Pattern Recognition 126 (2022) 108588

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Robust image matching via local graph structure consensus

Xingyu Jiang[a,1], Yifan Xia[a,1], Xiao-Ping Zhang[b], Jiayi Ma[a,*]

[a] Electronic Information School, Wuhan University, Wuhan 430072, China
[b] Department of Electrical, Computer, and Biomedical Engineering, Ryerson University, Toronto, ON M5B 2K3, Canada

ARTICLE INFO

Article history:
Received 19 August 2021
Revised 25 November 2021
Accepted 11 February 2022
Available online 13 February 2022

Keywords:
Image matching
Feature matching
Mismatch removal
Outlier
Image registration

ABSTRACT

Image matching plays a vital role in many computer vision tasks, and this paper focuses on the mismatch removal problem of feature-based matching. We formulate the problem into a general yet effective optimization framework based on graph matching by combining integer quadratic programming with a compensation term for discouraging matches, termed as *Local Graph Structure Consensus* (LGSC). Considering the local area similarity of those potential true matches, we design a local graph structure for preserving geometric topology, which contains a local indicator vector and a local affinity vector for each correspondence. The local indicator vector is utilized for edge construction, while the local affinity vector represents the match correctness of the nodes and edges between two graphs. In particular, the ranking shift with scale and rotation invariance is exploited to represent the node affinity. Ultimately, we derive a closed-form solution with linearithmic time and linear space complexity. Moreover, a multi-scale and iterative graph construction strategy is proposed to promote the performance of our method in terms of robustness and effectiveness. Extensive experiments on various real image datasets demonstrate that our LGSC can achieve superior performance over current state-of-the-art approaches.

1. Introduction

As a fundamental and important problem in vision-based applications, image matching attempts to identify then construct reliable correspondences between two images of the same object or scene. Relevant computer vision tasks include image registration [1,2] and fusion [3], 3D reconstruction [4], content-based image retrieval [5], panoramic image mosaic [6], object identification and tracking [7], simultaneous localization and mapping (SLAM) [8], etc. A robust and efficient feature matching algorithm can give full guarantee and support to the performance of these applications.

Regarding the combinatorial nature of feature matching, there is a high computation complexity problem in non-Pareto criterion complex optimization. Without considering outliers, even a simple problem of matching N points to another N points is accompanied by $N!$ permutations [9]. Currently point set registration and graph matching are devised to solve this problem, where the former aims to estimate the spatial transformation like the iterative closed point (ICP) method [10] and nonrigid registration method using the thin-plate spline (TPS-RPM) [11], and the latter commonly solves for correspondences by constructing an integer programming problem (IQP). Most graph matching methods provide approximate solutions by relaxing the stringent constraints, such as binary constraint preserving graph matching [12], common visual pattern discovery via spatially coherent correspondences (GS) [13], and recent graph learning-matching [14]. However, these direct matching methods are prone to suffer from a high computational burden and unstable matching performance.

To resolve the above issues, a common strategy of existing methods is an indirect two-step matching approach. Firstly construct relatively reliable putative matches depending on the similarity of feature descriptors, typical ones of which include scale-invariant feature transform (SIFT) [15], oriented FAST and rotated BRIEF (ORB) [16], and speeded up robust features (SURF) [17]. Nevertheless, the use of only local feature descriptors inevitably tends to cause a large of mismatches (i.e., outliers) in the image pairs. Therefore after that, extra local and/or global geometrical constraints ought to be imposed to identify correct matches (i.e., inliers) and reject outliers in the putative match set. Although recently proposed end-to-end learning-based methods [18,19] can construct correct correspondences from raw image pairs, wherein the inlier number and inlier rate are both extremely high thus largely outperforming traditional SIFT, an advanced outlier rejection method is still of great significant in real applications. On the one hand, recently proposed deep feature matchers may have obvious limitations in terms of generality, and would also create a high number and ratio of outliers even fail to address those 'un-

* Corresponding author.
E-mail addresses: jiangx.y@whu.edu.cn (X. Jiang), xiayifan@whu.edu.cn (Y. Xia), xzhang@ee.ryerson.ca (X.-P. Zhang), jyma2010@gmail.com (J. Ma).
[1] Authors contributed equally.

https://doi.org/10.1016/j.patcog.2022.108588
0031-3203/

seen' data, such as multi-model images [20]. On the other hand, these deep methods inherently include a strategy of mismatch removal and geometric verification, thus studying an effective mismatch removal method can both further improve their matching performance and provide useful theoretical guarantees for future research.

To this end, there are multiple algorithms designed. The most representative methods would be random sample consensus (RANSAC) [21], which tries to find a smallest inlier set to fit a given model under a hypothesize-and-verify paradigm and random resampling strategy. A large number of follow-ups have improved plain RANSAC in terms of both efficiency and accuracy [22–25]. To be specific, a universal framework in this pipeline, namely USAC [23], combines multiple advancements together into a unified framework, and shows superior performance. MAGSAC++ [25] applies δ-consensus with a new scoring function and has improvements in speed and robustness. But these methods experience an exponential increase in runtime as the outliers grows, and even fail when applied to address non-rigid cases. This can be mitigated by some specifically designed non-rigid methods, such as identifying point correspondences by correspondence function (ICF) [26], or performing on the reproducing kernel Hilbert space with Tikhonov regulation like vector field consensus (VFC) [27] and locally linear transforming [28]. These methods have shown satisfying performance in feature matching of arbitrary geometrical deformations. But they still affect a lot in case of wide-baseline image pairs or multiple independent motions in image scenes, since the slow-and-smooth assumptions they used would be violated to some extent. Besides, there has been a marked development in the relaxed methods, which impose less strict yet general geometric constraints for various scenarios. The representative ones of such methods include considering piece-wise affine transformation then seeking for deformation consistency [29], likelihood function estimation [30], local preserving consensus [31] with its improved methods like [32,33], grid-based motion statistics (GMS) [34], and clustering to achieve motion-consistent clusters [35–37]. However, these methods are quite efficient but highly rely on a good initialization to assure the accurate construction of local structure, thus showing coarse in their results.

Recently, the learning manner is increasingly being used in inlier and outlier classification, which is used to substitute traditional methods in correspondence testing or model regression. Aiming to label correspondences as inliers or outliers, learning to find good correspondences (LFGC) [38] trains a network containing imaging intrinsic with epipolar geometric constraint while outputting camera motion parameters. Nonetheless, its effectiveness needs to be improved in some matching dilemmas and it tends to sacrifice correct matches to estimate the motion parameters. Various modifications of LFGC have been proposed to improve the performance, such as using an order-aware strategy [39], attentive context normalization [40] or graph attention strategy [41]. Innovatively, a two-class classifier for mismatch removal called LMR [42] has been proposed, which exhibits decent matching performance in linearithmic time complexity. However, it may still be confined to challenging data or scenarios due to its limited matching expression. Learning-based methods have shown great potential for geometric estimation and feature correspondence classification with deep learning-based modules.

Although with significant development in the past few years, the mismatch removal problem still expects an accurate, robust, and efficient algorithm for practical use, whereby challenges mainly exist on the following three aspects. Firstly, the geometric transformation between an image pair is usually unknown in advance, and hence a general algorithm that can deal with various transformations is urgently desired. Secondly, the feature matching problem in real computer vision tasks typically suffers from object deformation, low-quality imaging, repeated structure, etc., which results in a large number of mismatches thus increasing the difficulty of establishing accurate correspondences. Thirdly, image pairs for practical applications are commonly not simply parametric model transformations, but rather complex nonrigid models, prone to high computational burden and poor matching results.

To address the above issues, this paper proposes a robust and effective image feature matching algorithm, termed as *Local Graph Structure Consensus* (LGSC), which is capable of removing mismatches from putative feature correspondence set and does not rely on any predefined transformation model. First and foremost, two graphs are constructed according to an image pair with a putative feature set. With a compensation term for discouraging matches, the mismatch removal is expressed by a global objective function to be maximized. We observe that for an image pair of the same object or scene, under scaling and rotation, or even nonrigid deformation, the local neighborhood structures of true correspondences are generally well preserved, as shown in Fig. 1. As a result, we design a local graph structure to represent the local topology information and hence resolve the objective on the principle of preserving the consensus of the local graph structure. The tolerance of this model allows it to cope with rigid and nonrigid transformations even when the images encounter severe deformations. Furthermore, we derive a simple closed-form solution with linearithmic time and linear space complexity concerning the scale of the putative set. It is demonstrated by qualitative and quantitative experiments that our LGSC has satisfactory properties in terms of robustness to large geometrical deformations, universality to various descriptors, and convenience in applications.

Concretely, our contributions in this work include the following three aspects:

- We introduce a general yet powerful mathematical optimization model based on graph matching for robust image matching. Compared with most existing methods, our method exploiting the consistency of local geometric information does not require a specific parametric or nonparametric model to resolve global image transformation, thus is robust for various image transformations including complex nonrigid.
- We derive a closed-form solution from the perspective of the consensus of local graph structure, which has linearithmic time and linear space complexity, and hence ensures the efficiency of our method in feature matching.
- We further propose a multi-scale and iterative strategy for local graph construction, which enables our method to efficiently filter out mismatches and retain correct matches with high robustness.

The remainder of this paper is organized as follows. In Section 2, our LGSC for robust feature matching is described in detail. Subsequently, we present various qualitative and quantitative comparison experiments on feature matching and vision-based applications in Section 3. Finally, Section 4 states some concluding remarks of this paper.

2. Method

This section will describe our proposed algorithm for robust image feature matching in detail. In the first place, a set of putative matches is constructed based on the similarity of feature descriptors (e.g., SIFT [15]), where matches whose descriptor vectors are apparently different have been filtered out. Subsequently, extra geometric constraints are required to remove mismatches included in the putative set. In the following, our focus is on the mismatch removal stage, with our method based on the consensus of the local graph structure.

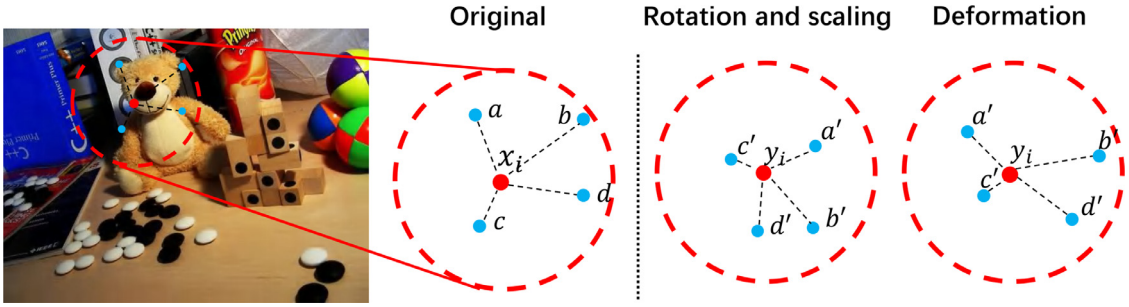

Fig. 1. Consensus of local topology structure under scaling, rotation and nonrigid deformation.

