MGCA-Net: Multi-Graph Contextual Attention Network for Two-View Correspondence Learning

Shuyuan Lin[1], **Mengtin Lo**[1], **Haosheng Chen**[2*], **Yanjie Liang**[3*], **Qiangqiang Wu**[4]
[1]College of Cyber Security, Jinan University, Guangzhou, China
[2]Chongqing Key Laboratory of Image Cognition, College of Computer Science and Technology,
Chongqing University of Posts and Telecommunications, Chongqing, China
[3]Peng Cheng Laboratory, Shenzhen, China
[4]Department of Computer Science, City University of Hong Kong, Hong Kong, China
swin.shuyuan.lin@gmail.com, ekyp1025@gmail.com, chenhs@cqupt.edu.cn, liangyj@pcl.ac.cn,
qiangqwu2-c@my.cityu.edu.hk

Abstract

Two-view correspondence learning is a key task in computer vision, which aims to establish reliable matching relationships for applications such as camera pose estimation and 3D reconstruction. However, existing methods have limitations in local geometric modeling and cross-stage information optimization, which make it difficult to accurately capture the geometric constraints of matched pairs and thus reduce the robustness of the model. To address these challenges, we propose a Multi-Graph Contextual Attention Network (MGCA-Net), which consists of a Contextual Geometric Attention (CGA) module and a Cross-Stage Multi-Graph Consensus (CSMGC) module. Specifically, CGA dynamically integrates spatial position and feature information via an adaptive attention mechanism and enhances the capability to capture both local and global geometric relationships. Meanwhile, CSMGC establishes geometric consensus via a cross-stage sparse graph network, ensuring the consistency of geometric information across different stages. Experimental results on two representative YFCC100M and SUN3D datasets show that MGCA-Net significantly outperforms existing SOTA methods in the outlier rejection and camera pose estimation tasks. Source code is available at http://www.linshuyuan.com.

1 Introduction

Two-view correspondence is critical for applications like visual localization [Chen *et al.*, 2024], SfM [Schonberger and Frahm, 2016], SLAM [Placed *et al.*, 2023], and 3D reconstruction [Schmied *et al.*, 2023]. By matching features between two images, a reliable geometric relationship is established, which lays the foundation for estimating camera pose and enhancing robust localization in complex scenes. However, factors such as occlusion, illumination variations, and descriptor inaccuracies often introduce a significant number of outliers (i.e., incorrect matches) [Lin *et al.*, 2024b]. These outliers not only reduce matching precision but also propagate errors into downstream tasks, such as camera pose estimation and 3D reconstruction. Consequently, outlier rejection is a critical step to improve the accuracy and robustness of two-view correspondence [Brachmann and Rother, 2019].

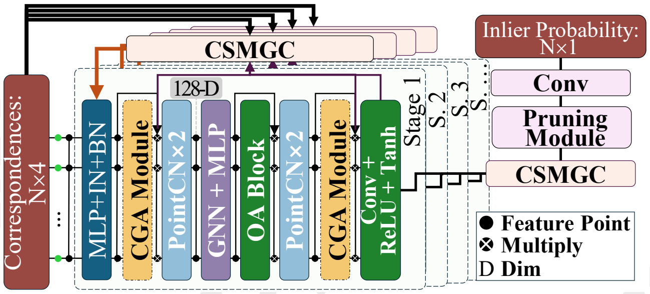


Figure 1: Overall architecture of the proposed MGCA-Net.

Traditional outlier rejection methods, such as RANSAC [Fischler and Bolles, 1981] and their variants [Barath and Matas, 2018], show high robustness and accuracy when dealing with low outlier ratios. These methods typically rely on randomly sampling minimal subsets to fit geometric models and validate their validity based on the number of inliers. However, their performance significantly degrades with increasing outlier ratios, especially in complex geometric scenes or under challenging conditions such as severe illumination and viewpoint changes [Ma *et al.*, 2021; Lin *et al.*, 2022]. While traditional methods remain competitive in certain tasks, their efficiency and robustness are limited in high-outlier scenes. Additionally, these methods often depend on user-defined thresholds or specific prior assumptions, restricting their application in real-world environments.

In contrast, deep learning-based outlier rejection methods have demonstrated considerable performance improvements. PointCN [Yi *et al.*, 2018], as a pioneering work, modeled the outlier rejection task as a binary classification problem. It utilized multi-layer perceptrons (MLPs) to process unordered keypoints and effectively capture global information through

[*]Corresponding Author

contextual normalization. However, MLPs are inherently limited in capturing local geometric information.

In addition, most deep learning-based methods [Zhao *et al.*, 2021] rely on simple convolutional operations or fixed neighborhood clustering, which fail to fully capture the complex geometric relationships between keypoints. In challenging scenes such as severe illumination changes or significant viewpoint variations, these models struggle to align global semantics with local geometric information, making it difficult to adequately capture and represent the intricate relationships between them.

To overcome these limitations, solutions based on convolutional neural networks (CNNs) and Transformers have been proposed. For example, ConvMatch [Zhang and Ma, 2023a] combines dense motion regularization with local convolutional operations to extract contextual information effectively. VSFormer [Liao *et al.*, 2024] fuses visual and geometric features across modalities to enhance representation, while PT-Net [Gong *et al.*, 2024] employs a pyramid Transformer architecture with sparse attention mechanisms to integrate multi-scale motion field information. Although these methods have improved outlier rejection, most focus solely on processing features from the previous stage while neglecting feature consistency and cross-stage interactions.

To address these challenges, we propose a Multi-Graph Contextual Attention Network (MGCA-Net) , which effectively integrates spatial features and geometric consensus to enhance outlier rejection. As illustrated in Fig. 1, MGCA-Net consists of two core modules: Contextual Geometric Attention (CGA) module and Cross-Stage Multi-Graph Consensus (CSMGC) module, designed to address the limitations of existing methods in local geometric modeling and cross-stage information optimization. Specifically, CGA consists of a Context Position Attention (CPA) and a Multi-Branch Feed Forward Network (MB-FFN). CPA dynamically fuses spatial and contextual information to effectively balance global semantics and local geometric details, while MB-FFN integrates multi-scale features to enhance feature representation in complex scenes. Furthermore, CSMGC uses a cross-stage sparse graph neural network to establish geometric consensus across different stages, enhancing feature interaction and ensuring geometric consistency throughout the process. The main contributions of this paper are summarized as follows:

- We propose a novel MGCA-Net that integrates global and local information with multi-stage feature fusion and geometric modeling, enhancing robustness and representation in high-outlier scenes.
- We propose CGA, which comprises CPA and MB-FFN. CPA effectively balances global semantics and local geometry by combining spatial and contextual information, while MB-FFN enhances feature representation in complex scenes by integrating multi-scale features.
- We propose CSMGC, which establishes geometric consistency across stages by incorporating geometric priors with multi-sparse graph neural networks, significantly improving robustness in outlier rejection.

