

MGNet: Learning Correspondences via Multiple Graphs

Luanyuan Dai¹, Xiaoyu Du¹, Hanwang Zhang², Jinhui Tang^{1*}

¹Nanjing University of Science and Technology, China

²Nanyang Technological University, Singapore

{dailuanyuan, duxy, jinhuitang}@njjust.edu.cn, hanwangzhang@ntu.edu.sg

Abstract

Learning correspondences aims to find correct correspondences (inliers) from the initial correspondence set with an uneven correspondence distribution and a low inlier rate, which can be regarded as graph data. Recent advances usually use graph neural networks (GNNs) to build a single type of graph or simply stack local graphs into the global one to complete the task. But they ignore the complementary relationship between different types of graphs, which can effectively capture potential relationships among sparse correspondences. To address this problem, we propose MGNet to effectively combine multiple complementary graphs. To obtain information integrating implicit and explicit local graphs, we construct local graphs from implicit and explicit aspects and combine them effectively, which is used to build a global graph. Moreover, we propose Graph Soft Degree Attention (GSDA) to make full use of all sparse correspondence information at once in the global graph, which can capture and amplify discriminative features. Extensive experiments demonstrate that MGNet outperforms state-of-the-art methods in different visual tasks. The code is provided in <https://github.com/DAILUANYUAN/MGNet-2024AAAI>.

Introduction

Finding high-quality pixel-wise correspondences is the precondition for many important computer vision and robotics tasks, *e.g.*, visual localization (Sattler et al. 2018), image stitching (Ma, Ma, and Li 2019), image registration (Ma et al. 2015; Liu et al. 2022), point cloud registration (Bai et al. 2021; Qin et al. 2022; Wang et al. 2023), Simultaneous Location and Mapping (SLAM) (Mur-Artal, Montiel, and Tardos 2015), Structure from Motion (SfM) (Schonberger and Frahm 2016), etc. A standard pipeline depends on off-the-shelf detector-descriptors (Lowe 2004; DeTone, Malisiewicz, and Rabinovich 2018) to obtain putative correspondences, which have excessive incorrect correspondences (*i.e.*, outliers) due to the challenging cross-image variations, such as rotations, illumination changes and viewpoint changes.

Hence, outlier rejection is an essential step to preserve correct correspondences as well as reject false ones. Initial

*Corresponding author.

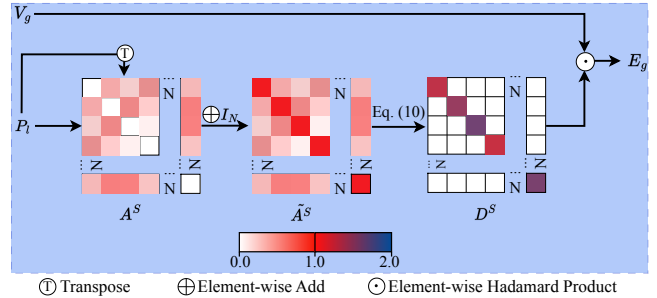


Figure 1: Graph Soft Degree Attention, in which A^S , \tilde{A}^S and D^S represent Soft Adjacent Matrix, the final Soft Adjacent Matrix and Soft Degree Matrix, respectively. Combining with the like-probability value (white to red and then to blue is from 0 to 1 and then to 2), it can prove that Soft Degree Matrix D^S can capture and amplify discriminative features.

correspondences are spread unevenly over an image pair, due to densely detected keypoints in textured areas but almost no keypoints in textureless areas. Hence, some networks (Zhang et al. 2019; Liu et al. 2021; Zhao et al. 2021; Dai et al. 2022; Li, Zhang, and Ma 2023) view sparse correspondences as graph data, in which there is no order and unified structure. OA-Net (Zhang et al. 2019), U-Match (Li, Zhang, and Ma 2023) and MS²DG-Net (Dai et al. 2022) only construct graphs in the local region without considering the global, where the first two implicitly construct local graphs and the other adopts an explicit approach. CL-Net (Zhao et al. 2021) simply stacks explicit local graphs into the global one, which is coped with a plain spectral graph convolutional layer (GCN) (Kipf and Welling 2016). At the same time, LMC-Net (Liu et al. 2021) only builds global graph Laplacian based on standard Laplacian matrix and decomposes it to solve the proposed formulation. They fail to consider potential relationships among different types of graphs and how to effectively use Laplacian matrix on graph data.

These networks have made certain progress in handling sparse correspondences, but there are still some problems. Firstly, no one uses GNNs to construct graphs from implicit and explicit perspectives at the same time, and explore their relationships and complementary advantages. Secondly, the ability of the plain spectral graph convolutional layer (GCN)

(Zhao et al. 2021; Kipf and Welling 2016) is not strong enough to capture discriminative feature in the global graph. That is to say, mainstream methods do not make full use of GNNs on sparse correspondences. Therefore, we propose a network, named MGNet, which effectively combines multiple graphs, to handle these sparse correspondences. Firstly, we build local graphs through implicit and explicit perspectives at the same time by GNNs, and explore potential relationships between them. Then, we propose Graph Soft Degree Attention (GSDA) to obtain and amplify discriminative features in the global graph. As shown in Figure 1, Soft Adjacent Matrix A^S does not consider its own information, and the final Soft Adjacent Matrix \hat{A}^S pays little attention to relationships between the selected sparse correspondence and others. In Soft Degree Matrix D^S , inspired by Laplace matrix, each selected correspondence fuses relationships between itself and all other correspondences. Hence, GSDA can capture and amplify discriminative features, as shown in Figure 1.