2.1. Problem formulation

Given a set of N putative image feature point correspondences $T = \{(\boldsymbol{y}_i, \boldsymbol{z}_i)\}_{i=1}^{N}$, where $\boldsymbol{y}_i$ and $\boldsymbol{z}_i$ are attribute vectors indicating the spatial position of a pair of corresponding feature points, which apply to both 2D and 3D matching problems. Our task is to remove mismatches from putative correspondence set T.

For two graphs formed by two feature point sets of an image pair, we introduce an objective S, which will be maximized to obtain the optimal solution of the unknown inlier set $\mathcal{I}$, i.e.,

$$\mathcal{I}^* = \arg\max_{\mathcal{I}} S(\mathcal{I}; T, \lambda), \tag{1}$$

with the objective S defined as follows:

$$S(\mathcal{I}; T, \lambda) = \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} + \lambda (N - |\mathcal{I}|), \tag{2}$$

where the first term is the graph matching representation IQP for inliers, composed by an indicator vector $\boldsymbol{x} \in \{0,1\}^{N\times 1}$ denoting the correspondence solution of $\mathcal{I}$ and an affinity matrix $\boldsymbol{W}$ describing the matching degree of nodes and edges. The second compensation term discourages matches and can avoid the trivial solution of all-inlier (i.e., $\boldsymbol{x} = \mathbf{1}^{N\times 1}$), where parameter $\lambda > 0$ controls the trade-off with the first term, and $|\cdot|$ indicates the cardinality of a set.

To define the affinity matrix $\boldsymbol{W}$, we construct two attributed relation graphs $G(V, E)$ and $G'(V', E')$ based on putative feature correspondence set T, where each node $v_i \in V$ refers to point $\boldsymbol{y}_i \in T$ with $v'_i \in V'$ for point $\boldsymbol{z}_i \in T$ similarly, and each edge $e_{ij} = (v_i, v_j) \in E$ connects node v_i to node v_j in graph G with $e'_{ij} = (v'_i, v'_j) \in E'$ connecting node v'_i to node v'_j in graph G'. Different from the one-to-one or many-to-one programming problem in graph matching, our ultimate goal is to figure out the inliers from $\{(v_i, v'_i)\}_{i=1}^{N}$.

For each node correspondence (v_i, v'_i), a node affinity score $s(v_i, v'_i)$ is defined to measure how well node v_i matches node v'_i. Likewise, for each correspondence pair (v_i, v'_i) and (v_j, v'_j), we define an edge affinity score $s(e_{ij}, e'_{ij})$ that represents how compatible the edge $e_{ij} = (v_i, v_j) \in E$ is with the edge $e'_{ij} = (v'_i, v'_j) \in E'$. Consequently, an affinity matrix $\boldsymbol{W}$ can be constructed, in which the diagonal term W_{ii} represents the node affinity score $s(v_i, v'_i)$ and the non-diagonal term W_{ij} represents the edge affinity score $s(e_{ij}, e'_{ij})$. That is

$$W_{ii} = s(v_i, v'_i), \quad W_{ij} = s(e_{ij}, e'_{ij}). \tag{3}$$

A binary vector $\boldsymbol{x}$ is determined as the solution of this matching problem, denoted by $\boldsymbol{x} = (x_1, \ldots, x_i, \ldots, x_N)^\top$, $x_i \in \{0,1\}$, and $x_i = 1$ signifies that correspondence (v_i, v'_i) is classified as an inlier and vice versa for an outlier. Clearly, $|\mathcal{I}| = \boldsymbol{x}^\top \boldsymbol{x}$, and hence we derive and obtain a concise formulation of this mismatch removal problem, i.e.,

$$\begin{aligned} S(\boldsymbol{x}; T, \lambda) &= \boldsymbol{x}^\top (\boldsymbol{W} - \lambda \boldsymbol{I}) \boldsymbol{x} + \lambda N \\ &\Leftrightarrow \boldsymbol{x}^\top (\boldsymbol{W} - \lambda \boldsymbol{I}) \boldsymbol{x} \end{aligned} \tag{4}$$

where λN is a predetermined constant term that can be omitted, and $\boldsymbol{I}$ is the N-dimensional identity matrix.

We disassemble the objective to innovatively solve for each node correspondence, and obtain:

$$\widetilde{S}(\boldsymbol{x}; T, \lambda) = \sum_{i=1}^{N} x_i \left(\boldsymbol{w}_i^\top \boldsymbol{x} - \lambda\right), \tag{5}$$

where $\boldsymbol{w}_i$ is the ith column of $\boldsymbol{W}$.

2.1.1. Local graph structure

Based on the observation that the local topology structure of an inlier will get well preserved even if under nonrigid transformation, we devise a novel local graph structure describing topology information to solve for each correspondence. First and foremost, we construct the neighborhood by searching K nearest neighbors (K-NN) under Euclidean distance, which is an efficient distance measurement and is enough to be used in feature matching problem. Subsequently, two local graphs are constructed by imposing physical constraints on the K neighborhoods. As a consequence, the objective $\widetilde{S}$ is transformed into $\widetilde{S}_K$, i.e.,

$$\widetilde{S}_K(\boldsymbol{x}; T, \lambda) = \sum_{i=1}^{N} x_i \left(\boldsymbol{w}_i^{K\top} \boldsymbol{x}_i^K - \lambda\right), \tag{6}$$

where $\boldsymbol{w}_i^K$ and $\boldsymbol{x}_i^K$ describe local graph structure centered on correspondence (v_i, v'_i), and they are termed as the local affinity vector and the local indicator vector, respectively.

Local indicator vector $\boldsymbol{x}_i^K$ is constructed as a binary vector for each correspondence (v_i, v'_i), which denotes the edge construction of local graphs. With $\boldsymbol{x}_i^K \in \{0,1\}^{(K+1)\times 1}$, its first term is set to 1 and other K items indicate whether the neighbors are connected to the center node v_i thus forming the graph edges. In other words, $x_{i_j}^K = 1$ in Eq. (7) denotes that two edges from two corresponding neighbors (v_{i_j}, v'_{i_j}) to their center correspondence (v_i, v'_i) are constructed in two local graphs G_{v_i} and $G_{v'_i}$, respectively.

On the purpose of maximizing the consensus of local topology structure, indicator vector $\boldsymbol{x}_i^K$ is determined by:

$$\begin{aligned} &\boldsymbol{x}_i^K = \{1, x_{i_1} \ldots, x_{i_j} \ldots, x_{i_K}\}^\top \\ &\text{s.t. } v_{i_j} \in \mathcal{N}_{v_i}^K, \quad j = 1, \ldots, K, \end{aligned} \tag{7}$$

where

$$x_{i_j} = \begin{cases} 1, & v'_{i_j} \in \mathcal{N}_{v'_i}^K, \\ 0, & \text{otherwise}, \end{cases} \tag{8}$$

and $\mathcal{N}_{v_i}^K$ denotes the K-nearest neighbors of node v_i. That is to say, common corresponding elements in the two neighborhoods $\mathcal{N}_{v_i}^K$ and $\mathcal{N}_{v'_i}^K$ are identified by local indicator vector, thus determining the edges of two local graph structures G_{v_i} and $G_{v'_i}$.

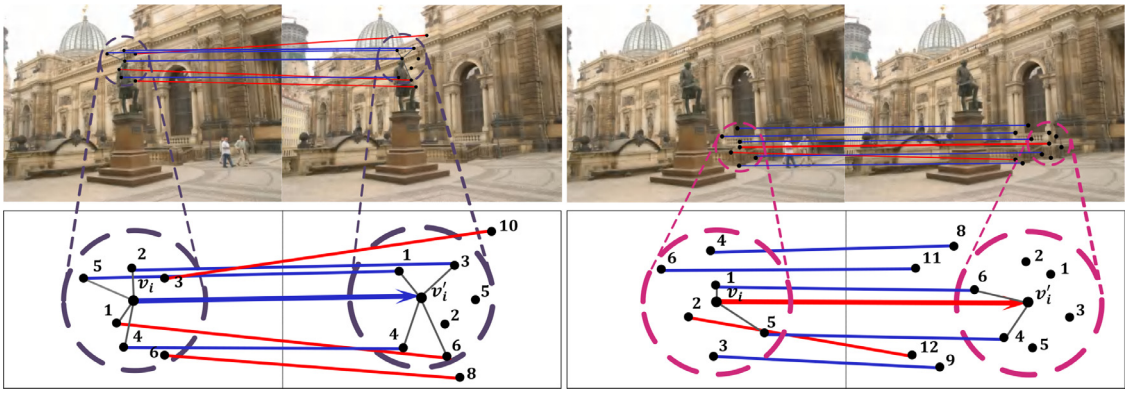

Fig. 2. Schematic of the consensus of the local graph structure. The given correspondence (v_i, v_i') is an inlier in the left group and an outlier in the right group. In each group, the top figure illustrates the putative correspondence (v_i, v_i') with its K neighbors, and the bottom figure shows its local graph structure with gray edges and black nodes. The circles denote the K ($K = 6$) neighborhood and the number around each node represents its distance ranking relative to its center node. Blue: inlier; Red: outlier. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

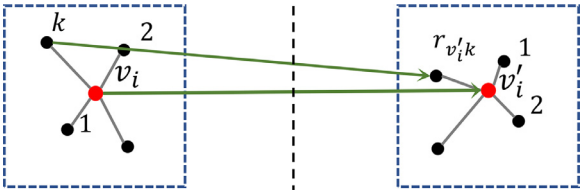

Fig. 3. Schematic of the ranking shift about correspondence (v_i, v_i'). The numbers and symbols around the neighbors denote the ranking numbers relative to the center node.

Local affinity vector A local affinity vector $\boldsymbol{w}_i^K$ contains a node affinity score and K edge affinity scores, and is the combination of $K+1$ non-zero entries of $\boldsymbol{w}_i$, i.e.,

$$\boldsymbol{w}_i^K = \{s(v_i, v_i'), s(e_{ii_1}, e'_{ii_1}) \ldots, s(e_{ii_j}, e'_{ii_j}) \ldots, s(e_{ii_K}, e'_{ii_K})\}^\top, \quad \text{s.t. } v_{i_j} \in \mathcal{N}_{v_i}^K, \quad j = 1, \ldots, K. \tag{9}$$

Next, we consider the definitions of the node affinity score $s(v_i, v_i')$ and the edge affinity score $s(e_{ii_j}, e'_{ii_j})$.

The node affinity score $s(v_i, v_i')$ represents the consistency between each correspondence (v_i, v_i'). It is not sufficient to consider only the preservation of intersections within the neighborhoods, so we use the ranking structure to measure the matching degree between two corresponding nodes, because ranking information is better suited for the mismatch removal problem compared to traditional descriptors.

We observe that the *ranking shift* (i.e., the change of rankings for the neighboring points in $\mathcal{N}_{v_i}^K$ and $\mathcal{N}_{v_i'}^K$) of correct matches is surprisingly robust compared to mismatches, as illustrated in Fig. 2, which means little change in relative position for K neighbors of correct correspondences. For example, neighbors of node v_i tend to remain around it after image transformation in case $(v_i, v_i') \in \mathcal{I}$. Therefore, we use the ranking shift to represent the node affinity score.

