ISPRS Journal of Photogrammetry and Remote Sensing 219 (2025) 38–51

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

MGCNet: Multi-granularity consensus network for remote sensing image correspondence pruning

Fengyuan Zhuang [a,b,1], Yizhang Liu [c,1], Xiaojie Li [a,b], Ji Zhou [d,e], Riqing Chen [a,b], Lifang Wei [a,b], Changcai Yang [a,b,f,*], Jiayi Ma [g]

[a] College of Computer and Information Sciences, Fujian Agriculture and Forestry University, Fuzhou 350002, China
[b] Center for Agroforestry Mega Data Science, School of Future Technology, Fujian Agriculture and Forestry University, Fuzhou 350002, China
[c] College of Computer and Data Science, Fuzhou University 350108, China
[d] Cambridge Crop Research, National Institute of Agricultural Botany (NIAB), Cambridge CB3 0LE, UK
[e] State Key Laboratory of Crop Genetics & Germplasm Enhancement, academy for Advanced Interdisciplinary Studies, Nanjing Agricultural University, Nanjing 210095, China
[f] Key Laboratory of Smart Agriculture and Forestry (Fujian Agriculture and Forestry University), Fujian Province University, Fuzhou 350002, China
[g] Electronic Information School, Wuhan University, Wuhan 430072, China

ARTICLE INFO

Keywords:
Correspondence pruning
Image matching
Remote sensing image
Group consensus
Multi-granularity consensus
Image registration

ABSTRACT

Correspondence pruning aims to remove false correspondences (outliers) from an initial putative correspondence set. This process holds significant importance and serves as a fundamental step in various applications within the fields of remote sensing and photogrammetry. The presence of noise, illumination changes, and small overlaps in remote sensing images frequently result in a substantial number of outliers within the initial set, thereby rendering the correspondence pruning notably challenging. Although the spatial consensus of correspondences has been widely used to determine the correctness of each correspondence, achieving uniform consensus can be challenging due to the uneven distribution of correspondences. Existing works have mainly focused on either local or global consensus, with a very small perspective or large perspective, respectively. They often ignore the moderate perspective between local and global consensus, called group consensus, which serves as a buffering organization from local to global consensus, hence leading to insufficient correspondence consensus aggregation. To address this issue, we propose a multi-granularity consensus network (MGCNet) to achieve consensus across regions of different scales, which leverages local, group, and global consensus to accomplish robust and accurate correspondence pruning. Specifically, we introduce a GroupGCN module that randomly divides the initial correspondences into several groups and then focuses on group consensus and acts as a buffer organization from local to global consensus. Additionally, we propose a Multi-level Local Feature Aggregation Module that adapts to the size of the local neighborhood to capture local consensus and a Multi-order Global Feature Module to enhance the richness of the global consensus. Experimental results demonstrate that MGCNet outperforms state-of-the-art methods on various tasks, highlighting the superiority and great generalization of our method. In particular, we achieve 3.95% and 8.5% mAP5° improvement without RANSAC on the YFCC100M dataset in known and unknown scenes for pose estimation, compared to the second-best models (MSA-LFC and CLNet). Source code: https://github.com/1211193023/MGCNet.

1. Introduction

Accurate feature matching is a crucial prerequisite for many computer vision tasks, *e.g.*, 3D reconstruction (Sun et al., 2022), image fusion (Ma et al., 2019b; Zhang et al., 2020; Ma et al., 2022b; Xu et al., 2020), remote sensing image registration (Liu et al., 2021a), point set registration (Wang et al., 2023a), simultaneous localization and mapping (SLAM) (Huang et al., 2020), etc. However, despite the significant progress in feature descriptor-based matching techniques in computer vision, the complexity of image types and degradations can often result in ambiguous feature descriptors. Consequently, the nearest neighbor matching strategy can lead to a considerable number of false correspondences. This situation can be particularly challenging due to the complex properties of remote sensing images. For instance, remote sensing images often exhibit severe noise and local distortions, as well as non-rigid transformations, which can further

* Correspondence to: No.15 Shangxiadian Road, Cangshan District, Fuzhou City, Fujian Province, China
E-mail address: changcaiyang@gmail.com (C. Yang).
[1] Fengyuan Zhuang and Yizhang Liu contributed equally to this work.

https://doi.org/10.1016/j.isprsjprs.2024.11.011
Received 23 April 2024; Received in revised form 15 November 2024; Accepted 19 November 2024
Available online 28 November 2024

Fig. 1. Illustrations of (a) local consensus and (b) global consensus. The lines in the figures denote inliers.

deteriorate the matching results. To alleviate this problem, correspondence pruning techniques have been extensively investigated as an essential component of image feature matching.

Fortunately, in image pairs, inliers often exhibit spatial consensus, whereas outliers tend to be randomly distributed. This property is a powerful tool for distinguishing between inliers and outliers. Traditional methods construct k-nearest neighbors for each correspondence and adopt a geometric or algebraic constraint to calculate the consensus score, which can effectively remove outliers and perform well in certain scenarios. For instance, Locality preserving matching (LPM) (Ma et al., 2019c) infers inliers by computing the common neighbor information, which is a model of local consensus. Learning for mismatch removal (LMR) (Ma et al., 2019a) builds local neighboring structures according to multiple K-nearest neighbors to capture multi-scale local consensuses. An advanced consensus of neighborhood topology (Liu et al., 2021b) is proposed to identify inliers and the neighborhood construction is constructed through a subset with a high percentage of inliers with a guided matching strategy from the putative matches. Locality-guided global-preserving optimization (LOGO) (Xia and Ma, 2022) enhances the robustness of correspondence pruning by employing a local topology-guided global preservation optimization strategy. However, it is vital to note that to achieve good performance, the putative correspondence set should not contain a large number of outliers. Otherwise, the performance of traditional methods may significantly degrade as the constructed local neighborhoods become unreliable and are susceptible to interference from outliers.

In recent years, learning-based methods have received increasing attention in exploring local and global consensus. Such methods possess powerful and flexible representation ability, allowing them to learn more complex patterns and offering great advantages over handcrafted methods. For example, Consensus learning framework (CLNet) (Zhao et al., 2021) proposes an annular convolution to aggregate local neighborhood information and uses a graph convolutional network (GCN) block to get a global feature embedding. Laplacian motion coherence network (LMCNet) (Liu et al., 2021c) obtains local motion consensus by fusing k-nearest neighbor feature maps through max pooling and exploits a smooth function to fit global motion consensus. MS^{2}DGNet (Dai et al., 2022) uses a transformer-like structure to aggregate local k nearest neighbor feature maps to gain local neighborhood features.

Previous works have achieved promising results by focusing on either local or global consensus, with a very small perspective or large perspective, respectively. However, they ignore the group consensus which is a moderate perspective between local and global consensus, serving as a buffering organization from local to global consensus, hence leading to a lack of medium-granularity consensus. To better illustrate this issue, we show an example in Fig. 1. Specifically, local consensus indicates that within a small area, inliers tend to exhibit similar motion behaviors at the local level (e.g., similar length and angles). In contrast, global consensus shows that inliers conform to a specific geometric transformation at the global level. It is worth noting that in Fig. 2(b), those inliers that satisfy geometric consensus at the global level but are spatially distant fail to meet local consensus, as they exhibit distinct lengths and angles. To address this limitation, we introduce a GroupGCN module that randomly divides the initial correspondences into m groups and then explores consensus within each group. Due to the randomness of grouping, correspondences within each group cover a broader range, which helps to mitigate the large discrepancy between local and global consensus, thus creating a buffering effect.

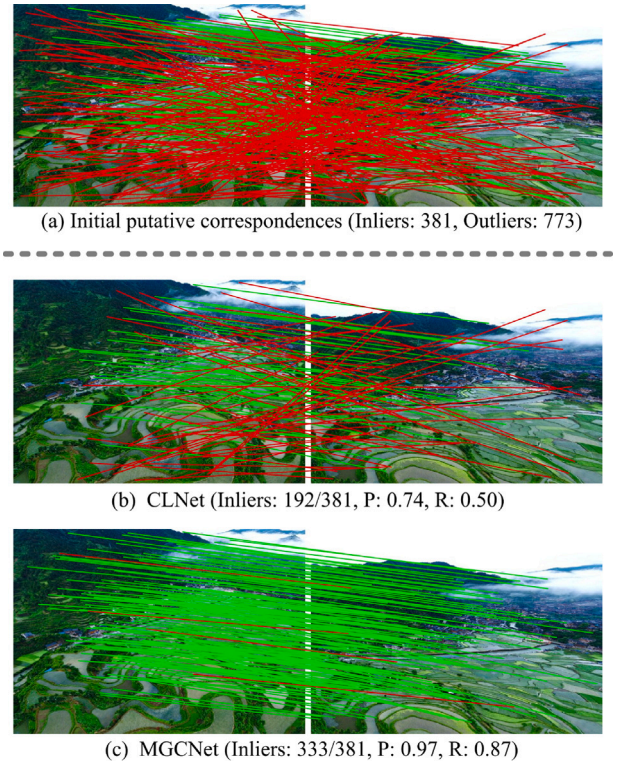

Fig. 2. Visual representation of (a) initial correspondences, (b) the baseline CLNet considering local consensus and global consensus and (c) our MGCNet introducing multi-granularity consensus including multi-level local consensus, group consensus and multi-order global consensus. The red/green lines indicate false/true correspondences, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Besides, our method takes a more holistic view and is designed to enhance consensus across multiple granularities. Considering that initial putative correspondences tend to be unevenly distributed across the image, with numerous key points in textured regions and very few in textureless regions, this uneven distribution of correspondences makes it challenging to achieve a uniform consensus. Most existing methods directly adopt a fixed local feature extraction module (Dai et al., 2022; Zhao et al., 2021; Liu et al., 2021c), such as a fixed scale and feature aggregation method. For example, CLNet sets a fixed number of k-nearest neighbors, which is not conducive to solving the above problem. To overcome this limitation, we propose a **multi-level local feature aggregation module (MLL Module)** that adapts to the size of the local neighborhood to capture local consensus among correspondences. In addition, we introduce a **multi-order global feature aggregation module (MOG Module)** to further improve the performance of our method by enhancing the richness of global consensus.

To sum up, building on the state-of-the-art method (Zhao et al., 2021), we propose a **multi-granularity consensus network (MGCNet)** to achieve consensus across regions of different scales and demonstrate its effectiveness on various tasks through extensive experiments. Compared with the baseline model, as shown in Fig. 2, our MGCNet can effectively solve the aforementioned challenges, resulting in the superior performance of the correspondence pruning.

In summary, our contributions can be summarized as follows:

(1) We introduce a GroupGCN module focusing on a novel consensus, group consensus, which acts as a buffer organization from local to global consensus, to enhance consensus learning.

(2) We propose a multi-level local feature aggregation module to adapt to the size of the local neighborhood to capture local consensus and a multi-order global feature aggregation module to enhance the richness of global consensus by retaining multi-order global features.

(3) We design a new network, the Multi-Granularity Consensus Network, which effectively mines consensus among correspondences from a local-to-group and group-to-global perspective for correspondence pruning and achieves state-of-the-art performance compared to other methods on various challenging tasks.

The rest of the paper is organized as follows. Section 2 describes the related work. Section 3 introduces a general learning pipeline for the image correspondence pruning problem and the details of our multi-granularity consensus network and its implementation details. Section 4 illustrates the experimental results on the tasks of correspondence pruning, camera pose estimation, and remote sensing image registration to demonstrate the superiority of our method. The concluding remarks are presented in Section 5.

