Information Fusion 77 (2022) 81–89

Contents lists available at ScienceDirect

Information Fusion

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

Full length article

PESA-Net: Permutation-Equivariant Split Attention Network for correspondence learning

Zhen Zhong [a], Guobao Xiao [a,*], Shiping Wang [b], Leyi Wei [c], Xiaoqin Zhang [d]

[a] College of Computer and Control Engineering, Minjiang University, Fuzhou, 350121, China
[b] College of Mathematics and Computer Science, Fuzhou University, Fuzhou, 350108, China
[c] School of Software, Shandong University, Jinan, 350108, China
[d] College of Computer Science and Artificial, Wenzhou University, Wenzhou, 350108, China

ARTICLE INFO

Keywords:
Feature matching
Wide-baseline stereo
Deep learning

ABSTRACT

Establishing reliable correspondences by a deep neural network is an important task in computer vision, and it generally requires permutation-equivariant architecture and rich contextual information. In this paper, we design a Permutation-Equivariant Split Attention Network (called PESA-Net), to gather rich contextual information for the feature matching task. Specifically, we propose a novel "Split–Squeeze–Excitation–Union" (SSEU) module. The SSEU module not only generates multiple paths to exploit the geometrical context of putative correspondences from different aspects, but also adaptively captures channel-wise global information by explicitly modeling the interdependencies between the channels of features. In addition, we further construct a block by fusing the SSEU module, Multi-Layer Perceptron and some normalizations. The proposed PESA-Net is able to effectively infer the probabilities of correspondences being inliers or outliers and simultaneously recover the relative pose by essential matrix. Experimental results demonstrate that the proposed PESA-Net relative surpasses state-of-the-art approaches for pose estimation and outlier rejection on both outdoor scenes and indoor scenes (i.e., YFCC100M and SUN3D). Source codes: https://github.com/x-gb/PESA-Net.

1. Introduction

Feature matching is a fundamental and important problem for a variety of applications in computer vision [1,2], such as Image Retrieval [3], Image Fusion [4], Image Registration [5] and Structure from Motion (SfM) [6,7].

Given two images of the same or similar scenes, the aim of feature matching is to establish reliable feature correspondences. Note that matching N feature points to another N feature points may require solving an NP-hard assignment problem. To deal with the complex problem, a common strategy of feature matching is typically solved in a two-step manner, that is, generating a set of putative correspondences by picking out point pairs with sufficiently similar feature descriptors and establishing reliable correspondences from the generated putative ones. For the first step, the putative correspondences are usually extracted by a robust extractor, such as scale invariant feature transform (SIFT) [8]. However, the brute-force putative correspondences often contain a large number of false matches (i.e., outliers), due to the low-quality images and the constraint of local descriptor information. Thus, it is critical to design a robust approach, for establishing reliable correspondences in the second step.

Recently, learning-based methods, e.g., LGC-Net [9], DFE-Net [10], OA-Net [11] and ACNe-Net [12], have been extensively proposed for feature matching, due to the excellent performance of deep neural network. However, LGC-Net, DFE-Net and OA-Net, rely on PointCN, a PointNet-like architecture with Context Normalization, which normalizes the feature maps according mean and variance. Therefore, Context Normalization can be expressed as the solution of a least-squares problem which is not robust to outliers. To deal with the problem, ACNe-Net is proposed to capture the context information in both global and local manner, by a normalization operation. However, the normalization operation neglects channel-wise correspondence contextual information, which may lead to sub-optimal performance for feature matching.

In this paper, we propose a novel attention mechanism called "*Split–Squeeze–Excitation–Union*" (SSEU) module, which extracts the contextual information in a channel-wise manner, to improve the matching performance. Comparing with other state-of-the-art approaches on YFCC100M unknown scenes [13], our network introduces very few additional parameters and negligible computations while bringing notable performance gain, as shown in Fig. 1. Specifically, the SSEU module

[*] Corresponding author.
E-mail address: gbx@mju.edu.cn (G. Xiao).

https://doi.org/10.1016/j.inffus.2021.07.018
Received 3 March 2021; Received in revised form 1 July 2021; Accepted 25 July 2021
Available online 2 August 2021
1566-2535/© 2021 Elsevier B.V. All rights reserved.

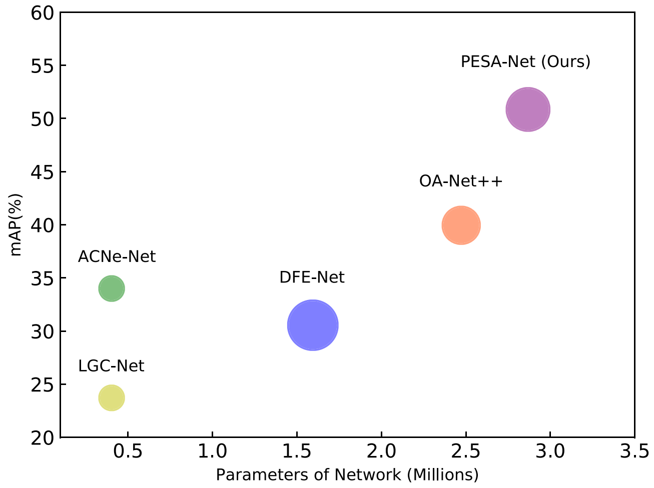

Fig. 1. Comparison of various learning based feature matching networks (i.e., LGC-Net [9], DFE-Net [10], OA-Net++ [11], ACNe-Net [12] and our proposed PESA-Net) in terms of mAP (at 5° error threshold), network parameters and FLOPs, indicated by radiuses of circles.

consists of four operations: *Split*, *Squeeze*, *Excitation* and *Union*, to gather the channel-wise global information from different aspects for feature matching. The *Split* operation generates multiple paths to exploit the geometrical context of putative correspondences from different aspects. The *Squeeze* operation aggregates feature-maps to produce a channel descriptor. The *Excitation* operation adopts the channel dependence to learn for each channel by a Multi-Layer Perceptron (MLP), to achieve the excitation of each channel. The *Union* operation combines and aggregates the geometrical context information from multiple paths. Note that, the SSEU module not only uses a channel-wise manner, but also includes a local and global manner.

To handle the unordered correspondence features, we follow the existing learning-based feature matching methods to build the network based on a Multi-Layer Perceptron (MLP), which is able to provide permutation equivariance, which is not feasible with neither convolutional nor fully-connected [12]. Then, we construct a Permutation-Equivariant Split Attention (PESA) block, which is fused by the MLP, SSEU module, and some normalizations. After that, by stacking the PSEA blocks together, we build our network called PESA-Net. We show the overview of our PESA-Net in Fig. 2. Note that we add Context Normalization after each MLP to enrich contextual information. In addition, we also insert the *Geometric Attention* Block, which contains a Differentiable Pooling Layer [14], Order Aware Filtering Block, and Differentiable Unpooling Layer [11], in the middle of each iterative sub-network to extract the local information and global information of correspondences due to the effective performance.

