Neurocomputing 431 (2021) 137–147

Contents lists available at ScienceDirect

Neurocomputing

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

SCSA-Net: Presentation of two-view reliable correspondence learning via spatial-channel self-attention

Xin Liu [a,b], Guobao Xiao [b,*], Luanyuan Dai [a], Kun Zeng [b], Changcai Yang [a], Riqing Chen [a,*]

[a] Digital Fujian Research Institute of Big Data for Agriculture and Forestry, College of Computer and Information Science, Fujian Agriculture and Forestry University, Fuzhou 350002, China

[b] Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, College of Computer and Control Engineering, Minjiang University, Fuzhou 350108, China

ARTICLE INFO

Article history:
Received 7 August 2020
Revised 29 October 2020
Accepted 10 December 2020
Available online 23 December 2020
Communicated by Zidong Wang

Keywords:
Correspondences
Attention modules
Outlier removal
Pose estimation

ABSTRACT

Seeking reliable correspondences between pairwise images is non-trivial in feature matching. In this paper, we propose a novel network, called the Spatial-Channel Self-Attention Network (SCSA-Net), to capture abundant contextual information of correspondences for obtaining reliable correspondences and estimating accurate camera pose of the matching images. In our proposed SCSA-Net, we introduce two types of attention modules, *i.e.*, the spatial attention module and the channel attention module. The two types of attention modules are able to capture complex global context of the feature maps by selectively aggregating mutual information in the spatial dimension and channel dimension, respectively. Meanwhile, we combine the outputs of two modules to generate rich global context and obtain feature maps with strong representative ability. Our SCSA-Net is able to effectively remove outliers, and simultaneously estimate accurate camera pose between pairwise images. These reliable correspondences and camera pose are vital for many computer vision tasks, such as SfM, SLAM and stereo matching. The tremendous experiments on outlier removal and pose estimate tasks have shown the better performance improvements of our SCSA-Net over current state-of-the-art methods on both outdoor and indoor datasets. Especially, our SCSA-Net outperforms the recent state-of-the-art OANet++ by 5.55% mAP5°on unknown outdoor datasets. Code is available at https://github.com/x-gb/SCSA-Net.

1. Introduction

Feature matching between pairwise images is a fundamental problem in computer vision, such as Structure from Motion (SfM) [1], visual Simultaneous Localization and Mapping (SLAM) [2] and stereo matching [3]. In [4], the related knowledge on feature matching is surveyed for the past two decades. Generally, feature matching contains three steps, *i.e.*, extracting and describing a number of feature points, estimating the initial correspondence set according to the similarity constraint of feature points and removing false correspondences (*i.e.*, outliers). The initial correspondence set usually contains a large number of outliers due to the existence of large scale variations, illumination changes, occlusions and blurs. Therefore, the outlier removal is provided as a key post-processing step to seek reliable correspondences (*i.e.*, inliers) from the initial correspondence set. The inliers have great effect on estimating accurate camera pose between two images. In this paper, we mainly focus on the research of outlier removal.

Recently, some learning-based methods [5–10] are developed as well as shown remarkable performance on outlier removal. Using Convolutional Neural Network (CNN) to process the correspondences is usually impractical, since the correspondences are unordered and irregular. Therefore, LFGC [5] proposes a PointNet-like architecture [11] to infer the probability of each correspondence as an inlier, and it primarily utilizes Multi-Layer Perceptrons (MLPs) to process each individual correspondence. LFGC also introduces a simple non-parametric Context Normalization (CN) over the entire correspondence set to capture global context. ACNe [6] develops the learning-based Attentive Context Normalization (ACN) to obtain useful context. However, the non-parametric normalized operation only utilizes the mean and variance of the correspondences, and processes each correspondence indiscriminately. Other learning-based methods [7–10] also equivalently treat each correspondence. However, this indiscriminate operation may limit the performance of networks, especially when outliers are dominant in the initial correspondence set.

In order to address the above problem, we propose an effective outlier removal network, called SCSA-Net, to discriminatively process each correspondence and channel map by selectively

[*] Corresponding authors.
E-mail addresses: x-gb@163.com (G. Xiao), riqing.chen@fafu.edu.cn (R. Chen).

https://doi.org/10.1016/j.neucom.2020.12.052
0925-2312/© 2020 Elsevier B.V. All rights reserved.

aggregating mutual information according to the information similarity in a global manner. That is, the feature maps with similar information can obtain mutual gains regardless of their positions. Therefore, our network is able to capture rich global contextual information and enhance the representation capacity of inliers and important channel maps because their information is usually similar, which is vital for network learning. Specifically, the overview of our SCSA-Net is shown in Fig. 1. Our SCSA-Net is able to infer the probability of each correspondence being an inlier and estimate the accurate camera pose of two matching images according to reserved inliers. SCSA-Net includes three primary parts, *i.e.*, the PointCN Block, the DiffPool & DiffUnpool Layer and the Spatial-Channel Self-Attention Block (SCSA Block). Specifically, the PointCN Block is proposed to process the unordered and irregular correspondences by LFGC, which is a pioneering work for learning-based outlier removal methods. The DiffPool & DiffUnpool Layer is proposed to capture local and global contexts by OANet [9].

The Self-attention mechanism [12] has attracted tremendous attention due to its powerful capacity for capturing rich contextual information over the global scope, which is essential for the learning of a deep network. It obtains the satisfying performance improvements in many fields [13–17]. Therefore, we extend the self-attention mechanism to the learning-based outlier removal network and introduce two types of attention modules (*i.e.*, the spatial attention module and channel attention module) for capturing rich contextual information in the spatial and channel dimensions, respectively. The spatial attention module is used to obtain global context of each correspondence. Specifically, we firstly utilize the dot-product operation to get a spatial attention matrix, which indicates the spatial similarity between any two correspondences. The values of spatial attention matrix represent the similarity weight of each two correspondences. Then we use a weighted sum operation to aggregate global spatial context of all correspondences. Therefore, any two correspondences with similar feature are able to obtain mutual gains regardless of their spatial positions. It enables the module to pay more attention to the potential inliers, since inliers usually have similar feature information and outliers usually are irregular. The channel attention module is used to capture global context of each channel. We use the similar operation with spatial attention module. This module is able to pay more attention to some important channels, which are useful for the outlier removal. In the end, we concatenate the outputs of these two modules along the channel dimension to obtain a new feature map with strong representative ability.

Furthermore, we add the SCSA Block to different positions of SCSA-Net, which has different types of feature maps. The first SCSA Block is used to capture the global context of N correspondences. In the DiffPool & DiffUnpool Layer, we map the N correspondences to M clusters for capturing local context. Therefore, the second SCSA Block is used to capture the global context of M clusters. We use the two SCSA Blocks to capture more useful contextual information of different types of feature maps for promoting their representative ability. In addition, we iteratively use the SCSA-Net twice, the outputs of the first one and relevant residual information as additional inputs of the second one. The results of the second one are able to achieve the large performance improvements than other state-of-the-art methods for the tasks of outlier removal and pose estimation. The quantitative and visual results well demonstrate the effectiveness of our SCSA-Net for seeking reliable correspondences.

