Multi-Stage Network With Geometric Semantic Attention for Two-View Correspondence Learning

Shuyuan Lin, *Member, IEEE*, Xiao Chen, Guobao Xiao, *Senior Member, IEEE*,
Hanzi Wang, *Senior Member, IEEE*, Feiran Huang, *Member, IEEE*,
and Jian Weng, *Senior Member, IEEE*

***Abstract*— The removal of outliers is crucial for establishing correspondence between two images. However, when the proportion of outliers reaches nearly 90%, the task becomes highly challenging. Existing methods face limitations in effectively utilizing geometric transformation consistency (GTC) information and incorporating geometric semantic neighboring information. To address these challenges, we propose a Multi-Stage Geometric Semantic Attention (MSGSA) network. The MSGSA network consists of three key modules: the multi-branch (MB) module, the GTC module, and the geometric semantic attention (GSA) module. The MB module, structured with a multi-branch design, facilitates diverse and robust spatial transformations. The GTC module captures transformation consistency information from the preceding stage. The GSA module categorizes input based on the prior stage's output, enabling efficient extraction of geometric semantic information through a graph-based representation and inter-category information interaction using Transformer. Extensive experiments on the YFCC100M and SUN3D datasets demonstrate that MSGSA outperforms current state-of-the-art methods in outlier removal and camera pose estimation, particularly in scenarios with a high prevalence of outliers. Source code is available at** https://shuyuanlin.github.io.

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

I. Introduction

TWO view correspondence, also known as image matching, is an essential task in computer vision [1], [2]. Its purpose is to establish point-to-point correspondences between different images of the same scene. This task plays a crucial role in various vision applications, including 3D reconstruction [3], [4], simultaneous localization and mapping [5], [6], and image retrieval [7], [8], among others. The conventional process of two-view correspondence involves three essential steps: feature point extraction, matching, and outlier removal. In this paper, our main focus is on outlier removal. The sparsity and irregular distribution of feature points extracted from images often result in a significant number of false correspondences in the initial matches [9]. Therefore, removing outliers is of paramount importance to ensure accurate image matching.

Currently, two main factors contribute to the challenges in outlier removal. Firstly, when the initial correspondence contains a large number of outliers (e.g., 90%) distinguishing inliers becomes a significant challenge. Secondly, the depiction of the initial correspondence as coordinate combinations for corresponding feature points hinders the extraction of meaningful insights from the data. To address the first challenge, previous methods have utilized iterative multi-stage networks for gradual outlier removal. However, their effectiveness in removing outliers is limited due to insufficient consideration of inter-stage relationships. The preceding stage can provide valuable priors for the subsequent stage, which should be taken into account. Regarding the second challenge, considering the characteristics of the initial correspondence set is crucial. Firstly, the relationship between the coordinates of feature points in each match is important, as correct matches satisfy geometric constraints (e.g., epipolar constraints), while incorrect matches do not exhibit such constraints. Secondly, correct matches demonstrate geometric transformation consistency (GTC) under geometric transformations since they all satisfy the same geometric constraints. Lastly, it is observed that geometrically and semantically neighboring matching pairs may not be adjacent in space, highlighting the importance of incorporating neighbor information in deep learning. Therefore, mining neighbor information, specifically geometric semantic information, becomes highly necessary. In summary, addressing these challenges requires leveraging inter-stage relationships and capturing GTC and geometric semantic information. However, existing methods have not adequately addressed these aspects.

This work was supported in part by the National Natural Science Foundation of China under Grant U22A2095, Grant U21A20514, Grant U22B2028, Grant 62272200, and Grant 61825203; in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515011740; in part by Guangdong Key Laboratory of Data Security and Privacy Preserving under Grant 2023B1212060036; and in part by Guangdong-Hong Kong Joint Laboratory for Data Security and Privacy Preserving under Grant 2023B1212120007. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Andrea Fusiello. *(Corresponding authors: Feiran Huang; Jian Weng.)*

Shuyuan Lin, Xiao Chen, Feiran Huang, and Jian Weng are with the College of Cyber Security/College of Information Science and Technology, Jinan University, Guangzhou 510632, China (e-mail: huangfr@jnu.edu.cn; cryptjweng@gmail.com).

Guobao Xiao is with the School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China.

Hanzi Wang is with Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, Xiamen 361005, China.

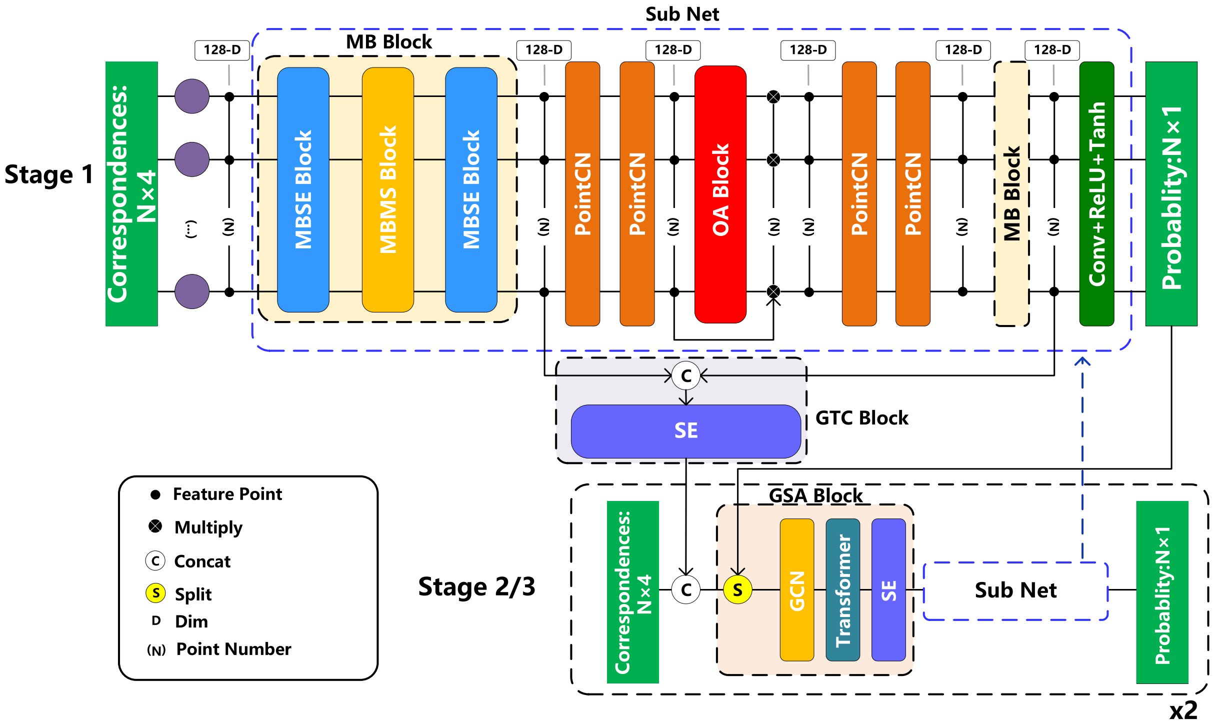


Fig. 1. Overall architecture of the proposed MSGSA.

In the past decades, traditional outlier removal methods such as RANSAC [10], VFC [11], PROSAC [12], NAPSAC [13] and some model fitting methods [14], [15], [16] have demonstrated good performance in scenarios with a small number of outliers. However, their performance significantly deteriorates when the number of outliers exceeds a certain threshold. In recent years, deep learning-based methods have been proposed for outlier removal [17], [18], [19], [20], [21], [22], [23]. These methods aim to transform the outlier removal problem into a binary classification or essential matrix regression problem. These deep learning-based methods can be broadly categorized into two types: single-stage network-based methods and multi-stage network-based methods. Single-stage network-based methods, such as LFGC [17] and ACNe [18], have shown improvements in outlier removal for scenarios with a high number of outliers compared to traditional methods. However, their improvement is still limited. On the other hand, multi-stage network-based methods, for example, OANet [19], T-Net [20], PESA-Net [21], MSA-Net [24], and MS^{2}DG-Net [25], have demonstrated significantly better feature extraction capabilities and robustness against outliers. However, the relationship between different stages in multi-stage networks has not been fully explored. For example, in a multi-stage network structure, the information from the previous stage can be considered as prior information for the next stage. If the next stage can obtain more useful prior information from the previous stage, then the difficulty of the task for the next stage will be significantly reduced. These methods rely on concatenating the output of the previous stage with the input of the next stage, which constrains their robustness in removing outliers.

To leverage inter-stage relationships and extract geometric transformation consistency (GTC) information and geometric semantic information effectively, we propose a novel network called Multi-Stage Geometric Semantic Attention (MSGSA). As shown in Fig. 1, MSGSA consists of three key modules: the multi-branch (MB) module, the geometric semantic attention (GSA) module, and the GTC module. The MB consists of two types of MB networks, as illustrated in Fig. 2. The first type, known as Multi-Branch with Squeeze-and-Excitation (MBSE), combines the squeeze-and-excitation (SE) module [26] with a local information extraction module. The second type of MB network is referred to as Multi-Branch with Multi-Scale (MBMS). It is composed of three parallel branches: a global spatial transformation branch based on global average pooling, a global spatial transformation branch based on global max pooling, and a local spatial transformation branch. The MB module combines global and local spatial transformations along with channel refinement to provide diverse and robust spatial transformations, enabling effective discrimination between inliers and outliers. The GTC module is positioned between two stages and takes advantage of intermediate features from the previous stage. It extracts geometric transformation consistency information from these features and combines them with the input of the next stage. The GSA module exploits geometric semantic neighboring information using the probabilities of correct matching pairs generated by the previous stage. It incorporates graph neural networks [27] and transformer modules [28], [29] to aggregate information within different semantic classes and facilitate information interaction between classes.

In summary, the main contributions of our proposed method can be summarized as follows:

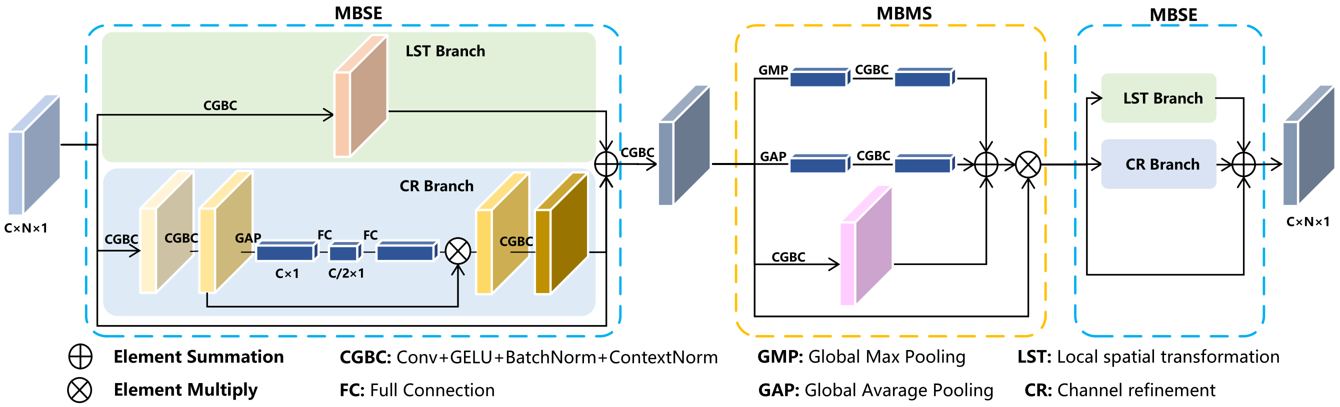


Fig. 2. Structure of the multi-branch module.

