# MSA-Net: Establishing Reliable Correspondences by Multiscale Attention Network

Linxin Zheng, Guobao Xiao, *Member, IEEE*, Ziwei Shi, Shiping Wang, *Member, IEEE*, and Jiayi Ma, *Senior Member, IEEE*

***Abstract*— In this paper, we propose a novel multi-scale attention based network (called MSA-Net) for feature matching problems. Current deep networks based feature matching methods suffer from limited effectiveness and robustness when applied to different scenarios, due to random distributions of outliers and insufficient information learning. To address this issue, we propose a multi-scale attention block to enhance the robustness to outliers, for improving the representational ability of the feature map. In addition, we also design a novel context channel refine block and a context spatial refine block to mine the information context with less parameters along channel and spatial dimensions, respectively. The proposed MSA-Net is able to effectively infer the probability of correspondences being inliers with less parameters. Extensive experiments on outlier removal and relative pose estimation have shown the performance improvements of our network over current state-of-the-art methods with less parameters on both outdoor and indoor datasets. Notably, our proposed network achieves an 11.7% improvement at error threshold 5° without RANSAC than the state-of-the-art method on relative pose estimation task when trained on YFCC100M dataset.**

***Index Terms*— Outlier removal, deep learning, wide-baseline stereo.**

## I. Introduction

THE feature matching method, which aims to establish feature correspondences between two groups of feature points [1], [2], acts as a premise to solve a series of tasks in the computer vision area, such as Simultaneous Localization and Mapping [3], Panoramic Stitching [4], Structure from Motion [5], [6], and Stereo Matching [7]. As shown in Fig. 1, a feature matching method involves three steps, i.e. obtaining keypoints and the corresponding descriptors, establishing the initial correspondence set, and removing outliers (*i.e.*, false correspondences) [8], [9]. Specifically, the initial features are generated by off-the-shelf methods, such as SIFT [10] or SuperPoint [11]. Then, we use the nearest neighbor matching method to generate the initial correspondence set. But unfortunately, above methods are far from satisfactory, which is inevitable to generate the outliers due to the possible challenging image variants, *e.g.*, light changes, repetitive structures, blurs, occlusions and lack of texture. Therefore, outlier removal plays an important role in the feature matching problem.

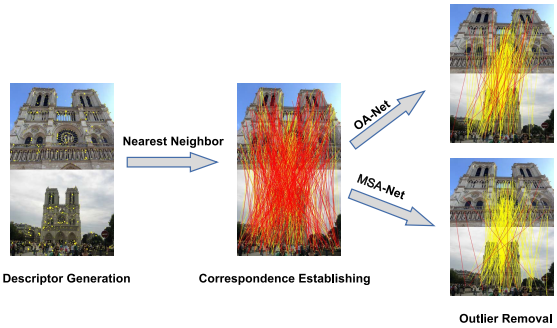



Fig. 1. Establishing reliable correspondences via MSA-Net. Given a pair of images, we first obtain the keypoints and descriptors from the images and establish the initial correspondences by nearest-neighbor matching method. At last, we use MSA-Net and OA-Net to infer the probability of correspondences being inliers and remove the outliers, respectively. It explicitly finds that MSA-Net removes more outliers than OA-Net.

Outlier removal methods have been developed for many years and gradually developed into two factions, *i.e.*, traditional methods and learning-based methods. RANSAC [12] and its variants [13], [14], which mainly adopt a hypothesize-and-verify strategy, are the main representatives of traditional methods. However, their theoretical running time often grows exponentially with the increase of ratio between outliers and all correspondences. It is adverse to our task because the outlier ratio in our validation dataset is often over 90%.

The advent of deep learning provides a new intuition for outlier removal methods. The majority of advances based on deep learning treat the outlier removal task as a binary classification task, which shows great performance and potential. However, the deep learning technology based outlier removal methods have many problems: (1) the unordered and irregular property of input requires the network to be permutation-equivariant when dealing with the correspondences; (2) the scarcity of available information, that is only spatial positions of matched feature points, largely influences the deep information acquisition and seriously damages the performance of network [8].

For example, LFGC-Net [8], OA-Net [9] and ACNet [15] adopt the PointNet-like architecture, which adopts Multi-Layer

Manuscript received 20 February 2022; revised 14 June 2022; accepted 17 June 2022. Date of publication 1 July 2022; date of current version 12 July 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62072223 and in part by the Natural Science Foundation of Fujian Province under Grant 2020J01829. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Lucio Marcenaro. *(Corresponding author: Guobao Xiao.)*

Linxin Zheng is with the College of Computer and Control Engineering, Minjiang University, Fuzhou 350108, China, and also with the College of Computer and Data Science and the College of Software, Fuzhou University, Fuzhou 350108, China.

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

Shiping Wang is with the College of Computer and Data Science and the College of Software, Fuzhou University, Fuzhou 350108, China.

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

Digital Object Identifier 10.1109/TIP.2022.3186535

Perceptrons (MLPs) to process each correspondence individually, to handle the unordered property problem. Then, they utilize Context Normalization (CN) to embed global context in each correspondence, and this will be beneficial to invariance under transformations $e.g.$, rotating and translating. CN is a technique to normalize all correspondences equally according to mean and variance order moments, which can be treated as the solution of a least-squares problem. As we know, the solution of a least-squares problem is not robust.

To address the problem, we try to introduce the attention mechanism (i.e., Squeeze and Excitation (SE) block [16]) to focus on the inlier information and reduce the influences of outliers. SE block is a popular attention mechanism and has been widely used in many fields due to its effectiveness. Note that, SE block squeezes each feature map into a scalar. This rough descriptor tends to emphasize global information and may ignore the most of local information. But, the local information involves the local motion coherence, which is beneficial to removing the false correspondences from putative correspondences. Thus, we propose the multi-scale attention (MSA) block to enhance the ability of local information acquirement, and rescale the original feature map by new scalar which fuses the scalar contained local information and the scalar contained global information. Specifically, we firstly squeeze the feature map and extract the global information to generate the global scalar. Then, we adopt the point-wise convolution layer with the bottleneck structure to generate the local scalar with the local context information. At last, we fuse the two scalars to form a multi-scale scalar containing multi-scale context information, and use the multi-scale scalar to softly select the feature map. Thus, the MSA block can distinctively process the feature map with the multi-scale context.

In addition, currently existing networks adopt PointCN blocks, which comprise Context Normalization, Batch Normalization, ReLU and MLPs, to exploit the information of putative correspondences. But, the PointCN block costs many parameters and the consumption of calculation. To improve the efficiency of our network, we propose the context channel refine block and the context spatial refine block to extract the context in channel and spatial aspects, respectively. Specifically, the proposed context channel refine block and the context spatial refine block firstly extract the context in channel and spatial respects with less parameters. Then, they refine the feature map from the past layer. After that, the semantic information of the obtained feature map is enhanced. At last, we concatenate the feature maps from the different layers as the final feature map. Compared with the PointCN block, the proposed blocks are able to help cost half parameters to extract the context and achieve the better performance. After that, we further combine the context channel refine block and multi-scale attention block into the attentional correspondence learning block, which distinctively extracts the context information. Finally, with all the proposed blocks, we propose a novel network for feature matching.

