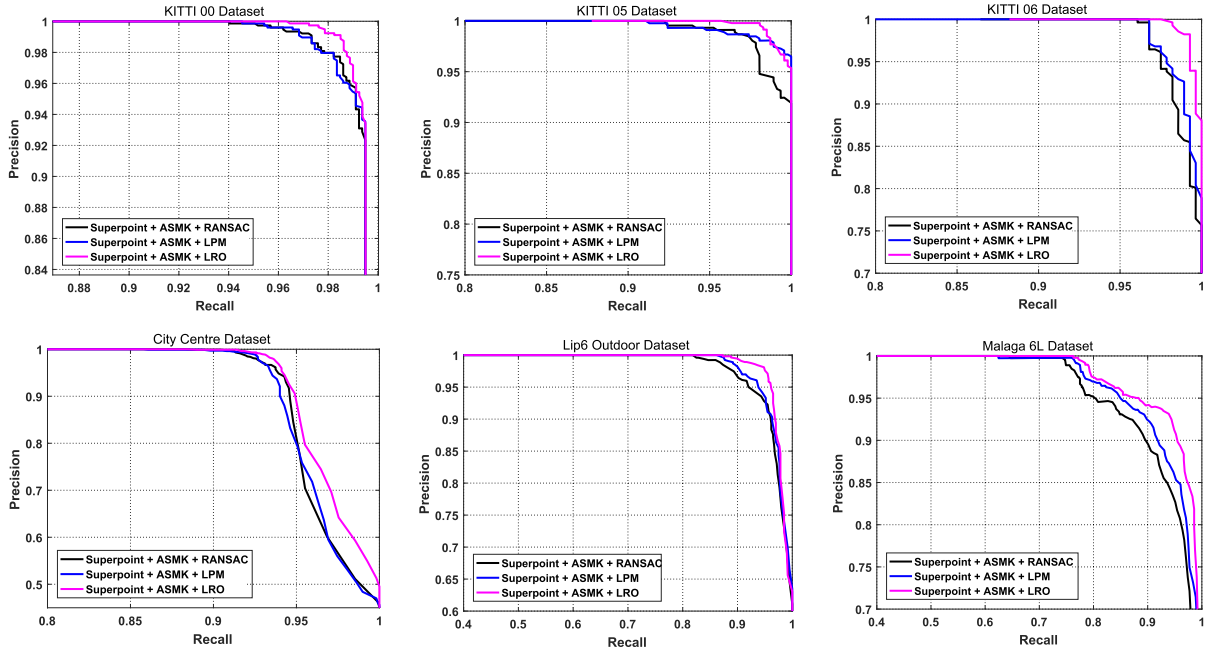

Fig. 10. PR-curves of LCD using different geometric checking methods on the six datasets. From top to bottom, left to right: KITTI 00, KITTI 05, KITTI 06, City Centre, Lip6 Outdoor, and Malaga 2009 Parking 6L.

D. Results on Loop-Closure Detection

1) LCD With Different Feature Descriptors and Image Representation Frameworks: To demonstrate the effectiveness of SuperPoint features and the ASMK framework in our approach, we conduct a comparative experiment for different LCD methods with the same LRO feature matching and summarize the results in Table III. In particular, we compare the LCD performance of the popular BoW model DBoW3, VLAD and our adopted ASMK framework under SuperPoint [17] features and test the LCD performance under VLAD and ASMK with SURF [22] features. The experiment adopts the area under the curve (AUC) of the precision-recall curve and the maximum recall (MR) rate when the precision rate is 100% as the evaluation index for LCD performance. From the table, we see that under the same feature and different retrieval frameworks, the methods that use ASMK have more satisfying results. Moreover, the performances of both VLAD and ASMK are better than that of BoW, because VLAD and ASMK use more accurate visual word vectors so that reduce the effect of quantizing noise. ASMK also outperforms VLAD on five datasets since ASMK employs a selectivity function to focus on more significant matches. Besides, we see that under the same retrieval framework and different features, the methods using SuperPoint features can achieve better LCD performance than SURF on most of the datasets.

2) LCD With Different Feature Matching: In order to confirm the validity of our LRO that is used to verify candidate frames for LCD tasks, we compare the precision-recall curves of three LCD schemes on six datasets in Fig. 10, which include applying three geometric verification algorithms: RANSAC [15], LPM [37], and LRO. From the curves, it can be observed that our LRO can significantly improve the LCD performance compared with other methods on these six datasets. In most cases, RANSAC has equivalent performance with LPM, sometimes over LPM.