- We propose an MB module that enhances the discrimination between correct and incorrect matching pairs through channel refinement and the combination of global and local spatial transformations.
- We propose a GTC module that leverages geometric transformation consistency information from the previous stage to guide feature extraction in the subsequent stage, thereby effectively mitigating the impact of outliers.
- We present a GSA module that effectively explores neighboring information associated with semantic relationships. This module not only enhances the performance of outlier removal but also significantly improves the accuracy of camera pose estimation.

The remaining sections of this paper are structured as follows: In Section II, we provide a comprehensive overview of the relevant literature on handcrafted methods, learning-based methods, and attention mechanisms. Next, in Section III, we present the specific details of our proposed method. Subsequently, in Section IV, we provide the experimental results we have obtained. Finally, we conclude the paper in Section V by summarizing our findings.

II. Related Work

In this section, we review the methods for removing outliers based on handcrafted features, learning-based methods, and attention mechanisms.

A. Handcrafted Methods

Outlier removal methods based on handcrafted features can be divided into two groups: hypothesis-based and validation-based resampling algorithms (e.g., RANSAC [10] and its variants [12], [13], [30], [31], [32]) and non-parametric model-based algorithms (e.g., VFC [11], MRCVF [33], L_2E [34], ICF [35]). RANSAC employs random sampling to find the best model through multiple verifications. However, random sampling often requires a large number of iterations to achieve significant results. To address this issue, PROSAC [12] enhances RANSAC by calculating the quality score of matching pairs using an evaluation function, sorting the pairs, and sampling from the progressively matching set to reduce computational costs. NAPSAC [13] assumes that correct matching pairs are closer to other correct matching pairs in the set than to incorrect matching pairs. It achieves this by projecting data onto an n-dimensional hypersphere and gradually reducing the radius of hypotheses and samples to eliminate matching pairs with outliers. USAC [31] introduces a unified sampling framework that combines the strengths of several RANSAC algorithms and their variants. However, the effectiveness of resampling algorithms heavily relies on the quality of initial matches and diminishes significantly with an increasing number of mismatched pairs. On the other hand, non-parametric model-based algorithms, such as VFC [11] make assumptions about shared properties among correct matching pairs and use consistency measures to remove false matches. However, VFC performs worse than resampling algorithms in scenarios involving image deformation. While handcrafted-based outlier removal algorithms work well in scenarios with a relatively small number of outliers, their effectiveness significantly decreases when the proportion of outliers is much larger than inliers. This limitation prompts a shift in focus towards deep learning-based methods, which have demonstrated robustness to outliers and align with the method adopted in this paper.

B. Learning-Based Methods

To enhance the robustness to outliers, several deep learning-based methods have been proposed. For example, DSAC [36] mimics the iterative resampling behavior of RANSAC using deep learning but fails to achieve significant improvement in outlier removal performance. Conversely, LFGC [17] introduces context normalization to capture global information, neglecting local information between samples. In response to this, ACNe [18] enhances context normalization through an iterative least squares method and introduces local attention. However, the localized attention mechanism in ACNe focuses only on a single matching pair, which is insufficient for capturing neighborhood information effectively. It is important to note that these methods adopt a single-stage network-based structure, which limits their feature extraction capabilities in scenes with a large number of outliers. To address these constraints, contemporary methodologies have pivoted towards a multi-stage network-based structure. For example, OANet [19] integrates global and local information effectively using an ordered "aggregation-segmentation" strategy and a geometry-aware filtering module. T-Net [20] utilizes a T-shaped structure to integrate output features from each stage, enabling the extraction of meaningful information. MSA-Net [24] incorporates a multi-scale attention mechanism into a multi-stage network. MS^{2}DG-Net [25] utilizes a sparse semantic dynamic graph network to extract sparse semantic information for matching pairs with similar semantics. Furthermore, contemporary methodologies such as CSR-Net [22] reframe the matching problem into a holistic evaluation of dynamic local structure consensus in an end-to-end fashion. Shape-Former [23] combines convolutional neural network (CNN) and transformer capabilities to enhance the representational prowess of structure consensus. PGF-Net [37] proposes an iterative filtering structure to progressively extract more reliable candidates from initial correspondences. GCA-Net [38] introduces a graph context attention mechanism to generate structure-aware graphs with high resilience to outliers. SGA-Net [39] incorporates a graph attention block to capture global and per-graph contextual information from dynamic graphs. While these multi-stage network-based structures improve the feature representation ability of the network, they simply concatenate the output of the previous stage with the input of the next stage, overlooking geometric transformation consistency information and failing to effectively extract and utilize semantic neighboring information. These limitations restrict their effectiveness in outlier removal. In contrast, the proposed method in this paper takes into account both geometric transformation consistency information and geometric semantic neighboring information. Additionally, the MB module proposed in this paper provides a more comprehensive and robust spatial transformation in comparison to previous models. These advantages significantly improve the effectiveness of outlier removal and the accuracy of camera pose estimation.

C. Attention Mechanisms

Attention mechanisms [28], [40] are closely linked to learning-based methods, specifically, the MSGSA method, which integrates two attention mechanisms: (1) channel attention and (2) self-attention. Channel attention, a commonly used technique in convolutional neural networks, enhances network performance by evaluating the significance of each channel. Notably, the SE-Net (Squeeze-and-Excitation Network) [26] is a prevalent implementation of channel attention, learning inter-channel relationships and assigning importance to each channel. Variations such as the SK-Net (Selective Kernel Network) [41] introduce dynamic selection mechanisms, while MobileNetV3 [42] utilizes SE blocks and hard-swish activation functions for lightweight attention modules. ECA-Net (Efficient Channel Attention Network) [43] replaces the fully connected layer in the SE with an adaptive kernel size. In this paper, we propose the MBSE and GTC modules leveraging SE blocks to learn channel relationships.

Self-attention, on the other hand, computes the relationships between each element and other elements in a sequence. It allows for interactions between different elements, with their importance determined through weighted calculations. The Transformer [28], initially introduced for natural language processing, has been successfully applied to image processing tasks as well, as seen in the Vision Transformer (ViT) [29]. The Transformer is powerful in capturing global information, but computationally expensive for long sequences. To address this, we proposed the GSA module, which employs classification aggregation for matching pairs and leverages the Transformer for information exchange between category nodes.

III. Methods

This study addresses the problem of two-view correspondence learning, focusing on outlier removal and camera pose estimation. In this section, we present the framework designed to achieve these objectives.

A. Problem Formulation

For a pair of images, denoted as I_1 and I_2, obtained from the same scene, their features are represented by feature points. Initially, either a traditional algorithm [44], [45], [46], [47] or a deep learning-based algorithm [35], [48] is used to acquire a set of feature point coordinates for the images, along with their corresponding descriptors. Subsequently, initial matching pairs between the images are generated using the nearest neighbor algorithm. Mathematically, the normalized coordinates of these matching pairs can be represented as $S = [s_1, s_2, s_3, \ldots, s_N] \in \mathbb{R}^{N\times4}$, where N is the number of matching pairs, and $s_i = (x_i, y_i, x_i', y_i')$ represents the normalized coordinates of a matching pair. Here, (x_i, y_i) corresponds to the normalized coordinates of a feature point in the i-th matching pair in the image I_1, and (x_i', y_i') corresponds to the normalized coordinates of the corresponding feature point in the i-th matching pair in image I_2.

In the context of two-view correspondence learning, our objective is to address two tasks: outlier removal and camera pose estimation. To address these tasks, as suggested by [17], we break down the problem into two subproblems: (1) binary classification of matching pairs and (2) essential matrix regression for camera pose estimation. We propose the application of a neural network to handle the binary classification problem. The neural network predicts the probability of correctness for each matching pair, and these probabilities are utilized to calculate weights for each pair. For the essential matrix regression problem, we combine these weights with the eight-point algorithm [49] to compute the essential matrix E, defined as follows:

$$E = g(S, tanh(ReLU(f_\varphi(S)))), \tag{1}$$

where f_φ represents the proposed neural network framework (MSGSA) and φ denotes the parameter of MSGSA; g represents the improved eight-point weighting algorithm; $tanh$ and $ReLU$ are activation functions.

B. Multi-Branch Module

To effectively leverage the geometric characteristics embedded in the initial correspondence set data, we introduce the

MB module. This module is designed to offer comprehensive and robust spatial transformations by incorporating channel refinement in the channel dimension and a combination of global and local spatial transformations in multiple dimensions. As depicted in Fig. 2, the MB module comprises two distinct types of MB networks: 1) MBSE and 2) MBMS. The input data undergoes a primary transformation via the MBSE module, followed by a secondary transformation through the MBMS module, and ultimately experiences a tertiary transformation through the MBSE module.

1) Multi-Branch With Squeeze-and-Excitation: The MBSE module encompasses two branches: (i) the channel refinement branch and (ii) the local spatial transformation branch. The output features from these branches are combined through an addition operation and are concurrently linked to the input features via a residual connection.

(i) Channel refinement. The channel refinement branch employs the SE module to learn and assign weights to inter-channel dependencies and individual channel features, thereby refining the features in the channel dimension. The input features undergo two Convolution + GELU + Batch Normalization + Context Normalization (CGBC) layers, denoted as $F_b(a) = Conv(GELU(BN(CN(a))))$, where a represents the input of CGBC. Following this, the SE module is applied to enhance channel refinement. The SE module can be mathematically defined as follows [26]:

$$v = F_{sq}(u) = \frac{1}{N \times 1}\sum_{i=1}^{N}\sum_{j=1}^{1} u(i, j), \tag{2}$$

$$h = F_{ex}(v) = \sigma(W_2\delta(W_1 v)), \tag{3}$$

where $u \in \mathbb{R}^{N\times 1\times C}$ represents the input of SE module; N is the number of matching pairs; C is the number of channels for each matching pair; and $v \in \mathbb{R}^{1\times 1\times C}$ represents the output of the squeeze operation F_{sq}. The excitation operation is denoted by F_{ex}, which consists of two linear layers represented by $W_1 \in \mathbb{R}^{\frac{C}{r}\times C}$ and $W_2 \in \mathbb{R}^{C\times\frac{C}{r}}$, where r is the channel reduction ratio. The sigmoid function σ and the $ReLU$ activation function δ are applied during the excitation operation. The excitation operation yields the channel weight parameter h, which is multiplied element-wise with the input feature u to achieve channel refinement.

