Pattern Recognition 159 (2025) 111120

Contents lists available at ScienceDirect

Pattern Recognition

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

TransMatch: Transformer-based correspondence pruning via local and global consensus

Yizhang Liu [a,b,1], Yanping Li [c,1], Shengjie Zhao [a,d,*]

[a] School of Software Engineering, Tongji University, Shanghai 201804, China
[b] College of Computer and Data Science, Fuzhou University, China
[c] Department of Computer Science and Technology, Tongji University, Shanghai 201804, China
[d] Engineering Research Center of Key Software Technologies for Smart City Perception and Planning, Ministry of Education, Shanghai, China

ARTICLE INFO

Keywords:
Correspondence pruning
Transformer
Local and global consensus
Camera pose estimation

ABSTRACT

Correspondence pruning aims to filter out false correspondences (a.k.a. outliers) from the initial feature correspondence set, which is pivotal to matching-based vision tasks, such as image registration. To solve this problem, most existing learning-based methods typically use a multilayer perceptron framework and several well-designed modules to capture local and global contexts. However, few studies have explored how local and global consensuses interact to form cohesive feature representations. This paper proposes a novel framework called TransMatch, which leverages the full power of Transformer structure to extract richer features and facilitate progressive local and global consensus learning. In addition to enhancing feature learning, Transformer is used as a powerful tool to connect the above two consensuses. Benefiting from Transformer, our TransMatch is surprisingly effective for differentiating correspondences. Experimental results on correspondence pruning and camera pose estimation demonstrate that the proposed TransMatch outperforms other state-of-the-art methods by a large margin. The code will be available at https://github.com/lyz8023lyp/TransMatch/.

1. Introduction

Establishing accurate point-wise correspondences (i.e., inliers) are fundamental requirements for many high-level computer vision tasks, including image stitching [1], image registration [2–4], image fusion [5,6], 3D reconstruction [7], Simultaneous Location and Mapping (SLAM) [8], etc. Initial correspondences can be established by using traditional feature extraction methods (e.g., SIFT [9]) or learning-based ones (e.g., SuperPoint [10]). However, due to the ambiguity of feature points (caused by repetitive structures, illumination changes, wide baseline, etc.), the initial correspondence set inevitably comprises vast amounts of false correspondences (i.e., outliers), making correspondence pruning become a crucial step for downstream tasks [11,12].

Traditional methods have shown satisfactory results in certain scenarios, e.g., the correspondence set contains only a small proportion of outliers, or the transformation between two images is relatively simple. Nevertheless, either a high proportion of outliers or complex transformations would lead to poor performance. Fortunately, deep learning has been introduced into the correspondence pruning task, inspired by its success in various computer vision tasks [13,14]. Compared to traditional methods, learning-based networks have demonstrated stronger abilities to extract features and achieve better generalization results. CNe [15] pioneered a Multi-layer perception (MLP) architecture for correspondence pruning, independently processing each correspondence and predicting the probabilities of correspondences as inliers. Subsequent work [16–19] attempted to design various modules to better capture local and global contexts for each correspondence. Although these methods have achieved promising results, they typically suffer from the following limitations.

- The MLP architecture with shared weights for feature learning cannot model the global context well. Although some researchers have proposed augmenting the feature learning with self-attention mechanisms or their variants as auxiliary modules [18,20], the full Transformer architecture's potent ability to model long-range dependencies has not been fully explored.
- How to incorporate local and global consensuses into more cohesive feature representations of all correspondences, has not been studied yet. Now, the common practice is to independently

[*] Corresponding author at: School of Software Engineering, Tongji University, Shanghai 201804, China.
E-mail address: shengjiezhao@tongji.edu.cn (S. Zhao).
[1] Co-first authors, contribute equally to this work.

https://doi.org/10.1016/j.patcog.2024.111120
Received 8 September 2023; Received in revised form 20 August 2024; Accepted 24 October 2024
Available online 2 November 2024

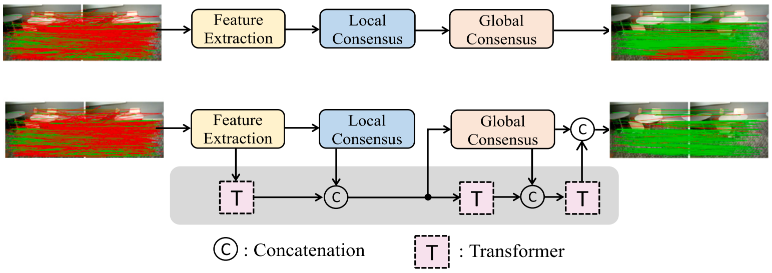

Fig. 1. Architecture comparison of the baseline model and our TransMatch. First row: the baseline model independently learns local and global consensuses at distinct layers, with simple cascading interaction between these two consensuses. Second row: our TransMatch uses a Transformer structure to model the relationship between any two correspondences and Transformer-based hierarchical aggregation is proposed to explore interactions between local and global consensuses and produce more cohesive feature representations.

learn local and global consensuses at distinct layers, with simple cascading interactions between them. This way will lead to suboptimal information utilization.

The recently proposed CLNet [21] has demonstrated remarkable superiority over the previous state-of-the-art methods. Specifically, it searched K nearest neighbors for each correspondence in the feature space and aggregated them to obtain the local context. Then, it used a graph connecting different local contexts to establish the global context. The convolutional operation on the graph enables it to capture the interdependence among nodes (i.e., local context). Additionally, the pruning strategy in CLNet for robust model estimation further contributes to its state-of-the-art performance. Nevertheless, it also suffers from the above limitations. While the global graph can provide the global context to some extent, it mainly explores the interdependence between different local contexts instead of relationships between any two correspondences. Furthermore, simple cascading interactions between local and global consensuses pose significant challenges to the network in learning cohesive feature representations. More specifically, there are significant dependencies between local and global consensuses, and simple cascading interactions fail to capture these complex interrelationships. This also leads to insufficient information fusion, making it difficult to fully leverage cross-scale information. As shown in Fig. 1, a high proportion of outliers in the initial correspondence set can lead to unreliable local consensus. Such interactions can adversely affect global consensus, resulting in poor matching performance.

To address the aforementioned problems, we build on state-of-the-art CLNet [21], and propose a simple yet surprisingly effective network termed as TransMatch. Specifically, inspired by the effectiveness of Transformer in modeling global dependencies, TransMatch utilizes a full Transformer structure to achieve two goals, i.e., richer feature extraction and progressive local and global consensus learning. In contrast to the baseline model exploring interdependence between different local contexts, TransMatch uses a Transformer structure to model relationships between any two correspondences, which is a more comprehensive and robust global context. In addition, a Transformer-based hierarchical aggregation module is proposed to explore interactions between local and global consensuses, resulting in more cohesive feature representations, which is crucial to transformation estimation tasks [22] (e.g., camera pose estimation). The comparison of the baseline model and our proposed TransMatch is shown in Fig. 1. The proposed TransMatch based on the Transformer structure is able to learn progressive local and global consensuses and combine them in a more comprehensive manner, and it outperforms the baseline model by a large margin (see experimental results). Our contributions can be summarized as follows.