3) Qualitative Illustration: The robot's trajectories on five datasets are shown in the black lines of Fig. 11. Temporal evolutions are marked by colored arrows according to the color-bar, while our results on the LCD task are indicated by red circles. In other words, we recognize when robot travels to the same route, and where overlap occurs on the navigated path. Such information can help the robot calibrate

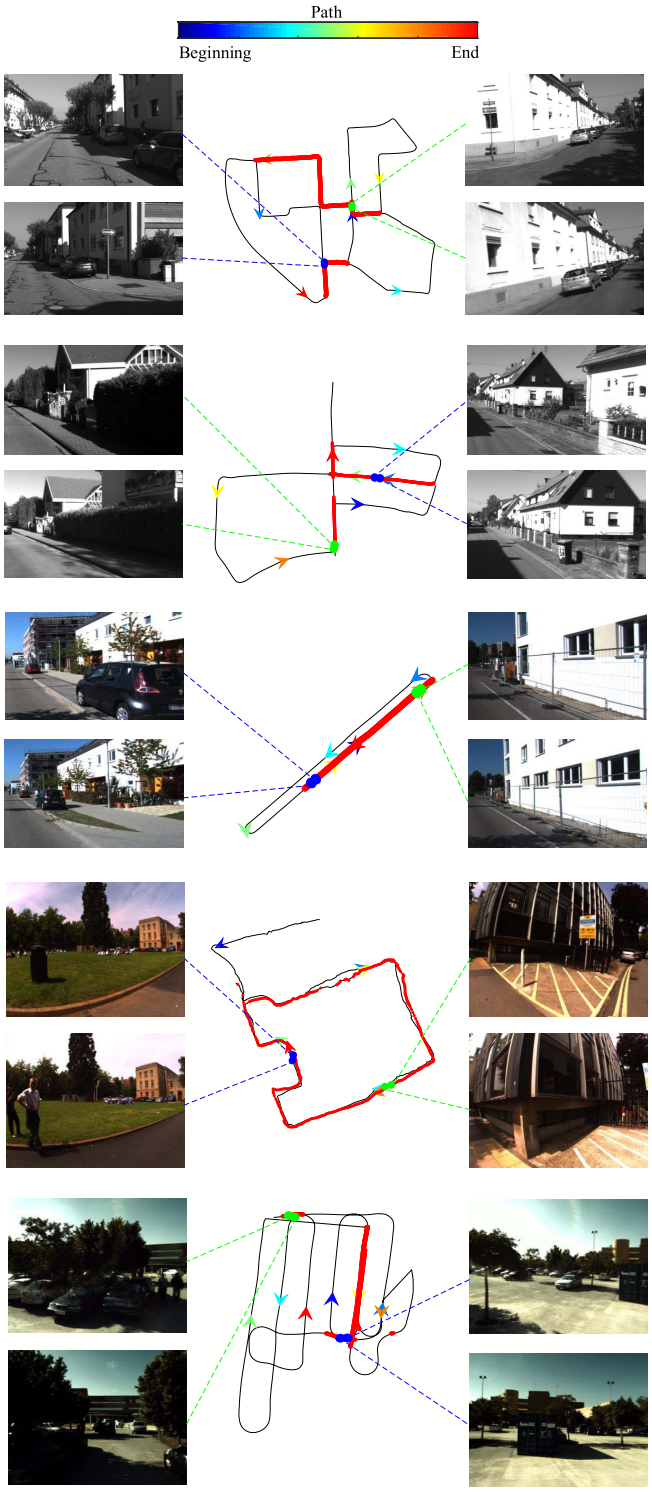

Fig. 11. The qualitative illustration of LCD tasks on five datasets. From top to bottom: KITTI 00, KITTI 05, KITTI 06, City Centre, and Malaga 2009 Parking 6L. In the middle of the figure, the black trajectory is taken from camera poses of datasets, while colored arrows illustrate the temporal evolution according to the color-bar. Red circle marks that a loop-closure event appears, and blue and green points represent true positive examples.

its location and construct a more accurate map about the current environment. As we can see in Fig. 11, our LCD approach can find almost all loop closing events under having no false positive.