To streamline the arithmetic, we use a binary sequence $\boldsymbol{\phi}_{v_i}$ to quantitatively represent the ranking shift of neighbors around node v_i. The kth term of $\boldsymbol{\phi}_{v_i}$ (i.e., $\phi_{v_i k}$) indicates the ranking shift of the kth neighbor. For instance, in the left image of an image pair, node v_j is the kth ($k \le K$) neighbor of node v_i, and the ranking of its corresponding node v_j' centered on v_i' is denoted by $r_{v_i' k}$, as shown in Fig. 3, then $\phi_{v_i k}$ is defined as below:

$$\phi_{v_i k} = \begin{cases} 1, & r_{v_i' k} > k, \\ 0, & r_{v_i' k} \le k. \end{cases} \tag{10}$$

The ranking shift should be symmetrical between an image pair, as left-to-right has equivalent significance to right-to-left, and hence it is necessary to calculate $\boldsymbol{\phi}_{v_i'}$. As a consequence, the node affinity score can be defined as follows:

$$s(v_i, v_i') = 1 - \frac{1}{2K} \left\| \boldsymbol{\phi}_{v_i} + \boldsymbol{\phi}_{v_i'} \right\|_1, \tag{11}$$

where $\|\cdot\|_1$ is ℓ_1-norm, and $\frac{1}{2K}$ is used for normalization.

As for the edge affinity score, $s(e_{ii_j}, e'_{ii_j})$ literally aims to denote the affinity between two corresponding edges $e_{ii_j} = (v_i, v_{i_j})$ and $e'_{ii_j} = (v_i', v'_{i_j})$ in two local graphs G_{v_i} and $G_{v_i'}$. The node affinity score introduced above makes use of the ranking shift, but has not fully exploit the local geometrical information. Hence, the edge affinity is required, and represented by the distance differences, i.e.,

$$s(e_{ii_j}, e'_{ii_j}) = \frac{1}{K} \exp\left(- \frac{\left| d(e_{ii_j}) - d(e'_{ii_j}) \right|}{\max\left(d(e_{ii_j}), d(e'_{ii_j}) \right)} \right), \tag{12}$$

where coefficient $\frac{1}{K}$ is set to balance K edge affinity scores with the above node affinity score, $d(\cdot)$ implies the distance metric, such as Euclidean distance, i.e., $d(e_{ii_j}) = \|\boldsymbol{y}_i - \boldsymbol{y}_{i_j}\|_2$.

It is worth noting that, our ranking shift strategy that measures the similarity between two local graph structure can be regarded as a relaxed form of subgraph isomorphism. In particular, subgraph isomorphism is a classical topic in graph theory, which is a general definition to measure if there exists a sub- or pattern graph in the target that has the same attributes in nodes and edges as given graph. While our proposed ranking shift is designed for feature matching problem by measuring the similarity. The ranking shift has excellent geometrical properties of scale and angle invariance and sufficiently exploits local topology in the coordinate space, thus is more suitable and flexible for handling putative matching set and screening feature correspondence.

2.1.2. Multi-scale strategy

In our formulation, the local graph structure is obtained from the neighborhood based on K-NN. However, a fixed K inevitably suffers from the following two problems: i) different feature correspondence sets generally have different inlier ratios and numbers, and hence the degree of correlation between nodes varies, and ii) the distribution of inliers and outliers is usually not uniform for all local graphs even in a single feature correspondence set.

To address the above issues, we adopt a multi-scale strategy, where local graphs are constructed under multiple scales with a

set of $\boldsymbol{K} = \{K_m\}_{m=1}^{M}$. In this case, $\mathcal{N}_{v_i}^{K_m}$ denotes the neighborhood structure centered on node v_i by searching K_m nearest neighbors, and $G_{v_i}^{K_m}$ is the corresponding local graph structure. Moreover, the objective in Eq. (6) becomes:

$$\begin{aligned}\widetilde{S}(\boldsymbol{x}; T, \lambda) &= \frac{1}{M}\sum_{m=1}^{M}\widetilde{S}_{K_m}(\boldsymbol{x}; T, \lambda) \\ &= \frac{1}{M}\sum_{m=1}^{M}\sum_{i=1}^{N} x_i\left(\boldsymbol{w}_i^{K_m\top}\boldsymbol{x}_i^{K_m} - \lambda\right),\end{aligned} \tag{13}$$

where $1/M$ acts on normalizing the contribution of local graphs with various levels. Apparently, the final objective function has outstanding properties in terms of scale and rotation invariance as well as no dependence on the image transformation model.

2.2. Solution

To resolve the objective function Eq. (13), its form has been readjusted, i.e.,

$$\widetilde{S}(\boldsymbol{x}; T, \lambda) = \sum_{i=1}^{N} x_i(s_i - \lambda), \tag{14}$$

where

$$s_i = \frac{1}{M}\sum_{m=1}^{M}\boldsymbol{w}_i^{Km\,T}\boldsymbol{x}_i^{K_m} \tag{15}$$

represents how well the ith correspondence (v_i, v_i') satisfies the consensus of local graph structure. Generally, the potential true matches have a high value s_i and thus capable of producing substantial contributions to the objective function while mismatches have a low value s_i and are therefore discouraged.

In the mismatch removal problem, with a given set of N feature correspondences, $\{s_i\}_{i=1}^{N}$ can be determined in advance. Since λ is predefined, the only variable that needs to be resolved in Eq. (14) is $\boldsymbol{x} = \{x_i\}_{i=1}^{N}$. Clearly, a correspondence with a value of s_i larger than λ will produce a positive item thus increasing the objective function, while a value of s_i smaller than λ will produce a negative item. Accordingly, to maximize Eq. (14), the optimal solution of $\boldsymbol{x}$ can be formulated as follows:

$$x_i = \begin{cases} 1, & s_i \geq \lambda, \\ 0, & s_i < \lambda, \end{cases} \quad i = 1, \ldots, N. \tag{16}$$

And then the optimal inlier set $\mathcal{I}^*$ is determined by the following formula:

$$\mathcal{I}^* = \{i | x_i = 1, i = 1, \ldots, N\}. \tag{17}$$

According to Eq. (16), parameter λ has the ability to evaluate the match correctness of each correspondence as a threshold. Hence, the objective in Eqs. (1) and (2) have been reduced to a binary classification problem for each correspondence, and a closed-form solution is derived.

Furthermore, to verify how well the multi-scale strategy works, we randomly collect a total of 30 image pairs as a test set, which is from various types of transformation, including piece linear, wide baseline, nonrigid, etc. The average inlier number and inlier percentage for total test image sets after SIFT matching are 425 and 59.8%, respectively. The metric for evaluating the performance of our method is F1-score, which is defined as F1-score $=$ $2 \times \text{Precision} \times \text{Recall}/(\text{Precision} + \text{Recall})$, where Precision is defined as the ratio of identified correct match number and the total retained match number, and Recall is defined as the ratio between identified correct match number and the number of actual correct matches. The F1-score performance of our method is reported in Fig. 4, which clearly demonstrates the effectiveness of the multi-scale strategy.

2.3. Iterative graph construction

The local graph structures $\{(G_{v_i}^{K_m}, G_{v_i'}^{K_m})\}_{m=1,i=1}^{M,N}$ are constructed based on the neighborhoods $\{(N_{v_i}^{K_m}, N_{v_i'}^{K_m})\}_{m=1,i=1}^{M,N}$, and the neighborhoods are constructed by the whole putative correspondence set that involves a certain percentage of outliers. Clearly, this may produce poor reference values for the similarity measurement of node affinity and edge affinity. Ideally, the local graph structures are established from the neighborhoods formed by only the inlier set $\mathcal{I}$.

Based on the fact that it is impossible to obtain the true inlier set $\mathcal{I}$ beforehand, we propose an iterative strategy for local graph construction to enhance the effectiveness of mismatch removal. Although with only the whole feature set at the beginning, after our objective maximization, the generated correspondence set could be considered as a desirable approximation of the true inlier set $\mathcal{I}$, named as $\mathcal{I}_1$, i.e., $\mathcal{I}_1 = \arg\max_{\mathcal{I}} \widetilde{S}(\mathcal{I}; T, \lambda)$, where the local graphs are constructed by the whole feature set T.

Subsequently, $\mathcal{I}_1$ is used to construct the neighborhoods thus local graph structures for each correspondence in T. In Fig. 5, the distributions of s_i for inliers and outliers are displayed based on the local graphs formed by $\mathcal{I}_0 = T$ and iterated set $\mathcal{I}_1$. Apparently, the iterative graph construction strategy contributes to distinguishing the outliers and inliers. Hence the optimal solution $\mathcal{I}^*$ is resolved, i.e.,

$$\mathcal{I}^* = \arg\max_{\mathcal{I}} \widetilde{S}(\mathcal{I}; \mathcal{I}_1, T, \lambda), \tag{18}$$

which can be seen as an iterative manner to gradually approach the optimal solution. In our experiments, we find that two iterations are empirically enough.

Considering that our feature matching algorithm is customized by the consensus of local graph structure, so we name it *Local Graph Structure Consensus*. The complete process is summarized in Algorithm 1 .

Algorithm 1: The LGSC algorithm.

Input: Putative set $T = \{(\boldsymbol{y}_i, \boldsymbol{z}_i)\}_{i=1}^{N}$, parameters $\boldsymbol{K}$, λ
Output: Inlier set $\mathcal{I}^*$

1. Construct $\{\mathcal{N}_{v_i}^{K_m}, \mathcal{N}_{v_i'}^{K_m}\}_{m=1,i=1}^{M,N}$ from T by K-NN;
2. Compute indicator vectors $\{\boldsymbol{x}_i^{K_m}\}_{m=1,i=1}^{M,N}$ by Eqs.~(7), (8) to establish local graphs $\{G_{v_i}^{K_m}, G_{v_i'}^{K_m}\}_{m=1,i=1}^{M,N}$;
3. Compute affinity vectors $\{\boldsymbol{w}_i^{K_m}\}_{m=1,i=1}^{M,N}$ using Eqs.~(11), (12) and (9);
4. Obtain inlier set $\mathcal{I}_1$ from Eqs.~(15), (16) and (17);
5. Construct neighborhood $\{\mathcal{N}_{v_i}^{K_m}, \mathcal{N}_{v_i'}^{K_m}\}_{m=1,n=1}^{M,N}$ from $\mathcal{I}_1$ using K-NN;
6. Compute indicator vectors $\{\boldsymbol{x}_i^{K_m}\}_{m=1,i=1}^{M,N}$ by Eqs.~(7), (8) to establish local graphs $\{G_{v_i}^{K_m}, G_{v_i'}^{K_m}\}_{m=1,i=1}^{M,N}$;
7. Compute affinity vectors $\{\boldsymbol{w}_i^{K_m}\}_{m=1,i=1}^{M,N}$ using Eqs.~(11), (12) and (9);
8. Obtain inlier set $\mathcal{I}^*$ from Eqs.~(15), (16) and (17).

2.4. Difference from LPM

This work is related to our previously developed LPM [31]. Inspired by LPM, we resolve each correspondence by preserving the local topological consensus. However, based on the graph matching theory, this work proposes a different optimization framework to solve the matching problem.