We summarize the contributions as follows:

- We design a novel effective network with less parameters, to establish the correct correspondences. As we know, we are the first ones to introduce the multi-scale attention mechanism to handle the feature matching problem.
- We propose an innovative multi-scale attention block to discriminatively treat correspondences. The proposed block, which softly selects feature maps by a scalar combined local and global information, can suppress useless context information and simultaneously improve the important context information.
- We design the context channel refine block and context spatial refine block to capture context information along channel and spatial aspects, respectively. The proposed blocks are able to help the network extract more useful information with less parameters.

It is worth pointing out that, the proposed MSA-Net has a great improvement over current state-of-the-art methods on the feature matching accuracy for the tasks of outlier removal and relative pose estimation. Our network achieves an 11.7% improvement at error threshold 5° without RANSAC compared to the state-of-the-art method on relative pose estimation task when trained on YFCC100M dataset. In addition, our network also achieves 2.92% and 1.95% improvement at the correspondence removal task when trained on YFCC100M dataset and SUN3D dataset, respectively.

The rest of the paper is organized as follows: we first review the related feature matching and attention mechanism literatures in Sec. II. Then, we describe the details of the proposed method in Sec. III and present the experimental results in Sec. IV. Finally, we draw conclusions in Sec. V.

## II. Related Work

In this section, we will introduce the recent outlier removal works and different attention mechanisms.

### A. Outlier Removal

Recent years, the outlier removal method has been developed two factions, $i.e.$ traditional methods and learning-based methods. Most traditional methods are based on the verifies-hypothesis strategy, such as RANSAC [12], EVSAC [17], MAGSAC [14], Contrario RANSAC [18], PROSAC [13], $etc$. RANSAC repeatedly selects random subsets of the input point set to fit a model and selects the model with the max score as the best model. PROSAC selects the point which has the high score of the similarity to be the subset. EVSAC adopts the extreme value theory to modify the sampling strategy, which accelerates the process of inferring an all-inlier sample probability. MAGSAC proposes the $\sigma$-consensus to eliminate the user-defined inlier-outlier threshold in RANSAC. Contrario RANSAC optimizes each model by selecting the most likely noise scale to reduce the dependency on $\sigma$. RANSAC and its variants are powerful solutions for the initial input set with proper inliers. However, their performance depends on the effectiveness of the sampled subsets, and it is not an easy task to sample effective subsets when outliers are dominant in data.

As the pioneer of learning-based outlier removal methods, LFGC [8] adopts PointNet-like architecture to solve the irregular and unordered input problem by processing the correspondence individually. In addition, the proposed Context Normalization (CN) is adopted to embed the global context into each correspondence. Although LFGC is important, there

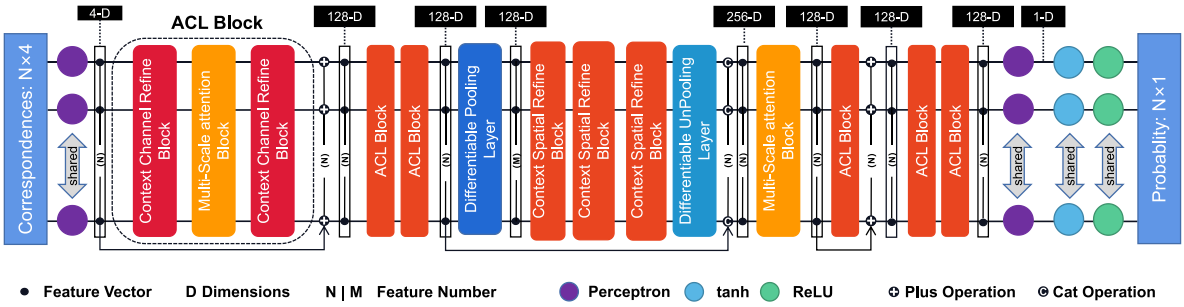



Fig. 2. Structure diagram of our multi-scale attention network.

is a little discussion about local context and robustness [15]. Many learning-based methods attempt to improve it by achieving local context and robustness. For example, OANet [9] adopts the differential pooling and unpooling layer to achieve the local context and predicts the probability of each correspondence being an inlier, respectively. ACNet [15] proposes the attentive context normalization (ACN), which is robust to outliers, by weighted mean and variance with global and local attention mechanism. LMCNet [19] learns the motion coherence property for correspondence pruning. Although existing methods have shown satisfactory performance, they still suffer from the issue of dominant outliers in the putative correspondences.

Thus, we propose the attentive correspondence learning block to discriminatively treat each correspondence and enhance the feature map quality according to removing redundancy context information.

### B. Attention Mechanism

Recently, researchers propose different attention mechanisms to show the benefits in a series of tasks, such as machine translating [20], salient object detection [21], [22], semantic segmentation [23], *etc*. In particular, [20] first proposes the self-attention to obtain the long dependency from the input sequences and applies it on machine translating. Reference [24] proposes an efficient criss-cross attention to obtain full-image contextual information for Semantic Segmentation task and reduces the parameters in a very effective and efficient way. The attention mechanism cannot only use the similarity between different features for allocation of the most informative feature expressions, but also use pooling operator to make the model focus on the important channels and positions. Reference [16] proposes the channel attention mechanism, which is an squeeze-and-excitation block to recalibrate feature maps, to achieve rich feature map. CBAM [25] proposes an effective attention module including channel attention and spatial attention to achieve significant context information and improve the performance of the network. In addition, some methods attempt to enhance the feature expressions. SKNet [26] introduces dynamic attention selection mechanism that allows each neuron to fuse multiple branches with different receptive field size. ResNeSt [27] based on Resnet [28] proposes a similar Split-Attention block that enables cross-group attention function mapping. In this paper, we propose a multi-scale attention block to capture local and global context of feature maps, and it helps our network treat correspondences discriminately to enhance the robustness to outliers.

## III. Multi-Scale Attention Network

In this section, we design a Multi-Scale Attention Network to infer the correspondences correctly and recover the pose estimation, as shown in Fig. 2. In the following, we first introduce the problem formulation in Sec. III-A. Then, we introduce the network framework in Sec. III-B. Then, we introduce the multi-scale attention block and context channel refine block in Sec. III-C and Sec. III-D, respectively. Subsequently, we introduce the context spatial refine block in Sec. III-E. At last, we introduce the loss function in Sec. III-F.

### A. Problem Formulation

We divide the two-view geometry estimation task into an inlier/outlier classification task and an essential matrix regression task. Specifically, given a pair of images $(I, I')$, we first use the traditional method [10] or learning-based methods [11] to extract the key-point coordinates and corresponding descriptors. After that, we search for the nearest neighbors from coordinates of keypoints in the other images to establish initial correspondences. We describe $N$ correspondences as follows:

$$S = [s_1, s_2, s_3, \ldots, s_N] \in R^{N\times 4}, s_i = (x_i, y_i, x_i', y_i'), \quad (1)$$

where $S$ stands for the input correspondences of the outlier removal method. $s_i$ represents the $i$-th correspondence. $(x_i, y_i)$ and $(x_i', y_i')$ are the corresponding coordinates of $i$-th keypoints in two images, respectively. Here, we use camera intrinsics to normalize it.