By fusing cross-stage information, MGCA-Net can accurately recognize outliers and effectively reject them even in highly outlier scenes, while dynamically refining correspondence reliability through progressive consensus learning.

2 Related Work

Outlier rejection in two-view correspondence tasks is a key research topic in computer vision. Various methods have been proposed to improve correspondence accuracy, which can be broadly categorized into three main types: traditional methods, learning-based methods, and attention-based methods.

2.1 Traditional Methods

Traditional methods play a crucial role in two-view correspondence learning, particularly in the outlier rejection and geometric consistency modeling. These methods often rely on hypothesis-validation strategies, with RANSAC (Random Sample Consensus) [Fischler and Bolles, 1981] being the most classic and widely applied approach. RANSAC iteratively samples minimal subsets of data, fits geometric models, and evaluates the number of inliers to identify the best model. However, RANSAC is inefficient as it typically requires a large number of iterations to produce meaningful results. To address this, USAC (Universal Sample Consensus) [Raguram *et al.*, 2012] integrates several RANSAC enhancements, including dynamic sampling, hypothesis test optimization, and model validation strategies, providing a unified sampling framework. In addition to hypothesis-validation methods, non-parametric models, such as VFC [Ma *et al.*, 2014], distinguish inliers from outliers by establishing sparse vector field models and improve adaptability to complex scenes through regularization constraints. LPM (Locality Preserving Matching) [Ma *et al.*, 2019] introduces local consistency constraints to efficiently eliminate erroneous matches with linear-logarithmic complexity. Despite significant advancements in outlier rejection, traditional methods still face limitations in handling high outlier ratios and complex geometric scenes.

2.2 Learning-Based Methods

To effectively address the robustness challenges posed by high outlier ratios and complex geometric scenes, deep learning approaches have become a primary solution for outlier rejection tasks. LFGC-Net [Yi *et al.*, 2018] is a pioneering work that introduces a simple yet effective PointCN-like structure, reformulating the feature point matching task as an inlier classification problem. However, the single-stage network architecture adopted by LFGC-Net fails to effectively leverage local contextual information, resulting in suboptimal performance in high-outlier scenes.

To overcome these limitations, some methods have introduced iterative networks and pruning strategies to alleviate class imbalance while capturing geometric information. For instance, MSA-Net [Zheng *et al.*, 2022] integrates multi-scale attention mechanisms into a multi-stage network to improve robustness, and MS2DG-Net [Dai *et al.*, 2022] uses a sparse semantic dynamic graph to dynamically update matching features, enhancing the semantic consistency. Additionally, GCA-Net [Guo *et al.*, 2023] and SGA-Net [Liao *et al.*, 2023] use graph attention mechanisms to effectively combine local and global information, improving accuracy and robustness.

Following this, NCMNet [Liu *et al.*, 2024] further incorporates a neighborhood consistency mining module, capturing geometric relationships between local and global neighborhoods via a sparse graph structure, optimizing matching performance in noisy environments.

However, they still rely on local consistency scores and lack a deep understanding of global geometric information. Furthermore, during outlier filtering, pruning strategies may incorrectly eliminate inliers, especially in high-outlier scenes, which significantly reduces the number of remaining inliers. To address these challenges, we propose CSMGC, which dynamically integrates geometric information across stages and enhances feature consistency modeling.

2.3 Attention-Based Methods

Attention mechanisms have become a crucial component in deep learning models, particularly in computer vision tasks, where they play an essential role in enhancing models' performance. In the task of two-view correspondence learning, attention mechanisms are commonly used for feature extraction. For instance, SuperGlue [Sarlin *et al.*, 2020] combines self-attention and cross-attention, where self-attention focuses on the representation of features in a single image, while cross-attention compares feature similarities between different images to improve matching accuracy. LoFTR [Sun *et al.*, 2021] employs a detector-free strategy, using attention mechanisms to capture complex geometric information and generate high-quality matching results. T-Net [Zhong *et al.*, 2021] employs a Permutation-Equivariant Context Squeeze-and-Excitation module, dynamically adjusting feature weights using channel attention to strengthen global context modeling. However, existing attention-based methods face limitations in balancing local and global features, fusing multi-scale information, and handling complex geometric structures, which limits their performance in challenging scenes. To address these issues, we propose CGA, which combines position encoding with attention mechanisms to preserve spatial relationships between features, thereby enhancing the capture of local geometric features.

3 Methodology

3.1 Problem Formulation

Given a pair of images I_1 and I_2 from the same scene, feature points and descriptors are initially extracted from both images using existing methods (e.g., SIFT [Lowe, 2004], SuperPoint [DeTone *et al.*, 2018]). Then, an initial set of correspondences $S = s_1, s_2, ..., s_N \in \mathbb{R}^{N\times 4}$ is generated using a nearest-neighbor matching strategy, where N is the number of initial correspondences. Each correspondence $s_i = (x_i, y_i, x'_i, y'_i)$ consists of the coordinates of a keypoint in I_1 and its corresponding point in I_2. Finally, the network outputs logit values and the initial correspondence set S, which are input into the eight-point algorithm g, to estimate the fundamental matrix $\hat{E}$.

3.2 Contextual Geometric Attention Module

As traditional feature extraction methods have difficulty in representing both global semantics and local geometric features, we propose the CGA Module, which enhances feature representation by leveraging spatial and contextual information. As illustrated in Fig. 2, CGA consists of two key components: CPA and MB-FFN.

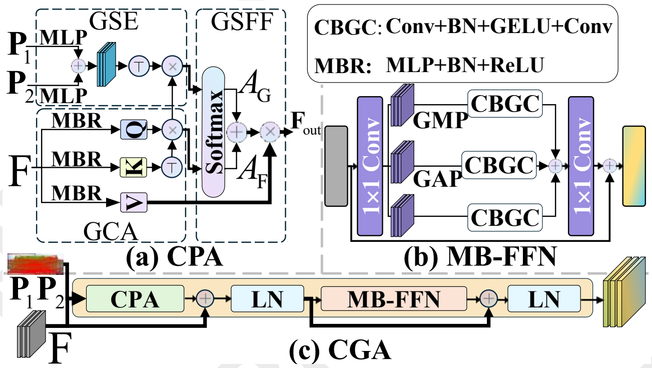


Figure 2: Pipeline of CGA.

3.2.1 Context Position Attention

CPA aims to improve the accuracy and robustness of feature point matching in two-view tasks by fusing global context and local geometric relationships. To achieve this, CPA incorporates a dual-attention mechanism (i.e., content attention and positional attention) and achieves dynamic integration of global semantics and local geometric information through the collaborative functioning of three key components: Global Context Awareness (GCA), Geometric Semantic Extraction (GSE), and Geometric Semantic Feature Fusion (GSFF).