We summarize the contributions as follows:

- We develop a simple and effective attention mechanism, named "*Split–Squeeze–Excitation–Union*" (SSEU) module, which generates multiple paths and adopts channel-wise dependence to capture rich contextual information from different aspects in a permutation invariant manner. To the best of our knowledge, we are the first one to introduce the split-attention mechanism to handle feature matching problems.
- We construct a permutation-equivariant block, which consists of the SSEU module, Multi-Layer Perceptron and some normalizations, to exploit the complex global context of sparse and unordered correspondence data. In addition, we also design an iterative permutation-equivariant network by stacking the PESA block and *Geometric Attention* block together, for feature matching.
- The proposed PESA-Net achieves the state-of-the-art performance on relative pose estimation and outlier rejection tasks on both two challenging outdoor and indoor benchmarks (i.e., YFCC100M [13] and SUN3D [15]).

The rest of the paper is organized as follows: We first review the related feature matching literatures in Section 2. Then, we describe the details of the proposed method in Section 3 and present the experimental results in Section 4. Finally, we draw conclusions in Section 5.

2. Related work

In the section, we briefly introduce the learning-based feature matching methods highly related to our paper. In addition, we also review some related work of attention mechanisms.

2.1. Traditional handcrafted matching

The traditional handcrafted methods for feature matching between two images use the descriptors, such as SIFT [8], rotated BRIEF (ORB) [16], LIFT [17] or SuperPoint [18] to generate the rough initial correspondence set. The initial correspondence set often contains a mass of false correspondences (i.e., outliers). Thus, outlier rejection plays a core role in feature matching. Random sampling consensus (RANSAC) [19] and its variances, e.g., a maximum likelihood estimation sample consensus (MLESAC) [20], progressive sample consensus (PROSAC) [21], a universal framework for random sample consensus (USAC) [22], degeneracy check using homographies (DEGENSAC) [23] and marginalizing sample consensus (MAGSAC) [24], are the classical traditional handcrafted matching methods. These methods employ a hypothesize-and-verify framework for an attempt to obtain the largest inlier correspondence set that conforms to a provided parametric model by re-sampling. MLESAC is adept in solving image geometry problems. PROSAC shows improvements when reducing the time of the estimation process. USAC integrates multiple advancements into a unified framework. DEGENSAC can estimate the epipolar geometry from point correspondences in the possible presence of a dominant scene plane. MAGSAC eliminates the threshold by marginalizing over the noise. Although RANSAC and its variants still remain to be regarded as a standard solution in the traditional handcrafted matching, these methods considerably rely on a predefined parametric model.

In addition, to deal with the rigid and non-rigid scenarios problem, the non-parametric fitting methods are introduced as well, such as locality preserving matching (LPM) [25], guided locality preserving feature matching (GLPM) [26], and vector field consensus (VFC) [27]. LPM exploits that the image pair of the same scene or object has a similar spatial neighborhood relationship. Based on this observation, LPM adopts spatial neighbor relationships to remove outliers to retain inliers. GLPM formulates the neighborhood structures of accurate potential matches between two images into a mathematical model and gets quick results. VFC defines a reproducing kernel Hilbert space with Tikhonov regularization for smoothness constraint and introduces a new framework for non-rigid point matching.

2.2. Learning-based feature matching methods

Recently, a mass of learning-based methods gets huge success in a wide range of computer vision tasks, such as image classification, image segmentation, 3D object detection, point cloud classification, etc. Analogously, they also achieve great success in feature matching domain. For example, Graph Neural Networks [28,29] treat feature matching as an assignment problem or an optimal transport problem. However, they are extremely time-consuming and memory-costly consumption. Additionally, LGC-Net [9] proposes a simple normalization for embedding contextual information to each correspondence. DFE-Net [10] treats feature matching as a series of weighted homogeneous

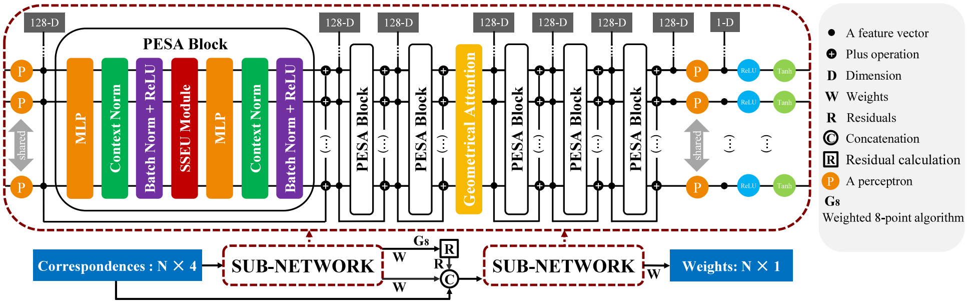

Fig. 2. The proposed PESA-Net architecture.

least-squares problems and employs a different loss function and an iterative network. OA-Net [11] employs a cluster-recover frame to reject outliers. Nevertheless, all of LGC-Net, DFE-Net and OA-Net do not perform very well for feature matching due to the constraints of the simple PointCN block, which is an important part of their network and it is not very robust to outliers. NM-Net [30] proposes a compatibility-specific distance to define neighbors of correspondence and integrates each correspondence with its neighbors. But NM-Net requires complex index preprocessing. ACNe-Net [12] also employs a normalization to extract global information and local information simultaneously, but it neglects latent channel relations of correspondences which are critical for rejecting outlier. In this work, we develop a novel module to gather channel-wise global information from different aspects, for boosting the matching performance.

2.3. Attention mechanism

The attention mechanism has been widely used in natural language processing and computer vision. For example, RMA [31] adopts the attention mechanism on the RNN model for image classification. RNNsearch [32] adopts a similar attention mechanism for simultaneous translation and alignment on machine translation tasks. Transformer [33] firstly proposes a self-attention mechanism for machine translation. After that, the attention mechanism is widely used both in natural language processing and computer vision. For instance, SE-Net [34] introduces a channel-attention mechanism to learn the contextual information adaptively. SK-Net [35] and ResNeSt [36] employ a split-attention mechanism to deal with the image classification issue.

In this paper, we try to introduce the novel attention mechanism to better capture contextual information for feature matching. However, the above attention mechanisms are designed with highly regular input data formats such as image grids. Note that, the feature matching task requires the network to be permutation-equivariant, since the input correspondences for feature matching are sparse and unordered. Thus, we design a novel "*Split–Squeeze–Excitation–Union*" (SSEU) module for feature matching.

It is worth pointing out that, although SK-Net and ResNeSt adopt similar split feature-map operation as our network, they are significantly different. Specifically, SK-Net and ResNeSt adopt a split–merge operation for making attention across the feature-map group while our network employs a split–union operation to learn contextual information from different aspects. Then, SK-Net and ResNeSt are designed for image data, thus, their networks involve the convolution layer and pooling layer to extract and compress image information. Note that, these two layers require the input order information provided by their networks (while the input data for feature matching is unordered). In contrast, our network is based on MLP and only contains a global pooling layer to capture the contextual information. That is, our network is permutation-equivariant and does not have the requirement. In addition, our network contains Context Normalization, which follows each MLP to further exploit the geometrical context of putative correspondences. Therefore, our network is much more effective than SK-Net and ResNeSt, to provide a clear leap in dealing with sparse and unordered data, for the feature matching tasks.