Our contribution is threefold. Firstly, implicit and explicit graphs are constructed at the same time by GNNs, and potential relationships between them have been discussed at length. After that, motivated by Laplacian matrix, Graph Soft Degree Attention (GSDA) is proposed and applied to effectively handle global information at once in the global graph, which can capture and amplify discriminative features. Finally, the proposed MGNet obtains state-of-the-art results on camera pose estimation, homography estimation, and visual localization with a relatively small number of parameters.

Related Work

Outlier Rejection

Traditional RANSAC (Fischler and Bolles 1981) and its variants (Torr and Zisserman 1998; Chum, Werner, and Matas 2005; Barath and Matas 2018; Barath, Matas, and Nuskova 2019; Barath et al. 2020) capture correct correspondences via the largest subset, so they may conform to specific scenarios. Thus, with the increasing of general dataset scale and outlier ratio, nearly all of them no longer work. Hence, using deep learning-based networks to handle irregular and unordered characteristics among sparse correspondences has emerged. First, CNe (Moo Yi et al. 2018) and DFE (Ranftl and Koltun 2018) only take correspondence coordinates as input and achieve great success. After that, some networks introduce the thought of attention mechanism (Vaswani et al. 2017) to enhance network performance. ACNe (Sun et al. 2020) and LAGA-Net (Dai et al. 2021) exploit attention mechanisms from local and global perspectives, but use different approaches. ANA-Net (Ye et al. 2023) provides the idea of second-order attention and proves its existence. To search more reliable correspondences, LFLN-Net (Wang et al. 2020) and NM-Net (Zhao et al. 2019) redefine neighborhood from different aspects. Next, LMC-Net (Liu et al. 2021) utilizes consistency constraint to remove outliers. CL-Net (Zhao et al. 2021) introduces a pruning operation to obtain inlier identification.

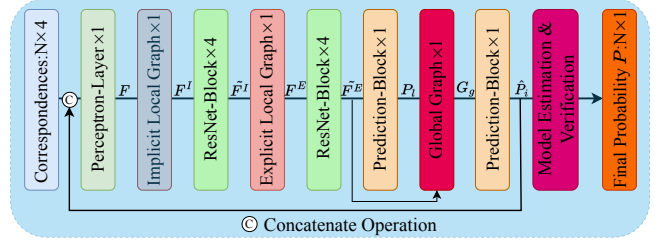


Figure 2: Network architecture of MGNet. The input is a putative correspondence set C , and the output is the final probability set P . $i = 1, 2$.

Graph Neural Network in Correspondences

Recently, Graph Neural Networks (GNNs) have emerged in correspondence learning, due to their powerful feature extraction abilities. To our knowledge, OA-Net (Zhang et al. 2019) is the first one introducing GNNs to remove outliers in sparse correspondences, which is inspired by DIFF-POOL (Ying et al. 2018) and improve DiffUnpool from plain to order-aware by a soft assignment manner. In LMC-Net (Liu et al. 2021), Liu et al. introduce graph Laplacian to decompose a new proposed formulation of motion coherence among sparse correspondences. In CL-Net (Zhao et al. 2021), Zhao et al. rely on dynamic graphs to obtain local and global consensus scores to progressively choose correct correspondences, where an annular convolutional operation is proposed to fuse local features. In MS²DG-Net (Dai et al. 2022), Dai et al. combine dynamic graphs and attention mechanism to capture local topology through similar sparse semantics information in each image pair. U-Match (Li, Zhang, and Ma 2023) combine a U-shaped network and GNNs, which can better utilize hierarchical graph information, to increase network ability to capture features.

Proposed Method

Problem Formulation

We use local features (SIFT (Lowe 2004), SuperPoint (DeTone, Malisiewicz, and Rabinovich 2018), etc.) followed by a NN matcher to build an initial correspondence set C .

$$C = \{c_1; c_2; \dots; c_N\} \in \mathbb{R}^{N \times 4} \quad (1)$$

where $c_i = (x_i, y_i, u_i, v_i)$ is a correspondence between two keypoints (x_i, y_i) and (u_i, v_i) , which are normalized under camera intrinsics. The initial correspondence set is polluted by excessive outliers, which can bring negative impact on downstream tasks.

Hence, we propose MGNet to reject outliers. Motivated by (Fischler and Bolles 1981; Zhao et al. 2021), we use a verification framework, but without the pruning operation (Zhao et al. 2021). That is because the pruning operation may reduce data abundance, and we prove it in Table 9. As shown in Figure 2, we iteratively use our main network twice to obtain the final probability set $P = \{p_1; p_2; \dots; p_N\}$ with $p_i \in [0, 1)$, in which elements present probabilities of the whole correspondences as inliers. From the first iteration, we can obtain the first estimated inlier probability set \hat{P}_1 .

After that, we use \hat{P}_1 and C to obtain \hat{P}_2 through the second iteration. Next, we use a weighted eight-point algorithm (Moo Yi et al. 2018) to estimate an essential matrix \hat{E} . Finally, a verification operation is used to test and verify the estimated essential matrix \hat{E} on the correspondence set C and obtain the final probability set P .

$$\begin{aligned} \hat{P}_1 &= f_{1\phi}(C), & \hat{P}_2 &= f_{2\psi}(\hat{P}_1, C) \\ \hat{E} &= g(\hat{P}_2, C), & P &= Ver(\hat{E}, E), \end{aligned} \quad (2)$$

where $f_{1\phi}(\cdot)$ and $f_{2\psi}(\cdot, \cdot)$ represent the first and second iterations with learnable parameter ϕ and ψ , respectively; \hat{P}_1 and \hat{P}_2 are the estimated inlier probability sets in the first and second iterations, respectively; $g(\cdot, \cdot)$ is the weighted eight-point algorithm; $Ver(\cdot, \cdot)$ is the verification operation.