In the outlier removal task, the proposed end-to-end network first obtains the global and local context from the initial correspondences. Then, the global and local context are used to infer the probability set of correspondences being inliers $P = \{p_1, \ldots, p_i, \ldots, p_N\}$, where $p_i$ represents the probability of $i$-th correspondences being inliers or outliers. It is worth noting that, $p_i = 0$ means the $i$-th correspondence is an outlier. Above processes can be formulated as:

$$o = f_\phi(S), \quad (2)$$

$$P = tanh(ReLU(o)), \quad (3)$$

where $o$ means the logit values for classification. $f$ denotes the end-to-end network which extracts the global and local context

and infers the probability of correspondences being inliers. In addition, the end-to-end network is permutation-equivariant, *i.e.*, the output result of the network will be not affected by the unordered input problem. $\phi$ stands for parameters of the network. $ReLU$ and $tanh$ are $ReLU$ and $tanh$ activation functions, respectively. Note that, they can be used to remove outliers in the classification problem [9].

Then, we employ the probability set $P$ and corresponding correspondence set $S$ as input, and adopt the weighted 8-points algorithm to regress the essential matrix. Different from the traditional 8-points algorithm [1], the weighted 8-points algorithm is robust to outliers, because it is adopted in conjunction with the outlier removal network, which avoids largely the effect of outliers. We formulate the process as:

$$\widehat{E} = g(P, S), \tag{4}$$

where $g(\cdot, \cdot)$ represents the weighted 8-points algorithm. The output $\widehat{E}$ is the predicted essential matrix.

### B. Network Framework

In this subsection, we introduce our framework as shown in Fig. 2. It mainly contains three key parts, *i.e.*, multi-scale attention block, context channel refine block, and context spatial refine block. Given the correspondence set $S \in \mathbb{R}^{N\times 4}$, we first use a MLP with 128 neurons to process it. Then, the generated feature map passes through 3 attentional correspondence learning (called ACL) blocks, including two context channel refine blocks and a multi-scale attention block, to discriminately treat each feature. Context spatial refine block is used to extract and refine context in spatial dimensions. Specifically, differentiable pooling layer introduced by [9] maps the $N$ correspondences to $M$ clusters, ($N$>$M$, *e.g.*, $N = 2000$, $M = 500$). Differentiable unpooling layer recovers the feature map to initial spatial size. In addition, 3 context spatial refine blocks between differentiable pooling layer and differentiable unpooling layer are used to deal with the clusters. Then, we use multi-scale attention block to fuse the feature map which is concatenated by context channel refine block and context spatial refine block. Subsequently, the feature map is processed by 3 attentional correspondence learning blocks. At last, we use MLP, tanh and ReLU to generate the probability of correspondences being inliers. By the way, our network adopts an iteration strategy which means that the network has two same structures (Fig. 2) in the form of serial connection, and the second structure output is the final output (i.e. the probability of correspondences being inliers). The first structure adopts the coordinate as the input, and the second structure adopts the coordinate points, the Epipolar distance and the probability of correspondences being inliers inferred by the first structure. It is worth noting that the epipolar distance is computed by the essential matrix and the coordinate points, and the essential matrix is computed by the coordinate points and probability of correspondences being inliers. Note that, we follow most of deep learning based feature matching methods, e.g., OANet++ and ACNet, to repeat the proposed structure. The reason behind this is that, the iteration strategy is able to detach the gradients from the latter stage, and this will improve the performance of a network for feature matching [9].

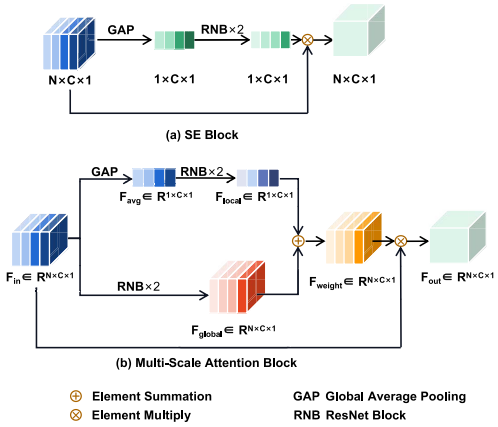



Fig. 3. Structure diagram of SE and our multi-scale attention block.

### C. Multi-Scale Attention Block

The percentage of inliers in our dataset is often around 10%, because of light changes, blurs, *etc*. Note that this is a critical challenge for a network because the CN layer treats all correspondences equally, which is not robust to outliers [15]. That is, CN will largely damage the performance of a network and wash out the important context.

Squeeze and excitation (SE) block, as shown in Fig. 3, a representative of attention mechanism, can effectively process the problem. However, SE block squeezes each feature map into a scalar. This rough descriptor tends to emphasize global information and may ignore most of the image signals in local context. The local context involves local motion consistency, which is important for feature matching.

To address this problem, we design the multi-scale attention block (see Fig. 3) to softly select feature maps with a weight descriptor, which is combined with the multi-scale context. The multi-scale context, which contains local context and global context, aggregated by the attention module helps our network better remove the outliers (i.e. false correspondences). Specifically, we first embed the global context information by a simply global average pooling operation to generate the global feature maps $\mathcal{F}_{avg} \in \mathbb{R}^{C\times 1}$. To generate the $k$-th element of $\mathcal{F}_{avg} \in \mathbb{R}^{C\times 1}$, we squeeze the $k$-th $\mathcal{F}_{in}$:

$$\mathcal{F}_{avg}^{k} = \frac{1}{N}\sum_{k=1}^{N} \mathcal{F}_{in}^{k}. \tag{5}$$

Then, we process the global feature map into a channel descriptor. Specifically, we adopt the point-wise convolution layer to aggregate the channel information, which exploits the interaction among features. In addition, to balance the parameters and efficiency, we adopt the bottleneck structure to compute $\mathcal{F}_{global}$:

$$\mathcal{F}_{global} = \beta(Conv_2(\theta(\beta(Conv_1(\mathcal{F}_{avg}))))), \tag{6}$$

where $Conv$ denotes point-wise convolution. The weight of $Conv_1$ and $Conv_2$ are the $C \times C/r \times 1 \times 1$ and $C/r \times C \times 1 \times 1$ weight, respectively. $C$ and $r$ means the number

of channels and the channel reduction ratio, respectively. In addition, $\beta$ represents the Batch Normalization layer and $\theta$ is the ReLU layer. We use them to obtain the context feature map. However, a single context feature map collected by global average pooling and Eq. (6) ignores different needs across scales and locations. Thus, we adopt the same method to extract the local context $\mathcal{F}_{local}$:

$$\mathcal{F}_{local} = \beta(Conv_2(\theta(\beta(Conv_1(\mathcal{F}_{in}))))). \tag{7}$$

Then, we combine $\mathcal{F}_{local}$ and $\mathcal{F}_{global}$ by a simply element-wise summation operator. And $Sigmoid$ is used to generate the weight $\mathcal{F}_{weight}$. Thus, the weight $\mathcal{F}_{weight}$ not only leverages the global context information but also captures complementary context information from the local aspect. As we know, the more global and local context information acquisition, the better for the task with image transformation and blurs:

$$\mathcal{F}_{weight} = Sigmoid(\mathcal{F}_{local} + \mathcal{F}_{global}). \tag{8}$$

At last, we adopt the weight $\mathcal{F}_{weight}$ to soft select the input $\mathcal{F}_{in}$ by element-wise multiplication operator, which suppresses the useless features and enhance the important features.

$$\mathcal{F}_{out} = \mathcal{F}_{weight} \cdot \mathcal{F}_{in}. \tag{9}$$

### D. Context Channel Refine Block

PointCN is adopted in many works [8], [9]. It stacks a CN, a BN, a ReLU activation function and a MLP. However, PointCN achieves less context information with many parameters, which decreases the quality of feature map and consumes the excess computer resources. To address this problem, we propose the context channel refine block to mine comprehensive context information with less parameters.