3. Method

In this section, we design an iterative permutation-equivariant network (called PESA-Net) to handle the outlier rejection and geometry estimation problem. In the following, we introduce the problem formulation in Section 3.1, describe the proposed SSEU module in Section 3.2, and discuss our network architecture in Section 3.3.

3.1. Problem formulation

Given image pairs, our goal is to reject outliers from putative correspondences and recover the relative pose. More specifically, we firstly adopt handcrafted features (e.g., SIFT [8], ORB [16], EIR [37]) or learning-based features (e.g., Lift [17], SuperPoint [18]) to establish putative correspondences. After that, we remove outliers by our network and establish geometrically consistent correspondences by weighted eight-point algorithm. The input data is N putative correspondences established by handcraft or learned features:

$$D = [d_1; \dots ; d_i; \dots ; d_N], d_i = [x_1^i, y_1^i, x_2^i, y_2^i], \tag{1}$$

where d_i represents a correspondence; (x_1^i, y_1^i) and (x_2^i, y_2^i) are keypoint coordinates in the range of $[-1, 1]$ normalized by camera intrinsics for two images.

In our work, the outlier rejection and two-view geometric estimation tasks are treated as the inlier/outlier classification problem and the essential matrix regression problem. Specifically, given the input correspondence set D, our network outputs the probability set $W_{PR} = [w_1, \dots, w_i, \dots, w_N]$. The probability $w_i \in [0, 1]$ is assigned to correspondence d_i, where d_i is predicted an outlier if $w_i = 0$; Otherwise d_i is predicted an inlier. Then, we adopt a weighted eight-point algorithm [9] to directly regress the essential matrix. That is, given a set of Q image pairs with the corresponding putative sets $\{D_q\}_{q=1}^{Q}$, we design a deep neural network to learn a function f, and we write the architecture as:

$$\begin{aligned} &\forall q, 1 \le q \le Q, NW_q = f_\varphi(D_q), \\ &\quad w_q = tanh(ReLU(NW_q)), \\ &\qquad \hat{E}_q = g(w_q, D_q), \end{aligned} \tag{2}$$

where NW_q represents the logit values for classification; $f_\varphi(\cdot)$ denotes a permutation-equivariant neural network; φ is the network parameters; w_q denotes the weight of correspondence D_q; $tanh$ and $ReLU$ are two activation functions for helping remove outliers [9]; $g(\cdot, \cdot)$ is the weighted eight-point algorithm which takes the corresponding weight w_q and the corresponding putative correspondence D_q to compute the essential matrix $\hat{E}_q$.

3.2. Split–squeeze–excitation–union module

In this subsection, we propose a novel attention mechanism, i.e., the SSEU module, to capture the rich global contextual information of sparse correspondences. Specifically, we implement the SSEU module

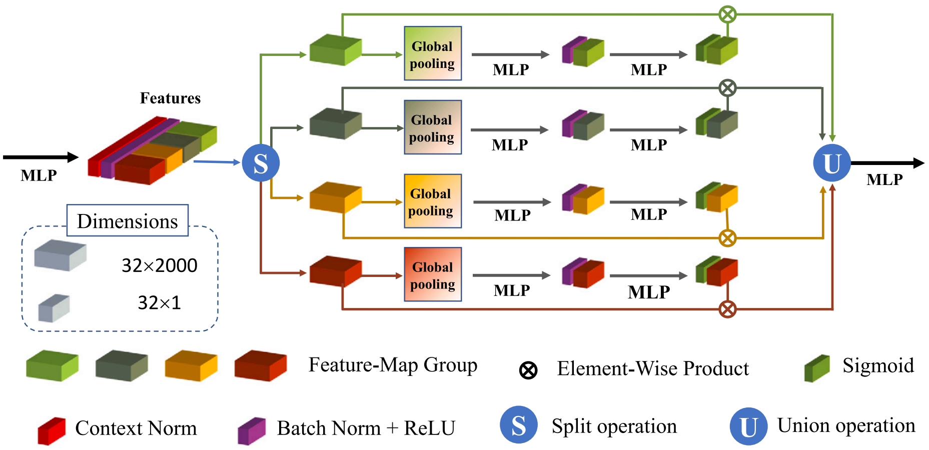

Fig. 3. The architecture of the proposed SSEU module.

via four operations, i.e., *split*, *squeeze*, *excitation* and *union*, as illustrated in Fig. 3, where the four-path case in shown.

Split: We first split feature-maps of putative correspondences, to capture the rich global contextual information. That is, for giving feature-map $F \in \mathbb{R}^{N\times C}$, we first split it into S groups of feature-map $FG \in \mathbb{R}^{N\times C/S}$, where C denotes the feature-map size of channel dimension. In this manner, we can generate multiple paths to exploit the geometrical context of putative correspondences from different aspects.

Squeeze: As mentioned in the Introduction, establishing reliable correspondence requires rich contextual information. Although, Context Normalization follows MLP in our SSEU module, it is not enough to capture full latent spatial relations (i.e., contextual information). To address this issue, in each feature-map group, we employ the squeeze manner to achieve global spatial information of correspondence features. More specifically, for each feature-map group, the global average pooling is adopted to generate channel-wise contextual descriptor. Note that, SE-Net and SK-Net also adopt a similar operation to capture channel-wise information, for image stylization; while our operation differs from them since our SSEU module is followed by MLP and the feature-map is permutation-equivariant for correspondences, that is, our operation is used to extract global contextual information of each correspondence. In contrast, the operation of SE-Net or SK-Net is followed by convolution units, where latent semantic information of each correspondence is destroyed. Thus, they only can employ the global average pooling to capture channel-wise information of feature maps. Therefore, for the *squeeze* operation, given the sth FG and the cth layer feature-map, we obtain the global contextual embedded channel-wise information gc_s^c as follows:

$$gc_s^c = \frac{1}{N}\sum_{i=1}^{N} FG_i^c, \tag{3}$$

Excitation: To exploit the information obtain by the squeeze operation in each feature-map group, we follow it with two MLPs to learn the weight of global contextual:

$$gc_s^{\prime c} = M(M^{\beta\gamma}(gc_s^c)), \tag{4}$$

where M denotes single MLP. $M^{\beta\gamma}$ is the learning strategy with Batch Normalization and ReLU function. $gc_s^{\prime c}$ is the learning result. After that, we use a sigmoid to enhance weight gc' and the final feature-map group V_s is:

$$V_s^c = \sum_{s=1}^{S} gc_s^{\prime c} \times FG_s^c, \tag{5}$$

Union: The *split* operation is used to divide feature-maps into several feature-map groups, and the *squeeze* and *excitation* operation are used to capture global contextual on feature-map groups. Then, to combine and aggregate the geometrical context information from multiple paths, we design an effective operation to unite information of all feature-map groups. SK-Net and ResNeSt adopt a *summation* operation to do that, but it is ill-suit for correspondences due to a mass of existing outliers. Thus, in this step, we adopt a *union* manner to aggregate all the feature-map groups. Formulation, we concatenate all the feature-map groups and get the final feature-maps:

$$V = Cat\{V_1, V_2, \ldots, V_S\}, \tag{6}$$

3.3. Network architecture

In this subsection, we describe the proposed PESA network, which consists two important blocks, i.e., the Permutation-Equivariant Split Attention (PESA) block and *Geometric Attention* block, as shown in Fig. 2.