(ii) Local spatial transformation branch. This branch captures detailed local information to provide spatial transformations across different dimensions. It applies CGBC operations exclusively to the input features.

2) Multi-Branch With Multi-Scale: The MBMS module consists of three branches: the global spatial transformation branch based on global max pooling, the global spatial transformation branch based on global average pooling, and the local spatial transformation branch. Global average pooling is effective in aggregating global information, but it is not robust to outliers. To address this, we introduce global max pooling to emphasize high response values and compensate for the limitations of global average pooling. The output features of these three branches are combined through an addition operation. The formulation can be described as follows:

$$q = F_b(GAP(e)) + F_b(GMP(e)) + F_b(e), \tag{4}$$

where $F_b(\cdot)$ represents the CBGC layer; $GAP(\cdot)$ represents global average pooling; $GMP(\cdot)$ represents global max pooling; and e represents the input of MBMS. The final step involves merging the output feature q with the input feature of MBMS through a multiplication operation.

C. Geometric Transformation Consistency Module

The GTC module is designed to capture the influence of geometric constraints, such as epipolar constraints, on correct match pairs. The transformation process of these pairs exhibits GTC under spatial transformations. To effectively extract this information, the GTC module operates in two stages, capturing intermediate transformation process information before and after spatial transformations in the previous stage. The obtained transformation process information is then fused into the input of the next stage.

Initially, the matching set space has a dimensionality of 4, and to enhance the feature capacity, a 1 × 1 convolution operation is applied, increasing it to 128 dimensions. Subsequently, using the MB module, we obtain the features in the feature space Z_1. Through a series of spatial transformations (PointCN×2, OA module, PointCN×2, and MB module) denoted as F_{tr}, these features are further transformed from Z_1 to the feature space Z_2, represented as $F_{tr} : Z_1 \rightarrow Z_2$. The GTC module incorporates a basic SE module, and its operation process can be described as follows:

$$f_c^l = SE(concat(f_{Z_1}^l, f_{Z_2}^l)), \tag{5}$$

$$f^{l+1} = concat(f_c^l, f_{in}^{l+1}), \tag{6}$$

where, $f_{Z_1}^l$ represents the features of the feature space Z_1; $f_{Z_2}^l$ represents the features of the feature space Z_2; and $l \in [1, 2]$ denotes the stage. The *concat* operation refers to concatenation along the channel dimension. Furthermore, f_c^l represents the extracted GTC information; f_{in}^{l+1} represents the input for the following stage; and f^{l+1} represents the fused features that incorporate the GTC information.

D. Geometric Semantic Attention Module

To effectively capture semantically adjacent neighbor information, we propose the GSA module, depicted in Fig. 3. This module operates during the second and third stages of the model and consists of five steps: (a) neighbor selection, (b) category graph construction, (c) neighbor feature aggregation, (d) category information interaction, and (e) feature fusion.

(a) Neighbor selection: Probability maps obtained from the final output of each stage guide the selection of neighbors. These maps indicate that the probabilities between geometrically and semantically adjacent matching pairs are close. Utilizing these probability values, we partition the input matching pairs of the next stage into k classes, where each class consists of mutually adjacent pairs.

(b) Category graph construction: For each class, we construct a category graph $G_i = (H_i, D_{ij})$ by creating a category

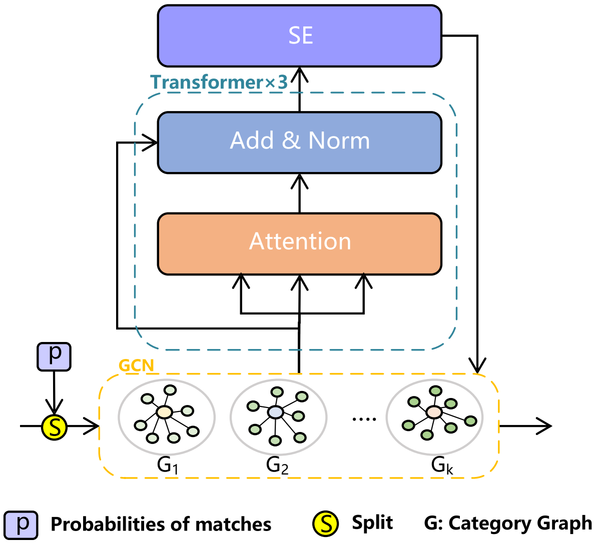


Fig. 3. Structure of the geometric semantic attention module.

node connected to its corresponding matching pairs. Here, G_i represents the category graph for the i-th class, H_i represents the category node of the i-th class, and D_{ij} represents the edge between category node i and matching pair j. The category nodes act as intermediaries for information exchange among the matching pairs in the same class. Each category graph (G_i) represents a set of matching pairs that share similar characteristics, with the nodes within each category graph being neighbors to one another. By grouping, refining, and aggregating matching pair nodes based on different features, the model acquires a more nuanced comprehension of the intrinsic architecture of the data.

(c) Neighbor feature aggregation: We obtain the initial feature of each category node by fusing the features of all matching pairs within that category. This fusion is accomplished through average pooling, which can be seen as a one-hop convolution in graph neural networks(GCN).

(d) Category information interaction: To further leverage the information from different categories, we apply transformer-based attention to all category nodes. The attention process can be mathematically defined as follows:

$$ATT_{ij}(\mathbf{Q}_i, \mathbf{K}_j, \mathbf{V}_j) = softmax(\frac{\mathbf{Q}_i\mathbf{K}_j^T}{\sqrt{d}})\mathbf{V}_j, \tag{7}$$

$$f_{H_i} = \sum_{j=1}^{k} ATT_{ij}(\mathbf{Q}_i, \mathbf{K}_j, \mathbf{V}_j), \tag{8}$$

where ATT_{ij} represents the attention of the i-th category node to the j-th category node; $\mathbf{Q}_i$ represents the query vector of the i-th category node, while $\mathbf{K}_j$ represents the key vector of the j-th category node; d represents the temperature coefficient; and $\mathbf{V}_j$ represents the value vector of the j-th category node. The output f_{H_i} represents the result of the i-th category node after being attended to by all category nodes. Finally, we refine the channel information of each category node using the SE module.

(e) Feature fusion: Following the aforementioned steps, each category node collects information from neighboring matching pairs and integrates global information. Then, we fuse the features of each category node with the features of all matching pairs in the corresponding category. This process enables each matching pair to possess both neighbor information and global information.

E. Loss Function

To formulate two-view correspondence learning as a multi-task problem, encompassing binary classification and essential matrix regression, we introduce a hybrid loss function that integrates both classification and regression losses. The hybrid loss function is defined as follows [50] and [51]:

$$Loss = l_c(W, L) + \gamma l_e(\hat{E}, E), \tag{9}$$

where l_c represents the loss function for the dichotomous classification problem with cross-entropy loss and l_e represents the loss function for the essential matrix regression problem; γ is a hyperparameter that balances the two loss functions. The hyperparameter γ is set to 0.5, following the previous works [19], [20], [38]. The l_e is specified as follows:

$$l_e(\hat{E}, E) = \frac{(p'^T \hat{E} p)^2}{\|Ep\|_{[1]}^2 + \|Ep\|_{[2]}^2 + \|Ep'\|_{[1]}^2 + \|Ep'\|_{[2]}^2}, \tag{10}$$

where p and p' are sets of correspondences of matching pairs; E and $\hat{E}$ represent the predicted essential matrix and ground truth essential matrix, respectively; and $\| \cdot \|_{[i]}$ denotes the element at the i-th position in the vector.

IV. Experiments

A. Implementation Details

As shown in Fig. 1, the input data comprises N (approximately 2000) initial matching pairs, where each pair contains 4 channels. In the first stage, the initial matching pairs undergo a dimensional expansion operation through a 1×1 convolution, resulting in a transformation into a 128-dimensional space. Subsequently, they are sequentially processed through the MB module, PointCN×2, OA module, PointCN×2, and MB module. The PointCN and OA modules are adopted from OANet. The second and third stages of the network differ from the first stage in that they incorporate geometric transformation consistency information extracted from the previous stage, following the expansion of the input to 128 dimensions. This integrated information is then processed using the GSA module. Finally, the output features of the three stages are fused using two MBSE modules for feature integration. The network is implemented using PyTorch, with a batch size of 32. The network parameters are optimized using the Adam optimizer with a learning rate of 10^{-3}. Following an iterative network method [19], the network undergoes training for 500k iterations. All experiments are conducted on NVIDIA GTX 3090 GPUs.

B. Datasets

To evaluate our method, we used two datasets: the YFCC100M dataset [52]) for indoor scenes, and the SUN3D dataset [53]) for outdoor scenes.

The YFCC100M dataset, introduced in the work [52], is a comprehensive collection of online media objects. It comprises a vast number of images and videos, including approximately 99.2 million images and 800,000 videos. We utilized a subset of this dataset consisting of 72 different indoor scene sequences, which were classified into various categories. During our experiments, we considered 68 of these scene sequences as known scenes, while the remaining four sequences were treated as unknown scenes. We divided the known scene sequences into training, validation, and testing sets, with proportions of 60%, 20%, and 20%, respectively. The training set was used to train our model, and the testing set, composed of the four unknown scene sequences, was used to evaluate the generalization performance of our method.

The SUN3D dataset, introduced in the work [53], is a large-scale RGB-D video database primarily focusing on outdoor scenes. The dataset we used in this study comprises a diverse collection of scene sequences captured using RGB-D sensors. Specifically, the dataset includes 253 scene sequences, of which we selected 239 sequences as known scenes for our experiments. Similar to the YFCC100M dataset, we divided the known scene sequences into training, validation, and testing sets, with proportions of 60%, 20%, and 20%, respectively. We evaluated the generalization capacity of our method using a testing set consisting of four sequences representing unknown scenes.

C. Evaluation Metrics

Our method is evaluated on two dimensions: outlier removal and camera pose estimation. Therefore, we employ different evaluation metrics to assess the performance in each aspect. For outlier removal, we use *Precision* (*P*), *Recall* (*R*), and *F-score* (*F*) as evaluation metrics; for camera pose estimation, we use the average accuracy (mAP) calculated under the error thresholds of 5° and 20° as evaluation metrics.