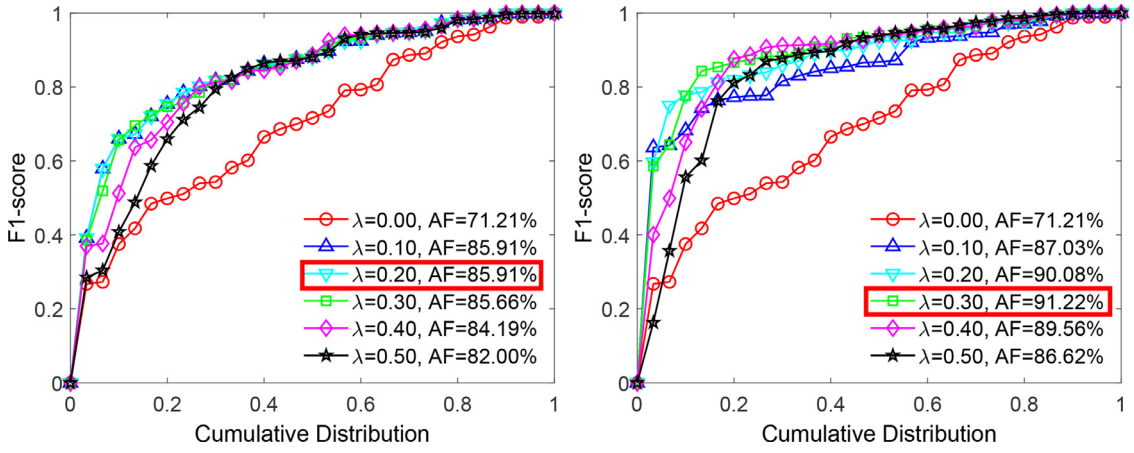

Fig. 4. F1-score on the test set under various λ. Left: result without using multi-scale strategy, where $K = 10$; right: result using multi-scale strategy, where $\boldsymbol{K} = [7, 10, 13]$. A point on the curve with coordinate (x, y) denotes that there are $100 * x$ percents of image pairs that have F1-score no more than y. The best average F1-scores for its λ are indicated by red boxes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

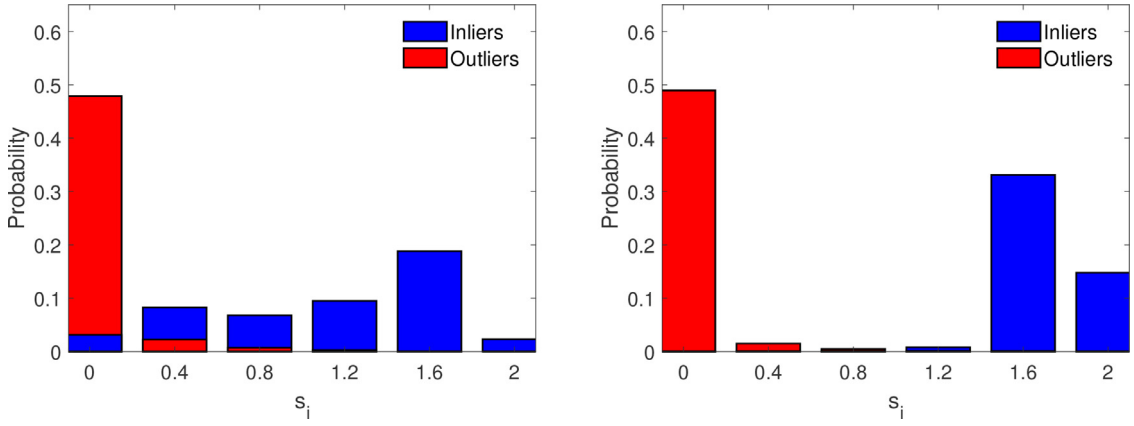

Fig. 5. Distribution of s_i in Eq. (15) about putative matches of 30 image pairs by using the whole initial feature set $\mathcal{I}_0 = T$ (left) and the iterated match set $\mathcal{I}_1$ (right) for neighborhood construction.

Firstly, based on the IQP, the mismatch removal problem is formulated as a graph matching model with a compensation term to discourage matches. Unlike the heuristic point distance measure-based modeling in LPM, our LGSC has a novel model formulation. Secondly, we propose to represent the topology information by devising and constructing a local graph structure. Compared to LPM where only intersections within K-NN are computed, we impose stricter physical constraints from the affinity scores of nodes and edges of the local graphs to preserve the consensus of topology structure. Thirdly, we introduce the ranking shift to transform the putative feature set from the two-dimensional geometric space to a ranking shift space with scale and rotation invariance, and hence making LGSC more stable.

In short, our LGSC proposes an innovative optimization framework capable of expressing local topological consensus more rigorously, with more robust and accurate feature matching performance. This will be validated in the subsequent experimental results.

2.5. Computational complexity

In the first step of neighborhood construction, we choose the K-D tree [43] to work on N feature correspondences, and this step has a time complexity of $O((K+N)\log N)$. Due to the multi-scale strategy, the time complexities of Lines 1 and 5 are no more than $O((K_M+N)\log N)$, which are the major time consumption of our LGSC. Moreover, in each iteration step, the K_m $(K_m < K_M)$ neighborhood $\mathcal{N}_{v_i}^{K_m}$ can be obtained from $\mathcal{N}_{v_i}^{K_M}$.

Subsequently, to construct the local graph structures, $O(MN(K_M \log K_M + 2K_M))$ time complexity needs to be consumed by determining the local indicator vectors in Lines 2 and 6. Later, the maximum for the time complexities of Lines 3 and 7 is $O(MK_MN)$ to compute affinity vectors $\{\boldsymbol{w}_i^{K_M}\}_{i=1}^N$, where the ranking shift sequences under K_m $(K_m < K_M)$ can be from $\boldsymbol{\phi}_i$ in K_M neighborhood. Finally, determining $\boldsymbol{x}$ and $\mathcal{I}^*$ in Lines 4 and 8 costs about $O(MK_MN)$ time complexity. Therefore, the total time complexity is roughly $O((K_M+N)\log N + MN(K_M \log K_M + 2K_M) + MK_MN + MK_MN)$.

Moreover, the space complexity of our algorithm is $O(MK_MN)$ based on the memory requirement for the local graph structures. In most cases, $MK_M \ll N$, thus the time and space complexities are approximate $O(N \log N)$ and $O(N)$, respectively. Consequently, our LGSC has linearithmic time and linear space complexity concerning the number of correspondences in the putative feature set.

2.6. Parameters setting

In our method, there are two parameters: $\boldsymbol{K}$ and λ. Parameter $\boldsymbol{K}$ denotes the number of nearest neighbors for multi-scale local graph structure construction. As shown in Table 1, the effect of our method has been significantly improved at $\boldsymbol{K} = [7, 10, 13]$. However, when the scale setting is larger, the consuming time and average F1-score tend to deteriorate due to computational load and

Table 1
Average F1-score (%) and average run time (ms) for our LGSC under different parameter $\boldsymbol{K}$ on the 30 image pairs in Fig. 4.

K	[2, 5, 8]	[7, 10, 13]	[12, 15, 18]	[17, 20, 23]
AF1	96.12	97.69	97.67	97.28
ART	136.6	154.1	208.1	263.6

unreliable consensus of insufficiently small neighborhoods, respectively. Moreover, parameter λ is the trade-off between inliers and outliers, and directly determines the optimal inlier set as a threshold. A smaller value of λ intends to preserve more inliers thus enhancing Recall and decreasing Precision, while a larger value does the opposite. From Fig. 4, parameter λ of the first iteration should be set as 0.3. Due to the higher inlier ratio in the second iteration, we set a higher λ, i.e., 0.45, empirically. In summary, the default values of these parameters in this paper are empirically set as $\boldsymbol{K} = [7, 10, 13]$ and $\lambda = 0.3, 0.45$ in the two iterations, respectively.

3. Experimental results

With the purpose of evaluating our LGSC, we firstly design qualitative and quantitative experiments on different kinds of real image pair sets and compare LGSC with other representative mismatch removal methods, e.g., RANSAC [21], MAGSAC++ [25], GS [13], LLT [28], LPM [31], GMS [34], LMR [42], and LFGC [38], and then we test the robustness and universality of LGSC in terms of image deformation and feature descriptors. In the end, we apply LGSC to the image registration task for the application. The open-source VLFeat toolbox [44] is applied for SIFT detector and descriptor as well as K-NN searching using K-D tree. All experiments are conducted on a laptop with a 2.71 GHz Intel Core CPU, 16 GB memory, and MATLAB R2020a code.

3.1. Datasets

To evaluate the performance of our LGSC reasonably and fairly, we select the following four image datasets:

- *RS* [45]. It is a remote sensing dataset consisting of 161 image pairs from color-infrared, SAR, and panchromatic photographs, the applied feature matching tasks of which usually include image mosaic, positioning and navigating, change detection, and so on. These image pairs usually agree with the parametric transformation model.
- *Daisy* [46]. This dataset mainly contains wide baseline image pairs with ground-truth depth maps, including two short image sequences and a few individual images. 52 image pairs are created by all the individual image pairs and two sequences for experimental evaluation.
- *VGG* [47]. This dataset is composed of 40 image pairs, which are either planar images (e.g., bark and wall) or roadside scenes (e.g., Leuven and trees). Due to the fixed position of the camera during acquisition, the image pairs in this dataset basically all conform to homography, and the ground truth is supplied by this dataset. From which we create in total 124 data, wherein the putative matches are obtained by using different ratio tests in SIFT.
- *Nonrigid*. This dataset includes two datasets, i.e., 720YUN [48] and FE [49], which both coincide complex nonrigid transformation models. 720YUN is a cloud dataset involving 20 image pairs from terrain, roads, buildings, terraces, etc. The raw images of which are video panoramic suffering from ground surface fluctuation and imaging view variations. And FE dataset is taken from a fish-eye camera including two scenes University and Urban with 32 image pairs. All 52 image pairs in the Nonrigid dataset undergo the nonrigid transformation thus easily causing a high computational burden.

The first dataset is created by ourselves, the ground truth of which is identified based on the benchmark established in advance for experimental objectivity, and inspected manually. For the last three datasets, the correctness of putative feature correspondences is determined by the ground truth information contained in the datasets.

3.2. Image feature matching

3.2.1. Comparison on representative image pairs

We select 12 respective image pairs for qualitative evaluation of our LGSC, the feature matching results of which are presented in Fig. 6. These image pairs contain a wide variety of transformation models to comprehensively evaluate our method. The two image pairs *Patch* and *Land* in the first row are from the RS dataset, undergoing only linear (e.g., rigid or affine) transformation. The second row includes image pairs *Fountain* and *Frustum* from Daisy and DTU datasets respectively, which have been subjected to wide-baseline imaging and large viewpoint change. The latter three image pairs *Wall, Bark* and *Trees* all come from the VGG dataset and obey homography. The image pair *Book* in the middle of the fourth row is classified as piece linear transformation, often used in image retrieval. The rest 4 image pairs *Retinal, YUN1, YUN2* and *University* belong to Retinal and Nonrigid datasets, and all involve nonrigid transformation. For each group of qualitative results, the left image pair denotes a feature matching result, and the right motion field shows the correctness of every correspondence in the putative feature set, visualized in color.