Implicit and Explicit Local Graphs

Build Implicit Local Graph. Firstly, the input correspondence set C is encoded into a S -dimensional feature set $F = \{f_i\}_{i=1}^N \in \mathbb{R}^{S \times N \times 1}$ by a Perceptron Layer. After that, DiffPooling operation (Zhang et al. 2019) is used to coarsen F into a M -dimensional coarse-grained graph set $G^I = \{g_i^I\}_{i=1}^M \in \mathbb{R}^{S \times M \times 1}$ via an implicit way. Following OANet (Zhang et al. 2019), we choose OA Filtering operation to process coarse-grained graphs, so that global information among them can be attained. Finally, DiffUnpooling operation (Zhang et al. 2019) is used to restore data to its original size by a soft way. These can be written as:

$$G^I = DiffPooling(F) \quad (3)$$

$$F^I = DiffUnpooling(F, OA(G^I)) \quad (4)$$

where $DiffPooling(\cdot)$, $DiffUnpooling(\cdot, \cdot)$ and $OA(\cdot)$ represent DiffPooling, DiffUnpooling and OA Filtering operations, respectively; F^I is denoted as an implicit local graph feature set.

Build Explicit Local Graph. First, we use ResNet blocks to extract an implicit local graph feature vector \tilde{F}^I from F^I and use it to construct explicit local graphs. Second, k -nearest neighbors are chosen in \tilde{F}^I in feature space. After that, an edge set E_j^E is constructed by concatenating the selected correspondence feature map and the residual ones in \tilde{F}^I , just like (Dai et al. 2022; Zhao et al. 2021). Next, an explicit graph set $G^E = \{g_i^E\}_{i=1}^N \in \mathbb{R}^{S \times N \times k}$ is built on \tilde{F}^I with its k -nearest neighbors, to capture local topology among sparse correspondences. Finally, we choose maximum pooling and MLPs to aggregate information to obtain an explicit local graph feature set $F^E \in \mathbb{R}^{S \times N \times 1}$. The above operations can be recorded as:

$$E^E = [\tilde{F}_j^I || \tilde{F}_j^I - \tilde{F}_j^I], j = 1, 2, \dots, k \quad (5)$$

$$G^E = (V^E, E^E) \quad (6)$$

$$F^E = maxpooling(MLPs(G^E)) \quad (7)$$

where $[\cdot || \cdot]$ presents concatenation; \tilde{F}^I , \tilde{F}_j^I and $\tilde{F}^I - \tilde{F}_j^I$ are correspondence, neighborhood and residual feature sets, respectively; V^E indicates a \tilde{F}^I 's neighbor set; E^E denotes an explicit edge set.

Relationship between them. The coarsening process of implicit local graphs is automatically learned and sparse correspondences can be automatically grouped. That is, the information of nodes (sparse correspondences) can be learned, and the local structural information (relationship among sparse correspondences) can also be learned at the same time. In addition, constructing explicit local graphs on the implicit local graph feature vector \tilde{F}^I , allows us to intuitively obtain the more accurate local explicit graphs, as shown in Table 10. Comparing with Table 10 and the third, fourth and fifth lines in Table 11, we find that building implicit graphs first performs best. This may because mining the information among sparse correspondences from an implicit perspective first and then using the captured information to construct graphs from an explicit aspect can more fully explore the potential information and relationships among sparse correspondences.

Construct Global Graph

First, F^E is put into ResNet blocks and an explicit local graph feature vector \tilde{F}^E is obtained. We put \tilde{F}^E into a Prediction layer so that we can obtain a local probability set P_l . Next, \tilde{F}^E can be denoted as the global graph node set V^g . After that, we propose a novelty yet simple (without additional parameters) approach, named Graph Soft Degree Attention (GSDA) to construct the global edge set E^g , as shown in Figure 1. Specifically, we explore relationships in every two members in the local probability set P_l to produce Soft Adjacent Matrix $A^S \in \mathbb{R}^{N \times N}$ (see Theorem 1.), which cannot consider its own information, so a self-loop is created on top of it. The above operations can be written as:

$$A^S = softmax(P_l \cdot P_l^T) \quad (8)$$

$$\tilde{A}^S = A^S + I_N \quad (9)$$

where I_N is a $N \times N$ unit matrix; $\tilde{A}^S = \{\tilde{A}_{i,j}^S\}_{i,j=1}^N \in \mathbb{R}^{N \times N}$ is the final Soft Adjacent Matrix.

After that, we construct Soft Degree Matrix $D^S = \{D_{i,j}^S\}_{i,j=1}^N \in \mathbb{R}^{N \times N}$, in which one diagonal element is the sum of the corresponding rows on the final Soft Adjacent Matrix \tilde{A}^S , and the remains are zeros. (See Theorem 2&3.) One element on the Soft Degree Matrix D^S diagonal represents the sum of relationships between the selected correspondence and others in an image pair, which can make full use of all sparse correspondence information at once and at a long distance. Comparing with A^S , \tilde{A}^S and D^S visualizations in Figure 1, we find the proposed GSDA can capture and amplify discriminative features. That can be written as:

$$D_{i,j}^S = \begin{cases} \sum_{j=1}^N \tilde{A}_{i,j}^S, & i = j \\ 0, & else \end{cases} \quad (10)$$

Next, an element-wise Hadamard product is performed between Soft Degree Matrix D^S and global graph node set

V^g . Finally, a global graph is built by integrating implicit and explicit local graph information. These are defined as:

$$E^g = D^S \odot V^g \quad (11)$$

$$G^g = (V^g, E^g) \quad (12)$$

where \odot is the element-wise Hadamard product.