D. Outlier Removal

As shown in Table I, we conducted a comparison between our method MSGSA and 12 alternative methods: RANSAC [10], PointNet++ [54], LFGC-Net [17], DEF-Net [50], ACNe-Net [18], OANet [19], T-Net [20], PESA-Net [21], MSA-Net [24], MS^{2}DG-Net [25], GCA-Net [38], and PGF-Net [37]. Among them, RANSAC, PointNet++, LFGC-Net, DEF-Net, and ACNe-Net are earlier methods, and their experimental data are adopted from T-Net. For the remaining methods, the experimental results are obtained by running the authors' code on the same local server and the same random image pairs. It can be observed that the deep learning-based methods (i.e., PointNet++, LFGC-Net, DEF-Net, ACNe-Net, OANet, T-Net, PESA-Net, MSA-Net, MS^{2}DG-Net, GCA-Net and PGF-Net) outperform the handcrafted-based method RANSAC in terms of precision, recall, and F-score. Notably, deep learning-based methods generally achieve a recall metric over 30% higher than RANSAC. The low recall metric of RANSAC can be attributed to the fact that, in scenarios where the number of outliers (nearly 90%) is significantly higher than the number of inliers, RANSAC tends to remove more inliers when selecting the model. Deep learning-based methods, with their stronger feature representation capabilities, effectively learn the characteristics of matching pairs and can better differentiate between correct and incorrect matches. For the known scene of the YFCC100M dataset, MSGSA outperforms RANSAC by 16.13%, 38.65% and 25.33% in the P, R, and F metrics, respectively. Furthermore, MSGSA significantly outperforms other deep learning-based methods. For instance, compared with the representative state-of-the-art method GCA-Net, in the known scene, MSGSA has improved by 0.60% and 0.26%, respectively, in the P and F metrics. In the unknown scene, MSGSA has improved by 0.66% and 0.32%, respectively, in the P and F metrics. For the SUN3D dataset, the performance of MSGSA is slightly lower than the first-ranked methods in the three metrics (i.e., P, R and F). This is because the SUN3D scenes are relatively blurry, lack texture, and contain numerous similar structures, making outlier removal more difficult in such challenging scenarios. Fig. 4 visualizes the outlier removal performance of MSGSA on the YFCC100M dataset, demonstrating MSGSA's stronger capability compared to current methods.

To examine the contributions of different modules at each stage, we employed *Principal Component Analysis* (*PCA*) to reduce the dimensionality of multiple intermediate features in the three stages and visualized them, as shown in Fig. 5. It can be observed that the initial matching pairs contain a mixture of inliers and outliers. After undergoing processing by the 1 × 1 convolutional layer, the inliers start to form distinct clusters. After each stage of the MB module, the separation between inliers and outliers becomes more pronounced. Successive stages of the MB module accentuate the distinction between inliers and outliers, particularly in the last two stages, where they are distinctly separated and clustered.

E. Camera Pose Estimation

As shown in Table II, we conducted a comparison between our proposed MSGSA and 12 other methods in the camera pose estimation task. We additionally compared SGA-Net [39], which uses an improved eight-point algorithm (including a modified weight calculation strategy and an additional dynamic weight loss function) to compute the essential matrix. It is evident that the handcrafted-based RANSAC algorithm performs significantly worse compared to deep learning-based methods. This is not only due to its lower accuracy in outlier removal but also because RANSAC retains fewer inliers (lower recall) after outlier removal. In contrast, deep learning-based methods achieve notably better results in the camera pose estimation task, primarily due to their higher recall when compared to RANSAC.

The proposed MSGSA method demonstrates superior performance on both the YFCC100M and SUN3D datasets, surpassing the other 13 competing methods. For example, for the YFCC100M dataset, MSGSA outperforms RANSAC by

TABLE I

Comparison of Outlier Removal Results on the YFCC100M and SUN3D Datasets. The Optimal Indicator Values Are in Bold, and the Second-Best Indicators Are Underlined

Dataset	YFCC100M (%)						SUN3D (%)					
Method	Known Scene			Unknown Scene			Known Scene			Unknown Scene		
	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
RANSAC [10]	47.35	52.39	49.47	43.55	50.65	46.83	51.87	56.27	53.98	44.87	48.82	46.76
PointNet++ [54]	49.62	86.19	62.98	46.39	84.17	59.81	52.89	86.25	65.57	46.30	82.72	59.37
LFGC-Net [17]	54.43	86.88	66.93	52.84	85.68	65.37	53.70	87.03	66.42	46.11	83.92	59.52
DFE-Net [50]	56.72	87.16	68.72	54.00	85.56	66.21	53.96	87.23	66.68	46.18	84.01	59.60
ACNe-Net [18]	60.02	88.99	71.69	55.62	85.47	67.39	54.11	88.46	67.15	46.16	84.01	59.58
OANet [19]	59.31	88.42	71.00	55.78	85.09	67.39	55.70	87.82	68.17	47.75	83.50	60.76
T-Net [20]	61.76	89.62	73.12	57.63	86.57	69.19	55.35	88.43	68.08	47.17	84.05	60.43
PESA-Net [21]	61.43	89.63	72.90	58.02	87.01	69.62	55.08	88.56	67.92	47.29	84.81	60.72
MSA-Net [24]	59.27	90.82	71.73	56.23	89.10	68.95	53.45	88.65	66.69	45.72	**85.02**	59.46
MS^{2}DG-Net [25]	62.31	90.37	73.75	58.67	87.67	70.28	55.93	**88.74**	68.61	47.97	84.33	61.15
PGF-Net [37]	60.36	89.88	72.22	57.46	87.87	69.48	54.69	88.55	67.62	46.62	84.74	60.15
GCA-Net [38]	62.88	**91.50**	74.54	59.77	**89.47**	71.66	**56.51**	88.17	**68.88**	**48.54**	83.98	**61.52**
MSGSA	**63.48**	91.04	**74.80**	**60.43**	89.01	**71.98**	55.92	88.56	68.55	47.99	84.32	61.22

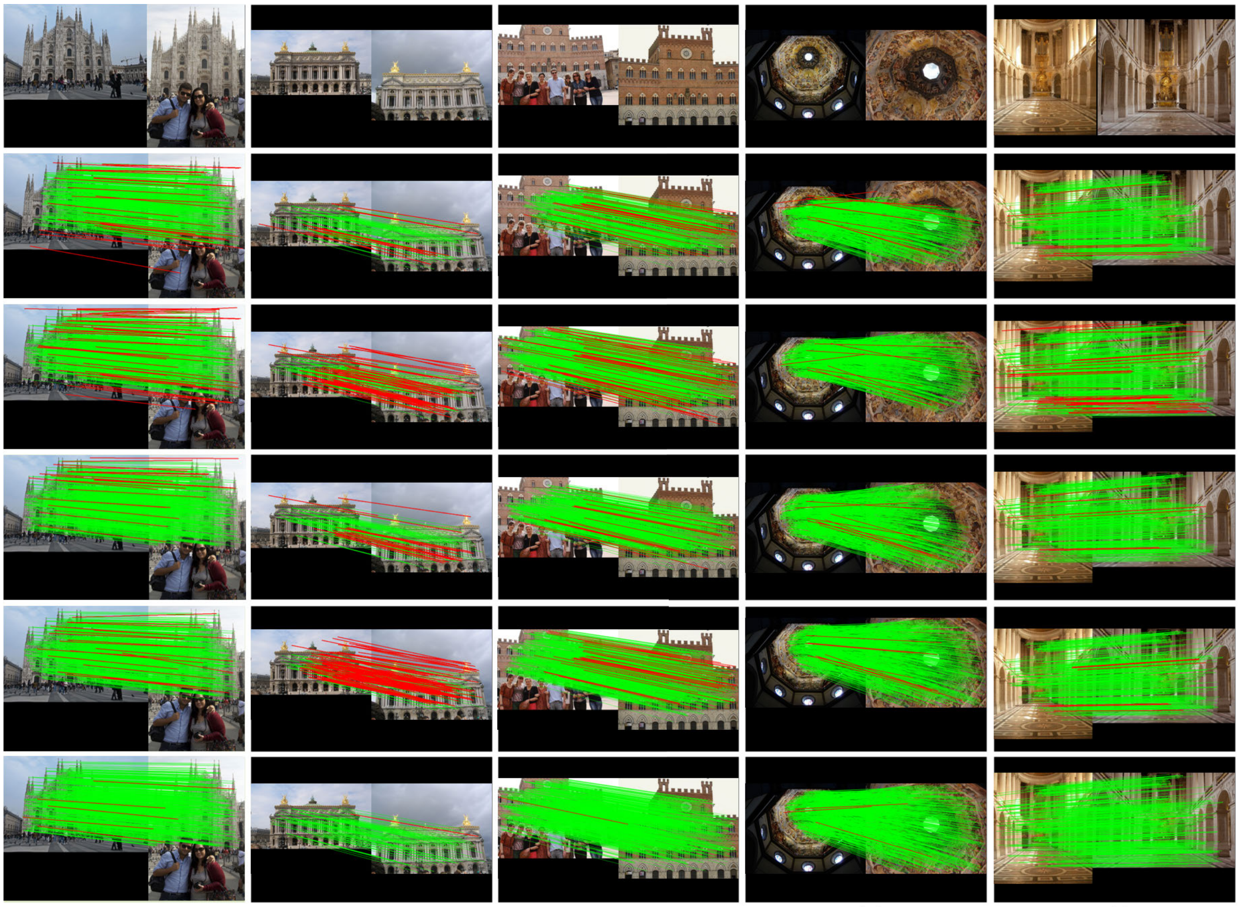

(a) (b) (c) (d) (e)

Fig. 4. Some visualization results of the proposed MSGSA on the YFCC100M dataset. (a) *Milan Cathedral*, (b) *Paris Opera*, (c) *Palazzo Pubblico*, (d) *Florence Cathedral Dome Interior*, and (e) *Palace of Versailles Chapel*. The first to sixth rows are the input image pairs and the visualization results obtained by OANet, T-Net, PESA-Net, MS^{2}DG-Net, and our proposed MSGSA, respectively.

40.2% and 51.00% in the known scenes, and by 48.88% and 56.07% in the unknown scenes, under error thresholds of 5° and 20°, respectively. For the SUN3D dataset, MSGSA outperforms RANSAC by 22.04% and 38.88% in known scenes, and by 18.16% and 36.20% in unknown scenes, under error thresholds of 5° and 20°, respectively. Compared with current state-of-the-art methods, MSGSA also shows significant improvements. For the YFCC100M dataset, compared with

