Locality-Guided Global-Preserving Optimization for Robust Feature Matching

Yifan Xia and Jiayi Ma, *Senior Member, IEEE*

***Abstract*— Feature matching is a fundamental problem in many computer vision tasks. This paper proposes a novel effective framework for mismatch removal, named *LOcality-guided Global-preserving Optimization* (LOGO). To identify inliers from a putative matching set generated by feature descriptor similarity, we introduce a fixed-point progressive approach to optimize a graph-based objective, which represents a two-class assignment problem regarding an affinity matrix containing global structures. We introduce a strategy that a small initial set with a high inlier ratio exploits the topology of the affinity matrix to elicit other inliers based on their reliable geometry, which enhances the robustness to outliers. Geometrically, we provide a locality-guided matching strategy, *i.e.,* using local topology consensus as a criterion to determine the initial set, thus expanding to yield the final feature matching set. In addition, we apply local affine transformations based on reference points to determine the local consensus and similarity scores of nodes and edges, ensuring the validity and generality for various scenarios including complex nonrigid transformations. Extensive experiments demonstrate the effectiveness and robustness of the proposed LOGO, which is competitive with the current state-of-the-art methods. It also exhibits favorable potential for high-level vision tasks, such as essential and fundamental matrix estimation, image registration and loop closure detection.**

***Index Terms*— Feature matching, mismatch removal, locality-guided, graph optimization, global structure.**

I. Introduction

IMAGE matching aims to establish a set of reliable correspondences between two related images of the same object or scene, which plays a basic and important role in many vision-based tasks, such as image registration and fusion, 3D reconstruction, and structure-from-motion [1]–[4]. Feature matching, characterized by extracting the feature structure of an image, performs great stability and efficiency in matching images, and has received widespread attention in recent years.

Feature matching is subjected to complex computational nature, and hence has a huge requirement for matching space. Even without considering outliers, finding correct one-to-one correspondences between two sets of N points requires up to $N!$ permutations [5]. As a common alternative to this problem, Graph Matching (GM) directly constructs correspondences between two feature point sets based on a graph model. Most GM methods relax the stringent constraints and provide an approximate solution [6]–[10]. However, this type of direct matching of two point sets usually has a very high time complexity, space burden, and poor scalability, thus is difficult to handle texture-rich images and large data.

To address the above issues, a popular indirect matching strategy is proposed. Firstly, construct a group of putative matches according to the similarity of feature descriptors, representative ones of which include Scale-Invariant Feature Transform (SIFT) [11], Oriented fast and Rotated Brief (ORB) [12], and Speeded Up Robust Features (SURF) [13]. Subsequently, extra geometric constraints are required to remove the mismatches (outliers) and preserve the correct matches (inliers) from the putative correspondence set, *i.e.,* determining the correctness of each pair of feature points.

The mismatch removal problem has been addressed by numerous methods in recent years. Classic RANdom SAmple Consensus (RANSAC) [14] and its variants [15], [16] estimate the parameters of transformation models in an iterative method, but rely on the predefined rigid transformation types. In addition, these methods typically perform badly when the image transformation conforms to complex nonrigid, and they would degrade in case of high proportional outliers. Unlike simple parametric models, non-parametric model-based methods can deal with degenerated scenarios, such as Vector Field Consensus (VFC) [17] and Locally Linear Transforming (LLT) [18]. These methods use a mixture model and construct a robust vector field, assuming that the motion field is smooth. However, they typically perform badly in cases of independent motions or discontinuous depths.

Relaxed geometric constraints are exploited by many methods to deal with various scenarios, which is a popular trend in recent years. Lin *et al.* [19] proposed to use nonlinear regression for the estimation of likelihood functions. Based on the preservation of small areas, Bian *et al.* devised a Grid-based Motion Statistics (GMS) method. Further, Ma *et al.* [20] presented Locality Preserving Matching (LPM) with high efficiency, some variants (*e.g.,* [21]) of which have made enhancements mainly in the expression of local topology. Jiang *et al.* [22] converted the feature matching into a spatial clustering problem with DBSCAN based on the local consensus, and Shao *et al.* [23] also used DBSCAN to calculate the minimum relative motion entropy thus improving the accuracy of matching. Recently, Cavalli *et al.* [24] proposed a hierarchical pipeline for effective outlier detection based

Manuscript received 23 January 2022; revised 20 June 2022 and 12 July 2022; accepted 13 July 2022. Date of publication 27 July 2022; date of current version 3 August 2022. This work was supported by the Key Research and Development Program of Hubei Province under Grant 2020BAB113. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Baojiang Zhong. *(Corresponding author: Jiayi Ma.)*

The authors are with the Electronic Information School, Wuhan University, Wuhan 430072, China (e-mail: jyma2010@gmail.com).

Digital Object Identifier 10.1109/TIP.2022.3192993

on the local affine consensus. These methods usually achieve high processing speed. However, these methods are subject to spatial local structure, and hence repetitive texture or high outlier ratio is prone to seriously interfere with their matching performance. Alternatively, part of the graph-based methods via relaxed constraints are also available for the mismatch removal problem, such as Graph Shift (GS) [25]. Nevertheless, they are highly sensitive to scale changes and severe deformation, and have a particularly high time loss.

Furthermore, learning-based matching methods have been applied in mismatch removal based on their strengths in learning and expressing deep features. Representatively, Yi *et al.* [26] trained a multi-layer perception-based deep network architecture to identify inliers and recover the relative camera pose simultaneously. However, it is dependent on the input of camera intrinsic parameters. Recent efforts have enhanced network architecture by exploiting more local contexts, such as neighborhood interaction structures [27] and global context clustering [28]. In particular, SuperGlue [29] combines the establishment of putative matches and outlier removal based on Graph Neural Networks (GNN), which in combination with SuperPoint [30] can perform amazingly well. However, the reliance on training data makes the generality of these learning-based methods questionable. Ma *et al.* [31] innovatively devised a two-class classifier for mismatch removal with linearithmic time complexity regarding the data scale. Nevertheless, it is prone to preserve bizarre outliers due to its local topology expressions when images undergo structure deformations and a high outlier ratio.

Despite that many methods have been developed recently for mismatch removal, practical vision tasks still require a robust and efficient approach, whereby challenges mainly exist in the following aspects. On the one hand, the transformation type of an image pair is usually not predictable in advance and is often complex and nonrigid in practice. Therefore, a general feature matching method is required. On the other hand, locality-based methods struggle to cope with repetitive texture and high outlier ratio, and global model-based methods perform poorly in the face of local distortion and independent motion. Hence a robust approach that can handle a wide range of terrible scenarios is of great importance.

To address the above challenges, this paper proposes a novel, robust, and effective model framework for the mismatch removal problem, named as *LOcality-guided Global-preserving Optimization* (LOGO). Unlike classical resampling-based methods, our method constructs an objective that does not rely on the image transformation model as prior information, which represents a two-class assignment problem based on the dynamic correspondence graph. To eliminate the restriction of merely local or global structure in previous methods, we devise a matching strategy that is guided by local topological consistency and gradually explores global structural information. Using the locality consensus to determine an initial match set with high precision,[1] we introduce the Fixed-point Progressive Optimization (FPO) method to expand other inliers for final correspondence set with high recall[2] while not sacrificing precision, which takes advantage of the fact that inliers have reliable geometric reference values and outliers have misleading geometric nature. Hence, our LOGO has great robustness to outliers. In addition, to ensure the matching accuracy, we use local affine transformations as a basis to account for the local consistency and global similarity. This is highly robust and versatile because of topology preservation properties within a small region under various scenarios including nonrigid transformations. Extensive experiments on a wide variety of datasets have demonstrated the advantages of our method over the current state-of-the-art methods in terms of effectiveness, accuracy, and robustness for feature matching, as well as promising applicability for high-level vision tasks, such as essential and fundamental matrix estimation, image registration, and loop closure detection.

The major contributions of this paper are as follows:

- We design a novel graph-based two-class assignment objective for mismatch removal problem, free from reliance on a prior specific transformation model. To solve it, we introduce the fixed-point optimization method with the climbing and convergence properties, which improves the robustness of our method to outliers.
- We devise a guided matching strategy based on the consensus of local affine transformations, where a correspondence set from local topology consensus guides the matching by a global-preserving affinity matrix. The sufficient excavation of topology contributes to the stable and accurate matching performance of our method, even in complex nonrigid transformations.

II. Methodology

Our proposed method is intended to establish one-to-one reliable feature point matching between two images with similar objects or contents. In our feature matching method, a set of putative correspondences is firstly constructed according to the similarity of feature descriptors (*e.g.*, SIFT [11]), where the matching strategy is generally the Nearest Neighbor Distance Ratio (NNDR) [11]. In the following, we concentrate on the mismatch removal stage, *i.e.,* eliminating the outliers and preserving the inliers adequately from the putative set.

A. Problem Formulation

Given a set of N putative feature point correspondences $\mathcal{S} = \{(\boldsymbol{u}_i, \boldsymbol{u}'_i)\}_{i=1}^{N}$ based on the descriptor similarity, where $\boldsymbol{u}_i = (a_i, b_i, 1)^\top$ and $\boldsymbol{u}'_i = (a'_i, b'_i, 1)^\top$ are the spatial homogeneous coordinates of two corresponding feature points, our goal is to distinguish inliers and remove outliers from $\mathcal{S}$. For convenience, $\mathcal{I}$ is denoted as the unknown inlier set, and can be derived from a binary indicator vector $\boldsymbol{x}$, *i.e.,*

$$\mathcal{I} = \{i | x_i = 1, i = 1, \ldots, N\}, \tag{1}$$

and $x_i = 0$ indicates that correspondence $(\boldsymbol{u}_i, \boldsymbol{u}'_i)$ is an outlier.

[1]Precision is the percentage of the number of identified inliers among that of all identified feature correspondences based on the ground truth.

[2]Recall is defined as the percentage of identified inliers among whole inliers based on the ground truth.

In order to retain the correct matches, we devise the following objective, which requires to be maximized to obtain optimum indicator vector $\boldsymbol{x}^*$ of inliers:

$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}} \ \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} - \lambda\|\boldsymbol{x}\|_1, \quad \text{s.t. } \boldsymbol{x} \in \{0,1\}^N, \tag{2}$$

where the first term $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$ is the quadratic score representing the resemblance of the feature correspondence set, and the second term is to discourage matching as a compensation term and avoid the trivial solution of all-inlier (*i.e.,* $\boldsymbol{x} = 1^{1\times N}$). Specifically, $\boldsymbol{A}$ can be considered as an affinity matrix of a graph, and the factor $\lambda > 0$ controls the trade-off between the quadratic score and the compensation term.

Clearly, $\|\boldsymbol{x}\|_1 = \boldsymbol{x}^\top \boldsymbol{x}$, we hence reconstruct the objective Eq. (2) as follows:

$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}} \ \boldsymbol{x}^\top \widetilde{\boldsymbol{A}}\boldsymbol{x}, \quad \text{s.t. } \boldsymbol{x} \in \{0,1\}^N, \tag{3}$$

where $\widetilde{\boldsymbol{A}} = \boldsymbol{A} - \lambda \boldsymbol{I}$ is a non-positive definite matrix, termed as *Full Affinity Matrix*. It is our devised integer quadratic programming formulation of mismatch removal.

B. Fixed-Point Progressive Optimization

Due to the NP-hard nature of Eq. (3), we introduce a Fixed-point Progressive Optimization (FPO) solution, inspired by the classical Integer Projected Fixed Point (IPFP) [7]. To ensure the robustness of LOGO to outliers, FPO primarily embraces an ideology that a set of matches with a high inlier ratio guides to identify other inliers iteratively, thus obtaining the optimum match set. This process is intended to utilize the geometric structure of inliers while avoiding the influence of outliers.