Similar to the local probability set, the global probability set P_g ($\hat{P}_i, i = 1, 2$) is defined by encoding the aggregated features by a ResNet block and a Prediction layer.

Related Theorem

Theorem 1. Adjacency Matrix $A \in \mathbb{R}^{N \times N}$ represents connections between any two nodes in graph data. If there is a connection between nodes v_i and v_j , an edge (v_i, v_j) will form and the corresponding element of Adjacency Matrix $A_{ij} = 1$, otherwise $A_{ij} = 0$. In addition, the diagonal element of Adjacency Matrix A is usually set to 0.

Theorem 2. Degree of a node refers to the total number of edges connected to it. $d(v)$ is usually used to present degree of a node.

Theorem 3. Degree Matrix $D = \{d_{i,j}\}_{i,j=1}^N$ of graph G is an $N \times N$ diagonal matrix, and an element on the diagonal is degree of the corresponding node, represented as:

$$d_{i,j} = \begin{cases} d(v_i), & i = j \\ 0, & else \end{cases} \quad (13)$$

Loss Function

Following OA-Net++ (Zhang et al. 2019) and CL-Net (Zhao et al. 2021), we choose a hybrid loss function:

$$L = L_c + \beta L_e(E, \hat{E}) \quad (14)$$

where L_c is defined as a binary classification loss with a proposed adaptive temperature, provided by CL-Net (Zhao et al. 2021); the later is a geometric loss between the ground truth E and an predicted estimated model \hat{E} ; β is a weighting factor to balance both of them.

Implementation Details

Network input is $N \times 4$ initial correspondences by SIFT or SuperPoint, and typically N is up to 2000. Cluster number m , neighbor number k and channel dimension S are 100, 24 and 128. Batchsize and β in Equation 14 are set to 32 and 0.5, respectively. Adam (Paszke et al. 2017) optimizer is used with a learning rate of 10^{-3} and we choose a warmup strategy. Clearly, a linearly growing rate is used for the first $10k$ iterations, after that the learning rate begins to decrease and reduce for every $20k$ iterations with a factor of 0.4. Experiments are performed on NVIDIA GTX 3090 GPUs.

Experiments

Evaluation Protocols

Main Datasets. Yahoo’s YFCC100M (Thomee et al. 2016) and SUN3D (Xiao, Owens, and Torralba 2013)

Method	Outdoor(%)		Indoor(%)	
	Known	Unknown	Known	Unknown
RANSAC	5.81	9.07	4.52	2.84
Point-Net++	10.49	16.48	10.58	8.10
DFE	19.13	30.27	14.05	12.06
CNe	13.81	23.95	11.55	9.30
OA-Net++	32.57	38.95	20.86	16.18
ACNe	29.17	33.06	18.86	14.12
SuperGlue	35.00	48.12	22.50	17.11
LMC-Net	33.73	47.50	19.92	16.82
CL-Net	39.16	53.10	20.35	17.03
MS ² DG-Net	38.36	49.13	22.20	17.84
U-Match	46.78	60.53	24.98	21.38
MGNet	51.43	64.63	25.96	21.27

Table 1: Evaluation on outdoor and indoor datasets with SIFT for camera pose estimation. The mAP5°(%) is reported and best result in each column is bold.

datasets are chosen as outdoor and indoor scenes, respectively. Following OA-Net++ (Zhang et al. 2019), 68 sequences are selected as training sequences and the rest 4 sequences are regarded as unknown scenes in outdoor scenes, and 239 sequences are chosen as training sequences, and the rest 15 sequences are unknown scenes in indoor scenes. Incidentally, we divide training sequences into three parts, consisting of training (60%), validation(20%) and testing (20%), and the last one is used as known scenes.

Main Evaluation Metrics. The error metrics can be defined by angular differences between calculated rotation/translation vectors (recovered from the essential matrix) and the ground truth. mAP5° and mAP20° are selected as the default metrics in the camera pose estimation task.

Main Baselines

We choose a traditional method (RANSAC (Fischler and Bolles 1981)) and ten learning-based networks (Point-Net++ (Qi et al. 2017), DFE (Ranftl and Koltun 2018), CNe (Moo Yi et al. 2018), OA-Net++ (Zhang et al. 2019), ACNe (Sun et al. 2020), SuperGlue (Sarlin et al. 2020), LMC-Net (Liu et al. 2021), CL-Net (Zhao et al. 2021), MS²DG-Net (Dai et al. 2022) and U-Match (Li, Zhang, and Ma 2023)) as main baselines. The official SuperGlue model is directly used to test.

Camera Pose Estimation

Camera pose estimation, referring to utilize the identified inliers to accurately excavate the relative position relationship (rotation and translation) between different camera views, is an important foundation for many computer vision tasks.

