PGFNet: Preference-Guided Filtering Network for Two-View Correspondence Learning

Xin Liu, Guobao Xiao, *Member, IEEE*, Riqing Chen, and Jiayi Ma, *Senior Member, IEEE*

***Abstract*— Accurate correspondence selection between two images is of great importance for numerous feature matching based vision tasks. The initial correspondences established by off-the-shelf feature extraction methods usually contain a large number of outliers, and this often leads to the difficulty in accurately and sufficiently capturing contextual information for the correspondence learning task. In this paper, we propose a Preference-Guided Filtering Network (PGFNet) to address this problem. The proposed PGFNet is able to effectively select correct correspondences and simultaneously recover the accurate camera pose of matching images. Specifically, we first design a novel iterative filtering structure to learn the preference scores of correspondences for guiding the correspondence filtering strategy. This structure explicitly alleviates the negative effects of outliers so that our network is able to capture more reliable contextual information encoded by the inliers for network learning. Then, to enhance the reliability of preference scores, we present a simple yet effective Grouped Residual Attention block as our network backbone, by designing a feature grouping strategy, a feature grouping manner, a hierarchical residual-like manner and two grouped attention operations. We evaluate PGFNet by extensive ablation studies and comparative experiments on the tasks of outlier removal and camera pose estimation. The results demonstrate outstanding performance gains over the existing state-of-the-art methods on different challenging scenes. The code is available at https://github.com/guobaoxiao/PGFNet.**

***Index Terms*— Correspondence selection, feature matching, outlier removal, camera pose estimation.**

I. Introduction

FEATURE matching aims to establish accurate feature point correspondences of two images with the same or similar scene, and then recover their camera poses, *e.g.*, rotation and translation and it is vital for a series of computer vision tasks, such as image registration [1], image fusion [2], Structure from Motion [3] and visual Simultaneous Localization and Mapping [4].

The performance of feature matching mainly depends on the quality of correspondences. It usually consists of two key steps: establishing initial correspondences and selecting correct ones (*i.e.*, inliers). To be specific, initial correspondences can be established according to the similarity constraint of feature descriptors obtained by off-the-shelf feature extraction methods [5], [6]. However, due to the existence of various cross-image variations (*e.g.*, viewpoint and illumination changes, repetitive structures, occlusions and blurs), false correspondences (*i.e.*, outliers) are inevitable, and this severely hinders the downstream tasks. Therefore, correspondence selection as a crucial post-processing step can be employed to further identify inliers from initial correspondences contaminated by numerous outliers.

There are a number of correspondence selection methods that have been proposed to solve this problem over the past few decades [2]. Recently, learning-based correspondence selection methods [7], [8], [9] have received increasing attention, since neural network can provide much stronger feature representation and learning capacity in comparison with traditional methods. As the pioneering work, LFGC [7] formulates the correspondence selection problem as an inlier/outlier classification problem. It designs an end-to-end permutation-equivariant neural network based on Multi-Layer Perceptrons (MLPs) because of the irregular and unordered properties of correspondences. LFGC also proposes a context normalization to embed global contextual information (*i.e.*, the latent relations among correspondences) and a weighted eight-point algorithm for recovering camera poses.

The follow-up works use different ways to improve the correspondence selection performance. Several methods utilize well-designed network structures to capture contextual information. For example, NM-Net [10] exploits reliable correspondence neighbors for local context; OANet [11] proposes three novel operations to capture both local and global contextual information; LMCNet [12] extracts the motion coherence property of correspondences from both local and global aspects. Moreover, ACNe [13], T-Net [14], MS^{2}DG-Net [15] and MSA-Net [16] develop different attention modules for enhancing the representation ability of important features. These methods can achieve decent performance gains, however, they still suffer from some limitations. On the one hand,

Manuscript received 1 September 2022; revised 23 November 2022 and 19 December 2022; accepted 30 January 2023. Date of publication 9 February 2023; date of current version 22 February 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62072223 and Grant 61972093, and in part by the Fujian University Industry University Research Joint Innovation Project under Grant 2022H6006. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Yulan Guo. *(Corresponding authors: Guobao Xiao; Riqing Chen.)*

Xin Liu is with the College of Computer and Control Engineering, Minjiang University, Fuzhou 350108, China, and also with the College of Computer and Information Science, Fujian Agriculture and Forestry University, Fuzhou 350002, China.

Guobao Xiao is with the College of Computer and Control Engineering, Minjiang University, Fuzhou 350108, China (e-mail: gbx@mju.edu.cn).

Riqing Chen is with the College of Computer and Information Science, Fujian Agriculture and Forestry University, Fuzhou 350002, China (e-mail: riqing.chen@fafu.edu.cn).

Jiayi Ma is with the Electronic Information School, Wuhan University, Wuhan 430072, China.

Digital Object Identifier 10.1109/TIP.2023.3242598

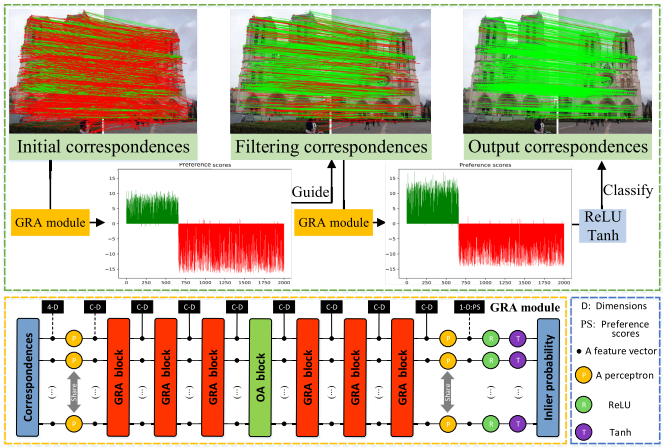


Fig. 1. An example showing the iterative process of PGFNet. We exhibit the visualization results of initial correspondences (top-left), filtering correspondences (top-middle), and out correspondences (top-right). The inliers and outliers are marked by green and red lines, respectively. The preference scores, which are responsible for guiding our correspondence filtering strategy and classifying initial correspondences, are derived from the GRA module. Each GRA module contains several GRA blocks and an OA block.

MLPs and normalization operations used in previous methods treat each correspondence indiscriminately, which makes it difficult to accurately distinguish both inliers and outliers. This is because initial correspondences established by feature extraction methods contain a large number of outliers (see the top-left image in Fig. 1), especially for some challenging scenes, such as our experimental datasets with around 90% of outliers. These randomly distributed outliers will seriously hinder neural network from capturing accurate and sufficient contextual information encoded by the inliers. Although most of previous works [13], [14], [15], [16] have attempted to implicitly learn the attention weights to mitigate this problem, they still suffer from the negative influence of numerous outliers during network learning, resulting in sub-optimal performance. On the other hand, previous works generally utilize a PointNet-like structure [17], [18] (denoted as PointCN block) as their network backbone. However, this structure is inadequate, since it only uses a single sequential structure to process correspondences. Meanwhile, it ignores the potential channel-wise and spatial-wise relations of feature maps, which limits the feature representation ability. Therefore, designing a suitable network backbone for gathering rich contextual information is necessary to further improve the performance of learning-based correspondence selection methods.

In order to address the above problems, we propose a novel and effective network, called Preference-Guided Filtering Network (PGFNet) as shown in Fig. 1. The proposed PGFNet explicitly avoids the negative influence of a large number of outliers for correspondence learning. To be specific, we design an iterative filtering structure to progressively distill more reliable candidates from initial correspondences for the subsequent network learning. We first utilize a network module to learn the preference score of each initial correspondence, which indicates the degree of an initial correspondence being an inlier, *i.e.*, a correspondence with a higher preference score is more likely to be an inlier, and vice versa. Then, the network can filter out the majority of outliers according to the preference scores. Next, the subsequent network module takes the filtering correspondences, which contain fewer outliers (see the top-middle image in Fig. 1) compared with the initial ones, as new inputs to largely alleviate the effects of outliers. Thus, it could improve the model robustness of the subsequent network module, since potential inliers can be highlighted, and more accuracy and sufficient contextual information can be captured. Finally, we combine the outputs of all network modules to obtain the final correspondence classification results. Noteworthy, the preference score is the core of our iterative filtering structure, and it plays two important roles: as the preliminary inlier probability to guide the correspondence filtering strategy, and as the final logit value of network to classify initial correspondences. Therefore, the performance of our network depends on the reliability of learned preference scores.