1) Global Context Awareness: To model the geometric relationships among global features, CPA processes the input features $\mathbf{F} \in \mathbb{R}^{N\times d}$ through a Multi-Layer Perceptron (MLP), Batch Normalization (BN), and the non-linear activation function ReLU, mapping the features into three separate spaces, query ($\mathbf{Q}$), key ($\mathbf{K}$) and value ($\mathbf{V}$), as follows:

$$\begin{aligned} \mathbf{Q} &= \mathrm{ReLU}(\mathrm{BN}(\mathrm{MLP}(\mathbf{F}))), \\ \mathbf{K} &= \mathrm{ReLU}(\mathrm{BN}(\mathrm{MLP}(\mathbf{F}))), \\ \mathbf{V} &= \mathrm{ReLU}(\mathrm{BN}(\mathrm{MLP}(\mathbf{F}))). \end{aligned} \tag{1}$$

The correlation matrix is then calculated as:

$$A_{\mathrm{F}} = \mathrm{Softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{d}}\right), \tag{2}$$

where $\sqrt{d}$ is a scaling factor that balances the numerical range of the attention scores ensuring training stability. The Softmax function normalizes the correlation matrix to produce attention weights, effectively capturing long-range dependencies among feature points.

2) Geometric Semantic Extraction: To address the limitations of existing methods in modeling local geometric relationships, CPA explicitly captures pairwise geometric relationships through a positional attention mechanism. The input feature point coordinates $\mathbf{P}_1, \mathbf{P}_2 \in \mathbb{R}^{N\times 2}$ are mapped into a high-dimensional space and encoded into geometric features $\mathbf{P}$ via MLP. Subsequently, the geometric features $\mathbf{P}$ are combined with the query features $\mathbf{Q}$ to achieve deep fusion of geometric and semantic information, enabling effective modeling of geometric-semantic relationships between feature pairs. Specifically, the calculation for geometry-enhanced attention is as follows:

$$\mathbf{P} = \mathrm{MLP}(\mathbf{P}_1) + \mathrm{MLP}(\mathbf{P}_2), \tag{3}$$

$$\mathbf{A}_{\mathrm{G}} = Softmax\left(\mathbf{Q}\mathbf{P}^T\right). \tag{4}$$

This component ensures an explicit and robust encoding of pairwise geometric relationships while integrating them with semantic information.

3) Geometric Semantic Feature Fusion: To address the separation of geometric semantics and global features in different feature spaces, CPA introduces a fusion mechanism. Specifically, CPA achieves deep integration of geometric features and semantic information by combining the geometry-enhanced attention weights $\mathbf{A}_{\mathrm{G}}$ and the global context attention weights $\mathbf{A}_{\mathrm{F}}$. The fused output are calculated as follows:

$$\mathbf{F}_{\mathrm{out}} = (\mathbf{A}_{\mathrm{G}} + \mathbf{A}_{\mathrm{F}})\mathbf{V}, \tag{5}$$

where $\mathbf{A}_{\mathrm{G}}$ represents the geometry-enhanced attention, $\mathbf{A}_{\mathrm{F}}$ denotes the global context attention, and $\mathbf{V}$ is the value feature. The fused output $\mathbf{F}_{\mathrm{out}}$ integrates geometric and semantic information, encapsulating both the geometric constraints between feature pairs and the dependencies on global contextual information. Based on the above collaboration, CPA achieves deep fusion of geometric and semantic information within the feature space, significantly enhancing the representational capacity and flexibility of MGCA-Net across features of varying scales.

3.2.2 Multi-Branch Feed Forward Network

Traditional Feedforward Neural Networks (FFNs) can perform nonlinear transformations on input features but they struggle to effectively capture both global and local information under complex geometric contexts. To effectively integrate multi-scale feature information and enhance the generalization capability of CPA, we propose MB-FFN. MB-FFN introduces a multi-branch structure that processes features in parallel through different branches, fusing them in the final stage, fully extracting the semantic information of multi-scale features. As illustrated in Fig. 2 (b), MB-FFN comprises three components: 1) Local Convolutional Branch; 2) Global Average Pooling Branch; and 3) Max Pooling Branch. The outputs from these branches are transformed through the Convolution-BatchNorm-GELU-Convolution (CBGC) module and fused using a summation operation. The final output $\mathcal{H}$ is calculated as:

$$\begin{aligned}\mathcal{H} = {} & CBGC(GAP(\mathrm{LN}(\mathbf{F}_{\mathrm{out}}))) \\ & + CBGC(GMP(\mathrm{LN}(\mathbf{F}_{\mathrm{out}}))) \\ & + CBGC(\mathrm{LN}(\mathbf{F}_{\mathrm{out}})),\end{aligned} \tag{6}$$

where LN represents Layer Normalization, $CBGC = Conv(GELU(BN(Conv(x))))$, and GAP represents Global Average Pooling, and GMP represents Global Max Pooling. These features are progressively transformed through the CBGC module and fused into the output $\mathcal{H}$ via addition. This design effectively captures global and local features from different perspectives and enhances the model's ability to represent multi-scale features in complex scenes.

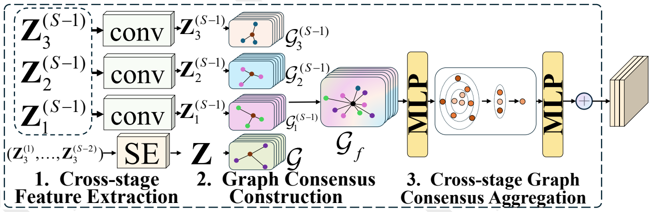


Figure 3: Overall architecture of the proposed CSMGC.

3.3 Cross-Stage Multi-Graph Consensus Module

In order to enhance the local geometric constraints and global information fusion between stages, we propose the CSMGC module, which consists of three components: 1) cross-stage feature extraction, 2) graph consensus construction, and 3) cross-stage graph consensus aggregation.