The three main contributions of this paper are summarized as follows: Firstly, we propose the Spatial-Channel Self-Attention Block to capture rich global context by selectively aggregating mutual information in the spatial dimension and channel dimension, respectively. The two types of modules cooperatively enhance the representative ability of feature maps. As far as existing literature is concerned, we are pilot to utilize these two attention modules for the learning-based outlier removal task. Secondly, we add the SCSA Block to different positions of the SCSA-Net, where they have different types of features. This operation is able to capture useful global context of different feature maps and enhance their representative ability. Thirdly, we iteratively use the SCSA-Net twice, to obtain better performance improvements for the outlier removal task and the camera pose estimation task on different challenging datasets.

The rest of the paper is structured as follows: In Section 2, we briefly introduce the relevant work in the fields of outlier removal and attention mechanism. We introduce the learning-based SCSA-Net and describe the two types of attention modules in Section 3. In Section 4, we exhibit the performance improvements of our SCSA-Net compared with the state-of-the-art outlier removal methods. In the end, we give the discussion and conclusion in Section 5 and Section 6, respectively.

2. Related work

In this section, we briefly introduce some traditional and learning-based outlier removal methods, followed by the popular attention mechanism.

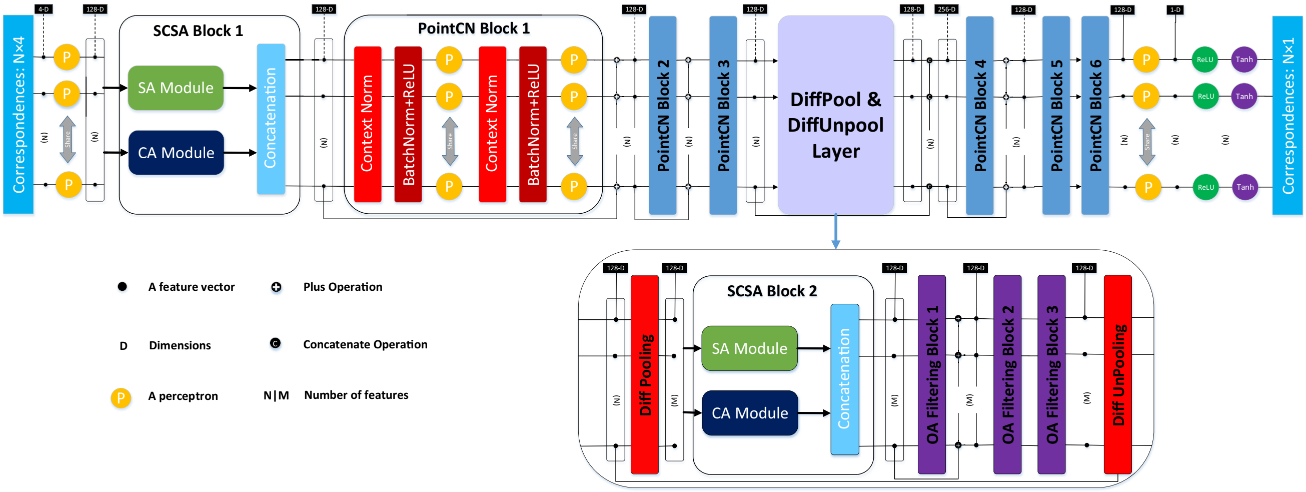


Fig. 1. The architecture of our SCSA-Net.

2.1. Traditional outlier removal

For feature matching of pairwise images, we firstly utilize some off-the-shelf feature methods to extract their keypoints and descriptors, such as SIFT [18] or newer LIFT [19]. Then, the initial correspondence set is obtained by the similarity constraint of descriptors. However, the initial correspondence set usually contains extensive false correspondences due to the ambiguities of descriptor matching. Therefore, outlier removal becomes an important step for feature matching. The most well-known resampling-based outlier removal methods are RANSAC [20] and its modified variants [21–24]. These methods iteratively utilize the hypothesize-and-verify strategy to get a subset with the largest number of inliers. They have shown great influence on the robust model estimation. However, the performance of these methods depends on the sampled subsets of inliers. Recently, some consistency-based methods [25–27] address this problem according to geometric attribute of correspondences. Such as, GMS [25] removes outliers by a grid-based motion statistical manner. LPM [26] overcomes the difficulty by using local domain consistency of inliers. These methods usually have a good performance with low computational complexity. However, if the initial correspondence set contains a large number of outliers, the performance of preceding traditional methods cannot perform well.

2.2. Learning-based outlier removal

In recent years, deep learning has attracted extensive attention in many fields [28–31], owning to its powerful learning and representation capacity. In Learning-based methods, some feature learning methods [19,32–34] use convolutional neural network (CNN) to obtain more reliable keypoints and descriptors. Although these methods have shown better results than handcrafted method [18], they cannot resolve the problem that initial correspondence set has a large amount of outliers. Therefore, the learning-based outlier removal method is proposed as a post-processing step to solve the problem. LFGC [5] is the first one that uses an end-to-end learning-based manner to label each correspondence as an inlier or an outlier. Meanwhile, it uses weighted eight-point algorithm to estimate an essential matrix for recovering the camera pose of two matching images. LFGC utilizes the MLPs architecture to process irregular correspondences and obtains remarkable performance. OA-Net [9] introduces cluster-based pooling and unpooling operations to extract useful local context and the order-aware filtering block to get crucial global context. ACNe [6] utilizes the attentive context normalization, which contains two types of attention normalization to capture local and global context of correspondences. The [10] proposes an improved method to establish robust neighborhood context by maximizing local neighborhood consensus, which also is vital for other works [35–37]. These learning-based outlier removal methods have achieved the significant performance improvements over traditional methods for seeking reliable correspondences. However, these methods equally process each correspondence. Different from these methods, our SCSA-Net is able to capture global context by selectively aggregating mutual information according to the information similarity and boost the representation capacity of inliers and important channel maps.

2.3. Attention mechanism

The attention mechanism has widely utilized by many researchers in various tasks [12–17], because it enables the network to pay attention to key parts of inputs. The work [12] firstly proposes the self-attention mechanism to capture the reliable semantic context of input sequences, and it obtains impressive performance based on the attention mechanism in the natural language processing. Based on the proposed attention mechanism, SENet [13] presents an effective squeeze-and-excitation block to learn channel attention and achieves significant performance improvements. CCNet [14] proposes an efficient criss-cross attention block for obtaining contextual information to process semantic segmentation task in a high computational efficiency. DANet [16] effectively utilizes two types of attention modules for capturing rich contextual information to further improve the feature representation in the scene segmentation. These researches show the remarkable performance of attention mechanism for processing different tasks.

In this paper, we extend the self-attention mechanism to the task of learning-based outlier removal and introduce two types of attention modules for capturing rich contextual information of feature maps in a global manner. The quantitative and visual results well demonstrate the effectiveness of our method.

3. Method

In the section, we first introduce the problem formulation in Section 3.1. Then we describe the proposed spatial attention module and channel spatial module in Sections 3.2 and 3.3, respectively. After that, the network framework and loss function are introduced in Sections 3.4 and 3.5, respectively.

3.1. Problem formulation

In order to obtain reliable correspondences and accurate camera pose of the given two images (I, I′), we first use the handcrafted SIFT [18] to find their keypoints and corresponding descriptors. Then, we obtain an initial correspondence set C according to the nearest-neighbor similarity constraint of descriptors:

$$C = \{c_1, c_2, \ldots, c_N\} \in \mathbb{R}^{N\times 4},\ \ c_i = \left(x_1^i, y_1^i, x_2^i, y_2^i\right) \tag{1}$$

where N is the number of correspondences and $\left(x_1^i, y_1^i\right)$ and $\left(x_2^i, y_2^i\right)$ are the normalized coordinates of keypoints in the given two images.