To enhance the reliability of preference scores, we design a simple yet effective Grouped Residual Attention (GRA) block as our network backbone. The GRA block not only includes an effective feature grouping manner to process correspondences for gathering sufficient contextual information, but also contains a hierarchical residual-like manner to increase the diversity of feature groups. In addition, it also has two effective grouped attention operations (*i.e.*, a grouped spatial attention operation and a grouped channel attention operation) to efficiently capture the potential spatial-wise and channel-wise relations of feature maps. In particular, we first evenly divide the input feature map along the channel dimension into several smaller feature groups. Each feature group has been processed by a PointCN block to exploit the contextual information from different aspects. Moreover, to increase the diversity and inlier representation ability of feature groups, we use a hierarchical residual-like manner and the grouped spatial attention operation to process each feature group. Then, we concatenate all feature groups and utilize the grouped channel attention operation to enhance the representation ability of important channels. Finally, we adopt a channel shuffle operation proposed by ShuffleNet [19] to further fuse mutual information of feature groups without bringing extra network parameters. Our GRA block can be used as a novel network backbone to replace the PointCN block of previous works, and it can achieve significant performance improvements with reasonable computational overhead.

We summarize the main contributions of this paper as follows:

- We propose a novel and effective network to guide the correspondence filtering strategy by learning the preference scores of correspondences. The proposed network explicitly avoids the negative effects of large amounts of outliers and sufficiently capture accurate contextual information for correspondence learning.
- We design a simple yet effective GRA block to improve the reliability of learned preference scores. The GRA block as the network backbone is able to further gather rich contextual information and enhance the representative ability of important features by our well-designed a series of novel operations.

- Our proposed PGFNet is able to achieve significant performance improvements compared with existing state-of-the-art methods by the qualitative and quantitative results of outlier removal and camera pose estimation tasks on both challenging outdoor and indoor datasets.

The remainder of this paper is organized as follows: We first review the related literature in Sec. II. Then, we describe the details of our proposed PGFNet in Sec. III. Finally, we give the experimental results in Sec. IV and the conclusions in Sec. V.

II. Related Work

In this section, we briefly introduce the related work about traditional and learning-based correspondence selection methods, followed by the attention mechanism.

A. Traditional Methods

The most representative traditional methods for correspondence selection are resampling-based RANSAC [20] and its variations. They use a hypothesize-and-verify structure to gradually generate the parametric model based on the sampled inlier set. In particular, RANSAC first hypothesizes a parametric model estimated by the randomly sampled data subset, and then verifies its reliability. Based on the idea, its variations have been proposed to improve the performance and efficiency by adopting different strategies. For example, DEGENSAC [21] proposes the concept of H-degeneracy by an epipolar geometry consistent. USAC [22] combines some important advancements and computational considerations by developing a universal framework. MAGSAC [23] presents a novel termination criterion based on the proposed marginalization method, and improves the model accuracy by using σ-consensus. This type of approach is still considered as the standard solution for selecting accurate inliers and estimating robust models over the past decades. However, the effectiveness and efficiency largely depend on the reliability of sampled subsets.

In addition, some non-parametric methods [24], [25], [26] are also designed to filter outliers by the geometric relationship of correspondences. VFC [24] interpolates a vector field between correspondences to estimate the consensus of inliers. GMS [25] converts the motion smoothness constraints into the grid-based motion statistic measures for rejecting outliers. LPM [26] removes outliers by the local neighborhood relationship of correspondences. Although these methods have shown promising results, they cannot perform well when outliers are dominant in the initial correspondences.

B. Learning-Based Methods

Recently, learning-based correspondence selection methods achieve competitive performance due to the powerful learning and representation capacity of neural network. Unlike some existing methods [27], [28], [29] that use Convolutional Neural Network or Transformer to process matching images directly, these methods take initial correspondences established by feature extraction methods as network inputs with competitive results and low computational overhead. As a pioneering work, LFGC [7] proposes a correspondence classification network to label the initial correspondences as inliers or outliers, and then recover the camera pose of two matching images according to the inlier weights. NM-Net [10] develops a compatibility-specific mining method to extract more reliable neighbors of each correspondence for local context. OANet [11] presents the differentiable pooling and unpooling operations to exploit local context. ACNe [13] employs an attentive context normalization to improve the network capacity for capturing local and global contexts. T-Net [14] designs a simple T-structure to adequately exploit the feature information of each sub-network. Furthermore, NM-Net [10], LMCNet [12] and MS^{2}DG-Net [15] leverage the neighbor consistency constraint, and devise different manners to aggregate the neighbor information for better distinguishing correspondences. However, the presence of a large number of outliers severely hinders the search of consistent neighbors and the learning of neural networks. In this paper, we propose an iterative filtering structure to explicitly avoid the negative effects of outliers. This structure is also useful for these previous works. For example, when the iterative filtering structure is equipped, [10], [12], and [15] can further search for reliable neighbors due to the explicit decrease of outliers.

C. Attention Mechanism

The attention mechanism helps a network pay attention to vital elements of the inputs [30]. In recent years, it has been widely used in the field of computer vision. For example, SENet [31] introduces a channel attention block to adaptively recalibrate channel-wise feature responses by modeling the dependency of feature channels. NLNet [32] presents a non-local operation based on self-attention mechanism to calculate the feature position weights for all image pixels. CBAM [33] infers attention maps along channel and spatial dimensions by different pooling operators. DANet [34] designs two types of attention modules to model semantic interdependencies for scene segmentation.

Some methods [13], [14], [15], [16] also introduce the attention mechanism to correspondence learning problem. They devise different attention mechanism modules to capture contextual information from the channel dimension or spatial dimension of feature maps. In this paper, we design a simple yet effective GRA block from both channel and spatial aspects to adequately capture contextual information and enhance the representation ability of important features.

III. Method

In this section, we describe the proposed two-view correspondence learning method in detail. In the following, we first introduce the problem formulation, and then describe the details of our proposed PGFNet.

A. Problem Formulation

Given a pair of matching images (I, I′), our goal is to establish reliable feature point correspondences and recover the

accurate camera pose accordingly. For this purpose, we first utilize off-the-shelf feature extraction methods (*e.g.*, SIFT [5] or SuperPoint [6]) to detect image keypoints and construct corresponding descriptors. Then, by using the nearest-neighbor matching strategy of feature descriptors, we establish an initial correspondence set S:

$$S = \{s_1, s_2, \ldots, s_N\} \in \mathbb{R}^{N\times 4}, \quad s_i = (u_i, v_i, u_i', v_i'), \quad (1)$$

where N is the number of correspondences. s_i is the i-th correspondence. $q_i = (u_i, v_i)$ and $q_i' = (u_i', v_i')$ are two keypoint coordinates normalized by known camera intrinsics:

$$s_i = \frac{s_i^r - q^c}{f}, \quad (2)$$

where s_i^r and q^c are the raw coordinates of correspondence s_i and central points in two matching images, respectively. f is the focal points.

Following previous works [7], [11], we formulate the two-view correspondence learning task as an inlier/outlier classification problem and an essential matrix regression problem. The whole architecture of the proposed network is outlined in Fig. 1. Specifically, our PGFNet takes the initial correspondence set S as input, and finally outputs the corresponding probability set W:

$$W = \{w_1, w_2, \ldots, w_N\}, \quad w_i \in [0, 1) \quad (3)$$

where w_i denotes the probability of correspondence s_i being an inlier. Subsequently, we utilize a weighted eight-point algorithm [7] to regress the essential matrix, which is used for recovering the camera pose of two matching images. The whole process of our method can be formulated as:

$$W = \tanh(ReLU(ps)), \quad ps = f_\theta(S), \quad \hat{E} = g(S, W), \quad (4)$$

where ps is the final output of our network (also denoted preference scores in this paper) for correspondence classification. $tanh(\cdot)$ and $ReLU(\cdot)$ are two activation functions to compute inlier weights. $f_\theta(\cdot)$ represents our neural network with relevant parameters θ. $\hat{E}$ is the predicted essential matrix according to the weighted eight-point algorithm $g(\cdot)$.