- We present a novel correspondence pruning network, called TransMatch, to solve the problem that existing methods ignore the importance of interactions between local and global consensuses.
- TransMatch leverages the full power of the Transformer structure to extract richer features and facilitate progressive local and global consensus learning. Meanwhile, the proposed Transformer-based hierarchical aggregation can produce more cohesive feature representations.
- TransMatch outperforms the existing state-of-the-art methods on correspondence pruning and camera pose estimation tasks. For example, TransMatch shows a significant improvement over the second-best method by a large margin of 11.98% and 13.19% mAP5° on the unknown and known outdoor scene of YFCC-100M dataset without RANSAC, respectively.

2. Related work

Traditional Correspondence Pruning. RANSAC [23] and its variants [24–26] utilized a generation-verification framework for correspondence pruning. They iteratively estimated parametric models to fit correct correspondences (a.k.a., inliers), and pruned the correspondences that did not conform to the obtained model. These methods have proven to be effective, particularly when the initial inlier ratio is high. On the contrary, when the initial inlier ratio is extremely low, reliable results cannot be obtained within a limited number of iterations. More recently, LPM [27] and its variants [28–30] and GMS [31] have formulated motion consistency into a mathematical problem, resulting in a significant acceleration of the correspondence pruning process. Nonetheless, their performance suffers greatly in scenarios with a low inlier ratio.

Learning-based Correspondence Pruning. Drawing inspiration from the remarkable advancements of MLP architecture in point cloud processing [32], CNe [15] proposed a pioneering network that utilized an MLP-based permutation invariant network to formulate correspondence pruning as a binary classification task. However, this network lacked designs to effectively capture the local and global contexts of each correspondence. As a result, subsequent studies have focused on developing modules to extract context information. For instance, OANet [16] presented differentiable pooling and unpooling layers to capture the local context, coupled with an order-aware filtering block for the global context. ACNe [33] designed an attentive context normalization as an auxiliary module to enhance context information learning. NMNet [34] incorporated feature descriptors into the network learning, which aimed to search for consistent neighbors globally. Moreover, recent studies like CLNet [21], LMCNet [35], and MS^{2}DG-Net [18] have all exploited the motion consistency of inliers to accomplish correspondence pruning. Graph Laplacian [36] has also been proven effective in mining relationships between nodes and extracting features of graph data, which are unordered and unstructured. MSA-Net [37] devised a multi-scale attention network to establish reliable correspondences. CSDA-Net [38] selectively aggregated information from spatial and channel dimensions by designing a channel-spatial attention module. PGFNet [17] proposed a preference-guided filtering network that learned the preference scores of correspondences for guiding the subsequent filtering process to alleviate the negative effects of outliers. JRA-Net [39] proposed a joint representation attention network to

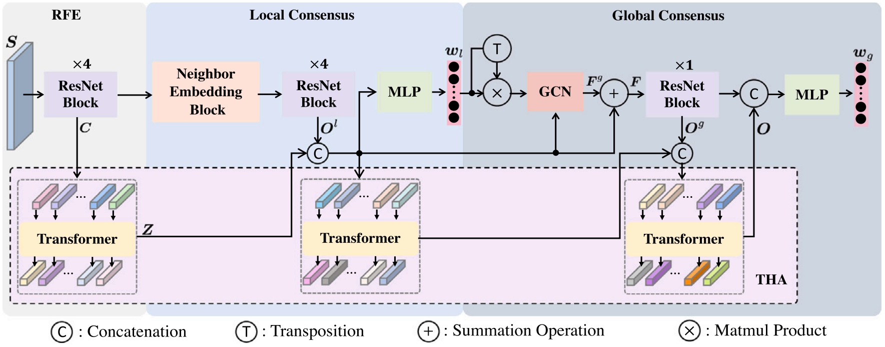

Fig. 2. The architecture of TransMatch. It consists of richer feature extraction (RFE), local consensus (LC), global consensus (GC), and Transformer-based hierarchical aggregation (THA).

learn the global context of different scales. U-Match [40] proposes an attention-based graph neural network to establish implicit local context at multiple levels. NCMNet [41] proposes to build three types of neighbors for correspondence and learn the spatial, feature, and graph-space consistency. While the above-mentioned methods have made significant efforts to either introduce well-designed modules for local and global contexts, or use auxiliary information (e.g., feature descriptors) for better generalization. However, they often ignore the importance of interactions between the local and the global consensus, resulting in inadequate information utilization.

Transformer for Vision. Transformer [42] comprises three main modules, i.e., input embedding, attention mechanism, and channel MLP, which was initially developed as a more advanced alternative to Recurrent Neural Network (RNN) for processing sequential data in Natural Language Processing (NLP). Due to its remarkable ability to model long-range dependencies, it has been applied to many vision tasks [14,43]. Transformer has also been introduced into the point cloud processing since it is inherently permutation invariant. [13] proposed a versatile framework based on Transformer, called Point Cloud Transformer (PCT), which presented an upgraded module over self-attention, called offset-attention. Meanwhile, a neighbor embedding module was proposed to enhance the locality-awareness. Since initial correspondences are unordered and unstructured, and Transformer, with permutation invariant property, is well suited for achieving correspondence pruning.

It is important to highlight that our proposed method uses a Transformer architecture that sets it apart from others. We leverage the full Transformer architecture to effectively build relationships between any two correspondences, while most previous methods have employed the self-attention mechanism as an auxiliary module. Specifically, in SuperGlue [20], self-attention refines keypoint features within each image, while cross-attention aligns and matches keypoints between the images. ACNe [33] incorporates the attention mechanism into the normalization process, i.e., the attention weights are then used to normalize features in a context-aware manner. MS^{2}DG-Net [18] adopts the self-attention to aggregate dynamic graph along the edge set. Our Transmatch only follows the original plain Transformer design, without introducing any additional or novel attention blocks, but can still yield promising results compared with other state-of-the-art methods.

3. Method

In this section, we provide a detailed overview of the proposed TransMatch for correspondence pruning. As shown in Fig. 2, building on the state-of-the-art CLNet [21], we introduce the design philosophy behind TransMatch and its novel components, which not only generate rich local and global consensuses, but also fully integrate them to form cohesive feature representations.

3.1. Problem formulation

Given an image pair $(\boldsymbol{I}, \boldsymbol{I}')$, a feature detection and description method (e.g., SIFT [9] or SuperPoint [10]) is firstly adopted to obtain descriptors of keypoints. Using a specific matching strategy (e.g., nearest-neighbor matching), we establish an initial correspondence set $\boldsymbol{S} = \left\{s_1, s_2, \dots, s_N\right\} \in \mathbb{R}^{N\times 4}$ based on the similarities of descriptors. Each correspondence $s_i = (x_i, y_i, x'_i, y'_i)$ consists of two matched keypoint coordinates (x_i, y_i), and (x'_i, y'_i). However, as mentioned above, the initial correspondence set $\boldsymbol{S}$ typically contains a large ratio of outliers. Therefore, our focus is on identifying inliers while rejecting outliers.