TABLE IV

COMPARATIVE RESULTS OF THE MAXIMUM RECALL RATE WITH 100% ACCURACY ON THE SIX DATASETS

Dataset	Approachs	Precision (%)	Recall (%)
KITTI 00 [40]	An *et al.* (2019) [25]	100	91.23
	Gehrig *et al.* (2017) [50]	100	92.00
	Bampis *et al.* (2016) [51]	100	81.54
	Proposed	100	**96.33**
KITTI 05 [40]	An *et al.* (2019) [25]	100	85.15
	Gehrig *et al.* (2017) [50]	100	94.00
	Bampis *et al.* (2016) [51]	100	84.80
	Proposed	100	**95.71**
KITTI 06 [40]	iBoW-LCD (2018) [7]	100	95.53
	SeqSLAM (2012) [52]	100	64.68
	FAB-MAP2.0 (2011) [22]	100	55.34
	Proposed	100	**97.49**
City Centre [43]	PREVIeW (2018) [28]	100	71.14
	Bampis *et al.* (2016) [51]	100	68.49
	FAB-MAP2.0 (2011) [22]	100	38.77
	Proposed	100	**88.38**
Lip6 Outdoor [29]	Tsintotas *et al.* (2019) [53]	100	54.00
	PREVIeW (2018) [28]	100	58.32
	iBoW-LCD(2018) [7]	100	85.24
	Proposed	100	**87.33**
Malaga 6L [42]	An *et al.* (2019) [25]	100	80.54
	Tsintotas *et al.* (2018) [8]	100	**87.99**
	Bampis *et al.* (2016) [51]	100	76.78
	Proposed	100	76.29

4) Quantitative Comparison: If an LCD module provides mistaken information for the SLAM system, the system would have irretrievable performance degradation, especially when the LCD module regards a false loop closing pair as a true match. Meanwhile, an excellent LCD method should detect as many loop closing pairs as possible, which means the recall rate should be high. Therefore, when the precision rate is 100%, the maximum recall rate is a significant index for evaluating the LCD performance.

Table IV shows the comparative results of our LCD method in contrast to the state-of-the-art solutions, such as FAB-MAP 2.0 [22], Bampis *et al.* [51], Gehrig *et al.* [50], PREVieW [28], Tsintotas *et al.* [8], [53], An *et al.* [25] and iBoW-LCD [7]. The result of each method is taken from the original papers. From this table, we can observe that the proposed approach can achieve the highest recall rates for 100% precision on five of the six datasets. In the KITTI datasets, our method outperforms other competitors with over 90% recall. City Centre and Lip6 Outdoor contain many high dynamic outdoor scenes, and their frame rates are low, which make it difficult to detect loop closure. However, the proposed method still exhibits over 80% of recall results in these two datasets. In the case of Malaga 2009 Parking 6L, the system encounters some low texture images and many similar objects like trees and roads, resulting in an unfavorably recall rate compared to the rest algorithms.

In Fig. 12, we report the execution time of each image in the KITTI 00 dataset. When a query image inputs our LCD method, it needs to pass three components: feature extraction using the SuperPoint Network, visual words (VWs) calculation using ASMK, and candidate frames verification using LRO. For these three components, on average, a frame of the KITTI 00 dataset costs 45.7ms, 91.9ms and 35.0ms, which enables our algorithm to operate in real-time.

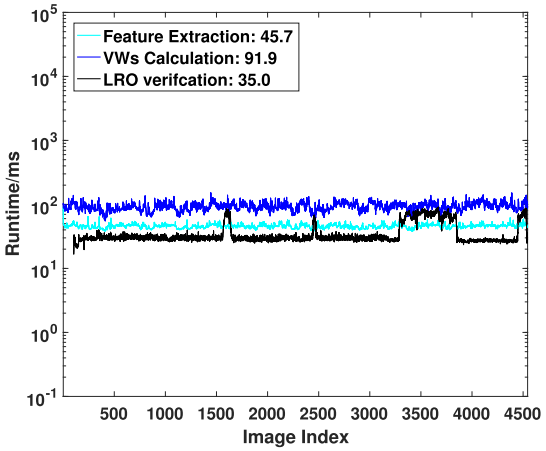

Fig. 12. The computation time of three parts in our LCD method over the KITTI 00 dataset. The legend in the upper left corner shows the average runtime of each part.

In summary, the proposed method can achieve promising results on different datasets, and demonstrates significant robustness against viewpoint and illumination changes, dynamic object occlusions, or complex surroundings, *e.g.*, scenes with many repetitive patterns.