Loss Function: We utilize a hybrid loss function (*i.e.*, a classification loss and an essential matrix loss) to optimize our neural network as follows:

$$\mathcal{L} = \mathcal{L}_c(ps, L) + \beta \mathcal{L}_e(E, \hat{E}), \quad (5)$$

where β is a weight to balance these two loss functions. $\mathcal{L}_c$ indicates the classification loss based on the binary cross entropy:

$$\mathcal{L}_c(ps, L) = \frac{1}{N}\sum_{i=1}^{N} \gamma_i H(ps, L_i), \quad (6)$$

where γ_i (set 0.5 for simplicity) is the per-label weight to balance the outlier/inlier ratios. H represents the binary cross entropy function. The ground-truth correspondence labels L is calculated by a geometric distance as follows:

$$d(s, E) = \frac{(q'^T E q)^2}{\|Eq\|_{[1]}^2 + \|Eq\|_{[2]}^2 + \|E^T q'\|_{[1]}^2 + \|E^T q'\|_{[2]}^2}, \quad (7)$$

where $s = (q, q')$ represents the initial correspondences including two keypoint coordinates. E is the ground truth essential matrix, and $t_{[i]}$ denotes the i-th item of vector t. The labels are determined by a threshold of 10^{-4}. This is a weakly supervised manner to avoid exhaustive annotations.

The essential matrix loss $\mathcal{L}_e$ is a geometry loss [9] based on the above geometric distance:

$$\mathcal{L}_e(E, \hat{E}) = \frac{(p'^T \hat{E} p)^2}{\|Ep\|_{[1]}^2 + \|Ep\|_{[2]}^2 + \|E^T p'\|_{[1]}^2 + \|E^T p'\|_{[2]}^2}. \quad (8)$$

Noteworthy, p and p' are virtual correspondence coordinates generated by the ground truth essential matrix E.

B. Preference-Guided Filtering Network

Recall that existing works adopt MLPs and normalization operations to indiscriminately process initial correspondences or implicitly learn the attention weights to alleviate the negative influence of outliers. This is difficult to distinguish inliers and outliers well when outliers are dominant in the initial correspondences.

To avoid the negative effect of highly outlier-contaminated correspondences as much as possible, our PGFNet utilizes an iterative filtering structure, and it progressively distills a more reliable candidate set $\hat{S}$ from the initial correspondence set S by exploiting the preference scores. Here, we give the example of two iterations as illustrated in Fig. 1. Specifically, we first use a GRA module taking the set $S \in \mathbb{R}^{N\times 4}$ as input to learn the preference score (*i.e.*, the preliminary inlier probability) of each correspondence for guiding the progressive correspondence filtering strategy. Then, we get a candidate set $\hat{S} \in \mathbb{R}^{\hat{N}\times 4}$ by distilling the top $\hat{N}$ correspondences according to the preference scores, where $\hat{N} < N$. By taking the filtering correspondence set $\hat{S}$ as input of the subsequent GRA module, we expect to obtain a more robust parametric model for correspondence learning. The filtering correspondences have fewer outliers compared with the initial correspondences so that the subsequent GRA module can capture accuracy and sufficient contextual information. In addition, the weights and residuals calculated by the previous GRA module have been added as the additional inputs of the subsequent GRA module (we omit the two terms for simplicity). In the end, we combine the preference scores of two GRA modules as the final outputs of our PGFNet, which is used for computing the final inlier probability set. The whole structure can be formulated as:

$$W = \tanh(ReLU(ps)), \quad ps = \alpha ps_1 + ps_2 \quad (9)$$

$$ps_1 = f_{\theta_1}(S), \quad \hat{ps_2} = f_{\theta_2}(\hat{S}), \quad ps_2 = com(\hat{ps_2}, idx) \quad (10)$$

where $f_{\theta_1}(\cdot)$ and $f_{\theta_2}(\cdot)$ denote two GRA modules with learnable parameters θ_1 and θ_2, respectively. ps_1 and $\hat{ps_2}$ are preference scores learned from two GRA modules, respectively. We use a customized function $com(\cdot)$ to complement and align the preference scores of the second GRA module (we set 0 for removed correspondences). idx is the indexes of $\hat{N}$ correspondences. α is a learned weight, which has been initialized as 0 and gradually learns a more suitable value, to control the influence of the first GRA module.

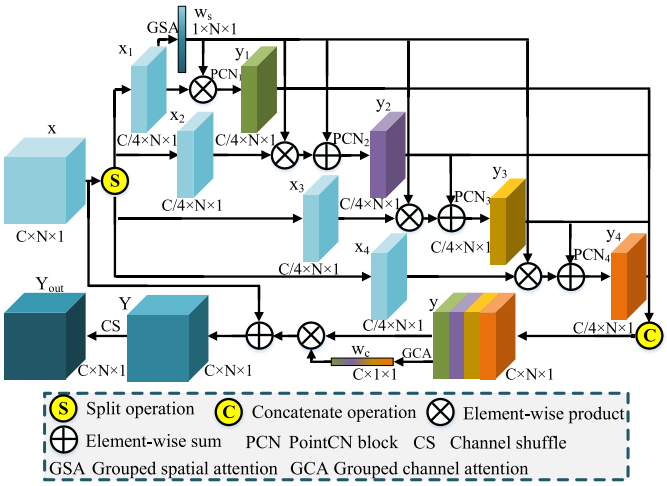


Fig. 2. The structure diagram of our Grouped Residual Attention block (the size of group $g = 4$). It evenly splits the feature map x into g groups, and obtains the final output Y_{out} by the grouped processing. The hierarchical residual-like manner increases the diversity of feature groups. The grouped spatial attention operation and the grouped channel attention operation enhance the representation ability of inliers and important channels, respectively. The channel shuffle operation fuses mutual information of feature groups.

Our PGFNet is able to obtain remarkable results due to the iterative filtering structure. We further demonstrate its effectiveness in Sec. IV. It is worth noting that the preference scores are not only the guided data to distill the reliable candidate set, but also the final outputs of network to predict the inlier probability. Therefore, we design a novel GRA block to learn reliable preference scores.

C. Grouped Residual Attention Block

PointCN block as network backbone used in previous works has some limitations as described in Sec. I. Therefore, we construct a novel GRA module as shown in Fig. 1, where the GRA block is the network backbone. Moreover, to capture local context, we also insert an Order-Aware (OA) block [11], and it consists of a DiffPool layer to cluster correspondences, some Order-Aware Filtering layers to exploit global context and a DiffUnpool layer to recover the original size of correspondences. Each network module mainly contains several GRA blocks and an OA block, which provides adequate contextual information and strong feature maps for learning reliable preference scores.

Our proposed GRA block adopts a grouped residual-like manner to adequately capture contextual information from different aspects, and utilizes the attention mechanism to enhance the representative ability of inliers and important channels. The structure of GRA block is illustrated in Fig. 2. Specifically, given an input feature map $x \in \mathbb{R}^{C\times N\times 1}$, where C and N are the numbers of channels and correspondences, respectively, we evenly split it along the channel dimension into g groups, denoted by $x_i, i \in \{1, 2, \ldots, g\}$. Therefore, each feature map subset $x_i \in \mathbb{R}^{C/g\times N\times 1}$ has the same feature size. Firstly, we apply a spatial attention operation for the first feature map subset x_1 to generate the spatial-wise relations. For efficiently computing the spatial weights, we adopt average-pooling and max-pooling operations along the channel dimension of x_1 to aggregate the channel information. We then concatenate them to generate a spatial feature $F_s \in \mathbb{R}^{2\times N\times 1}$:

$$F_s = concat(AvgPool_0(x_1), MaxPool_0(x_1)), \tag{11}$$

where $concat(\cdot)$ is the concatenation operation. $AvgPool_0(\cdot)$ and $MaxPool_0(\cdot)$ are the average-pooling operation and max-pooling operation along the 0-th dimension (*i.e.*, the channel dimension), respectively. The two pooling operations not only aggregate channel information from different aspects, but also squeeze feature maps for lightweight computation. F_s is then passed through an MLP layer followed by a batch normalization layer and a sigmoid activation function to obtain the spatial attention weight $w_s \in \mathbb{R}^{1\times N\times 1}$:

$$w_s = GSA(x_1) = \sigma(BN(MLP_0(F_s))), \tag{12}$$

where $MLP_0(\cdot)$ denotes an MLP layer with the size of $\mathbb{R}^{2\times 1\times 1\times 1}$ to decrease the channel dimension to 1. $BN(\cdot)$ represents a batch normalization layer and $\sigma(\cdot)$ is the sigmoid activation function. $GSA(\cdot)$ denotes the whole grouped spatial attention operation.