Given an initial indicator vector $\boldsymbol{x}_0$ denoting a set with few outliers but a certain amount of missing inliers, the initial quadratic score is defined as $S^* = \boldsymbol{x}_0 \widetilde{\boldsymbol{A}} \boldsymbol{x}_0$. After initialization, the objective (3) would be solved iteratively. In the k-th iteration, to acquire a discrete solution of Eq. (3) intuitively, we need to derive $\boldsymbol{y}_k$ as the following:

$$\boldsymbol{y}_k = \arg\max_{\boldsymbol{y}} \ \boldsymbol{y}^\top \widetilde{\boldsymbol{A}}\boldsymbol{x}_k, \quad \text{s.t. } \boldsymbol{y} \in \{0,1\}^N, \tag{4}$$

which can be easily found in linear time. Since FPO starts with an indicator vector $\boldsymbol{x}_0$ with high inlier ratio, $\widetilde{\boldsymbol{A}}\boldsymbol{x}_0$ implies the reliable use of topological information. In detail, $\widetilde{\boldsymbol{A}}\boldsymbol{x}_0$ is the calculation of the unary (*i.e.*, the diagonal elements) and pairwise (*i.e.*, the non-diagonal elements) similarity scores associated with the correspondences indicated as inliers by $\boldsymbol{x}_0$. Therefore, this tends to be less susceptible to outliers with low confidence. Hence, $\boldsymbol{y}_0$ is a reliable guide. Likewise, $\boldsymbol{y}_k$ can effectively act on the subsequent iterative process.

Within the continuous domain, the quadratic score $\boldsymbol{x}^\top \widetilde{\boldsymbol{A}}\boldsymbol{x}$ has such second order Taylor expansion around the current solution $\boldsymbol{x}_k$:

$$\boldsymbol{x}^\top \widetilde{\boldsymbol{A}}\boldsymbol{x} \approx \boldsymbol{x}_k^\top \widetilde{\boldsymbol{A}}\boldsymbol{x}_k + 2\boldsymbol{x}_k^\top \widetilde{\boldsymbol{A}}(\boldsymbol{x} - \boldsymbol{x}_k) + (\boldsymbol{x} - \boldsymbol{x}_k)^\top \widetilde{\boldsymbol{A}}(\boldsymbol{x} - \boldsymbol{x}_k). \tag{5}$$

To estimate the last two terms of the above equation around $\boldsymbol{x}_k$, these two values can be found by $\boldsymbol{y}_k$:

$$B = \boldsymbol{x}_k^\top \widetilde{\boldsymbol{A}}(\boldsymbol{y}_k - \boldsymbol{x}_k), \quad C = (\boldsymbol{y}_k - \boldsymbol{x}_k)^\top \widetilde{\boldsymbol{A}}(\boldsymbol{y}_k - \boldsymbol{x}_k). \tag{6}$$

If C is non-negative, $\boldsymbol{y}_k$ may mean a higher quadratic score than $\boldsymbol{x}_k$ and then is assigned to $\boldsymbol{x}_{k+1}$. Otherwise, owing to the restriction of discrete constraints, a choice in the continuous domain between $\boldsymbol{x}_k$ and $\boldsymbol{y}_k$ may be a more appropriate point of convergence than $\boldsymbol{y}_k$. Hence, by setting

$$r = \frac{\|\boldsymbol{x} - \boldsymbol{x}_k\|_2}{\|\boldsymbol{y}_k - \boldsymbol{x}_k\|_2}, \tag{7}$$

we have the following function:

$$f(r) = \boldsymbol{x}^\top \widetilde{\boldsymbol{A}}\boldsymbol{x} \approx \boldsymbol{x}_k^\top \widetilde{\boldsymbol{A}}\boldsymbol{x}_k + 2rB + r^2 C. \tag{8}$$

To find the right point, we get $r^* = -B/C$ by calculating its convexity. Combined with the above, $\boldsymbol{x}_{k+1}$ is decided by

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + r^*(\boldsymbol{y}_k - \boldsymbol{x}_k). \tag{9}$$

Concretely, $\boldsymbol{y}_k$ acts to indicate the direction of the largest possible increase in the discrete domain, as well as the continuous domain. For the sake of rigor, the fact that current $\boldsymbol{y}_k$ can increase the objective quadratic score requires to be verified, and if so, current discrete optimum solution $\boldsymbol{x}^*$ and current maximum quadratic score S^* would be updated. Practically, since FPO initializes a small set with a high inlier ratio, $\boldsymbol{x}_{k+1}$ is normally a set containing more inliers than $\boldsymbol{x}_k$. Therefore, the process of iterations is also a process of expansion of the feature matching set identified by $\boldsymbol{x}_k$, which contributes to the reduction of interference from outliers.

For a given step k, if $\boldsymbol{x}_{k+1}$ is extremely close to $\boldsymbol{x}_k$, we conclude that FPO has converged, with the convergence condition as follows:

$$\frac{\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|_2}{\|\boldsymbol{x}_k\|_2} < \Phi, \tag{10}$$

where Φ is a tiny threshold, empirically set to 10^{-4}. In practice, FPO converges within about ten iterations, and the maximum iterative number is therefore set to 10.

Next, we analyze our proposed method from the perspective of dynamic correspondence graph. To this end, we construct a graph G to represent the topological relationships based on the indicator vector $\boldsymbol{x}$, which is fully connected and directed. Given that $\boldsymbol{x}$ indicates M inliers, *e.g.,* $\{(\boldsymbol{u}_i, \boldsymbol{u}_i')\}_{i=1}^M$, G has M nodes and $M(M-1)$ edges, where each node v_i refers to a feature correspondence $(\boldsymbol{u}_i, \boldsymbol{u}_i')$, and each edge $e_{ij} = (v_i, v_j)$ connects from a node v_i to another node v_j. As G varies according to $\boldsymbol{x}$, we call it a dynamic correspondence graph. In the k-th iteration of FPO, referring to Eq. (4), $\boldsymbol{x}_k$ denotes the correspondence graph G_k, and $\widetilde{\boldsymbol{A}}\boldsymbol{x}_k$ indicates the exploitation of node and edge affinities (*i.e.*, the unary and pairwise similarity scores) within G_k. Fig. 1 provides an illustration of the main idea of the proposed FPO. The feature match set from initial correspondence graph G_0 indicated by $\boldsymbol{x}_0$ has a high precision 99.68% but a low recall 74.34%, which means too many inliers are lost. Through the fixed-point progressive optimization process, our method leans on the reliable guidance from $\boldsymbol{x}_0$ and iteratively mines the affinity matrix $\widetilde{\boldsymbol{A}}$ to expand the dynamic correspondence graph G, and eventually obtains a feature correspondence set with high recall 98.78% at basically no sacrifice to precision. The entire procedure of FPO is outlined in Alg. 1.

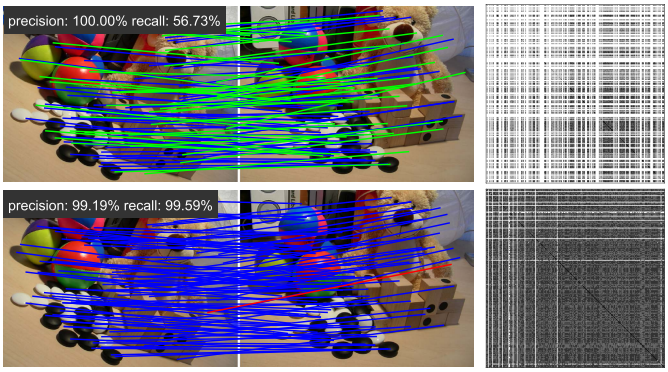

Fig. 1. Illustration of fixed-point progressive optimization. The top row shows an image matching pair and matrix representation of graph structure based on the initial indicator $\boldsymbol{x}_0$, while the bottom row is for the final indicator vector $\boldsymbol{x}$. In each row, the left plot includes feature point correspondences (blue = true positive, green = false negative, and red = false positive), where only 100 randomly selected matches are presented for visibility; the right plot is a part of the visual representation of matrix $\boldsymbol{x}\boldsymbol{x}^\top \odot \boldsymbol{A}$ ($\odot$ denotes the element-by-element product of two matrices), where diagonal gray boxes indicate the nodes contained in graph G, and non-diagonal gray boxes indicate the edges, with black indicating high affinity score and vice versa.

C. Analysis of Growth and Convergence

FPO is fundamentally a series of linear solvers, and thus has linear time complexity with high efficiency, where the next solution of indicator vector $\boldsymbol{x}_{k+1}$ is derived based on the $\boldsymbol{x}_k$ of the previous iteration. Apparently, $\boldsymbol{y}_k$ guides the binary solution $\boldsymbol{x}^*$ in the discrete domain as Eq. (4). However, in the continuous domain, since the first-order Taylor expansion about the objective quadratic score is $\boldsymbol{x}^\top \widetilde{\boldsymbol{A}}\boldsymbol{x} \approx \boldsymbol{x}_k^\top \widetilde{\boldsymbol{A}}\boldsymbol{x}_k + 2\boldsymbol{x}_k^\top \widetilde{\boldsymbol{A}}(\boldsymbol{x}-\boldsymbol{x}_k)$, the discrete solution of maximizing it is equivalent to Eq. (4). Consequently, $\boldsymbol{y}_k$ also provides a maximum possible direction of the iteration in continuous domain, and Lines 10–12 in Alg. 1 further guarantee the growth of quadratic score for $\boldsymbol{y}_k$ in each iteration.

The next argument is that the quadratic score $\boldsymbol{x}_k^\top \widetilde{\boldsymbol{A}}\boldsymbol{x}_k$ is increasing in each iteration while $\boldsymbol{x}_k$ tends to converge. Due to Eq. (4), $B \geq 0$ significantly. In the case of $C \geq 0$, we set $\boldsymbol{x}_{k+1} = \boldsymbol{y}_k$, due to that

$$\begin{aligned} \boldsymbol{x}_{k+1}^\top \widetilde{\boldsymbol{A}}\boldsymbol{x}_{k+1} &= \boldsymbol{y}_k^\top \widetilde{\boldsymbol{A}}\boldsymbol{y}_k \\ &\approx \boldsymbol{x}_k^\top \widetilde{\boldsymbol{A}}\boldsymbol{x}_k + 2B + C \geq \boldsymbol{x}_k^\top \widetilde{\boldsymbol{A}}\boldsymbol{x}_k. \end{aligned} \tag{11}$$

If $C < 0$, $f''(r) = 2C < 0$ based on Eq. (8). Meanwhile, since $f'(0) = 2B \geq 0$, $f(r)$ is concave and $f(r)$ reaches its maximum value when $r^* = -B/C$. But r should be less than 1 to ensure the tendency of convergence to the discrete solution, because $\boldsymbol{x}_{k+1} = \boldsymbol{y}_k$ when $r = 1$. Thus, $r = \min\{-B/C, 1\}$ assures the growth of $\boldsymbol{x}_k \widetilde{\boldsymbol{A}}\boldsymbol{x}_k$ in each iteration. Simultaneously, it is bounded in the feasible domain while monotonically increasing, and hence it must converge.

All in all, by increasing the quadratic score $\boldsymbol{x}_k \widetilde{\boldsymbol{A}}\boldsymbol{x}_k$, the next solution $\boldsymbol{x}_{k+1}$ enables to solve the original problem Eq. (2) better than the previous $\boldsymbol{x}_k$, and eventually our method converges to the optimal solution $\boldsymbol{x}^*$ of the objective Eq. (3).

FPO, as an optimization strategy for LOGO, enables efficient and robust derivation of optimum inlier set. On the other hand, to sufficiently exploit the geometry of feature matches for matching effectiveness, LOGO uses the local affine consensus as a guide to mine reliable information within the global affinity matrix, thus resolving the mismatch removal problem, which is shown in detail in the following sections.

Algorithm 1 : Guided Fixed-Point Progressive Solution

Input: Initial indicator vector $\boldsymbol{x}_0$, full affinity matrix $\widetilde{\boldsymbol{A}}$, maximum iterative number $MaxIterNum$
Output: Final inlier indicator vector $\boldsymbol{x}^*$

```
1  Initialize x* = x_0, S* = x_0^T Ã x_0, k = 0;
2  repeat
3      Calculate y_k using Eq. (4);
4      Update B and C by Eq. (6);
5      if C ≥ 0 then
6          x_{k+1} = y_k;
7      else
8          r = min{−B/C, 1} and set x_{k+1} by Eq. (9);
9      end
10     if y_k Ã y_k > S* then
11         x* = y_k and S* = y_k Ã y_k;
12     end
13     if Eq. (10) then
14         stop and return the solution x* as x;
15     end
16     k = k + 1;
17 until k > MaxIterNum;
```

D. Local Affine Consensus

The effectiveness of LOGO is affected by the calculation of the initialization $\boldsymbol{x}_0$, which can strongly contribute to the matching effect. In actual vision-related tasks, image transformations are in many cases complex and do not globally have a simple correspondence relationship in terms of distance and rotation. However, even under complex nonrigid transformations, the topology of small areas in image pairs tends to be somewhat maintained. Consequently, we use multiple local affine transformations to approximate the image corresponding relationship and thus calculate $\boldsymbol{x}_0$.