As mentioned before, SIFT is firstly used to construct the putative feature matching sets, yet they contain a large number and a high percentage of mismatches. For the representative 12 image pairs, the inlier numbers and ratios of them are (303, 37.27%), (388, 15.76%), (34, 49.28%), (124, 40.52%), (1089, 54.45%), (168, 58.33%), (865, 43.25%), (565, 75.74%), (44, 36.97%), (179, 23.37%), (165, 15.36%) and (457, 79.34%), respectively. Then our LGSC acts on the 12 putative feature matching sets to filter out outliers and preserve inliers as fully as possible. The accuracy and recall of the matching results for 12 image pairs respectively are (99.67%, 98.68%), (99.74%, 100%), (96.97%, 94.12%), (97.62%, 99.19%), (98.10%, 99.45%), (99.41%, 100%), (98.40%, 99.77%), (99.28%, 98.05%), (100%, 100%), (98.33%, 98.88%), (98.19%, 98.79%), and (99.56%, 100%), respectively. The image matching results are displayed in Fig. 6, where our LGSC performs well in feature matching for various image pairs.

To objectively evaluate the matching performance of our LGSC on 12 image pairs, we also designed a quantitative comparison with eight state-of-the-art methods, i.e., RANSAC, MAGSAC++, GS, LLT, LPM, GMS, LMR, and LFGC. For the comparison approaches, in brief, RANSAC is the classic resampling-based algorithm, MAGSAC++ is an advanced RANSAC-like method, GS is a graph matching method, LLT is a non-parametric model-based approach, LPM is a local area preserving-based method, GMS is a grid-based motion statistics method, LMR is a classification method based on the learning strategy, and LFGC is a learning-based method for feature matching problem. All implementations of these methods are from original papers and the related GitHub codes, and we do our best to tune their parameters for the best performance. Furthermore, the parameters of each method are fixed during all experiments. The matching performance, expressed by F1-score, is shown in Table 2. It can be observed that LGSC shows the best matching performance among 9 methods when processing these typical image pairs except for *Wall* image pair,

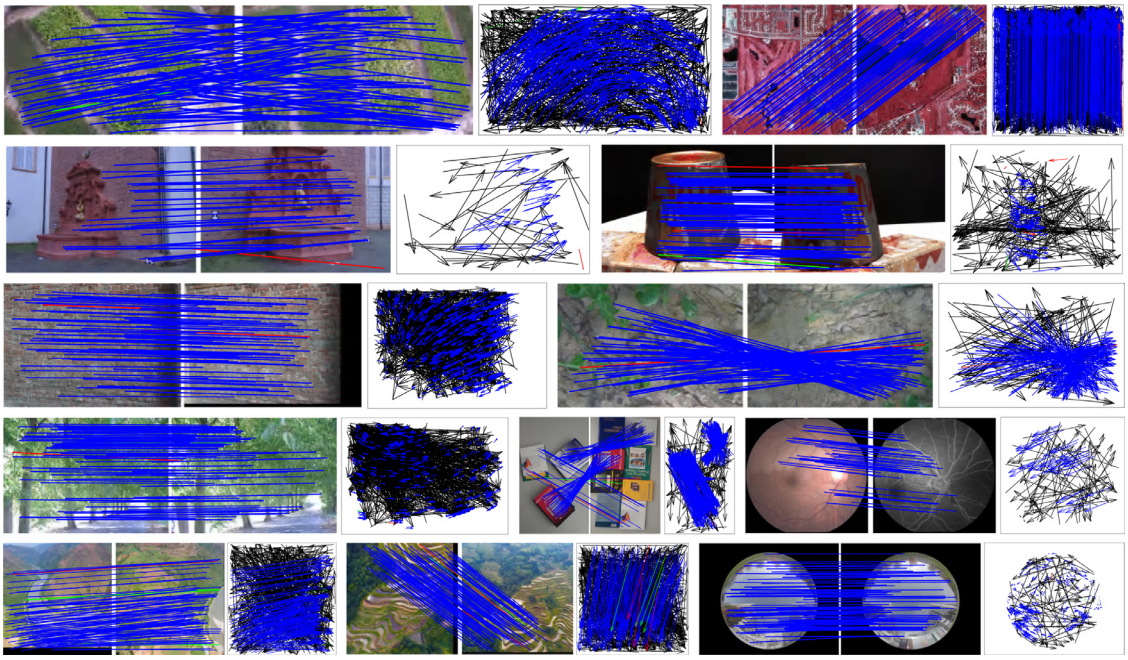

Fig. 6. Image feature matching results of our LGSC on 12 representative image pairs. From top to bottom, left to right: *Patch, Land, Fountain, Frustum, Wall, Bark, Trees, Book, Retinal, YUN1, YUN2* and *University*. The inlier ratios of the 12 image pairs are 37.27%, 15.76%, 49.28%, 40.52%, 54.45%, 58.33%, 43.25%, 75.74%, 36.97%, 23.37%, 15.36%, and 79.34%. The head and tail of each arrow in the motion field indicate the positions of feature points in two images (blue = true positive, black = true negative, green = false negative, red = false positive). For visibility, in the image pairs, at most 100 randomly selected matches are presented, and the true negatives are not shown. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Matching performance (F1-score, %) of LGSC and 8 comparative algorithms on 12 representative image pairs, where the highest F1-score regarding the same image pair is shown in bold.

	RANSAC	MAGSAC+	GS	LLT	LPM	GMS	LMR	LFGC	LGSC
Patch	99.01	96.71	93.93	99.17	98.36	70.91	98.20	62.38	**99.17**
Land	77.37	97.61	66.44	90.84	99.86	98.03	99.74	15.90	**99.87**
Fountain	95.62	88.57	74.73	77.42	90.63	85.00	82.76	80.95	**96.52**
Frustum	96.80	95.55	91.53	81.25	92.78	85.04	84.02	76.49	**98.40**
Wall	97.98	98.07	77.18	98.78	97.90	95.06	**98.91**	90.86	98.83
Bark	97.60	96.34	98.52	98.49	99.10	21.16	99.11	49.40	**99.70**
Trees	98.02	95.20	45.79	99.09	97.31	93.27	98.38	77.92	**99.30**
Book	60.84	63.40	90.50	85.37	97.90	72.82	97.50	72.98	**98.66**
Retinal	**100.0**	80.49	97.13	97.67	94.12	77.78	86.97	83.02	**100.0**
YUN1	69.93	98.33	93.04	96.40	95.37	80.50	96.76	55.26	**98.61**
YUN2	64.54	93.81	93.83	77.69	89.32	81.53	94.29	10.12	**98.49**
University	93.20	66.18	70.34	96.66	95.79	93.66	96.26	94.47	**99.78**
Average	87.58	89.19	82.75	91.57	95.70	79.56	94.41	64.15	**98.84**

even for which the F1-score of LGSC is less than 1 percentage away from the best LMR. Therefore, our LGSC has satisfying mismatch removal performance facing multi-type images.

3.2.2. Comparison on datasets

To comprehensively and reasonably assess the performance of our LGSC, we design quantitative experiments on four image datasets and compare our method with the eight state-of-the-art methods. For the four datasets, the average number of putative matches are 1475.60, 545.99, 742.44, and 378.75. The average inlier ratio of each dataset is shown in Fig. 7.

As shown in Fig. 7, the metrics of experimental results include precision, recall, F1-score, and runtime, from top to bottom. From left to right, each column represents the statistic results on datasets *RS, Daisy, VGG*, and *Nonrigid*, respectively. Similarly, GS usually has high precision but low recall, because it is not affine-invariant thus difficult to precisely estimate affinity parameters. RANSAC is limited by the parametric model thus reducing its effectiveness on the Nonrigid dataset. MAGSAC++, as a fast, reliable, and accurate robust estimator, would sacrifice substantial inliers for high precision. In addition, LFGC requires knowledge of the camera's intrinsic parameters, which are not provided in our experiments and therefore limited in performance. LLT performs poorly when it comes to nonrigid image pairs, as does LMR, which is due to their geometrical structure without strict constraints. Similarly, LPM is restricted by unstrict local topology representations when faced with complex image scenarios, such as sparse point sets and nonrigid transformation, and hence is prone to losing a large number of correct feature matches. Since GMS is designed with a large amount of low-quality matches, it has not achieved the best performance even with the highest computing speed. Clearly, we can observe that LGSC achieves the highest F1-score and relatively high computing efficiency. In summary, our LGSC has the dominant performance of mismatch removal in various scenes and has a superior performance over other different state-of-the-art methods.

In fact and for some specific cases, a higher precision (or recall) from 98% to 99% cannot bring real improvement for high-level applications, such as model fitting, pose estimation or localization.