Then, each feature map subset is processed by a corresponding PointCN block, denoted by $PCN_i(\cdot)$. We use y_i to denote the output of $PCN_i(\cdot)$. Meanwhile, we utilize the spatial attention weight w_s to enhance the inlier representation ability of each feature map subset. This operation is effective and avoids repeated calculations of attention weights. To increase the diversity of subsets that output features can represent, we adopt a novel hierarchical residual-like manner to connect all feature map subsets. To be specific, except for $PCN_1(\cdot)$, each $PCN_i(\cdot)$ also receives the output of previous PointCN block as an additional input. Thus, y_i is formulated as:

$$y_i = \begin{cases} PCN_i(w_s \otimes x_i), & i = 1; \\ PCN_i(w_s \otimes x_i + y_{i-1}), & 1<i \le g, \end{cases} \tag{13}$$

where $\otimes$ denotes the element-wise product. Note that we utilize the feature grouping manner to exploit contextual information from different aspects, which is beneficial for neural network learning. Furthermore, the hierarchical residual-like manner can make that each $PCN_i(\cdot)$ is able to receive feature information of previous PointCN blocks, and then generates the output features with strong representation ability due to the combinatorial explosion effect.

Next, we concatenate all feature map subsets along the channel dimension and get the output feature y:

$$y = concat(y_1, y_2, \ldots, y_g). \tag{14}$$

We then exploit the channel-wise relations by employing the grouped channel attention operation for feature y. Similarly, we adopt both average-pooling and max-pooling operations along the spatial dimension of output feature y to aggregate its spatial information. We then add them to generate a channel feature $F_c \in \mathbb{R}^{C\times 1\times 1}$:

$$F_c = AvgPool_{12}(y) + MaxPool_{12}(y), \tag{15}$$

where $AvgPool_{12}(\cdot)$ and $MaxPool_{12}(\cdot)$ are the average pool operation and maximum pool operation along the spatial

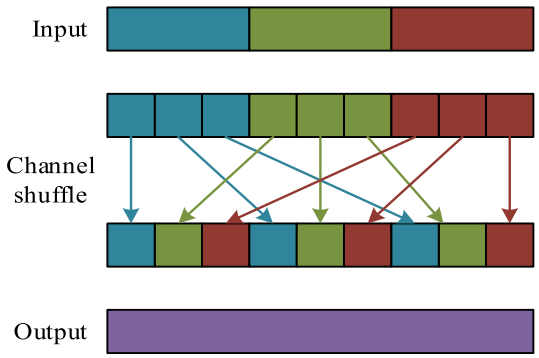


Fig. 3. The example of channel shuffle operation for three feature map subsets.

dimension (*i.e.*, both the 1-th dimension and the 2-th dimension). Here, we use an element-wise sum operator to fuse the information of two channel features for parameter-efficient. F_c is then passed through a bottleneck structure to compute the channel attention weight $w_c \in \mathbb{R}^{C\times1\times1}$:

$$\begin{aligned} w_c &= GCA(y) \\ &= \sigma(BN(MLP_2(ReLU(BN(MLP_1(F_s)))))), \end{aligned} \tag{16}$$

where $MLP_1(\cdot)$ and $MLP_2(\cdot)$ with the size of $\mathbb{R}^{C\times\frac{C}{r}\times1\times1}$ and $\mathbb{R}^{\frac{C}{r}\times C\times1\times1}$ (we empirically set the reduction ratio $r = g$ in this paper) are two MLP layers. $GCA(\cdot)$ denotes the whole grouped channel attention operation. The channel attention weight is used for enhancing the representation ability of important channels of feature y. We then utilize an element-wise sum operation with the input x to obtain the output Y:

$$Y = x + w_c \otimes y. \tag{17}$$

Finally, to increase the communication of feature map subsets and fuse their mutual information, we introduce the channel shuffle operation proposed by ShuffleNet [19] as shown in Fig. 3 to obtain a new feature map:

$$Y_{out} = CS(Y), \tag{18}$$

where $CS(\cdot)$ is the channel shuffle operation, which will not bring extra learning parameters. Y_{out} is the final output of our GRA block.

IV. Experiments

In this section, we evaluate the performances of our PGFNet by the tasks of outlier removal and camera pose estimation. In the following, we first describe the used datasets and implementation details, and then introduce the evaluation metrics. Finally, we give the ablation studies and comparative results with some state-of-the-art methods.

A. Datasets

We utilize both outdoor scenes and indoor scenes to train and test networks.

1) Outdoor Scenes: We use the Yahoo's YFCC100M dataset [35], which is a tourist collection of 100 million images from Internet. Following OANet [11], the dataset has been divided into 71 image sequences according to different tourist spots. We use 4 sequences as unknown scenes to test network generalization ability, and the remaining 67 sequences for network training.

2) Indoor Scenes: We use the SUN3D dataset [36], which is a collection of image frames from some RGBD videos. The dataset has been divided into 254 image sequences, where 15 sequences as unknown scenes to test network generalization ability, and the remaining 239 sequences for network training. Note that, the indoor scenes usually contain extensive repetitive structures or texture-less objects. Therefore, it is a more challenging scene than the outdoor scenes.

In this paper, we test network performance on both known and unknown scenes. The known scenes are the above training sequences. Thus, we divide them into some disjoint subsets, including training set (60%), validation set (20%) and testing set (20%).

B. Implementation Details

The baseline networks [7], [11] utilize 12 PointCN blocks as their network backbones, OANet [11] uses 6 Order-Aware Filtering layers in an OA block to capture global context. Therefore, for a fair comparison, we use the same number of network blocks as OANet. SIFT [5] has been used to establish $N \times 4$ (typically $N = 2000$) initial correspondences as the network inputs. We adopt the two iteration structure as the default setting for PGFNet. After the first GRA module, the number of correspondences is filtered to $\hat{N} = N/2$, which gives the best performance. In our GRA block, the numbers of channels and subsets are set to 256 and 4, respectively. We will discuss the details in Sec. IV-D. For OA block, we keep the same settings as OANet. Our method and compared learning-based methods have been implemented by Pytorch. To optimize network models, we utilize Adam optimizer with the learning rate of 10^{-3}. The batchsize is set as 32 and the training period is set as 500k iterations. The weight β in our loss function is initialized as 0 during the first 20k iterations, and then set as 0.5 in the rest iterations.

C. Evaluation Metrics

For both outlier removal task and camera pose estimation task, we use different evaluation metrics to reflect the performance of methods.

1) Outlier Removal: Outlier removal task is aimed at measuring the accuracy of methods for correspondence classification. We adopt $Precision$ (P), $Recall$ (R) and $F\text{-}score$ (F) as the evaluation metrics. In particular, $Precision$ is defined as the number ratio between the recognized inliers and the preserved correspondences. $Recall$ is defined as the number ratio between the recognized inliers and practical inliers in the initial correspondences. $F\text{-}score$ is calculated by $2PR/(P+R)$.

TABLE I

ABLATION STUDIES ABOUT THE EFFECTIVENESS OF ITERATIVE FILTERING STRUCTURE ON YFCC100M. WE REPORT MAP5° (%) AND MAP20° (%) **WITH/WITHOUT** RANSAC ON THE UNKNOWN OUTDOOR SCENES. **POINTCN**: USING THE POINTCN BLOCK. **OA**: USING THE OA BLOCK. **GRA**: USING THE GRA BLOCK. **2-ITER**: USING THE TWO ITERATIONS. **3-ITER**: USING THE THREE ITERATIONS

PointCN	GRA	OA	2-Iter	3-Iter	mAP5°	mAP20°
✓					50.00/25.95	70.20/55.46
✓		✓			51.18/33.08	72.09/60.72
	✓				51.75/34.90	72.55/62.99
	✓	✓			53.20/37.88	73.34/65.52
	✓	✓	✓		57.83/53.70	76.91/75.97
	✓	✓		✓	56.65/49.78	76.56/74.14

2) Camera Pose Estimation: Camera pose estimation task aims to measure the accuracy of methods for predicted essential matrix. We adopt rotation and translation vectors recovered from the essential matrix to encode the camera pose of two matching images. Here, the angular differences between the predicted and the ground truth vectors are used as the error metrics. We summarize the results in terms of mean average precision (mAP) of the angular differences as the evaluation metrics. Furthermore, RANSAC with a threshold of 0.001 as a post-processing step has been adopted to further improve the accuracy. In this paper, we use mAP under 5° and 20° (*i.e.*, mAP5° and mAP20°) as the main evaluation metrics.