For a feature map $F \in \mathbb{R}^{N\times C}$, compared with the feature map directly generated by PointCN, the new feature map generated by the context channel refine block contains $h$ heads with different semantic information. Specifically, we first adopt PointCN with the $C \times C/h \times 1 \times 1$ weight to downsample the feature map and remove redundancy context information. Then, we sequentially adopt a corresponding PointCN block to process each feature map generated by previous PointCN block, denote as $PointCN_i$, where $i = \{2, 3, \ldots, h\}$. As the feature map is refined, the semantic information is gradually increased. To keep the channel of feature map, each head has the $C/h$ channels. At last, we concatenate all heads to the new feature map. As shown in Fig 4, the structure can be written as:

$$\mathcal{F}^1 = PointCN_1(\mathcal{F}) \tag{10}$$

$$\mathcal{F}^i = PointCN_i(\mathcal{F}^{i-1}) \tag{11}$$

$$\mathcal{F}_{new} = Concat(F^1, F^2, \ldots, F^h). \tag{12}$$

We can see that, each PointCN block $PointCN_i$ could receive the feature information of previous PointCN blocks such that it can obtain more context information and sematic information from previous blocks for network learning. Therefore, the output feature map has stronger representation ability

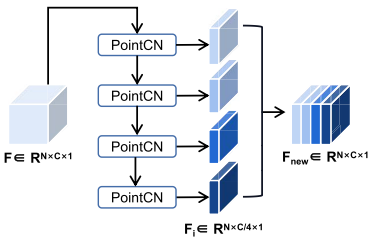



Fig. 4. Structure diagram of our context channel refine block.

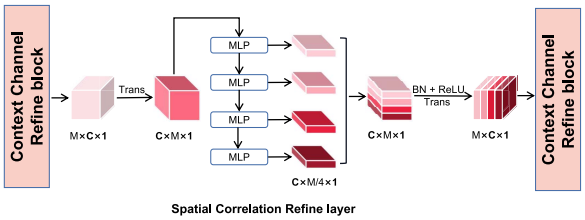



Fig. 5. Structure diagram of our context spatial refine block.

for the correspondence removal. In addition, we propose the attentional correspondence learning block, which is made up of a context channel refine block, a multi-scale attention block and a context channel refine block in order. It uses two context channel refine blocks to help the different parts exchange the information. It helps MSA-Net infer better results.

It is worth pointing out that, the proposed context channel refine block uses a multi-level structure as CCL [29], U-Net [30], and FPN [31]. However, the details and effects are different. Specifically, CCL, U-Net and FPN adopt the convolution kernels with different sizes for detecting the object. That is, they adopt different receptive fields to achieve the different scale object information. In contrast, our block uses point-wise convolution kernels to preserve the permutation-equivariant property for the unordered input problems of feature matching, and our block aims to extract the channel information to help the network infer correspondences being inliers/outliers. In addition, CCL, U-Net and FPN adopt skip connection, the top-down and bottom-up architecture, but our block just connects the $h$ PointCN blocks and concatenates the feature map along channel dimension. Thus, our block can include less parameters.

### E. Context Spatial Refine Block

The context channel refine block just focuses on channel information. Thus, we propose the context spatial refine block to mine spatial relationship to make the network accurate.

As shown in Fig. 5, the context spatial refine block contains spatial correlation refine layer and two context channel refine blocks introduced in Sec. III-D. Specifically, the spatial correlation refine layer is a simple but more effective layer which is implemented by a series of MLPs and the transpose operation. The aim of transpose operation is extracting spatial context by transposing the spatial dimension and the channel dimension. In addition, we stack four MLPs to effectively learn feature representations in the spatial aspect. We insert the spatial correlation refine layer between two context channel refine blocks. They are able to aggregate complementary information, since context channel refine block extracts channel information and

spatial correlation refine layer extracts spatial information. Thus, the generated feature map have more comprehensive context information. It is worth noting that the context spatial refine block only can be used after the differentiable pooling layer [9] because of the correspondence unordered problem.

### *F. Loss Function*

We adopt a hybrid loss function to supervise our network. The goal of our network is to minimize the value of hybrid loss function in the training process as follows:

$$\mathcal{L}(\phi) = \mathcal{L}_{cls}(P, L_k) + \lambda \mathcal{L}_{ess}(\hat{E}, E), \tag{13}$$

where $\mathcal{L}_{\phi}$ is the hybrid loss function. $\mathcal{L}_{cls}$ denotes the classification loss function which computes between the probability set $P$ and the label set $L_k$. $\mathcal{L}_{ess}$ represents the essential matrix loss function constrained by geometry distance. $\mathcal{L}_{ess}$ computes between the ground-truth essential matrix $\hat{E}$ and the matrix $\hat{E}$ predicted by our proposed network. $\lambda$ denotes a hyper-parameter to balance the classification loss function and essential matrix loss function. We formulate the classification loss function and essential matrix loss function as follows:

$$\mathcal{L}_{cls}(w, L_k) = \frac{1}{N_k} \sum_{i=1}^{N_k} \mu_k^i \mathcal{G}(w_k^i, L_k^i), \tag{14}$$

$$\mathcal{L}_{ess}(\hat{E}, E) = \frac{(p\prime^T \hat{E} p)^2}{(Ep)_{[1]}^2 + (Ep)_{[2]}^2 + (E^T p\prime)_{[1]}^2 + (E^T p\prime)_{[2]}^2}, \tag{15}$$

where $\mathcal{G}(\cdot, \cdot)$ denotes the binary cross entropy which computes the probability set and label set in the $k$-th pair of images. In addition, $\mu_k^i$ denotes hyper-parameter to balance the ratio of inlier/outlier. In Eq. (15), $p = (x, y, 1)$ and $p' = (x', y', 1)$ are derived from the keypoint locations of a correspondence. $()_{[i]}$ represents the $i$-th element of vector.

## IV. Experiments

In this section, we first introduce our datasets in Sec. IV-A. Then, we introduce the evaluation metrics and implementation details in Sec. IV-B and Sec. IV-C, respectively. Then, we analyse the proposed network performance on outlier removal task and pose estimation task in Sec. IV-D and Sec. IV-E, respectively. Subsequently, we design ablation studies and the ablation study of each module in Sec. IV-F and Sec. IV-G, respectively. At last, we analyse the parameter of our network in Sec. IV-H.

### *A. Datasets*

To verify our network effectiveness, we conduct many experiments on several benchmark visual datasets, including YFCC100M and SUN3D datasets. We use the YFCC100M and SUN3D datasets as outdoor and indoor datasets, respectively. To have a fair comparison, we train all models at the same training setting. Specifically, we split all datasets into known and unknown scenes. The known scenes are divided into several disjoint subsets, *i.e.*, training (60%), validation (20%) and testing (20%). The training (60%) subset is used to train our network, and then the validation (20%) subset is used to verify the network for adjusting the parameters of the network. After that, the testing (20%) subset is used to test the performance of our network. For unknown scenes, all datasets are used to test the performance of a network. The testing sets from known and unknown scenes are different: The testing set from known scenes shares the same scene with the training subset while the scene of testing set from unknown scenes is different from the scene of training subset. Thus, the test on unknown scenes is able to show the generalization ability of a network.