Our network takes the initial correspondence set C as the input, and obtain the corresponding probability set $w = \{w_1, w_2, \ldots, w_N\}$ with $w_i \in [0, 1)$, which indicates the likelihood of correspondence c_i being an inlier, *e.g.*, $w_i = 0$ indicates the correspondence c_i is an outlier. Therefore, we formulate our learning-based outlier removal network as:

$$w = \tanh(ReLU(o)), o = f_\theta(C) \tag{2}$$

where o is the logit values of network output. *ReLU* and *tanh* are two activation functions to obtain the probability set w. $f_\theta(.)$ is a function that formulates our network, and θ is relevant network parameters.

When we have obtained the probability set w, we can utilize a weighted eight-point algorithm, proposed by LFGC, to estimate the accurate essential matrix. The method only considers the inliers and avoids the effect of outliers. Therefore, we can obtain a more accurate essential matrix to estimate the camera pose than traditional eight-point algorithm, which only considers all correspondences. We formulate the method as:

$$\widehat{E} = g(C, w) \tag{3}$$

where $g(.)$ is a function of weighted eight-point algorithm. The initial correspondence set C and the probability set w as the inputs. The output $\widehat{E}$ is our estimated essential matrix.

3.2. Spatial attention module

In this section, we introduce our spatial attention module. As we known, the contextual information is particularly crucial for network learning, and it indicates the geometrical relevance of feature maps. The self-attention mechanism is proposed by Vshish [12], which is used for relating different positions of a sequence to capture long-range contextual information. Inspired by the self-attention mechanism, we design a spatial attention module to capture the rich contextual information of each correspondence. The spatial attention module is able to capture the similarity of any two correspondences, and focuses on the potential inliers, since their geometrical relevance is usually consistent. Thus, we can utilize spatial attention module for capturing the global context of each correspondence to boost the representation capability of potential inliers and inhibit the representation capability of potential outliers.

As shown in Fig. 2, for a feature map $F \in \mathbb{R}^{C\times N\times 1}$(where C is the number of channels, and N is the number of correspondences). We firstly obtain three feature maps $Q, K, V \in \mathbb{R}^{C\times N}$ by using three different PointCN modules (described in Section 3.4) and reshape operations. Then we get the spatial similarity matrix by performing a dot-product operation between the transpose of Q and V, and obtain the spatial attention matrix $SA \in \mathbb{R}^{N\times N}$by using a softmax function. The spatial attention matrix indicates similarity weight of any two correspondences regardless of their positions. Therefore, the two correspondences with similar information will be obtained mutual gains. The acquisition of the spatial attention matrix is formulated as follows:

$$SA = Softmax\left(D\left(Q^T, K\right)\right) \tag{4}$$

where $D(.)$ is the dot-product operation, and $Softmax(.)$ is the softmax function. Finally, we perform a dot-product operation between V and the transpose of SA to aggregate global spatial context of all correspondences and reshape the result to $\mathbb{R}^{C\times N\times 1}$. We formulate a series of operations as:

$$F_s = SAtt(Q, K, V) = D\left(V, SA^T\right) \tag{5}$$

where $SAtt(.)$ is the formulation of our spatial attention module. $F_s \in \mathbb{R}^{C\times N\times 1}$ is the output of spatial attention module. It integrates global context of each correspondence and boosts representation capability for potential inliers.

Noteworthily, we add the spatial attention module to two different positions in our SCSA-Net, where they have different types and numbers of features as shown in Fig. 1. Therefore, the size of obtaining spatial attention matrix is different, but their operations are the same.

3.3. Channel attention module

In this section, we introduce the channel attention module and the information fusion operation between two modules. Channel maps can be treated as the multi-spatial responses, and each channel map represents a class-specific spatial response. Some relative important channels usually have similar spatial response, and they are beneficial for the outlier removal. Therefore, in order to explore similarity of any two channels and enhance the representation capability of important channel maps, we propose a channel attention module to capture rich global context for each channel map. Our channel attention module is able to focus on some relative important channel maps. Thus, we can utilize channel attention module for capturing global context of each channel to enhance the representation capability of some important channel maps.

We illustrate the channel attention module in Fig. 3. Different from the spatial attention module, for a feature map $F \in \mathbb{R}^{C\times N\times 1}$, we directly reshape F to $\mathbb{R}^{C\times N}$ and perform a dot-product operation between F and the transpose of F to obtain channel similarity matrix, then we use a softmax function to obtain the channel attention matrix $CA \in \mathbb{R}^{C\times C}$. The channel attention matrix indicates similarity weight of any two channels. Therefore, the two channels with similar information will be obtained mutual gains. The acquisition of channel attention matrix is formulated as follows:

$$CA = Softmax\left(D\left(F, F^T\right)\right) \tag{6}$$

where $D(.)$ is the dot-product operation, and $Softmax(.)$ is the softmax function. Finally, we perform a dot-product operation between CA and F to aggregate global channel context of all channels and reshape the result to $\mathbb{R}^{C\times N\times 1}$. We formulate a series of operations as:

$$F_c = CAtt(F) = D(CA, F) \tag{7}$$

where $CAtt(.)$ is the formulation of our channel attention module. $F_c \in \mathbb{R}^{C\times N\times 1}$ is the output of channel attention module. It integrates global context of each channel map and boosts representation capability for some important channel maps.

Concatenation: The spatial and channel attention modules are able to capture spatial and channel global contexts of the feature map. For obtaining more strong feature maps, we combine the outputs of two attention modules. Here, we adapt the concatenate operation between the outputs of the two modules along their channel dimensions. Then we use a PointCN module to process the concatenate feature map for recovering C channels. The fusion operation has formulated as follows:

$$F' = PointCN(Concat(F_s, F_c)) \tag{8}$$

where $F' \in \mathbb{R}^{C\times N\times 1}$ is the output of information fusion operation. $PointCN(.)$ and $Concat(.)$ are the PointCN module and the concatenation operation, respectively. Then we use an element-wise sum operation between F' and the input feature map F of SCSA module. We formulate this process as follows:

$$F_{out} = F + \alpha F' \tag{9}$$

where F_{out} is the output feature map of SCSA module. α is a learned hyper-parameter, which is initialized as 0 and gradually learns to be a more suitable weight.

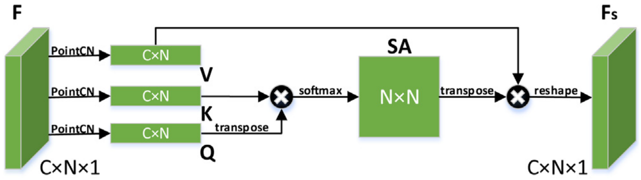


Fig. 2. Spatial attention module.

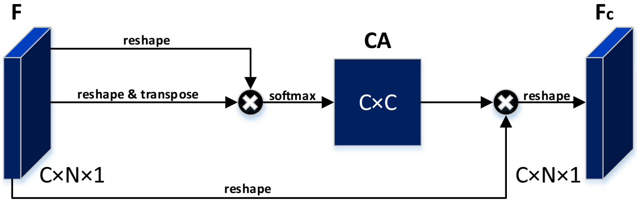


Fig. 3. Channel attention module.