Camera Pose Estimation Results with SIFT. As shown in Table 1, we present the quantitative results of the proposed MGNet and main baselines on indoor and outdoor scenes with SIFT. Clearly, learning-based networks generally perform much better than traditional RANSAC. Besides, the proposed MGNet achieves the optimal value on all evaluation indicators. Our MGNet has increased 5.32% on mAP5°

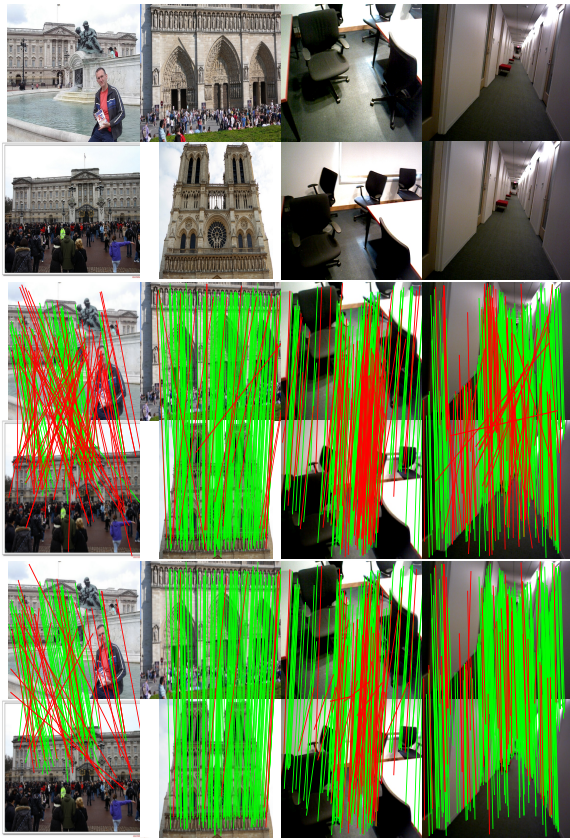


Figure 3: Partial typical visualization results on YFCC100M and SUN3D datasets with SIFT. From top to bottom: input image pairs, results of CLNet and our MGNet. The green lines describe inliers, the red lines otherwise.

than the second best network (U-Match) in unknown outdoor scenes. Comparing to CL-Net, our MGNet gains performance increments of 31.33%, 21.711%, 27.57% and 24.90% on the known outdoor, unknown outdoor, known indoor and unknown indoor scenes, respectively. And meanwhile, partial typical visualization results of CLNet and our MGNet in Figure 3 prove that the proposed MGNet can perform better under wide baseline, large viewpoint changes, illumination changes and textureless region scenes. This is because our MGNet can effectively combine multiple graphs to capture and amplify discriminative features.

Camera Pose Estimation Results with SuperPoint. In addition, we choose a popular learning-based feature extractor, named SuperPoint (DeTone, Malisiewicz, and Rabinovich 2018), to instead of SIFT to build putative correspondences. From Table 2, we can find that our MGNet still achieves the best results in all situations. Comparing with Table 1 and Table 2, there is a phenomenon that almost all methods (except for RANSAC and Point-Net++) perform better on the datasets preprocessed with SIFT than with SuperPoint. Besides, for some performing poor methods (RANSAC and Point-Net++), more correct correspondences (SuperPoint) may have more advantageous. But, those net-

Method	Outdoor(%)		Indoor(%)	
	Known	Unknown	Known	Unknown
RANSAC	12.85	17.47	14.93	12.15
Point-Net++	11.87	17.95	11.40	9.38
DFE	18.79	29.13	13.35	12.04
CNe	12.18	24.25	12.63	10.68
OA-Net++	29.52	35.27	20.01	15.62
ACNe	26.72	32.98	18.35	13.82
CL-Net	29.35	38.99	15.89	14.03
MS ² DG-Net	30.40	37.38	20.28	16.08
U-Match	35.12	45.72	22.73	18.87
MGNet	41.53	49.37	24.58	20.65

Table 2: Evaluation on outdoor and indoor datasets with SuperPoint for camera pose estimation. mAP5° (%) is reported.

	YFCC100M(%)		PhotoTourism(%)	
	ORB	SP	SIFT	SP
CNe	7.40	14.78	20.17	5.89
OA-Net++	12.05	19.40	40.39	8.99
CL-Net	14.75	21.00	45.54	9.41
MS ² DG-Net	11.38	21.05	45.53	12.91
U-Match	16.70	28.38	54.43	11.48
MGNet	20.00	32.88	57.64	20.41

Table 3: Generalization ability test on YFCC100M and PhotoTourism with different feature extractors, including ORB, SuperPoint (SP), and SIFT. mAP5° (%) is reported.

works, performing good enough, combine with SIFT better. That is explained in (Dai et al. 2022), in which Dai et al. prove that although SuperPoint obtains more correct correspondences than SIFT, but its average logit value is much lower.

Generalization Ability Test. To evaluate the generalization ability of networks, we compare our MGNet and part of main baselines in different datasets with different feature extractors. Clearly, we introduce PhotoTourism (Jin et al. 2021) and ORB (Rublee et al. 2011) in the work, in which the former is a challenging photo-tourism dataset and the later is a fast yet accurate detector-descriptor method to be used as a preprocessing technique. We train all models on YFCC100M with SIFT and directly test them on different datasets with different extractors. As summarized in Table 3, MGNet performs best in all setting, because it can effectively combine multiple different types of graphs to extract potential relationships among sparse correspondences. This can prove that MGNet has strong robustness and generalization abilities.

Homography Estimation

The purpose of homography estimation is to find a linear image-to-image mapping in the homogeneous space, which is the basis for many subsequent computer vision tasks. We compare the proposed MGNet and part of main baselines on HPatches benchmark (Balntas et al. 2017) with Direct Linear Transform (DLT). HPatches benchmark has 696 images and 116 scenes, each of which is composed of 1 reference image and 5 query images. That is, there are 580 image

Method	HPatches(%)		
	ACC.3px	ACC.5px	ACC.10px
CNe	38.97	51.55	65.34
OA-Net++	39.83	52.76	62.93
CL-Net	43.10	55.69	68.10
MS ² DG-Net	41.21	50.17	62.59
U-Match	48.90	59.41	70.83
MGNet	52.08	61.53	71.23

Table 4: Evaluation homography estimation on HPatches.