D. Ablation Studies

In this section, we provide ablation studies about the core part of our PGFNet on the unknown outdoor scenes for the camera pose estimation task. We select LFGC [7] and OANet [11] as the baselines.

1) Iterative Filtering Structure: In order to verify the effectiveness of our proposed iterative filtering structure, we test the performance gains of PGFNet with different combinations as shown in Table I. When we replace the PointCN block of LFGC with the GRA block (the subset number and each subset channel dimension are fixed to 4 and 64), it can achieve better performance improvements in terms of mAP5° and mAP20°. Our PGFNet without the iterative operation (GRA + OA) outperforms OANet by 4.80% in terms of mAP5° without RANSAC. Noteworthy, when we adopt the iterative filtering structure, it obtains a large margin of improvements over baselines. To be specific, when using the two iterations (each GRA module contains 6 GRA blocks and 3 Order-Aware Filtering layers in an OA block), our PGFNet (GRA + OA + 2-Iter) achieves significant performance gains over OANet with 20.62% and 15.25% in terms of mAP5° and mAP20° without RANSAC as the post-processing. However, we find that the performance of using the three iterations (each GRA module contains 4 GRA blocks and 2 Order-Aware Filtering layers in an OA block) drops compared with using the two iterations. This is because the redundant filtering operation also discards many important inliers, leading to poor performance.

In other words, the camera pose estimation requires sufficient inliers. Therefore, we test the network's generalization ability for different inlier ratios, which indicate the percentage of inliers in the initial correspondences. Here, we utilize Lowe's ratio test [5] to control the inlier ratio of the testing set. Different thresholds (*i.e.*, 1, 0.95, 0.9, 0.85, 0.8, and 0.75) of the ratio test are used to obtain the corresponding testing sets. As shown in Fig. 4, the performance of the network will decrease when the inlier ratio increases. Although the ratio test can reduce outliers in the input correspondences, it also involuntarily discards many important inliers. When the threshold is 1 (*i.e.*, without the ratio test), our method can achieve the best performance, which also demonstrates that our network is better suited for challenging situations. Furthermore, the performance of camera pose estimation may also depend on the number of input correspondences. We therefore use SIFT to establish different numbers of initial correspondences as network inputs. As shown in Table II, the network performs better when using more initial correspondences as inputs. On the contrary, using fewer correspondences as inputs will severely degrade the performance of the network, which further indicates that the network needs sufficient inliers to estimate reliable camera poses. Thus, considering the computational resource constraints and network performance, we adopt the two iterations in our PGFNet, and use 2000 initial correspondences as network inputs.

The main parameter of our iterative filtering structure is the number of distilled correspondences $\hat{N}$, which controls the input size of the second iterative network module. A larger value may result in the presence of too many outliers, and a smaller value may lead to poor generalization ability. To verify this inference, we test the performance of our PGFNet with

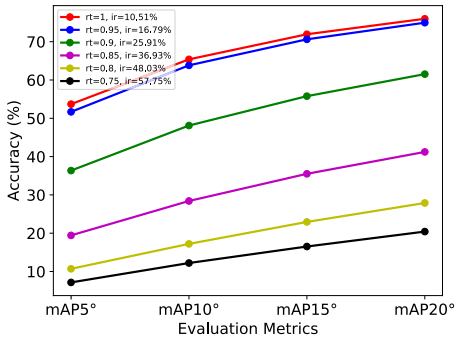
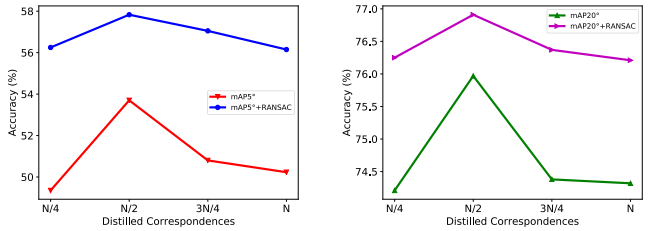


Fig. 4. Ablation studies of the performance for different inlier ratios (ir) on the unknown outdoor scenes. The ratio test (rt) with different thresholds is used to control the inlier ratio. Results without RANSAC under error thresholds of 5°, 10°, 15° and 20° are reported.

TABLE II

ABLATION STUDIES OF NETWORK'S GENERALIZATION ABILITY FOR USING DIFFERENT NUMBERS OF INITIAL CORRESPONDENCES AS NETWORK INPUTS. WE SHOW MAP5° (%) AND MAP20° (%) **WITH/WITHOUT** RANSAC AS A POST-PROCESSING STEP ON THE UNKNOWN OUTDOOR SCENES

Input Number	mAP5°	mAP20°
500	27.65/22.75	45.42/43.51
1000	46.10/41.48	64.78/63.86
2000	57.83/53.70	76.91/75.97
4000	59.93/53.30	80.38/78.42
8000	57.08/47.25	77.39/73.72



(a) mAP5° (b) mAP20°

Fig. 5. Parameter analysis about the size of $\hat{N}$ on YFCC100M datasets. We show mAP5° (%) and mAP20° (%) with/without RANSAC as post-processing on the unknown outdoor scenes.

TABLE III

ABLATION STUDIES ABOUT THE EFFECTIVENESS OF GROUPED RESIDUAL ATTENTION BLOCK ON YFCC100M. WE REPORT MAP5° (%) AND MAP20° (%) WITH/WITHOUT RANSAC ON THE UNKNOWN OUTDOOR SCENES. **POINTCN**: USING THE POINTCN BLOCK. **GSA**: USING THE GROUPED SPATIAL ATTENTION OPERATION. **HR**: USING THE HIERARCHICAL RESIDUAL-LIKE MANNER. **GCA**: USING THE GROUPED CHANNEL ATTENTION OPERATION. **CS**: USING THE CHANNEL SHUFFLE OPERATION

PointCN	GSA	HR	GCA	CS	mAP5°	mAP20°
✓					50.00/25.95	70.20/55.46
✓		✓	✓	✓	50.95/33.15	71.09/62.19
✓	✓		✓	✓	50.63/32.60	71.62/61.67
✓	✓	✓		✓	51.93/33.58	72.38/62.92
✓	✓	✓	✓		50.88/32.45	70.22/59.63
✓	✓	✓	✓	✓	51.75/34.90	72.55/62.99

TABLE IV

ABLATION STUDIES ABOUT THE DIFFERENT MODEL SIZES OF GRA BLOCK ON YFCC100M. MAP5° (%) AND MAP20° (%) ON THE UNKNOWN OUTDOOR SCENES ARE REPORTED WITH/WITHOUT RANSAC AS A POST-PROCESSING STEP. PARAMETER B IS THE NUMBER OF BLOCK, AND g IS THE NUMBER OF SUBSETS, AND C IS THE CHANNEL DIMENSIONS. **PARAMS(M):** THE NUMBER OF NETWORK PARAMETERS

Method	Setting	mAP5°	mAP20°	Params (M)
PointCN	12B-0g-128C	50.00/25.95	70.20/55.46	0.4032
PointCN	12B-0g-256C	51.53/31.88	71.53/61.27	1.5928
GRA (Increased block)	6B-4g-256C	49.78/31.30	70.71/59.62	0.4098
	12B-4g-256C	51.75/34.90	72.55/62.99	0.8180
	18B-4g-256C	51.48/35.48	73.13/64.05	1.2262
GRA (Increased subset)	12B-2g-128C	49.70/32.28	70.47/60.43	0.4102
	12B-4g-256C	51.75/34.90	72.55/62.99	0.8180
	12B-8g-512C	50.80/33.78	70.97/61.43	1.6336
GRA (Increased channel)	12B-4g-128C	49.05/25.73	69.54/52.95	0.2124
	12B-4g-256C	51.75/34.90	72.55/62.99	0.8180
	12B-4g-512C	54.00/42.60	73.81/68.99	3.2088
GRA (Similar Params)	24B-2g-128C	53.28/35.40	73.09/63.83	0.8196
	12B-4g-256C	51.75/34.90	72.55/62.99	0.8180
	6B-8g-512C	51.40/33.55	71.77/61.91	0.8183

TABLE V

ABLATION STUDIES ABOUT THE EFFECTIVENESS OF ITERATIVE FILTERING STRUCTURE (IFS) AND GROUPED RESIDUAL ATTENTION (GRA) BLOCK. MAP5° (%) AND MAP20° (%) WITH/WITHOUT RANSAC ON THE UNKNOWN OUTDOOR SCENES ARE REPORTED

Method	mAP5°	mAP20°
OANet	51.18/33.08	72.09/60.72
OANet++	52.59/38.95	72.99/66.85
OANet & IFS	54.10/42.73	75.24/69.51
OANet++ & GRA	56.08/50.98	75.35/74.58
PGFNet	57.83/53.70	76.91/75.97

different numbers of $\hat{N}$. The results are shown in Fig. 5. The $\hat{N} = N$ indicates that the second GRA module uses the same number of correspondences as the first GRA module. The results demonstrate that our PGFNet with $\hat{N} = N/2$ gains the best performance compared to other settings.