1) Cross-stage Feature Extraction: To effectively capture geometric features and contextual semantic information from the previous and cross stages, we extract three key features $\mathbf{Z}_1^{(M-1)}, \mathbf{Z}_2^{(M-1)}, \mathbf{Z}_3^{(M-1)}$ from different modules to represent the feature distribution of the first stage network. Among them, $\mathbf{Z}_1^{(M-1)}$ refers to features extracted by the first CPA module of the previous stage, representing the global semantic relationship; $\mathbf{Z}_2^{(M-1)}$ refers to the processing obtained by the second CPA module after PointCN, OANet and PointCN, representing richer local and global information; $\mathbf{Z}_3^{(M-1)}$ refers to the final combination of MLP and residual information, representing the comprehensive features of the current stage. In addition, to capture the geometric consistency across stages, we perform feature extraction on the network outputs across stages $\mathbf{Z}_3^{(1)}, \mathbf{Z}_3^{(2)}, \ldots, \mathbf{Z}_3^{(M-2)}$, which can be expressed as:

$$\mathbf{Z} = SE\left(concat\left(\mathbf{Z}_3^{(1)}, \mathbf{Z}_3^{(2)}, \ldots, \mathbf{Z}_3^{(M-2)}\right)\right), \tag{7}$$

where $\mathbf{Z}_3^{(1)}, \mathbf{Z}_3^{(2)}, \ldots, \mathbf{Z}_3^{(M-2)}$ represent the feature representations extracted from stages 1 to $M-2$ across stages, respectively. These features are concatenated through the $concat(\cdot)$ operation to form a global representation containing multi-stage features. Subsequently, the fusion module $SE(\cdot)$ adjusts the weights and compresses the concatenated features to dynamically model the relative importance of the features, thereby generating the final fused feature $\mathbf{Z}$.

2) Graph Consensus Construction: To capture geometric relationships and enhance the consistency of features across stages, for each stage of feature representation $\mathbf{Z}, \mathbf{Z}_1^{(M-1)}, \mathbf{Z}_2^{(M-1)}, \mathbf{Z}_3^{(M-1)}$, we separately construct a k-nearest neighbor graph $\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i)$ for each feature point , where the set of nodes $\mathcal{V}_i = \{c_1^i, ..., c_k^i\}$ denotes the feature points and their neighborhoods, the set of edges $\mathcal{E}_i = \{e_1^i, ..., e_k^i\}$ represents the geometric relationships between feature points. With this sparse graph structure, the geometric correlations between feature points can be dynamically captured, thus providing higher robustness for feature matching in complex scenes.

Furthermore, to improve the consistency of features across stages, we perform high-dimensional feature alignment on

Dataset	YFCC100M (%)						SUN3D (%)					
	Known Scene			Unknown Scene			Known Scene			Unknown Scene		
Method	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
RANSAC [Fischler and Bolles, 1981]	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++ [Qi *et al.*, 2017]	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 [Yi *et al.*, 2018]	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 [Ranftl and Koltun, 2018]	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 [Sun *et al.*, 2020]	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 [Zhang *et al.*, 2019]	61.14	88.16	69.73	57.90	85.07	66.53	54.43	88.08	63.72	46.50	83.83	56.32
T-Net [Zhong *et al.*, 2021]	61.18	89.94	70.47	57.18	87.01	66.73	55.01	88.36	64.18	46.50	83.98	56.33
PESA-Net [Zhong *et al.*, 2022]	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 [Zheng *et al.*, 2022]	59.27	90.28	68.92	56.49	88.60	66.46	56.09	87.57	64.71	48.64	83.81	57.89
MS²DG-Net [Dai *et al.*, 2022]	64.24	89.31	72.49	60.38	86.71	68.96	55.58	89.01	64.63	47.42	84.50	57.12
U-Match [Li *et al.*, 2023]	63.29	92.12	72.56	61.02	90.64	70.59	55.29	**89.35**	64.53	47.69	**85.6**	57.53
PGFNet [Liu *et al.*, 2023]	59.21	90.22	68.77	56.38	88.15	66.20	55.80	87.80	64.68	48.37	84.03	57.89
GCA-Net [Guo *et al.*, 2023]	63.82	91.36	72.87	60.44	88.92	69.74	55.00	89.08	64.21	46.88	85.14	56.79
ConvMatch [Zhang and Ma, 2023a]	63.14	91.2	72.21	60.22	89.48	69.65	55.79	89.23	64.89	48.13	85.55	57.87
ConvMatch $^{+}$[Zhang and Ma, 2023b]	62.75	92.05	72.11	59.41	90.12	69.13	55.51	89.30	64.80	47.60	85.51	57.53
NCM-Net [Liu *et al.*, 2024]	77.92	81.41	79.25	76.83	78.61	77.45	66.15	74.59	69.38	60.92	68.94	64.04
PT-Net [Gong *et al.*, 2024]	65.69	90.61	74.14	62.14	89.22	71.16	55.41	89.17	64.66	47.45	85.52	57.39
DeMatch [Zhang *et al.*, 2024]	61.33	**92.77**	71.19	58.74	**91.02**	68.91	56.00	88.90	65.10	48.27	85.24	58.07
BCLNet [Miao *et al.*, 2024]	78.36	82.23	79.87	77.90	80.07	78.73	66.20	74.12	69.19	61.14	68.33	63.92
MSGSA [Lin *et al.*, 2024a]	63.29	91.37	72.40	60.03	89.34	69.53	55.92	88.56	65.08	47.99	84.53	57.81
CGR-Net [Yang *et al.*, 2024]	78.31	82.34	79.90	77.22	79.61	78.14	66.46	74.46	69.51	61.24	68.85	64.18
MGCA-Net	**84.84**	84.13	**83.83**	**83.62**	81.07	**81.82**	**74.91**	74.29	**74.31**	**70.03**	69.86	**69.63**

Table 1: Quantitative comparison results of the outlier removal task on the YFCC100M and SUN3D datasets are presented as Precision (%), Recall (%) and F-score (%), with the optimal and suboptimal indicators highlighted in bold and underlined, respectively.

the sparse graphs $\mathcal{G}, \mathcal{G}_1^{(M-1)}, \mathcal{G}_2^{(M-1)}, \mathcal{G}_3^{(M-1)}$ generated at each stage. Specifically, it is completed through the deep concatenation operation of node features $\mathcal{V}$ and edge features $\mathcal{E}$, as follows:

$$\mathcal{G}_{\mathrm{f}} = concat([\mathcal{G}, \mathcal{G}_1^{(M-1)}, \mathcal{G}_2^{(M-1)}, \mathcal{G}_3^{(M-1)}]), \qquad (8)$$

where $concat(\cdot)$ represents the concatenation operation, which is used to combine the graph features of different stages into a unified representation space.

Finally, the generated fusion graph $\mathcal{G}_{\mathrm{f}}$ not only integrates the multi-scale information of features at each stage, but also strengthens the global geometric consistency between nodes and edges through joint representation.

3) Cross-stage Graph Consensus Aggregation: The cross-stage graph consensus aggregation strategy enhances the global geometric consistency and suppresses the interference of redundant information on subsequent tasks by dynamically weighting and aggregating the feature graphs of each stage at multiple levels.