*1) YFCC100m Dataset:* YFCC100M (i.e., Yahoo Flicker creative commons 100 Million) dataset is created by Yahoo, which contains 100 millon images from Internet. Following [9], we use the subset of the YFCC100M dataset to generate 71 sequences. Then, we divide the 71 sequences into 67 sequences and 4 sequences. We utilize the 67 sequences as known scenes. The remaining 4 sequences are utilized as unknown scenes to test the generalization ability.

*2) SUN3D Dataset:* We select SUN3D dataset as the indoor dataset, which is an RGBD video dataset of entire room and the relative ground-truth information calculated by generalized bundle adjustment. We split the indoor dataset into 254 sequences, which are selected 239 sequences for known sequences and 15 sequences for unknown sequences, respectively. Note that, the indoor scenes have extensive blurs, occlusions and repetitive structures. Thus, it is more challenging to the task of outlier removal than the outdoor datasets.

### *B. Evaluation Metrics*

The task of our proposed network is divided into two tasks, which are the outlier removal task and the relative pose estimation task. For explicitly evaluating the performance of the outlier task, we adopt the *Precision* ($P$), *Recall* ($R$) and *F-score* ($F$) as the evaluation metrics. Specifically, the *Precision* is defined as the ratio between the identified correct correspondence number and the preserved correspondence number. The *Recall* is defined as the ratio between the identified correct correspondence number and the number of all correct correspondence in initial correspondence set. The *F-score* is defined as $2 * P * R/(P + R)$. We select the mean average precision (mAP) of the angular differences under different error thresholds as the evaluation metric of camera pose estimation task. Note that, the angular difference is calculated by the ground truth and the predicted rotation and translation vectors.

### *C. Implementation Details*

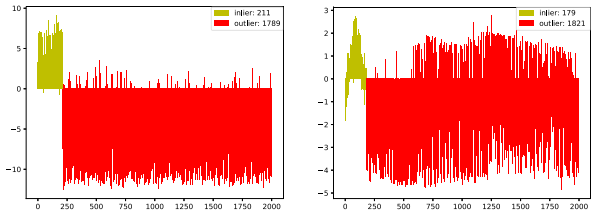
The structure of our network is shown in Fig 2, and we discuss its components in Sec. III. The inputs and outputs of these components are 128 channels. For the network input, we utilize $N \times 4$ initial correspondences, generally $N = 2000$. In ACL block, to balance the effectiveness and parameters, the reduced channel ratio of multi-scale attention bottleneck and the number of PointCN in context channel refine is 4. In DiffPool layer, we map the $N$ correspondences into fixed number 500. In addition, we utilize the multi-scale

TABLE I

RESULTS OF OUR NETWORK AND OTHER MODELS ABOUT THE PRECISION, RECALL AND F-SCORE 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 |
| PointNet++ | 49.84 | 86.41 | 63.22 | 46.60 | 84.17 | 59.99 | 51.40 | 86.73 | 64.54 | 44.79 | 83.23 | 58.23 |
| LFGC-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 | 56.61 | 87.04 | 68.60 | 53.93 | 85.52 | 66.15 | 53.88 | 87.20 | 66.61 | 46.21 | 84.07 | 59.64 |
| ACNet | 59.99 | 88.81 | 71.61 | 55.54 | 85.38 | 67.30 | 54.03 | 88.45 | 67.08 | 45.97 | 83.94 | 59.40 |
| OANet++ | 60.03 | 89.31 | 71.80 | 55.78 | 85.93 | 67.65 | 54.30 | **88.54** | 67.32 | 46.15 | **84.36** | 59.66 |
| Ours | **61.98** | **90.53** | **73.58** | **58.70** | **87.99** | **70.42** | **55.38** | 87.51 | **67.83** | **48.10** | 83.81 | **61.12** |

channel attention module to fuse the context generated by context channel refine block and context spatial refine block. At last, the 128 channel feature map is processed by MLPs to generate a probability of correspondences being inliers/outliers as the output. We utilize Pytorch to implement our network with batchsize 32 and apply the Adam optimizer with the learning rate of $10^{-3}$ to optimize the parameters. In addition, we train the network $500k$ iterations and adopt the iterative network [9]. Note that all experiments are implemented on NVIDIA TESLA P100 GPUs.




(a) (b)

Fig. 6. The output weight of our network. (a) is implemented on YFCC100M. (b) is implemented on SUN3D. The x-axis represents the number of correspondence. The y-axis denotes weight of correspondence. In addition, yellow lines and red lines represents inliers and outlier, respectively.

### D. Outlier Removal Task

As shown in Table I, we compare our method with six methods, including RANSAC [12], PointNet++ [32], LFGC-Net [8], DFE [33], OA-Net++ [9], ACNet [15]. Inspired by PointNet, LFGC is a deep permutation-equivariant network, which extracts global context from the input set to classify the correspondences as inliers or outliers. In addition, we choose PointNet++, an improved version of PointNet, as a comparative method. DFE is a specialized network to recover the pose estimation. OA-Net++ is the official improved version of OA-Net, which clusters the correspondences into fixed clusters to extract local context by the affine transformation and infers the probability of correspondences being inliers. ACNet adopts the global and local attention to improve robustness ability to outliers. Note that SuperGlue [34] is not compared since it focuses on another task, i.e. predicting high-quality initial correspondences rather than outlier removal.

From the results in Table I, we can find that all learning-based methods are better than RANSAC, since RANSAC is suit to specific constraints, *e.g.*, high inlier ratio in initial correspondence set. The inlier rate of our initial correspondence set is often below 10%, thus all learning-based methods perform better than RANSAC. In addition, our MSA-Net also achieves the best results among all learning-based methods in *Precision* and *F-score*, but the *Recall* is slightly low in the SUN3D dataset. Specifically, our method improve 2.92% and 1.95% than OANet++ on the unknown scene and known scene in YFCC100M datasets, respectively. In SUN3D, our method improve 1.95% and 1.08% than OANet++ on the unknown scenes and known scenes. For the *F-score*, we obtain better improvements than other outlier removal methods. We visualize our network weight in Fig. 6. We can find that our network is able to discriminately treat each correspondence on YFCC100M dataset, but the performance of our network is not well on SUN3D due to extensive blurs, repetitive structures and lack of texture. In addition, as shown in Fig. 7, we visualize RANSAC, ACNet, OA-Net++ and our network results in two datasets. We can find that our network is able to establish correct correspondences better than other methods.

### E. Relative Pose Estimation

In this section, we discuss the result of the relative pose estimation and the impact of different feature extraction methods. The high-quality keypoints and descriptors are able to make the performance of the deep network obtain a large improvement in most scenes. In addition to using SIFT as a feature extraction method, we also use a learning-based feature extraction method to establish initial correspondences, i.e., SuperPoint [11]. Since it extracts keypoints and descriptors based on a self-supervised network, it is able to be applied a large number of multi-view geometry tasks in computer vision.