To represent topological structure in a rigorous yet universal way, we calculate local affine transformation matrix $\boldsymbol{H}_i$ for each feature point pair $(\boldsymbol{u}_i, \boldsymbol{u}_i')$. For an inlier $(\boldsymbol{u}_i, \boldsymbol{u}_i')$, its mapping relationship should be consistent with the affine matrix $\boldsymbol{H}_i$ obtained from the surrounding feature correspondences, and the opposite for outliers. Therefore, the similarity score between the correspondence $(\boldsymbol{u}_i, \boldsymbol{u}_i')$ can be acquired as:

$$S_i = \frac{2}{1 + e^{\delta \|\boldsymbol{u}_i' - \boldsymbol{H}_i \boldsymbol{u}_i\|_2^2}}, \tag{12}$$

where $\|\cdot\|_2$ is the l_2 norm, and δ is a constant factor affecting the downward trend of the local affine difference $\|\boldsymbol{u}_i' - \boldsymbol{H}_i\boldsymbol{u}_i\|_2^2$ versus the similarity score S_i.

Thereafter, the initial indicator vector $\boldsymbol{x}_0 = \{\boldsymbol{x}_0^1, \boldsymbol{x}_0^2, \cdots, \boldsymbol{x}_0^N\}$ can be determined as follows:

$$\boldsymbol{x}_0^i = \begin{cases} 1, & S_i > \epsilon, \\ 0, & S_i \leq \epsilon, \end{cases} \quad i = 1, \cdots, N, \tag{13}$$

where ϵ acts as a threshold, and its value is relatively large for the high inlier ratio of $\boldsymbol{x}_0$.

To accurately estimate the local affine transformations, we derive $\boldsymbol{H}_i$ based on the neighborhood consisting of reference points. Since the putative matching set inherently contains a proportion of outliers, directly selecting adjacent point pairs as reference points would inevitably utilize the wrong information from outliers, thus prone to obtain a wrong local affine matrix $\boldsymbol{H}_i$. Therefore, it is significant to devise a concise treatment to get reliable feature points as reference points.

For an inlier $v_i = (\boldsymbol{u}_i, \boldsymbol{u}_i')$, the neighborhoods of two feature points $\boldsymbol{u}_i$ and $\boldsymbol{u}_i'$ have many common elements owing to the local consistency. But for an outlier, the neighborhoods in the two images may be dissimilar and have few common elements. Hence, the proportion n_i/K can be simply used as a criterion for the reference points, where n_i is the number of common elements within two K-neighborhoods of feature points $\boldsymbol{u}_i$ and $\boldsymbol{u}_i'$, and the K-neighborhood is derived by the K nearest neighbors (K-NN) method. Consequently, the reference correspondence set $\mathcal{R}$ can be obtained as follows:

$$\mathcal{R} = \left\{ (\boldsymbol{u}_i, \boldsymbol{u}_i') | \frac{n_i}{K} > \tau,\ i = 1, \cdots, N \right\}. \tag{14}$$

Given that the affine matrix has six degrees of freedom up to scale, choosing four reference point pairs can provide more robustness to noises and outliers compared to using only three. Specifically, for a pair of feature nodes $(\boldsymbol{u}_i, \boldsymbol{u}_i')$, we choose four pairs of nearest feature points from the reference point set $\mathcal{R}$ to derive the local affine matrix $\boldsymbol{H}_i$ using the least squares method, and hence $\boldsymbol{H}_i$ is capable of representing the local topological structure about the corresponding point pair $(\boldsymbol{u}_i, \boldsymbol{u}_i')$ accurately.

E. Global Affinity Matrix

Nevertheless, the utilization of merely small area information is undoubtedly one-sided, and drastic changes in the local area can easily affect the matching result. Therefore, we construct the affinity matrix $\boldsymbol{A}$ containing global topological structure. Considering that the correlation between different correspondences is not only related to the similarity, but also to the weight of influence between them, thus $\boldsymbol{A}$ can be yielded from the weight matrix $\boldsymbol{W}$ and similarity matrix $\boldsymbol{C}$.

For the weight matrix $\boldsymbol{W}$, the non-diagonal term W_{ij} as the weight of each edge $e_{ij} = (v_i, v_j)$ is assigned to quantify the influence of the tail correspondence v_j on the head correspondence v_i. The higher the edge weight W_{ij} is, the higher the reference value v_j has for the v_i, which is intuitively related to the distance between two correspondences v_i and v_j. In other words, the closer the two correspondences are to each other, the higher the reference value. The detailed formula is as follows:

$$\begin{aligned} d_{ij} &= \frac{\left\| \boldsymbol{u}_i - \boldsymbol{u}_j \right\|_2^2}{\| \max(\boldsymbol{u}) - \min(\boldsymbol{u}) \|_2^2} + \frac{\left\| \boldsymbol{u}_i' - \boldsymbol{u}_j' \right\|_2^2}{\| \max(\boldsymbol{u}') - \min(\boldsymbol{u}') \|_2^2}, \\ W_{ij} &= \frac{2}{1 + e^{\frac{d_{ij}}{\sum_{j=1}^{N} d_{ij}}}}, \text{ and } W_{ii} = 1,\ i, j = 1, \cdots, N, \end{aligned} \tag{15}$$

where d_{ij} is derived from the normalized distance between two correspondences, and the edge weight W_{ij} also requires a normalization, thus avoiding interference of image size or correspondence number and distribution for generality.

The similarity matrix $\boldsymbol{C}$ denotes the consistency of geometric topology between nodes and edges. Given that the local affine transformation $\boldsymbol{H}_i$ well indicates the transforming feature of a correspondence $v_i = (\boldsymbol{u}_i, \boldsymbol{u}_i')$, the second-order edge similarity score S_{ij} that measures the topological consistency between two feature correspondence $v_i = (\boldsymbol{u}_i, \boldsymbol{u}_i')$ and $v_j = (\boldsymbol{u}_j, \boldsymbol{u}_j')$ can be written as:

$$\begin{aligned} S_{ij} &= \frac{2}{1 + e^{\delta(\|\boldsymbol{u}_i' - \boldsymbol{u}_j'\|_2^2 - \|\boldsymbol{H}_i \boldsymbol{u}_i - \boldsymbol{H}_j \boldsymbol{u}_j\|_2^2)}}, \\ &\text{s.t.,} \quad i \neq j, \quad i, j = 1, \ldots, N. \end{aligned} \tag{16}$$

S_{ij} is the result from the difference between the actual distance of two mapping points (*i.e.,* $\boldsymbol{u}_i'$ and $\boldsymbol{u}_j'$) and the distance after local affine mapping (*i.e.,* $\boldsymbol{H}_i \boldsymbol{u}_i$ and $\boldsymbol{H}_j \boldsymbol{u}_j$), and a larger value of S_{ij} indicates a higher affine consensus regarding edge e_{ij}.

According to Eq. (16), the edge similarity C_{ij} can be formulated with a predefined threshold ζ as follows:

$$C_{ij} = \begin{cases} 1, & S_{ij} \geq \zeta, \\ 0, & S_{ij} < \zeta. \end{cases} \tag{17}$$

Since the first-order similarity score S_i denotes how compatible the feature point $\boldsymbol{u}_i$ is with its corresponding point $\boldsymbol{u}_i'$ by Eq. (12), the diagonal elements of similarity matrix $\boldsymbol{C}$ can be defined as:

$$C_{ii} = S_i. \tag{18}$$

Each node affinity score represents the consistency of a pair of feature points, and the edge affinity score denotes the similarity reference value of one correspondence to another. As a consequence, the affinity matrix $\boldsymbol{A}$ is constructed as follows:

$$\boldsymbol{A} = \boldsymbol{W} \odot \boldsymbol{C}, \tag{19}$$

where $\odot$ is the element-by-element product of two matrices.

Evidently, our affinity matrix covers the geometric consistency score within the global topology and is local affine-invariant such that it can handle complex nonrigid images.

The procedure of local affine consensus for initialization and global affinity matrix construction is described in Alg. 2.

F. Differences From Related Mismatch Removal Methods

The proposed LOGO is related to the recently developed locality-based and graph-based matching methods such as LPM [20] and LGSC [21], but LOGO has its technical novelty. Both LPM and LGSC are based on local topological consistency, with the former pioneering a mathematical model that preserves local neighborhood structure and the latter proposing a novel local graph structure to represent neighborhood geometry. However, the solutions of both degenerate from the global to the local domain, *i.e.,* in essence the correctness of a putative feature point pair is determined by only two corresponding

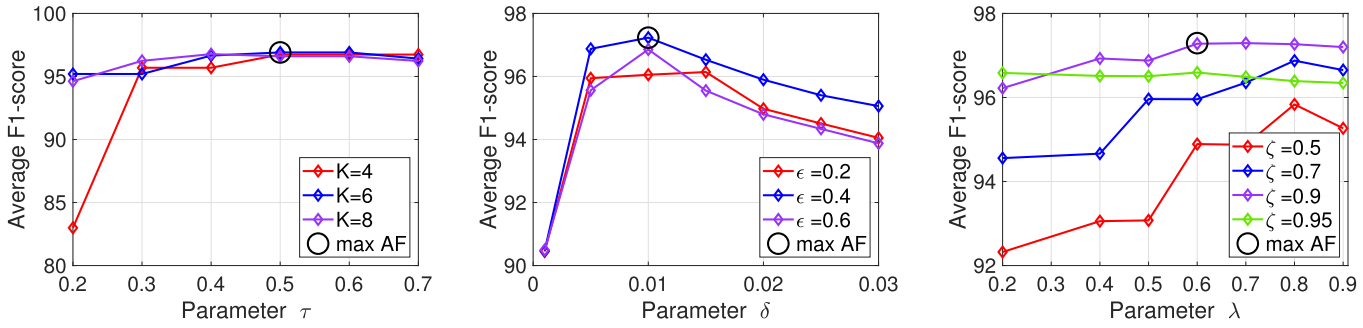

Fig. 2. Average F1-score of LOGO regarding different values of parameters, *i.e.*, τ, K, δ, ϵ, λ, and ζ. Three groups of experiments are conducted by adjusting two parameters (as in the first plot, τ and K), while fixing other parameters (as δ, ϵ, λ, and ζ) to proper values.

Algorithm 2 : Local Affine Consensus for Initialization and Global Affinity Matrix Construction

Input: Putative set $\mathcal{S}$, parameters τ, δ_a, ϵ, δ_d, ζ
Output: Initial indicator vector $\boldsymbol{x}_0$, affinity matrix $\boldsymbol{A}$
1 Find reference correspondence set $\mathcal{R}$ using Eq. (14);
2 Select four correspondences from $\mathcal{R}$ using K-NN on $\{\boldsymbol{u}_i\}_{i=1}^N$;
3 Calculate affine matrices $\{\boldsymbol{H}_i\}_{i=1}^N$ via least squares;
4 Calculate similarity score S_i for each correspondence $(\boldsymbol{u}_i, \boldsymbol{u}_i')$ using Eq. (12);
5 Determine initial indicator vector $\boldsymbol{x}_0$ using Eq. (13);
6 Calculate weight matrix $\boldsymbol{W}$ using Eq. (15);
7 Calculate similarity matrix $\boldsymbol{C}$ via Eqs. (16)-(18);
8 Calculate affinity matrix $\boldsymbol{A}$ based on Eq. (19).

neighborhoods, which has limited generality and robustness despite computational simplifications.