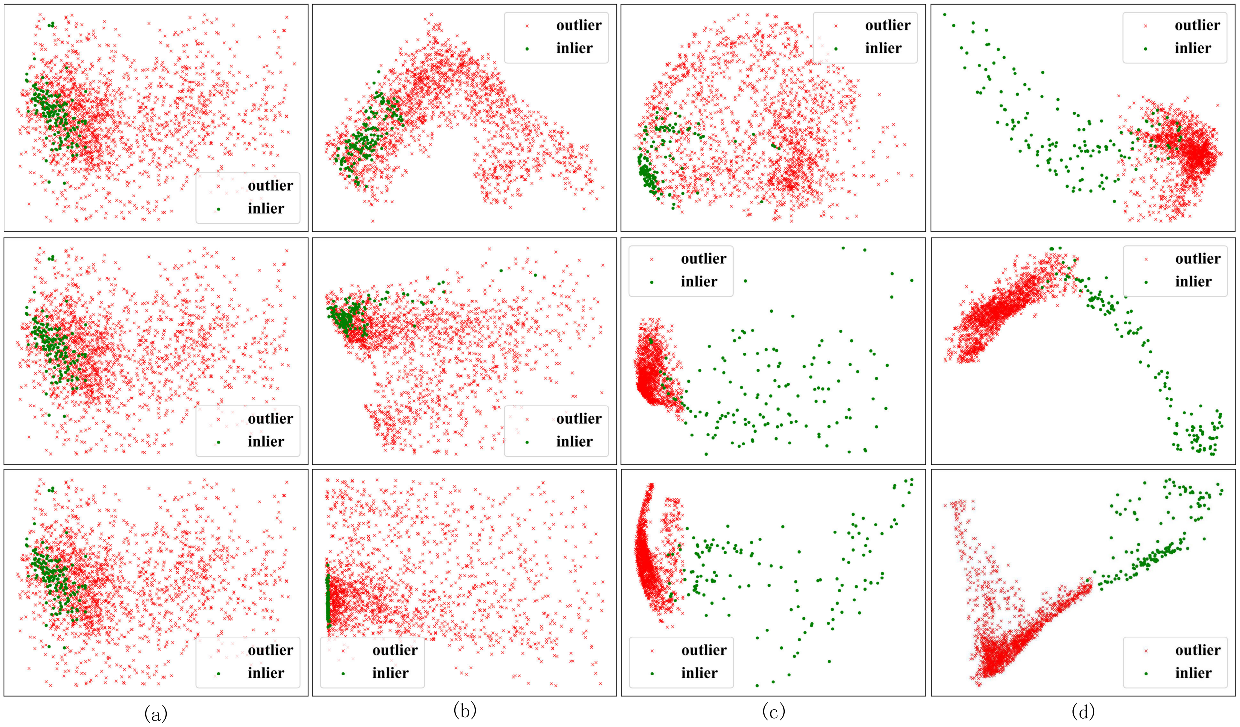


Fig. 5. Some visualization results of the proposed MSGSA. From top to bottom, three rows represent three stages of MSGSA, respectively. (a) The initial feature of the matching pairs, (b) the features after being dimensionally increased through a 1 × 1 convolution, (c) the features after passing through the first multi-branch module, and (d) the features after passing through the second multi-branch module.

TABLE II

Performance Comparison of Camera Pose Estimation on the YFCC100M and SUN3D Datasets Using Error Thresholds of 5° and 20°. The Optimal Indicator Values Are in Bold, and the Second-Best Indicators Are Underlined

Dataset	YFCC100M (%)				SUN3D (%)			
Method	Known Scene		Unknown Scene		Known Scene		Unknown Scene	
	5°	20°	5°	20°	5°	20°	5°	20°
RANSAC [10]	5.74	16.67	9.05	22.71	4.43	15.38	2.85	11.23
PointNet++ [54]	11.88	32.86	15.98	44.82	8.78	31.02	7.22	29.77
LFGC-Net [17]	14.51	35.82	23.71	50.57	11.93	36.03	9.73	33.09
DFE-Net [50]	19.27	42.14	30.55	59.15	14.18	39.14	12.13	26.26
ACNe-Net [18]	29.63	52.71	34.00	62.98	19.08	46.32	14.27	39.29
OANet [19]	33.50	57.53	41.33	68.79	22.41	49.23	17.57	42.61
T-Net [20]	40.86	63.81	46.74	73.11	23.55	50.99	17.69	44.03
PESA-Net [21]	37.15	59.76	45.03	71.95	22.67	50.02	18.00	44.10
MSA-Net [24]	37.40	60.16	48.45	73.23	18.51	45.74	15.26	41.00
MS^{2}DG-Net [25]	37.78	62.78	46.98	75.13	22.93	50.67	17.34	43.41
PGF-Net [37]	37.01	59.63	55.60	75.32	23.97	51.33	17.45	40.06
GCA-Net [38]	43.82	67.21	55.07	$\underline{79.07}$	23.13	50.78	18.05	43.84
MSGSA	$\underline{45.94}$	$\underline{67.67}$	$\underline{57.93}$	78.78	**26.47**	**54.26**	**21.01**	**47.43**
SGA-Net$^+$ [39]	43.66	65.39	57.33	78.74	22.69	48.36	18.78	43.66
MSGSA$^+$	**54.94**	**74.32**	**64.95**	**82.71**	$\underline{25.28}$	$\underline{52.29}$	$\underline{20.41}$	$\underline{46.12}$

"+" indicates that the method uses the improved eight-point weighting algorithm.

the representative state-of-the-art method GCA-Net, MSGSA shows improvements of 2.12% in the known scenes and 2.86% in the unknown scenes, under the error threshold of 5°. For the SUN3D dataset, compared with the representative state-of-the-art method PGF-Net, MSGSA showcases improvements of 2.50% and 2.90% in known scenes and 3.56% and 7.37% in unknown scenes, under the error thresholds of 5° and 20°, respectively. Following modifications to the loss function and eight-point weighting algorithm, SGA-Net experienced a substantial performance boost. In order to ensure a fair comparison, we applied identical modifications to MSGSA. On the YFCC100M dataset, MSGSA demonstrated noteworthy performance enhancements compared to SGA-Net in known scenes, registering increases of 11.5% and 9.29% at 5° and 20° thresholds, respectively. In unknown scenes, MSGSA exhibited improvements of 6.25% and

TABLE III

Statistical Analysis of Representative Methods for Outlier Removal Metrics on the yfcc100m Datasets. μ and σ^2 Indicate the Mean and Variance, Respectively. The Optimal Indicator Values Are in Bold

Dataset	YFCC100M (%)											
Method	Known Scene						Unknown Scene					
	P		R		F		P		R		F	
	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2
T-Net	61.76	0.49	89.62	0.18	73.12	0.13	57.63	0.47	86.57	0.34	69.19	0.11
MS^{2}DG-Net	62.31	1.37	90.37	0.10	73.75	0.53	58.67	1.82	87.67	0.33	70.28	0.65
GCA-Net	62.88	0.39	**91.50**	**0.04**	74.54	0.15	59.77	0.50	**89.47**	0.21	71.66	0.16
MSGSA	**63.48**	**0.26**	91.04	0.11	**74.80**	**0.06**	**60.43**	**0.30**	89.01	**0.14**	**71.98**	**0.09**

TABLE IV

Statistical Analysis of Representative Methods for Camera Pose Estimation Metrics on the yfcc100m Datasets. μ and σ^2 Indicate the Mean and Variance, Respectively. The Optimal Indicator Values Are in Bold

Dataset	YFCC100M (%)							
Method	Known Scene				Unknown Scene			
	5		20		5		20	
	μ	σ^2	μ	σ^2	μ	σ^2	μ	σ^2
T-Net	40.86	**0.07**	63.81	**0.03**	46.74	1.54	73.11	0.23
MS^{2}DG-Net	37.78	4.83	62.78	2.11	46.98	10.55	75.13	2.64
GCA-Net	43.82	1.32	67.21	0.91	55.07	2.22	**79.07**	0.43
MSGSA	**45.94**	0.40	**67.67**	0.24	**57.93**	**0.07**	78.78	**0.06**

3.66% at 5° and 20° thresholds, respectively, compared to SGA-Net.

While MSGSA may not surpass all current state-of-the-art methods in every metric for outlier removal on the SUN3D dataset, its standout performance in the camera pose estimation task is significant. This underscores a crucial insight: the accuracy of camera pose estimation relies not only on the precision and recall of outlier removal but also on the quality of the match set obtained post-outlier removal. Notably, our proposed method consistently generates a higher-quality match set, contributing to enhanced accuracy in camera pose estimation.

F. Statistical Difference Analysis

To comprehensively evaluate the statistical differences between the current state-of-the-art technologies, several representative existing techniques, including T-Net, MS^{2}DG-Net, GCA-Net and the proposed MSGSA, were selected for independent experiments on the YFCC100M dataset. These methods were retrained five times on the same training set, and then the means and variances of various evaluation metrics obtained by each method on the testing set were reported for statistical analysis. From the perspective of mean difference analysis, the experimental results presented in Tables III and IV show that, compared to all other three methods, the proposed MSGSA obtains the best means in terms of the P and F metrics for the outlier removal task in both known and unknown scenes. Furthermore, MSGSA also obtains the best means under the error thresholds of 5° and 20° for the camera pose estimation task in the known scene. In contrast, MSGSA is only slightly lower than GCA-Net in terms of the R metric for the outlier removal task across both scenes, and under the error threshold of 20° for the camera pose estimation task in the known scene. From the perspective of variance analysis, as shown in Tables III and IV, the variance of the R metric obtained by MSGSA is only slightly higher than those of MS^{2}DG-Net and GCA-Net for the outlier removal task in the known scene. Meanwhile, the variance under the error thresholds of 5° and 20° obtained by MSGSA is slightly higher than that of T-Net for the camera pose estimation task in the known scene. However, MSGSA still achieves the best variances in 7 out of a total of 10. Overall, MSGSA exhibits lower variance than some other methods, indicating its higher stability and reliability. Regarding the issue of stability, we attribute the primary reasons to the randomness of parameter initialization and the balance between outlier removal loss and camera pose estimation loss. In future work, we plan to conduct a more in-depth study on stability.

G. Downstream Applications

To demonstrate the versatility of our method, we conducted experiments across diverse downstream tasks, including remote sensing image registration, 3D point cloud registration, and 3D fusion. For the remote sensing image registration, we compared MSGSA against OANet, T-Net, MS^{2}DG-Net on a remote sensing dataset [55]. We showcased the qualitative results on several representative scenes in Fig. 6. From the experimental results, it can be observed that some methods do not perform well in scenes with significant viewpoint changes. For instance, OANet had poor registration results in scenes Figs. 6 (a) and (b), while T-Net showed unsatisfactory