3.4. Network framework

This section, we introduce our network framework as shown in Fig. 1. It mainly contains three key parts, *i.e.*, the PointCN Block, the DiffPool & DiffUnpool Layer and the Spatial-Channel Self-Attention (SCSA) Block. The PointCN Block is proposed by LFGC, and it consists of two same sequential modules, *i.e.*, a Context Normalization layer to capture global context, a Batch Normalization layer with a ReLU activation function to accelerate the training of network, a Multi-Layer Perceptrons layer with 128 neurons for network learning. We call this sequential module as a PointCN module. We utilize 6 PointCN Blocks to deal with unordered and irregular correspondences. The DiffPool & DiffUnpool Layer is proposed by OA-Net to capture the local and global context. It firstly uses a differentiable pooling layer to map the N correspondences to M clusters, *i.e.*, the feature map $F_{row} \in \mathbb{R}^{C\times N\times 1}$ becomes $F_{pool} \in \mathbb{R}^{C\times M\times 1}$ ($N > M, e.g., N = 2000, M = 500$). The differentiable pooling layer contains a PointCN module and a softmax layer. Then it proposes an Order-Aware Filtering Block to capture global context of clusters. The block contains three PointCN modules, and it transposes spatial dimension and channel dimension of the feature map in the middle of the PointCN module. We use 3 Order-Aware Filtering Blocks to deal with these clusters. Finally, we utilize a differentiable unpooling layer to recover initial spatial size of the feature map, and the differentiable unpooling layer also contains a PointCN module and a softmax layer. Note that the unpooling matrix is calculated by the input feature map of DiffPool & DiffUnpool layer, *i.e.*, F_{row}, because it must keep the order information of correspondences.

Our Spatial-Channel Self-Attention (SCSA) Block is used for capturing rich global context of feature maps and selectively aggregating mutual information. These two detailed structures are introduced in Sections 3.2 and 3.3, respectively. We use two SCSA Blocks in front of PointCN Block1 and Order-Aware Filtering Block1 to capture the global context of correspondences and their clusters, respectively. In the end of our SCSA-Net, we use the activation function of *ReLU* and *tanh* to get the probability of each correspondence being an inlier.

3.5. Loss function

Following OANet [9], we adopt a hybrid loss function to supervise our network training. The first loss function is a classification loss, and it is used to reject the outliers. The second loss function is an essential matrix loss, and it is used to predict the essential matrix. Specifically, the loss function is formulated as follows:

$$L = L_c(w, Y) + \beta L_e\left(E, \hat{E}\right) \tag{10}$$

where β is a weight parameter to balance the two loss function terms. $L_c(w, Y)$ is the classification loss, and $L_c(.)$ is a binary cross entropy loss function; w is the predicted probability value of our network; Y is weakly supervised label of each correspondence, and it is calculated by using a geometric distance constraint; We formulate the geometric distance constraint as follows:

$$d(c, E) = \frac{\left(p'^T E p\right)^2}{\|Ep\|_{[1]}^2 + \|Ep\|_{[2]}^2 + \|E^T p'\|_{[1]}^2 + \|E^T p'\|_{[2]}^2} \tag{11}$$

where E is the ground-truth essential matrix. p and p' are two keypoint coordinates forming a correspondence c, *i.e.*, $c = \left(p^T, p'^T\right)^T$. $v_{[k]}$ denotes the kth entry of vector v. We use a threshold of 10^{-4} to determine the label of each correspondence, *i.e.*, the geometric distance of a correspondence under 10^{-4} is an inlier.

For the essential matrix loss, we use a geometry loss based on above geometric distance constraint. Therefore, the essential matrix loss is formulated as:

$$L_e\left(E, \hat{E}\right) = \sum_{i=1}^{N_{in}} \frac{\left(p_i'^T \hat{E} p_i\right)^2}{\|Ep_i\|_{[1]}^2 + \|Ep_i\|_{[2]}^2 + \|E^T p_i'\|_{[1]}^2 + \|E^T p_i'\|_{[2]}^2} \tag{12}$$

where p_i and p_i' are keypoint coordinates from ground-truth inlier set, and i is the index of inliers and N_{in} is the number of ground-truth inliers. $\hat{E}$ is our predicted essential matrix according to the weighted eight-point algorithm.

4. Experiments

In this section, we test the effectiveness of our proposed SCSA-Net for handling the tasks of outlier removal and camera pose estimation on both outdoor and indoor scenes. We compare our SCSA-Net with the recent state-of-the-art methods. We choose RANSAN [20], PointNet++ [38], DFE [39], LFGC [5], ACNe [6], OANet [9] as comparative state-of-the-art methods. All these methods are under the same settings and the experimental environment to train and test. In the following, in Section 4.1, we firstly introduce the two types of datasets, *i.e.*, the outdoor and indoor scenes. Then, we detailedly show our experimental and comparative results on outlier removal task and pose estimation task in Section 4.2. Finally, the implementation details and ablation studies are introduced in Sections 4.3 and 4.4, respectively.

4.1. Datasets

Following OANet, we select both outdoor and indoor scenes as main datasets for network training and testing.

Outdoor scenes: For outdoor scenes, we mainly select Yahoo's YFCC100M dataset [40], which is a collection of 100 million tourist photos from Internet. We divide these tourist photos into 71 image sequences. We select 67 sequences to train network and the remaining 4 sequences as unknown scenes to test the generalization ability of network.

Indoor scenes: For indoor scenes, we mainly select SUN3D dataset [41], which is an RGBD video dataset and the relative ground-truth information is calculated by generalized bundle adjustment. We divide the indoor scenes into 254 sequences, and select 239 sequences for training network and 15 sequences as unknown scenes for testing generalization ability of network. These indoor scenes have little distinctive features, many repetitive structures and extensive occlusions, so the tasks of outlier removal and pose estimation are greatly challenging in this scenes.

We use both known and unknown scenes to test generalization ability of the network. We select the aforementioned training sequences as known scenes. Specifically, we split the training sequences into some disjoint subset, *i.e.*, training (60%), validation (20%) and testing (20%). Noteworthily, the performance of network usually depends on the unknown scenes. The known scenes are only as a reference.

4.2. Evaluation metrics and results

This section, we mainly introduce our experimental evaluation metrics and compared results with other state-of-the-art methods for outlier removal task and camera pose estimation task on different challenging datasets. We select several state-of-the-art methods as comparisons, including RANSAC [20], PointNet++ [38], DFE [39], LFGC [5], ACNe [6], OANet [9] and its iterative variation OANet++. The first one method is the most well-known traditional method, and the last six methods are the learning-based methods.

RANSAC is a classical resampling method, and it has widely applied in various fields, *e.g.*, computer vision and pattern recognition. PointNet++ is an improved version of PointNet [11], which is a pioneer in point cloud classification and segmentation by using deep learning. DFE is a robust approach to estimate fundamental matrix from noisy data. LFGC is the first one of deep learning-based method for outlier removal. ACNe is a simple yet effective technique to remove outliers, and it introduces two types of attention to capture local and global contexts. OANet is a recent learning-based work with state-of-the-art performance, and it is able to capture useful local and global contexts of correspondences. To be fair, all methods are trained and tested on the same setting and experimental environment.