To this end, we propose a surprisingly effective method called TransMatch. Specifically, the initial correspondence set $\boldsymbol{S}$ as the input is passed through a feature encoding process, mapping the input from a low-dimensional space (4D) to a high-dimensional feature space (128D). We then propose a richer feature extraction block using a Transformer structure, followed by progressive local and global consensus learning with explicit interactions between them. The whole training process is iterated twice. Under the supervision of the loss function, the network will produce local and global consensus scores in the first iteration, which combine with $\boldsymbol{S}$ and are used as an additional input of the network in the second iteration to obtain the inlier probability of each correspondence. Then, a parametric model $\hat{\boldsymbol{E}}$ can be estimated based on the predicted inlier probability and $\boldsymbol{S}$. We perform the verification on $\boldsymbol{S}$ to get the final inlier probability of each correspondence. The above process can be formulated as

$$\begin{aligned} &\hat{\boldsymbol{w}} = f_\phi(\boldsymbol{S}), \quad \hat{\boldsymbol{w}}_p = f_\psi(\hat{\boldsymbol{w}}, \boldsymbol{S}) \\ &\hat{\boldsymbol{E}} = g(\hat{\boldsymbol{w}}_p, \boldsymbol{S}), \quad \boldsymbol{w} = h(\hat{\boldsymbol{E}}, \boldsymbol{S}), \end{aligned} \tag{1}$$

where $f_\phi(\cdot)$ and $f_\psi(\cdot)$ are neural networks with learnable parameter ϕ and ψ, respectively; $\hat{\boldsymbol{w}}$ denotes local and global consensus scores and $\hat{\boldsymbol{w}}_p$ indicates the estimated inlier probability of each correspondence in the second iteration; $g(\cdot)$ is the weighted eight-point algorithm [15] to estimate the essential matrix. Notably, the essential matrix $\hat{\boldsymbol{E}}$ encodes the rotation and translation between two camera views, and it can be expressed as $\hat{\boldsymbol{E}} = [\hat{\boldsymbol{t}}] \times \hat{\boldsymbol{R}}$, where $\hat{\boldsymbol{R}}$ is the rotation matrix, and $[\hat{\boldsymbol{t}}]\times$ is the skew-symmetric matrix of the translation vector $[\hat{\boldsymbol{t}}]$. $h(\cdot,\cdot)$ uses the estimated model $\hat{\boldsymbol{E}}$ to conduct the verification on $\boldsymbol{S}$, obtaining the final inlier probability of each correspondence.

3.2. TransMatch

Richer Feature Extraction. In previous studies [18,21,44], ResNet Block is adopted to accomplish feature extraction, which consists of two shared MLP layers and some normalization operations. Obviously, although weight-shared MLP can reduce the size of these models, its feature extraction ability is heavily constrained by a limited channel number. Considering that the feature extraction of each correspondence is fundamental to subsequent local and global consensus learning, we enhance the feature extraction by adding Transformer after ResNet

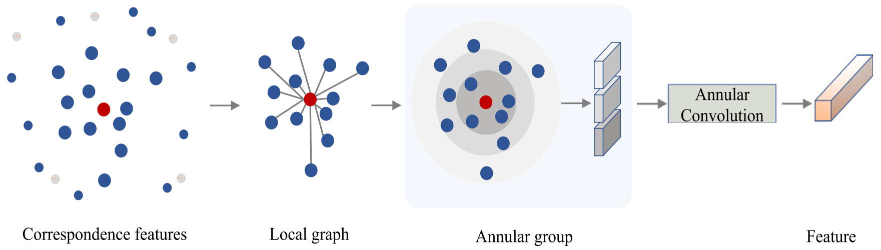

Fig. 3. Illustration of the Neighbor Embedding module.

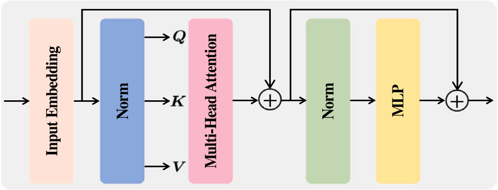

Fig. 4. The structure of the original plain Transformer.

Block due to its superior modeling ability for long-range dependencies, and Transformer structure is shown in Fig. 4. We denote the output feature map of ResNet Block as $\boldsymbol{C} = \{c_1, c_2, \ldots, c_N\} \in \mathbb{R}^{N\times D}$. The feature map $\boldsymbol{C}$ first passes through input embedding, which is a fully connected layer and can be expressed as

$$\hat{\boldsymbol{C}} = \boldsymbol{InputEmb}(\boldsymbol{C}), \tag{2}$$

where $\hat{\boldsymbol{C}} \in \mathbb{R}^{N\times M}$ denotes the embedding tokens with embedding dimension M. $\boldsymbol{InputEmb}(\cdot)$ is the input embedding operation. Since correspondences are unordered and unstructured, the positional embedding can be omitted. The embedded feature map $\hat{\boldsymbol{C}}$ is then passed to the normalization layer and the multi-head self-attention layer (MSA). For each head, three different linear projections are applied to $\hat{\boldsymbol{C}}$, generating query $\boldsymbol{Q} \in \mathbb{R}^{N\times d}$, key $\boldsymbol{K} \in \mathbb{R}^{N\times d}$, and value $\boldsymbol{V} \in \mathbb{R}^{N\times d}$, where $d = \frac{M}{m}$ and m is the number of heads. The attentive feature map $\boldsymbol{C}'$ can be obtained as follows

$$\boldsymbol{C}' = \boldsymbol{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}}\right)\boldsymbol{V}. \tag{3}$$

In this way, each correspondence will interact with each other in a global view. Then the residual connection between $\hat{\boldsymbol{C}}$ and the output of MSA is conducted and we denote the feature map obtained as $\hat{\boldsymbol{Z}}$. Then another normalization layer and channel MLP are applied to $\hat{\boldsymbol{Z}}$. Similarly, the residual connection between $\hat{\boldsymbol{Z}}$ and the output of channel MLP is conducted, which can be formulated as

$$\boldsymbol{Z} = \hat{\boldsymbol{Z}} + \boldsymbol{MLP}(\boldsymbol{Norm}(\hat{\boldsymbol{Z}})). \tag{4}$$

The outputs of ResNet Block and Transformer, namely, $\boldsymbol{C}$ and $\boldsymbol{Z}$, are high-level feature representations of the original correspondence set $\boldsymbol{S}$.