First, for the multi-stage information in the consensus feature graph $\mathcal{G}_{\mathrm{f}}$, we introduce MLP to perform the feature alignment, which aligns the cross-stage feature information and unifies it into a shared feature space, as follows:

$$\mathbf{c}_{\mathrm{f}} = MLP\left(\mathcal{G}_{\mathrm{f}}\right), \qquad (9)$$

where $\mathbf{c}_{\mathrm{f}}$ is the fused feature graph after MLP processing. Then, to aggregate features with geometric consensus relationships of interior points, we employ Annular Convolution (AC) to preserve the relationships between nodes in parallel. Annular Convolution divides k neighbors into k/p annular regions based on their affinity to the anchor point and aggregates the contextual information of each annular region using a $1 \times p$ convolution kernel to maintain the relative relationships between neighbors as follows:

$$\tilde{e}_n^i = XW_n e_j^i + b_n, \quad (n-1)p \leq j \leq np, \qquad (10)$$

where $\tilde{e}_n^i$ denotes the aggregated features of the n-th annular region, W_n and b_n are learnable parameters, and e_j^i is the neighbor information of the i-th feature point. Finally, we process the aggregated features by MLP. The strategy ensures that the ring convolutionally aggregated features can fully express local geometric and global semantic information.

3.4 Loss Function

In the two-view correspondence task, we aim to classify inliers and outliers while estimating the camera pose via fundamental matrix regression. To optimize both tasks simultaneously, we design a hybrid loss function that combines classification and regression objectives, as follows:

$$Loss = l_c(W, L) + \gamma l_e(\hat{E}, E), \qquad (11)$$

where $l_c(W, L)$ represents the classification loss, which is used to distinguish inliers and outliers; γ is the weight hyperparameter used to balance the classification loss and regression loss, which is set to 0.5, and $l_e(\hat{E}, E)$ represents the fundamental matrix regression loss, as follows:

$$l_e(\hat{E}, E) = \frac{(p^T \hat{E} p')^2}{\|Ep\|_{[1]}^2 + \|Ep\|_{[2]}^2 + \|Ep'\|_{[1]}^2 + \|Ep'\|_{[2]}^2}, \qquad (12)$$

where $\hat{E}$ and E represent the estimated and true fundamental matrices respectively; p and p' are the corresponding feature point coordinates in the two images; the numerator $(p^T \hat{E} p')^2$ represents the squared geometric error of the feature point coordinates under the estimated fundamental matrix; $\| \cdot \|_{[i]}$ represents the sum of the squares of the elements at the i-th position in the vector, which ensures the normalization and regularization of the geometric errors at different scales.

4 Experiments

4.1 Implementation Details

The number of initial correspondences N for MGCA-Net is set to 2000, with a network dimension of 128. In addition, the number of input neighbors k of each stage of CSMGC is set to 3, and the number of clusters in the Order-Aware block is set to 500. The training process follows the training strategies from previous benchmarks [Zhang *et al.*, 2019] and CGR-Net [Yang *et al.*, 2024], and it is trained for a total of 500k iterations on Ubuntu 18.04 with an NVIDIA GTX 3090.

4.2 Datasets and Evaluation Metrics

To evaluate the proposed MGCA-Net, we use representative outdoor and indoor datasets for training and testing.

Outdoor Dataset: The YFCC100M dataset [Thomee *et al.*, 2016] contains 99.2 million images and 0.8 million videos with rich metadata. We select 72 outdoor scene sequences, with 68 used for training, validation, and testing, and 4 as unknown scenes for generalization evaluation.

Indoor Dataset: The SUN3D dataset [Xiao *et al.*, 2013] consists of 254 RGB-D video scenes. We use 239 for training, validation, and testing, and 15 for evaluating generalization.

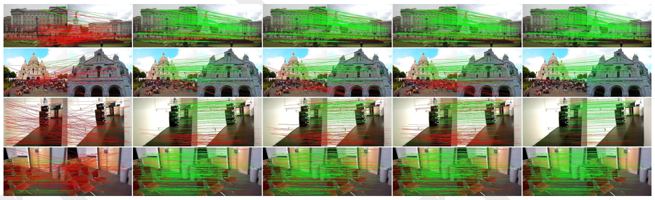
Outlier Rejection: To evaluate the effectiveness of the proposed MGCA-Net in the outlier rejection task, we used three common metrics: Precision (P), Recall (R), and F-score (F), which reflect the effectiveness of MGCA-Net in outlier identification and removal.

Camera Pose Estimation: For camera pose estimation task, we assess the performance of MGCA-Net using Mean Average Precision (mAP) and Area Under the Curve (AUC) at 5° and 20° thresholds.

4.3 Outlier Rejection

To evaluate the effectiveness of MGCA-Net, we compare it with 21 representative methods, including traditional approaches such as RANSAC [Fischler and Bolles, 1981]; MLP-based deep learning methods such as PointNet++ [Qi *et al.*, 2017], LFGC-Net [Yi *et al.*, 2018], OANet [Zhang *et al.*, 2019], NCM-Net [Liu *et al.*, 2024], and CGR-Net [Yang *et al.*, 2024]; and CNN-based methods such as ConvMatch$^+$ [Zhang and Ma, 2023b], PT-Net [Gong *et al.*, 2024], and DeMatch [Zhang *et al.*, 2024]. Experimental results for RANSAC [Fischler and Bolles, 1981], PointNet++ [Qi *et al.*, 2017], LFGC-Net [Yi *et al.*, 2018], DFE-Net [Ranftl and Koltun, 2018], and ACNe-Net [Sun *et al.*, 2020] are cited from T-Net [Zhong *et al.*, 2021], while other results are obtained using publicly available code under consistent experimental settings for fair comparison.

As reported in Table 1, MGCA-Net outperforms other competing methods on both the YFCC and SUN3D datasets, demonstrating its significant advantage in the two-view outlier removal task. For the YFCC100M dataset, in known scenes, MGCA-Net's P and F scores reach 84.84% and 83.83%, respectively, which are 6.48% and 3.93% higher than those of the second-ranked BCLNet. In unknown scenes, MGCA-Net's P and F scores are respectively 5.72% and 3.09% higher than those of BCLNet. For the SUN3D dataset, in known scenes, MGCA-Net's P and F scores are 8.45% and 4.8% higher than those of the second-ranked CGR-Net, while in unknown scenes, they achieve improvements of 8.79% and 5.45%. Experimental results demonstrate that MGCA-Net achieves higher accuracy and robustness than existing methods across different scenes.



(a) BCLNet (b) MSGSA (c) ConvMatch (d) PT-Net (e) MGCA-Net

Figure 4: Qualitative results of outlier removal. The first and second rows show outdoor scenes from YFCC100M, while the third and fourth rows depict indoor scenes from SUN3D. False matches are marked in red and correct matches are marked in green.

Although MGCA-Net achieves high Precision and F-score, its Recall is slightly lower than DeMatch and ConvMatch in some cases, mainly due to different logit threshold settings. Lower thresholds improve Recall but reduce Precision and F-score. As illustrated in Fig. 4, MGCA-Net yields fewer incorrect matches and superior overall matching in both indoor and outdoor scenes, highlighting its robustness and accuracy.