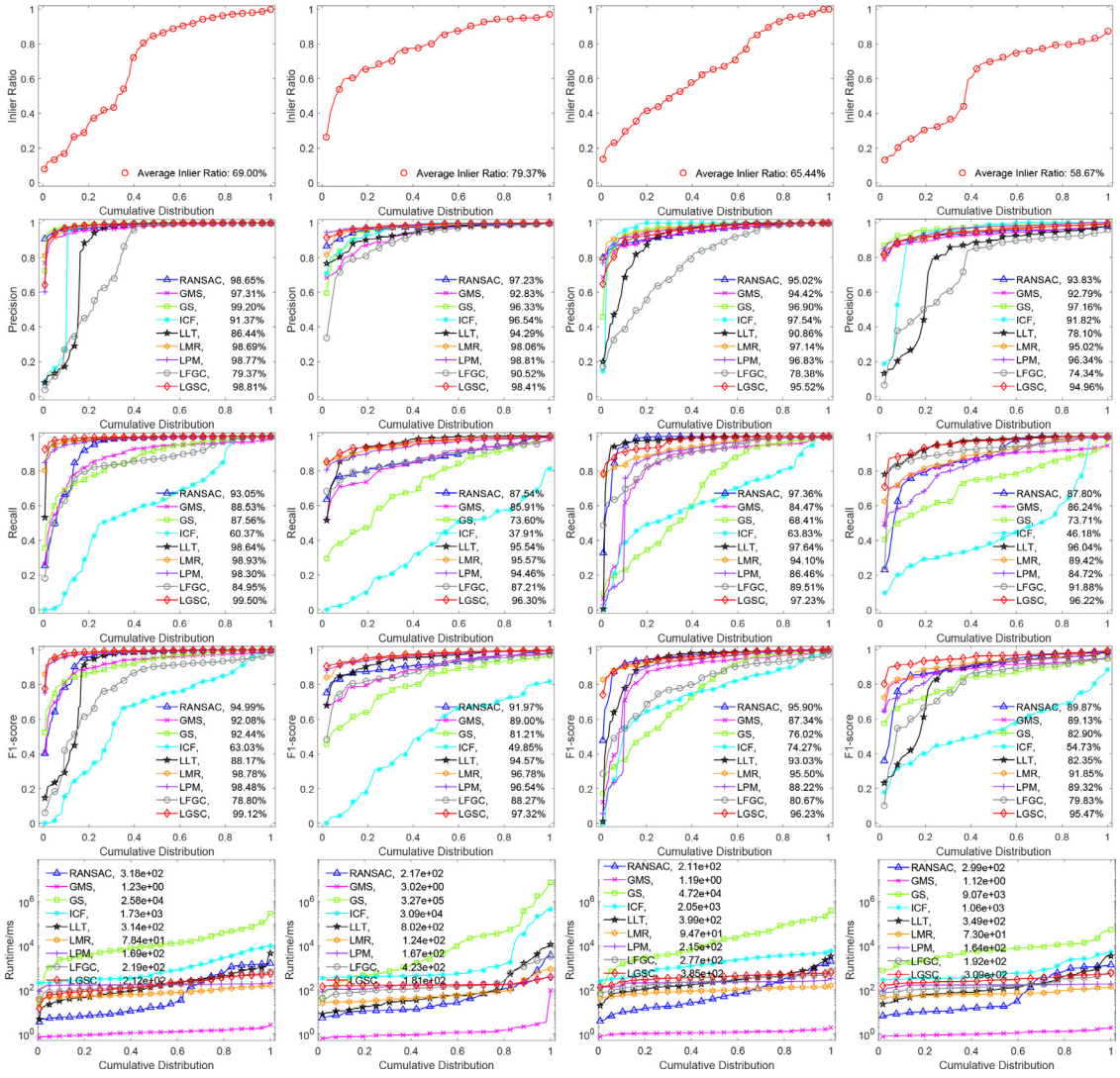

Fig. 7. Quantitative comparisons of RANSAC, MAGSAC++, GS, LLT, LPM, GMS, LMR, LFGC, and our LGSC on four image sets. From left to right: *RS, Daisy, VGG* and *Nonrigid* dataset. From top to bottom: initial inlier ratio, precision, recall, F1-score, and run time concerning the cumulative distribution. A point on the curve with coordinate (x, y) denotes that there are $(100 * x)\%$ percent of image pairs which have the performance value (i.e., Inlier Ratio, Precision, Recall, F1-score, and Runtime, respectively from top to bottom) no more than y, and the average performance for each comparing methods is shown in the legend accordingly.

That is because those tasks merely request a reliable inlier subset to support their parameter estimations, such as using four correspondences for homography estimation or seven for fundamental estimation [50]. Thus, inlier ratio is higher to a certain degree would be sufficient to guarantee the accuracy of subsequent robust estimators. But this is not true for non-rigid scenarios, only both higher inlier number and inlier ratio can support the accurate parameter estimation of non-rigid deformations, or can better measure the similarity of two images in image retrieval or loop closure detection (LCD) [42]. This will be demonstrated by image registration and LCD tasks in Sections 3.3 and 3.4.

3.2.3. Robustness and universality testing

To further test the robustness and universality of LGSC, two sets of comparative experiments are designed from the perspective of different degrees of deformation and different descriptors.

The first is a robustness testing experiment based on images with different deformation levels, as shown in Fig. 8, which contains six scenes with six images of different deformation levels in each scene. According to the degree of image deformations, the first image in each scene is combined with the other five images with increasing degrees of deformation to form five pairs of testing images, and hence each dataset contains six images of different scenes in the same deformation level, as each column in Fig. 8.

We use F1-score as a metric of matching performance as before. The inlier ratios of the five image pair datasets in five degrees of deformations are presented in Fig. 9, as well as the F1-scores of our LGSC and eight state-of-the-art methods on each degree of deformation. It can be seen that our LGSC can maintain a stable and accurate matching performance for different scenes under different degrees of deformation, with more robustness than other methods.

Subsequently, we conduct a quantitative test on the universality of descriptors by selecting three common feature descriptors (i.e., SIFT [15], ORB [16], and SURF [17]), and hence compare the experimental results of our LGSC with eight comparative methods. The VGG dataset, which is rich in image types, is selected as the testing image set and processed with different descriptors. In particular, each image pair is processed by SIFT and a similarity threshold of 1.0 or 1.5 is set to construct the putative feature matching set, called *VGG-SIFT*, the *VGG-ORB* is constructed by matching the top 1000 and 2000 feature points of ORB, and the *VGG-SURF* putative feature set is constructed using the top 40% and 60% correspondences of similarity of SURF, where the ground-truth labels of each dataset are determined based on the homographies sup-

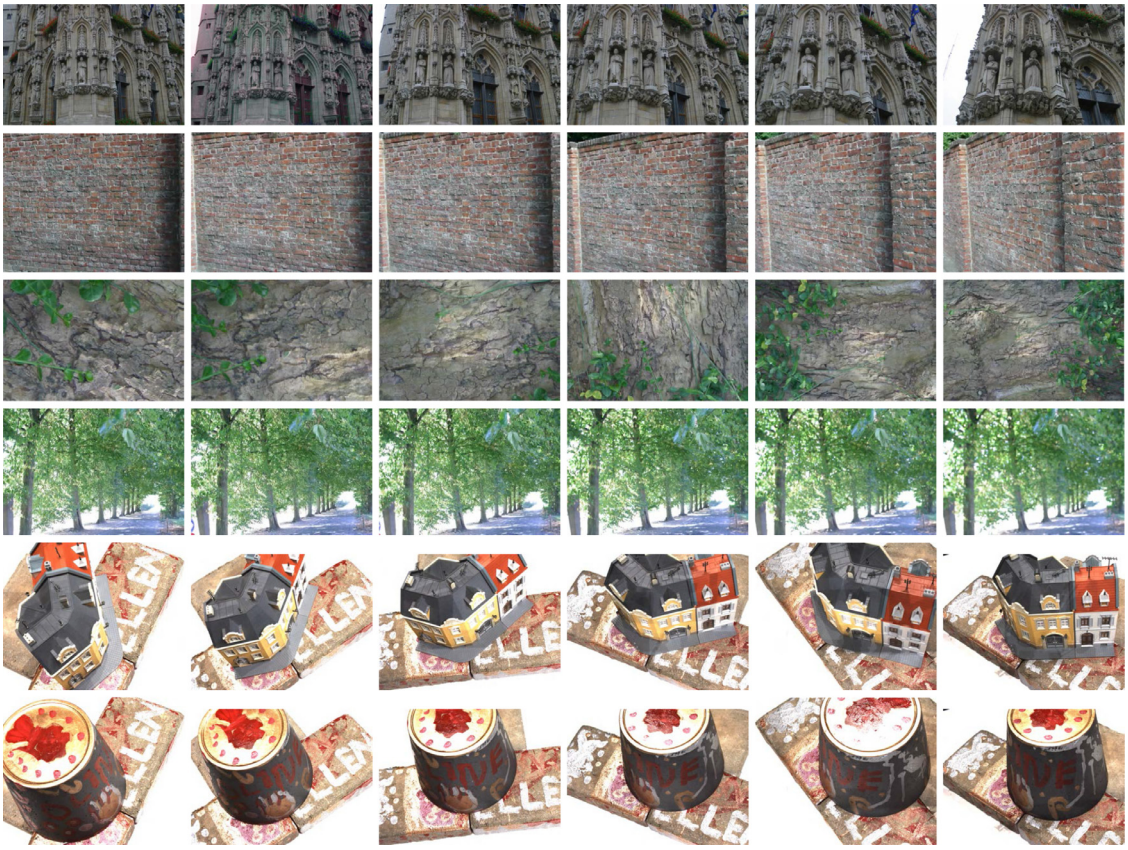

Fig. 8. Image dataset of 6 scenes with different degrees of deformation, including: architecture, wall, bark, trees, house, and frustum.

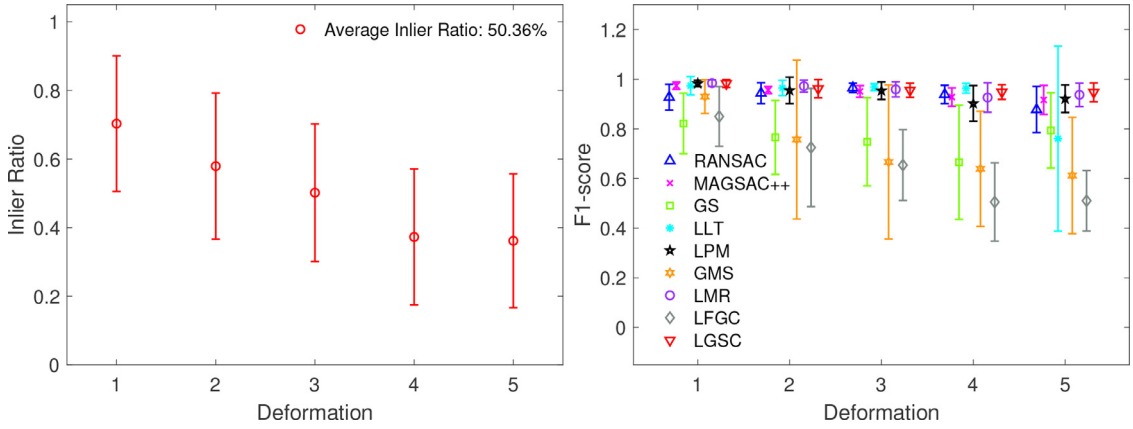

Fig. 9. Matching results between LGSC and 8 comparative algorithms on a dataset of image pairs consisting of five different degrees of deformation. The left image shows the mean and standard deviation of the putative matching inlier ratio for each degree of deformation, and the right image presents the mean and standard deviation of F1-score for LGSC and other eight comparative methods.

plied by VGG. The average inlier numbers and ratios of putative feature correspondence sets, corresponding to the descriptors SIFT, ORB, and SURF, are (693.18, 0.8810), (625.78, 0.5761) and (853.34, 0.5224), respectively. The matching resultant F1-scores of comparative algorithms in the universality experiments with different descriptors are shown in Table 3. Our LGSC is suitable for putative feature sets with various descriptors and has better matching accuracy than other state-of-the-art methods.

3.3. Image registration and mosaic

To exploit the practical value of LGSC, we apply it to the image registration task, i.e., maximizing the overlapped area between the reference image and the transformed sensed image. Firstly, LGSC processes the image pair containing a reference image and a sensed image, and obtains a reliable set of feature correspondences, after which thin-plate spline (TPS) with generality and smoothness in functional mapping is chosen to resolve the smooth fitting step of feature matching, and it estimates the image transformation function. Ultimately, each pixel of the sensed image is mapped to a transformed image domain coordinate using the estimated transform function, and the intensity of that coordinate in the reference image is then calculated using a bicubic interpolation algorithm.

For visualization of the effect of LGSC in the image registration task, we select 8 representative image pairs, and registration re-

Table 3
The F1-score (%) of LGSC and eight comparative methods on three descriptor-extracted feature sets VGG-SIFT, VGG-SURF, and VGG-ORB. The average result of each method is shown in the bottom row. Bold represents the best result on each feature set.