Local Consensus Learning. For each individual correspondence, it is hard to determine its correctness without any prior knowledge. Fortunately, based on the motion consistency theory, inliers share similar motions in a small region, while outliers are randomly distributed [45]. Thus, the neighborhood information with respect to correspondences helps to identify inliers. Constructing a local graph that consists of K nearest neighbors of each correspondence in feature space has been proven effective in capturing the local consensus. Thus, we propose a Neighbor Embedding block, as shown in Fig. 3. First, we define the local graph of c_i as:

$$\mathcal{G}_i^l = \left(\mathcal{V}_i^l, \mathcal{E}_i^l\right), \quad 1 \le i \le N, \tag{5}$$

where $\mathcal{V}_i^l = \{c_i^1, c_i^2, \ldots, c_i^K\}$ is the vertex set, consisting of K nearest neighbors of c_i, and $\mathcal{E}_i^l$ is the edge set containing connections between c_i and its neighbors in $\mathcal{V}_i^l$. The local context of each correspondence can be expressed as

$$e_i^j = \boldsymbol{Comb}(c_i, \Delta c_i^j), \quad 1 \le j \le K, \tag{6}$$

where $\boldsymbol{Comb}(\cdot,\cdot)$ means concatenating by channels and $\Delta c_i^j = c_i - c_i^j$. We further divide e_i^j into several annular groups according to their spatial relationship. An annular convolutional layer aggregates features of each correspondence in the annular groups [21]. We use a series of ResNet Blocks for further feature extraction to obtain features with local context information, denoted as $\boldsymbol{O}^l$, which only derives from $\boldsymbol{C}$ (the output of ResNet Block). However, $\boldsymbol{Z}$ is an upgrade over $\boldsymbol{C}$, which contains richer context information and has not been considered yet. Motivated by the principle of residual connections, we aggregate $\boldsymbol{Z}$ and $\boldsymbol{O}^l$ by channels, followed by an MLP layer, to obtain the local consensus score

$$\boldsymbol{w}_l = \boldsymbol{MLP}(\boldsymbol{Aggr}(\boldsymbol{Z}, \boldsymbol{O}^l)), \tag{7}$$

where $\boldsymbol{Aggr}(\cdot,\cdot)$ denotes the operation of concatenation and dimensionality reduction to the same as inputs.

Global Consensus Learning. The graph Laplacian has been proven effective in mining the relationship between nodes and extracting features of graph data [36]. Similar to CLNet [21], we also utilize the graph Laplacian to capture the global context of each correspondence. Specifically, we define the adjacent matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ as

$$\boldsymbol{A} = \boldsymbol{w}_l \cdot \boldsymbol{w}_l{}^T, \tag{8}$$

which encodes the affinity of each correspondence. The features with global context information, denoted as $\boldsymbol{F}^g$, can be calculated as

$$\boldsymbol{F}^g = \boldsymbol{L} \cdot \boldsymbol{Aggr}(\boldsymbol{Z}, \boldsymbol{O}^l)\boldsymbol{W}, \tag{9}$$

$$\boldsymbol{L} = \boldsymbol{D}^{-\frac{1}{2}}\widetilde{\boldsymbol{A}}\boldsymbol{D}^{-\frac{1}{2}}, \tag{10}$$

where $\boldsymbol{L}$ is the graph Laplacian [36]; $\boldsymbol{W}$ is a learnable matrix; $\widetilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{I}_N$ for numerical stability; $\boldsymbol{D} \in \mathbb{R}^{N\times N}$ is the diagonal degree matrix of $\widetilde{\boldsymbol{A}}$. Like the residual connection, an upgrade global context can be obtained

$$\boldsymbol{F} = \boldsymbol{F}^g + \boldsymbol{Aggr}(\boldsymbol{Z}, \boldsymbol{O}^l). \tag{11}$$

$\boldsymbol{F}$ then passes through a ResNet Block for further feature learning, which can be denoted as $\boldsymbol{O}^g$. Since local and global consensuses are both essential to identify inliers, an explicit and tightly coupled connection between them will help to form more cohesive feature representations. To this end, we propose a Transformer-based hierarchical aggregation module to integrate richer features and local and global consensuses, which can mitigate rank collapse problem encountered by only use the self-attention mechanisms [42]. In other words, relying solely on the self-attention can lead to similar attention outputs for certain inliers and outliers (those with similar motion vectors), reducing the efficacy of the model. In contrast, the multi-head attention module, along with layer normalization, and residual connections within the Transformer architecture, works with the hierarchical aggregation module to preserve the diversity of learned representations, thereby enlarging the distributions between inliers and outliers. For simplicity, we denote $\boldsymbol{Trans}(\cdot)$ as the same operations in Eqs. (2)–(4). The aggregated local and global consensus can be formulated as

$$\boldsymbol{O} = \boldsymbol{Trans}(\boldsymbol{Aggr}(\boldsymbol{Trans}(\boldsymbol{Aggr}(\boldsymbol{Z}, \boldsymbol{O}^l)), \boldsymbol{O}^g)). \tag{12}$$

The nested $\boldsymbol{Trans}(\cdot)$ and $\boldsymbol{Aggr}(\cdot,\cdot)$ can not only effectively integrate local and global consensuses but also the former can be served as the global prior for the latter. Similarly, the global consensus score can be obtained

$$\boldsymbol{w}_g = \boldsymbol{MLP}(\boldsymbol{Aggr}(\boldsymbol{O}^g, \boldsymbol{O})). \tag{13}$$

Finally, the inlier probabilities for all correspondences can be determined by

$$\boldsymbol{w} = \tanh(\mathrm{ReLU}(\boldsymbol{w}^g)). \tag{14}$$

3.3. Training phase

Motivated by [16], we adopt an iterative strategy. Specifically, we take local and global consensus scores obtained in the first iteration as the additional input to the second one. Different from the baseline CLNet [21], we consistently utilize the initial correspondence set for network training, instead of pruning the initial correspondence set $\boldsymbol{S}$ for essential matrix estimation, to ensure adequate feature learning. We argue that a pruning strategy in the training phase will decrease data diversity, opposite to the effect of data augmentation. (See Table 4 for an ablation test.)

Following the learning-based methods [16,21,44], the training objective can be expressed as

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{reg}(\hat{\boldsymbol{E}}, \boldsymbol{E}), \tag{15}$$

$$\mathcal{L}_{cls} = \sum_{i=1}^{L} \left(\boldsymbol{H}(\sigma(\boldsymbol{\tau}_i \odot \boldsymbol{w}_l^i), y) + \boldsymbol{H}(\sigma(\boldsymbol{\tau}_i \odot \boldsymbol{w}_g^i), y) \right), \tag{16}$$

$$\boldsymbol{\tau}_i = \exp\left(-\frac{\left\| d_i - d_{\text{thr}} \right\|_1}{d_{\text{thr}}} \right), (d_i \le d_{thr}) \tag{17}$$

where $\mathcal{L}_{cls}$ denotes a classification loss and $\mathcal{L}_{reg}$ is a regression loss [16]; λ is a hyper-parameter to balance these two terms. $\boldsymbol{\tau}_i$ is an adaptive temperature vector proposed by [21] for mitigating the influence of label ambiguity; Specifically, d_i is the epipolar distance of s_i; d_{thr} is an ad-hoc threshold (typically set to 1e-4) determining the correctness of each correspondence. We can find that for an inlier s_i with a smaller d_i would enforce larger regularization on the model optimization via a smaller $\boldsymbol{\tau}_i$. For $d_i \ge d_{thr}$, s_i is an outlier, and we set $\boldsymbol{\tau}_i = 1$. $\boldsymbol{w}_l^i$ and $\boldsymbol{w}_g^i$ are local and global consensus scores obtained in the ith iteration, respectively. $\odot$ is the Hadamard product; $\sigma(\cdot)$ represents the sigmoid function that converts the local and global consensus scores $\boldsymbol{w}_l^i$ and $\boldsymbol{w}_g^i$ into a probabilistic form for the subsequent calculation of binary cross-entropy loss. y is the binary ground truth labels, indicating the correctness of all correspondences; $\boldsymbol{H}(\cdot,\cdot)$ denotes a binary cross entropy function; L is the number of iterations; $\hat{\boldsymbol{E}}$ is the estimated essential matrix with weighted eight-point algorithm [15] and $\boldsymbol{E}$ is the ground truth essential matrix.