Here, we first employ SuperPoint or SIFT to detect keypoints and extract local descriptors, and then we use extracted descriptors to generate the putative correspondences set with the nearest-neighbor matching algorithm. The result of the feature extraction method using SuperPoint and SIFT is shown in Table II. Note that we conduct outlier removal without RANSAC. As in the outlier removal experiment, we select RANSAC, PointNet++, LFGC-Net, DFE, OA-Net++, ACNet to compare our method with the mAP at error thresholds 5° and 20°. The experiments are conducted on known and unknown scenes the YFCC100M dataset, our method achieves the best results compared with other methods except the results at error thresholds 20° on the known scene with the SuperPoint method. From the results in Table II, the result of SuperPoint is worse than SIFT. This is mainly because we use the dataset in this paper to train the network, resulting in low accuracy of the detected keypoints. Therefore, if we

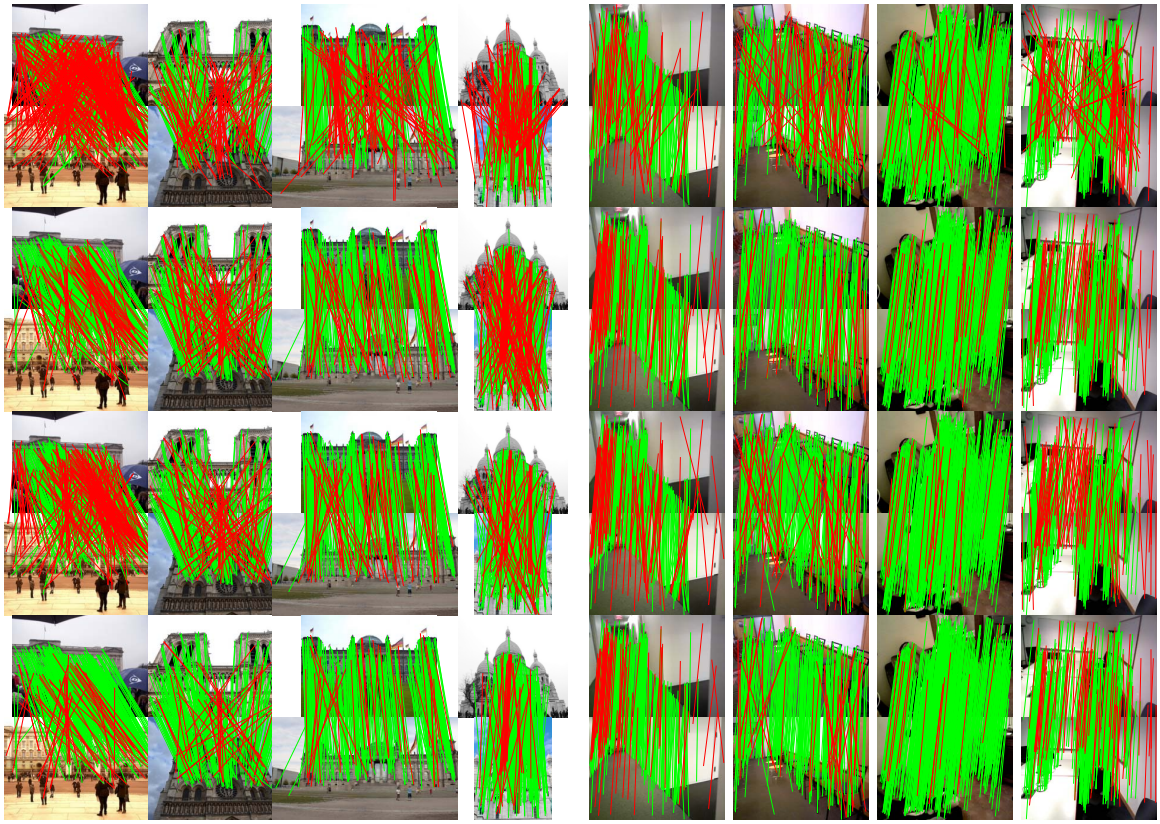


Fig. 7. Visualization results. From top to bottom: results of RANSAC, ACNet, OA-Net++ and our network. Different colors denote different feature consistencies and red represents the false-positive. To better comparison, RANSAC uses the ratio test to improve performance.

TABLE II

RESULTS OF OUR NETWORK AND OTHER MODELS WHEN USING DIFFERENT LOCAL FEATURE EXTRACTION METHODS ON THE YFCC100M DATASET

| | Matcher | Known Scene | | Unknown Scene | |
|---|---|---|---|---|---|
| | | 5° | 20° | 5° | 20° |
| SIFT | RANSAC | 5.81 | 16.88 | 9.07 | 22.92 |
| | PointNet++ | 10.49 | 31.17 | 16.48 | 42.09 |
| | LFGC-NET | 13.81 | 35.20 | 23.95 | 52.44 |
| | DFE | 19.13 | 42.03 | 30.27 | 59.18 |
| | ACNet | 29.17 | 52.59 | 33.06 | 62.91 |
| | OA-Net++ | 32.57 | 56.89 | 38.95 | 66.85 |
| | Ours | **39.53** | **61.75** | **50.65** | **77.99** |
| SuperPoint | RANSAC | 12.85 | 31.22 | 17.47 | 38.83 |
| | PointNet++ | 11.87 | 33.35 | 17.95 | 49.32 |
| | LFGC-NET | 12.18 | 34.75 | 24.25 | 52.70 |
| | DFE | 18.79 | 40.53 | 29.13 | 58.41 |
| | ACNet | 26.72 | 49.29 | 32.98 | 62.68 |
| | OA-Net++ | 29.52 | **53.76** | 35.27 | 66.81 |
| | Ours | **30.63** | 53.74 | **38.53** | **68.56** |

use Superpoint to generate the putative correspondences set, it will contain a lower ratio of inliers, which is leading to worse results in the outlier removal task for all networks. Specifically, our method improves 11.7% and 6.96% than OA-Net++ on the unknown scene and known scene at error thresholds 5° about the SIFT method. For the SuperPoint method, our method also achieves better results than other method at error thresholds 5°.

### *F. Ablation Studies*

In this section, we provide some ablation studies about the network architecture on the YFCC100M dataset. As shown in Table III, we test the performance of different combinations for relative pose estimation task and outlier removal task on the YFCC100M dataset.

In Table III, the first row is OANet++. We treat OANet++ as the baseline. We can find that the performance of our all iterative combinations outperforms the baseline. Specifically, the second row of table (Context Spatial Refine block (CSR) + PointCN) and the third row of table (CSR + Context Channel Refine block (CCR)) obtain a 2.68% and a 3.25% improvement than baseline on unknown scenes at error threshold 5° without RANSAC, respectively. In addition, they decrease 0.96M and 1.2M parameters. They indicate that CSR and CCR are able to capture richer context information with less parameters. Then, the multi-scale attention block inserted in context channel refine block (CSR + CCR + Att1) obtains an 6.95% improvement over the baseline on unknown scenes without RANSAC. In addition, the multi-scale attention block fuses the global context and local context (CSR + CCR + Att1 + Att2) achieves an 11.7% and a 2.7% improvement than the baseline at error thresholds 5° and *Precision*, respectively. It means that our multi-scale attention block discriminately treats the feature map in an effective manner and improves the interaction among correspondences.