Data	RANSAC	MAGSAC+	GS	LLT	LPM	GMS	LMR	LFGC	LGSC
VGG-SIFT	97.01	96.89	94.91	96.30	95.95	85.58	95.42	87.07	**97.39**
VGG-ORB	93.76	93.73	52.52	94.34	94.77	92.24	94.18	78.34	**94.37**
VGG-SURF	93.17	94.01	84.55	94.15	91.54	81.49	93.32	76.85	**94.71**
Average	94.65	94.88	77.33	94.93	94.69	86.44	94.31	80.75	**95.49**

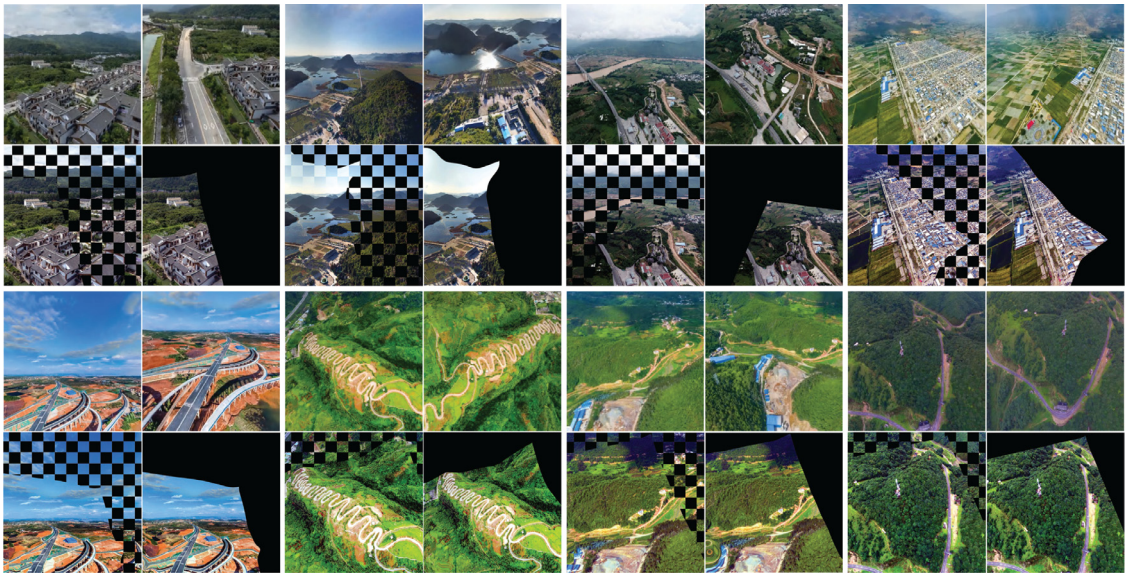

Fig. 10. Image registration results of LGSC applied on 8 typical image pairs. The first and third rows represent the original input images, where the left and right in each group are the reference and sensed images. The second and fourth rows present the registration results of LGSC, where the left and right in each group are checkboard results and the warped sensed images, respectively.

sults are shown in Fig. 10. The first and third rows represent the original input image pairs, in each group, the left image is the reference image and the right is the sensed image to be transformed. The second and fourth rows represent the image registration results of LGSC, where the left and right images of each group denote checkerboard results and the warped sensed images, respectively. It is shown that LGSC can match the overlapped regions of the image pair well, including the difficult edge regions.

In order to quantitatively evaluate the image registration performance of LGSC, a total of 181 image pairs of RS and 720YUN are selected as the objects to be aligned. We compare the image registration result of LGSC on these image pairs with those of other comparative methods. The average numbers of feature matches and inlier ratios are 958.53 and 27.95%, respectively. Moreover, for each image pair, 20 pairs of landmark pixel values $\{r_i^c, s_i^c\}_{i=1}^L$ are randomly selected to calculate the metrics of image registration, including the root mean square error (RMSE), maximum error (MAE) and median error (MEE), with the three metrics defined as follows:

$$\mathrm{RMSE} = \sqrt{1/L \sum_{i=1}^{L} \left(r_i^c - \mathcal{F}\left(s_i^c\right)\right)^2}, \tag{19}$$

$$\mathrm{MAE} = \max \left\{ \sqrt{\left(r_i^c - \mathcal{F}\left(s_i^c\right)\right)^2} \right\}_{i=1}^{L}, \tag{20}$$

$$\mathrm{MEE} = \mathrm{median} \left\{ \sqrt{\left(r_i^c - \mathcal{F}\left(s_i^c\right)\right)^2} \right\}_{i=1}^{L}. \tag{21}$$

The results of quantitative image registration experiments for 181 image pairs are presented in Table 4. Clearly, our LGSC has the best registration performance in RMSE, MAE, and MEE metrics, and is stable for various images.

Table 4
The results of image registration quantitative experiments. The average values and standard deviations of RMSE, MAE, and MEE are used for evaluation. Bold indicates the best.

Method	RMSE	MAE	MEE
RANSAC	18.99 (±36.77)	44.53 (±81.73)	23.22 (±46.80)
MAGSAC+	10.12(±23.32)	23.71 (±47.22)	11.19 (±25.11)
GS	10.24 (±26.88)	25.87 (±63.47)	12.00 (±32.61)
LLT	24.29 (±19.64)	70.27 (±54.17)	26.83 (±22.84)
LPM	20.92 (±31.69)	46.19 (±65.61)	26.59 (±41.73)
GMS	106.7 (±138.1)	216.7 (±270.7)	142.8 (±190.9)
LMR	35.46 (±63.28)	72.36 (±117.6)	47.26 (±87.86)
LFGC	107.3 (±133.2)	228.8 (±271.4)	141.2 (±181.2)
LGSC	**8.300**(±**15.18**)	**20.52**(±**35.64**)	**10.00**(±**19.17**)

Moreover, by means of image registration, our proposed LGSC can be directly applied to the image mosaic task, which aims to stitch two or more images with overlapping parts into a large-scale image or single panorama. Taking a pair of images as an example, our LGSC processes them to obtain a reliable set of feature matches, and then calculates the transformation matrix $\mathcal{F}$. Based on $\mathcal{F}$, the images are transformed into the same coordinate system and the image stitching result is obtained through steps such as erosion and alpha blending. If more than two images are required to be stitched, we adopt a multi-image matching strategy with bundle adjustment to obtain a panoramic [51]. In fact, the core to evaluate image mosaic performance is intrinsically to evaluate image registration [52,53]. That means, our image registration results can directly reflect the mosaic performance. In this regard,

Table 5
Results on loop closure detection. The precision at 100% recall is used for evaluation. Bold indicates the best.

	RANSAC	MAGSAC+	LLT	LPM	GMS	LMR	LFGC	LGSC
K00	0.9112	0.9105	0.8964	0.9115	0.8252	0.8820	0.8236	**0.9162**
K02	0.7632	0.7757	0.7477	0.7427	0.7118	0.7750	0.7332	**0.7768**

Fig. 11. Image mosaic results of LGSC applied on 4 typical image pairs. In each group of images, the left two images are a pair of images to be stitched, while the right image containing black artifacts is the result of image stitching using LGSC.

Fig. 12. Panoramic image mosaic results of RANSAC (top) and our LGSC (bottom).

we just show some qualitative results to demonstrate the practical utility of our method, including two image pairs chosen from *RS* dataset and two from *Nonrigid* dataset, as shown in Fig. 11. The result of panoramic image mosaic is shown in Fig. 12. It reveals that our LGSC can well perform the image mosaic task, especially around the stitching lines due to the accuracy of LGSC in image registration.

3.4. Loop closure detection

To further signify the practical value, we also exploit two sequences from KITTI vision suite [54] (i.e., *K00, K02*) to test the performance of our method in loop closure detection (LCD) task, i.e., recognizing re-observations during the navigation of a robot. As DELG [55] is considered state-of-the-art for LCD-related tasks such as image retrieval and landmark recognition, we choose it to simultaneously extract global and local features of images to perform LCD in a hierarchical way. Specifically, given a query image, the similarity of global features under L2-metric is first used to select its candidate frame, followed by performing model fitting between them based on local features. Only when sufficient matches are preserved would a loop-closing event be triggered. In this task, the maximum recall rate at 100% precision is regarded as the evaluation metric, where precision is the number of true positives over the total identifications, while recall is the ratio of true positives and all true examples indicated by ground truth (GT) [56,57]. Here GT is provided in the form of binary matrices and preserves the image-wise correspondence of datasets. Meanwhile, an identification would be regarded as a true positive when it locates within 10 neighboring frames of true loop-closing examples indicated by GT.

Quantitative results of LCD are reported in Table 5, where dataset *K02* is much more challenging than *K00* due to its complex scenes. In LCD task, only using more inliers but less outliers, i.e., both higher precision and recall, can better measure the similarity of two images thus recognizing the loop-closing event. From the table, we find that our LGSC can achieve the best even in challenging *K02*, due to its robustness and accuracy on inlier-outlier classification task (as shown in Fig. 7). RANSAC-like methods have obtained promising results, as they perform on global geometrical model estimation with sufficient iterations.

4. Discussion and conclusion

In this paper, we construct a novel, general, and efficient optimization framework to solve mismatch removal. Based on the consensus of local graph structure, we have devised an innovative objective thus classifying each correspondence as an inlier or outlier. Our method named *Local Graph Structure Consensus* (LGSC) has a simple closed-form solution with linearithmic time and lin-

ear space complexity, ensuring its high computational efficiency. Moreover, a multi-scale strategy is introduced to improve the robustness of our LGSC, and we also have proposed an iterative local graph construction strategy to further enhance the image matching performance. The superiority of our algorithm compared to other state-of-the-art methods has been verified in extensive experiments. Eventually, our LGSC can efficiently filter out mismatches and preserve correct correspondences fully, even in extreme cases of image distortion and nonrigid transformation. Furthermore, LGSC has a reliable and stable matching performance in terms of robustness to image deformations and universality to various descriptors. Besides, with few parameters to be preset manually and a low computational burden, our LGSC shows excellent application significance in vision-based tasks, such as image registration.

Albeit local graph structure consensus can universally and effectively complete feature matching, it still has certain limitations. Evidently, LGSC is reliant on local topological expression, and hence excessive outlier rates would cause failures for the calculation of consistency even with the iteration strategy. In addition, the speed of LGSC does not offer a significant advantage over state-of-the-art algorithms like GMS, leaving room for improvement. These will be studied in our future work.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant no. 61773295, and in part by the Natural Sciences and Engineering Research Council of Canada under Grant no. RGPIN-2020-04661.