4. Experiments

In this section, we first introduce datasets and evaluation metrics. Then, the implementation details of our proposed TransMatch are presented. After that, we show a qualitative analysis and a quantitative comparison with other state-of-the-art methods on correspondence pruning and camera pose estimation tasks. Finally, we present an analysis, which can demonstrate the effectiveness of components in the proposed TransMatch.

4.1. Datasets and evaluation metrics

Datasets. Following [16,18,21], we conduct experiments on both outdoor scene YFCC100M [46] and indoor scene SUN3D [47]. For YFCC100M, the image pairs mainly suffer from wide baseline, repetitive structures, and large viewpoint changes. For SUN3D, the image pairs mainly undergo many textureless regions and repetitive patterns. The training set, test set, known scene, and unknown scene are generated as the protocols proposed by [16].

Evaluation Metrics. For the correspondence pruning task, we choose the commonly used Precision, Recall, and F-scores for evaluation. For the camera pose estimation, we use the mean average precision under 5° (mAP5°), the same as used in [16,18], which can reflect the accuracy of the estimated essential matrix.

4.2. Implementation details

As shown in Fig. 2, TransMatch builds on state-of-the-art CLNet [21] and the configuration of the same part remains unchanged. Specifically, we also use SIFT [9], combined with the nearest neighbor matching strategy, to construct the initial correspondence set $\boldsymbol{S}$ up to 2000 correspondences for each image pair. We set the feature channel dimension D and the embedding dimension M as 128. The multi-head self-attention includes m heads, which is set to 2 in the work. TransMatch performs two iterations in total, in which the number of nearest neighbors search K is set to 9 and 6, respectively. The hyper-parameter λ is 0.5.

Instead of using a fixed learning rate, TransMatch uses a warmup strategy with a linearly growing rate for the first $10k$ iterations. The learning rate begins to decrease at the $10k$th iteration, and reduces for every $20k$ iterations with a factor of 0.4. We utilize Adam [48] as the optimizer for a total of $50k$ iterations.

4.3. Correspondence pruning

We first evaluate TransMatch on the correspondence pruning task, which is the basis for subsequent high-level tasks. For this qualitative analysis of TransMatch, we select five image pairs that present a range of challenges, including wide baseline, repetitive structure, large viewpoint changes, textureless region and extremely low initial inlier ratio, respectively. We present visualization results of three models: the most recently proposed MS^{2}DG-Net [18], baseline CLNet [21] and our TransMatch in Fig. 5 . The visualizations show that TransMatch outperforms CLNet [21] and MS^{2}DG-Net [18] in correspondence pruning, retaining a higher proportion of inliers.

We then present quantitative comparison results between our TransMatch and other state-of-the-art methods, including traditional methods such as RANSAC [23], GMS [31], LPM [27], MAGSAC [49] and learning-based methods LFGC [15], OANet++ [16], ACNe [33], T-Net [21], CLNet [21], MS^{2}DG-Net [18], MSA-Net [37], PGFNet [17], and JRA-Net [39]. As shown in Table 1, traditional methods do not perform well, since they are sensitive to high ratio outliers in the initial correspondence set. All learning-based competitors perform better, most of them achieve high Recall but low Precision to a certain extent, except for CLNet and our TransMatch, which indicates that they cannot effectively differentiate inliers from the initial correspondence set. CLNet greatly improves Precision with a certain sacrifice in Recall, and achieves better performance in F-scores. In addition, our TransMatch further improves Precision and Recall by introducing a full Transformer structure for richer feature extraction and explicit interaction between local and global consensuses, to form more cohesive feature representations. Actually, lower Recall in CLNet and TransMatch can be attributed to the verification strategy proposed in CLNet, i.e., the final inlier probability of each correspondence is not directly determined by the network but by conducting verification on $\boldsymbol{S}$ using the estimated essential matrix. As a result, our TransMatch achieves the best precision and F-scores among all the compared methods. However, it should be noted that TransMatch benefits are not fully realized due to noises in the correspondence labels caused by a weakly-supervised label generation.

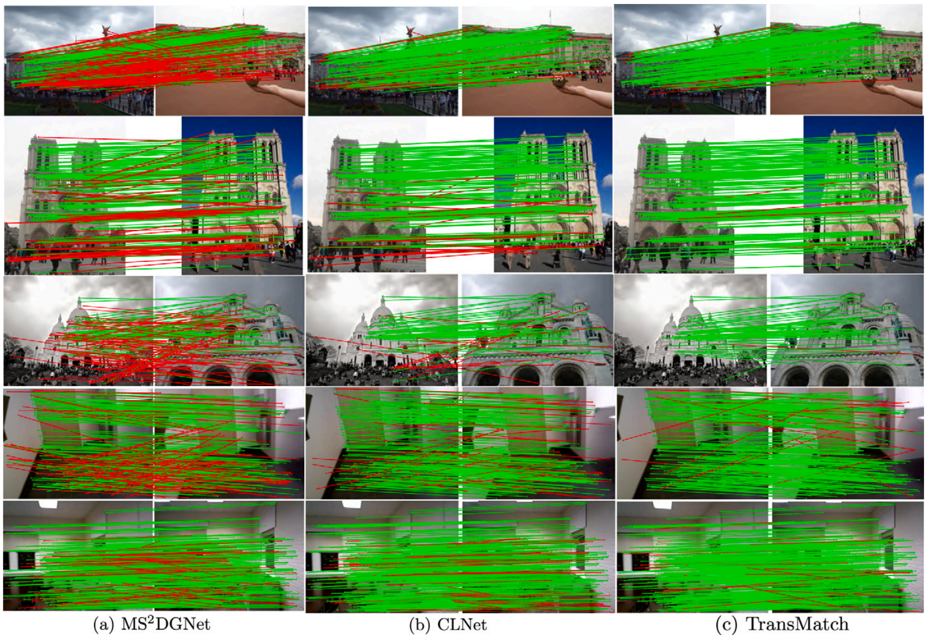
(a) MS^{2}DGNet (b) CLNet (c) TransMatch

Fig. 5. Correspondence pruning results of our TransMatch and two state-of-the-art methods on five representative image pairs (From top to down, image pairs suffer from wide baseline, repetitive structure, large viewpoint changes, textureless region and extremely low initial inlier ratio, respectively.). Inliers and outliers are marked in green and red lines, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Quantitative comparison with state-of-the-art methods on the correspondence pruning task. **Bold** indicates the best result.