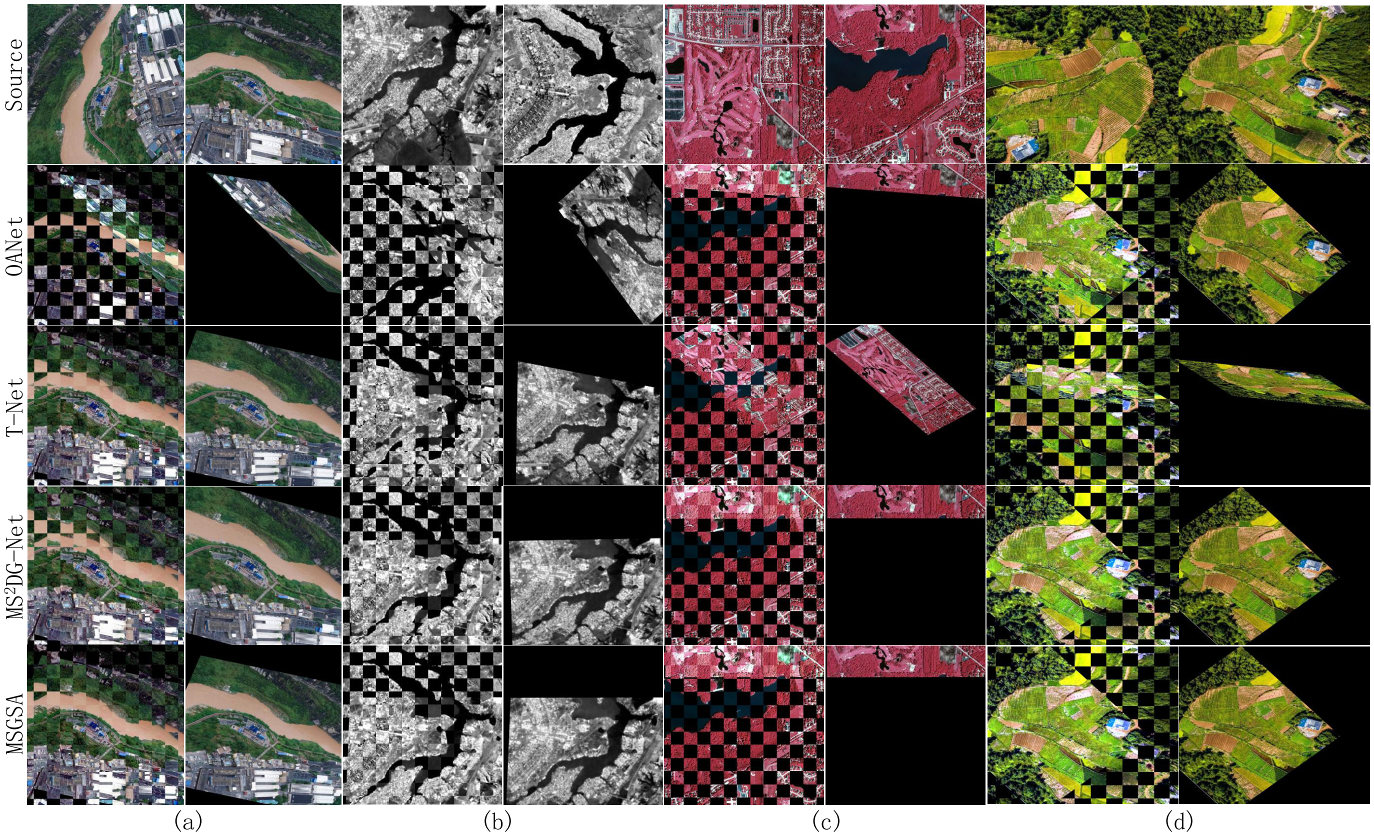


Fig. 6. These visualizations demonstrate remote sensing image registration. The initial row displays pairs of images for alignment, with the left as the source and the right as the target. Subsequent rows show results from the registration methods: OANet, T-Net, MS^{2}DG-Net, and our MSGSA network. Each column exhibits a checkerboard image on the left and the warped sensed image on the right.

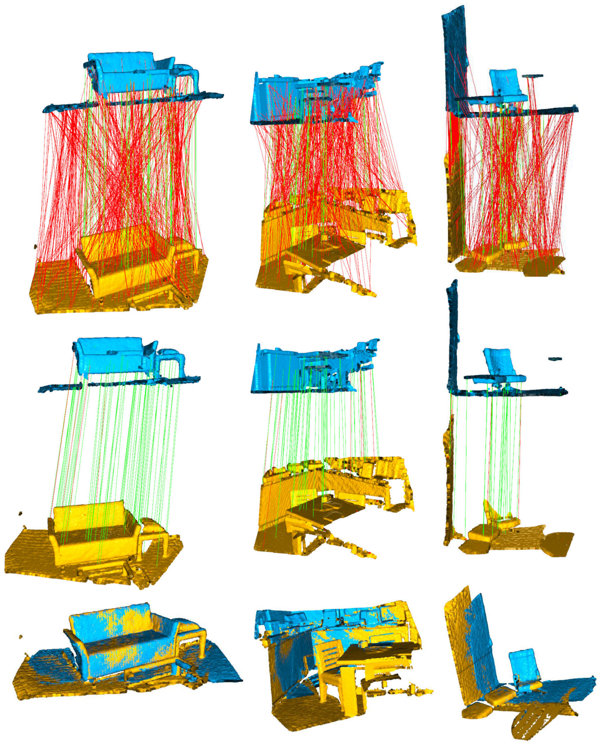
performance in scene Fig. 6 (d). On the other hand, both MS^{2}DG-Net and our MSGSA method demonstrated better registration performance in these scenes. Furthermore, upon closer inspection, MSGSA outperformed MS^{2}DG-Net in terms of capturing finer details, highlighting its superior generalization ability.

To assess the generalization ability of our method in 3D scenes, we conducted experiments on the 3DMatch dataset [56]. Using the PointDSC framework [57], we trained our model on the 3DMatch training set and validated it on the test set. The test set included 1623 pairs of partially overlapping point cloud fragments from eight scenes, each labeled with ground truth. As shown in Fig. 7, by employing our method, the majority of misaligned points (near 97%) were effectively eliminated. Visualization of selected scenes demonstrated the effectiveness of our method, successfully eliminating misaligned points and achieving highly satisfactory 3D fusion. This capability allows seamless integration of point cloud fragments, contributing to generating more complete and accurate 3D scenes.

H. Ablation Experiments

MSGSA consists of three main components: the MB backbone network, the GTC module, and the GSA module. The MB backbone network provides rich and robust spatial transformations, while the GTC module helps suppress the effects of outliers. Lastly, the GSA module improves the quality of the matched set after outlier removal. To assess the effectiveness of each component, we conducted ablation experiments on both the YFCC100M and SUN3D datasets. In Table V, it is evident that even when solely employing the MB module without the GTC and GSA modules, MSGSA outperforms the majority of existing methods. The introduction of the GTC module yields a significant performance improvement, further enhanced by the GSA module. However, the impact of the GSA module is relatively less pronounced on the YFCC100M dataset. To provide a clearer demonstration of the effectiveness of the GSA module, we conducted ablative experiments on the SUN3D dataset. As shown in Table VI, the GSA module achieved a performance improvement of 2.02% and 1.45% in the known and unknown scenarios, respectively, with an error threshold of 5°. These experiments provide compelling evidence that all three modules in MSGSA effectively enhance its performance.

1) The Impact of Local Spatial Transformation Branch in MBSE: To assess the effectiveness of the local spatial transformation branch in MBSE, we conducted experiments where this branch was removed. The results, as shown in Table VII and Table VIII, demonstrate that using the local spatial transformation branch improves the performance of both the outlier removal task and the camera pose estimation task, compared to not using it. The improvement is particularly prominent in the camera pose estimation task. Especially,



(a) (b) (c)

Fig. 7. Visualization of 3D point cloud registration and fusion. From top to bottom, the images depict the initial matching, outlier removal, and 3D fusion.

TABLE V

Results From Ablation Experiments on the Three Proposed Modules on the YFCC100M Dataset. **MB**: Multi-Branch Module. **GTC**: Geometric Transformation Consistency Module. **GSA**: Geometric Semantic Attention Module

MB	GTC	GSA	Known	Unknown
✓			43.40	53.40
✓	✓		46.22	57.70
✓	✓	✓	**46.47**	**58.35**

in the camera pose estimation task with an error threshold of 5°, the MSGSA with MBSE's local spatial transformation branch exhibited a performance improvement of 1.62% in known scenes and 2.60% in unknown scenes compared to the MSGSA without this branch. This improvement is attributed to the recognition that the geometric properties of the initial matching data manifest not only in the matching pair channels but also in the relationships between the matching pairs. Relying solely on the channel refinement branch is insufficient to capture the geometric relationships between matching pairs. With the inclusion of the local spatial transformation branch, the MBSE module effectively captures information in both the channel and spatial dimensions, enabling a more comprehensive transformation.

TABLE VI

The Geometric Semantic Attention Module Showed More Significant Improvement on the SUN3D Dataset

MB	GTC	GSA	Known	Unknown
✓	✓		24.45	19.56
✓	✓	✓	**26.47**	**21.01**

TABLE VII

Result of Outlier Removal With and Without Local Spatial Transformation Branch in MBSE. **MSGSA***: Without Local Spatial Transformation Branch

Dataset	YFCC100M (%)					
Method	Known Scene			Unknown Scene		
	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
MSGSA*	62.90	91.04	74.40	59.87	88.79	71.52
MSGSA	**63.28**	**91.29**	**74.75**	**60.03**	**89.34**	**71.81**

TABLE VIII

Result of Camera Pose Estimation With and Without Local Spatial Transformation Branch in MBSE. **MSGSA***: Without Local Spatial Transformation Branch

Dataset	YFCC100M(%)			
Method	Known Scene		Unknown Scene	
	5°	20°	5°	20°
MSGSA*	44.85	66.80	55.75	78.07
MSGSA	**46.47**	**67.99**	**58.35**	**79.13**

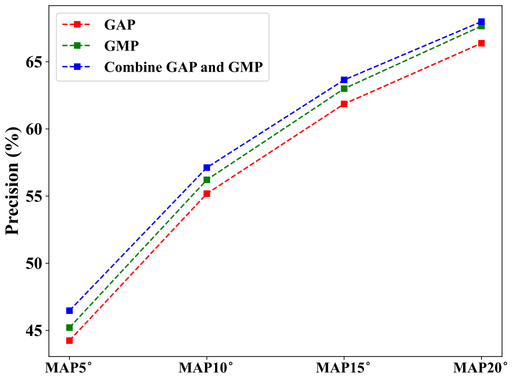


Fig. 8. The impact of global average pooling (GAP) and global max pooling (GMP) in camera pose estimation.

2) Global Max Pooling and Global Average Pooling: To evaluate the impact of global average pooling and global max pooling in the MBMS module, we conducted three ablation experiments. In the first experiment, we employed only global average pooling, while the second experiment utilized only global max pooling. The third experiment combined both global average pooling and global max pooling. Fig. 8 illustrates the camera pose estimation accuracy for the three experiments at the error thresholds of 5°, 10°, 15°, and 20°. The results demonstrate that the combination of global average pooling and global max pooling achieved the best performance, followed by using only global max pooling while using only global average pooling resulted in the lowest performance. The reason behind these observations is as follows: Global average pooling captures global information without losing valuable details, but it is less robust to outliers. In scenarios where outliers significantly outnumber inliers, using only global average pooling can weaken the influence of inliers, as their information is overshadowed by that of outliers. On the other hand, global max pooling emphasizes high response values, effectively extracting inlier information.

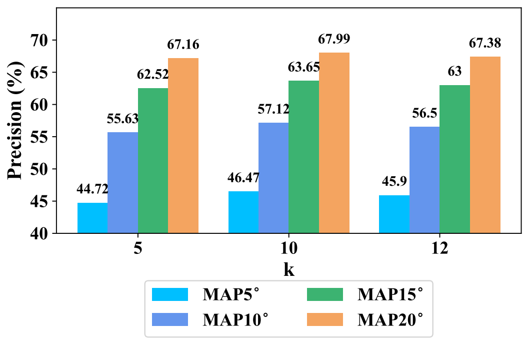


Fig. 9. Results of camera pose estimation under different k-values.