References

[1] J. Ma, X. Jiang, A. Fan, J. Jiang, J. Yan, Image matching from handcrafted to deep features: asurvey, Int. J. Comput. Vis. 129 (1) (2021) 23–79.
[2] C. Min, Y. Gu, Y. Li, F. Yang, Non-rigid infrared and visible image registration by enhanced affine transformation, Pattern Recognit. 106 (2020) 107377.
[3] J. Ma, W. Yu, P. Liang, C. Li, J. Jiang, FusionGAN: a generative adversarial network for infrared and visible image fusion, Inf. Fusion 48 (2019) 11–26.
[4] Y. Furukawa, J. Ponce, Accurate, dense, and robust multiview stereopsis, IEEE Trans. Pattern Anal. Mach. Intell. 32 (8) (2009) 1362–1376.
[5] Y. Liu, D. Zhang, G. Lu, W.-Y. Ma, A survey of content-based image retrieval with high-level semantics, Pattern Recognit. 40 (1) (2007) 262–282.
[6] R. Xie, M. Xia, J. Yao, L. Li, Guided color consistency optimization for image mosaicking, ISPRS J. Photogramm. Remote Sens. 135 (2018) 43–59.
[7] S. Nebiker, N. Lack, M. Deuber, Building change detection from historical aerial photographs using dense image matching and object-based image analysis, Remote Sens. 6 (9) (2014) 8310–8336.
[8] R. Mur-Artal, J.M.M. Montiel, J.D. Tardos, ORB-SLAM: a versatile and accurate monocular slam system, IEEE Trans. Robot. 31 (5) (2015) 1147–1163.
[9] C. Wang, L. Wang, L. Liu, Progressive mode-seeking on graphs for sparse feature matching, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 788–802.
[10] P.J. Besl, N.D. McKay, A method for registration of 3-D shapes, IEEE Trans. Pattern Anal. Mach. Intell. 14 (2) (1992) 239–256.
[11] H. Chui, A. Rangarajan, A new point matching algorithm for non-rigid registration, Comput. Vis. Image Underst. 89 (2–3) (2003) 114–141.
[12] B. Jiang, J. Tang, C. Ding, B. Luo, Binary constraint preserving graph matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4402–4409.
[13] H. Liu, S. Yan, Common visual pattern discovery via spatially coherent correspondences, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 1609–1616.
[14] B. Jiang, P. Sun, B. Luo, GLMNet: graph learning-matching convolutional networks for feature matching, Pattern Recognit. 121 (2022) 108167.
[15] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[16] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, ORB: an efficient alternative to sift or surf, in: Proceedings of the International Conference on Computer Vision, 2011, pp. 2564–2571.
[17] H. Bay, T. Tuytelaars, L. Van Gool, Surf: speeded up robust features, in: Proceedings of the European Conference on Computer Vision, 2006, pp. 404–417.
[18] K. Yi, E. Trulls, V. Lepetit, P. Fua, Lift: learned invariant feature transform, in: Proceedings of the European Conference on Computer Vision, 2016, pp. 467–483.
[19] J. Sun, Z. Shen, Y. Wang, H. Bao, X. Zhou, LoFTR: detector-free local feature matching with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8922–8931.
[20] X. Jiang, J. Ma, G. Xiao, Z. Shao, X. Guo, A review of multimodal image matching: methods and applications, Inf. Fusion 73 (2021) 22–71.
[21] M.A. Fischler, R.C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM 24 (6) (1981) 381–395.
[22] P.H. Torr, A. Zisserman, MLESAC: a new robust estimator with application to estimating image geometry, Comput. Vis. Image Underst. 78 (1) (2000) 138–156.
[23] R. Raguram, O. Chum, M. Pollefeys, J. Matas, J.-M. Frahm, USAC: a universal framework for random sample consensus, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2012) 2022–2038.
[24] D. Barath, J. Matas, J. Noskova, MAGSAC: marginalizing sample consensus, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10197–10205.
[25] D. Barath, J. Noskova, M. Ivashechkin, J. Matas, MAGSAC++, a fast, reliable and accurate robust estimator, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 1304–1312.
[26] X. Li, Z. Hu, Rejecting mismatches by correspondence function, Int. J. Comput. Vis. 89 (1) (2010) 1–17.
[27] J. Ma, J. Zhao, J. Tian, A.L. Yuille, Z. Tu, Robust point matching via vector field consensus, IEEE Trans. Image Process. 23 (4) (2014) 1706–1721.
[28] J. Ma, H. Zhou, J. Zhao, Y. Gao, J. Jiang, J. Tian, Robust feature matching for remote sensing image registration via locally linear transforming, IEEE Trans. Geosci. Remote Sens. 53 (12) (2015) 6469–6481.
[29] Y. Lipman, S. Yagev, R. Poranne, D.W. Jacobs, R. Basri, Feature matching with bounded distortion, ACM Trans. Graph. 33 (3) (2014) 1–14.
[30] W.-Y. Lin, F. Wang, M.-M. Cheng, S.-K. Yeung, P.H. Torr, M.N. Do, J. Lu, Code: coherence based decision boundaries for feature correspondence, IEEE Trans. Pattern Anal. Mach. Intell. 40 (1) (2017) 34–47.
[31] J. Ma, J. Zhao, J. Jiang, H. Zhou, X. Guo, Locality preserving matching, Int. J. Comput. Vis. 127 (5) (2019) 512–531.
[32] X. Jiang, J. Jiang, A. Fan, Z. Wang, J. Ma, Multiscale locality and rank preservation for robust feature matching of remote sensing images, IEEE Trans. Geosci. Remote Sens. 57 (9) (2019) 6462–6472.
[33] J. Jiang, Q. Ma, X. Jiang, J. Ma, Ranking list preservation for feature matching, Pattern Recognit. 111 (2021) 107665.
[34] J. Bian, W.-Y. Lin, Y. Matsushita, S.-K. Yeung, T.-D. Nguyen, M.-M. Cheng, GMS: grid-based motion statistics for fast, ultra-robust feature correspondence, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2828–2837.
[35] X. Jiang, J. Ma, J. Jiang, X. Guo, Robust feature matching using spatial clustering with heavy outliers, IEEE Trans. Image Process. 29 (2020) 736–746.
[36] G. Wang, Y. Chen, Robust feature matching using guided local outlier factor, Pattern Recognit. 117 (2021) 107986.
[37] G. Xiao, S. Wang, H. Wang, J. Ma, Mining consistent correspondences using co-occurrence statistics, Pattern Recognit. 119 (2021) 108062.
[38] K.M. Yi, E. Trulls, Y. Ono, V. Lepetit, M. Salzmann, P. Fua, Learning to find good correspondences, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2666–2674.
[39] J. Zhang, D. Sun, Z. Luo, A. Yao, L. Zhou, T. Shen, Y. Chen, L. Quan, H. Liao, Learning two-view correspondences and geometry using order-aware network, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 5845–5854.
[40] W. Sun, W. Jiang, E. Trulls, A. Tagliasacchi, K.M. Yi, Acne: attentive context normalization for robust permutation-equivariant learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11286–11295.
[41] P.-E. Sarlin, D. DeTone, T. Malisiewicz, A. Rabinovich, Superglue: learning feature matching with graph neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4938–4947.
[42] J. Ma, X. Jiang, J. Jiang, J. Zhao, X. Guo, LMR: learning a two-class classifier for mismatch removal, IEEE Trans. Image Process. 28 (8) (2019) 4045–4059.
[43] J.L. Bentley, Multidimensional binary search trees used for associative searching, Commun. ACM 18 (9) (1975) 509–517.
[44] A. Vedaldi, B. Fulkerson, VLFeat: an open and portable library of computer vision algorithms, in: Proceedings of the ACM International Conference on Multimedia, 2010, pp. 1469–1472.
[45] J. Ma, J. Jiang, H. Zhou, J. Zhao, X. Guo, Guided locality preserving feature matching for remote sensing image registration, IEEE Trans. Geosci. Remote Sens. 56 (8) (2018) 4435–4447.
[46] E. Tola, V. Lepetit, P. Fua, Daisy: an efficient dense descriptor applied to wide-baseline stereo, IEEE Trans. Pattern Anal. Mach. Intell. 32 (5) (2009) 815–830.
[47] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, L. Van Gool, A comparison of affine region detectors, Int. J. Comput. Vis. 65 (1) (2005) 43–72.

[48] L. Liang, W. Zhao, X. Hao, Y. Yang, K. Yang, L. Liang, Q. Yang, Image registration using two-layer cascade reciprocal pipeline and context-aware dissimilarity measure, Neurocomputing 371 (2020) 1–14.
[49] M. Horst, R. Möller, Visual place recognition for autonomous mobile robots, Robotics 6 (2) (2017) 9.
[50] A. Fan, J. Ma, X. Jiang, H. Ling, Efficient deterministic search with robust loss functions for geometric model fitting, IEEE Trans. Pattern Anal. Mach. Intell. (2021), doi:10.1109/TPAMI.2021.3109784.
[51] M. Brown, D.G. Lowe, Automatic panoramic image stitching using invariant features, Int. J. Comput. Vis. 74 (1) (2007) 59–73.
[52] Y. Zhang, Z. Wan, X. Jiang, X. Mei, Automatic stitching for hyperspectral images using robust feature matching and elastic warp, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13 (2020) 3145–3154.
[53] L. Deng, X. Yuan, C. Deng, J. Chen, Y. Cai, Image stitching based on nonrigid warping for urban scene, Sensors 20 (24) (2020) 7050.
[54] A. Geiger, P. Lenz, C. Stiller, R. Urtasun, Vision meets robotics: the KITTI dataset, Int. J. Robot. Res. 32 (11) (2013) 1231–1237.
[55] B. Cao, A. Araujo, J. Sim, Unifying deep local and global features for image search, in: European Conference on Computer Vision, Springer, 2020, pp. 726–743.
[56] K. Zhang, X. Jiang, J. Ma, Appearance-based loop closure detection via locality-driven accurate motion field learning, IEEE Trans. Intell. Transp. Syst. (2021), doi:10.1109/TITS.2021.3086822.
[57] S. An, H. Zhu, D. Wei, K.A. Tsintotas, Fast and incremental loop closure detection with deep features and proximity graphs, arXiv preprint arXiv:2010.11703(2020).

Xingyu Jiang received the B.E. degree from the Department of Mechanical and Electronic Engineering, Huazhong Agricultural University, Wuhan, China, in 2017, and the M.S. degree from the Electronic Information School, Wuhan University, Wuhan, China, in 2019. He is currently a Ph.D. student with the Electronic Information School, Wuhan University. His research interests include computer vision, machine learning, and pattern recognition.

Yifan Xia received the B.E. degree from the Electronic Information School, Wuhan University, Wuhan, China, in 2021. He is currently a Ph.D. student with the Electronic Information School, Wuhan University. His current research interests include computer vision and image processing.

Xiao-Ping Zhang received B.S. and Ph.D. degrees from Tsinghua University, in 1992 and 1996, respectively, both in Electronic Engineering. He holds an MBA in Finance, Economics and Entrepreneurship with Honors from the University of Chicago Booth School of Business, Chicago, IL. Since Fall 2000, he has been with the Department of Electrical, Computer and Biomedical Engineering, Ryerson University, Toronto, ON, Canada, where he is currently a Professor and the Director of the Communication and Signal Processing Applications Laboratory. He has served as the Program Director of Graduate Studies. He is cross-appointed to the Finance Department at the Ted Rogers School of Management, Ryerson University. He was a Visiting Scientist with the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, USA, in 2015 and 2017. He is a frequent consultant for biotech companies and investment firms. His research interests include image and multimedia content analysis, machine learning, statistical signal processing, sensor networks and IoT, and applications in big data, finance, and marketing. Dr. Zhang is Fellow of the Canadian Academy of Engineering, Fellow of the Engineering Institute of Canada, Fellow of the IEEE, a registered Professional Engineer in Ontario, Canada, and a member of Beta Gamma Sigma Honor Society. He is Senior Area Editor for IEEE TIP and IEEE TSP. He was Associate Editor for IEEE TIP, IEEE TMM, IEEE TCSVT, IEEE TSP, and IEEE SPL. He received 2020 Sarwan Sahota Ryerson Distinguished Scholar Award – the Ryerson University highest honor for scholarly, research and creative achievements. He is selected as IEEE Distinguished Lecturer by the IEEE Signal Processing Society for the term 2020 to 2021, and by the IEEE Circuits and Systems Society for the term 2021 to 2022.

Jiayi Ma 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 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 robotics. 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, an Associate Editor of Neurocomputing, Sensors and Entropy, and a Guest Editor of Remote Sensing. He is a Senior Member of IEEE.