Outlier removal: For explicitly reflecting the performance of methods for outlier removal task, we adopt the *Precision* (P), *Recall* (R) and *F-score* (F) as the evaluation metrics. Specifically, the *Precision* is the ratio between the identified correct match number and the preserved match number. The *Recall* is the ratio between the identified correct match number and the all correct match number in the initial correspondence set. The *F-score* is calculated by $2 * Precision * Recall/(Precision + Recall)$. For simplicity, we select the unknown scenes as the testing sequences of outlier removal on outdoor and indoor scenes.

Quantitative compared results are shown in Table 1. These learning-based methods significantly outperform traditional RANSAC method on the two datasets, since the traditional RANSAC method only suits the specific constraints and scenarios, *e.g.*, higher ratio inliers in the initial correspondence set. For the challenging datasets (the inlier ratio in the initial correspondence set is often about 10%, *e.g.*, our outdoor and indoor scenes), these learning-based methods are able to obtain better performance on outlier removal task. In learning-based methods, our SCSA-Net achieves best results of *Precision* and *F-score* on the two datasets, but the *Recall* is slightly lower than some methods. In particularly, our method accomplishes *Precision* improvements of 3.07% and 1.19% than OANet++ on both outdoor and indoor datasets, respectively. For *F-score*, we also achieve better improvements than other methods. Fig. 4 also shows some visualization results of our method and two learning-based methods (LFGC and OANet++). We can see that our method is able to obtain the best performance on the two challenging datasets, which usually have large viewpoint variations, large illumination changes, occlusions and blurs. At the top left corner of each image pair, we show the information of inliers (the ratio between the identified correct match number and the preserved match number, *i.e.*, *precision*(P)) and *recall*(R), and our method is able to obtain better results than other methods.

The quantitative and visual results well demonstrate the effectiveness of our proposed SCSA-Net for processing the outlier removal task.

Pose estimation: For explicitly reflecting the performance of methods for camera pose estimation task, we define the angular differences between ground truth and our predicted rotation and translation vectors, which are calculated by our predicted essential matrix. We select the mean average precision (mAP) of the angular differences under different error thresholds as evaluation metrics. Here, we mainly use the mAP under 5° error as our default evaluation metrics (*i.e.*, mAP5°) because it is more vital for some computer vision applications. The camera pose estimation is a challenging task, since it requires reliable correspondences to recover the essential matrix between the matching images. We also select the same comparative state-of-the-art methods with outlier removal task. Furthermore, for learning-based methods, we adopt RANSAC as a post-processing step to help them obtain more accurate pose estimation results. We use both known and unknown scenes to test generalization ability of network for camera pose estimation task on the different datasets.

The quantitative comparative results are reported in Table 2. We can see that our method significantly outperforms these comparative state-of-the-art methods. For unknown scenes, our method without RANSAC as the post-processing step obtains the mAP5° improvements of 5.55% and 1.27% than the second best OANet++ on both outdoor and indoor datasets. When using RANSAC as the post-processing step, our method still improves the mAP5° by 3.29% and 0.23% over the OANet++ on both outdoor and indoor datasets. For known scenes, our method also achieves the better performance improvements over these state-of-the-art methods on the outdoor and indoor scenes. For better showing the performance improvements of our SCSA-Net with other methods, Figs. 5 and 6 illustrate the improvements of mAP with different error thresholds (mAP5°, mAP10°, mAP15°, mAP20°) on both outdoor and indoor scenes. We can see that our SCSA-Net has achieved significant improvements against LFGC and OANet++ on different error thresholds.

Comprehensive benchmark evaluations well demonstrate the effectiveness of our SCSA-Net for addressing the pose estimation task. Noteworthily, the performance of outlier removal and pose estimation is a significant positive correlation as shown in Tables 1 and 2.

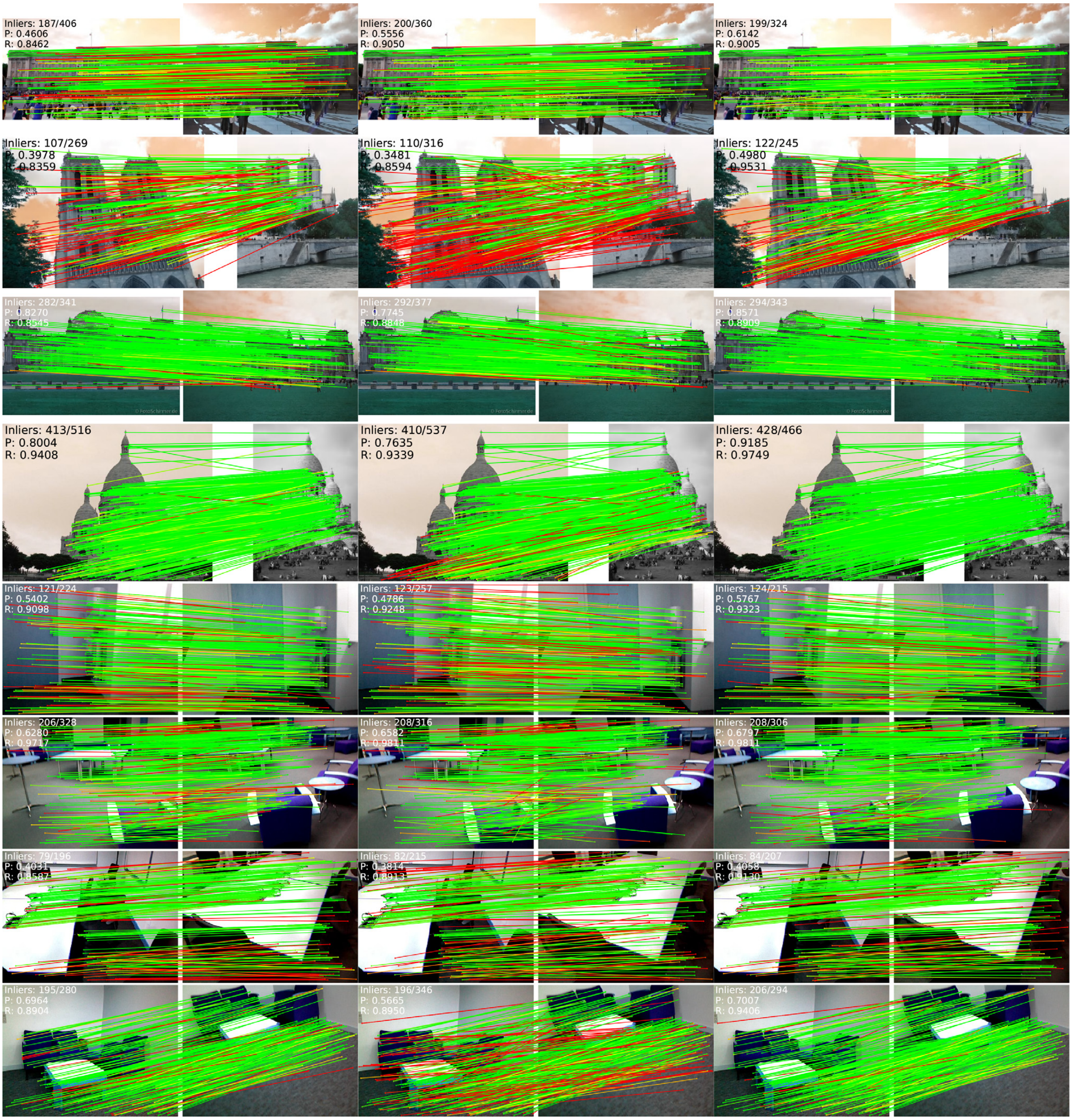
4.3. Implementation details

The architecture of network is shown in Fig. 1, and it contains three primary parts as discussed in Section 3.4. These three primary parts all have 128 channels. We take the $N \times 4$ initial correspondences as the input of network, typically $N = 2000$. In DiffPool & DiffUnpool layer, we map the N correspondences to M clusters, typically $M = 500$. Finally the 1-D logit value is processed by the activation functions of *ReLU* and *tanh* to get 1-D probability values as the output of network. All the implementation of network is by using Pytorch. We adopt the Adam optimizer with the learning rate of 10^{-3} to optimize network parameters. The size of batch is 32 and

Table 1
Quantitative comparative results of outlier removal on the YFC100M (left) and SUN3D (right) datasets.