4.4 Camera Pose Estimation

To assess the performance of MGCA-Net in camera pose estimation, we compared MGCA-Net with 16 representative methods, including traditional approaches such as RANSAC [Fischler and Bolles, 1981]; MLP-based deep learning methods such as PointNet++ [Qi *et al.*, 2017], LFGC-Net [Yi *et al.*, 2018], DFE-Net [Ranftl and Koltun, 2018], and ACNe-Net [Sun *et al.*, 2020]; and CNN-based methods. The results for RANSAC, PointNet++, LFGC-Net, DFE-Net, and ACNe-Net are cited from T-Net [Zhong *et al.*, 2021].

As reported in Table 2, MGCA-Net achieves substantial gains over the second-best methods on YFCC100M and SUN3D, covering both indoor and outdoor scenes. On YFCC100M, MGCA-Net outperforms traditional methods [Fischler and Bolles, 1981] in mAP@5° and mAP@20° by 59.44% and 64.73% for known scenes, and by 68.05% and 66.52% for unknown scenes. On SUN3D, the improvements reach 28.63% and 43.84% (known), and 22.64% and 40.31% (unknown). Compared with SOTA deep learning approaches such as MSGSA [Lin *et al.*, 2024a], BCLNet [Miao *et al.*, 2024], and DeMatch [Zhang *et al.*, 2024], MGCA-Net achieves higher mAP@5°, mAP@20°, AUC@5°, and AUC@20° by 11.31%, 7.36%, 8.31%, and 7.72% in known YFCC100M scenes, and by 9.25%, 4.66%, 7.4%, and 4.92% in unknown scenes; for SUN3D, the gains are 4.57%, 3.63%, 2.84%, and 3.43% (known), and 2.03%, 1.7%, 1.23%, and 1.6% (unknown), respectively.

The above experimental results show that MGCA-Net performs well in different datasets and scenes, especially in complex scenes. This is due to the fact that MGCA-Net effectively combines the feature representations of content and location through CGA, which in turn improves the feature extraction capability for local geometric consistency and

Dataset	YFCC100M (%)				SUN3D (%)			
Method	Known Scene		Unknown Scene		Known Scene		Unknown Scene	
	5°	20°	5°	20°	5°	20°	5°	20°
RANSAC [Fischler and Bolles, 1981]	5.74 / -	16.67 / -	9.05 / -	22.71 / -	4.43 / -	15.38 / -	2.85 / -	11.23 / -
PointNet++ [Qi *et al.*, 2017]	11.88 / -	32.86 / -	15.98 / -	44.82 / -	8.78 / -	31.02 / -	7.22 / -	29.77 / -
LFGC-Net [Yi *et al.*, 2018]	14.51 / -	35.82 / -	23.71 / -	50.57 / -	11.93 / -	36.03 / -	9.73 / -	33.09 / -
DFE-Net [Ranftl and Koltun, 2018]	19.27 / -	42.14 / -	30.55 / -	59.15 / -	14.18 / -	39.14 / -	12.13 / -	26.26 / -
ACNe-Net [Sun *et al.*, 2020]	29.63 / -	52.71 / -	34.00 / -	62.98 / -	19.08 / -	46.32 / -	14.27 / -	39.29 / -
OANet [Zhang *et al.*, 2019]	33.75 / 14.49	57.13 / 48.04	40.80 / 17.00	69.26 / 58.61	21.54 / 8.04	48.91 / 40.14	16.37 / 6.09	41.82 / 34.08
T-Net [Zhong *et al.*, 2021]	41.46 / 18.20	64.14 / 54.48	48.40 / 20.70	73.92 / 62.94	23.63 / 9.03	51.04 / 42.08	18.04 / 6.62	43.65 / 35.66
PESA-Net [Zhong *et al.*, 2022]	37.15 / 16.09	59.76 / 50.32	45.03 / 19.73	71.95 / 61.23	22.67 / 9.07	50.02 / 42.46	18.00 / 7.26	44.10 / 37.47
MSA-Net [Zheng *et al.*, 2022]	36.04 / 15.69	58.81 / 49.68	48.70 / 20.95	73.07 / 62.31	19.58 / 7.24	47.03 / 38.54	16.77 / 6.07	42.10 / 34.32
MS²DG-Net [Dai *et al.*, 2022]	36.46 / 15.36	61.94 / 52.18	46.88 / 18.84	74.84 / 63.27	22.57 / 8.51	50.89 / 41.81	17.19 / 6.24	43.27 / 35.22
U-Match [Li *et al.*, 2023]	46.22 / 21.73	67.67 / 57.90	60.15 / 29.59	79.69 / 69.03	26.45 / 10.20	53.56 / 44.38	22.41 / 8.39	48.65 / 40.07
PGFNet [Liu *et al.*, 2023]	33.54 / 14.17	56.73 / 47.57	46.70 / 20.48	72.26 / 61.62	22.63 / 8.64	49.26 / 40.55	19.02 / 7.14	44.70 / 36.61
GCA-Net [Guo *et al.*, 2023]	44.32 / 19.93	67.24 / 57.16	55.70 / 24.90	79.58 / 68.21	22.57 / 8.61	50.39 / 41.49	18.53 / 6.85	44.27 / 36.08
ConvMatch [Zhang and Ma, 2023a]	43.25 / 20.02	65.61 / 55.81	55.45 / 26.69	77.53 / 66.87	27.45 / 10.84	54.65 / 45.35	22.52 / 8.68	48.64 / 40.09
ConvMatch+ [Zhang and Ma, 2023b]	45.79 / 21.19	67.69 / 57.72	58.07 / 27.79	78.88 / 68.00	27.22 / 10.81	54.97 / 45.70	22.67 / 8.62	49.13 / 40.51
NCM-Net [Liu *et al.*, 2024]	50.24 / 25.02	71.27 / 61.31	62.65 / 32.40	82.30 / 71.82	24.99 / 9.96	51.87 / 42.88	20.41 / 7.92	46.42 / 38.16
PT-Net [Gong *et al.*, 2024]	49.16 / 20.23	71.07 / 56.23	61.62 / 27.06	81.21 / 67.45	27.23 / 10.67	54.38 / 45.11	22.88 / 8.59	48.52 / 39.85
DeMatch [Zhang *et al.*, 2024]	47.56 / 22.51	69.14 / 59.17	60.00 / 30.01	80.02 / 69.45	28.49 / 11.34	55.59 / 46.24	23.46 / 9.20	49.84 / 41.20
BCLNet [Miao *et al.*, 2024]	53.21 / 27.48	73.48 / 63.63	67.85 / 37.22	84.57 / 74.44	24.32 / 9.64	51.24 / 42.37	20.06 / 7.72	45.83 / 37.62
MSGSA [Lin *et al.*, 2024a]	53.87 / 21.99	74.04 / 58.26	63.78 / 28.23	82.23 / 68.31	25.28 / 10.43	52.29 / 44.99	20.41 / 7.81	46.12 / 39.00
CGR-Net [Yang *et al.*, 2024]	53.56 / 26.45	73.98 / 62.97	66.47 / 35.01	84.14 / 73.17	26.48 / 10.58	53.41 / 44.28	21.69 / 8.51	47.94 / 39.53
MGCA-Net	**65.18 / 35.79**	**81.40 / 71.35**	**77.10 / 44.62**	**89.23 / 79.36**	**33.06 / 14.18**	**59.22 / 49.67**	**25.49 / 10.43**	**51.54 / 42.80**