2. Related work

2.1. Traditional correspondence pruning methods

Correspondence pruning methods have been widely applied in many fields such as computer vision, pattern recognition, image analysis, and particularly in the field of remote sensing. Traditional correspondence pruning methods can be classified into three categories: resampling-based, non-parametric model-based, and relaxed methods (Ma et al., 2021) . Resampling-based methods, exemplified by random sample consensus (RANSAC) (Fischler and Bolles, 1981), introduce a hypothesize-and-verify framework to identify inliers from the initial correspondence set. RANSAC has been widely used for automatic correspondence pruning of remote sensing images (Ma et al., 2022a). This method, relying on a predefined parametric model, becomes less efficient when the underlying image transformation is nonrigid, especially in the context of remote sensing images. Several non-parametric model-based methods have been proposed to address these issues, for instance, Pilet et al. (2008) model the deformation by the triangulated 2-D mesh, which can reduce the negative impact of outliers. Vector field consensus (VFC) (Ma et al., 2014) constrains the deformation function in the reproducing kernel Hilbert space with Tikhonov regularization, and estimates in a Bayesian model. Relaxed methods, such as Bian et al. (2017) and Ma et al. (2019c, 2015), employ less strict geometric constraints to handle complex scenes and the complex properties of remote sensing images. Grid-based motion statistics (GMS) (Bian et al., 2017) segments the image pair into $n \times n$ meshes, assigns correspondences to specific meshes, and selects inliers based on a predefined threshold. Locality preserving matching (LPM) (Ma et al., 2019c) determines the correctness of each correspondence based on the number of neighboring correspondences. NMRC (Ma et al., 2022a) uses the manifold learning to preserve the consensus of the local neighborhood structures of the potential inliers. Progressive motion coherence (PMC) (Liu et al., 2022) proposes two novel coherence constraints, namely efficient neighborhood element coherence and relative order-aware motion coherence, to design a mathematical model for correspondence pruning. While these handcrafted methods are effective for specific scenes, they may not be suitable for general datasets with extremely low inlier ratios.

2.2. Learning-based correspondence pruning methods

Deep learning-based networks are commonly used to deal with correspondence pruning problems while maintaining permutation-equivariance, allowing them to leverage the rich potential relationships within data. Differentiable RANSAC (DSAC) (Brachmann et al., 2017) alleviates the non-differentiable defect of RANSAC (Fischler and Bolles, 1981), and applies its idea to a deep learning network for correspondence pruning. Motivated by Point-Net (Qi et al., 2017a), Learning to find good correspondences (LFGC) (Yi et al., 2018) introduces PointNet-like architectures, named ResNet Block, to process input data independently and predict the probability of each correspondence as an inlier. While it performs well on public datasets, ignoring the relationship among correspondences can lead to detrimental effects. Hence, to deal with the problem, some networks introduce local information in various ways while maintaining permutation-equivariance. For instance, Order-aware network (OANet) (Zhang et al., 2019) proposes a Diffpool & Diffunpool Block, including a combination of differentiable pooling and unpooling operation and an order-aware operation, to capture local and global contextual information of sparse correspondences. Attentive context networks (ACNe) (Sun et al., 2020) introduces both local and global attention operations that leverage the attentive context normalization (ACN) mechanism, to effectively capture both local and global contextual information. Compared to traditional methods, deep learning-based approaches have demonstrated better performance due to their superior feature representation ability, especially when there is a large number of outliers in the putative correspondence set. However, counting solely on designing diverse modules to incorporate both local and global contextual information into feature learning has a limited impact on performance improvement. Therefore, it has become a research trend to introduce correspondence consensus as an integral part of network learning.

2.3. Correspondence consensus

Inliers are known to be consistent with epipolar geometry or under the homography constraint (Hartley and Zisserman, 2003), whereas outliers exhibit inconsensus due to their random distribution. The correspondence consensus has been extensively investigated in the correspondence pruning for several decades (Liu et al., 2021c; Ma et al., 2021). For example, various methods such as bilateral functions (BF) (Lin et al., 2014) and coherence-based decision boundaries (CODE) (Lin et al., 2017) propose global motion consensus to distinguish correspondences. Other studies (Bian et al., 2017; Ma et al., 2019c) constraint the model by leveraging geometric consensus. Although the above traditional methods have made progress, their performance is still unsatisfactory and requires careful parameter tuning, particularly when dealing with challenging datasets with a low percentage of inliers. In recent years, the consensus in learning-based methods has drawn increasing attention. For instance, Mining reliable neighbors network (NM-Net) (Zhao et al., 2019) proposes a compatibility-specific neighbor mining method by using local affine information of feature descriptors as prior information to find more reasonable neighbors. Learning for mismatch removal (LMR) (Ma et al., 2019a) builds local neighborhood structures for each correspondence according to multiple k-nearest neighbors to capture multi-scale local consensuses. CLNet (Zhao et al., 2021) progressively removes outliers through a local-to-global consensus learning strategy. LMCNet (Liu et al., 2021c) fuses local neighborhood motion consensus and exploits a smooth function to fit global motion consensus to find inliers. Multiple sparse semantics dynamic graph network (MS^{2}DGNet) (Dai et al., 2022) build a sparse dynamic graph to capture local topology among correspondences. MSA with Local Feature Consensus (MSA-LFC) (Wang et al., 2023b) enhances local feature consensus by incorporating mutual neighborhood consensus among neighboring features and aggregating them. Sparse semantic learning network (SSLNet) (Chen et al.,

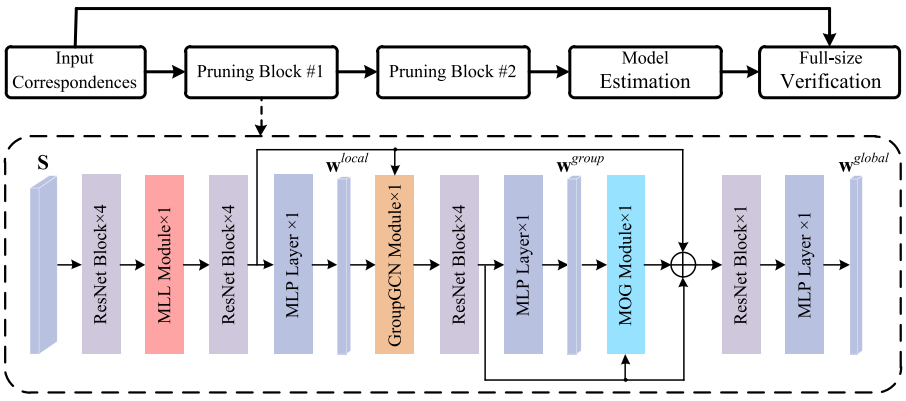

Fig. 3. The architecture of the Multi-granularity Consensus Network for correspondence pruning. It includes three novel modules to learn three-granularity consensuses: (1) The multi-level local feature aggregation module (MLL Module), which adapts to the size of the local neighborhood and captures local consensus. (2) The GroupGCN module focuses on group consensus and acts as a buffer organization from local to global consensus. (3) The multi-order global feature aggregation module (MOG Module) enhances the richness of global consensus. Besides, $\mathbf{S} \in \mathbb{R}^{N\times 4}$ indicates the initial putative correspondence set. $\mathbf{w}^{local}, \mathbf{w}^{group}, \mathbf{w}^{global} \in \mathbb{R}^{N\times 1}$ represent local consensus, group consensus, and global consensus scores, respectively.

2024) enhances region-to-whole graph consensus learning through attention mechanisms to measure the confidence of nodes in the sparse graph. To avoid the negative impact of outliers, Consistency guided ResFormer network (CGR-Net) (Yang et al., 2024) uses consistent correspondences to guide model perspective focusing. Rotation-invariant sequence-aware consensus (RoSe) (Liu et al., 2024) is derived to better characterize the topological structure of inliers and enlarge the distribution between inliers and outliers. Most existing learning-based methods tend to focus on local and global consensus, which represent a very small and large perspective, respectively. However, the considerable disparity between these two perspectives may result in challenges for the stable optimization of the network. Meanwhile, only correspondences that strictly satisfy both local and global consensus are preserved. Nevertheless, local distortions and non-rigid transformations are commonly found in remote sensing images, which means that even inliers cannot guarantee strict adherence to both local and global consensus. To address this issue, we propose a group consensus, which acts as a buffer organization from local to global consensus. We further consider consensus with multiple granularities, i.e., local, group, and global consensus, and propose a novel consensus network that more comprehensively learns consensus among correspondences, enabling better performance on complex tasks.

3. Proposed method

In this section, we first introduce the problem formulation. Next, we describe the proposed MLL Module, GroupGCN Module, MOG Module, and loss function. Following that, we will elaborate on implementation details.

3.1. Problem formulation

Given a pair image $(\mathrm{I}, \mathrm{I}')$, we first extract feature points for I and I' using off-the-shelf feature detection and description methods (e.g., handcrafted scale invariant feature transform (SIFT) (Lowe, 2004) or learned SuperPoint (DeTone et al., 2018)), respectively. The putative correspondence set $\mathbf{S}$ can then be established based on the nearest neighbor matching strategy:

$$\mathbf{S} = \{s_i\} \in \mathbb{R}^{N\times 4}, i = 1, 2, \ldots, N, s_i = \left(x_1^i, y_1^i, x_2^i, y_2^i\right), \tag{1}$$

where s_i is the ith correspondence; N is the number of correspondences; (x_1^i, y_1^i) and (x_2^i, y_2^i) refer to the normalized coordinates of the ith correspondence with respect to the camera internal parameters. Our objective is to identify inliers while rejecting outliers in $\mathbf{S}$.

Similar to previous works (Zhao et al., 2021; Dai et al., 2022), we also cast the correspondence pruning into a binary classification and an essential matrix regression problem. To this end, building on the state-of-the-art iterative pruning framework (Zhao et al., 2021), as shown in Fig. 3, the architecture of the proposed MGCNet consists of three key steps: correspondence pruning (with two pruning blocks), model estimation, and full-size verification. Specifically, the initial putative correspondence set $\mathbf{S}$ as input first passes through the first pruning block, yielding two outputs, including pruned correspondence set $\mathbf{S}_1$ and inlier weight $\mathbf{O}_1$. Next, we combine $\mathbf{S}_1$ and $\mathbf{O}_1$ and feed them into the second pruning block. This block produces a subset with a higher inlier ratio, denotes as $\mathbf{S}_2$, and a more accurate inlier weight, denoted as $\mathbf{O}_2$. The above processes can be expressed as :

$$(\mathbf{S}_1, \mathbf{O}_1) = f_{\boldsymbol{\Phi}}(\mathbf{S}), \tag{2}$$

$$(\mathbf{S}_2, \mathbf{O}_2) = f_{\Psi}([\mathbf{S}_1 \parallel \mathbf{O}_1]), \tag{3}$$

where $f_{\boldsymbol{\Phi}}$ and f_{Ψ} are the first and second pruning blocks with learnable parameters $\boldsymbol{\Phi}$ and Ψ. $[\cdot\|\cdot]$ presents concatenation operation. $\mathbf{S}_1 \in \mathbb{R}^{N_1\times 4}$ and $\mathbf{S}_2 \in \mathbb{R}^{N_2\times 4}$ are two pruned correspondence subsets, where $N_2 < N_1 < N$. $\mathbf{O}_1$ and $\mathbf{O}_2$ represent the final logit values of two pruning blocks, in which $\mathbf{O}_2$ is additionally processed by a ResNet block and an MLP layer. After that, we use the weighted eight-point algorithm to calculate essential matrix $\hat{\mathbf{E}}$. Finally, the estimated $\hat{\mathbf{E}}$ is adopted to accomplish a full-size verification on the initial putative correspondence set $\mathbf{S}$ to recover those inliers that are erroneously pruned. The above processes can be formulated as:

$$\hat{\mathbf{E}} = g(\mathbf{S}_2, \mathbf{O}_2), \tag{4}$$

$$\mathbf{P} = h(\hat{\mathbf{E}}, \mathbf{S}), \tag{5}$$

where $g(\cdot,\cdot)$ is the weighted eight-point algorithm; $h(\cdot,\cdot)$ is to compute the epipolar distances for a full-size verification. $\mathbf{P}$ is the epipolar distance set that indicates the correctness of each correspondence in $\mathbf{S}$.