Method	Day	Night
	(0.25m, 2°)/(0.5m, 5°)/(1.0m, 10°)	(1.0m, 10°)
CNe	81.3/91.4/95.9	68.4/78.6/87.8
OA-Net++	82.3/91.9/96.5	71.4/79.6/90.8
CL-Net	83.3/92.4/ 97.0	71.4/80.6/ 93.9
MS ² DG-Net	82.8/92.1/96.8	70.4/82.7/ 93.9
U-Match	85.3 /92.6/96.8	72.4/82.7/90.8
MGNet	85.3 / 92.7 / 97.0	72.6 / 82.9 / 93.9

Table 5: Evaluation visual localization on Aachen Day-Night.

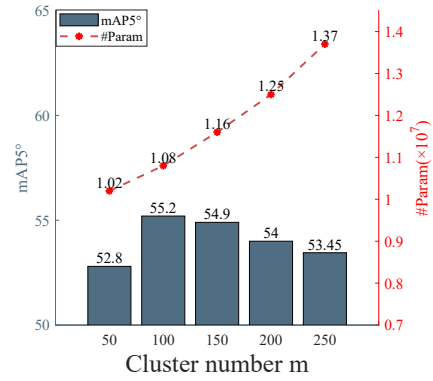
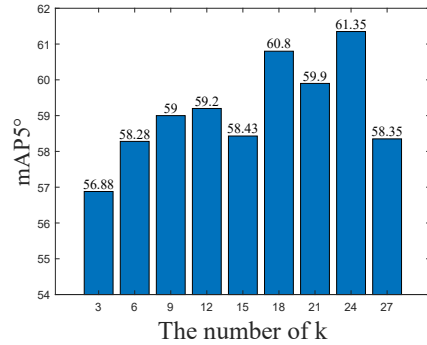
pairs in HPatches benchmark, in which some are collected in viewpoint changes and others have different illumination. In our work, each image pair is detected up to 4000 keypoints with SIFT followed by a NN matcher. Following (DeTone, Malisiewicz, and Rabinovich 2018), we choose homography error to evaluate them and present results that their average error is below 3/5/10 pixels. Table 4 shows that MGNet performs best at all thresholds, especially obtains an absolute 3.18 percentage point increase at the lowest threshold.

Visual Localization

Visual localization is intended to estimate the 6-degree of freedom (DOF) camera pose of a given image relative to its 3D scene model, which is a fundamental problem in many computer vision and robotic tasks. Specifically, we integrate our MGNet and other comparative networks into the official HLoc (Sarlin et al. 2019). Aachen Day-Night benchmark (Sattler et al. 2018) is chosen as a tested dataset, where 922 query images (824 daytime and 98 nighttime) are captured by mobile phones and 4328 reference ones are from a European ancient town. We extract up to 4096 feature points with SIFT on each image, followed by a NN matcher. After that, a SfM model is triangulated from day-time images with known poses, and registers night-time query images with 2D-2D matches obtained from correspondence learning networks and COLMAP (Schonberger and Frahm 2016). Following HLoc (Sarlin et al. 2019), the percentage of correctly localized queries at specific distances and orientation thresholds is regarded as the evaluation matrix. Results in Table 5 shows that our MGNet performs best in all situations and demonstrates that MGNet is suitable for visual localization.

Ablation Studies

In this section, ablation studies about implicit graphs, explicit graphs, global graphs, the relationship among them,

Figure 4: Relationship between mAP5° (%) and network parameter number with different cluster number m .Figure 5: Parametric analysis of k in the explicit local graph.

the pruning operation and the verification framework on the unknown outdoor scene with SIFT are reported.

How to choose m ? Cluster number m , determining the coarsening degree of implicit graphs, is pretty important. As shown in Figure 4, in which coordinate axes on the left and right represent mAP5° under unknown YFCC100M scenes and the parameter number, respectively. And we can find that with cluster number m increasing and after $m = 100$, the network performance gradually decreases while the parameter quantity still rises. Hence, we choose cluster number $m = 100$ to complete subsequent experiments.

How to choose k ? The neighbor number k , determining how much information in each explicit local graph, is of vital importance. As shown in Figure 5, as k increases from 3 to 27, the network performance first increases and then decreases. When $k = 24$ is the turning point and also the best time for model performance, so we choose $k = 24$ to construct explicit local graphs.

How to aggregate information in explicit graphs? We choose three ways (average pooling, maximum pooling and annular convolution (Zhao et al. 2021)) to fuse information. And Table 6 denotes that maximum pooling method performs best on mAP5° (%) and mAP20° (%), which is simple yet effective. It is worth mentioning that average pooling and maximum pooling have the same number of parameters, but

	mAP5°	mAP20°	Size(MB)
Annular Conv	61.00	81.13	1.54
Avg-pooling&MLPs	58.35	79.62	1.18
Max-pooling&MLPs	61.35	82.78	1.18

Table 6: Quantitative comparisons of different aggregation methods in explicit local graphs.

	plain GCN	GAA	GAIA	GSDA
mAP5°	63.08	63.60	64.00	64.63
mAP20°	82.72	83.05	83.43	83.76

Table 7: Evaluate the plain GCN (Zhao et al. 2021), GAA, GAIA and GSDA to construct global edge.

its results are worse. Annular convolution (Zhao et al. 2021) not only has more parameters than maximum pooling, but also performs poorer. Hence, we select maximum pooling method.