2) Grouped Residual Attention Block: To enhance the reliability of preference scores, we design a GRA block as the network backbone to replace the PointCN block used in previous methods. Our GRA block mainly contains four parts: the grouped spatial attention operation, the hierarchical residual-like manner, the grouped channel attention operation, and the channel shuffle operation. We use LFGC (PointCN) as the baseline, and do not adopt the iterative filtering structure for a fair comparison. Table III reports the performance gains of the four parts. We can see that, the performance of methods slightly drops when any one of the four parts is not applied, demonstrating the effectiveness of our four parts in GRA block. Our whole GRA block achieves the best performance gains over the baseline with 8.95% and 7.53% in terms of mAP5° and mAP20° without RANSAC.

The main parameters of our GRA block are the subset number and the subset channel dimension. The more subsets and subset channel dimensions mean it may have a stronger capacity to capture contextual information but needs more computational cost. Therefore, to analyze the relationship between network performance and different model sizes (*i.e.*, block number (B), subset number (g) and channel dimension (C)) of GRA block, we conduct ablation experiments as shown in Table IV. We select the LFGC with 128 and 256 channels as the baselines. When we keep the same or lower network parameters than baselines, our GRA block can achieve outstanding performance improvements. Meanwhile, we further evaluate the performance gains of increasing the numbers of blocks and channels. We can see that the performance of GRA block will be further improved, but this requires more network parameters. However, when we increase the grouped subsets with the same blocks and subset channels, the GRA block with 12B-8g-512C will slightly drop compared with the GRA block with 12B-4g-256C. We infer that redundant subsets may lead to information redundancy. Moreover, when we keep similar network parameters, their results all outperform the baselines. Noteworthily, when each subset has small channel dimensions (*e.g.*, the GRA block with 12B-4g-128C), its performance has a significant drop since each subset cannot capture sufficient contextual information. Therefore, considering the balance between the effectiveness and efficiency, we set the GRA block with 12B-4g-256C as the optimal setting in this paper.

Furthermore, we add the proposed iterative filtering structure and GRA block into the strong baseline [11]. The results are reported in Table V. Compared to the raw iterative operation of OANet++, our iterative filtering structure (OANet & IFS) achieves better performance gains. When we use the proposed GRA block to replace the PointCN block of OANet++, it (OANet++ & GRA) can obtain significant performance improvements compared with OANet++, *e.g.*, the 12.03% mAP5° gain without RANSAC as post-processing.

TABLE VI
QUANTITATIVE COMPARATIVE RESULTS OF OUTLIER REMOVAL TASK ON BOTH UNKNOWN YFCC100M (OUTDOOR) AND SUN3D (INDOOR) DATASETS. **BOLD** INDICATES THE BEST

Datasets	YFCC100M			SUN3D		
Matcher	Precision (%)	Recall (%)	F-score	Precision (%)	Recall (%)	F-score
RANSAC [20]	41.83	57.08	48.28	44.11	46.42	45.24
MAGSAC [23]	45.15	62.36	50.26	44.41	54.46	50.01
GMS [25]	47.75	47.92	47.83	41.84	47.91	44.67
LPM [26]	43.75	65.65	51.72	44.28	55.42	50.63
PointNet++ [18]	48.42	82.93	61.54	45.64	83.43	59.00
LFGC [7]	52.84	85.68	65.37	46.11	83.92	59.52
DFE [9]	51.68	83.49	63.84	44.09	84.00	57.83
ACNe [13]	55.62	85.47	67.39	46.16	84.01	59.58
OANet++ [11]	55.78	85.93	67.65	46.15	84.36	59.66
PGFNet	**57.54**	**88.77**	**69.82**	**47.05**	**85.02**	**60.58**

This demonstrates that our GRA block is a very effective backbone compared with the original PointCN block. Our PGFNet delivers the best performance at all settings.

E. Comparative Results

We compare our method with some state-of-the-art feature matching methods [7], [9], [11], [12], [13], [14], [15], [16], [18], [20], [21], [23], [25], [26], [27], [37], [38]. Noteworthy, for the input of traditional methods, we clean most of the initial correspondences by additionally utilizing the ratio test [5] with a threshold of 0.8, which can significantly improve their performance. For most learning-based methods, we retrain them by the released code and keep the same training setting, including the batchsize, the learning rate, and the training period, for a fair comparison. PointNet++ [18] is an improved version of PointNet [17], which improves network performance by capturing local context of points. Here, we adopt the 4D Euclidean space (*i.e.*, correspondence coordinate space in 2D images) to seek the neighbors for each correspondence. DFE [9] is a robust method to estimate the fundamental matrix, and we use the essential matrix to replace the fundamental matrix for pose estimation. LFGC [7] is a pioneering correspondence learning work. ACNe [13] designs a learned attentive context normalization to capture local and global contexts. OANet [11] presents a cluster operation for local context, and we employ the iterative version (*i.e.*, OANet++) for comparison. LMCNet [12] learns the motion coherence property of putative correspondences with a smooth function to distinguish inliers. T-Net [14] develops two novel structures, *i.e.*, a "−" structure and a "|" structure, to integrate the useful feature information of each iteration. MS^{2}DG-Net [15] captures the local topology of correspondences by utilizing a graph neural network, and we directly quote the original results. MSA-Net [16] designs three different blocks based on the attention mechanism to capture contextual information. SuperGlue [27] utilizes attention mechanism and graph neural network to solve the correspondence assignment optimization problem.

1) Results of Outlier Removal: Outlier removal is a key step for recovering accurate two-view camera poses. In this step, we show the comparative results of some state-of-the-art outlier removal methods [7], [9], [11], [13], [18], [20], [23], [25], [26]. Table VI reports the quantitative comparative results on both unknown outdoor and indoor scenes. We see that PGFNet outperforms all comparative methods in terms of P, R and F for all scenes. In particular, our PGFNet outperforms OANet++ on F by 2.17% and 0.92% on both unknown outdoor and indoor scenes, respectively. For the known scenes, our PGFNet achieves better F gains over OANet++ with 2.74% and 1.57% on both outdoor and indoor scenes, respectively. In addition, the learning-based methods consistently obtain better performance than traditional ones, which demonstrates the effectiveness of learning-based methods for challenging matching datasets. Moreover, Fig. 6 visualizes some output correspondence results of networks. Clearly, our method can achieve the best performance on several challenging scenes, such as illumination variations, large viewpoint changes, occlusions and repetitive structures.

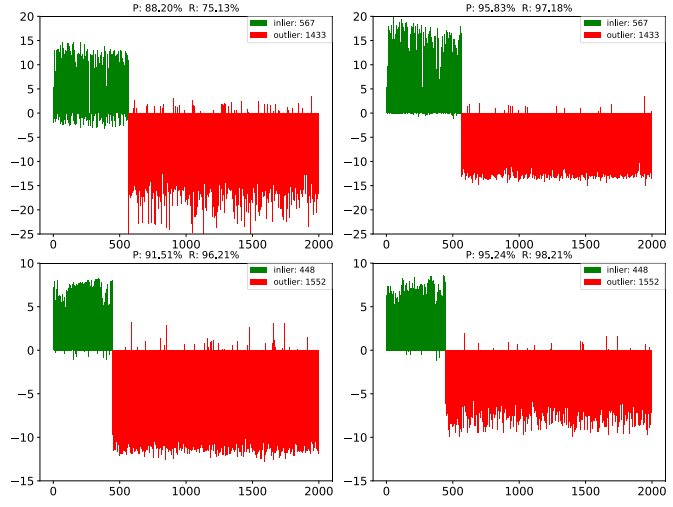
In addition, we visualize the final outputs of networks in Fig. 7. A positive value indicates that the correspondence is treated as an inlier, and vice versa. We can find that our PGFNet not only obtains more accurate correspondence classification results, but also significantly distinguishes inliers and outliers, especially for inliers (*e.g.*, the values of inliers estimated by PGFNet are obviously higher than OANet++), which further demonstrates the effectiveness of our PGFNet.