Table 2: Quantitative comparison results of the relative pose estimation task on the YFCC100M and SUN3D datasets are presented in the form of mAP AUC at 5° and 20°, with the best and second-best results highlighted in bold and underlined, respectively.

global contextual information. Secondly, based on CSMGC, MGCA-Net is able to integrate multi-scale geometric features across stages and construct cross-stage geometric consensus relations through multiple different sparse graphs.

4.5 Ablation Experiments

MGCA-Net consists of two main modules: CGA, which serves as the core for contextual and geometric feature extraction, and CSMGC, which enhances cross-stage geometric consistency. To evaluate their effectiveness, we conducted ablation experiments on unknown scenes from YFCC100M, as shown in Table 3. The results demonstrate that each module incrementally improves performance, each component contributes to the network performance as follows:

Baseline: When only the basic module is used for feature extraction, mAP@5° and mAP@20° are 56.70% and 79.33%, respectively, highlighting the limitations without contextual and geometric consistency information.

Iter: After adding the multi-stage iterative structure, mAP5° and mAP20° reach 58.20% and 80.46%, respectively, demonstrating that iterative optimization improves feature extraction, though performance remains limited without effective use of intermediate information.

CGA: After adding CGA, the network performance is further improved, and mAP5° and mAP20° reach 66.57% and 84.58%, respectively, which shows that the fusion of context features and graph coordinate features contributes significantly to the model performance.

Iter + CGA: After combining the iterative network and CGA module, mAP5° and mAP20° reach 69.88% and 86.39% respectively, which reflects the importance of the collaboration between them.

Iter + CSMGC: After combining the iterative network and CSMGC, mAP5° and mAP20° reach 73.70% and 88.26% respectively, which verifies the importance of cross-stage feature fusion and geometric consistency modeling.

Baseline	Iter	CGA	CSMGC	mAP5° (%)	mAP20° (%)
✓				56.70	79.33
✓		✓		58.20	80.46
✓	✓			66.57	84.58
✓	✓	✓		69.88	86.39
✓	✓		✓	73.70	88.26
✓	✓	✓	✓	77.10	89.23

Table 3: Ablation study on unknown scenes with different modules.

Full Model: With CGA, CSMGC, and multi-stage iteration, mAP@5° and mAP@20° reach 77.10% and 89.23%, significantly outperforming other settings. These results confirm that combining CGA and CSMGC with multi-stage structure fully unlocks MGCA-Net's potential for geometric feature representation and matching.

5 Conclusion

In this paper, we propose MGCA-Net, a simple yet effective framework for two-view geometric correspondence learning that captures deep geometric relationships by integrating contextual and positional information. MGCA-Net comprises the CGA and CSMGC modules, which jointly enhance feature representation and establish geometric consensus across different stages. Experimental results on the YFCC100M and SUN3D datasets demonstrate that MGCA-Net achieves robust performance in both indoor and outdoor scenes, showing stable outlier rejection and significant improvements in camera pose estimation. Compared to traditional and state-of-the-art deep learning methods, MGCA-Net exhibits superior adaptability in terms of matching accuracy and robustness, particularly in complex scenarios. In future work, we will explore extending MGCA-Net to unsupervised or self-supervised learning frameworks to reduce reliance on labeled data and enhance its ability to generalize in different domains.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. U22A2095, 62476112, 62202249); in part by the Guangdong Basic and Applied Basic Research Foundation (Grant Nos. 2024A1515011740, 2025A1515010181); in part by the Fundamental Research Funds for the Central Universities (Grant Nos. 21624404, 23JNSYS01); in part by the China Postdoctoral Science Foundation (Grant Nos. GZC20233362, 2024MD754043) and Chongqing Municipal Education Commission (Grant No. KJQN202400648); in part by the Major Key Project of PCL (Grant No. PCL2024A04-4); in part by Guangdong Key Laboratory of Data Security and Privacy Preserving (Grant No. 2023B1212060036); and in part by Guangdong-Hong Kong Joint Laboratory for Data Security and Privacy Preserving (Grant No. 2023B1212120007).

References

[Barath and Matas, 2018] Daniel Barath and Jiří Matas. Graph-cut ransac. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6733–6741, 2018.

[Brachmann and Rother, 2019] Eric Brachmann and Carsten Rother. Neural-guided ransac: Learning where to sample model hypotheses. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4322–4331, 2019.

[Chen *et al.*, 2024] Changhao Chen, Bing Wang, Chris Xiaoxuan Lu, Niki Trigoni, and Andrew Markham. Deep learning for visual localization and mapping: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(12):17000–17020, 2024.

[Dai *et al.*, 2022] Luanyuan Dai, Yizhang Liu, Jiayi Ma, Lifang Wei, Taotao Lai, Changcai Yang, and Riqing Chen. Ms2dg-net: Progressive correspondence learning via multiple sparse semantics dynamic graph. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8973–8982, 2022.

[DeTone *et al.*, 2018] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 224–236, 2018.

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

[Gong *et al.*, 2024] Zhepeng Gong, Guobao Xiao, Ziwei Shi, Shiping Wang, and Riqing Chen. Pt-net: Pyramid transformer network for feature matching learning. *IEEE Transactions on Instrumentation and Measurement*, 73:1–11, 2024.

[Guo *et al.*, 2023] Junwen Guo, Guobao Xiao, Zhimin Tang, Shunxing Chen, Shiping Wang, and Jiayi Ma. Learning for feature matching via graph context attention. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023.

[Li *et al.*, 2023] Zizhuo Li, Shihua Zhang, and Jiayi Ma. U-match: Two-view correspondence learning with hierarchy-aware local context aggregation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1169–1176, 2023.

[Liao *et al.*, 2023] Tangfei Liao, Xiaoqin Zhang, Yuewang Xu, Ziwei Shi, and Guobao Xiao. Sga-net: A sparse graph attention network for two-view correspondence learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12):7578–7590, 2023.