Method	Outdoor(%)			Indoor(%)		
	P	R	F	P	R	F
RANSAC	41.83	57.08	48.28	44.11	46.42	45.24
PointNet++	48.42	82.93	61.54	45.64	83.43	59.00
DFE	51.68	83.49	63.84	44.09	84.00	57.82
LFGC	53.12	85.51	65.53	47.24	83.45	60.32
ACNe	54.56	86.92	67.04	46.44	84.23	59.87
OANet	55.65	85.80	67.51	46.54	83.43	59.74
OANet++	54.55	86.67	66.96	46.95	83.77	60.17
SCSA-Net	**57.62**	85.79	**68.93**	**48.14**	83.52	**61.08**

The bold values indicate the best relusts of our method compared with other methods.



(a) LFGC (b) OANet++ (c) Our

Fig. 4. Visualization results using (a) LFGC [5] (left), (b) OANet++ [9] (middle) and (c) our method (right). The top four images come from unknown test set of the YFCC100M dataset and the rest four images come from unknown test set of the SUN3D dataset. We exhibit the more correct matches (green lines) and false matches (red lines) according to the geometric distance constraint. Meanwhile, we show the information of precision and recall at the top left corner of each image pair.

the number of training iterations is 500k. The weight parameter β is initialized as 0 and then to 0.5 after 20 k training iterations. Furthermore, we iteratively use the SCSA-Net twice, the outputs of the first one and relevant residual information as additional inputs of the second one.

4.4. Ablation studies

In this section, we provide some ablation studies about the proposed SCSA module and the network architecture on the YFCC100M dataset.

To demonstrate the effectiveness of our SCSA-Net, we test the different combinations about the SCSA module. As shown in Table 3, we test the performance of different combinations for camera pose estimation task on the YFCC100M dataset. The first row of the table is OANet++, and it is the recent state-of-the-art learning-based outlier removal method. Therefore, we take OANet++ as the comparative baseline. We can see that the performance of our all iterative combinations outperforms the baseline. Specifically, the spatial attention module (SCSA1 + SCSA2 + SA + Iter) achieves an improvement of 3.15% over the baseline on unknown scenes when without RANSAC. The channel attention

Table 2
Quantitative comparison results of camera pose estimation on the YFC100M (left) and SUN3D (right) datasets. The mAP5° with/without RANSAC as a post-processing step are reported.

mAP5°	Outdoor(%)		Indoor(%)	
	Known	Unknown	Known	Unknown
RANSAC	5.82/–	9.08/–	4.38/–	2.86/–
PointNet++	34.69/11.49	45.85/15.75	21.00/11.80	18.79/10.29
DFE	35.17/12.52	49.80/21.78	20.34/10.08	15.68/08.81
LFGC	37.19/16.77	49.93/26.13	20.85/13.62	16.35/11.96
ACNe	39.08/25.55	51.62/35.40	21.08/13.44	16.40/11.62
OANet	41.40/31.00	51.45/35.07	22.29/19.22	16.95/13.69
OANet++	42.06/34.04	51.65/38.95	22.45/21.44	17.48/16.38
SCSA-Net	**43.67/35.66**	**54.94/44.50**	**22.98/22.93**	**17.71/17.65**

The bold values indicate the best relusts of our method compared with other methods.

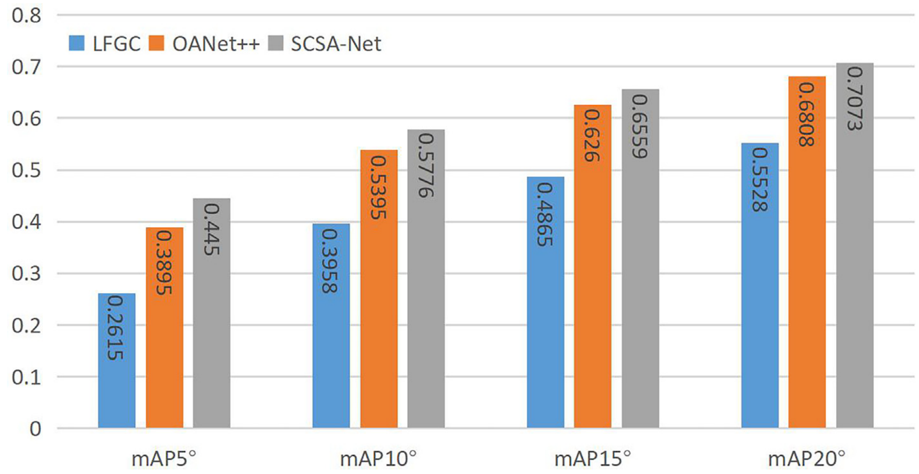


Fig. 5. Visualization result comparisons of mAP with different error thresholds on the unknown outdoor datasets.

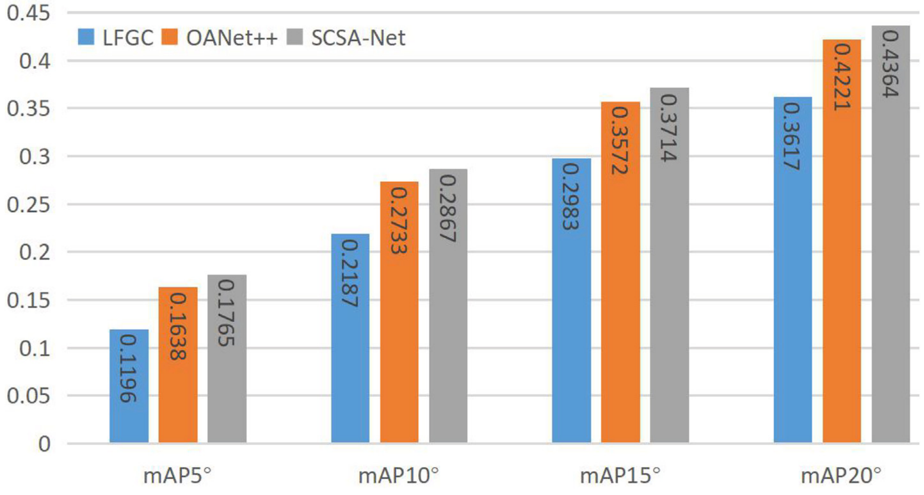


Fig. 6. Visualization result comparisons of mAP with different error thresholds on the unknown indoor datasets.

module (SCSA1 + SCSA2 + CA + Iter) obtains the 2.45% improvement over the baseline on unknown scenes when without RANSAC. The two terms indicate the spatial attention module and channel attention module are able to capture rich contextual information of feature maps to promote the performance of network for the camera pose estimation task. For only using the SCSA Block1 (SCSA1 + SA + CA + Iter), we obtain a significantly promotion of performance over the baseline, achieving an improvement of 3.63% over baseline on unknown scenes without RANSAC as a post-processing step. For only using the SCSA Block2 (SCSA2 + SA + CA + Iter), we also obtain 1.87% improvement over the baseline on unknown scenes when without RANSAC as a post-processing step. The two terms indicate that adding the SCSA block to different positions of the SCSA-Net is effective.