In contrast, the objective function of LOGO is constructed based on a globally scoped dynamic correspondence graph, and the FPO is specifically devised to optimize this objective. In terms of topological geometry, we use local affine consistency to guide the mining of global topology, ensuring the generality of matching; in terms of solution strategy, we introduce a high-precision-set-guided matching strategy to enhance the robustness to outliers. Thereby, innovations in modeling and solving allow LOGO to generally deal with a variety of scenarios, including difficult wide baselines, multiple motions, and complex nonrigid and can cope with high outlier matching requirements. In the experimental section, we will demonstrate the superiority of the proposed LOGO over previous related methods qualitatively and quantitatively.

G. Implementation Details

In the implementation of LOGO, there are six parameters, *i.e.*, K, τ, δ, ϵ, ζ, and λ. To determine the values of parameters and verify their universality, we randomly select 45 pairs of images as a test set, which includes various transformations such as piecewise linear, wide baseline, complex nonrigid, *etc*. The average inlier number and percentage of testing images after SIFT matching are about 855 and 57.7%, respectively.

As shown in Fig. 2, we apply LOGO with different values of parameters to the test set, with F1-score[3] as the metric. In each figure, values of two parameters are adjustable and analyzed, while others are fixed to the suitable values. Parameter K is the number of nearest neighbors by K-NN, and together with τ as a threshold determining the reference correspondence set $\mathcal{R}$. Clearly, according to Eq. (14), when K is large or τ is small, the number of reference correspondences will increase. In the first plot of Fig. 2, LOGO is relatively robust to the choice of K value, a small τ value can lead to a too sparse set $\mathcal{R}$ and thus is not conductive to matching, and the average F1-score reaches its maximum at $K = 6$ and $\tau = 0.5$. Parameter ϵ determines the initial indicator vector $\boldsymbol{x}_0$ as a threshold, and a high value of ϵ would cause sparse but pure correspondence set in $\boldsymbol{x}_0$. Parameter δ affects the calculation of similarity scores based on local affine consensus. In the second plot, the case that $\delta = 0.01$ and $\epsilon = 0.4$ assists LOGO to achieve the best matching performance. The third plot shows the effect of parameters λ and ζ on the proposed LOGO, where LOGO is stable for the value of $\lambda = 0.6$ and has the highest F1-score when $\zeta = 0.9$. Parameter ζ solves the edge similarity as a threshold, and parameter λ controls the balance between quadratic score and compensation term in the objective as Eq. (2) and determines the full affinity matrix $\widetilde{\boldsymbol{A}}$ in Eq. (3). A large λ inclines to improve the precision and suppress the recall. In summary, we empirically set the default values as $K = 6$, $\tau = 0.5$, $\delta = 0.01$, $\epsilon = 0.4$, $\zeta = 0.9$, and $\lambda = 0.6$.

H. Computational Complexity

The guided fixed-point optimization method in Alg. 1 has $O(N)$ time complexity due to the linear operations during each iteration. In Alg. 2, Line 1 using K-D tree [32] costs the time complexity $O((K+N)\log N)$. Furthermore, Lines 2–6 have the time complexity $O(N)$. To construct global affinity matrix $\boldsymbol{A}$, Lines 7-9 cost the time complexity $O(N^2)$. Additionally, the space complexity of our method lies mainly in the affinity matrix construction, which requires $O(N^2)$ memory space.

III. Experiments

In this section, we carry out sufficient experiments for testing and analysis regarding our proposed LOGO. Firstly,

[3]F1-score $= 2 \times$ Precision $\times$ Recall$/$(Precision $+$ Recall).

TABLE I

THE TESTING RESULTS OF FPO ON RIGID, MULTI-MOTION, AND NONRIGID TESTING IMAGES. PRECISION (%) AND RECALL (%) ARE REGARDED AS THE METRICS TO EVALUATE THE MATCHING RESULTS. **BOLD** INDICATES THE BEST

Algorithm		Rigid		Multi-motion		Nonrigid	
		P/R	F	P/R	F	P/R	F
RANSAC [14]		97.7/90.0	93.7	95.0/70.7	81.1	91.7/88.0	89.8
RANSAC-FPO		97.0/92.1	94.5	93.1/83.4	88.0	90.9/91.7	91.3
LPM [20]		96.8/97.2	97.0	95.0/95.4	95.1	91.2/92.4	91.7
LPM-FPO		97.1/94.3	95.7	95.2/94.7	94.9	90.8/91.7	91.2
VFC [17]		97.8/90.1	93.8	85.3/68.6	76.0	**96.0**/89.5	92.6
VFC-FPO		97.1/92.7	94.8	83.4/81.0	82.2	92.8/91.9	92.3
LOGO	No RPS	94.3/77.5	85.1	93.5/58.3	71.8	93.1/54.6	68.8
	Rand. Init.	93.1/91.7	92.4	94.2/93.3	93.7	91.6/91.3	91.4
	No FPO	**98.3**/78.8	87.5	**95.7**/90.7	93.1	93.7/86.1	89.7
	Completed	**98.3/97.4**	**97.8**	95.6/**95.5**	**95.5**	93.7/**94.0**	**93.8**

TABLE II

THE QUANTITATIVE COMPARISON OF CONSENSUS MAXIMIZATION POST-PROCESSING METHODS ON RIGID, MULTI-MOTION, AND NONRIGID TESTING IMAGES. #INLIERS (*i.e.*, $|\mathcal{I}|$) AND RUNTIME (SECOND) ARE REGARDED AS THE METRICS TO EVALUATE THE PROCESSING RESULTS. **BOLD** INDICATES THE BEST

Algorithm		IRLS [34]	EP [35]	ADMM [36]	IBCO [37]	FPO
Rigid	$\lvert\mathcal{I}\rvert$	353	117	340	356	**364**
	Time (s)	**0.011**	15.43	284.0	0.784	0.014
Multi-motion	$\lvert\mathcal{I}\rvert$	96	71	158	106	**178**
	Time (s)	0.008	4.498	955.7	1.181	**0.001**
Nonrigid	$\lvert\mathcal{I}\rvert$	112	57	285	171	**327**
	Time (s)	**0.010**	28.29	1410	7.172	0.024
Average	$\lvert\mathcal{I}\rvert$	187	82	261	211	**290**
	Time (s)	**0.010**	16.07	883.2	3.046	0.013

we demonstrate and analyze the role of each block of LOGO. In particular, we carry out the comparisons of inlier maximization methods. Secondly, extensive experimental evaluation of our LOGO for feature matching are presented, as well as the comparisons with other state-of-the-art methods, robustness to outliers, and generality to different descriptors. Ultimately, we apply LOGO to vision-based applications, *i.e.*, essential and fundamental matrix estimation, image registration, and loop closure detection. The open-source VLFeat toolbox [33] is applied for SIFT and NNDR with threshold 1.5, as well as K-NN searching based on K-D tree. All experiments are performed on a desktop with a 2.90 GHz Intel Core CPU, 16 GB memory, and MATLAB R2022a code.

A. Ablation Study

The innovative blocks of our LOGO mainly include reference point set, initialization of local consistency, and fixed-point progressive optimization. To evaluate in detail the role of each technical design, we conduct the ablation experiments on various types of image pairs, *i.e.,* rigid, multi-motion, and nonrigid. With the precision, recall, and F1-score as the evaluation metrics, each image type has 30 pairs of images. Experiment results are presented in Table I.

Reference Point Set (RPS) is a design imposed in order to calculate local affine transformations accurately. Estimating local geometric properties based on the set of reference points gets rid of the interference of outliers inside the neighborhood, thus guaranteeing the reliability of affine matrices constructed. In Table I, by comparing the first and fourth rows, we can find that the absence of reference points causes a reduction in precision and recall, especially the latter. This is due to the increased error of the local affine matrix caused by the outliers, leading to a reduction in the perceived inliers.

Initialization of Local Consistency is that an initial indicator vector is determined by a threshold for local affine consensus before optimization (*i.e.,* Eq. (13)), the aim of which is to accelerate convergence and avoid falling into unreasonable local optima. In Table I, the second row shows the matching results of our method using the initialization based on a random indicator vector, which is assumed as uniformly distributed in [0, 1]. Compared with the fourth row, the initialization based on the threshold method can facilitate the optimization and thus improve the matching effect.

Fixed-point Progressive Optimization (FPO) is specifically devised to handle our proposed objective (*i.e.,* Eq. (3)), finding the optimal solution of the indicator $\boldsymbol{x}$ by an affinity matrix that holds global information. To concretely demonstrate the effectiveness of FPO, we evaluate the feature matching results of our method without and with FPO. In practice, when there is no FPO, the feature matching result of our method is donated by the initial indicator $\boldsymbol{x}_0$, which is determined by local affine consensus. The testing results are presented in Table I. According to it, FPO is little effective for multi-motion image pairs, which is because even two adjacent objects can present vastly different motions so global similarity scores draw little significance. However, for both rigid and nonrigid transformations, FPO exerts a marked improvement.

To further assess the effectiveness of our method concerning the FPO, three representative comparative methods are selected and the matching results with FPO are compared, with the results reported in Table I. It can be seen from it that FPO is not a significant improvement over the other methods. This is because the optimization of FPO lies mainly in the geometric constraint of the affinity matrix, and it is suitable to be paired with our initialization based on local affine consistency, which ensures the effectiveness of our method.

FPO enables an increase in the number of inliers in linear time, but there are many consensus maximization methods, including standard Iterative Re-weighted Least Squares (IRLS) [34], Exact Penalty method (EP) [35], Alternating Direction Method of Multipliers (ADMM) [36], and Iterative BiConvex Optimization (IBCO) [37]. Using the local affine consensus-based initialization of our LOGO, we later use different post-processing methods to increase inliers, including IRLS, EP, IBCO, ADMM, and our FPO. It is worth noting that we have used the *gurobi* solver as suggested in the original paper for best performance. With the three types of images as testing objects, the evaluation metrics include the number of inliers (#inliers) and runtime, and the experimental results are shown in Table II. Since the maximum consensus methods

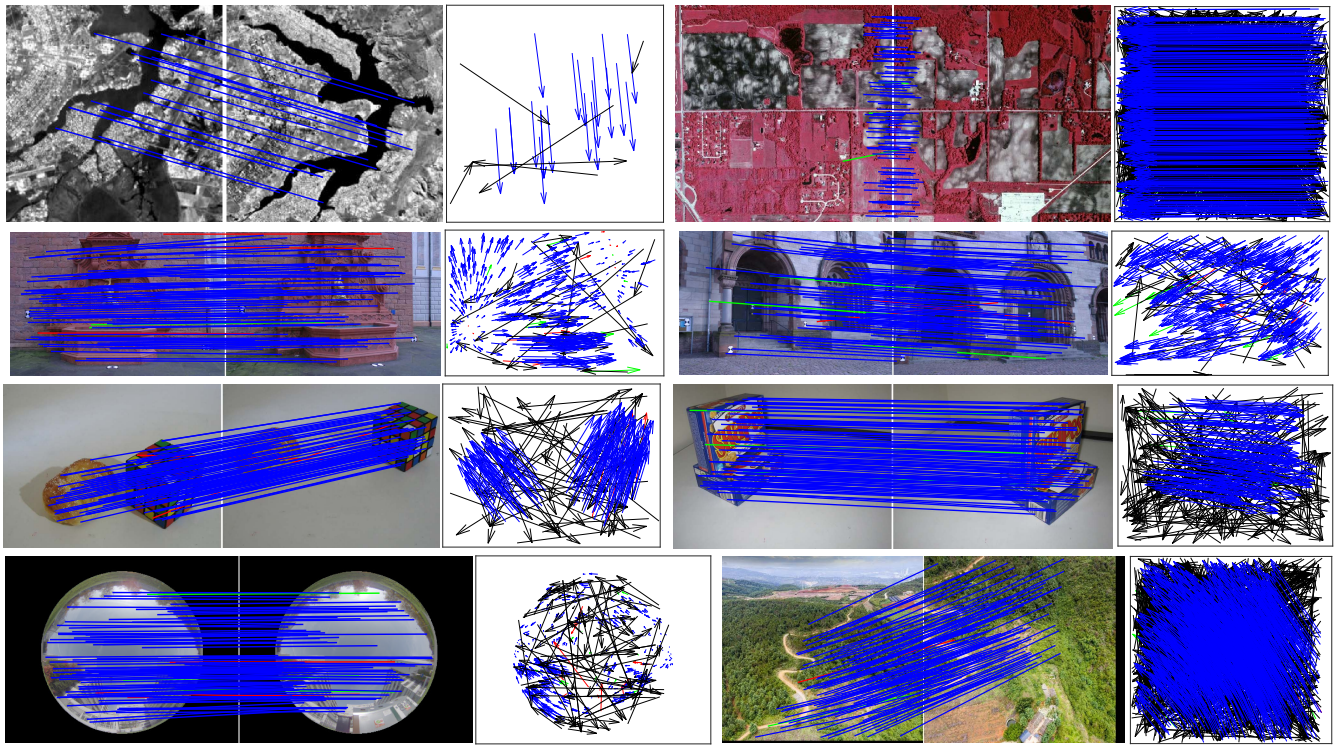

Fig. 3. Feature matching results of LOGO on 8 representative image pairs. From top to bottom, left to right: *Patch*, *Land*, *Fountain*, *Architecture*, *Breadcube*, *Biscuit*, *University*, and *Yun*. The inlier ratios of the 8 image pairs are 76.92%, 13.32%, 87.56%, 86.53%, 65.96%, 81.57%, 79.34%, and 41.46%. 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.

all rely on fixed model types (*e.g.,* linear, homography, and fundamental matrix), they have difficulty in dealing with multi-motion and nonrigid situations in feature matching, and tend to fall into the trap of confusing solutions with long runtime. It can be seen that our FPO can maximize the inlier consensus while ensuring high operational efficiency.