### *G. Ablation Study of Each Module*

*1) Impact of Context Channel Refine Arrangement:* In order to study the arrangement of different numbers, we test the multi-head structure and context channel refine block, which have same channels and heads. They are implemented on

TABLE III

ABLATION STUDY ON THE YFCC100M DATASET. MAP ( PROCESSED WITHOUT/WITH RANSAC IN TESTING. P&O&UNP: USING THE POOLING AND UNPOOLING LAYER, ORDER-AWARE FLITER BLOCK COME FROM OANET. CSR: USING CONTEXT SPATIAL REFINE BLOCK. CCR: USING CONTEXT CHANNEL REFINE BLOCK. ATT1: MULTI-SCALE ATTENTION BLOCK IS APPLIED IN CONTEXT CHANNEL REFINE BLOCK. ATT2: MULTI-SCALE ATTENTION BLOCK IS APPLIED IN POOLING AND UNPOOLING

| PointCN | P&O&UnP | CSR | CCR | Att1 | Att2 | P (%) | R (%) | F (%) | mAP (%) 5° | mAP (%) 20° | Param (M) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| √ | √ | | | | | 55.44 | 86.31 | 65.71 | 38.95/52.59 | 66.85/72.99 | 2.47 |
| √ | | √ | | | | 56.38 | 86.36 | 68.22 | 41.63/52.58 | 68.94/73.31 | 1.51 |
| | | √ | √ | | | 56.02 | 87.40 | 68.27 | 42.20/53.78 | 69.93/74.59 | 1.27 |
| | | √ | √ | √ | | 57.44 | 88.24 | 69.58 | 45.38/55.93 | 70.96/75.47 | 1.48 |
| | | √ | √ | √ | √ | 58.14 | 88.92 | 70.42 | 50.65/56.28 | 77.99/78.46 | 1.62 |

TABLE IV

ARCHITECTURES FOR ABLATION STUDY ON IMPACT OF CONTEXT CHANNEL REFINE BLOCK ARRANGEMENT

| Method | mAP (%) | | | Params(M) | GFlops |
|---|---|---|---|---|---|
| | 5° | 10° | 20° | | |
| OANet++ | 38.95 | 55.23 | 66.85 | 2.47 | 1.81 |
| Multi-H+OANet++ | 41.13 | 55.23 | 68.99 | 2.47 | 1.82 |
| CCR+OANet++ | 43.38 | 56.81 | 70.09 | 2.22 | 1.33 |

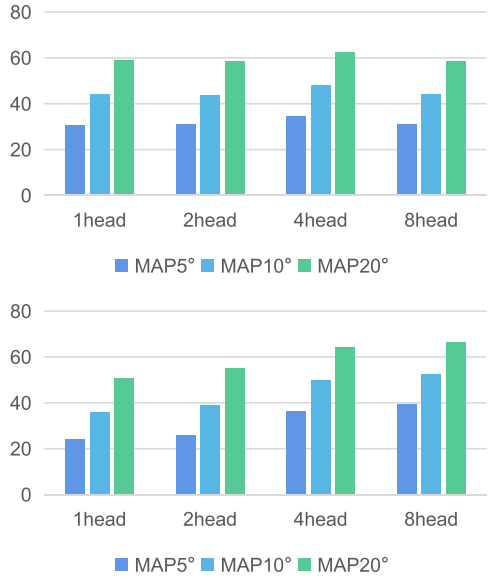



Fig. 8. Performance of context channel refine head impact on the YFCC100M dataset. The top has the same parameters, but each head has different channels. The bottom denotes 64 channels in each head.

YFCC100M dataset. The only difference is their arrangement of structures. Note that Multi-Head structure mimics the perception structure, but all kernel size are $1 \times 1$, due to the unordered correspondence problem. As shown in Table IV, Multi-Head structure and Multi-Level structure are better than the OANet++ about the mAP at error thresholds 5°, 10° and 20°. In addition, context channel refine block increases 2.25%, 1.58% and 1.1% than Multi-Head structure about the mAP at error thresholds 5°, 10° and 20°. It means that the proposed context channel refine block mines the available information more comprehensively than other structures.

*2) Impact of Feature Extractor Head:* As shown in Fig. 4, the number of heads influences the depth of semantic information. To verify the strategy, we test our block on YFCC100M dataset with different number of heads. For fairness, we test all models at the same parameters. As shown in Fig. 8, we can find the 4 heads has the best results at the same parameters. The 2 heads and 8 heads have the sub-optimal performance than 4 heads due to the small channels. When we use the same

TABLE V

ARCHITECTURES FOR ABLATION STUDY ON THE IMPACT OF COMBINE OPERATOR

| Method | mAP (%) | | | P(%) | R(%) | F(%) |
|---|---|---|---|---|---|---|
| | 5° | 10° | 20° | | | |
| Concat | 43.38 | 70.09 | 72.35 | 58.10 | 87.06 | 69.69 |
| Sum | 41.13 | 68.99 | 72.08 | 58.77 | 85.80 | 69.75 |

TABLE VI

ARCHITECTURES FOR ABLATION STUDY ON THE IMPACT OF FEATURE INTEGRATION STRATEGIES

| Method | mAP (%) | | | P(%) | R(%) | F(%) |
|---|---|---|---|---|---|---|
| | 5° | 10° | 20° | | | |
| $global + global$ | 43.45 | 57.00 | 70.61 | 57.35 | 86.33 | 68.91 |
| $local + local$ | 45.38 | 59.24 | 71.90 | 57.71 | 87.01 | 69.39 |
| $global + local$ | 46.52 | 59.64 | 72.02 | 58.10 | 87.06 | 69.69 |

channels in each head, we can find that 8 heads achieve the best results. Specifically, the block of 8 heads obtains a 15.28% improvement than the 1 head. However, the computational cost increases linearly with the increase of the head number.

*3) Impact of Combine Operator :* To study the impact of input combine methods, we conduct two experiments to compare the effectiveness of concat operator and sum operator. All experiments are implemented on YFCC100M dataset, as shown in Table V. In addition, the channel reduction ratio $r$ and the number of head $h$ in the concat and sum operator are same. The only difference is their input combine method. The results suggest that the concat operator outperforms the sum operator except for the parameter, because the concat operator preserves the context structure and is suitable to the classification problem.

*4) Impact of Multi-Scale Attention Block :* To study the impact of Multi-Scale Attention Block, we design three ablation studies to prove performance of our proposed method, as shown in Table VI. For fair comparison, the three ablation studies (*i.e.* Global + Local, Global + Global and Local + Local) have the same parameter, module width and channel reduction ratio $r$. It can be seen that the Global + Local outperforms single-scale ones in all settings. Compare with the single-scale ones, Global + Local obtains around 1% improvements at *Precision*, *Recall* and *F-score*. In addition, Global + Local raises 3.07% and 1.14% over Global + Global and Local + Local at error thresholds 5°, respectively. The results suggest that, compared with the single context, the network adopts multi-scale context is vital for seeking the reliable correspondences.

### *H. Parameter Analysis*

As shown in Table III and Fig. 9, we can find that the proposed network has not the lowest number of parameters and

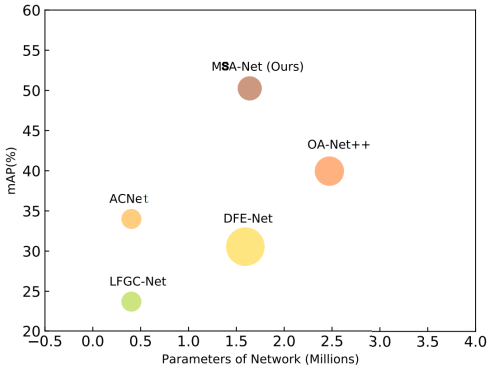



Fig. 9. The parameter analysis about deep learning networks.