VI. Conclusion

In this paper, we propose a novel approach for appearance-based loop-closure detection. Features extracted by the SuperPoint network are used for the ASMK framework and geometric verification of candidate loop closures. Due to the robustness of SuperPoint features, the ASMK framework can represent images properly and select candidate frames precisely from the database. To further confirm the correctness of candidate frames, we present a new feature matching algorithm called LRO for real-time geometric verification. LRO aims to efficiently establish reliable correspondences between image pairs based on preserving local topological structures, and it can deal with both common scenes and complex scenes, such as involving many repetitive patterns and dynamic objects. Qualitative and quantitative comparative experiments on six publicly available datasets have demonstrated the advantages of our LCD approach over the current state-of-the-art.

References

[1] F. I. Pereira, J. A. Luft, G. Ilha, and A. Susin, "A novel resection–intersection algorithm with fast triangulation applied to monocular visual odometry," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 11, pp. 3584–3593, Nov. 2018.

[2] D.-D. Nguyen, A. Elouardi, S. A. R. Florez, and S. Bouaziz, "HOOFR SLAM system: An embedded vision SLAM algorithm and its hardware-software mapping-based intelligent vehicles applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 11, pp. 4103–4118, Nov. 2019.

[3] Y. Dong *et al.*, "A novel texture-less object oriented visual SLAM system," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 1, pp. 36–49, Jan. 2021, doi: 10.1109/TITS.2019.2952159.

[4] J. Han, J. Kim, and D. H. Shim, "Precise localization and mapping in indoor parking structures via parameterized SLAM," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 12, pp. 4415–4426, Dec. 2019.

[5] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: Part I," *IEEE Robot. Autom. Mag.*, vol. 13, no. 2, pp. 99–110, Jun. 2006.

[6] C.-T. Li and W.-C. Siu, "Fast monocular visual place recognition for non-uniform vehicle speed and varying lighting environment," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1679–1696, Mar. 2021, doi: 10.1109/TITS.2020.2975710.

[7] 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, Oct. 2018.

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

[9] E. Garcia-Fidalgo and A. Ortiz, "Hierarchical place recognition for topological mapping," *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1061–1074, 2017.

[10] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015.

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

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

[13] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2564–2571.

[14] X. Yang and K.-T. Cheng, "Local difference binary for ultrafast and distinctive feature description," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 188–194, Jan. 2014.

[15] 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, Jun. 1981.

[16] H. Yue, J. Miao, Y. Yu, W. Chen, and C. Wen, "Robust loop closure detection based on bag of SuperPoints and graph verification," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 3787–3793.

[17] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 224–236.

[18] 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, Feb. 2016.

[19] C. Siagian and L. Itti, "Biologically inspired mobile robot vision localization," *IEEE Trans. Robot.*, vol. 25, no. 4, pp. 861–873, Aug. 2009.

[20] J. Kosecka, L. Zhou, P. Barber, and Z. Duric, "Qualitative image based localization in indoors environments," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jul. 2003, p. 3.

[21] I. Ulrich and I. Nourbakhsh, "Appearance-based place recognition for topological localization," in *Proc. IEEE Int. Conf. Robot. Automat. Symp. Proc.*, 2005, pp. 1023–1029.

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

[23] X. Jiang, J. Ma, G. Xiao, Z. Shao, and X. Guo, "A review of multimodal image matching: Methods and applications," *Inf. Fusion*, vol. 73, pp. 22–71, 2021.

[24] M. Mohan, D. Galvez-Lopez, C. Monteleoni, and G. Sibley, "Environment selection and hierarchical place recognition," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2015, pp. 5487–5494.

[25] S. An, G. Che, F. Zhou, X. Liu, X. Ma, and Y. Chen, "Fast and incremental loop closure detection using proximity graphs," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 3787–3793.

[26] A. Khaliq, S. Ehsan, Z. Chen, M. Milford, and K. McDonald-Maier, "A holistic visual place recognition approach using lightweight CNNs for significant viewpoint and appearance changes," *IEEE Trans. Robot.*, vol. 36, no. 2, pp. 561–569, Apr. 2020.

[27] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, p. 12.