Methods	References	YFCC100M(%)			SUN3D(%)		
		Precision	Recall	F-score	Precision	Recall	F-scores
RANSAC [23]	Commun. ACM 1981	41.83	57.08	48.28	44.11	46.42	45.24
GMS [31]	CVPR2017	47.75	47.92	47.83	41.84	47.91	44.67
LPM [27]	IJCV2019	43.75	65.65	51.72	44.28	55.42	50.63
MAGSAC [49]	CVPR2019	45.15	62.36	50.26	44.41	54.46	50.01
LFGC [15]	CVPR2018	52.84	85.68	65.37	46.11	83.92	59.52
OANet++ [16]	ICCV2019	55.78	85.93	67.65	46.15	84.36	59.66
ACNe [33]	CVPR2020	55.62	85.47	67.39	46.16	84.01	59.58
T-Net [44]	ICCV2021	58.21	86.38	69.55	47.27	84.16	60.54
CLNet [21]	ICCV2021	74.94	77.39	75.84	59.43	67.86	62.67
MS^{2}DG-Net [18]	CVPR2022	59.11	88.40	70.85	46.95	84.55	60.37
MSA-Net [37]	TIP2022	58.70	87.99	70.42	48.10	83.81	61.12
PGFNet [17]	TIP2023	57.54	**88.77**	69.82	47.05	**85.02**	60.58
JRA-Net [39]	PR2023	58.56	85.88	69.64	48.10	83.26	60.97
TransMatch	–	**77.83**	79.38	**78.60**	**61.11**	69.09	**64.86**

4.4. Camera pose estimation

Camera pose estimation refers to the process of determining the position and orientation of a camera relative to the 3D world. It is a critical aspect of computer vision tasks, including visual slam, object tracking, and structure from motion. Accurate camera pose estimation can enable them to function effectively and produce reliable results. We conduct a comparison between the proposed TransMatch and several state-of-the-art traditional and learning-based methods. We choose RANSAC [23], GC-RANSAC [50], MAGSAC [49], and MAGSAC++ [51] as traditional methods, which are well-suited for robust estimation and model fitting. In addition, learning-based networks include LFGC [15], OANet++ [16], ACNe [33], SuperGlue [20], LMCNet [35], T-Net [44], CLNet [21], MS^{2}DG-Net [18], MSA-Net [37], CSDA-Net [38], PGFNet [17], and JRA-Net [39]. For the traditional methods, the ratio test [9] strategy with a threshold of 0.8 is adopted to eliminate the negative impact of high ratio outliers and render reasonable results. In addition, RANSAC with a threshold of 0.001 is adopted as the post-processing step for learning-based methods.

Table 2 shows that our TransMatch outperforms other state-of-the-art methods by a significant margin in both outdoor and indoor scenarios. Specifically, for outdoor data, without RANSAC for post-processing, our method achieves mAP5° of 67.63% and 56.18% for unknown and known scenes, respectively. Our method shows a significant improvement over the second-best method by a margin of 13.7% and 11.98%, respectively. Similarly, for indoor data, our TransMatch achieves mAP5° of 22.38% and 26.89% for unknown and known scenes without RANSAC, with an improvement over the second-best method by a margin of 3.23% and 3.06%, respectively. Compared with correspondence pruning, the performance of camera pose estimation depends not only on the number of preserved inliers, but on the correlation formed between inliers, in which the latter are not paid enough attention. To alleviate this problem, we use richer feature extraction to obtain progressive local and global consensuses. Additionally, we use Transformer-based hierarchical aggregation to facilitate interaction between these two consensuses, allowing the network to prioritize the essential interdependence among inliers and ultimately improving the estimated essential matrix accuracy.

Notably, for other learning-based methods, RANSAC as a postprocessing step typically results in improved performance in most cases. But, we have found that the performance of our TransMatch actually declines, compared with these networks. This suggests that TransMatch

Table 2
Quantitative comparison with state-of-the-art methods on the camera pose estimation task. mAP5° (%) on both known and unknown scenes (SIFT as feature descriptor) are reported. **Bold** indicates the best result.

Methods	References	YFCC100M (%)				SUN3D (%)			
		Known		Unknown		Known		Unknown	
		–	RANSAC	–	RANSAC	–	RANSAC	–	RANSAC
RANSAC [23]	Commun. ACM 1981	–	31.58	–	41.09	–	20.15	–	11.49
GC-RANSAC [50]	CVPR2018	–	30.43	–	41.58	–	18.86	–	14.14
MAGSAC [49]	CVPR2019	–	32.80	–	41.61	–	20.35	–	16.24
MAGSAC++ [51]	CVPR2020	–	30.48	–	40.95	–	18.90	–	14.19
LFGC [15]	CVPR2018	17.45	36.75	25.95	50.00	11.55	20.60	9.30	16.40
OANet++ [16]	ICCV2019	32.57	41.53	38.95	52.59	20.86	22.31	16.18	17.18
ACNe [33]	CVPR2020	29.17	40.32	33.06	50.89	18.86	22.12	14.12	16.99
SuperGlue [20]	CVPR2020	35.00	43.17	48.12	55.06	22.50	23.68	17.11	18.23
LMCNet [35]	CVPR2021	33.73	40.39	47.50	55.03	19.92	21.79	16.82	17.38
T-Net [44]	ICCV2021	42.99	45.25	48.20	55.85	22.38	22.96	17.24	17.57
CLNet [21]	ICCV2021	39.60	45.12	53.93	59.75	19.66	23.73	15.83	18.88
MS^{2}DG-Net [18]	CVPR2022	38.36	45.34	49.13	57.68	22.20	23.00	17.84	17.79
MSA-Net [37]	TIP2022	39.53	44.57	50.65	56.28	18.64	22.03	16.86	17.79
CSDA-Net [38]	PR2022	39.53	44.57	50.65	56.28	18.64	22.03	16.86	17.79
PGFNet [17]	TIP2023	44.20	46.28	53.70	57.83	23.66	23.87	19.32	18.00
JRA-Net [39]	PR2023	38.88	44.21	45.75	54.13	22.50	22.45	17.44	17.26
TransMatch	–	**56.18**	**51.22**	**67.63**	**63.15**	**26.89**	**25.91**	**22.38**	**20.64**

Table 3
Quantitative comparison with state-of-the-art methods on the unknown scenes of YFCC100M. mAP5° (%) is reported. **Bold** indicates the best result.

Methods	References	SuperPoint	
		–	RANSAC
RANSAC [23]	Commun. ACM 1981	–	34.38
LFGC [15]	CVPR2018	24.25	42.57
OANet++ [16]	ICCV2019	35.27	45.45
ACNe [33]	CVPR2020	32.98	45.34
T-Net [44]	ICCV2021	40.08	47.83
CLNet [21]	ICCV2021	39.19	48.15
MS^{2}DG-Net [18]	CVPR2022	37.38	46.48
PGFNet [17]	TIP2023	42.03	47.30
JRA-Net [39]	PR2023	38.85	47.28
TransMatch	–	**55.56**	**53.45**

Table 4
Ablation studies of key components in our TransMatch on YFCC100M dataset. **BASE**: baseline model; **RFE**: richer feature extraction; **LC**: local consensus; **GC**: global consensus; **THA**: Transformer-based Hierarchical Aggregation; **ITER**: iterative strategy.