3.2. Multi-level local feature aggregation module

In remote sensing image pairs, correspondences often have a non-uniform density across various regions, which poses a significant challenge for correspondence pruning. Some existing networks (Zhao et al., 2021; Liu et al., 2021c; Dai et al., 2022) directly adopt a fixed local feature extraction module, such as a fixed scale and feature aggregation method. However, this approach may not be reasonable for solving the aforementioned challenge. To address this issue, we propose a multi-level local feature aggregation module (MLL Module, as shown in Fig. 4), which can adapt to the size of the local neighborhood to

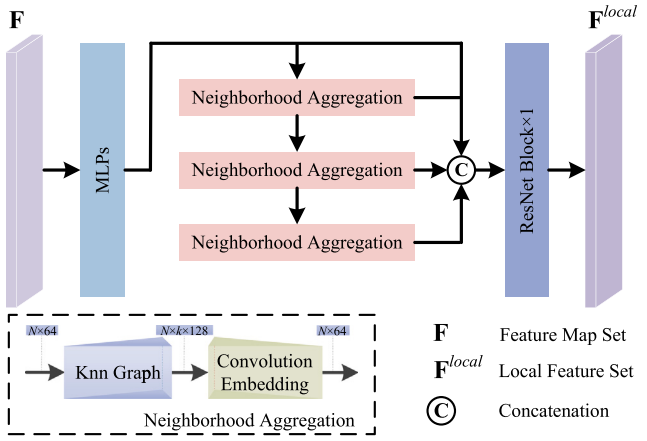

Fig. 4. The multi-level local feature aggregation module (MLL Module). Neighborhood aggregation is implemented by building a k nearest neighbor graph for each correspondence in feature space, followed by a convolution embedding layer, similar to that of CLNet (Zhao et al., 2021).

capture local consensus among correspondences flexibly for different scenes.

Specifically, for each correspondence s_i, it first passes through a series of ResNet Blocks and an MLP layer for feature embedding, where an MLP layer is implemented by a convolutional operation with a kernel size of 1 and a stride of 1. We denote the embedded feature vector with respect to s_i as f_i. After that, we search k nearest neighbors for f_i and obtain a feature map $\{f_{ij}\} \in \mathbb{R}^{N\times k\times C}, j = 1, \dots, k, 1 < k < N$. Based on this, a k nearest neighbor graph $G_i^{local} = (V_i^{local}, E_i^{local})$ can be constructed to capture local context among correspondences, where $V_i^{local} = \{f_{ij} | j = 1, 2, \dots, k\}$ are k-nearest neighbors of f_i and E_i^{local} indicates the set of directed edges from f_i to its k nearest neighbors. Following Dai et al. (2022), the local neighborhood features of each correspondence can be described as:

$$e_{ij} = [f_i \parallel f_i - f_{ij}], j = 1, 2, \dots, k, \tag{6}$$

where $f_i - f_{ij}$ is the residual feature; $[\cdot\|\cdot]$ presents concatenation operation along the channel dimension.

Compared to a simple pooling operation, this approach is more flexible in extracting contextual features. Furthermore, it is evident that the size of the contextual receptive field depends on the value of k. Although increasing k can effectively enlarge the receptive field, significantly increasing k would result in huge computations and become time-consuming, which is not conducive to real-time applications. Inspired by the fact that in multi-layer convolutional neural networks, a large kernel can be divided into several small kernel cascades for reducing the parameters while increasing the nonlinear transformation ability of the networks, we propose the MLL module that progressively aggregates features from cascaded KNN Graphs to capture multi-scale local features and avoid complex computation for a large-scale KNN Graph. Assuming that there are n Neighborhood Aggregation blocks cascade, the multi-level aggregation can be formulated as:

$$f_i^r = \begin{cases} f_i & \text{if } r = 1 \\ \mathbf{NA}\left(f_i^{r-1}\right) & \text{if } 1 < r \le n, \end{cases} \tag{7}$$

where $\mathbf{NA}(\cdot)$ denotes Neighborhood Aggregation, including two steps: Obtaining the features between f_i^{r-1} and each neighbor as $\{e_{ij}^{r-1} | 1 < j < k\}$ from $G_i^{(local,r-1)} = (V_i^{(local,r-1)}, E_i^{(local,r-1)})$ by KNN Graph construction, where $e_{ij}^{r-1} = [f_i^{r-1} \parallel f_i^{r-1} - f_{ij}^{r-1}], j = 1, 2, \dots, k$, and aggregating the features $\{e_{ij}^{r-1} | 1 < j < k\} \to f_i^r$ by passing messages along graph edges $E_i^{(local,r-1)}$ with an annular convolutional layer. When $1 < r \le n$, f_i^r is not only the input feature of the rth Neighborhood Aggregation block, but also the output feature of the (r-1)th Neighborhood Aggregation block. Finally, we concatenate the resolution-like local features and obtain $f_i^{out} = concat(f_i^1, f_i^2, \dots, f_i^n)$, followed by a ResNet Block to fuse these features and obtain the local feature set $\mathbf{F}^{local} = \{f_i^{local}\}$.

3.3. GroupGCN

Most existing works (Zhao et al., 2021; Liu et al., 2021c) focus on local and global consensus, which represent a very small and large perspective, respectively. However, the considerable disparity between these two perspectives may result in challenges for the stable optimization of the network. Meanwhile, only correspondences that strictly satisfy both local and global consensus are preserved. Nevertheless, local distortions and non-rigid transformations are commonly found in remote sensing images, which means that even inliers cannot guarantee strict adherence to both local and global consensus. To address this issue, we propose group consensus, which focuses on a balanced perspective and acts as a buffer organization from local to global consensus, called group graph convolution neural network (GroupGCN) Module, shown in Fig. 5. Specifically, the local feature set $\mathbf{F}^{local}$ is encoded to obtain the local consensus set $\mathbf{w}^{local} \in \mathbb{R}^{N\times 1}$ through an MLP layer. Next, $\mathbf{F}^{local}$ and $\mathbf{w}^{local}$ are divided evenly into m groups along the spatial dimension, as shown in Fig. 6. The above grouping details can be formulated as:

$$\mathbf{F}^{local} = \left\{\mathbf{F}_i^{local}\right\}, \mathbf{F}_i^{local} \in \mathbb{R}^{\frac{N}{m}\times C}, i = 1, \dots, m, \tag{8}$$

$$\mathbf{w}^{local} = \left\{\mathbf{w}_i^{local}\right\}, \mathbf{w}_i^{local} \in \mathbb{R}^{\frac{N}{m}\times 1}, i = 1, \dots, m. \tag{9}$$

For each group, we construct a group graph: $G_i^{group} = \left(V_i^{group}, E_i^{group}\right)$, where V_i^{group} denotes all local features in $\mathbf{F}_i^{local}$, edges E_i^{group} connects any two local features within the group. Similar to Zhao et al. (2021), we also use a graph convolutional network (GCN) (Kipf and Welling, 2016) to learn the correlations among correspondences within the group. We define the adjacency matrix as $\mathbf{A}_i = \mathbf{w}_i^{local} \cdot \mathbf{w}_i^{local^T}$. Then, the group feature can be obtained as follows:

$$\mathrm{F}_i^{group} = \delta(\mathbf{L}_i^{local}\mathbf{F}_i^{local}\mathbf{W}_i^{group}), \tag{10}$$

where the Laplacian matrix $\mathbf{L}_i = \widetilde{\mathbf{D}}_i^{-\frac{1}{2}}\widetilde{\mathbf{A}}_i\widetilde{\mathbf{D}}_i^{-\frac{1}{2}}$, $\widetilde{\mathbf{A}}_i = \mathbf{A}_i + \mathbf{I}$ with $\mathbf{I}$ being an identity matrix. $\widetilde{\mathbf{D}}_i$ is the diagonal degree matrix of $\widetilde{\mathbf{A}}_i$. $\delta(\cdot)$ is an activation function. $\mathbf{W}_i^{group}$ is a learnable matrix. After obtaining all F_i^{group}, they are concatenated along the spatial dimension, followed by a series of ResNet Blocks to perform further feature learning. Thus, the group feature $\mathbf{F}^{group}$ is obtained for subsequent global feature learning.

Compared to using a KNN Graph for extracting local consensus, the random grouping strategy applied to all correspondences allows correspondences belonging to the same group to spread over a large region, thereby expanding the receptive field. The correlations learned between correspondences within each group using GCN are referred to as "group consensus". The receptive field of group consensus ranges between local and global consensus, thus acting as a buffer from local to global consensus.

3.4. Multi-order global feature aggregation module

The Laplacian matrix used in graph convolutional neural network (GCN) is an important tool for representing the topological structure of a graph and Laplacian matrices of different orders have different potential applications. Existing networks rely solely on first-order Laplacian matrix to transmit node features, representing the relationship between correspondence and its neighboring neighbors, which may not be sufficient for classification tasks in complex scenarios. Notably, a high-order Laplacian can increase the graph connectivity (Gao and Ji, 2019). For instance, a k-order Laplacian builds links between nodes whose distances are at most k hops (Chepuri and Leus, 2016).

To this end, we consider using multi-order Laplacian matrices to learn global consensus. Specifically, the group feature set $\mathbf{F}^{group}$ is

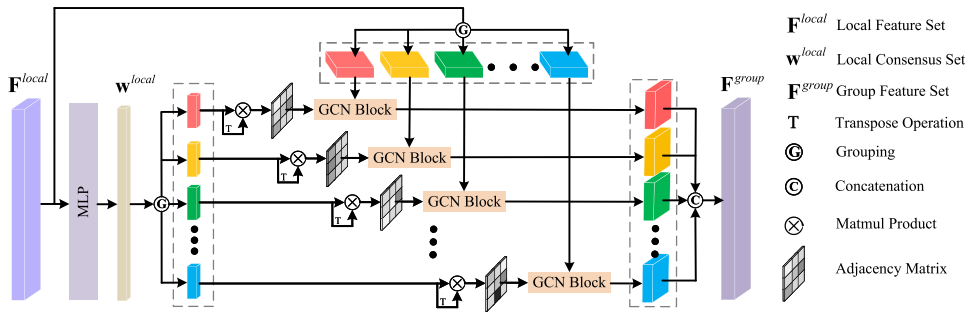

Fig. 5. The architectural details of the GroupGCN module. It uses GCN Blocks to learn the correlations among correspondences, which is permutation-equivariant.