We iteratively use the SCSA-Net twice to obtain better performance improvements. The iterative network is able to largely improve the performance of mAP5° from 29.18% to 44.50% without RANSAC as a post-processing step on the unknown scenes.

Table 3
Ablation studies about the proposed modules on YFCC100M. mAP5° (%) on both known and unknown scenes are reported with/without RANSAC as a post-processing step. **SCSA1**: using the SCSA1 module. **SCSA2**: using the SCSA2 module. **SA**: using the spatial attention module in SCSA modules. **CA**: using the channel attention module in SCSA modules. **iter**: using the iterative network.

SCSA1	SCSA2	SA	CA	Iter	Known	UnKnown
				✓	42.06/34.04	51.65/38.95
✓	✓	✓		✓	43.59/36.16	53.97/42.10
✓	✓		✓	✓	42.37/33.64	53.93/41.40
✓		✓	✓	✓	44.00/36.18	54.13/42.58
	✓	✓	✓	✓	43.01/34.94	52.95/40.82
✓	✓	✓	✓		38.70/23.67	50.73/29.18
✓	✓	✓	✓	✓	43.67/35.66	**54.94/44.50**

The bold values indicate the best relusts of our method compared with other methods.

Table 4
Comparison of OANet++ and our SCSA-Net in terms of network parameters (Param), floating point operations per second (FLOPs) and mAP5° on the unknown outdoor scenes.

Method	Param(M)	FLOPs(G)	mAP5° (%)
OANet++	2.47	1.81	38.95
Ours	2.68	2.08	44.50

Although it increases the double amount of parameters, the iterative network achieves a larger performance improvement. Furthermore, the results in Table 3 demonstrate our proposed SCSA-Net (SCSA1 + SCSA2 + SA + CA + Iter) is able to achieve the best performance improvement over the baseline. We obtain the mAP5° improvement of 5.55% on the unknown scenes when without RANSAC. As shown in Table 4, although the computational cost (*e.g.*, Param and FLOPs) of our network is slightly more than baseline due to the introduction of attention mechanism, our network achieves the impressive improvement (*e.g.*, mAP5°) in terms of effectiveness.

5. Discussion

The main advantages of our SCSA-Net are summarized as follows: First of all, previous methods equivalently treat each correspondence, which leads to sub-optimal results, especially when the initial correspondence set contains a large number of outliers. This indiscriminate operation may limit the performance of network for seeking reliable correspondences. Therefore, we propose a spatial attention module based on the attention mechanism to discriminatively process each correspondence. It is able to capture the similarity of any two correspondences according to their relevance, and focus on potential inliers, because the relevance of inliers usually is consistent. Besides, we propose a channel attention module based on the attention mechanism to enhance the representation capacity of important channels. Finally, we combine the outputs of two modules to obtain the enhanced feature maps, which have rich global contextual information and strong representation ability. Therefore, our SCSA-Net is able to capture rich global context and enhance the representation capacity of feature for boosting the network learning.

Our proposed SCSA Block is general. With the block, we can not only effectively process the computer vision task, but also apply to other fields (*e.g.*, economics [42] or diagnosis [43]). However, we use the attention mechanism in a global manner, the computational cost may be not friendly. Therefore, in our future work, we will pay attention to design an efficient and effective module to solve this problem.

6. Conclusions

In this paper, we propose the Spatial-Channel Attention Network (SCSA-Net) for two-view reliable correspondence learning. Specifically, We utilize the spatial attention module and the channel attention module, which enable our network to capture more useful global contextual information in the spatial dimension and channel dimension, respectively. Then, we combine the output of two types of attention modules to obtain the feature maps with strong representation ability. Different from state-of-the-art methods, our SCSA-Net based on the attention mechanism is able to selectively aggregate mutual information of feature maps, so we can capture rich global context and enhance the representation capacity of important feature maps. The quantitative and visual results well demonstrate the effectiveness of our method for addressing the outlier removal and pose estimation tasks on different challenging datasets. Notably, our SCSA-Net achieves state-of-the-art mAP5° of 44.50% and 17.65% on unknown outdoor and indoor datasets.

CRediT authorship contribution statement

Xin Liu: Conceptualization, Investigation, Methodology, Software, Validation, Writing - original draft, Writing - review & editing. **Guobao Xiao:** Funding acquisition, Supervision, Validation, Writing - original draft, Writing - review & editing. **Luanyuan Dai:** Software, Visualization. **Kun Zeng:** Writing - review & editing. **Changcai Yang:** Writing - review & editing. **Riqing Chen:** Funding acquisition, Formal analysis, Validation, Writing - original draft, Writing - review & 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.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grants 61972093, 62072223, 61702431, and by Natural Science Foundation of Fujian Province under Grants 2020J01829.

References

[1] M. Havlena, K. Schindler, Vocmatch: Efficient multiview correspondence for structure from motion, in: European Conference on Computer Vision, 2014, pp. 46–60.
[2] R. Mur-Artal, J.M.M. Montiel, J.D. Tardos, Orb-slam: a versatile and accurate monocular slam system, IEEE Transactions on Robotics 31 (5) (2015) 1147–1163.
[3] H. Hirschmuller, Stereo processing by semiglobal matching and mutual information, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2) (2007) 328–341.
[4] J. Ma, X. Jiang, A. Fan, J. Jiang, J. Yan, Image matching from handcrafted to deep features: a survey, International Journal of Computer Vision (2020) 1–57.

[5] K. Moo Yi, E. Trulls, Y. Ono, V. Lepetit, M. Salzmann, P. Fua, Learning to find good correspondences, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2666–2674.

[6] W. Sun, W. Jiang, E. Trulls, A. Tagliasacchi, K.M. Yi, Acne: Attentive context normalization for robust permutation-equivariant learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11286–11295.

[7] J. Ma, X. Jiang, J. Jiang, J. Zhao, X. Guo, Lmr: Learning a two-class classifier for mismatch removal, IEEE Transactions on Image Processing 28 (8) (2019) 4045–4059.

[8] C. Zhao, Z. Cao, C. Li, X. Li, J. Yang, Nm-net: Mining reliable neighbors for robust feature correspondences, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 215–224.

[9] J. Zhang, D. Sun, Z. Luo, A. Yao, L. Zhou, T. Shen, Y. Chen, L. Quan, H. Liao, Learning two-view correspondences and geometry using order-aware network, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 5845–5854.

[10] Y. Wang, X. Mei, Y. Ma, J. Huang, F. Fan, J. Ma, Learning to find reliable correspondences with local neighborhood consensus, Neurocomputing 406 (2020) 150–158.

[11] C.R. Qi, H. Su, K. Mo, L.J. Guibas, Pointnet: Deep learning on point sets for 3d classification and segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 652–660.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[13] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[14] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, W. Liu, Ccnet: Criss-cross attention for semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 603–612.