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

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

[30] Y. Hou, H. Zhang, and S. Zhou, "BoCNF: Efficient image matching with bag of ConvNet features for scalable and robust visual place recognition," *Autom. Robots*, vol. 42, no. 6, pp. 1169–1185, Aug. 2018.
[31] H. Jegou, M. Douze, C. Schmid, and P. Perez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3304–3311.
[32] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the Fisher vector: Theory and practice," *Int. J. Comput. Vis.*, vol. 105, no. 3, pp. 222–245, Dec. 2013.
[33] 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. (CVPR)*, Jun. 2016, pp. 5297–5307.
[34] S. Lowry *et al.*, "Visual place recognition: A survey," *IEEE Trans. Robot.*, vol. 32, no. 1, pp. 1–19, Feb. 2016.
[35] 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. 7, pp. 1–57, Jul. 2021.
[36] 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. (CVPR)*, Jun. 2020, pp. 4938–4947.
[37] 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, May 2019.
[38] 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," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4181–4190.
[39] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Commun. ACM*, vol. 18, no. 9, pp. 509–517, Sep. 1975.
[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.*, Jun. 2012, pp. 3354–3361.
[41] R. Arroyo, P. F. Alcantarilla, L. M. Bergasa, J. J. Yebes, and S. Bronte, "Fast and effective visual place recognition using binary codes and disparity information," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2014, pp. 3089–3094.
[42] 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, Nov. 2009.
[43] M. Cummins and P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance," *Int. J. Robot. Res.*, vol. 27, no. 6, pp. 647–665, Jun. 2008.
[44] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
[45] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[46] X. Li and Z. Hu, "Rejecting mismatches by correspondence function," *Int. J. Comput. Vis.*, vol. 89, no. 1, pp. 1–17, Aug. 2010.
[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, 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, Aug. 2019.
[49] K. Mikolajczyk *et al.*, "A comparison of affine region detectors," *Int. J. Comput. Vis.*, vol. 65, nos. 1–2, pp. 43–72, Nov. 2005.
[50] M. Gehrig, E. Stumm, T. Hinzmann, and R. Siegwart, "Visual place recognition with probabilistic voting," in *Proc. IEEE Int. Conf. Robot. Autom.*, Sep. 2017, pp. 3192–3199.
[51] 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. (IROS)*, Oct. 2016, pp. 4530–4536.
[52] M. J. Milford and G. F. Wyeth, "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 1643–1649.
[53] 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, Apr. 2019.

Jiayi Ma (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, Wuhan, China, in 2008 and 2014, respectively. He is currently a Professor with the Electronic Information School, Wuhan University. He has authored or coauthored more than 150 refereed journals and conference papers, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, *IJCV*, CVPR, ICCV, and ECCV. His research interests include computer vision, machine learning, and robotics. He has been identified in the 2020 and 2019 Highly Cited Researcher lists from the Web of Science Group. He is an Area Editor of *Information Fusion*, an Associate Editor of *Neurocomputing*, *Sensors* and *Entropy*, and a Guest Editor of *Remote Sensing*.

Xinyu Ye received the B.S. degree from the Electronic Information School, Wuhan University, Wuhan, China, in 2020. She is currently pursuing the Ph.D. degree with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University. Her current research interests include robotics, computer vision, and machine learning.

Huabing Zhou (Member, IEEE) received the B.S. and M.S. degrees in computer science and technology from the Wuhan Institute of Technology, Wuhan, China, in 2005 and 2008, respectively, and the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology, Wuhan, in 2012.

From 2009 to 2010, he was a Research Intern with the Chinese Academy of Surveying and Mapping. From 2018 to 2019, he was a Visiting Scholar with Temple University, Philadelphia, PA, USA. He is currently an Associate Professor with the School of Computer Science and Engineering, Wuhan Institute of Technology. His research interests include computer vision, remote sensing image analysis, and intelligent robot.

Xiaoguang Mei (Member, IEEE) received the B.S. degree in communication engineering from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2007, the M.S. degree in communications and information systems from Huazhong Normal University, Wuhan, in 2011, and the Ph.D. degree in circuits and systems from HUST, in 2016. From 2010 to 2012, he was a Software Engineer with the 722 Research Institute, China Shipbuilding Industry Corporation, Wuhan. He is currently an Associate Professor with the Electronic Information School, Wuhan University. His research interests include hyperspectral imagery, machine learning, and pattern recognition.

Fan Fan received the B.S. degree in communication engineering and the Ph.D. degree in electronic circuit and system from the Huazhong University of Science and Technology, Wuhan, China, in 2009 and 2015, respectively. He is currently an Assistant Professor with the Electronic Information School, Wuhan University, China. His current research interests include infrared thermal imaging, machine learning, and computer vision.