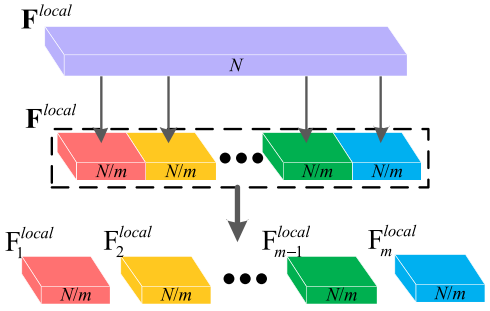

Fig. 6. The division of the local feature set $\mathbf{F}^{local}$.

encoded to obtain the group consensus set $\mathbf{w}^{group} \in \mathbb{R}^{N\times 1}$ through an MLP layer. The first-order global features $\mathbf{F}^{(1)}$ can be obtained with a GCN block, followed by a sigmoid activation function to enhance the nonlinear transformation, which can be expressed as:

$$\mathbf{F}^{(1)} = \delta(\mathbf{L}\mathbf{F}^{group}\mathbf{W}_1^{global}), \tag{11}$$

where the first-order Laplacian matrix $\mathbf{L}$ is determined based on the group consensus set $\mathbf{w}^{group}$. To obtain high-order global features, we stack GCN blocks for feature learning. The high-order Laplacian matrices can be achieved by using consensus sets derived from high-order global features. We show the second and third-order global feature learning as follows:

$$\mathbf{F}^{(2)} = \delta(\mathbf{L}\mathbf{F}^{(1)}\mathbf{W}_2^{global}), \tag{12}$$

$$\mathbf{F}^{(3)} = \delta(\mathbf{L}\mathbf{F}^{(2)}\mathbf{W}_3^{global}), \tag{13}$$

where $\mathbf{W}_i^{global}$ is a learnable matrix and $i \in \{1, 2, 3\}$. We concatenate the multi-order global features and use an MLP layer to fuse them, resulting in the comprehensive global feature $\hat{\mathbf{F}}^{global}$. The global consensus set $\mathbf{w}^{global} \in \mathbb{R}^{N\times 1}$ can be obtained by applying an MLP layer to $\hat{\mathbf{F}}^{global}$.

3.5. Loss function

Following CLNet (Zhao et al., 2021), we adopt a hybrid loss function to guide the training of our MGCNet, which includes a classification loss and an essential matrix regress loss $L_{ess}(\cdot,\cdot)$

$$L = L_{cls} + \alpha L_{ess}(\mathbf{E}, \hat{\mathbf{E}}), \tag{14}$$

where $\mathbf{E}$ and $\hat{\mathbf{E}}$ are the ground truth essential matrix and the predicted essential matrix, respectively; α is a weight parameter, which is used to balance these two losses. L_{cls} is formulated as

$$\begin{aligned} L_{cls} = \sum_{m=1}^{M} \big(& \ell_{bce}(\sigma(\tau_m \odot \mathbf{w}_m^{local}), y_m) \\ & + \ell_{bce}(\sigma\left(\tau_m \odot \mathbf{w}_m^{group}\right), y_m) \\ & + \ell_{bce}(\sigma(\tau_m \odot \mathbf{w}_m^{global}), y_m)\big) \\ & + \ell_{bce}\left(\sigma\left(\hat{\tau} \odot \hat{\mathbf{o}}\right), \mathbf{y}\right), \end{aligned} \tag{15}$$

where ℓ_{bce} is a binary cross entropy loss; $\mathbf{w}_m^{local}$, $\mathbf{w}_m^{group}$, $\mathbf{w}_m^{global}$ represent the consensus sets of three-granularity modules in the mth pruning block, respectively; $\hat{\mathbf{o}}$ is the last output in our MGCNet; $\mathbf{y}$ denotes the set of binary ground-truth labels; $\hat{\tau}$ is an adaptive temperature to alleviate the effect of label ambiguity; σ indicates the sigmoid function; $\odot$ represents the Hadamard product; M is the number of pruning blocks. The essential matrix loss is a geometry loss between the predicted essential matrix $\hat{\mathbf{E}}$ and the ground-truth essential matrix $\mathbf{E}$, $L_{ess}(\hat{\mathbf{E}}, \mathbf{E})$ is formulated as

$$L_{ess}(\hat{\mathbf{E}}, \mathbf{E}) = \frac{(\mathbf{p}_2^{\mathrm{T}}\hat{\mathbf{E}}\mathbf{p}_1)^2}{\|\mathbf{E}\mathbf{p}_1\|_{[1]}^2 + \|\mathbf{E}\mathbf{p}_1\|_{[2]}^2 + \|\mathbf{E}^{\mathrm{T}}\mathbf{p}_2\|_{[1]}^2 + \|\mathbf{E}^{\mathrm{T}}\mathbf{p}_2\|_{[2]}^2}, \tag{16}$$

where $\mathbf{p}_1$ and $\mathbf{p}_2$ are virtual correspondences generated by the ground truth $\mathbf{E}$ matrix. $\mathbf{t}_{[i]}$ refers to the ith element of vector $\mathbf{t}$.

3.6. Implementation details

As shown in Fig. 3, MGCNet comprises two pruning blocks designed to identify reliable candidates for estimating essential matrix and inlier probabilities. Each pruning block consists of ResNet Block, MLL Module, GroupGCN Module, MLG Module, and MLP Layer. Specifically, the ResNet Block includes Perceptron, Context Normalization, Batch Normalization, and ReLU activation function. Following CLNet (Zhao et al., 2021), we sample 50% correspondences as candidates in each pruning block, i.e., $N_2 = \frac{1}{2}N_1 = \frac{1}{4}N$. Additionally, we use SIFT to establish the initial correspondence set, which involves up to 2000 correspondences. The channel dimension C is set to 128. In the MLL, we halve the channel dimension C for k nearest neighbor search to balance effectiveness and computational cost. The number of groups in the MOG is set to 4. Note that we also add the group consensus to the loss function for networking training. MGCNet is trained on PyTorch, adopting a warmup strategy instead of a fixed learning rate. The learning rate starts with linearly increasing for the first 10k iterations, followed by decreasing every 20k iterations with a factor of 0.4. Adam optimization is used for a total of 50k iterations with a batchsize of 32. And the parameter α is set as 0 at the beginning, and after 20k iterations, it is changed to 0.5. All experiments are conducted on Ubuntu 18.04 with an NVIDIA GTX 3090 GPU.

4. Experiments

In this section, we first introduce the datasets and evaluation protocols. To validate the performance of the proposed MGCNet, we further conduct comparative experiments on correspondence pruning, camera pose estimation, remote sensing image registration. Finally, to demonstrate the effectiveness of each component in MGCNet, we also perform comprehensive ablation studies.

Table 1
Performance comparison of correspondence pruning on outdoor and indoor datasets in both known and unknown scenes. The best performance is highlighted in bold.

Datasets	Outdoor (%)						Indoor (%)					
Method	Known scene			Unknown scene			Known scene			Unknown scene		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
RANSAC (Fischler and Bolles, 1981)	47.35	52.39	49.74	43.55	50.65	46.83	51.87	56.27	53.98	44.87	48.82	46.76
MAGSAC (Barath et al., 2019)	45.15	62.36	50.26	44.41	54.46	50.01	–	–	–	–	–	–
GMS (Bian et al., 2017)	47.75	47.92	47.83	41.84	47.91	44.67	–	–	–	–	–	–
LPM (Ma et al., 2019c)	43.75	65.65	51.72	44.28	55.42	50.63	–	–	–	–	–	–
LMR (Ma et al., 2019a)	50.75	66.12	55.19	44.88	58.21	52.71	–	–	–	–	–	–
Point-Net++ (Qi et al., 2017b)	49.62	86.19	62.98	46.39	84.17	59.81	52.89	86.25	65.57	46.30	82.72	59.37
LFGC (Yi et al., 2018)	54.43	86.88	66.93	52.84	85.68	65.37	53.70	87.03	66.42	46.11	83.92	59.37
OANet++ (Zhang et al., 2019)	60.03	89.31	71.80	55.78	85.93	67.65	54.30	88.54	67.32	46.15	84.36	59.66
ACNe (Sun et al., 2020)	60.02	88.99	71.69	55.62	85.47	67.39	54.11	88.46	67.15	46.16	84.01	59.58
CLNet (Zhao et al., 2021)	76.01	79.56	77.74	74.89	76.79	75.83	65.06	73.78	69.15	59.97	68.35	63.89
OANet+++ (Zhang et al., 2022)	60.76	90.84	72.82	56.64	87.69	68.82	55.38	88.56	68.14	47.32	84.40	60.64
MS^{2}DGNet (Dai et al., 2022)	63.17	90.98	74.57	59.11	88.40	70.85	54.50	88.63	67.50	46.95	84.55	60.37
MSA-LFC (Wang et al., 2023b)	64.38	**91.85**	75.70	59.67	**88.42**	71.25	55.82	**88.78**	68.54	47.86	**84.84**	61.20
MGCNet	**77.97**	82.02	**79.94**	**76.94**	79.14	**78.02**	**65.81**	74.55	**69.91**	**60.78**	69.24	**64.73**

4.1. Datasets

YFCC100M (Thomee et al., 2016) is used for outdoor scene evaluation. The dataset consists of 72 sequences, of which 68 sequences are fixedly selected for training and the remaining 4 sequences are used for testing. SUN3D (Xiao et al., 2013) is used for indoor scene evaluation, which is a sequence of image frames from indoor RGB-D videos. The dataset comprises 254 sequences, of which 239 sequences are fixedly selected for training and the remaining 15 sequences for testing. The sequences for training and test are the same for all of the methods. Following Zhang et al. (2019), for both of the aforementioned datasets, the test set is considered as the unknown scene, while 20% of the training set is considered as the known scene. YFCC100M and SUN3D are used for the evaluation of correspondence pruning and camera pose estimation.

Remote Sensing dataset (Ma et al., 2022a) is primarily utilized for evaluating image registration. We select 60 pairs of remote sensing images with different scenes and deformations for evaluation. These selected image pairs suffer from severe noise, affine transformations, small overlaps, and large viewpoint changes.

UAV'Rice comprises a total of 257 color image pairs, which are captured by an unmanned aerial vehicle (UAV). The UAV images are of size 819×546. These image pairs are instrumental in addressing the challenge of automatic crop monitoring and involve intricate conditions such as large viewpoint changes, projection transformations, and a substantial amount of similar local structures.

4.2. Evaluation protocols

We evaluate the performance of the proposed network on various tasks. For correspondence pruning, we choose Precision (P), Recall (R), and F-score (F) as the metric. For camera pose estimation, we use the mean average precision under 5° (mAP5°) as the metric (Dai et al., 2022). For image registration, we select the root mean square error (RMSE), median error (MEE), and maximum error (MAE) as evaluation metrics (Ma et al., 2022a).

4.3. Correspondence pruning for YFCC100M and SUN3D

We first evaluate the proposed MGCNet on correspondence pruning task, which serves as the crucial prerequisite for many computer vision tasks. Specifically, we select four traditional handcrafted methods, including classical RANSAC (Fischler and Bolles, 1981), MAGSAC (Barath et al., 2019), GMS (Bian et al., 2017), LPM (Ma et al., 2019c), and ten leaning-based methods including LMR (Ma et al., 2019a), Point-Net++ (Qi et al., 2017b), LFGC (Yi et al., 2018), OANet++ (Zhang et al., 2019), ACNe (Sun et al., 2020), CLNet (Zhao et al., 2021), LMCNet (Liu et al., 2021c), OANet+++ (Zhang et al., 2022) and MS^{2}DGNet (Dai et al., 2022) for comparison. Note that, OANet+++ (Zhang et al., 2022) is an improved version of OANet++ (Zhang et al., 2019), which enhances the network performance by incorporating additional information as extra inputs, such as the mutual nearest neighbors and the test ratio computed from the descriptors.

Table 1 reports the quantitative comparison results on both outdoor and indoor scenes. It can be observed that learning-based methods consistently outperform traditional methods due to the high proportion of outliers present in both datasets. The proposed MGCNet outperforms all competitors in terms of Precision and F-score across all scenes. Specifically, compared to the second ranked method, the F-score of our MGCNet improves by 2.20% and 2.19% on the known and unknown outdoor scenes, respectively. For indoor scenes, our MGCNet also achieves better Precision and F-score, with improvements over the second-best of 0.81% and 0.76% for known scenes and 0.81% and 0.84% for unknown scenes, respectively. Compared to other methods except CLNet, MGCNet achieves a significant lead in Precision with a certain sacrifice in Recall. This implies that MGCNet is more effective in learning the consensuses among inliers, with only a few matches being misjudged. Compared with the baseline model CLNet, the proposed GroupGCN introduces a group consensus, which acts as a buffer organization from local to global consensus and helps find more consistent inliers (reflected in higher Recall values). Meanwhile, the proposed MLL and MLG modules can effectively boost local and global consensus learning (reflected in higher Precision values).