3.3.1. PESA block

PESA-Block is based on our proposed SSEU module in Section 3.2, and it also contains two MLP layers to extract the correspondence features in canonical order of initial correspondences. The Context Normalization formula is described as following:

$$CN(f_i^l) = \frac{f_i^l - u^l}{o^l}, \tag{7}$$

where f_i^l is the output of the ith correspondence feature in the lth layer of MLP. o^l and u^l represent standard deviation and mean, respectively.

Then, the Batch Normalization and ReLU activation function are followed by Context Normalization. The PESA block is able to exploit the complex global context of sparse and unordered correspondence data.

3.3.2. Geometric Attention block

The *Geometric Attention* block includes three parts, i.e., Differentiable Pooling Layer [14], Order and Aware Filtering block, and Differentiable Unpooling layer [11]. Specifically, the Differentiable Pooling Layer first adopts a soft assignment matrix to map the input putative correspondences into a set of clusters. Next, the Order-Aware Filtering block exploits the cluster relation with spatially-correlated operations. At last, the Differentiable Unpooling layer is employed to assign per-correspondence predictions. Note that the *Geometric Attention* block is permutation-equivariant for the input putative correspondences.

We employ 6 PESA blocks with a *Geometric Attention* block to build a sub-network. The inputs of the first iterative sub-network are $N \times 4$ putative correspondences established by using nearest-neighbor matching of SIFT feature descriptor, usually $N = 2000$. The inputs of the second iterative sub-network are $N \times 6$, which contains the putative correspondences, residuals and weights: the putative correspondences ($N \times 4$) are the same as the input data of the first sub-network; the residuals and weights, which inherit from the first iterative sub-network, are the 5 and 6 rows, respectively.

3.3.3. Loss function

We construct a hybrid loss function by a classification loss and a geometry loss [10,38]:

$$L(D_q) = l_c(w_q, L_q) + \beta l_g(\hat{E}_q, E_q), \tag{8}$$

where $L(\cdot)$ represents the hybrid loss. $l_c(\cdot,\cdot)$ is a binary cross entropy loss. w_q and L_q denote the corresponding weight and label, respectively. $l_g(\cdot,\cdot)$ represents a geometry loss between the predicted essential matrix $\hat{E}_q$ and the ground truth essential matrix E_q generated by VisualSFM [39]. β denotes the weight to balance the classification loss and the geometry loss.

The binary cross entropy loss is computed as:

$$l_c(w_q, L_q) = \frac{1}{N_q} \sum_{i=1}^{N_q} \kappa_q^i H(S(w_q^i), L_q), \tag{9}$$

where $S(\cdot)$ represents the logistic function used with the binary cross-entropy H. κ_q^i denotes a self-adaptive weight to balance the positive/negative ratios.

The weakly supervised L_q is defined by a geometric distance as follows:

$$\begin{aligned} dist(d_q, E_q) &= \frac{(p_q'^T E_q p_q)^2}{\|\gamma_q\|_{[1]}^2 + \|\gamma_q\|_{[2]}^2 + \|\gamma_q'\|_{[1]}^2 + \|\gamma_q'\|_{[2]}^2}, \\ \gamma_q &= E_q p_q, \gamma_q' = E_q^T p_q', \end{aligned} \tag{10}$$

where p_q and p_q' indicate two keypoint positions from the putative correspondence d_q. $A_{[i]}$ represents the ith element of vector A. More specifically, we set the geometric distance less 10^{-4} as $L_q = 1$, and $L_q = 0$ otherwise. The weakly supervised label L_q for a correspondence is able to avoid prohibitively expensive annotation. For the geometry loss $l_g(\hat{E}_q, E_q)$, we utilize the geometric based distance to define it:

$$l_g(\hat{E}_q, E_q) = \sum_{q=1}^{Q} \frac{(p_q'^T \hat{E}_q p_q)^2}{\|\gamma_q\|_{[1]}^2 + \|\gamma_q\|_{[2]}^2 + \|\gamma_q'\|_{[1]}^2 + \|\gamma_q'\|_{[2]}^2}. \tag{11}$$

It is worth pointing out that, inspired by Iteratively Reweighed least-squares algorithm and DFE-Net [10], our network adopts an iterative manner to establish reliable correspondences. As shown in Fig. 2, two sub-networks are involved in our network, and the second iterative sub-network inherits the weights and residuals from the first sub-network.

4. Experimental results

In the section, we compare the proposed PESA-Net with a de facto standard of handcraft method (i.e., RANSAC [19]) and several state-of-the-art methods, including Point-Net++ [40], LGC-Net [9], DFE-Net [10], OA-Net++ [11] and ACNe-Net [12] on two tasks. Two publicly available datasets (i.e., YFCC100M dataset [13] and SUN3D dataset [15]) are employed both in the camera pose estimation and outlier rejection task. In the following: we first introduce the details of two datasets. After that, we expound the evaluation metrics on camera pose estimation and outlier rejection tasks, respectively, and then we discuss the implementation details and experiment results on two tasks. Finally, we analyze the ablation study.

4.1. Datasets

4.1.1. Outdoor scenes

We evaluate the performance of the seven methods on camera pose estimation using the Yahoo YFCC100M dataset [13]. YFCC100M dataset contains 100 million publicly accessible images from internet. J. Heinly et al. [41] made it into 72 tourist landmarks image collections for Structure from Motion (SfM). Following the previous work in [11], we separate the YFCC100M dataset into 68 sequences for training model and 4 sequences (i.e., Reichstag, Buckingham palace, Notre dame front facade and sacre coeur) as the unknown scenes for testing generalization ability.

4.1.2. Indoor scenes

For the indoor scene dataset, we use the SUN3D dataset [15], which is an RGBD video dataset captured with a kinect. Following [11], we keep a sample interval per 10 frames. Finally, we get 253 image sequences about different indoor scenes, where 15 sequences are used as unknown scenes for testing and the remaining 238 sequences for training model.

In this work, we evaluate the performance on camera pose estimation both in outdoor scenes (i.e., YFCC100M dataset) and indoor scenes (i.e., SUN3D dataset). More specifically, we test both known scenes and unknown scenes. For the known scenes, the training sequences are split into three subsets (i.e., training (60%), validation (20%), and testing (20%)). The unknown sequences are the testing sequences mentioned above.