BASE	RFE	LC	GC	THA	ITER	mAP5°
✓						53.93
✓	✓					55.78
✓	✓	✓				57.32
✓	✓	✓	✓			62.44
✓	✓	✓	✓	✓		66.24
✓	✓	✓	✓	✓	✓	67.63

Table 5
Ablation studies of different combinations of local consensus and global consensuses.

Method	Precision	Recall	F-scores	mAP5°
-GC	59.45	84.51	69.80	58.92
-LC	75.69	77.24	76.46	64.87
TransMatch	77.83	79.38	78.60	67.63

Table 6
Efficiency evaluation. Average runtime (ART, unit: ms) of each image pair on YFCC100M with SIFT and parameter size (size, MB) are reported. **Bold** indicates the best result.

Methods	OANet++ [16]	CLNet [21]	MS^{2}DGNet [18]	TransMatch
Size(MB)	2.47	**1.27**	2.61	3.93
ART(ms)	**51.36**	58.45	55.62	62.81

is already highly effective, and additional post-processing steps such as RANSAC are unnecessary. In fact, TransMatch can be used as a standalone solution for essential matrix estimation.

4.5. Analysis

Analysis of the robustness. To test the robustness of different networks, we also consider using the learning-based feature extraction method SuperPoint [10] to establish an initial correspondence set. We conduct the experiment on the unknown scene of YFCC100M and report the results of our TransMatch and other state-of-the-art methods. As shown in Table 3, our method achieves the best results (55.56% and 53.45% mAP5° without and with RANSAC, respectively). Without RANSAC, TransMatch obtains 13.53% performance gains over the second-best method, which is consistent with the results obtained using SIFT as the feature extractor.

Ablation studies of key components. To demonstrate the effectiveness of key components in our TransMatch, we progressively incorporate them into the baseline. As shown in Table 4, "BASE" refers to the baseline model CLNet [21]; "RFE" denotes replacing the feature extraction in the baseline with richer feature extraction proposed in TransMatch; "LC" and "GC" indicate using newly proposed local and global consensus learning, respectively. "THA" is the Transformer-based hierarchical aggregation, replacing the sequential direct connection between local and global consensuses in the baseline. "ITER" denotes the iterative strategy, replacing the pruning strategy in the baseline. Upon analyzing Table 4, it becomes apparent that each key component has a significant impact, with particular emphasis on the local and global consensus learning, and the Transformer-based hierarchical aggregation. This observation confirms the necessity of the explicit interactions between local and global consensuses.

Ablation studies of different combinations of local and global consensuses. We further explore the importance of local and global consensuses by experimenting with different combinations. To be specific, "-GC" and "-LC" refer to the exclusion of original global and local consensus learning from the complete TransMatch model, respectively. Table 5 shows that global consensus learning plays a decisive role in essential matrix estimation, while local consensus learning has a comparatively lesser impact. However, incorporating both through the proposed Transformer-based hierarchical aggregation leads to the most optimal performance.

Analysis of efficiency. We provide the parameter size and average runtime per image pair on YFCC100M with SIFT in Table 6. We compare three methods, OANet++, CLNet, and MS^{2}DGNet. It can be observed that OANet++ is the fastest method among the competitors. CLNet has the smallest parameter size, thanks to the proposed pruning framework. Although our TransMatch has a larger parameter size and longer average runtime, the differences are not significant. Given the substantial performance improvement, these differences are acceptable.

5. Conclusion

This paper proposes TransMatch, a simple yet surprisingly effective method to accomplish correspondence pruning. TransMatch builds a full Transformer structure on a state-of-the-art network for richer feature extraction and progressive local and global consensus learning. The Transformer-based hierarchical aggregation module establishes interactions between local and global consensuses of correspondences, resulting in highly cohesive feature representations. Only employing the original plain Transformer can produce a remarkable improvement over the baseline model, particularly in the challenging task of camera pose estimation. Although TransMatch delivers superior performance, its efficiency is relatively lower due to its parameter-heavy architecture, which involves numerous transformer modules. In the future, a lightweight transformer design could be explored to enhance acceleration for real-time applications.

CRediT authorship contribution statement

Yizhang Liu: Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Yanping Li:** Conceptualization, Visualization, Writing – review & editing. **Shengjie Zhao:** Funding acquisition, Supervision,.

Declaration of competing interest

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

Acknowledgments

This work was supported in part by the National Key R&D Program of China 2023YFC3806000 and 2023YFC3806002, in part by the National Natural Science Foundation of China under Grant 61936014, in part by Shanghai Municipal Science and Technology Major Project, China No. 2021SHZDZX0100, and in part by Shanghai Science and Technology Innovation Action Plan Project 22511105300.

Data availability

Data will be made available on request.