Fig. 7 presents a comparison of the qualitative visualization results achieved by our MGCNet and several other representative methods. It is evident from the figure that MGCNet successfully removes most outliers and accurately identifies inliers despite encountering challenges such as a high proportion of outliers, the wide baseline, textureless regions, etc. By contrast, the other methods are considerably less robust to such challenges, leading to the existence of many outliers.

4.4. Camera pose estimation for YFCC100M and SUN3D

Camera pose estimation aims to recover a reliable camera pose in the world system, which is the basis of 3D reconstruction, and stereo vision. Compared to correspondence pruning, camera pose estimation is more challenging because the network not only needs to identify the inliers, but also needs to learn a geometric model that the inliers adhere to. We select traditional methods such as classical RANSAC (Fischler and Bolles, 1981) and its variants (DEGENSAC (Chum et al., 2005), GC-RANSAC (Barath and Matas, 2018), MAGSAC (Barath et al., 2019), and MAGSAC++ (Barath et al., 2020)), and learning-based methods, including Point-Net++ (Qi et al., 2017b), LFGC (Yi et al., 2018), OANet++ (Zhang et al., 2019), ACNe (Sun et al., 2020), SuperGlue (Sarlin et al., 2020), CLNet (Zhao et al., 2021), LMCNet (Liu et al., 2021c), OANet+++ (Zhang et al., 2022),

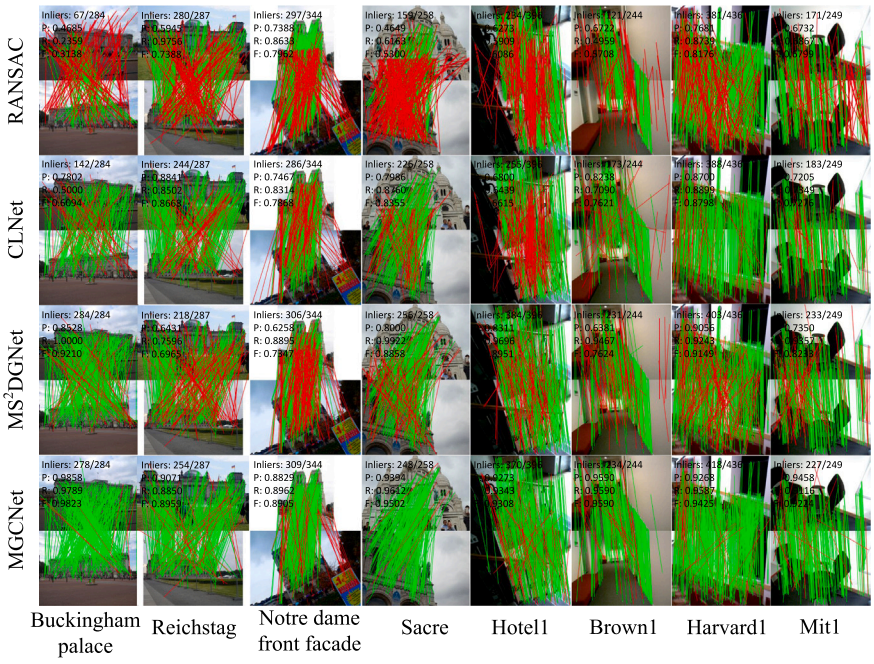

Fig. 7. Partial typical visualization results of RANSAC, MS^{2}DGNet, CLNet, and MGCNet (from top to bottom). The first four columns are from the YFCC100M dataset. The remaining images are from the SUN3D dataset. The green lines and red lines represent inliers and outliers, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Performance comparison of camera pose estimation on outdoor and indoor datasets in both known and unknown scenes with SIFT descriptor. The mAP5° (%) is reported without/with RANSAC. And model parameters of all deep methods are listed in the second column. The best performance is highlighted in bold.

Matcher	Param (MB)	Outdoor (%)				Indoor (%)			
		Known scene		Unknown scene		Known scene		Unknown scene	
		–	RANSAC	–	RANSAC	–	RANSAC	–	RANSAC
RANSAC (Fischler and Bolles, 1981)	–	–	05.81	–	09.07	–	04.52	–	02.84
DEGENSAC (Chum et al., 2005)	–	–	21.00	–	27.65	–	16.01	–	11.01
GC-RANSAC (Barath and Matas, 2018)	–	–	30.43	–	41.58	–	18.86	–	14.14
MAGSAC (Barath et al., 2019)	–	–	32.80	–	41.61	–	20.35	–	16.24
MAGSAC++ (Barath et al., 2020)	–	–	30.48	–	40.95	–	18.90	–	14.19
Point-Net++ (Qi et al., 2017b)	12.00	10.49	33.78	16.48	46.25	10.58	19.17	08.10	15.29
ACNe (Sun et al., 2020)	0.41	29.17	40.32	33.06	50.89	18.86	22.12	14.12	16.99
LFGC (Yi et al., 2018)	0.39	13.81	34.55	23.95	48.03	11.55	20.60	9.30	16.40
OANet++ (Zhang et al., 2019)	2.47	32.57	41.53	38.95	52.59	20.86	22.31	16.18	17.18
SuperGlue (Sarlin et al., 2020)	12.02	35.00	43.17	48.12	55.06	22.50	23.68	17.11	18.23
CLNet (Zhao et al., 2021)	1.27	39.00	45.22	54.05	59.70	20.62	24.15	16.95	18.87
LMCNet (Liu et al., 2021c)	0.97	33.73	40.39	47.50	55.03	19.92	21.79	16.82	17.38
OANet+++ (Zhang et al., 2022)	2.47	37.48	44.73	43.13	55.95	22.74	23.02	17.39	17.44
MS^{2}DGNet (Dai et al., 2022)	2.61	38.36	45.34	49.13	57.68	22.20	23.00	17.84	17.79
MSA-LFC (Wang et al., 2023b)	1.73	44.60	46.19	53.62	57.25	22.84	22.64	18.41	17.80
MGCNet	2.05	**48.55**	**48.79**	**62.55**	**63.22**	**23.22**	**25.13**	**19.30**	**19.96**

Table 3
Performance comparison of camera pose estimation on outdoor scenes with SuperPoint descriptor. The mAP5° (%) is reported without/with RANSAC. The best performance is highlighted in bold.

Matcher	Known scene		Unknown scene	
	–	RANSAC	–	RANSAC
RANSAC (Fischler and Bolles, 1981)	–	12.85	–	17.47
Point-Net++ (Qi et al., 2017b)	11.87	28.46	17.95	38.83
ACNe (Sun et al., 2020)	26.72	31.16	32.98	45.34
LFGC (Yi et al., 2018)	12.18	30.25	24.25	42.57
OANet++ (Zhang et al., 2019)	29.52	35.72	35.27	45.45
CLNet (Zhao et al., 2021)	27.93	32.75	38.48	45.02
OANet+++ (Zhang et al., 2019)	30.62	35.73	36.85	45.97
MS^{2}DGNet (Dai et al., 2022)	30.40	36.02	37.38	46.48
MSA-LFC (Wang et al., 2023b)	32.58	36.67	42.07	48.53
MGCNet	**34.33**	**36.83**	**44.49**	**50.33**

MS^{2}DGNet (Dai et al., 2022), and MSA-LFC (Wang et al., 2023b) for comparison.

Table 2 shows the quantitative comparison results on both outdoor and indoor scenes. Similar to the results of the correspondence pruning, learning-based methods demonstrate superior performance compared to traditional ones. Traditional methods employing the hypothesize-verification framework often fail to achieve satisfactory results in the presence of a high proportion of outliers. By contrast, learning-based methods benefit from the powerful feature representation and generalization ability of deep neural networks, enabling them to learn more compact correlations among inliers and fit more accurate essential matrices. Among all learning-based methods, our MGCNet gains the best results under all testing scenes. Specifically, our MGCNet obtains 3.95% and 8.50% improvement in terms of mAP5° on known and unknown scenes of outdoor scenes, compared to the second-best models (MSA-LFC and CLNet) without RANSAC. It is worth noting that other learning-based methods often require RANSAC as a post-processing step

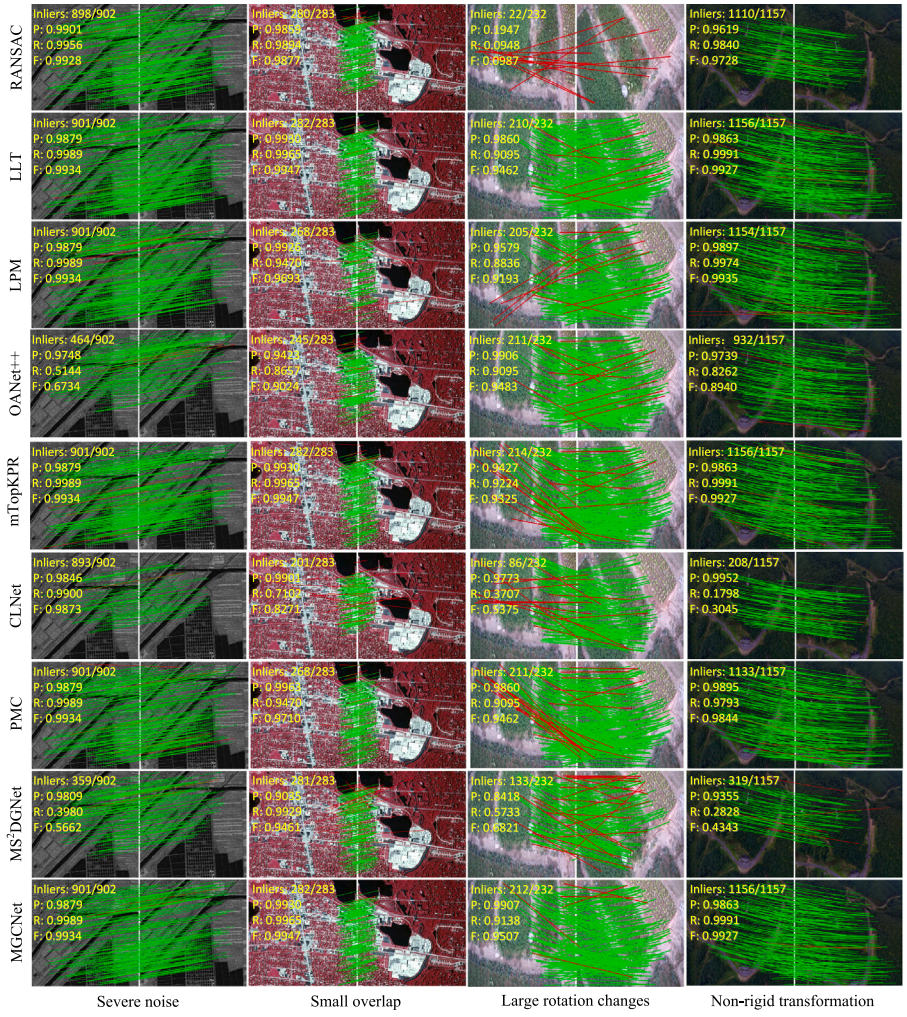

Fig. 8. Partial typical visualization results of correspondence pruning on remote sensing dataset. The image pairs suffer from severe noise, small overlap, large rotation changes, and non-rigid transformation, respectively. From top to bottom: results of RANSAC with $1k$ iterations, LLT, LPM, OANet++, mTopKRP, CLNet, PMC, MS^{2}DG-Net and MGCNet. The green lines and red lines represent inliers and outliers, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

to achieve satisfactory results, whereas MGCNet can achieve superior results independently. The good performance of MGCNet demonstrates that the proposed MLL, MOG, and GroupGCN can effectively perform multi-granularity consensus learning and enhance the interactions between inliers, producing accurate camera pose estimation results. Moreover, Table 2 also gives model parameters of all learning-based methods. We can see that our MGCNet gives the best performance with a reasonable computation complexity and a competitive advantage.