4.2. Evaluation metrics

For camera pose estimation, our goal is obtaining an essential matrix which is to the greatest extent with the ground truth. We use the angular difference between the predicted vectors and ground truths for both rotation and translation as the error metric. Thus, the mean average precision (mAP) on both rotation and translation is employed for evaluating the result. Moreover, we adopt precision (P), recall (R), and F-measure (F) as the evaluation criteria for outlier rejection.

4.3. Implementation details

The optimizing strategy of our network is Adam [42] with the learning rate of 10^{-3}. Moreover, we set the batch size of input data as 32. The parameter β of the geometry loss is set as 0 during the first $20k$ iterations and then we set it to 0.1 in the rest $480k$. All experiments are performed on Linux 3.10.0 with NVIDIA TESLA $P100$ GPUs.

4.4. Camera pose estimation

We compare our method with five state-of-the-art methods, (i.e., Point-Net++ [40], LGC-Net [9], DFE-Net [10], OA-Net++ [11] and ACNe-Net [12]) on outdoor and indoor datasets. We also use RANSAC [19] as baseline. All these methods are trained under the same settings. For Point-Net++, we adopt 4D Euclidean space as the underlying metric space. DFE-Net is designed for fundamental matrix estimation, and we employ essential matrix instead of fundamental matrix to recover camera pose, by setting the regression target as essential matrix. LGC-Net and OA-Net++ (an improved version of OA-Net) are the official implementations. For ACNe-Net, we re-implemented based on PyTorch with the help of authors.

We report the results in Table 1, where the mAP at error thresholds 5°, 10° and 20° are reported on two datasets with known and unknown scenes. Following [11], we mainly analyze the error thresholds with 5°. From Table 1, we can see that our network consistently achieves the best results under all testing, showing improvements of 28.04% and 27.14% over the baseline LGC-Net on both outdoor known and unknown scenes, and we also achieve significantly improvements on

Table 1
Performance comparison for camera pose estimation on YFCC100M and SUN3D datasets.

Datasets	YFCC100M (%)						SUN3D (%)					
Methods	Known scene			Unknown scene			Known scene			Unknown scene		
	5°	10°	20°	5°	10°	20°	5°	10°	20°	5°	10°	20°
RANSAC	5.74	9.75	16.67	9.05	14.19	22.71	4.43	8.26	15.38	2.85	5.61	11.23
Point-Net++	11.88	20.47	32.86	15.98	28.00	44.82	8.78	17.01	31.02	7.22	16.10	29.77
LGC-Net	14.51	23.15	35.82	23.71	36.37	50.57	11.93	22.16	36.03	9.73	19.51	33.09
DFE-Net	19.27	29.25	42.14	30.55	43.95	59.15	14.18	24.35	39.14	12.13	21.78	36.26
ACNe-Net	29.63	40.42	52.71	34.00	48.46	62.98	19.08	30.96	46.32	14.27	24.74	39.29
OA-Net++	33.54	45.09	57.75	39.95	53.96	67.79	20.91	32.80	48.09	16.88	27.37	41.87
PESA-Net	**42.55**	**53.93**	**65.61**	**50.85**	**63.70**	**75.02**	**24.22**	**36.48**	**51.50**	**19.21**	**30.59**	**45.23**
Point-Net++[a]	34.09	44.80	57.13	46.73	57.02	67.83	20.31	30.49	43.97	15.97	24.51	36.18
LGC-Net[a]	34.27	44.56	56.27	46.93	57.38	68.11	20.64	30.76	44.11	16.04	24.74	36.92
DFE-Net[a]	36.93	47.34	59.25	51.38	60.99	70.97	21.44	31.56	44.92	16.35	25.28	37.85
ACNe-Net[a]	41.60	52.54	64.15	52.05	62.19	72.34	22.33	33.10	46.80	17.03	26.51	39.11
OA-Net++[a]	42.49	53.02	64.85	53.53	63.51	73.92	22.44	33.35	47.30	17.24	26.66	39.41
PESA-Net[a]	**46.03**	**56.69**	**68.13**	**55.38**	**65.95**	**76.01**	**23.42**	**34.15**	**48.20**	**18.32**	**28.04**	**41.03**

[a] Methods means using RANSAC [19] for post-processing.

Table 2
Comparative results of outlier rejection on the YFCC100M and SUN3D datasets.

Datasets	YFCC100M (%)						SUN3D (%)					
Methods	Known scene			Unknown scene			Known scene			Unknown scene		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
RANSAC	47.44	52.64	49.90	43.51	50.68	46.82	51.82	56.43	54.03	44.89	48.68	46.71
Point-Net++	49.84	86.41	63.22	46.60	84.17	59.99	51.40	86.73	64.54	44.79	83.23	58.23
LGC-Net	56.64	86.30	68.39	54.67	84.76	66.47	51.64	88.51	65.23	43.95	83.71	57.64
DFE-Net	56.61	87.04	68.60	53.93	85.52	66.15	53.88	87.20	66.61	46.21	84.07	59.64
ACNe-Net	59.99	88.81	71.61	55.54	85.38	67.30	54.03	88.45	67.08	45.97	83.94	59.40
OA-Net++	60.96	88.79	72.29	56.70	85.49	68.18	54.59	88.60	67.56	46.84	84.44	60.26
PESA-Net	**63.09**	**90.08**	**74.21**	**59.18**	**86.61**	**70.31**	**55.75**	**88.68**	**68.46**	**48.22**	**84.59**	**61.42**

indoor dataset. It is worth pointing out that OA-Net++ is the current work with stat-of-the-art performance and ACNe-Net is the latest network. Compared with the OA-Net++ and ACNe-Net, our network also outperforms them at least 9.01% and 10.9% mAP increasing in YFCC100M known and unknown scenes.

We also evaluate our network with RANSAC for post-processing, which is the most classical method for outlier rejection. As reported in Table 1, the post-processing strategy can boost the performance of most of competing methods. Note that, our network has not been improved so obviously as other competing networks. The reason behind this is that our network has a stronger ability to remove outliers, thus, the post-processing strategy only has slight boosting or even declines the performance. In particular, we also observe that RANSAC harms the performance on indoor scenes, which is an extremely challenging dataset. Its performance with our network on known and unknown scenes drops about 0.8% and 0.89% on SUN3D dataset. This is because, through PESA Block and *Geometric Attention* Block, our network is able to effectively infer the relative importance of each correspondence. However, RANSAC is often interested in the maximum co-consist set and may remove some critical inliers.

4.5. Outlier rejection

In this subsection, we test all competing methods on the task of outlier rejection. As the setting in pose estimation, we evaluate our network on two challenging datasets (i.e., YFCC100M and SUN3D) and as mentioned above, we employ precision (P), recall (R), and F-measure (F) as the evaluation criteria.