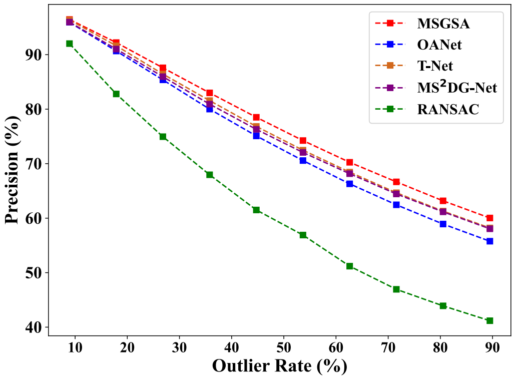


Fig. 10. The impact of various outlier rates on outlier removal.

However, it may suffer from information loss. Considering the strengths and weaknesses of global max pooling and global average pooling, it is evident that these two pooling techniques complement each other in practice. Therefore, in the third experiment, we combined both global average pooling and global max pooling to leverage the advantages of both methods, resulting in optimal performance.

3) The Value of k in GSA: The parameter "k" in the GSA module represents the number of category nodes, and it plays a crucial role in determining the semantic division. A larger value of "k" leads to a finer semantic division, allowing for more precise extraction of semantic neighbor information. However, excessively large values of "k" can result in missing semantic neighbors, which leads to inadequate extraction of neighbor information. On the other hand, a smaller "k" value can result in a broader semantic scope, which can affect the extraction of semantic neighbor information. To determine an appropriate "k" value, we conducted ablation experiments on the YFCC100M dataset, evaluating different values of "k". Fig. 9 illustrates the camera pose estimation accuracy at the error thresholds of 5°, 10°, 15°, and 20° for "k" values of 5, 10, and 12. The results indicate that neither a larger "k" value nor a smaller "k" value necessarily leads to better performance. Instead, a value of 10 emerges as a relatively balanced choice, optimizing the precision of semantic division and neighbor information extraction.

4) The Impact of Outlier Rates: To evaluate the effectiveness of various methods in eliminating mismatched points across different outlier rates, we conducted ablation experiments with selected methods, including OANet, T-Net, MS^{2}DG-Net, RANSAC, and our proposed method. OANet, T-Net, MS^{2}DG-Net, and our method are deep learning-based, while RANSAC is a traditional method. Fig. 10 presents the experimental results. Under low outlier rates, all methods achieved a mismatch removal accuracy of over 90%. The deep learning-based methods showed consistent performance, with marginal differences compared to RANSAC, the traditional method. However, as the outlier rate increased, our method's effectiveness in removing mismatches gradually surpassed the other methods, indicating superior robustness to outliers. A notable observation is the significant degradation in the performance of the traditional method as the outlier rate increased, emphasizing the superior robustness of deep learning-based methods in handling higher outlier rates.

TABLE IX

CAMERA POSE ESTIMATION RESULTS FOR NETWORKS WITH AND WITHOUT THE GTC MODULE ("^{a}" REPRESENTS USING THE GTC MODULE)

Dataset	YFCC100M(%)			
Method	Known Scene		Unknown Scene	
	5°	20°	5°	20°
OANet	33.50	57.53	41.33	68.79
OANet^a	**35.71**	**59.28**	**45.17**	**71.41**
T-Net	40.86	63.81	46.74	73.11
T-Net^a	**45.49**	**67.69**	**53.88**	**77.36**
MS^{2}DG-Net	37.78	62.78	46.98	75.13
MS^{2}DG-Net^a	**41.97**	**65.49**	**52.75**	**77.96**

5) Transferability of GTC: To evaluate the adaptability of the GTC module to other models, we integrated the GTC module into the OANet, T-Net, and MS^{2}DG-Net models individually and compared their performance with the original models. The results, presented in Table IX, demonstrate a significant enhancement in the camera pose estimation task upon incorporating the GTC module. The GTC module has demonstrated outstanding performance across various models, mainly due to its two core strengths: Firstly, the GTC module successfully extracts crucial geometric transformation consistency information by using the feature information of matching pairs during the transformation process. This not only reveals the common geometric constraints followed by correct matching pairs, but also provides a strong support for distinguishing correct and incorrect matches. Secondly, the GTC module significantly enhances the information transfer and interdependence between different stages of the network, thereby significantly improving the overall performance of the model. By incorporating the prior knowledge from the previous stage as input for subsequent stages, the accumulation and optimization of information are achieved, leading to a more efficient learning process within the multi-stage network architecture.

6) Transferability of GSA: To assess the applicability of the GSA module to other models, we integrated the GSA module into the OANet model and compared its performance to the original OANet model. The results, As shown in Tables X and XI, demonstrated a significant improvement in both the outlier removal and camera pose estimation tasks when the GSA module was incorporated. Especially, in the camera pose

TABLE X

OUTLIER REMOVAL RESULTS FOR NETWORKS WITH AND WITHOUT THE GSA MODULE ("*" REPRESENTS USING THE GSA MODULE)

Dataset	YFCC100M (%)					
Method	Known Scene			Unknown Scene		
	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
OANet	59.30	88.40	70.98	55.80	85.10	67.40
OANet*	**60.59**	**88.65**	**71.98**	**57.02**	**85.73**	**68.49**

TABLE XI

CAMERA POSE ESTIMATION RESULTS FOR NETWORKS WITH AND WITHOUT THE GSA MODULE ("*" REPRESENTS USING THE GSA MODULE)

Dataset	YFCC100M (%)			
Method	Known Scene		Unknown Scene	
	5°	20°	5°	20°
OANet	33.50	57.53	41.33	68.79
OANet*	**36.06**	**58.89**	**44.75**	**70.63**

estimation task with an error threshold of 5°, the OANet with the inserted GSA module exhibited a performance improvement of 2.56% in known scenes and 3.42% in unknown scenes compared to the original OANet. This improvement can be attributed to the GSA module's ability to leverage geometric semantic neighbor information, highlighting the importance of such information in achieving superior performance.

To provide a clearer visualization of the impact of the GSA module, we conducted experiments on image pairs with a substantial presence of similar semantics and compared the performance of the original OANet and OANet with the GSA module. As depicted in Fig. 11, we observed that the original OANet, without the GSA module, struggled to discriminate between inliers and outliers in such scenarios. In contrast, the OANet model enhanced with the GSA module exhibited significant performance improvement, successfully distinguishing between inliers and outliers. Moreover, the visualized performance of MSGSA in these scenarios was even more impressive, highlighting the synergistic effect of the various modules within MSGSA, which provide a robust foundation for capturing semantic neighbor information.

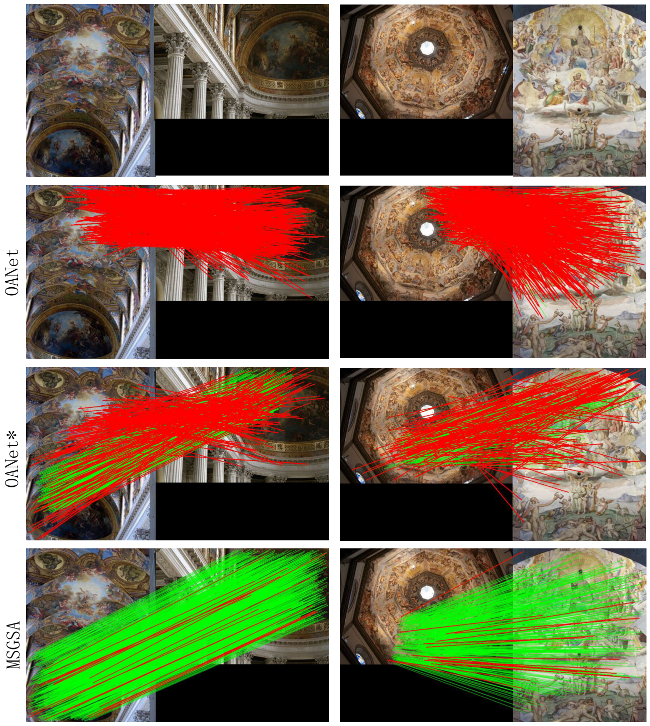


Fig. 11. In scenes with a large number of semantically similar features, OANet loses its ability to distinguish these similar matches. By incorporating our GSA module (OANet*), OANet achieves a significant improvement in its ability to discriminate among these semantically similar matches.

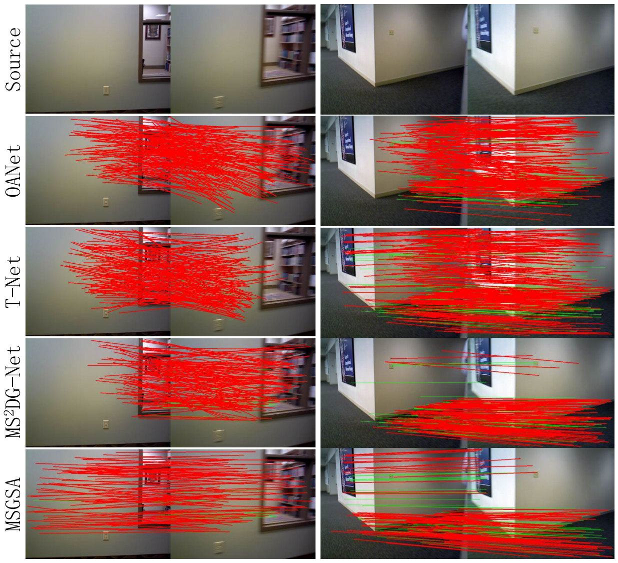


Fig. 12. Outlier removal effect in scenes with low light, large areas of solid color walls or floors, and presence of ghosting.

I. Limitations

Our method exhibits limitations, particularly in scenarios represented by the SUN3D dataset, where it struggles to effectively remove false matches. Analysis of the SUN3D dataset and visualization of results have revealed specific conditions leading to suboptimal performance. These conditions include low lighting, extensive featureless surfaces like plain walls or floors, and image artifacts induced by camera shake. With an anomaly rate exceeding 95% in such scenes, we visualized two representative instances in Fig. 12, featuring outlier rates of 98.3% and 95.85%, respectively. Notably, existing methods also face challenges in these scenarios.

To address these limitations, we identify the need for enhancements in the feature point extraction algorithm. Traditional algorithms heavily depend on local texture information, posing challenges in accurately extracting key points in scenes with limited texture, such as those with low lighting and large featureless surfaces. We propose that improving the feature point extraction algorithm is pivotal. This enhancement could involve incorporating additional visual information, such as depth and color, to bolster adaptability in specific challenging scenarios. Depth information can aid in feature point extraction by analyzing the depth map to identify geometric structures, while color information can augment extraction by leveraging color variations for diversity and robustness.

An alternative method to enhance the feature point extraction algorithm is the integration of machine learning techniques. Training models on abundant sample data can facilitate automatic learning and extraction of feature points tailored to specific challenging scenarios. Machine learning, leveraging relationships between samples, holds promise for extracting more discriminative feature points.