In addition to evaluating the robustness of MGCNet for camera pose estimation with different feature descriptors, we also use the learning-based descriptor SuperPoint (DeTone et al., 2018) to construct the initial correspondence set. Table 3 reports the performance comparison of camera pose estimation on outdoor scenes with the SuperPoint descriptor. Our MGCNet also achieves superior performance, which is consistent with the results using the SIFT descriptor, demonstrating the robustness of MGCNet.

4.5. Remote sensing image registration

The performance of MGCNet is further evaluated on remote sensing image registration. Image registration mainly focuses on maximizing the alignment of the overlapped area between the reference image and the transformed sense image. Although image registration in remote sensing is also achieved by estimating a transformation matrix, unlike the parametric models such as essential matrix, the remote sensing image pairs involve local distortion and non-rigid transformation. Thus, the transformation can only be modeled by methods such as TPS. To achieve better registration performance, the inliers identified by the network should be distributed as evenly as possible in various regions of the image instead of being concentrated only in a local area, which requires the network to fully explore the consensus among inliers.

The model trained on the YFCC100M dataset is chosen for evaluation. We first show the visualization results of correspondence pruning and image registration on several typical image pairs in Figs. 8 and 9, respectively. It can be found that the matching results of MGCNet can cover a larger overlap area than other competitors. Using multi-granularity consensus, MGCNet has the advantage of being able to handle unevenly distributed keypoints. Furthermore, as shown in Fig. 9, MGCNet produces impressive results for all challenges. In contrast to the baseline model CLNet, MGCNet can effectively handle image pairs with a small overlap area. In addition, we also present the cumulative distribution results for image registration, including the root mean square error (RMSE), mean Euclidean error (MEE), mean absolute error (MAE), and run time (RT) in Fig. 10 and their average values in

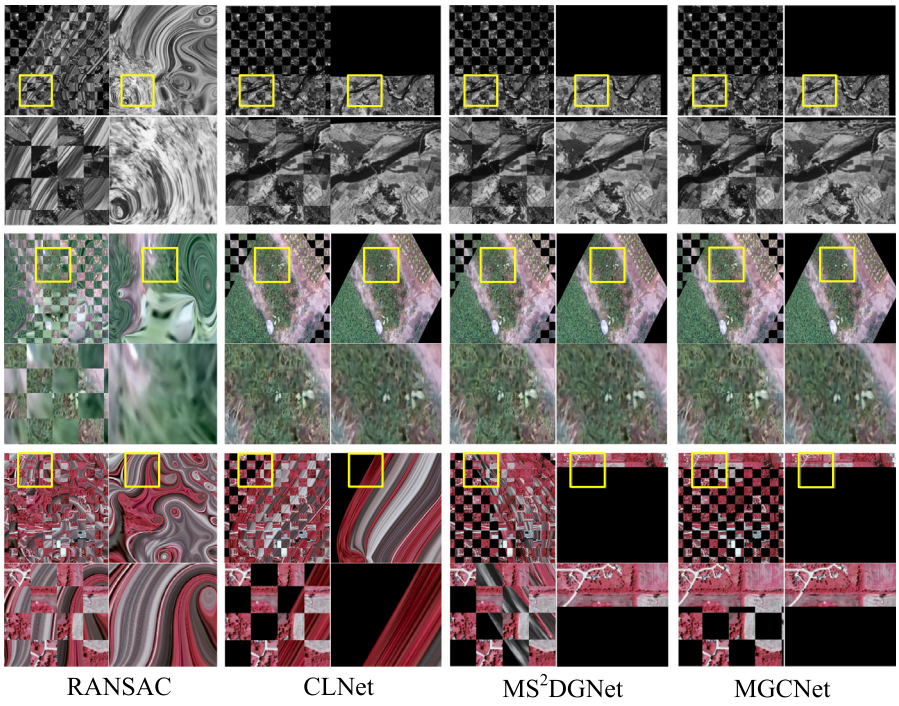

Fig. 9. Partial typical visualization results of image registration on remote sensing dataset. The image pairs suffer from severe noise, small overlap, large rotation changes, and non-rigid transformation, respectively. From left to right: the results of RANSAC with $1k$ iterations, CLNet, MS^{2}DG-Net, and MGCNet, where the left and right images in each group are checkerboard images and warped sensed results, respectively, and the second row ones show the local region.

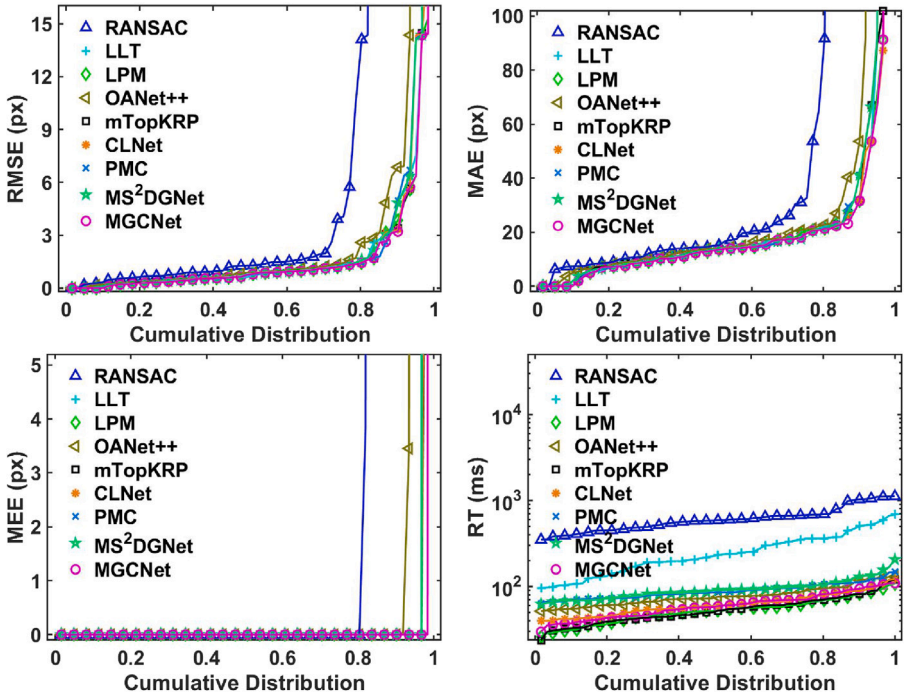

Fig. 10. Quantitative statistics of image registration. A point coordinate (x, y) on this curve presents that there are (100*x)% percent of image pairs whose performance values (i.e., RMSE, MAE, MEE, and RT) do not exceed y.

Table 4. From the quantitative results, we can find that RANSAC with $1k$ iterations fails to achieve satisfactory registration performance since it is sensitive to a high proportion of outliers and can only produce the parametric model, which is not suitable for non-rigid transformations between remote sensing image pairs. Among learning-based methods, MGCNet achieves nearly optimal performance across all metrics, except for a small amount of additional run time. In summary, the proposed MGCNet also performs well on remote sensing image registration.

4.6. UAV remote sensing image registration

To further evaluate the generalization ability of MGCNet, we test it on a new UAV remote sensing image dataset, i.e. the UAV'Rice dataset. We extract the SIFT features and construct putative correspondences based on the similarity of feature descriptors with the nearest neighbor matching strategy. Similar to Ma et al. (2022a), the ground truth is

Table 4
Performance comparison on remote sensing data with SIFT. The average RMSE, MAE, MEE, and RT are reported. The best performance is highlighted in bold.

Method	RMSE ↓	MAE ↓	MEE ↓	RT (ms) ↓
RANSAC (Fischler and Bolles, 1981)	67.69	226.85	71.48	571.51
LLT (Ma et al., 2015)	4.309	30.866	4.1506	264.89
LPM (Ma et al., 2019c)	1.47	22.79	**0.0004**	**58.67**
OANet++ (Zhang et al., 2019)	18.24	69.26	19.82	82.94
mTopKRP (Jiang et al., 2019)	1.456	22.82	**0.0004**	51.696
CLNet (Zhao et al., 2021)	5.045	30.07	6.1805	75.38
PMC (Liu et al., 2022)	1.448	22.76	**0.0004**	63.825
MS^{2}DGNet (Dai et al., 2022)	3.59	73.26	0.0475	86.18
MGCNet	**1.444**	**20.88**	**0.0004**	95.19

established concerning a benchmark prepared in advance for objectivity, and each putative correspondence in each image pair is checked manually.

To comprehensively evaluate the efficacy of the proposed network, we report the results of correspondence pruning (P, R, F) and image registration (RMSE, MAE, MEE, RT) in Table 5. We choose RANSAC and three learning-based methods for comparison. All models are trained on the YFCC100M dataset with SIFT descriptor. It can be observed that our MGCNet outperforms other competitors. MS^{2}DGNet, only constructing local graphs to obtain local consensus, is easily disturbed by a large amount of similar local structures. Therefore, many inliers cannot be identified, resulting in a poor registration model. Compared to CLNet, MGCNet can predict more inliers and gain a better registration model thanks to group consensus, which can compensate for the lack of medium-granularity consensus. This demonstrates that our MGCNet has a stronger generalization ability since it identifies correct correspondences by multi-granularity consensus learning, and shows better effectiveness for UAV remote sensing image correspondence pruning and registration.

4.7. Ablation studies

Ablation studies involve the modification of certain components or parameters to observe their impact on the models performance. To demonstrate the effectiveness of each component of the proposed MGCNet, we conduct ablation studies on camera pose estimation on the YFCC100M dataset. mAP5° (%) and mAP20° (%) on both known and unknown scenes are reported for evaluation.

Multi-granularity Consensus. In MGCNet, we design three modules to exploit three types of consensus of correspondences, i.e., multi-level local feature aggregation module (MLL) for local consensus, GroupGCN Module for group consensus, and multi-order global feature aggregation Module (MOG) for global consensus. Firstly, we add each of them to the baseline model CLNet, as shown in the 2nd–4th rows of Table 6. It can be found that each module significantly enhances performance. For instance, the GroupGCN module achieves a 5.06% and a 1.79% improvement compared to the baseline model on known scenes for mAP5° without and with RANSAC, respectively. Comparison between the 1st and 4th rows of Table VI demonstrates that, in contrast to the baseline model CLNet, which only utilizes a first-order Laplacian matrix to extract the global context of correspondences, our MOG module enriches the extracted global context by introducing multi-order Laplacian matrices. In this process, lower-order features provide a foundation for higher-order features. By integrating these multi-order features, our performance has been significantly improved. Moreover, the same conclusion can be drawn from the comparisons between the 2nd and 6th rows, the 3rd and 7th rows, and the 5th and 8th rows of Table 6. Subsequently, we conduct three additional experiments by combining any two modules and adding them to the baseline model, as shown in the 5th–7th rows of Table 6. We observe that the joint utilization of both modules yields better results compared to each individual module, which indicates that the proposed modules can interact with each other

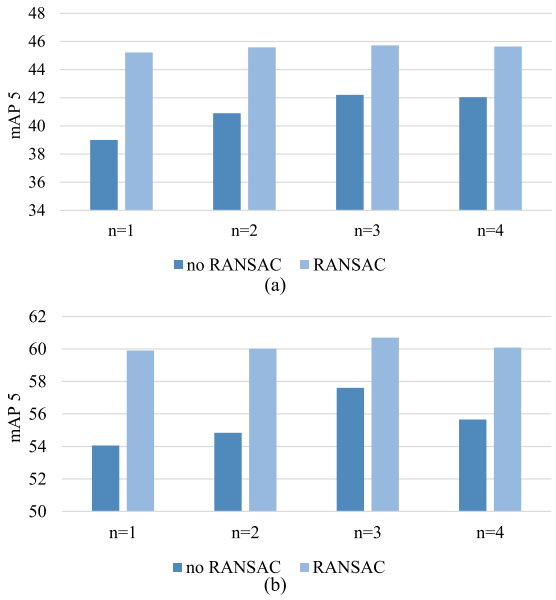

Fig. 11. Performance statistics of different iteration number n with respect to Neighborhood Aggregation in MLL on the YFCC100M dataset. (a) shows the results on the unknown scene and (b) refers to the known scene.

and have a complementary effect. Finally, we achieve the best results by adding all modules to the baseline, i.e. MGCNet shown in the last row of Table 6. This indicates that MGCNet learning Multi-granularity consensus is vital to identify inliers.