[15] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, Eca-net: Efficient channel attention for deep convolutional neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11534–11542.

[16] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3146–3154.

[17] J. Ma, H. Zhang, P. Yi, Z.-Y. Wang, Scscn: A separated channel-spatial convolution net with attention for single-view reconstruction, IEEE Transactions on Industrial Electronics 67 (2019) 8649–8658.

[18] D.G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.

[19] K.M. Yi, E. Trulls, V. Lepetit, P. Fua, Lift: Learned invariant feature transform, in: European Conference on Computer Vision, 2016, pp. 467–483.

[20] M.A. Fischler, R.C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Communications of the ACM 24 (6) (1981) 381–395.

[21] P.H. Torr, A. Zisserman, Mlesac: A new robust estimator with application to estimating image geometry, Computer Vision and Image Understanding 78 (1) (2000) 138–156.

[22] R. Raguram, O. Chum, M. Pollefeys, J. Matas, J.-M. Frahm, Usac: a universal framework for random sample consensus, IEEE transactions on Pattern Analysis and Machine Intelligence 35 (8) (2012) 2022–2038.

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

[24] R. Li, J. Sun, D. Gong, Y. Zhu, H. Li, Y. Zhang, Arsac: Efficient model estimation via adaptively ranked sample consensus, Neurocomputing 328 (2019) 88–96.

[25] J. Bian, W.-Y. Lin, Y. Matsushita, S.-K. Yeung, T.-D. Nguyen, M.-M. Cheng, Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4181–4190.

[26] J. Ma, J. Zhao, J. Jiang, H. Zhou, X. Guo, Locality preserving matching, International Journal of Computer Vision 127 (5) (2019) 512–531.

[27] Y. Liu, Y. Li, L. Dai, C. Yang, L. Wei, T. Lai, R. Chen, Robust feature matching via advanced neighborhood topology consensus, Neurocomputing (2020) 1, https://doi.org/10.1016/j.neucom.2020.09.047.

[28] S. Hassantabar, M. Ahmadi, A. Sharifi, Diagnosis and detection of infected tissue of covid-19 patients based on lung x-ray image using convolutional neural network approaches, Chaos, Solitons & Fractals 140 (2020) 110170.

[29] S. Hassantabar, N. Stefano, V. Ghanakota, A. Ferrari, G.N. Nicola, R. Bruno, I.R. Marino, N.K. Jha, Coviddeep: Sars-cov-2/covid-19 test based on wearable medical sensors and efficient neural networks, ArXiv Preprint ArXiv:2007.10497..

[30] S. Hassantabar, X. Dai, N.K. Jha, Steerage: Synthesis of neural networks using architecture search and grow-and-prune methods, ArXiv Preprint ArXiv:1912.05831..

[31] S. Hassantabar, Z. Wang, N.K. Jha, Scann: Synthesis of compact and accurate neural networks, ArXiv Preprint ArXiv:1904.09090..

[32] D. DeTone, T. Malisiewicz, A. Rabinovich, Superpoint: Self-supervised interest point detection and description, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 224–236.

[33] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, T. Sattler, D2-net: A trainable cnn for joint description and detection of local features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 8092–8101.

[34] J. Wu, Z. Gan, W. Guo, X. Yang, A. Lin, A fully convolutional network feature descriptor: Application to left ventricle motion estimation based on graph matching in short-axis mri, Neurocomputing 392 (2020) 196–208.

[35] C. Yan, Z. Li, Y. Zhang, Y. Liu, X. Ji, Y. Zhang, Depth image denoising using nuclear norm and learning graph model, ArXiv Preprint ArXiv:2008.03741..

[36] C. Yan, B. Shao, H. Zhao, R. Ning, Y. Zhang, F. Xu, 3d room layout estimation from a single rgb image, IEEE Transactions on Multimedia (2020) 1, https://doi.org/10.1109/TMM.2020.2967645.

[37] C. Yan, B. Gong, Y. Wei, Y. Gao, Deep multi-view enhancement hashing for image retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) 1, https://doi.org/10.1109/TPAMI.2020.2975798.

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

[39] R. Ranftl, V. Koltun, Deep fundamental matrix estimation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 284–299.

[40] 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, Communications of the ACM 59 (2) (2016) 64–73.

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

[42] M. Ahmadi, S. Jafarzadeh-Ghoushchi, R. Taghizadeh, A. Sharifi, Presentation of a new hybrid approach for forecasting economic growth using artificial intelligence approaches, Neural Computing and Applications 31 (12) (2019) 8661–8680.

[43] S. Dorosti, S.J. Ghoushchi, E. Sobhrakhshankhah, M. Ahmadi, A. Sharifi, Application of gene expression programming and sensitivity analyses in analyzing effective parameters in gastric cancer tumor size and location, Soft Computing 24 (13) (2020) 9943–9964.

Xin Liu received the B.S. degree in internet of things engineering form ZhouKou Normal University, Zhoukou, China, in 2019. He is currently a Master student with the School of computer and information, Fujian Agriculture and Forestry University, Fuzhou, China. His current research interests include computer vision and image matching.

Guobao Xiao 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 2016–2018, he was a Postdoctoral Fellow in the School of Aerospace Engineering at Xiamen University, China. He is currently a Full Professor at Minjiang University, China. He has published over 30 papers in the international journals and conferences including TPAMI/TIP/TITS/TGRS/TIE, IJCV, PR, ICCV, ECCV, ACCV, AAAI, ICIP, ICARCV, etc. His research interests include machine learning, computer vision, pattern recognition and bioinformatics. He has been awarded the best PhD thesis in Fujian Province and the best PhD thesis award in China Society of Image and Graphics (a total of ten winners in China). He also served on the program committee (PC) of CVPR, ICCV, ECCV, AAAI, IJCAI, etc. He was the General Chair for IEEE BDCLOUD 2019.

Luanyuan Dai received the B.S. degree in Geophysics from Northeast Petroleum University, Daqing, China in 2018. She is currently studying the master's degree with the School of computer and information, Fujian Agriculture and Forestry University, Fuzhou, China. Her current research interests mainly include computer vision and image matching.

Kun Zeng received Ph. D degrees from Department of Computer Science, Xiamen University, Xiamen, China, in 2015, where he was a Postdoctoral Fellow with the Department of Electronic Science, from 2016 to 2019. He is currently a Lecturer with Minjiang University, China. His current research interests include image processing, machine learning and medical image reconstruction.

Changcai Yang received M.S. degree in control theory and control engineering from China Three Gorges University, China, in 2008, and his Ph.D. in pattern recognition and intelligent systems at Huazhong University of Science and Technology (HUST), China, in 2012. From 2012 to 2014, he worked as Post Doctor at HUST, China. Since 2016, he has been an Associate Professor and MS Supervisor with the College of Computer and Information Science, Fujian Agriculture and Forestry University, Fuzhou, China. He has authored or coauthored more than 40 papers. His research interests include computer vision, image processing, point set registration.

Riqing Chen received the B.Eng. degree of communication engineering from Tongji University, China, in 2001, the M.Sc. degree of communications and signal processing from Imperial College London, U.K., in 2004, and the D.Phil. degree of engineering science from the University of Oxford, U.K., in 2010. Since 2014, he has been affiliated with 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 etc.