2) Results of Camera Pose Estimation: We further verify the performance of methods on the task of camera pose estimation. This task is used for evaluating the reliability of the estimated essential matrix, which is vital for various downstream vision tasks. In this section, we compare our proposed PGFNet with some state-of-the-art works. These methods include traditional RANSAC [20], DEGENSAC [21], GC-RANSAC [37], MAGSAC [23] and MAGSAC++ [38], and learning-based PointNet++ [18], LFGC [7], DFE [9], OANet++ [11], ACNe [13], SuperGlue [27], LMCNet [12], T-Net [14], MS^{2}DG-Net [15], and MSA-Net [16].

The quantitative comparative results on both outdoor and indoor scenes are reported in Table VII. The performance of RANSAC equipped with the ratio test strategy increases drastically. MAGSAC performs best among traditional methods. From the results of Table VII, we can see that learning-based methods are able to obtain better performance than traditional ones, in which our method achieves the best results under almost all testing scenes. Our PGFNet without RANSAC as a post-processing step shows significant performance improvements over these competitors without RANSAC, which demonstrates the effectiveness of our network. In addition, we observe that using RANSAC as a post-processing step may decrease the performance on indoor scenes. For example, the performance of our PGFNet with RANSAC on the unknown indoor scene drops by 1.32%. Since the indoor video dataset exists severe lack of texture, large repetitive patterns, and consistent scales, it is extremely difficult for RANSAC to further distill reliable correspondences for camera pose estimation. SuperGlue slightly outperforms our PGFNet on the unknown indoor scene when using RANSAC, but it needs more model parameters due to the use of feature descriptors and attention operations. We also exhibit the performance gains of our method with LFGC and OANet++ in terms of mAP with different error thresholds on unknown outdoor

Plain image pair (a) LFGC (b) OANet++ (c) PGFNet

Fig. 6. Visualization results using (a) LFGC (left), (b) OANet++ (middle) and (c) our PGFNet (right). The top three images come from unknown test set of the outdoor scenes and the rest images come from unknown test set of the indoor scenes. Inliers (green lines) and outliers (red lines) preserved by networks have been exhibited. Meanwhile, the information of precision and recall has been shown at the top left corner of each image pair.



(a) OANet++ (b) PGFNet

Fig. 7. Visualization results about the final output of (a) OANet++ and (b) PGFNet. The results are taken from the unknown test set of outdoor and indoor scenes. The x-axis denotes correspondences and y-axis denotes their logit values estimated by networks. The green line and red line represent the ground truth inlier and outlier, respectively. Meanwhile, we report the precision and recall at the top of each image.

TABLE VII

QUANTITATIVE COMPARISON RESULTS OF CAMERA POSE ESTIMATION TASK ON BOTH YFCC100M (OUTDOOR) AND SUN3D (INDOOR) DATASETS. MAP UNDER THE ERROR THRESHOLD OF 5° (*i.e.*, MAP5°) IS REPORTED. THOSE DEEP METHODS ALSO ADOPT RANSAC WITH THE INLIER THRESHOLD OF 0.001 AS A POST-PROCESSING STEP, IN WHICH THEIR MODEL PARAMETERS ARE LISTED IN THE SECOND COLUMN. RANSAC* MEANS DO NOT USE THE RATIO TEST STRATEGY. **BOLD** INDICATES THE BEST

Matcher	Datasets Size (MB)	YFCC100M (%) Known Scene -	YFCC100M (%) Known Scene RANSAC	YFCC100M (%) Unknown Scene -	YFCC100M (%) Unknown Scene RANSAC	SUN3D (%) Known Scene -	SUN3D (%) Known Scene RANSAC	SUN3D (%) Unknown Scene -	SUN3D (%) Unknown Scene RANSAC
RANSAC* [20]	-	-	05.72	-	09.05	-	04.43	-	02.85
RANSAC [20]	-	-	30.19	-	40.83	-	19.13	-	14.57
DEGENSAC [21]	-	-	21.00	-	27.65	-	16.01	-	11.01
GC-RANSAC [37]	-	-	30.43	-	41.58	-	18.86	-	14.14
MAGSAC [23]	-	-	32.80	-	41.61	-	20.35	-	16.24
MAGSAC++ [38]	-	-	30.48	-	40.95	-	18.90	-	14.19
PointNet++ [18]	12.00	10.54	33.66	16.65	45.72	07.51	20.31	07.93	15.63
LFGC [7]	0.39	17.45	36.75	25.95	50.00	11.55	20.60	09.30	16.40
DFE [9]	0.40	18.02	36.67	30.29	49.02	14.44	20.80	12.34	16.16
OANet++ [11]	2.47	32.57	41.53	38.95	52.59	20.86	22.31	16.18	17.18
ACNe [13]	0.41	29.17	40.32	33.06	50.89	18.86	22.12	14.12	16.99
SuperGlue [27]	12.02	35.00	43.17	48.12	55.06	22.50	23.68	17.11	**18.23**
LMCNet [12]	0.93	33.73	40.39	47.50	55.03	19.92	21.79	16.82	17.38
T-Net [14]	3.78	42.99	45.25	48.20	55.85	22.38	22.96	17.24	17.57
MS^{2}DG-Net [15]	2.61	38.36	45.34	49.13	57.68	22.20	23.00	17.84	17.79
MSA-Net [16]	1.45	39.53	44.57	50.65	56.28	18.64	22.03	16.86	17.79
PGFNet	2.99	**44.20**	**46.28**	**53.70**	**57.83**	**23.66**	**23.87**	**19.32**	18.00

and indoor scenes in Fig. 8. Our method achieves impressive improvements for all evaluation metrics.

Meanwhile, the feature extraction method is crucial for our correspondence learning task, since it determines the quality of initial correspondences. Therefore, we consider learning-based SuperPoint [6] to construct initial correspondences. SuperPoint uses a fully-convolutional neural network to compute SIFT-like 2D keypoints and corresponding descriptors in a self-supervised manner, which is used for multiple-view geometry

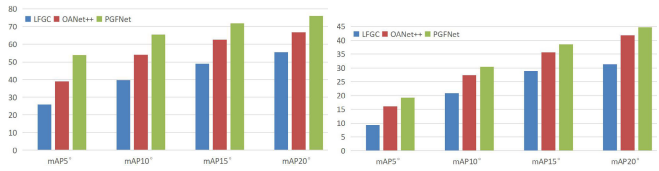


Fig. 8. Quantitative comparison of LFGC, OANet++ and PGFNet in terms of mAP (%) with different error thresholds on both the unknown (left) outdoor scene and (right) indoor scene.