Effect of Iteration Number of Neighborhood Aggregation in MLL. We also investigate the effect of the iteration number of Neighborhood Aggregation in MLL. The model trained on the YFCC100M dataset is chosen for evaluation. As shown in Fig. 11, we can find that with the increase of iteration number, MGCNet demonstrates improved performance on both known and unknown scenes, with and without RANSAC for post-processing, indicating that the multi-level local feature aggregation module can obtain rich and diverse local contextual information. However, the performance slightly declines when the iteration number reaches four, indicating that the network is somewhat overfitting. Consequently, to balance efficiency and performance, we opt for an iteration number $n = 3$ as the default setting.

Effect of Group Number of Correspondences in GroupGCN. The group number of correspondences in GroupGCN, denoted as m, significantly influences the learning of the group consensus. For both training and testing, we set m to 1, 2, 4, 6, and 8. Notably, $m = 1$ corresponds to the solution in CLNet, i.e., only a GCN module is applied to learn global features. From Fig. 12 and Table 7, it can be observed that an increasing m corresponds to continuous performance enhancement. Nevertheless, beyond a group number of 4, the performance improvement plateaus, possibly due to an insufficiency of correspondence with each group. In addition, the increase of m also leads to an augmentation in model parameters. Therefore, we set $m = 4$ as default.

Comparison of MGCNet and CLNet with a larger capacity. To explore the performance comparison between our MGCNet and the baseline CLNet under a similar scale of model capacity, we increased the model parameters of CLNet by duplicating its core modules, including the residual structures. As shown in Table 8, enhancing the model's parameters can improve performance to a certain extent. However, merely increasing network parameters without targeted optimization and design may lead to a large number of redundant parameters in the network. This not only complicates the optimization process but could also induce performance bottlenecks. In contrast, our MGCNet, even with a similar scale of network parameters as CLNet*, demonstrates

Table 5
Performance comparison results of correspondence pruning and image registration on the UAV'Rice dataset. The best performance is highlighted in bold.

Method	Correspondence pruning			Images registration			
	P (%) ↑	R (%) ↑	F (%) ↑	RMSE ↓	MAE ↓	MEE ↓	RT (ms) ↓
RANSAC (Fischler and Bolles, 1981)	78.39	55.83	65.21	51.79	141.62	62.31	593.41
OANet (Zhang et al., 2019)	79.23	75.11	77.12	8.42	27.55	9.70	**66.433**
CLNet (Zhao et al., 2021)	81.30	75.54	78.31	5.20	16.25	6.67	66.919
MS^{2}DGNet (Dai et al., 2022)	71.69	71.00	71.34	17.00	51.63	19.72	71.959
MGCNet	**83.50**	**82.26**	**82.88**	**2.07**	**9.76**	**1.54**	67.699

Table 6
Ablation study on the YFCC100M dataset with SIFT descriptor. The mAP5° (%) and mAP20° (%) on both known and unknown scenes without and with RANSAC are reported. MLL: Multi-level Local Feature Aggregation Module. GroupGCN: GroupGCN Module. MOG: Multi-order Global Feature Aggregation Module.

Baseline	MLL	GroupGCN	MOG	Known				Unknown			
				mAP5°		mAP20°		mAP5°		mAP20°	
				–	RANSAC	–	RANSAC	–	RANSAC	–	RANSAC
✓				39.00	45.22	61.95	67.90	54.05	59.7	76.37	79.65
✓	✓			42.21	45.73	64.92	68.45	57.60	60.70	79.16	80.24
✓		✓		44.06	47.01	66.76	69.68	56.00	59.98	78.79	80.00
✓			✓	42.51	47.18	64.73	69.24	57.35	60.92	78.78	80.41
✓	✓	✓		47.07	48.48	69.03	70.51	61.35	61.88	81.03	81.51
✓	✓		✓	44.03	47.37	66.50	69.68	59.60	61.62	79.77	80.09
✓		✓	✓	45.03	47.42	66.62	69.36	57.07	61.35	78.97	80.78
✓	✓	✓	✓	**48.55**	**48.79**	**70.48**	**71.09**	**62.55**	**63.22**	**82.09**	**81.94**

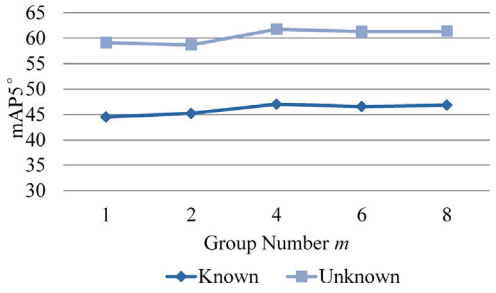

Fig. 12. Ablation study about the different number of groups in the GroupGCN module on the YFCC100M dataset.

Table 7
Ablation study about the different number of groups in the GroupGCN module on the YFCC100M dataset. The mAP5° (%) and mAP20° (%) on both known and unknown scenes without and with RANSAC are reported. Params (MB): the number of network parameters.

m	Known		Unknown		Params (MB)
	mAP5°	mAP20°	mAP5°	mAP20°	
1	44.52/47.84	67.22/69.89	59.10/60.77	79.94/80.45	1.72
2	45.22/47.49	67.26/69.77	58.70/61.52	80.09/80.09	1.75
4	47.04/48.06	69.58/70.70	61.75/62.85	81.76/81.72	1.82
6	46.57/48.01	69.03/70.49	61.30/61.65	81.34/81.01	1.88
8	46.84/47.98	69.36/70.51	61.35/61.88	81.03/81.51	1.95

Table 8
Performance comparison of MGCNet and CLNet* (CLNet with a larger capacity) on the YFCC100M dataset.

Method	Known		Unknown		Params (MB)
	mAP5°	mAP20°	mAP5°	mAP20°	
CLNet	39.00/45.22	61.95/67.90	54.05/59.70	76.37/79.65	1.26
CLNet*	44.76/47.21	66.98/69.18	57.43/61.38	78.85/80.63	2.07
MGCNet	48.55/48.79	70.48/71.09	62.55/63.22	82.09/81.94	2.05

significantly superior performance. This result highlights the rationality and effectiveness of the intrinsic design of MGCNet.

Table 9
Performance comparison under different weight parameter α on YFCC100M with SIFT.

α	Known		Unknown	
	mAP5°	F	mAP5°	F
0.0	36.97/45.31	75.95	54.25/60.50	75.83
0.1	46.85/48.01	79.66	60.85/61.65	78.85
0.2	47.58/48.23	79.99	61.58/62.84	78.45
0.3	47.74/48.32	79.92	60.91/62.13	78.05
0.4	47.68/48.59	79.97	62.13/62.46	78.23
0.5	48.55/48.79	80.13	62.55/63.22	78.11
0.6	47.55/48.48	79.98	62.52/62.92	78.37
0.7	47.24/48.60	79.96	61.52/62.00	78.10
0.8	48.06/48.37	79.90	62.62/63.10	78.16
0.9	47.99/48.75	80.04	60.82/62.42	77.98
1.0	48.43/48.76	79.99	62.75/63.00	78.24

Analysis of Weighting Parameter α. The loss function consists of two terms: one is the classification loss, and the other is the essential matrix loss. The weighting parameter α is used to trade off these two terms. To better determine the value of α, we conduct experiments on correspondence pruning and pose estimation on YFCC100M with SIFT by adjusting α from 0 to 1 with an interval of 0.1. As shown in Table 9, it can be observed that two losses are essential for the optimization process of our network. $\alpha = 0$ indicates that the essential matrix loss is not used, leading to poor performance. Different weight parameters have a slight bias towards correspondence pruning and pose estimation tasks, respectively. When setting $\alpha = 0.5$, the performance of both tasks is balanced. Therefore, we set $\alpha = 0.5$ as the default.

5. Conclusion

This paper proposes the multi-granularity consensus network (MGCNet) for remote sensing image correspondence pruning. In MGCNet, a group consensus is first proposed to alleviate the optimization difficulties caused by a large gap between local and global consensus, which acts as a buffer organization from local to global consensus. To accommodate the uneven distribution of putative correspondences, we design a Multi-level Local Feature Aggregation Module that adaptively

adjusts the size of the local neighborhood to capture local consensus. We further adopt a Multi-order Global Feature Aggregation Module to strengthen global consensus. Extensive experiments on popular benchmarks have demonstrated the effectiveness and generalization ability of the proposed MGCNet over state-of-the-art methods.

To reduce the gap between local and global consensus, we construct multi-granularity consensus features for the correspondences and design MGCNet to increase the receptive field. Compared with existing models, our proposed MGCNet may be more adept at handling large-scale transformations and non-rigid scenarios. However, when the overlapping region between two-view images is small, it may have a negative effect on the grouping of local feature sets. In the future, we will design a better network architecture by deeply studying the relationship between these multi-granularity consensus features.

CRediT authorship contribution statement

Fengyuan Zhuang: Writing – review & editing, Writing – original draft, Software, Methodology, Conceptualization. **Yizhang Liu:** Writing – review & editing. **Xiaojie Li:** Writing – review & editing, Writing – original draft, Visualization, Software. **Ji Zhou:** Writing – review & editing, Supervision, Formal analysis. **Riqing Chen:** Writing – review & editing, Supervision, Funding acquisition. **Lifang Wei:** Writing – review & editing, Project administration. **Changcai Yang:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Investigation, Funding acquisition, Conceptualization. **Jiayi Ma:** Writing – review & editing, Writing – original draft, Supervision, Formal analysis.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grant 62171130, in part by the Natural Science Fund of Fujian Province, China under Grant 2023J01470, in part by the Fujian Innovation Fund Project, China under Grant 2021C0101, in part by the Fujian Province Science and Technology Plan Guided Fund, China under Grant 2021H0013, in part by the Big Data in Agroforestry (Cross-Disciplinary) of Fujian Agriculture and Forestry University, China under Grant 712023030, and in part by the Science and Technology Innovation Special Fund Project of FAFU, China under Grant KFb22096XA.

References

Barath, D., Matas, J., 2018. Graph-cut RANSAC. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6733–6741. http://dx.doi.org/10.1109/CVPR.2018.00704.

Barath, D., Matas, J., Noskova, J., 2019. MAGSAC: marginalizing sample consensus. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10197–10205. http://dx.doi.org/10.1109/CVPR.2019.01044.