We report the quantitative results of different approaches in Table 2. PESA-Net clearly outperforms all existing methods on both two challenge datasets. Moreover, OA-Net++ and ACNe-Net surpass other three learning-based methods (i.e., Point-Net++, LGC-Net, DFE-Net) because they can capture more contextual information. Note that, our PESA-Net can extract latent spatial relations from different aspects. Therefore, it is able to perform better than OA-Net++ and ACNe-Net. We show some typical results of our network and other comparative methods in Fig. 4. Clearly, our network is able to achieve the best performance on several challenging scenes.

Table 3
Comparative results of our PESA-Net with different S on YFCC100M. The results of mAP (%) at error thresholds 5°, 10° and 20° on unknown scenes are reported. **#P** and **M** represent Parameters and Million, respectively.

PESA-Net	5°	10°	20°	#P	GFLOPs
$S = 1$	47.52	60.06	72.09	2.36M	1.49
$S = 2$	48.15	61.48	73.46	2.51M	1.79
$S = 4$	50.85	63.70	75.02	2.87M	2.40

4.6. Ablation study

4.6.1. Parameter analysis

The critical parameter of our method is the number of split feature-map groups, i.e., S. As shown in Fig. 3, the more split feature-map groups mean that they own stronger capacity to capture richer contextual information; while it also will cost more time. Here we test the performance of our network on the YFCC100M unknown scenes with different numbers of split feature-map groups i.e., $S = 1, 2$ and 4, and report the results in Table 3. We can see that, our network with $S = 4$ is able to improve at least 2% on the thresholds 5°, 10° and 20° over the version with $S = 1$. In addition, our network improves the mAP of the version with $S = 2$ from 48.15%, 61.48% and 73.46% to 50.85%, 63.7%, 75.02% on three thresholds, respectively. This can further show the effectiveness of using the SSEU module in our network.

Nevertheless, from Table 3, we can also see that our network inevitably introduces a slightly increase in parameter and computation with the number of split feature-map groups. However, we cannot ignore that the increased complexity does bring better accuracy. To balance the effectiveness and efficiency of our network, we set $S = 4$ in our experiment and do not adopt more split groups in our work.

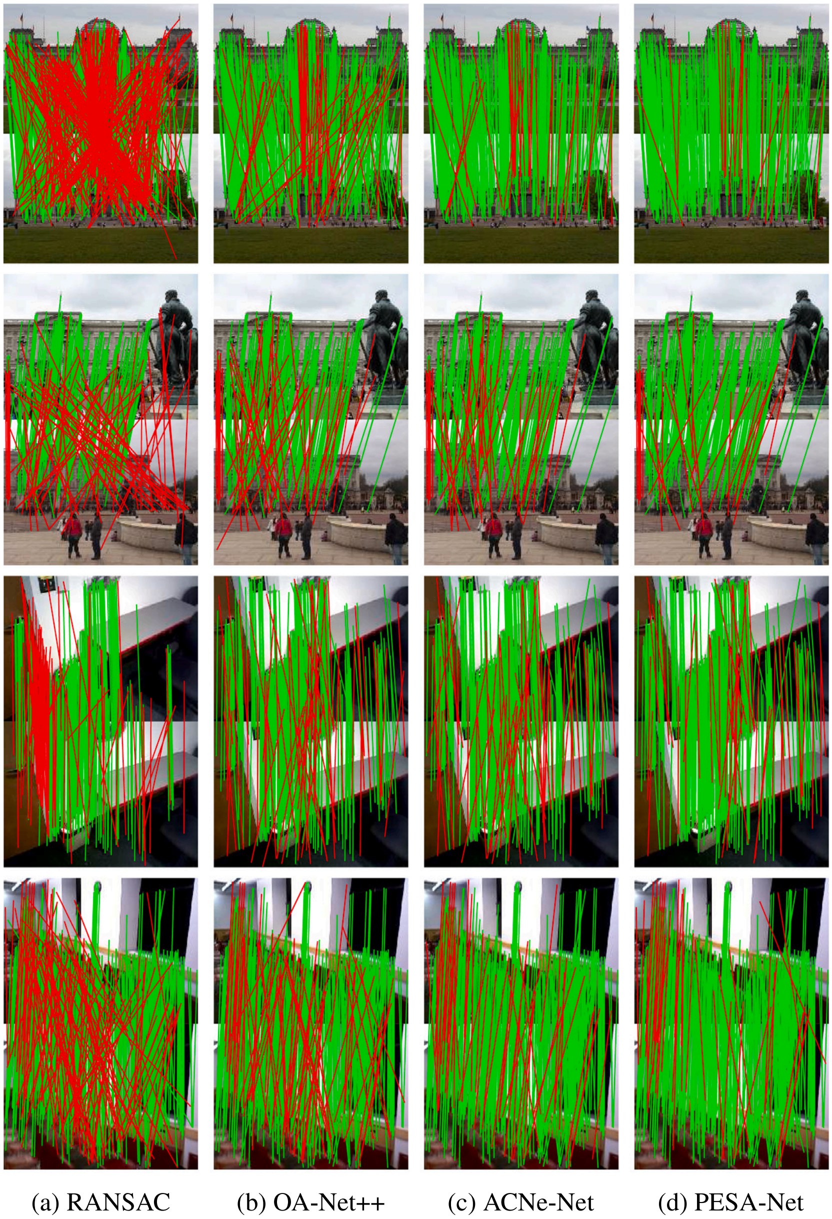
(a) RANSAC (b) OA-Net++ (c) ACNe-Net (d) PESA-Net

Fig. 4. Visualization results on two challenging datasets, i.e., YFCC100M dataset, SUN3D dataset. From top to bottom: Reichstag, Buckingham-palace and Te-harvard1. We draw the correspondences in green if they conform to the ground-truth epipolar geometry, and in red otherwise. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Ablation study on YFCC100M. The results of mAP (%) under error thresholds of 5° on both known and unknown scenes (**with/without** RANSAC post-processing) are reported. **Geo:** using the *Geometric Attention* Block. **Iter:** using the iterative network. **PE-S:** using PESA block with summation operation rather than PointCN blocks. **PE-U:** using PESA block with union operation rather than PointCN blocks.

PointCN	Geo	Iter	PE-S	PE-U	Known	Unknown
✓					14.51/34.27	23.71/46.93
✓	✓				25.18/40.36	32.36/51.22
✓	✓	✓			33.54/42.49	39.95/53.53
	✓	✓	✓		39.16/44.87	46.93/54.02
	✓	✓		✓	42.55/**46.03**	50.85/**55.38**

4.6.2. SSEU module analysis

In our SSEU module, we propose to use an *union* operation to gather geometrical context information from multiple paths. Note that, both of SK-Net [35] and ResNeSt [36] adopt a *summation* operation to do that. To show the effectiveness of our *union* operation, we test different versions of our network and show the results in Table 4. We can see that, when we adopt the *summation* operation, the network without using RANSAC achieves 46.93% with the mAP thresholds of 5° on the YFCC100M unknown scenes. In contrast, our network with the *union* operation is able to achieve 50.85%. Thus, the *union* operation can effectively improve the feature matching performance.