In summary, improving the feature point extraction algorithm is crucial for enhancing matching performance in scenes marked by low lighting and large featureless surfaces. This improvement contributes to overall accuracy and robustness in image processing and analysis. Consequently, this area stands out as a significant focus for our future research endeavors.

V. Conclusion

This paper proposes the MSGSA network for two-view correspondence learning, consisting of three key components: the MB module, the GTC module, and the GSA module. The MB module enhances spatial transformations, the GTC module focuses on GTC, and the GSA module leverages graph neural networks and transformers to mine semantic relationships, improving the matching pair set quality post-outlier removal. Experimental results on two public datasets underscore the superior performance of MSGSA over various state-of-the-art methods. In situations with a high number of outliers, outlier removal resembles a binary classification problem with significant class imbalance. To further refine our method, we plan to address the challenge of class imbalance in our future research efforts.

References

[1] Y. Jin et al., “Image matching across wide baselines: From paper to practice,” *Int. J. Comput. Vis.*, vol. 129, no. 2, pp. 517–547, Feb. 2021.

[2] 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.

[3] C. Wu, “Towards linear-time incremental structure from motion,” in *Proc. Int. Conf. 3D Vis.*, Jun. 2013, pp. 127–134.

[4] J. L. Schonberger and J.-M. Frahm, “Structure-from-motion revisited,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2016, pp. 4104–4113.

[5] 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.

[6] C. Cadena et al., “Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age,” *IEEE Trans. Robot.*, vol. 32, no. 6, pp. 1309–1332, Dec. 2016.

[7] G. Tolias, Y. Avrithis, and H. Jégou, “Image search with selective match kernels: Aggregation across single and multiple images,” *Int. J. Comput. Vis.*, vol. 116, pp. 247–261, Mar. 2016.

[8] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, “Large-scale image retrieval with attentive deep local features,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3456–3465.

[9] S. Lin, H. Luo, Y. Yan, G. Xiao, and H. Wang, “Co-clustering on bipartite graphs for robust model fitting,” *IEEE Trans. Image Process.*, vol. 31, pp. 6605–6620, 2022.

[10] M. A. Fischler and R. 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.

[11] J. Ma, J. Zhao, J. Tian, A. L. Yuille, and Z. Tu, “Robust point matching via vector field consensus,” *IEEE Trans. Image Process.*, vol. 23, pp. 1706–1721, 2014.

[12] O. Chum and J. Matas, “Matching with PROSAC-progressive sample consensus,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 220–226.

[13] D. R. Myatt, P. H. S. Torr, S. J. Nasuto, J. M. Bishop, and R. Craddock, “NAPSAC: High noise, high dimensional robust estimation—It’s in the bag,” in *Proc. Procedings Brit. Mach. Vis. Conf.*, 2002, p. 3.

[14] S. Lin, X. Wang, G. Xiao, Y. Yan, and H. Wang, “Hierarchical representation via message propagation for robust model fitting,” *IEEE Trans. Ind. Electron.*, vol. 68, no. 9, pp. 8582–8592, Sep. 2021.

[15] S. Lin, G. Xiao, Y. Yan, D. Suter, and H. Wang, “Hypergraph optimization for multi-structural geometric model fitting,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8730–8737.

[16] T. Lai, A. Sadri, S. Lin, Z. Li, R. Chen, and H. Wang, “Efficient sampling using feature matching and variable minimal structure size,” *Pattern Recognit.*, vol. 137, May 2023, Art. no. 109311.

[17] 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.

[18] 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.

[19] 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.

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

[21] Z. Zhong, G. Xiao, S. Wang, L. Wei, and X. Zhang, “PESA-Net: Permutation-equivariant split attention network for correspondence learning,” *Inf. Fusion*, vol. 77, pp. 81–89, Jan. 2022.

[22] J. Chen, S. Chen, X. Chen, Y. Dai, and Y. Yang, “CSR-Net: Learning adaptive context structure representation for robust feature correspondence,” *IEEE Trans. Image Process.*, vol. 31, pp. 3197–3210, 2022.

[23] J. Chen et al., “Shape-former: Bridging CNN and transformer via ShapeConv for multimodal image matching,” *Inf. Fusion*, vol. 91, pp. 445–457, Mar. 2023.

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

[25] L. Dai et al., “MS2DG-Net: Progressive correspondence learning via multiple sparse semantics dynamic graph,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8973–8982.

[26] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2018, pp. 7132–7141.

[27] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–14.

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

[29] A. Dosovitskiy et al., “An image is worth 16 × 16 words: Transformers for image recognition at scale,” in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–22.

[30] P. H. S. Torr and A. Zisserman, “MLESAC: A new robust estimator with application to estimating image geometry,” *Comput. Vis. Image Understand.*, vol. 78, no. 1, pp. 138–156, Apr. 2000.

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

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

[33] G. Wang, Z. Wang, Y. Chen, X. Liu, Y. Ren, and L. Peng, “Learning coherent vector fields for robust point matching under manifold regularization,” *Neurocomputing*, vol. 216, pp. 393–401, Dec. 2016.

[34] J. Ma, W. Qiu, J. Zhao, Y. Ma, A. L. Yuille, and Z. Tu, “Robust L_2E estimation of transformation for non-rigid registration,” *IEEE Trans. Signal Process.*, vol. 63, no. 5, pp. 1115–1129, Jan. 2015.

[35] X. Li and Z. Hu, “Rejecting mismatches by correspondence function,” *Int. J. Comput. Vis.*, vol. 89, no. 1, pp. 1–17, Aug. 2010.

[36] E. Brachmann et al., “DSAC—Differentiable RANSAC for camera localization,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2492–2500.

[37] X. Liu, G. Xiao, R. Chen, and J. Ma, “PGFNet: Preference-guided filtering network for two-view correspondence learning,” *IEEE Trans. Image Process.*, vol. 32, pp. 1367–1378, 2023.

[38] J. Guo, G. Xiao, Z. Tang, S. Chen, S. Wang, and J. Ma, "Learning for feature matching via graph context attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5102714.

[39] T. Liao, X. Zhang, Y. Xu, Z. Shi, and G. Xiao, "SGA-Net: A sparse graph attention network for two-view correspondence learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 12, pp. 7578–7590, 2023.

[40] S. Lin, A. Yang, T. Lai, J. Weng, and H. Wang, "Multi-motion segmentation via co-attention-induced heterogeneous model fitting," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 3, pp. 1786–1798, Mar. 2024.

[41] 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.

[42] A. Howard et al., "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324.

[43] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11534–11542.

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

[45] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Proc. 9th Eur. Conf. Comput. Vis. (ECCV)*, in Lecture Notes in Computer Science, 2006, pp. 404–417.

[46] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2564–2571.

[47] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2006, pp. 430–443.

[48] 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.

[49] H. C. Longuet-Higgins, "A computer algorithm for reconstructing a scene from two projections," *Nature*, vol. 293, no. 5828, pp. 133–135, Sep. 1981.

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

[51] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

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

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

[54] 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. Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 5099–5108.

[55] J. Ma, X. Jiang, J. Jiang, J. Zhao, and X. Guo, "LMR: Learning a two-class classifier for mismatch removal," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 4045–4059, Aug. 2019.

[56] A. Zeng, S. Song, M. Nießner, M. Fisher, J. Xiao, and T. Funkhouser, "3DMatch: Learning local geometric descriptors from RGB-D reconstructions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1802–1811.

[57] X. Bai et al., "PointDSC: Robust point cloud registration using deep spatial consistency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15859–15869.

Shuyuan Lin (Member, IEEE) received the Ph.D. degree in computer science and technology from Xiamen University, Fujian, China, in 2020. He is currently an Assistant Professor with the College of Cyber Security/College of Information Science and Technology, Jinan University, China. He has published more than 20 papers in international journals and conferences, including IJCV, IEEE Transactions on Image Processing/IEEE Transactions on Circuits and Systems for Video Technology/IEEE Transactions on Industrial Electronics/IEEE Signal Processing Letters, IEEE ICASSP/ICIP, AAAI, PR, EAAI, and CJC. His research interests include machine learning, computer vision, and pattern recognition. He was awarded the Best Ph.D. Thesis from Fujian Province.

Xiao Chen received the bachelor's degree from Fuzhou University in 2018. He is currently pursuing the M.S. degree with Jinan University. His research interests include computer vision, machine learning, and pattern recognition.

Guobao Xiao (Senior Member, IEEE) received the Ph.D. degree from Xiamen University, China. He is currently a Tenured Professor with Tongji University, China. He has published more than 50 papers in journals and conferences, including IEEE Transactions on Pattern Analysis and Machine Intelligence/IEEE Transactions on Image Processing, IJCV, ICCV, and ECCV. His research interests include machine learning, computer vision, and pattern recognition. He has been awarded the Best Ph.D. Thesis Award from China Society of Image and Graphics (a total of ten winners in China). He served on the Program Committee (PC) for CVPR, ICCV, and ECCV.

Hanzi Wang (Senior Member, IEEE) received the Ph.D. degree in computer vision from Monash University, Australia, in 2004. He is currently a Distinguished Professor with Minjiang Scholars, Fujian, and the Founding Director of the Centre for Pattern Analysis and Machine Intelligence (CPAMI), Xiamen University, China. He has published more than 150 academic papers in top-level international journals and conferences, including IEEE Transactions on Pattern Analysis and Machine Intelligence, *IJCV*, IEEE Transactions on Image Processing, IEEE Transactions on Multimedia, IEEE Transactions on Cybernetics, IEEE Transactions on Circuits and Systems for Video Technology, ICCV, ECCV, CVPR, and AAAI. His research interests include concentrating on computer vision and related fields. He was the General Chair of the IEEE Conference of Future Media Technology 2018 and ACM ICIMCS 2014. He was an Associate Editor of IEEE Transactions on Circuits and Systems for Video Technology from 2010 to 2015.

Feiran Huang (Member, IEEE) received the B.Sc. degree from Central South University, Changsha, China, in 2011, and the Ph.D. degree in computer software and theory from the School of Computer Science and Engineering, Beihang University, Beijing, China, in 2019. He is currently a Professor with the School of Information Science and Technology and Jinan University, Guangzhou, China. He has published more than 50 articles, such as IEEE Transactions on Image Processing (TIP), IEEE Transactions on Cybernetics (TCYB), KBS, ACM MM, CIKM, and ICMR. His research interests include social media analysis and multimodal learning.

Jian Weng (Senior Member, IEEE) received the Ph.D. degree in computer science and engineering from Shanghai Jiao Tong University in 2008. From 2008 to 2010, he held a postdoctoral position with the School of Information Systems, Singapore Management University. He is currently a Professor and the Vice President of Jinan University. He has published more than 100 papers in cryptography and security conferences and journals, such as *Cryptography*, EUROCRYPT, ASIACRYPT, ACM CCS, and USENIX Security. His research interests include cryptography, data security, and blockchain. He won the Innovation Award from Chinese Association for Cryptologic Research in 2015 and the National Science Fund for Distinguished Young Scholars in 2018. He served as the PC co-chair or a PC member for more than 50 international conferences.