Barath, D., Noskova, J., Ivashechkin, M., Matas, J., 2020. MAGSAC++, a fast, reliable and accurate robust estimator. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1304–1312. http://dx.doi.org/10.1109/CVPR42600.2020.00138.

Bian, J., Lin, W.-Y., Matsushita, Y., Yeung, S.-K., Nguyen, T.-D., Cheng, M.-M., 2017. Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4181–4190. http://dx.doi.org/10.1109/CVPR.2017.302.

Brachmann, E., Krull, A., Nowozin, S., Shotton, J., Michel, F., Gumhold, S., Rother, C., 2017. Dsac-differentiable ransac for camera localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6684–6692. http://dx.doi.org/10.1109/CVPR.2017.267.

Chen, S., Xiao, G., Shi, Z., Guo, J., Ma, J., 2024. SSL-Net: Sparse semantic learning for identifying reliable correspondences. Pattern Recognit. 146, 110039. http://dx.doi.org/10.1016/j.patcog.2023.110039.

Chepuri, S.P., Leus, G., 2016. Subsampling for graph power spectrum estimation. In: IEEE Sensor Array and Multichannel Signal Processing Workshop. SAM, pp. 1–5. http://dx.doi.org/10.1109/SAM.2016.7569707.

Chum, O., Werner, T., Matas, J., 2005. Two-view geometry estimation unaffected by a dominant plane. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Vol. 1, IEEE, pp. 772–779. http://dx.doi.org/10.1109/CVPR.2005.354.

Dai, L., Liu, Y., Ma, J., Wei, L., Lai, T., Yang, C., Chen, R., 2022. MS^{2}DG-Net: Progressive correspondence learning via multiple sparse semantics dynamic graph. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8973–8982. http://dx.doi.org/10.1109/CVPR52688.2022.00877.

DeTone, D., Malisiewicz, T., Rabinovich, A., 2018. Superpoint: Self-supervised interest point detection and description. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 224–236. http://dx.doi.org/10.1109/CVPRW.2018.00060.

Fischler, M.A., Bolles, R.C., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. Acm 24 (6), 381–395.

Gao, H., Ji, S., 2019. Graph U-Nets. In: International Conference on Machine Learning. pp. 2083–2092.

Hartley, R., Zisserman, A., 2003. Multiple View Geometry in Computer Vision. Cambridge University Press.

Huang, W., Zhang, G., Han, X., 2020. Dense mapping from an accurate tracking SLAM. IEEE/CAA J. Autom. Sin. 7 (6), 1565–1574. http://dx.doi.org/10.1109/JAS.2020.1003357.

Jiang, X., Jiang, J., Fan, A., Wang, Z., Ma, J., 2019. Multiscale locality and rank preservation for robust feature matching of remote sensing images. IEEE Trans. Geosci. Remote Sens. 57 (9), 6462–6472. http://dx.doi.org/10.1109/TGRS.2019.2906183.

Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.

Lin, W.-Y.D., Cheng, M.-M., Lu, J., Yang, H., Do, M.N., Torr, P., 2014. Bilateral functions for global motion modeling. In: Proceedings of the European Conference on Computer Vision. pp. 341–356.

Lin, W.-Y., Wang, F., Cheng, M.-M., Yeung, S.-K., Torr, P.H., Do, M.N., Lu, J., 2017. CODE: Coherence based decision boundaries for feature correspondence. IEEE Trans. Pattern Anal. Mach. Intell. 40 (1), 34–47. http://dx.doi.org/10.1109/TPAMI.2017.2652468.

Liu, Y., Li, Y., Dai, L., Lai, T., Yang, C., Wei, L., Chen, R., 2021a. Motion consistency-based correspondence growing for remote sensing image matching. IEEE Geosci. Remote Sens. Lett. 19, 1–5. http://dx.doi.org/10.1109/LGRS.2020.3048258.

Liu, Y., Li, Y., Dai, L., Yang, C., Wei, L., Lai, T., Chen, R., 2021b. Robust feature matching via advanced neighborhood topology consensus. Neurocomputing 421, 273–284. http://dx.doi.org/10.1016/j.neucom.2020.09.047.

Liu, Y., Liu, L., Lin, C., Dong, Z., Wang, W., 2021c. Learnable motion coherence for correspondence pruning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3237–3246.

Liu, Y., Zhao, B.N., Zhao, S., Zhang, L., 2022. Progressive motion coherence for remote sensing image matching. IEEE Trans. Geosci. Remote Sens. 60, 1–13. http://dx.doi.org/10.1109/TGRS.2022.3205059.

Liu, Y., Zhou, W., Li, Y., Zhao, S., 2024. RoSe: Rotation-invariant sequence-aware consensus for robust correspondence pruning. In: Proceedings of the 32st ACM International Conference on Multimedia.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60 (2), 91–110. http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94.

Ma, J., Jiang, X., Fan, A., Jiang, J., Yan, J., 2021. Image matching from handcrafted to deep features: A survey. Int. J. Comput. Vis. 129 (1), 23–79. http://dx.doi.org/10.1007/s11263-020-01359-2.

Ma, J., Jiang, X., Jiang, J., Zhao, J., Guo, X., 2019a. LMR: Learning a two-class classifier for mismatch removal. IEEE Trans. Image Process. 28 (8), 4045–4059. http://dx.doi.org/10.1109/TIP.2019.2906490.

Ma, J., Li, Z., Zhang, K., Shao, Z., Xiao, G., 2022a. Robust feature matching via neighborhood manifold representation consensus. ISPRS J. Photogramm. Remote Sens. 183, 196–209. http://dx.doi.org/10.1016/j.isprsjprs.2021.11.004.

Ma, J., Tang, L., Fan, F., Huang, J., Mei, X., Ma, Y., 2022b. SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. IEEE/CAA J. Autom. Sin. 9 (7), 1200–1217. http://dx.doi.org/10.1109/JAS.2022.105686.

Ma, J., Yu, W., Liang, P., Li, C., Jiang, J., 2019b. FusionGAN: A generative adversarial network for infrared and visible image fusion. Inf. Fusion 48, 11–26. http://dx.doi.org/10.1016/j.inffus.2018.09.004.

Ma, J., Zhao, J., Jiang, J., Zhou, H., Guo, X., 2019c. Locality preserving matching. Int. J. Comput. Vis. 127 (5), 512–531. http://dx.doi.org/10.1007/s11263-018-1117-z.

Ma, J., Zhao, J., Tian, J., Yuille, A.L., Tu, Z., 2014. Robust point matching via vector field consensus. IEEE Trans. Image Process. 23 (4), 1706–1721. http://dx.doi.org/10.1109/TIP.2014.2307478.

Ma, J., Zhou, H., Zhao, J., Gao, Y., Jiang, J., Tian, J., 2015. Robust feature matching for remote sensing image registration via locally linear transforming. IEEE Trans. Geosci. Remote Sens. 53 (12), 6469–6481. http://dx.doi.org/10.1109/TGRS.2020.3001089.

Pilet, J., Lepetit, V., Fua, P., 2008. Fast non-rigid surface detection, registration and realistic augmentation. Int. Conf. Comput. Vis. 76 (2), 109–122. http://dx.doi.org/10.1007/s11263-006-0017-9.

Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 652–660. http://dx.doi.org/10.1109/CVPR.2017.16.

Qi, C.R., Yi, L., Su, H., Guibas, L.J., 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Adv. Neural Inf. Process. Syst. 30, 5099–5108.

Sarlin, P.-E., DeTone, D., Malisiewicz, T., Rabinovich, A., 2020. Superglue: Learning feature matching with graph neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Rrecognition. pp. 4938–4947. http://dx.doi.org/10.1109/CVPR42600.2020.00499.

Sun, W., Jiang, W., Trulls, E., Tagliasacchi, A., Yi, K.M., 2020. ACNe: Attentive context normalization for robust permutation-equivariant learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11286–11295. http://dx.doi.org/10.1109/CVPR42600.2020.01130.

Sun, G., Lu, H., Zhao, Y., Zhou, J., Jackson, R., Wang, Y., Xu, L.-x., Wang, A., Colmer, J., Ober, E., Zhao, Q., Han, B., Zhou, J., 2022. AirMeasurer: open-source software to quantify static and dynamic traits derived from multiseason aerial phenotyping to empower genetic mapping studies in rice. New Phytol. 236 (4), 1584–1604. http://dx.doi.org/10.1111/nph.18314.

Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.-J., 2016. YFCC100M: The new data in multimedia research. Commun. ACM 59 (2), 64–73. http://dx.doi.org/10.1145/2812802.

Wang, J., Liu, X., Dai, L., Ma, J., Wei, L., Yang, C., Chen, R., 2023a. PG-Net: Progressive guidance network via robust contextual embedding for efficient point cloud registration. IEEE Trans. Geosci. Remote Sens. 60, 1–12.

Wang, L., Wu, J., Fang, X., Liu, Z., Cao, C., Fu, Y., 2023b. Local consensus enhanced siamese network with reciprocal loss for two-view correspondence learning. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 5235–5243. http://dx.doi.org/10.1145/3581783.3612458.

Xia, Y., Ma, J., 2022. Locality-guided global-preserving optimization for robust feature matching. IEEE Trans. Image Process. 31, 5093–5108. http://dx.doi.org/10.1109/TIP.2022.3192993.

Xiao, J., Owens, A., Torralba, A., 2013. Sun3d: A database of big spaces reconstructed using sfm and object labels. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1625–1632. http://dx.doi.org/10.1109/ICCV.2013.458.

Xu, H., Ma, J., Jiang, J., Guo, X., Ling, H., 2020. U2Fusion: A unified unsupervised image fusion network. IEEE Trans. Pattern Anal. Mach. Intell. 44 (1), 502–518. http://dx.doi.org/10.1109/TPAMI.2020.3012548.

Yang, C., Li, X., Ma, J., Zhuang, F., Wei, L., Chen, R., Chen, G., 2024. CGR-Net: Consistency guided ResFormer for two-view correspondence learning. IEEE Trans. Circuits Syst. Video Technol. 1. http://dx.doi.org/10.1109/TCSVT.2024.3439348.

Yi, K.M., Trulls, E., Ono, Y., Lepetit, V., Salzmann, M., Fua, P., 2018. Learning to find good correspondences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2666–2674. http://dx.doi.org/10.1109/CVPR.2018.00282.

Zhang, H., Ma, J., Chen, C., Tian, X., 2020. NDVI-Net: A fusion network for generating high-resolution normalized difference vegetation index in remote sensing. ISPRS J. Photogramm. Remote Sens. 168, 182–196. http://dx.doi.org/10.1016/j.isprsjprs.2020.08.010.

Zhang, J., Sun, D., Luo, Z., Yao, A., Chen, H., Zhou, L., Shen, T., Chen, Y., Quan, L., Liao, H., 2022. OANet: Learning two-view correspondences and geometry using order-aware network. IEEE Trans. Pattern Anal. Mach. Intell. 44 (6), 3110–3122. http://dx.doi.org/10.1109/TPAMI.2020.3048013.

Zhang, J., Sun, D., Luo, Z., Yao, A., Zhou, L., Shen, T., Chen, Y., Quan, L., Liao, H., 2019. Learning two-view correspondences and geometry using order-aware network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5845–5854. http://dx.doi.org/10.1109/ICCV.2019.00594.

Zhao, C., Cao, Z., Li, C., Li, X., Yang, J., 2019. Nm-net: Mining reliable neighbors for robust feature correspondences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 215–224. http://dx.doi.org/10.1109/CVPR.2019.00030.

Zhao, C., Ge, Y., Zhu, F., Zhao, R., Li, H., Salzmann, M., 2021. Progressive correspondence pruning by consensus learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6464–6473. http://dx.doi.org/10.1109/ICCV48922.2021.00640.