References

[1] Q. Dai, F. Fang, J. Li, G. Zhang, A. Zhou, Edge-guided composition network for image stitching, Pattern Recognit. 118 (2021) 108019.
[2] Y. Liu, B.N. Zhao, S. Zhao, L. Zhang, Progressive motion coherence for remote sensing image matching, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–13.
[3] X. Jiang, Y. Xia, X.-P. Zhang, J. Ma, Robust image matching via local graph structure consensus, Pattern Recognit. 126 (2022) 108588.
[4] Z. Li, Y. Ma, X. Mei, J. Ma, Two-view correspondence learning using graph neural network with reciprocal neighbor attention, ISPRS J. Photogramm. Remote Sens. 202 (2023) 114–124.
[5] J. Ma, Y. Ma, C. Li, Infrared and visible image fusion methods and applications: A survey, Inf. Fusion 45 (2019) 153–178.
[6] M. Xu, L. Tang, H. Zhang, J. Ma, Infrared and visible image fusion via parallel scene and texture learning, Pattern Recognit. 132 (2022) 108929.
[7] X. Ma, Z. Wang, H. Li, P. Zhang, W. Ouyang, X. Fan, Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6851–6860.
[8] X. Jiang, J. Ma, G. Xiao, Z. Shao, X. Guo, A review of multimodal image matching: Methods and applications, Inf. Fusion 73 (2021) 22–71.
[9] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[10] D. DeTone, T. Malisiewicz, A. Rabinovich, Superpoint: Self-supervised interest point detection and description, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 224–236.
[11] L. Cavalli, V. Larsson, M.R. Oswald, T. Sattler, M. Pollefeys, Adalam: Revisiting handcrafted outlier detection, 2020, arXiv preprint arXiv:2006.04250.
[12] J. Ma, X. Jiang, A. Fan, J. Jiang, J. Yan, Image matching from handcrafted to deep features: A survey, Int. J. Comput. Vis. 129 (2021) 23–79.
[13] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R.R. Martin, S.-M. Hu, Pct: Point cloud transformer, Comput. Vis. Media 7 (2) (2021) 187–199.
[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.
[15] K.M. Yi, E. Trulls, Y. Ono, V. Lepetit, M. Salzmann, P. Fua, Learning to find good correspondences, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2666–2674.
[16] J. Zhang, D. Sun, Z. Luo, A. Yao, L. Zhou, T. Shen, Y. Chen, L. Quan, H. Liao, Learning two-view correspondences and geometry using order-aware network, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 5845–5854.
[17] X. Liu, G. Xiao, R. Chen, J. Ma, Pgfnet: Preference-guided filtering network for two-view correspondence learning, IEEE Trans. Image Process. 32 (2023) 1367–1378.
[18] L. Dai, Y. Liu, J. Ma, L. Wei, T. Lai, C. Yang, R. Chen, MS2DG-Net: Progressive correspondence learning via multiple sparse semantics dynamic graph, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022, pp. 8973–8982.
[19] Z. Li, C. Su, F. Fan, J. Huang, J. Ma, MC-Net: Integrating multi-level geometric context for two-view correspondence learning, IEEE Trans. Circuits Syst. Video Technol. (2024).
[20] P.-E. Sarlin, D. DeTone, T. Malisiewicz, A. Rabinovich, Superglue: Learning feature matching with graph neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 4938–4947.
[21] C. Zhao, Y. Ge, F. Zhu, R. Zhao, H. Li, M. Salzmann, Progressive correspondence pruning by consensus learning, in: Proceedings of the IEEE International Conference on Computer Vision, 2021, pp. 6464–6473.
[22] J. Ma, A. Fan, X. Jiang, G. Xiao, Feature matching via motion-consistency driven probabilistic graphical model, Int. J. Comput. Vis. 130 (9) (2022) 2249–2264.
[23] M.A. Fischler, R.C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM 24 (6) (1981) 381–395.
[24] O. Chum, J. Matas, Matching with PROSAC-progressive sample consensus, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1, 2005, pp. 220–226.
[25] P. Torr, A. Zisserman, Robust computation and parametrization of multiple view relations, in: Proceedings of the IEEE International Conference on Computer Vision, 1998, pp. 727–732.
[26] E. Brachmann, C. Rother, Neural-guided RANSAC: Learning where to sample model hypotheses, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 4322–4331.
[27] J. Ma, J. Zhao, J. Jiang, H. Zhou, X. Guo, Locality preserving matching, Int. J. Comput. Vis. 127 (5) (2019) 512–531.
[28] Y. Liu, B.N. Zhao, S. Zhao, Rectified neighborhood construction for robust feature matching with heavy outliers, IEEE Geosci. Remote Sens. Lett. 19 (2022) 1–5.
[29] Y. Liu, Y. Li, L. Dai, T. Lai, C. Yang, L. Wei, R. Chen, Motion consistency-based correspondence growing for remote sensing image matching, IEEE Geosci. Remote Sens. Lett. 19 (2021) 1–5.
[30] Y. Liu, Y. Li, L. Dai, C. Yang, L. Wei, T. Lai, R. Chen, Robust feature matching via advanced neighborhood topology consensus, Neurocomputing 421 (2021) 273–284.
[31] J.-W. Bian, W.-Y. Lin, Y. Liu, L. Zhang, S.K. Yeung, M.-M. Cheng, I. Reid, GMS: Grid-based motion statistics for fast, ultra-robust feature correspondence, Int. J. Comput. Vis. 128 (6) (2020) 1580.
[32] C.R. Qi, H. Su, K. Mo, L.J. Guibas, Pointnet: Deep learning on point sets for 3d classification and segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 652–660.
[33] W. Sun, W. Jiang, E. Trulls, A. Tagliasacchi, K.M. Yi, Acne: Attentive context normalization for robust permutation-equivariant learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 11286–11295.
[34] C. Zhao, Z. Cao, C. Li, X. Li, J. Yang, Nm-net: Mining reliable neighbors for robust feature correspondences, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 215–224.
[35] Y. Liu, L. Liu, C. Lin, Z. Dong, W. Wang, Learnable motion coherence for correspondence pruning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 3237–3246.

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

[37] L. Zheng, G. Xiao, Z. Shi, S. Wang, J. Ma, MSA-Net: Establishing reliable correspondences by multiscale attention network, IEEE Trans. Image Process. 31 (2022) 4598–4608.

[38] S. Chen, L. Zheng, G. Xiao, Z. Zhong, J. Ma, CSDA-Net: Seeking reliable correspondences by channel-spatial difference augment network, Pattern Recognit. 126 (2022) 108539.

[39] Z. Shi, G. Xiao, L. Zheng, J. Ma, R. Chen, JRA-Net: Joint representation attention network for correspondence learning, Pattern Recognit. 135 (2023) 109180.

[40] Z. Li, S. Zhang, J. Ma, U-match: Two-view correspondence learning with hierarchy-aware local context aggregation, in: IJCAI, 2023, pp. 1169–1176.

[41] X. Liu, J. Yang, Progressive neighbor consistency mining for correspondence pruning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 9527–9537.

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[43] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: Proceedings of Machine Learning Research, 2021, pp. 10347–10357.

[44] Z. Zhong, G. Xiao, L. Zheng, Y. Lu, J. Ma, T-Net: Effective permutation-equivariant network for two-view correspondence learning, in: Proceedings of the IEEE International Conference on Computer Vision, 2021, pp. 1950–1959.

[45] A.L. Yuille, N.M. Grzywacz, A mathematical analysis of the motion coherence theory, Int. J. Comput. Vis. 3 (2) (1989) 155–175.

[46] B. Thomee, D.A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, L.-J. Li, YFCC100M: The new data in multimedia research, Commun. ACM 59 (2) (2016) 64–73.

[47] J. Xiao, A. Owens, A. Torralba, Sun3d: A database of big spaces reconstructed using sfm and object labels, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1625–1632.

[48] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.

[49] D. Barath, J. Matas, J. Noskova, MAGSAC: marginalizing sample consensus, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10197–10205.

[50] D. Barath, J. Matas, Graph-cut RANSAC, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6733–6741.

[51] D. Barath, J. Noskova, M. Ivashechkin, J. Matas, MAGSAC++, a fast, reliable and accurate robust estimator, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 1304–1312.

Yizhang Liu received the B.S. degree in electronic and information engineering and master's degree in computer science and technology from Fujian Agriculture and Forestry University, Fuzhou, China, in 2013 and 2017, respectively. He is currently pursing the Eng.D. degree with the school of software engineering, Tongji University, Shanghai. His current research interests include computer vision and image matching.

Yanping Li received the M.S. degree in computer science and technology from Hohai University, Nanjing, China, in 2020. She is currently pursuing the D.Eng.degree with the College of Electronics and Information Engineering, Tongji University. Her research interests include computer vision and person re-identification.

Shengjie Zhao received the B.S. degree in electrical engineering from the University of Science and Technology of China, Hefei, China, in 1988, the M.S. degree in electrical and computer engineering from the China Aerospace Institute, Beijing, China, in 1991, and the Ph.D. degree in electrical and computer engineering from Texas A&M University, College Station, TX, USA, in 2004. He is currently the Dean of the College of Software Engineering and a Professor with the College of Software Engineering and the College of Electronics and Information Engineering, Tongji University, Shanghai, China. In previous postings, he conducted research at Lucent Technologies, Whippany, NJ, USA, and the China Aerospace Science and Industry Corporation, Beijing. His research interests include artificial intelligence, big data, wireless communications, image processing, and signal processing. He is a fellow of the Thousand Talents Program of China and an academician of the International Eurasian Academy of Sciences.