TABLE VII

ARCHITECTURES FOR ABLATION STUDY ON THE IMPACT ON DIFFERENT ROBUST ESTIMATOR METHODS. NORMAL MEANS THAT WE USE THE NETWORK TO INFER THE RELATIVE POSE ESTIMATION TASK WITHOUT THE ROBUST ESTIMATOR METHODS. PROSAC, DEGENSAC AND RANSAC MEANS THAT THE NETWORK INFER THE RELATIVE POSE ESTIMATION TASK WITH PROSAC, DEGENSAC AND RANSAC METHODS, RESPECTIVELY

| Method | NORMAL | PROSAC | DEGENSAC | RANSAC |
|---|---|---|---|---|
| LFGC++ | 30.82 | 40.20 | 49.08 | 50.15 |
| OA-Net++ | 38.95 | 45.58 | 52.76 | 53.63 |
| Ours | 50.65 | 51.25 | 54.83 | 56.28 |

FLOPs among different networks. However, compared with OA-Net++, which has the highest error thresholds 5° among different networks except our method, we decrease $1.03M$ parameters and $0.6G$ FLOPs. In addition, compared with ACNet, which has the lowest parameters and FLOPs among different networks, we improve 16.41% at error thresholds 5°. The results show that we use less computation consume to extract richer context information than other methods.

### *I. Impact on Different Robust Estimator Methods*

The robust estimator method (a.k.a. post-processing) plays a crucial role in feature matching methods which determines the accuracy of relative pose estimation. In this section, we conduct the experiments about the relative pose estimation task with different robust estimator methods (i.e. DEGENSAC, PROSAC and RANSAC). We adopt the error thresholds 5° as evaluation criteria. In addition, we adopt the LFGC++ and OA-Net++ as baselines to compare the influence of different robust estimator methods. As shown in Table VII, the network obtains the best results by using RANSAC. Compared with RANSAC, our network decreases about 5.03 and 1.45 when using PROSAC and DEGENSAC, respectively. The main reason is that the restraint of a dominant plane is difficult to be held in the outdoor scenes due to its complex environment. Nevertheless, our method obtain the best results than other methods with different robust estimator methods.

## V. CONCLUSION

In this paper, we propose the Multi-Scale Attention Network (MSA-Net) to infer the probability of correspondences being inliers in an accurate and efficient way. Specifically, we propose a context channel refine block and a context spatial refine block with less parameters to mine the context information along channel and spatial dimensions, respectively. In addition, we design a multi-scale attention block to exchange the information among correspondences and enhance the representational ability according to extracting dependencies from all correspondences and discriminately selecting the features. Extensive experiments have shown that, compared with the state-of-the-art methods, the performance on the outlier removal and relative pose estimation task and computational consume of our network are more prominent.

## REFERENCES

[1] J. Ma, X. Jiang, A. Fan, J. Jiang, and J. Yan, "Image matching from handcrafted to deep features: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 1, pp. 23–79, Jan. 2021.
[2] X. Jiang, J. Ma, G. Xiao, Z. Shao, and X. Guo, "A review of multimodal image matching: Methods and applications," *Inf. Fusion*, vol. 73, pp. 22–71, Sep. 2021.
[3] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015.
[4] M. Brown and D. G. Lowe, "Automatic panoramic image stitching using invariant features," *Int. J. Comput. Vis.*, vol. 74, no. 1, pp. 59–73, Aug. 2007.
[5] N. Snavely, S. M. Seitz, and R. Szeliski, "Modeling the world from internet photo collections," *Int. J. Comput. Vis.*, vol. 80, no. 2, pp. 189–210, Nov. 2008.
[6] G. Xiao, J. Ma, S. Wang, and C. Chen, "Deterministic model fitting by local-neighbor preservation and global-residual optimization," *IEEE Trans. Image Process.*, vol. 29, pp. 8988–9001, 2020.
[7] H. Xu and J. Zhang, "AANet: Adaptive aggregation network for efficient stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1959–1968.
[8] K. M. Yi, E. Trulls, Y. Ono, V. Lepetit, M. Salzmann, and P. Fua, "Learning to find good correspondences," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2666–2674.
[9] J. Zhang *et al.*, "Learning two-view correspondences and geometry using order-aware network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5845–5854.
[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[11] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 224–236.
[12] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.
[13] O. Chum and J. Matas, "Matching with PROSAC–progressive sample consensus," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 220–226.
[14] D. Barath, J. Matas, and J. Noskova, "MAGSAC: Marginalizing sample consensus," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10197–10205.
[15] W. Sun, W. Jiang, E. Trulls, A. Tagliasacchi, and K. M. Yi, "ACNe: Attentive context normalization for robust permutation-equivariant learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11286–11295.
[16] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[17] V. Fragoso, P. Sen, S. Rodriguez, and M. Turk, "EVSAC: Accelerating hypotheses generation by modeling matching scores with extreme value theory," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2472–2479.
[18] L. Moisan, P. Moulon, and P. Monasse, "Automatic homographic registration of a pair of images, with a contrario elimination of outliers," *Image Process. Line*, vol. 2, pp. 56–73, May 2012.
[19] Y. Liu, L. Liu, C. Lin, Z. Dong, and W. Wang, "Learnable motion coherence for correspondence pruning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3237–3246.

[20] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[21] Q. Zhang *et al.*, "Dense attention fluid network for salient object detection in optical remote sensing images," *IEEE Trans. Image Process.*, vol. 30, pp. 1305–1317, 2021.
[22] Z. Zhang, Z. Lin, J. Xu, W.-D. Jin, S.-P. Lu, and D.-P. Fan, "Bilateral attention network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 1949–1961, 2021.
[23] F. Zhang *et al.*, "ACFNet: Attentional class feature network for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6798–6807.
[24] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.
[25] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[26] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.
[27] H. Zhang *et al.*, "ResNeSt: Split-attention networks," 2020, *arXiv:2004.08955*.
[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[29] H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang, "Context contrasted feature and gated multi-scale aggregation for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2393–2402.
[30] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.
[31] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.
[32] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5099–5108.
[33] R. Ranftl and V. Koltun, "Deep fundamental matrix estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 284–299.
[34] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperGlue: Learning feature matching with graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4938–4947.

**Linxin Zheng** received the bachelor's degree in information and computing science from Huaqiao University, China, in 2019. He is currently pursuing the M.S. degree with Fuzhou University. He is also attached to the Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University. His research interests include computer vision, machine learning, and pattern recognition.

**Guobao Xiao** (Member, IEEE) received the B.S. degree in information and computing science from Fujian Normal University, China, in 2013, and the Ph.D. degree in computer science and technology from Xiamen University, China, in 2016. From 2016 to 2018, he was a Postdoctoral Fellow with the School of Aerospace Engineering, Xiamen University. He is currently a Professor at Minjiang University, China. He has published over 50 papers in the international journals and conferences, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, IEEE TRANSACTIONS ON MULTIMEDIA, *IJCV*, *PR*, ICCV, and ECCV. His research interests include machine learning, computer vision, and pattern recognition. He has been awarded the Best Ph.D. Thesis in Fujian Province and the Best Ph.D. Thesis Award in China Society of Image and Graphics (a total of ten winners in China). He has served on the Program Committee (PC) for CVPR, ICCV, and ECCV.

**Ziwei Shi** received the B.S. degree in automation from Chongqing University, Chongqing, China, in 2018. He is currently pursuing the master's degree with the School of Mechanical and Electronic Engineering, Fujian Agriculture and Forestry University, Fuzhou, China. He is also attached to the Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University. His current research interests include computer vision and image matching.

**Shiping Wang** (Member, IEEE) received the Ph.D. degree from the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China, in 2014.

He was a Research Fellow at Nanyang Technological University, Singapore, from 2015 to 2016. He is currently a Full Professor and a Qishan Scholar with the College of Mathematics and Computer Science, Fuzhou University, Fuzhou, China. His research interests include machine learning, computer vision, and granular computing.

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