TABLE VIII

PERFORMANCE COMPARISON OF OUR METHOD WITH OTHER BASELINES WHEN USING DIFFERENT FEATURE EXTRACTION METHODS ON BOTH KNOWN AND UNKNOWN YFCC100M DATASETS. RESULTS **Without/With** RANSAC POST-PROCESSING UNDER ERROR THRESHOLDS OF 5° AND 20° (*i.e.*, MAP5° AND MAP20°) ARE REPORTED. **BOLD** INDICATES THE BEST

Features	Matcher	Known Scene				Unknown Scene			
		mAP5°		mAP20°		mAP5°		mAP20°	
		-	RANSAC	-	RANSAC	-	RANSAC	-	RANSAC
SuperPoint	RANSAC [20]	-	16.09	-	48.71	-	34.38	-	57.38
	LFGC [7]	12.18	30.25	34.75	52.13	24.25	42.57	52.70	66.89
	OANet++ [11]	29.52	35.72	53.76	57.75	35.27	45.45	66.81	70.99
	PGFNet	**33.96**	**37.09**	**56.14**	**59.54**	**42.03**	**47.30**	**70.43**	**72.35**
SIFT	RANSAC [20]	-	30.19	-	51.20	-	40.83	-	61.19
	LFGC [7]	17.45	36.75	39.75	58.91	25.95	50.00	55.46	70.20
	OANet++ [11]	32.57	41.53	56.89	63.91	38.95	52.59	66.85	72.99
	PGFNet	**44.20**	**46.28**	**66.67**	**68.11**	**53.70**	**57.83**	**75.79**	**76.91**

TABLE IX

QUANTITATIVE COMPARISON RESULTS OF BOTH THE PERFORMANCE AND COMPUTATION COMPLEXITY OF NETWORKS ON THE UNKNOWN OUTDOOR SCENES. RESULTS **WITH/WITHOUT** RANSAC AS POST-PROCEDURE UNDER THE ERROR THRESHOLDS OF 5° ARE REPORTED. THE MODEL PARAMETERS, THE FLOATING POINT OPERATIONS (FLOPS) OF EACH MATCHING PAIR AND THE RUNNING TIME (AT) OF TESTING **WITH/WITHOUT** RANSAC ARE ALSO REPORTED. **BOLD** INDICATES THE BEST

Method	mAP5°	Size (MB)	FLOPs (G)	AT (s)
OANet++ [11]	52.59/38.95	**2.47**	1.84	**151/80**
T-Net [14]	55.85/48.20	3.78	2.70	329/140
PGFNet	**57.83/53.70**	2.99	**1.64**	220/160

problems in the field of computer vision. We also utilize the nearest-neighbor descriptor matching strategy to establish the initial correspondences. Table VIII gives the comparison results between SIFT and SuperPoint, where results at error thresholds 5° and 20° on both known and unknown outdoor scenes are reported. We can see that our method still achieves good performance when using SuperPoint as the feature extraction method. In addition, we see that SIFT performs better than SuperPoint in all cases. We infer that SuperPoint has better feature descriptors but inaccurate keypoint locations, which leads to difficulties in recovering accurate camera poses.

Moreover, Table IX gives the comparative results of both the performance and computation complexity of networks. We can see that our PGFNet gives the best performance with a reasonable computation complexity. Our PGFNet has lower FLOPs but more running time compared with OANet++ due to the grouped operation in GRA block.

V. CONCLUSION

In this paper, we propose a Preference-Guided Filtering Network (called PGFNet) to explicitly alleviate the effects of outliers for two-view correspondence learning. To be specific, we develop a new iterative filtering structure to guide the correspondence filtering strategy by learning the preference scores of correspondences. Moreover, we design a simple yet effective Grouped Residual Attention block as the network backbone to capture rich contextual information and enhance the representative ability of important features. The experimental results demonstrate the advantages of our method for addressing both outlier removal and camera pose estimation tasks. Our PGFNet achieves remarkable performance improvements than state-of-the-art methods on challenging datasets. Noteworthily, the preference scores are vital in our PGFNet. Therefore, in future work, we will pay attention to designing a more effective metric for correspondence filtering and learning.

REFERENCES

[1] G. Xiao, J. Ma, S. Wang, and C. Chen, "Deterministic model fitting by local-neighbor preservation and global-residual optimization," *IEEE Trans. Image Process.*, vol. 29, pp. 8988–9001, 2020.

[2] X. Jiang, J. Ma, G. Xiao, Z. Shao, and X. Guo, "A review of multimodal image matching: Methods and applications," *Inf. Fusion*, vol. 73, pp. 22–71, Sep. 2021.

[3] G. Xiao, H. Wang, J. Ma, and D. Suter, "Segmentation by continuous latent semantic analysis for multi-structure model fitting," *Int. J. Comput. Vis.*, vol. 129, no. 7, pp. 2034–2056, Jul. 2021.

[4] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015.

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

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

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

[8] E. Brachmann et al., "DSAC-differentiable RANSAC for camera localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6684–6692.

[9] R. Ranftl and V. Koltun, "Deep fundamental matrix estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 284–299.

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

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

[12] Y. Liu, L. Liu, C. Lin, Z. Dong, and W. Wang, "Learnable motion coherence for correspondence pruning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3237–3246.

[13] W. Sun, W. Jiang, E. Trulls, A. Tagliasacchi, and K. M. Yi, "ACNe: Attentive context normalization for robust permutation-equivariant learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11286–11295.

[14] Z. Zhong, G. Xiao, L. Zheng, Y. Lu, and J. Ma, "T-net: Effective permutation-equivariant network for two-view correspondence learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1950–1959.

[15] L. Dai et al., "MS2DG-Net: Progressive correspondence learning via multiple sparse semantics dynamic graph," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 8973–8982.

[16] L. Zheng, G. Xiao, Z. Shi, S. Wang, and J. Ma, "MSA-Net: Establishing reliable correspondences by multiscale attention network," *IEEE Trans. Image Process.*, vol. 31, pp. 4598–4608, 2022.

[17] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 652–660.

[18] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," *in: Adv. Neural Inf. Process. Syst., 2017*, pp. 5099–5108.

[19] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.

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

[21] O. Chum, T. Werner, and J. Matas, "Two-view geometry estimation unaffected by a dominant plane," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 772–779.

[22] R. Raguram, O. Chum, M. Pollefeys, J. Matas, and J.-M. Frahm, "USAC: A universal framework for random sample consensus," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 2022–2038, Aug. 2013.

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

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

[25] J. Bian, W.-Y. Lin, Y. Matsushita, S.-K. Yeung, T.-D. Nguyen, and M.-M. Cheng, "GMS: Grid-based motion statistics for fast, ultra-robust feature correspondence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4181–4190.

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

[27] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 4938–4947.

[28] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "LoFTR: Detector-free local feature matching with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8922–8931.

[29] W. Jiang, E. Trulls, J. Hosang, A. Tagliasacchi, and K. M. Yi, "COTR: Correspondence transformer for matching across images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6207–6217.

[30] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[31] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[32] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[33] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[34] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3146–3154.

[35] B. Thomee et al., "YFCC100M: The new data in multimedia research," *Commun. ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[36] J. Xiao, A. Owens, and A. Torralba, "SUN3D: A database of big spaces reconstructed using SfM and object labels," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1625–1632.

[37] D. Barath and J. Matas, "Graph-cut RANSAC," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6733–6741.

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

Xin Liu received the B.S. degree in the Internet of Things engineering from Zhoukou Normal University, Zhoukou, China, in 2019. He is currently pursuing the master's degree with the School of Computer and Information, Fujian Agriculture and Forestry University, Fuzhou, China. He is also attached with the Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University. His current research interests include computer vision and image matching.

Guobao Xiao (Member, IEEE) received the B.S. degree in information and computing science from Fujian Normal University, China, in 2013, and the Ph.D. degree in computer science and technology from Xiamen University, China, in 2016. From 2016 to 2018, he was a Postdoctoral Fellow with the School of Aerospace Engineering, Xiamen University. He is currently a Professor at Minjiang University, China. He has published over 60 papers in the international journals and conferences, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, *International Journal of Computer Vision* (IJCV), ICCV, and ECCV. His research interests include machine learning, computer vision, and pattern recognition. He was awarded the Best Ph.D. Thesis in Fujian Province and the China Society of Image and Graphics (a total of ten winners in China). He has served on the Program Committee (PC) for CVPR, ICCV, and ECCV.

Riqing Chen received the B.Eng. degree in communication engineering from Tongji University, China, in 2001, the M.Sc. degree in communications and signal processing from Imperial College London, U.K., in 2004, and the D.Phil. degree in engineering science from the University of Oxford, U.K., in 2010. Since 2014, he has been affiliated with the Digital Fujian Institute of the Big Data for Agriculture and Forestry, Fujian Agriculture and Forestry University, Fuzhou, China. His research interests include computer vision, big data and visualization, cloud computing, consumer electronics, flash memory, and wireless sensor networking.

Jiayi Ma (Senior Member, IEEE) received the B.S. degree in information and computing science and the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2008 and 2014, respectively. He is currently a Professor with the Electronic Information School, Wuhan University. He has authored or coauthored more than 200 refereed journals and conference papers, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, *International Journal of Computer Vision* (IJCV), CVPR, ICCV, and ECCV. His research interests include computer vision, machine learning, and pattern recognition. He has been identified in the 2019–2021 Highly Cited Researcher lists from the Web of Science Group. He is an Area Editor of *Information Fusion* and an Associate Editor of *Neurocomputing*, *Sensors*, and *Entropy*.