B. Feature Matching

LOGO is intended to establish feature correspondences between a pair of images with similar contents or objects. To qualitatively and quantitatively examine the accuracy, robustness, and efficiency of our LOGO in feature matching, we apply LOGO to some representative image pairs and datasets and compare it with other state-of-the-art methods. The multi-type image datasets used are as follows.

- *RS* [20]. This dataset contains 156 pairs of remote-sensing images, including color-infrared, SAR, and panchromatic photographs. These image pairs all fulfill the parametric model transformations and are generally applied to image mosaic, objective positioning, change detection, *etc*.
- *Daisy* [38]. It has 52 wide baseline image pairs with ground-truth depth maps. These images are mainly about two short image sequences and some individual images.
- *AdelaideRMF* [39]. This dataset contains 38 image pairs, the first 19 of which conform to the homography transformation and the next 19 to the fundamental matrix. In addition, most image pairs satisfy multi-motion patterns, *i.e.*, multiple objects have different motions, which is challenging for many methods.
- *Nonrigid*. It consists of two nonrigid transformation subdatasets, which are *720YUN* [40] and *FE* [41]. The former involves 20 image pairs of terrain, roads, buildings, terraces, *etc*. FE was acquired by a fish-eye camera on university and urban scenes, with 32 image pairs in total. These 52 pairs of images are prone to a high computational burden due to complex transformation constraints.

1) Qualitative Results: Our LOGO processes eight representative image pairs undergoing various transformations, the feature matching results of which are displayed in Fig. 3. The *Patch* and *Land* in the first row are two aerial photograph pairs with only linear (*e.g.,* rigid or affine) transformation from the *RS* dataset, where *Patch* has a sparse putative feature set. The *Fountain* and *Architecture* in the second row are from the *Daisy* dataset, which have undergone wide-baseline imaging. *Breadcube* and *Biscuit* come from the *AdelaideRMF* dataset and fall into multi-motion transformations. In the bottom row, *University* and *Yun* belong to complex nonrigid transformation, which are from *FE* and *720YUN* datasets, respectively. For each group of qualitative illustrations, the left figure is the feature matching result on the image pair, and the right plot is the motion field which shows the correctness of each correspondence from the putative feature matching set, visualized in color.

As mentioned in Section II, SIFT is applied to construct the putative correspondence set, which however contains a

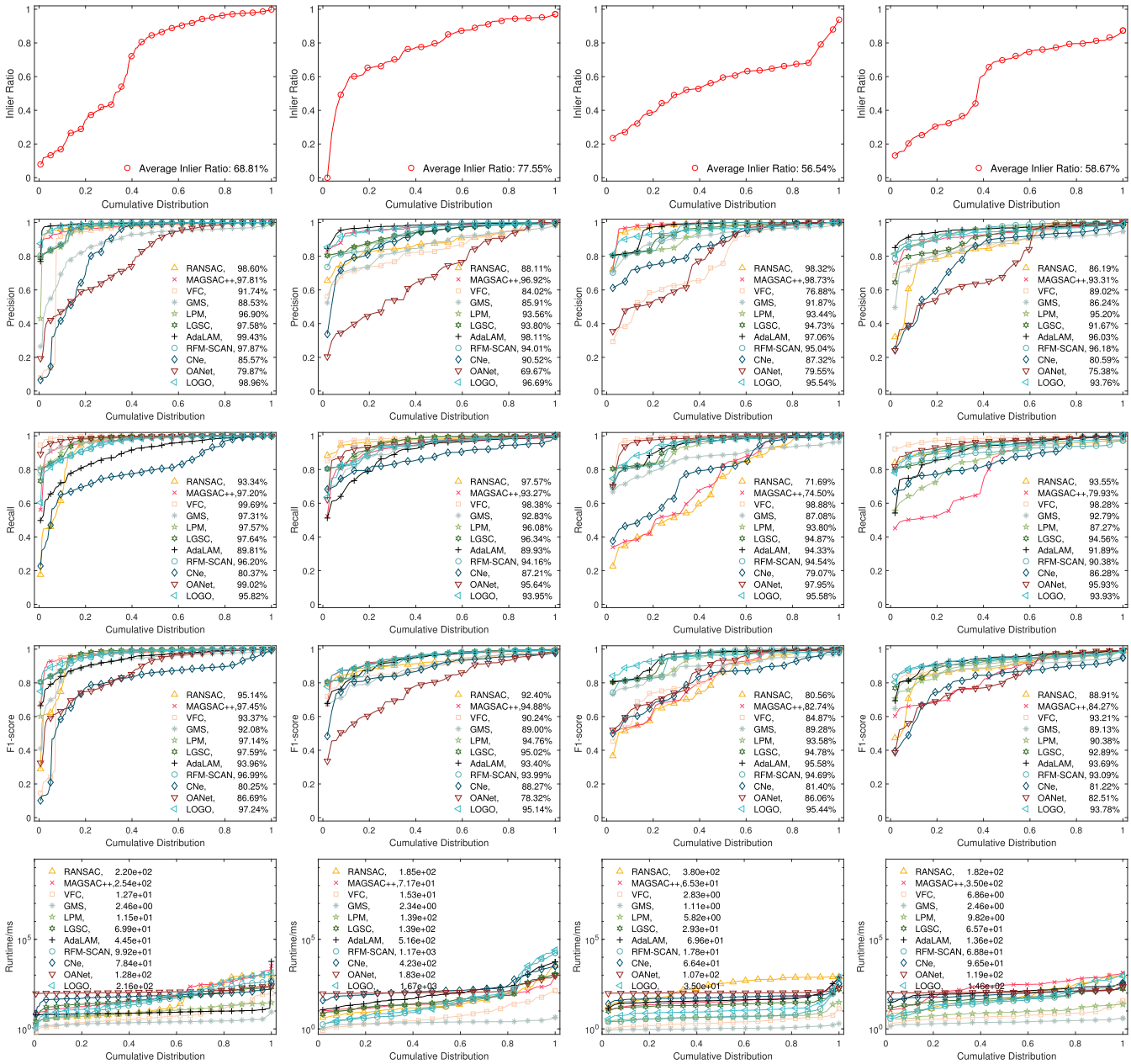

Fig. 4. Quantitative comparisons of RANSAC, MAGSAC++, VFC, GMS, LPM, LGSC, AdaLAM, RFM-SCAN, CNe, OANet, and our LOGO on four image sets. From left to right: *RS*, *Daisy*, *AdelaideRMF* and *Nonrigid* datasets. From top to bottom: initial inlier ratio, precision, recall, F1-score, and runtime concerning the cumulative distribution. A point on the curve with coordinates (x, y) denotes that there are $100 * x$ percent of image pairs that have the metric value (*i.e.*, Inlier Ratio, Precision, Recall, F1-score, and Runtime, respectively from top to bottom) no more than y, and the average values of the four image datasets for each comparing method are shown in the legend accordingly. The average values on the legend are best viewed zoomed in.

non-negligible number of mismatches due to the incomplete feature descriptions. From top to bottom, left to right, the inlier numbers and ratios of 8 representative image pairs are (20, 76.92%), (306, 13.32%), (401, 87.56%), (366, 86.53%), (310, 65.96%), (571, 81.57%), (457, 79.34%), and (638, 41.46%), respectively. These putative feature sets cover a wide range of sparse features, low inlier ratio, large data, *etc.*, which poses considerable challenges to the mismatch removal task. The proposed LOGO method processes the eight putative feature correspondence sets, eliminates the mismatches, and preserves the correct matches. Accordingly, the precision and recall of eight mismatch removal results are (100%, 100%), (100%, 97.06%), (97.03%, 97.76%), (98.34%, 97.27%), (97.63%, 100%), (100%, 97.26%), (97.14%, 96.72%), and (97.68%, 98.01%). The relevant feature matching results are displayed in Fig. 3, where our LOGO accurately finds and sufficiently preserves the correct correspondence between representative image pairs.

2) Quantitative Comparisons: To further evaluate the feature matching effectiveness of LOGO in a holistic and circumstantial way, we use LOGO to handle four different types of image dataset, *i.e.*, *RS*, *DAISY*, *AdelaideRMF*, and *Nonrigid*. For each dataset, the average inlier numbers are 1475.6, 445.3, 693.2, and 378.8 respectively, and the average

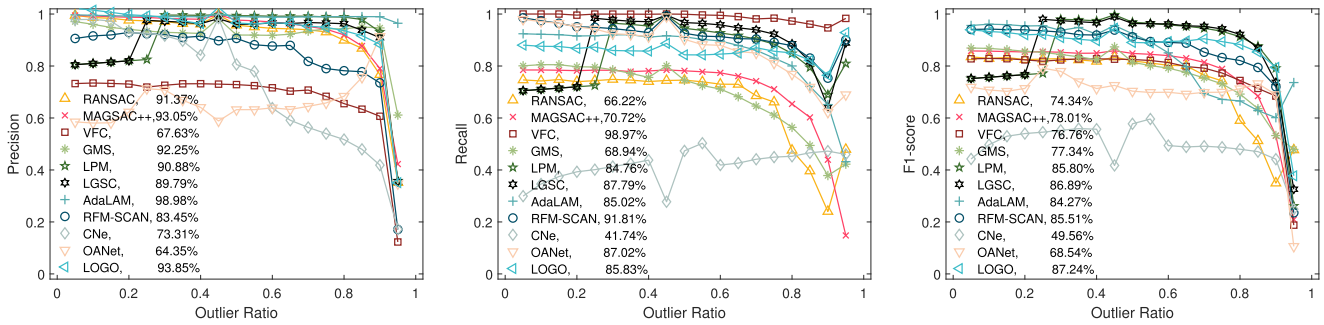

Fig. 5. Robustness testing results in the outlier ratio of putative matching sets. Precision, recall, and F1-score are used as metrics to evaluate the performance of these 11 comparing methods at different outlier ratios.

inlier ratios are shown in the first row of Fig. 4. In addition, we compare our LOGO with other 10 state-of-the-art methods, *i.e.*, RANSAC [14], MAGSAC++ [16], VFC [17], GMS [42], LPM [20], LGSC [21], AdaLAM [24], RFM-SCAN [22], CNe [26], and OANet [28]. These eleven methods are representative of the leading methods in the mismatch removal, in brief, RANSAC is the classic resample-based method, MAGSAC++ is its advanced version with a new model quality function, VFC is a non-parametric model-based approach for robust feature matching, GMS and LPM are locality-based methods with high speed, LGSC is to preserve the local topology as graph structure, RFM-SCAN solves the mismatch removal rejection problem by spatial clustering, AdaLAM is an efficient outlier removal method based on the local affine consensus, CNe and OANet are the pioneering deep learning-based methods. All implementations of these methods come from publicly available codes, and the parameters of models have been adjusted to suitable values for best performance. The precision, recall, F1-score, and runtime, as metrics, objectively assess the performance of each method on four datasets, and corresponding cumulative distributions are displayed in the last four rows of Fig. 4. In particular, precision directly describes the representation of mismatch removal, *i.e.,* the purity of the resulting match set. Too low a precision can lead to an accumulation of errors in high-level vision tasks, such as incorrect pose estimation and obvious alignment bias in image registration. Recall describes the ability of mismatch removal to preserve inliers. A low recall would undermine the representation of the match set in subsequent tasks, *e.g.,* many misaligned regions in image registration.