4.6.3. Camera pose estimation with learned features

Here, we use the state-of-the-art learned feature method, i.e., SuperPoint [18], to construct correspondences and report the result on two benchmarks (i.e., YFCC100M and SUN3D). Different from SIFT, SuperPoint is an end-to-end network for detecting keypoints and descriptions. We employ the pre-trained model provided by the author to gain keypoint positions and local descriptors and generate the putative correspondences by nearest-neighbor matching. The rest settings are set as the same as previous experiments in Section 4. As reported in Table 5, we observe that Superpoint gives better results on RANSAC or

Table 5
Performance comparison for camera pose estimation on YFCC100M and SUN3D datasets.

	Datasets	YFCC100M (%)		SUN3D (%)	
	Methods	Known	Unknown	Known	Unknown
SIFT	RANSAC	−/5.74	−/9.05	−/4.43	−/2.85
	LGC-Net	14.51/34.27	23.71/46.93	11.93/20.64	9.73/16.04
	ACNe-Net	29.63/41.60	34.00/52.05	19.08/22.33	14.27/17.03
	OA-Net++	33.54/42.49	39.95/53.53	20.91/22.44	16.88/17.24
	PESA-Net	42.55/**46.03**	50.85/**55.38**	**24.22**/23.42	**19.21**/18.32
SuperPoint	RANSAC	−/12.87	−/17.48	−/14.51	−/12.20
	LGC-Net	17.18/31.37	28.25/43.68	13.63/22.39	11.68/17.81
	ACNe-Net	28.90/35.43	34.08/46.73	20.72/23.66	14.83/18.36
	OA-Net++	29.66/35.19	35.35/45.53	20.12/23.55	15.88/18.73
	PESA-Net	35.70/**37.99**	43.58/**48.15**	21.21/**24.36**	16.33/**19.16**

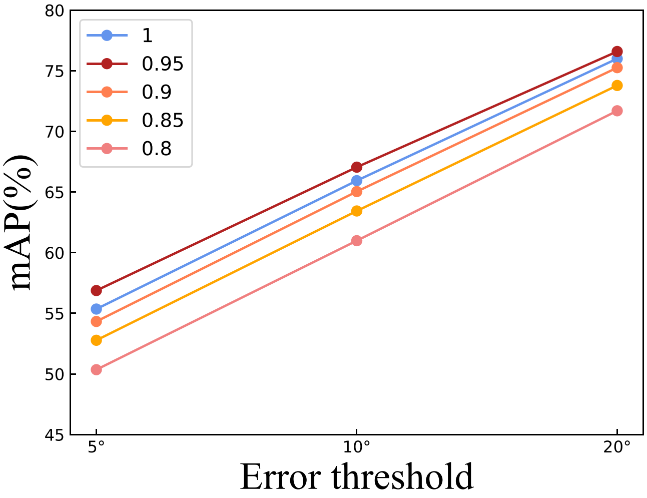

Fig. 5. The results of ratio test on YFCC100M unknown sequences under error thresholds of 5°, 10° and 20°.

LGC-Net and worse results than SIFT on our method. The main reason is that SuperPoint has better descriptors but is limited by the accuracy of keypoints. That is, SuperPoint can offer putative correspondence set with a higher inlier ratio. When inlier ratio large improvement, the key point array will become the main bottleneck. Nevertheless, for different local features, our method consistently achieves the best performance.

4.6.4. *Impact on the ratio test*

Lowe's ratio test can filter out non-discriminative matches, with a threshold $\in [0, 1]$. To test the impact of the threshold of ratio test, we set the ratio test threshold as 1, 0.95, 0.9, 0.85 and 0.8 in the PESA-Net with RANSAC, respectively. As reported in Fig. 5, the ratio test can boost the performance of PESA-Net when the ratio test threshold is suitable, degrade without. When the threshold is 0.95, PESA-Net is able to get the best performance.

4.6.5. *Impact on the post-processing*

Post-processing is a critical step in post estimation. In this subsection, we study RANSAC and its variants (i.e., DEGENSAC) as post-processing. We select three representative networks (i.e., LGC-Net, ACNe-Net, OA-Net++) and our network as the front-end network of RANSAC and its variants. In addition, we also adopt Lowe's ratio test to boost the performance of post estimation. As shown in Table 6, RANSAC can get superior performance as the post-processing when comparing with DEGENSAC.

Table 6
Performance of our network and baselines combining with different post-processing on the YFCC100M and SUN3D datasets. The mAP performance at error threshold 5° and 20° are reported. **R:** using RANSAC post-processing. **D:** using DEGENSAC post-processing.

Datasets	YFCC100M (%)		SUN3D (%)	
	5°	20°	5°	20°
Methods	R/D	R/D	R/D	R/D
LGC-Net	50.93/49.63	70.62/68.74	16.74/15.95	38.48/35.71
OA-Net++	54.08/54.63	74.75/73.76	17.71/17.12	39.89/37.50
ACN-Net	53.08/51.28	72.65/70.12	17.01/16.33	38.20/36.32
PESA-Net	**56.88**/**55.43**	**76.59**/**74.78**	**17.84**/**17.21**	**40.87**/**38.56**

Table 7
Comparison with Graph Neural Networks. The results of mAP (%) under error thresholds of 5° on both outdoor and indoor unknown scenes with 512 keypoints are reported.

Methods	YFCC100M	SUN3D	#P	GFLOPs
RANSAC	18.45	12.07	–	–
SuperGlue	43.17	16.09	12.02M	19.59
PESA-Net	**46.20**	**17.52**	2.87M	0.83

4.6.6. *Comparison with Graph Neural Networks*

Graph Neural Networks are also important methods to establish reliable correspondences. Here we compare our method with SuperGlue [29] which is one of the most popular Graph Neural Network for feature matching, and we also run RANSAC as a baseline. For SuperGlue, we employ the official pre-trained model of SuperGlue to test the performance, since the authors have not provided a training code. To have a fair comparison, we employ the pre-trained model train on Section 4.6.3 to evaluate the performance of PESA-Net. Following [29], we achieve the pose by estimating the essential matrix with OpenCV's findEssentialMat and RANSAC for post-processing. Additionally, SuperGlue gives suboptimal results on 2000 SuperPoint keypoints, thus, we detect 512 keypoints per image. From Table 7, we can see that our network not only uses significantly fewer parameters, but also works well with Graph Neural Networks under the same settings.

5. Conclusion

In this paper, we have designed a novel SSEU module, which is used to build a Permutation-Equivariant Split Attention Network (PESA-Net) for correspondence learning. The proposed SSEU module is able to gather rich contextual information from different aspects in a permutation invariant manner, by generating multiple paths and adopting channel-wise dependence. In addition, we also construct a permutation-equivariant block by fusing the SSEU module, Multi-Layer Perceptron and some normalizations, to solve problems on permutation-equivariant correspondence data. Our experiments have shown that PESA-Net achieves significant improvement over existing approaches for addressing the camera pose estimation and outlier removal tasks on outdoor and indoor datasets.