How to effectively construct global graph edge? We choose four ways to construct global graph edge. The first one is the plain GCN (Zhao et al. 2021), which is used in recent works. Combining with Laplacian matrix knowledge, we propose the remaining three ways. Specifically, the second one is based on the Soft Adjacent Matrix, named GAA, and next one is to add a self-loop (its own information) on the second, called GAIA. The last one is GSDA, which is based on Soft Degree Matrix. As shown in Table 7, GSDA performs best and we choose it.

Is additional local information helpful? As shown in Table 8, we can find that only using the global probability to verify is enough, which not only minimizes the parameter number but also performs best. Besides, we also find the more local information (implicit or explicit) is added, the worse the network performs. It is probably because local information has already been integrated into the global, and reusing it can cause overfitting to degrade performance.

Is the pruning operation helpful? As summarized in Table 9, the model performance deteriorates with the pruning ratio increasing. Interestingly, when the pruning ratio is 0.25, it performs worse than without the verification framework. That is probably because if a model is not complex enough (without a large number of parameters), the pruning operation will reduce data abundance (contrary to data augmentation), so that the model performance reduces.

	mAP5°	mAP20°	Size(MB)
+ P_{li}	61.83	82.04	1.32
+ P_{le}	63.15	83.09	1.32
+ $P_{li}+P_{le}$	61.23	82.52	1.42
MGNet	64.63	83.76	1.31

Table 8: Compare effect of adding different local information. MGNet represents only utilize the global probability to verify. "+" represents add other information on MGNet. P_{li} and P_{le} represents implicit and explicit local probabilities.

	mAP5°	mAP20°	Size(MB)
<i>wo</i> verification	61.17	81.64	1.60
<i>w</i> verification, $pr=0.75$	50.80	73.31	1.31
<i>w</i> verification, $pr=0.5$	61.68	81.69	1.31
<i>w</i> verification, $pr=0.25$	63.05	82.34	1.31
MGNet	64.63	83.76	1.31

Table 9: Parameter analysis of the pruning operation. pr presents the pruning ratio. *w/wo* represent with/without.

	mAP5°	mAP20°	Size(MB)
local explicit graph first	59.95	81.92	1.18
local implicit graph first	61.35	82.71	1.18

Table 10: Analysis about the order of building local graphs.

Relationship among them. Comparing with Table 10 and the third, fourth and fifth lines in Table 11, we find that building implicit graphs first is much better than building explicit graphs first, building implicit graphs twice and building explicit graphs twice. Observing the sixth, seventh and last lines in Table 11, we find that the combination we have been proposed (MGNet) performs much better than others, which can effectively combine local and global information.

Is the verification framework useful? From the eighth and last lines in Table 11, we find the verification framework is very useful and can improve 3.46 mAP5°(%) on the model without it. This is probably because using the verification framework can more fully explore potential relationships among sparse correspondences.

Conclusion

This work proposes MGNet for learning correspondences. There are two main improvements: 1) We construct local graphs from implicit and explicit aspects at the same time, and explore their potential relationship. 2) Motivated by Laplacian matrix, Graph Soft Degree Attention (GSDA) is proposed to capture and amplify discriminative features based on the whole sparse correspondence information in the global graph. Experiments on different tasks and datasets prove that MGNet has a great performance improvement compared to other state-of-the-art methods.

Ver	IG	EG	GG	mAP5°	mAP20°
✓				44.9	70.81
✓	✓			55.2	77.24
✓	✓	✓		61.35	82.71
✓	✓✓			51.18	75.06
✓		✓✓		59.78	81.26
✓	✓✓✓			51.43	75.47
✓		✓✓✓		63.95	83.29
✓	✓	✓	✓	61.17	81.64
✓	✓	✓	✓	64.63	83.76

Table 11: Ablation studies about network compositions. Ver, IG, EG and GG represent the verification framework, implicit local graph, explicit local graph and global graph.

Acknowledgments

This work was supported by National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-RP-2021-022), the National Natural Science Foundation of China (62172226), and the 2021 Jiangsu Shuangchuang (Mass Innovation and Entrepreneurship) Talent Program (JSSCBS20210200).