As shown in Fig. 4, RANSAC performs stably, retaining the inliers adequately based on parametric model estimation, but encounters significant limitations when faced with non-parametric transformations. MAGSAC++ is similar to RANSAC in terms of matching effects but has a notable improvement in efficiency with marginalizing samples, especially for high numbers of feature matches. Although VFC formulates the mismatch removal as a maximum-likelihood estimation for both parametric and non-parametric estimations, it is limited by the global modeling and therefore cannot handle images with multiple movements with poor performance for *AdelaideRMF*. GMS is the most efficient as a grid-based method, yet is less effective for complex transformations and low inlier ratio data, and therefore has low robustness. LPM establishes a simple but effective model based on locality preserving, but its topological representation is related and does not make use of global information. LGSC further optimizes the local topological constraint, but still struggles when faced with complex nonrigid transformations. CNe introduces deep learning, which is groundbreaking in terms of theoretical research, but the relevant geometric details are immature and rely on a prior of camera parameters, thus not yet competitive. OANet as a follow-up deep learning method shows some improvement with complex constraints, but the matching performance is still not good enough. The training requirements of deep learning methods for specific tasks and datasets lead to their low generality. In particular, our LOGO has the best feature matching performance in precision-recall trade-off with the highest F1-score and is robust for various image transformations and imaging scenarios.

3) Robustness to Outliers: As our FPO can promote the robustness of LOGO to outliers, we test our method on different outlier ratios and compare it to the other ten state-of-the-art methods. Three typical and complex image pairs that conform to homography, wide-baseline, and multi-motion transformations are chosen, and the numbers of inliers are 854, 565, and 71, respectively. For each image pair, we randomly remove or add outliers, so that the outlier ratios vary from 0.05 to 0.95 in intervals of 0.05 and there are 20 sets of feature points for one image pair at each outlier rate. The precision, recall and F1-score as metrics objectively assess the performance of each method under different outlier ratios, with the results shown in Fig. 5. From it, our LOGO has the highest F1-score, and hence has satisfactory robustness to outliers.

4) Generality to Different Descriptors: Our method concentrates on the rejection of mismatches as a part of feature matching, which primarily exploits the geometric structure between image transformations. Traditional feature descriptors such as SIFT [11], SURF [13], and KAZE [43] have proven the practical value in vision-based tasks and are popular and widely used, *e.g.,* the strong stability of SIFT-based feature descriptors for 3D reconstruction by structure-from-motion. Therefore, the matching experiments in this paper are usually set up to cope with the match set after SIFT processing. However, there have been a number of deep

TABLE III

GENERALITY TESTING FOR HANDCRAFTED AND LEARNED DESCRIPTORS. THE ACCURACY OF HOMOGRAPHY MATRIX ESTIMATION AND RUNTIME OF THE WHOLE PROCEDURE ARE REPORTED AS METRICS, WHERE THE FRONT IN PERCENT % AND THE LAST IN SECONDS. **BOLD** INDICATES THE BEST

Descriptors	Matcher	Accuracy	Runtime
SIFT [11]	NNDR+MAGSAC++ [16]	69.2	1.919
	NNDR+LPM [20]	63.1	**1.887**
	NNDR+AdaLAM [24]	69.2	2.147
	NNDR+OANet [28]	67.7	2.018
	NNDR+**LOGO**	**70.8**	2.208
SURF [13]	NNDR+MAGSAC++ [16]	61.5	0.351
	NNDR+LPM [20]	**64.6**	**0.307**
	NNDR+AdaLAM [24]	61.5	0.430
	NNDR+OANet [28]	56.9	0.432
	NNDR+**LOGO**	58.5	0.461
SuperPoint [30]	NNDR+MAGSAC++ [16]	55.4	2.582
	NNDR+LPM [20]	60.0	3.325
	NNDR+AdaLAM [24]	55.4	**2.759**
	NNDR+OANet [28]	58.5	3.460
	NNDR+**LOGO**	63.8	3.258
	SuperGlue [29]	**70.8**	12.52
HardNet [44]	NNDR+MAGSAC++ [16]	47.7	**1.717**
	NNDR+LPM [20]	46.2	3.151
	NNDR+AdaLAM [24]	41.5	2.189
	NNDR+OANet [28]	**49.2**	3.296
	NNDR+**LOGO**	46.2	3.054
SOSNet [45]	NNDR+MAGSAC++ [16]	46.2	**1.718**
	NNDR+LPM [20]	44.6	3.153
	NNDR+AdaLAM [24]	43.1	2.136
	NNDR+OANet [28]	**47.7**	3.305
	NNDR+**LOGO**	46.2	3.150

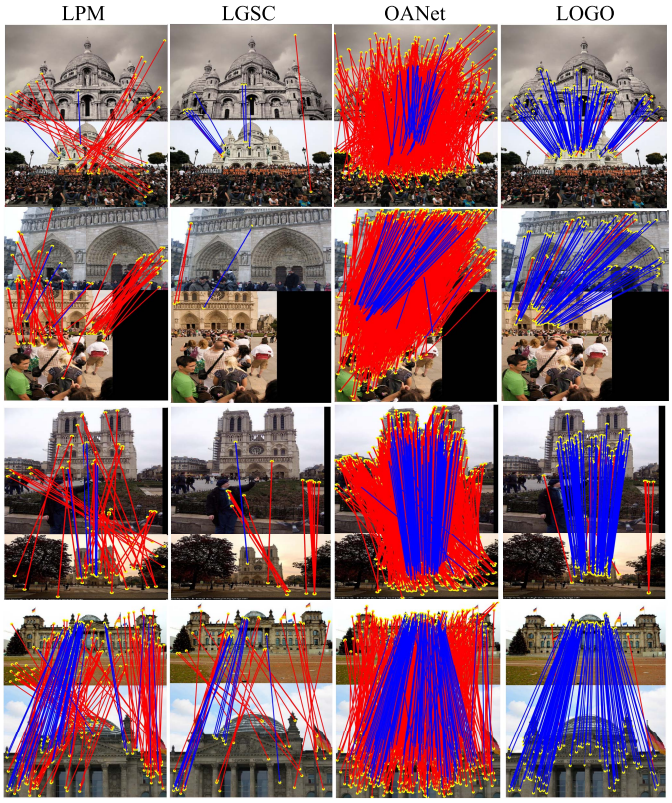

Fig. 6. Qualitative comparisons on the *YFCC100M* dataset for essential matrix estimation. From left to right: LPM, LGSC, OANet, and LOGO. There are four representative image pairs presented. In each pair of images, the yellow dots represent feature points, the blue lines are correctly identified matches, and the red lines are incorrect matches.

learning-based feature description methods recently, such as HardNet [44] and SOSNet [45] that have demonstrated distinctive description capabilities on experimental data. Moreover, the SuperPoint [30] + SuperGlue [29] pipeline has been proven significantly superior. Nevertheless, traditional mismatch removal methods still have important value for the matching problem due to the data-independent applicability and descriptor-independent generality.

We test two publicly available datasets, VGG [46] and Hannovor [47], which contain a total of 65 pairs of images and cover a wide range of deformation levels. Various combinations of feature descriptors and matching methods are used to design a generality testing experiment. The putative matches are established by the NNDR strategy with the threshold of 1.2. It is worth mentioning that for a unified computing environment, SuperGlue runs on the CPU. The datasets provide homography matrices as ground truth, so we estimate the homography matrices based on the matching results of each feature matching method, where the model estimation method is RANSAC with an inlier-outlier threshold as 4. The error is calculated based on the estimated homography matrix and the ground truth, and a match with an error of less than 4 pixels is considered accurate and the proportion is used as the evaluation metric, *i.e.,* Accuracy. In addition, we also consider the runtime of the whole procedure including descriptor construction, NNDR and outlier removal.

The matching results are shown in Table III, where our LOGO shows a competitive performance. The SuperPoint+SuperGlue pipeline has some dependence on training data, and the high demands on the computing environment constrain the runtime. Compared to the end-to-end matching of feature sets, the mismatch removal methods have high universality for local features. Compared to deep-learning-based methods, the traditional feature matching has no pressing need for computational resources and can perform efficiently with complex theoretical topological constraints.

C. Essential Matrix Estimation

Concerning the feature matching task, one of the most important applications is the recovery of information about the scene, such as 3D structure. In this process, feature matching methods are used in combination with robust estimators, such as RANSAC, in order to estimate the essential matrix and thus recover the relative pose of the camera. A robust feature matching algorithm will significantly improve the accuracy of the estimation and facilitate subsequent operations.

Yahoo's *YFCC100M* [48] collects 100 million photos from Internet. Heinly *et al.* [49] organized them into 72 scenes reconstructed with the VisualSfM software and provided bundle adjusted camera poses, intrinsics, and triangulated point clouds. In deep learning matching research, this dataset is typically divided into 68 sequences for training and 4 sequences

TABLE IV

THE QUANTITATIVE COMPARISON OF ESSENTIAL MATRIX ESTIMATION ON THE *YFCC100M* DATASET. **BOLD** INDICATES THE BEST

Method	AUC		P/R	F
	@5°	@10°		
RANSAC [14]	22.3	31.4	23.7/43.9	30.8
MAGSAC++ [16]	24.9	33.7	24.3/55.3	32.8
VFC [17]	36.8	44.4	51.2/48.0	48.1
GMS [42]	35.8	44.7	54.6/68.7	43.3
LPM [20]	40.3	49.2	51.4/43.6	45.9
LGSC [21]	28.9	35.9	56.1/33.8	34.8
AdaLAM [24]	51.5	62.1	79.1/49.2	60.7
RFM-SCAN [22]	39.4	48.9	45.4/76.4	56.1
CNe [26]	45.1	54.6	47.2/73.4	54.9
OANet [28]	**53.6**	**64.1**	51.2/**87.0**	**61.7**
LOGO	50.9	60.7	**79.6**/38.5	50.3

for testing [50]. These four sequences (*i.e.,* Buckingham palace, Sacre coeur, Reichstag, and Notre dame front facade) contain 4000 image pairs. And the camera poses and sparse models provided by [49] are used to generate ground truth.

At the beginning, the OpenCV SIFT [11] and Nearest Neighbor (NN) matching strategy is adopted to construct putative correspondences with the maximum number of 4000. Considering that AdaLAM uses the descriptor ratio information, other handcrafted mismatch removal methods also make use of this information for pre-processing, but deep-learning-based methods do not apply it because they are already consider it during training. To estimate the essential matrix quickly and accurately, we use USAC_DEFAULT [51] integrated in OpenCV RANSAC. Considering that the essential matrix can be decomposed into rotation and translation, we measure the errors of rotation and translation and take the maximum value, setting the exact Area Under the Curve (AUC) with thresholds of 5 and 10 degrees as the evaluation metrics. In addition, we evaluate the direct processing results of each mismatch removal method with the metrics of Precision (P), Recall (R), and F1-score (F).

To visualize the effect of our LOGO on *YFCC100M*, we compare it with three other related and advanced methods, and the qualitative results are shown in Fig. 6. From this figure we can see that LOGO can identify and preserve the inliers more accurately and robustly, compared to the previous related methods (*e.g.,* LPM [20] and LGSC [21]) and advanced deep learning method (*e.g.,* OANet [28]).

The results of quantitative experiments are shown in Table IV. As shown from it, many regular handcrafted mismatch removal methods cannot cope with severe scenarios such as large viewpoint transformations and massive outlier distribution, which include MAGSAC++, GMS, and LGSC. Our LOGO and AdaLAM show competitive estimation results. Considering that the model parameters of deep learning methods are trained on such a dataset, the performance of LOGO is general and already satisfactory.