In addition, PESA-Net adopts the iterative manner since it can extremely boost the performance for outlier rejection. Nevertheless, we find that a large amount of information in the previous iteration is not fully exploited, and only the last iteration result is used as the predicted weight. Therefore, for future research directions, we consider improving the iterative network to comprehensively utilize all the information of iterations.

CRediT authorship contribution statement

Zhen Zhong: Conceptualization, Methodology, Software, Writing – original draft. **Guobao Xiao:** Methodology, Writing – reviewing and editing, Supervision. **Shiping Wang:** Writing – reviewing and editing. **Leyi Wei:** Visualization, Investigation. **Xiaoqin Zhang:** Writing – reviewing and editing.

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 by the National Natural Science Foundation of China under Grants 62072223, and supported by the Natural Science Foundation of Fujian Province under Grant 2020J01131199.

References

[1] 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.

[2] 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.

[3] K.T. Ahmed, S. Ummesafi, A. Iqbal, Content based image retrieval using image features information fusion, Inf. Fusion 51 (2019) 76–99.

[4] Y. Liu, S. Liu, Z. Wang, Multi-focus image fusion with dense SIFT, Inf. Fusion 23 (2015) 139–155.

[5] X. Yuan, J. Zhang, B.P. Buckles, Evolution strategies based image registration via feature matching, Inf. Fusion 5 (4) (2004) 269–282.

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

[7] J.L. Schonberger, J.-M. Frahm, Structure-from-motion revisited, in: CVPR, 2016, pp. 4104–4113.

[8] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.

[9] K. Moo Yi, E. Trulls, Y. Ono, V. Lepetit, M. Salzmann, P. Fua, Learning to find good correspondences, in: CVPR, 2018, pp. 2666–2674.

[10] R. Ranftl, V. Koltun, Deep fundamental matrix estimation, in: ECCV, 2018, pp. 284–299.

[11] 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: ICCV, 2019, pp. 5845–5854.

[12] W. Sun, W. Jiang, E. Trulls, A. Tagliasacchi, K.M. Yi, ACNe: Attentive context normalization for robust permutation-equivariant learning, in: CVPR, 2020, pp. 11286–11295.

[13] 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, 2015, arxiv preprint arXiv:1503.01817.

[14] Z. Ying, J. You, C. Morris, X. Ren, W. Hamilton, J. Leskovec, Hierarchical graph representation learning with differentiable pooling, in: NIPS, 2018, pp. 4800–4810.

[15] J. Xiao, A. Owens, A. Torralba, Sun3d: A database of big spaces reconstructed using sfm and object labels, in: ICCV, 2013, pp. 1625–1632.

[16] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, ORB: An efficient alternative to SIFT or SURF, in: ICCV, 2011, pp. 2564–2571.

[17] K.M. Yi, E. Trulls, V. Lepetit, P. Fua, Lift: Learned invariant feature transform, in: ECCV, 2016, pp. 467–483.

[18] D. DeTone, T. Malisiewicz, A. Rabinovich, Superpoint: Self-supervised interest point detection and description, in: CVPR, 2018, pp. 224–236.

[19] 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.

[20] P.H. Torr, A. Zisserman, MLESAC: A new robust estimator with application to estimating image geometry, Comput. Vis. Image Underst. 78 (1) (2000) 138–156.

[21] O. Chum, J. Matas, Matching with PROSAC-progressive sample consensus, in: CVPR, 2005, pp. 220–226.

[22] R. Raguram, O. Chum, M. Pollefeys, J. Matas, J.-M. Frahm, USAC: A universal framework for random sample consensus, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2012) 2022–2038.

[23] O. Chum, T. Werner, J. Matas, Two-view geometry estimation unaffected by a dominant plane, in: CVPR, Vol. 1, 2005, pp. 772–779.

[24] D. Barath, J. Matas, J. Noskova, MAGSAC: marginalizing sample consensus, in: CVPR, 2019, pp. 10197–10205.

[25] J. Ma, J. Zhao, J. Jiang, H. Zhou, X. Guo, Locality preserving matching, Int. J. Comput. Vis. 127 (5) (2019) 512–531.

[26] J. Ma, J. Jiang, H. Zhou, J. Zhao, X. Guo, Guided locality preserving feature matching for remote sensing image registration, IEEE Trans. Geosci. Remote Sens. 56 (8) (2018) 4435–4447.

[27] J. Ma, J. Zhao, J. Tian, A.L. Yuille, Z. Tu, Robust point matching via vector field consensus, IEEE Trans. Image Process. 23 (4) (2014) 1706–1721.

[28] Z. Zhang, W.S. Lee, Deep graphical feature learning for the feature matching problem, in: ICCV, 2019, pp. 5087–5096.

[29] P.-E. Sarlin, D. DeTone, T. Malisiewicz, A. Rabinovich, Superglue: Learning feature matching with graph neural networks, in: CVPR, 2020, pp. 4938–4947.

[30] C. Zhao, Z. Cao, C. Li, X. Li, J. Yang, NM-Net: Mining reliable neighbors for robust feature correspondences, in: CVPR, 2019, pp. 215–224.

[31] V. Mnih, N. Heess, A. Graves, et al., Recurrent models of visual attention, in: NIPS, 2014, pp. 2204–2212.

[32] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, 2014, arXiv preprint arXiv:1409.0473.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: NIPS, 2017, pp. 5998–6008.

[34] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: CVPR, 2018, pp. 7132–7141.

[35] X. Li, W. Wang, X. Hu, J. Yang, Selective kernel networks, in: CVPR, 2019, pp. 510–519.

[36] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, Z. Zhang, H. Lin, Y. Sun, T. He, J. Mueller, R. Manmatha, et al., Resnest: Split-attention networks, 2020, arXiv preprint arXiv:2004.08955.

[37] K.T. Ahmed, S. Ummesafi, A. Iqbal, Content based image retrieval using image features information fusion, Inf. Fusion 51 (2019) 76–99.

[38] R. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge university press, 2003.

[39] C. Wu, Towards linear-time incremental structure from motion, in: 3DV, 2013, pp. 127–134.

[40] C.R. Qi, L. Yi, H. Su, L.J. Guibas, Pointnet++: Deep hierarchical feature learning on point sets in a metric space, in: NIPS, 2017, pp. 5099–5108.

[41] J. Heinly, J.L. Schonberger, E. Dunn, J.-M. Frahm, Reconstructing the world* in six days*(as captured by the yahoo 100 million image dataset), in: CVPR, 2015, pp. 3287–3295.

[42] P. Adam, G. Sam, C. Soumith, C. Gregory, Y. Edward, D. Zachary, L. Zeming, D. Alban, A. Luca, L. Adam, Automatic differentiation in PyTorch, in: NIPS, 2017.