References

- Bai, X.; Luo, Z.; Zhou, L.; Chen, H.; Li, L.; Hu, Z.; Fu, H.; and Tai, C.-L. 2021. Pointdsc: Robust point cloud registration using deep spatial consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15859–15869.
- Balntas, V.; Lenc, K.; Vedaldi, A.; and Mikolajczyk, K. 2017. HPatches: A benchmark and evaluation of hand-crafted and learned local descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5173–5182.
- Barath, D.; and Matas, J. 2018. Graph-cut RANSAC. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6733–6741.
- Barath, D.; Matas, J.; and Nuskova, J. 2019. MAGSAC: marginalizing sample consensus. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10197–10205.
- Barath, D.; Nuskova, J.; Ivashechkin, M.; and Matas, J. 2020. MAGSAC++, a fast, reliable and accurate robust estimator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1304–1312.
- Chum, O.; Werner, T.; and Matas, J. 2005. Two-view geometry estimation unaffected by a dominant plane. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, 772–779. IEEE.
- Dai, L.; Liu, X.; Liu, Y.; Yang, C.; Wei, L.; Lin, Y.; and Chen, R. 2021. Enhancing Two-View Correspondence Learning By Local-Global Self-Attention. *Neurocomputing*.
- Dai, L.; Liu, Y.; Ma, J.; Wei, L.; Lai, T.; Yang, C.; and Chen, R. 2022. MS2DG-Net: Progressive Correspondence Learning via Multiple Sparse Semantics Dynamic Graph. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8973–8982.
- DeTone, D.; Malisiewicz, T.; and Rabinovich, A. 2018. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 224–236.
- Fischler, M. A.; and Bolles, R. C. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6): 381–395.
- Jin, Y.; Mishkin, D.; Mishchuk, A.; Matas, J.; Fua, P.; Yi, K. M.; and Trulls, E. 2021. Image matching across wide baselines: From paper to practice. *International Journal of Computer Vision*, 129(2): 517–547.
- Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.
- Li, Z.; Zhang, S.; and Ma, J. 2023. U-Match: Two-view Correspondence Learning with Hierarchy-aware Local Context Aggregation. In *International Joint Conference on Artificial Intelligence (IJCAI)*.
- Liu, Y.; Li, Y.; Dai, L.; Lai, T.; Yang, C.; Wei, L.; and Chen, R. 2022. Motion Consistency-Based Correspondence Growing for Remote Sensing Image Matching. *IEEE Geoscience and Remote Sensing Letters*, 19: 1–5.
- Liu, Y.; Liu, L.; Lin, C.; Dong, Z.; and Wang, W. 2021. Learnable Motion Coherence for Correspondence Pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3237–3246.
- Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2): 91–110.
- Ma, J.; Ma, Y.; and Li, C. 2019. Infrared and visible image fusion methods and applications: A survey. *Information Fusion*, 45: 153–178.
- Ma, J.; Zhou, H.; Zhao, J.; Gao, Y.; Jiang, J.; and Tian, J. 2015. Robust feature matching for remote sensing image registration via locally linear transforming. *IEEE Transactions on Geoscience and Remote Sensing*, 53(12): 6469–6481.
- Moo Yi, K.; Trulls, E.; Ono, Y.; Lepetit, V.; Salzmann, M.; and Fua, P. 2018. Learning to find good correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2666–2674.
- Mur-Artal, R.; Montiel, J. M. M.; and Tardos, J. D. 2015. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE transactions on robotics*, 31(5): 1147–1163.
- Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch.
- Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30: 5099–5108.
- Qin, Z.; Yu, H.; Wang, C.; Guo, Y.; Peng, Y.; and Xu, K. 2022. Geometric transformer for fast and robust point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11143–11152.
- Ranftl, R.; and Koltun, V. 2018. Deep fundamental matrix estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 284–299.
- Rublee, E.; Rabaud, V.; Konolige, K.; and Bradski, G. 2011. ORB: An efficient alternative to SIFT or SURF. In *2011 International conference on computer vision*, 2564–2571. Ieee.
- Sarlin, P.-E.; Cadena, C.; Siegwart, R.; and Dymczyk, M. 2019. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12716–12725.
- Sarlin, P.-E.; DeTone, D.; Malisiewicz, T.; and Rabinovich, A. 2020. SuperGlue: Learning Feature Matching With Graph Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

- Sattler, T.; Maddern, W.; Toft, C.; Torii, A.; Hammarstrand, L.; Stenborg, E.; Safari, D.; Okutomi, M.; Pollefeys, M.; Sivic, J.; et al. 2018. Benchmarking 6dof outdoor visual localization in changing conditions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8601–8610.
- Schonberger, J. L.; and Frahm, J.-M. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.
- Sun, W.; Jiang, W.; Trulls, E.; Tagliasacchi, A.; and 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*, 11286–11295.
- Thomee, B.; Shamma, D. A.; Friedland, G.; Elizalde, B.; Ni, K.; Poland, D.; Borth, D.; and Li, L.-J. 2016. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2): 64–73.
- Torr, P.; and Zisserman, A. 1998. Robust computation and parametrization of multiple view relations. In *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*, 727–732. IEEE.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.
- Wang, J.; Liu, X.; Dai, L.; Ma, J.; Wei, L.; Yang, C.; and Chen, R. 2023. PG-Net: Progressive Guidance Network via Robust Contextual Embedding for Efficient Point Cloud Registration. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–12.
- Wang, Y.; Mei, X.; Ma, Y.; Huang, J.; Fan, F.; and Ma, J. 2020. Learning to find reliable correspondences with local neighborhood consensus. *Neurocomputing*, 406: 150–158.
- Xiao, J.; Owens, A.; and 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*, 1625–1632.
- Ye, X.; Zhao, W.; Lu, H.; and Cao, Z. 2023. Learning Second-Order Attentive Context for Efficient Correspondence Pruning. *arXiv preprint arXiv:2303.15761*.
- Ying, Z.; You, J.; Morris, C.; Ren, X.; Hamilton, W.; and Leskovec, J. 2018. Hierarchical graph representation learning with differentiable pooling. *Advances in neural information processing systems*, 31.
- Zhang, J.; Sun, D.; Luo, Z.; Yao, A.; Zhou, L.; Shen, T.; Chen, Y.; Quan, L.; and Liao, H. 2019. Learning two-view correspondences and geometry using order-aware network. In *Proceedings of the IEEE International Conference on Computer Vision*, 5845–5854.
- Zhao, C.; Cao, Z.; Li, C.; Li, X.; and Yang, J. 2019. Nmnet: Mining reliable neighbors for robust feature correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 215–224.
- Zhao, C.; Ge, Y.; Zhu, F.; Zhao, R.; Li, H.; and Salzmann, M. 2021. Progressive Correspondence Pruning by Consensus Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6464–6473.