D. Fundamental Matrix Estimation

Fundamental matrix estimation is also a basic and critical component of computer vision. We explore the performance

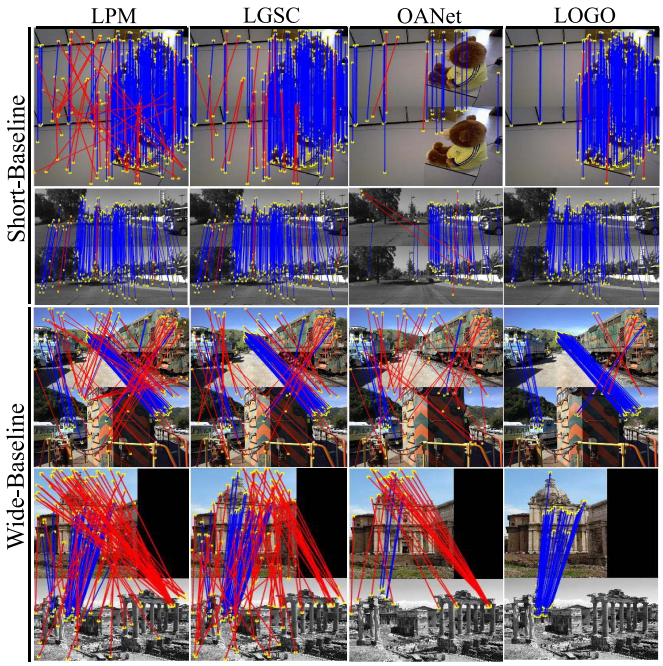

Fig. 7. Qualitative comparisons on short-baseline and wide-baseline image datasets. From left to right: LPM, LGSC, OANet, and LOGO. From top to bottom: matching results on *TUM*, *KITTI*, *T&T*, and *CPC*, respectively. The first two datasets are short-baseline and the last two are wide-baseline. In each pair of images, the yellow dots represent feature points, the blue line are correctly identified matches, and the red lines are incorrect matches.

of our proposed method on a feature match benchmark[4] [52], which includes four publicly available real-world datasets, *i.e., TUM* (1000 pairs), *KITTI* (1000 pairs), Tanks and Temples (*T&T*, 1000 pairs), and Community Photo Collection (*CPC*, 1000 pairs) for fundamental matrix. These data sets are described in detail below: i) The *TUM* SLAM dataset [53] is of indoor scenes and contains short-baseline image pairs with the resolution of 480 × 640. ii) The *KITTI* odometry dataset is in a driving scenario and the geometry between images is dominated by forwarding motion. It contains short-baseline image pairs with the resolution of 370 × 1226. iii) The *T&T* dataset provides many scans of scenes or objects for image reconstruction and contains wide-baseline image pairs with the resolution of 1080 × 2048 and 1080 × 1920. iv) The *CPC* dataset provides unstructured images of well-known landmarks around the world collected from Flicker, which are taken from arbitrary cameras at different times. It contains wide-baseline image pairs with varying resolutions. In this benchmark, the putative correspondences are established with SIFT [11] and NNDR matching strategy with the threshold as 1.25, where the maximum number of feature points is set to 4000. All datasets provide the fundamental matrices as the ground truth, which calibrates correct matches in the putative correspondence set.

All feature matching methods require to be followed by a model estimator to calculate the fundamental matrix based on the matching results, hence for convenience we adopt the classical RANSAC. The evaluation results are derived from the comparison of the estimated model with the ground-truth

[4] https://github.com/JiawangBian/FM-Bench

TABLE V

THE QUANTITATIVE COMPARISON OF FUNDAMENTAL MATRIX ESTIMATION ON SHORT-BASELINE IMAGE DATASETS (*i.e., TUM* AND *KITTI*). **BOLD** INDICATES THE BEST RESULTS

Method	TUM			KITTI		
	Acc.	P/R	F	Acc.	P/R	F
RANSAC [14]	66.0	94.7/95.2	94.5	90.8	98.2/92.6	95.1
MAGSAC++ [16]	**71.0**	78.6/94.6	90.5	89.3	95.6/**99.1**	97.0
VFC [17]	66.6	97.6/92.7	94.9	90.6	**99.0**/91.0	94.7
GMS [42]	64.0	93.4/91.4	92.1	90.5	96.7/90.9	93.6
LPM [20]	68.2	96.2/92.5	93.9	90.4	98.0/94.7	96.2
LGSC [21]	65.8	94.5/98.0	96.0	90.8	98.1/96.1	97.0
AdaLAM [24]	65.4	78.4/94.4	90.5	86.2	95.4/98.9	96.8
RFM-SCAN [22]	65.6	94.6/97.7	95.9	91.0	98.4/96.0	96.9
CNe [26]	63.1	93.2/78.2	84.1	87.6	95.1/90.5	92.7
OANet [28]	69.2	95.9/**98.8**	**97.0**	91.1	97.7/97.5	97.2
LOGO	66.8	**99.0**/89.0	93.4	**91.3**	98.6/96.6	**97.5**

TABLE VI

THE QUANTITATIVE COMPARISON OF FUNDAMENTAL MATRIX ESTIMATION ON WIDE-BASELINE IMAGE DATASETS (*i.e., T&T* AND *CPC*). **BOLD** INDICATES THE BEST RESULTS

Method	T&T			CPC		
	Acc.	P/R	F	Acc.	P/R	F
RANSAC [14]	83.7	72.5/78.8	75.5	52.7	89.5/87.7	88.1
MAGSAC++ [16]	**91.7**	73.3/91.4	81.4	63.4	84.4/93.5	88.7
VFC [17]	79.7	82.0/74.6	76.0	51.4	89.8/82.5	84.3
GMS [42]	88.7	77.1/83.0	78.9	55.8	81.3/82.2	80.7
LPM [20]	90.2	76.8/**93.0**	83.2	61.8	81.9/**96.2**	87.9
LGSC [21]	88.7	81.7/86.6	83.2	62.5	87.8/90.4	**89.0**
AdaLAM [24]	90.4	74.3/90.8	79.4	**68.1**	82.7/94.8	88.4
RFM-SCAN [22]	88.6	79.5/91.5	**83.9**	56.7	87.5/91.0	88.3
CNe [26]	85.3	72.9/79.3	74.8	54.7	82.3/85.9	83.9
OANet [28]	90.2	84.8/85.1	**83.9**	67.7	89.3/89.1	88.5
LOGO	89.7	**94.0**/75.1	82.4	62.4	**95.7**/80.4	86.5

model, and the Accuracy (Acc.) is the ratio of accurate estimates to all estimates. In detail, following the benchmark [52], an estimate is identified as accurate by the corresponding metric being less than 0.05, where *normalized symmetric geometry distance* (NSGD) is used as the metric and computed by dividing the SGD (in pixels) by the length of image diagonals. In particular, due to the different imaging situations (*e.g.,* short and wide baselines), RANSAC has different inlier-outlier thresholds to different datasets for best performance: 0.5 pixels for short-baseline image datasets (*TUM* and *KITTI*), and 2 pixels for wide-baseline image datasets (*T&T* and *CPC*).

Qualitative comparisons regarding representative image pairs are shown in Fig. 7. FM-Bench includes both short-baseline and wide-baseline images, which are composed of indoor and outdoor scenes. Comparing with previous related methods (LPM and LGSC) and advanced deep learning method (OANet), our LOGO can identify the inliers relatively accurately and preserve them adequately.

Table V shows the result statistics of the fundamental matrix estimation experiments on the short-baseline datasets of *TUM* and *KITTI*, while Table VI shows the results on the wide-baseline datasets of *T&T* and *CPC*. Our LOGO demonstrates versatility for imaging scenarios, accuracy in fundamental matrix estimation, and robustness in mismatch removal, which is competitive compared to other state-of-the-art methods.

TABLE VII

THE QUANTITATIVE RESULTS OF IMAGE REGISTRATION. THE AVERAGE VALUES AND STANDARD DEVIATIONS OF RMSE, MAE, AND MEE ARE USED FOR EVALUATION. **BOLD** INDICATES THE BEST

Method	RMSE	MAE	MEE
RANSAC [14]	18.99 (±36.77)	44.53 (±81.73)	23.22 (±46.80)
MAGSAC++ [16]	13.12 (±33.32)	31.71 (±77.22)	15.19 (±38.11)
VFC [17]	43.10 (±67.33)	94.31 (±142.5)	58.31 (±98.17)
GMS [42]	106.7 (±138.1)	216.7 (±270.7)	142.8 (±190.9)
LPM [20]	12.36 (±22.77)	24.13 (±46.79)	12.93 (±33.45)
LGSC [21]	8.924 (±16.77)	20.61 (±37.39)	11.87 (±23.70)
AdaLAM [24]	10.13 (±23.91)	22.39 (±43.17)	11.69 (±24.38)
CNe [26]	107.3 (±133.2)	228.8 (±271.4)	141.2 (±181.2)
OANet [28]	35.93 (±66.32)	80.03 (±139.8)	46.38 (±88.67)
LOGO	**8.113 (±14.58)**	**18.13 (±33.34)**	**9.390 (±19.15)**

E. Image Registration

The image registration task focuses on aligning the common areas of the sensed and reference images to the maximum extent possible. With this purpose, the proposed LOGO is used firstly to construct accurate feature correspondences between two images, and then thin plate spline (TPS) [54] is chosen to estimate the transform function $\mathcal{F}$ based on its mapping generality and smoothness, with no free parameters to be adjusted manually. Ultimately each pixel in the sensed image is mapped to its transformed coordinate by the estimated transform function $\mathcal{F}$, while the bilateral interpolation method calculates the intensity of that coordinate.

To visualize the image registration results, LOGO is applied to four representative image pairs, as shown in Fig. 8. From the smoothness of textures and the consistency of contents especially for difficult edges, LOGO assists in accurately matching the common regions of two related images, thus satisfactorily achieving the image registration tasks.

For quantitative evaluation of the image registration performance, a total of 181 pairs of images from *RS* and *720YUN* datasets are selected as targets to be aligned, and seven state-of-the-art methods mentioned before are used as comparative methods. The average number of feature matches and the average inlier ratios are 958.53 and 27.95%, respectively. In particular, 20 pairs of landmark pixel values $\{r_i, s_i\}_{i=1}^{L}$ are randomly chosen from each pair of images, which can be employed to calculate quantitative metrics, *i.e.*, root mean square error (RMSE), maximum error (MAE), and median error (MEE). The definitions of these metrics are as follows:

$$\mathrm{RMSE} = \sqrt{1/L\textstyle\sum_{i=1}^{L}\left(r_i - \mathcal{F}\left(s_i\right)\right)^2}, \tag{20}$$

$$\mathrm{MAE} = \max\left\{\sqrt{\left(r_i - \mathcal{F}\left(s_i\right)\right)^2}\right\}_{i=1}^{L}, \tag{21}$$

$$\mathrm{MEE} = \mathrm{median}\left\{\sqrt{\left(r_i - \mathcal{F}\left(s_i\right)\right)^2}\right\}_{i=1}^{L}. \tag{22}$$

As shown in Table VII, the proposed LOGO is capable of accurately constructing feature correspondences between two

Fig. 8. Image registration results of LOGO applied on 4 typical image pairs. The first row represents the original input images, where the left and right in each group are the sensed images and reference images. The second row presents the registration results of LOGO, where the left and right in each group are warped sensed images and the check-board results, respectively.