[Liao *et al.*, 2024] Tangfei Liao, Xiaoqin Zhang, Li Zhao, Tao Wang, and Guobao Xiao. Vsformer: Visual-spatial fusion transformer for correspondence pruning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3369–3377, 2024.

[Lin *et al.*, 2022] Shuyuan Lin, Hailing Luo, Yan Yan, Guobao Xiao, and Hanzi Wang. Co-clustering on bipartite graphs for robust model fitting. *IEEE Transactions on Image Processing*, 31:6605–6620, 2022.

[Lin *et al.*, 2024a] Shuyuan Lin, Xiao Chen, Guobao Xiao, Hanzi Wang, Feiran Huang, and Jian Weng. Multi-stage network with geometric semantic attention for two-view correspondence learning. *IEEE Transactions on Image Processing*, 33:3031–3046, 2024.

[Lin *et al.*, 2024b] Shuyuan Lin, Feiran Huang, Taotao Lai, Jianhuang Lai, Hanzi Wang, and Jian Weng. Robust heterogeneous model fitting for multi-source image correspondences. *International Journal of Computer Vision*, 132:2907–2928, 2024.

[Liu *et al.*, 2023] Xin Liu, Guobao Xiao, Riqing Chen, and Jiayi Ma. Pgfnet: Preference-guided filtering network for two-view correspondence learning. *IEEE Transactions on Image Processing*, 32:1367–1378, 2023.

[Liu *et al.*, 2024] Xin Liu, Rong Qin, Junchi Yan, and Jufeng Yang. Ncmnet: Neighbor consistency mining network for two-view correspondence pruning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):11254–11272, 2024.

[Lowe, 2004] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, pages 91–110, 2004.

[Ma *et al.*, 2014] Jiayi Ma, Ji Zhao, Jinwen Tian, Alan L Yuille, and Zhuowen Tu. Robust point matching via vector field consensus. *IEEE Transactions on Image Processing*, pages 1706–1721, 2014.

[Ma *et al.*, 2019] Jiayi Ma, Ji Zhao, Junjun Jiang, Huabing Zhou, and Xiaojie Guo. Locality preserving matching. *International Journal of Computer Vision*, 127:512–531, 2019.

[Ma *et al.*, 2021] Jiayi Ma, Xingyu Jiang, Aoxiang Fan, Junjun Jiang, and Junchi Yan. Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision*, pages 23–79, 2021.

[Miao *et al.*, 2024] Xiangyang Miao, Guobao Xiao, Shiping Wang, and Jun Yu. Bclnet: Bilateral consensus learning for two-view correspondence pruning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4225–4232, 2024.

[Placed *et al.*, 2023] Julio A Placed, Jared Strader, Henry Carrillo, Nikolay Atanasov, Vadim Indelman, Luca Carlone, and José A Castellanos. A survey on active simultaneous localization and mapping: State of the art and new frontiers. *IEEE Transactions on Robotics*, 39(3):1686–1705, 2023.

[Qi *et al.*, 2017] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Proceedings of the Advances in Neural Information Processing Systems*, pages 5099–5108, 2017.

[Raguram *et al.*, 2012] Rahul Raguram, Ondrej Chum, Marc Pollefeys, Jiri Matas, and Jan-Michael Frahm. Usac: A universal framework for random sample consensus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 2022–2038, 2012.

[Ranftl and Koltun, 2018] René Ranftl and Vladlen Koltun. Deep fundamental matrix estimation. In *Proceedings of the European Conference on Computer Vision*, pages 284–299, 2018.

[Sarlin *et al.*, 2020] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4938–4947, 2020.

[Schmied *et al.*, 2023] Aron Schmied, Tobias Fischer, Martin Danelljan, Marc Pollefeys, and Fisher Yu. R3d3: Dense 3d reconstruction of dynamic scenes from multiple cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3216–3226, 2023.

[Schonberger and Frahm, 2016] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016.

[Sun *et al.*, 2020] Weiwei Sun, Wei Jiang, Eduard Trulls, Andrea Tagliasacchi, and Kwang Moo Yi. Acne: Attentive context normalization for robust permutation-equivariant learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11286–11295, 2020.

[Sun *et al.*, 2021] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8922–8931, 2021.

[Thomee *et al.*, 2016] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, pages 64–73, 2016.

[Xiao *et al.*, 2013] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1625–1632, 2013.

[Yang *et al.*, 2024] Changcai Yang, Xiaojie Li, Jiayi Ma, Fengyuan Zhuang, Lifang Wei, Riqing Chen, and Guodong Chen. Cgr-net: Consistency guided resformer for two-view correspondence learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(12):12450–12465, 2024.

[Yi *et al.*, 2018] Kwang Moo Yi, Eduard Trulls, Yuki Ono, Vincent Lepetit, Mathieu Salzmann, and Pascal Fua. Learning to find good correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2666–2674, 2018.

[Zhang and Ma, 2023a] Shihua Zhang and Jiayi Ma. Convmatch: Rethinking network design for two-view correspondence learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3472–3479, 2023.

[Zhang and Ma, 2023b] Shihua Zhang and Jiayi Ma. Convmatch: Rethinking network design for two-view correspondence learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):2920–2935, 2023.

[Zhang *et al.*, 2019] Jiahui Zhang, Dawei Sun, Zixin Luo, Anbang Yao, Lei Zhou, Tianwei Shen, Yurong Chen, Long Quan, and Hongen Liao. Learning two-view correspondences and geometry using order-aware network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5845–5854, 2019.

[Zhang *et al.*, 2024] Shihua Zhang, Zizhuo Li, Yuan Gao, and Jiayi Ma. Dematch: Deep decomposition of motion field for two-view correspondence learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 20278–20287, 2024.

[Zhao *et al.*, 2021] Chen Zhao, Yixiao Ge, Feng Zhu, Rui Zhao, Hongsheng Li, and Mathieu Salzmann. Progressive correspondence pruning by consensus learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6464–6473, 2021.

[Zheng *et al.*, 2022] Linxin Zheng, Guobao Xiao, Ziwei Shi, Shiping Wang, and Jiayi Ma. Msa-net: Establishing reliable correspondences by multiscale attention network. *IEEE Transactions on Image Processing*, pages 4598–4608, 2022.

[Zhong *et al.*, 2021] Zhen Zhong, Guobao Xiao, Linxin Zheng, Yan Lu, and Jiayi Ma. T-net: Effective permutation-equivariant network for two-view correspondence learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1950–1959, 2021.

[Zhong *et al.*, 2022] Zhen Zhong, Guobao Xiao, Shiping Wang, Leyi Wei, and Xiaoqin Zhang. Pesa-net: Permutation-equivariant split attention network for correspondence learning. *Information Fusion*, pages 81–89, 2022.