TABLE VIII

THE QUANTITATIVE RESULTS OF LOOP CLOSURE DETECTION. THE MAXIMUM RECALL (%) AT 100% PRECISION OF 8 FEATURE MATCHING METHODS ON *K00* AND *K02* SEQUENCES. BOLD INDICATES THE BEST

	RANSAC [14]	MAGSAC++ [16]	VFC [17]	GMS [42]	LPM [20]	LGSC [21]	AdaLAM [24]	CNe [26]	OANet [28]	LOGO
K00	91.12	91.05	89.17	82.52	89.52	90.67	90.78	84.32	87.13	**91.53**
K02	76.32	77.57	73.27	71.18	71.18	76.38	74.59	71.96	75.17	**78.77**

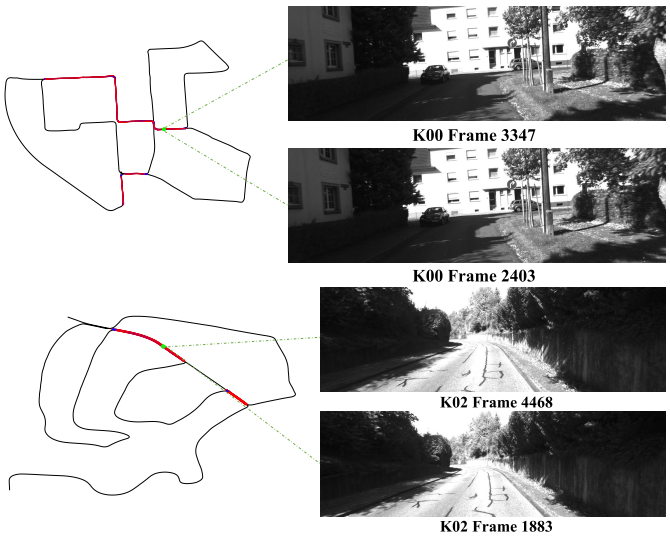

Fig. 9. LCD illustrations of LOGO applied on *K00* and *K02* datasets. From left to right: robot's trajectory and true positive example. On the left of each row, the black trajectory comes from GPS logs, the red hollow points denote loop closure pairs and are connected by a blue line, and the green solid points are the specific illustrations of true-positive detection.

related images with low error values, and performs relatively well for the image registration task.

F. Loop Closure Detection

To further exploit the practical utility of LOGO, we apply it to the loop closure detection (LCD) task, *i.e.*, robot surroundings recognition and verification during the navigation, where two sequences *K00* and *K02* from the *KITTI* dataset [55] are used for evaluation. DELG [56] is chosen to extract both local and global image features to perform LCD due to its superiority in LCD-related tasks like image retrieval. In detail, DELG uses the similarity of global features under the l_2 norm constraint to select the candidate frame of a query image, and then the model fitting is conducted between them according to local features.

1) Qualitative Result: Some qualitative results of LCD using our LOGO method are displayed in Fig. 9. Based on the data recorded by odometry, the trajectories of robots on each dataset are drawn by the black lines. The loop closure pairs are labeled as red hollow points while connecting them with blue lines, and the green solid points are the specific illustrations of true-positive detection.

2) Quantitative Comparison: For LCD, only if enough matches are retained would the loop closure event be triggered. Therefore, the maximum recall at 100% precision is usually treated as the assessment metric, where true samples are supplied by ground truth (GT) [57]. Regarding precision and recall, an identification is regarded as a true inlier when it is positioned within 10 adjacent frames of true loop-closing samples from GT.

The quantitative results of LCD are exhibited in Table VIII, where *K02* is considerably more demanding due to complex scenes resulting from violent dynamics. In LCD tasks, only feature matching methods with both high precision and high recall can ensure sufficient inliers and few outliers, and hence effectively trigger loop-closing events. From the table, RANSAC-related methods (RANSAC and MAGSAC++) perform well owing to appropriate global transformation models estimated in adequate iterations. Noteworthily, our LOGO is able to construct the largest number of feature correspondences with guaranteed purity even in challenging *K02*, which is of great assistance to the robot navigation.

IV. CONCLUSION AND DISCUSSION

This paper proposed an innovative mismatch removal method targeting robust feature matching, *i.e.*, *LOcality-guided Global-preserving Optimization* (LOGO). The proposed

locality-guided strategy can make appropriate use of local and global structures based on the fixed-point progressive optimization with high robustness. In addition, the node and edge similarity scores were calculated based on local affine transformations, facilitating the graph structure to express topology. In feature matching experiments, our LOGO demonstrated its superiority in comparison with other state-of-the-art methods with both high precision and high recall. As our LOGO performs well in both essential and fundamental matrix estimation, it can be of great assistance during the estimation of camera parameters in 3D reconstruction tasks, facilitating the subsequent sparse point clouds recovery. Furthermore, in image registration and loop closure detection, LOGO also had satisfactory performance, showcasing great significance for high-level vision-based tasks. In particular, for nonrigid image pairs, such as those taken by fish-eye cameras, our LOGO can perform the image registration tasks well.

Although the affinity metric containing global topological content allows our mismatch removal method to identify tricky inliers, it is constructed with a relatively high space burden. Even though the fixed-point progressive optimization method is a linear iterative process, LOGO would consume a lot of time when dealing with a set containing numerous putative matches. In the future, we plan to investigate the improvements of the ease of LOGO, with the aim of achieving robust feature matching with a low space complexity and a high speed.

REFERENCES

[1] J. Ma, X. Jiang, A. Fan, J. Jiang, and J. Yan, "Image matching from handcrafted to deep features: A survey," *Int. J. Comput. Vis.*, vol. 129, pp. 23–79, Aug. 2021.

[2] Y. Jin *et al.*, "Image matching across wide baselines: From paper to practice," *Int. J. Comput. Vis.*, vol. 129, pp. 517–547, Oct. 2021.

[3] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1362–1376, Aug. 2008.

[4] C. Min, Y. Gu, Y. Li, and F. Yang, "Non-rigid infrared and visible image registration by enhanced affine transformation," *Pattern Recognit.*, vol. 106, Oct. 2020, Art. no. 107377.

[5] C. Wang, L. Wang, and L. Liu, "Progressive mode-seeking on graphs for sparse feature matching," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 788–802.

[6] M. Leordeanu and M. Hebert, "A spectral technique for correspondence problems using pairwise constraints," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, 2005, pp. 1482–1489.

[7] M. Leordeanu, M. Hebert, and R. Sukthankar, "An integer projected fixed point method for graph matching and map inference," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1114–1122.

[8] M. Cho, J. Lee, and K. M. Lee, "Reweighted random walks for graph matching," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 492–505.

[9] T. Wang, H. Ling, C. Lang, and S. Feng, "Graph matching with adaptive and branching path following," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2853–2867, Dec. 2018.

[10] S. Lee, J. Lim, and I. H. Suh, "Progressive feature matching: Incremental graph construction and optimization," *IEEE Trans. Image Process.*, vol. 29, pp. 6992–7005, 2020.

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

[12] 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.

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

[14] M. A. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

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

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

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

[18] J. Ma, H. Zhou, J. Zhao, Y. Gao, J. Jiang, and J. Tian, "Robust feature matching for remote sensing image registration via locally linear transforming," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6469–6481, Dec. 2015.

[19] W.-Y. Lin *et al.*, "CODE: Coherence based decision boundaries for feature correspondence," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 34–47, Jan. 2018.

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

[21] X. Jiang, Y. Xia, X.-P. Zhang, and J. Ma, "Robust image matching via local graph structure consensus," *Pattern Recognit.*, vol. 126, Jun. 2022, Art. no. 108588.

[22] 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, 2019.

[23] F. Shao, Z. Liu, and J. An, "Feature matching based on minimum relative motion entropy for image registration," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.

[24] L. Cavalli, V. Larsson, M. R. Oswald, T. Sattler, and M. Pollefeys, "Handcrafted outlier detection revisited," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 770–787.

[25] H. Liu and S. Yan, "Common visual pattern discovery via spatially coherent correspondences," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1609–1616.

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

[27] C. Zhao, Z. Cao, C. Li, X. Li, and J. Yang, "NM-Net: Mining reliable neighbors for robust feature correspondences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 215–224.

[28] J. Zhang *et al.*, "Learning two-view correspondences and geometry using order-aware network," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 5845–5854.

[29] 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.

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

[31] 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.

[32] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Commun. ACM*, vol. 18, no. 9, pp. 509–517, Sep. 1975.

[33] A. Vedaldi and B. Fulkerson, "Vlfeat: An open and portable library of computer vision algorithms," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 1469–1472.

[34] Z. Zhang, "Determining the epipolar geometry and its uncertainty: A review," *Int. J. Comput. Vis.*, vol. 27, no. 2, pp. 161–195, 1998.

[35] H. Le, T.-J. Chin, and D. Suter, "An exact penalty method for locally convergent maximum consensus," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1888–1896.

[36] H. Le, T.-J. Chin, A. Eriksson, T.-T. Do, and D. Suter, "Deterministic approximate methods for maximum consensus robust fitting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 842–857, Mar. 2021.

[37] Z. Cai, T.-J. Chin, H. Le, and D. Suter, "Deterministic consensus maximization with biconvex programming," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 685–700.

[38] E. Tola, V. Lepetit, and P. Fua, "DAISY: An efficient dense descriptor applied to wide-baseline stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 815–830, May 2009.

[39] H. S. Wong, T.-J. Chin, J. Yu, and D. Suter, "Dynamic and hierarchical multi-structure geometric model fitting," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1044–1051.

[40] L. Liang *et al.*, "Image registration using two-layer cascade reciprocal pipeline and context-aware dissimilarity measure," *Neurocomputing*, vol. 371, pp. 1–14, Jan. 2020.
[41] M. Horst and R. Möller, "Visual place recognition for autonomous mobile robots," *Robotics*, vol. 6, no. 2, p. 9, 2017.
[42] J.-W. Bian *et al.*, "GMS: Grid-based motion statistics for fast, ultra-robust feature correspondence," *Int. J. Comput. Vis.*, vol. 128, no. 6, pp. 1580–1594, 2020.
[43] P. F. Alcantarilla, A. Bartoli, and A. J. Davison, "Kaze features," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 214–227.
[44] A. Mishchuk, D. Mishkin, F. Radenovic, and J. Matas, "Working hard to know your neighbor's margins: Local descriptor learning loss," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–12.
[45] Y. Tian, X. Yu, B. Fan, F. Wu, H. Heijnen, and V. Balntas, "SOSNet: Second order similarity regularization for local descriptor learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11016–11025.
[46] K. Mikolajczyk *et al.*, "A comparison of affine region detectors," *Int. J. Comput. Vis.*, vol. 65, no. 1, pp. 43–72, 2005.
[47] K. Cordes, B. Rosenhahn, and J. Ostermann, "High-resolution feature evaluation benchmark," in *Proc. Int. Conf. Comput. Anal. Images Patterns*, 2013, pp. 327–334.
[48] B. Thomee *et al.*, "YFCC100M: The new data in multimedia research," *Commun. ACM*, vol. 59, no. 2, pp. 64–73, 2016.
[49] J. Heinly, J. L. Schonberger, E. Dunn, and J.-M. Frahm, "Reconstructing the world* in six days*(as captured by the Yahoo 100 million image dataset)," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3287–3295.
[50] W. Sun, W. Jiang, E. Trulls, A. Tagliasacchi, and K. M. Yi, "ACNe: Attentive context normalization for robust permutation-equivariant learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 11286–11295.
[51] R. Raguram, O. Chum, M. Pollefeys, J. Matas, and J.-M. Frahm, "USAC: A universal framework for random sample consensus," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 2022–2038, Aug. 2012.
[52] J.-W. Bian *et al.*, "An evaluation of feature matchers for fundamental matrix estimation," in *Proc. Brit. Mach. Vis. Conf.*, 2019, pp. 1–14.
[53] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2012, pp. 573–580.
[54] G. Donato and S. Belongie, "Approximate thin plate spline mappings," in *Proc. Eur. Conf. Comput. Vis.*, 2002, pp. 21–31.
[55] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.
[56] B. Cao, A. Araujo, and J. Sim, "Unifying deep local and global features for image search," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 726–743.
[57] K. Zhang, X. Jiang, and J. Ma, "Appearance-based loop closure detection via locality-driven accurate motion field learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 2350–2365, Mar. 2022.

Yifan Xia received the B.E. degree in information and communication engineering from Wuhan University, Wuhan, China, in 2021, where he is currently pursuing the Ph.D. degree with the Electronic Information School. His current research interests include computer vision and image processing.

Jiayi Ma (Senior Member, IEEE) received the B.S. degree in information and computing science and the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology, 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 200 refereed journals and conference papers, including the 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 pattern recognition. He 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 Editorial Board Member of *Neurocomputing*, *Sensors*, and *Entropy*, and a Guest Editor of *Remote Sensing*.