BAG-Net: Bidirectional Receptive-Field Graph Network for Two-View Correspondence Pruning

Leyi Wang[1], **Hao Chen**[1,2*], **Changcai Yang**[1,2*]
[1]College of Computer and Information Sciences, Fujian Agriculture and Forestry University, Fuzhou 350002, China
[2]Agriculture and Forestry Artificial Intelligence Research Institute, Fujian Agriculture and Forestry University, Fuzhou 350002, China
wanglyfafu@163.com, chenhaox@outlook.com, changcaiyang@gmail.com

Abstract

Learning reliable two-view correspondences is essential for geometric computer vision applications. Existing graph-based pruning methods typically aggregate information unidirectionally, capturing only whether a correspondence receives neighborhood support while ignoring its structural contribution during aggregation. This makes it difficult to distinguish locally clustered outliers from globally critical inliers. To overcome this limitation, we propose a Bidirectional Receptive-Field Graph Network (BAG-Net), which introduces a bidirectional aggregation perspective for correspondence pruning. Specifically, we design Bidirectional Dynamic Graph Convolution (BiDGC) that jointly models forward consistency aggregation and reverse structural dependency, enabling the network to identify key inliers that are widely referenced during feature propagation. To further enhance robustness, we design a Stage-wise Graph Expansion-Contraction (SGEC) architecture that combines multi-scale receptive field evolution with Stage-wise Squeeze-and-Excitation (SSE) and Deep-Guided Modulation (DGM) to stabilize feature propagation across network stages. Extensive experiments on the YFCC100M and SUN3D datasets demonstrate that BAG-Net significantly outperforms existing SOTA methods in both outlier rejection and camera pose estimation, achieving 66.10% and 78.88% mAP5°(%) on YFCC100M in known and unknown scenes, respectively. Source code: https://github.com/LeyiWang13/BAG-Net.

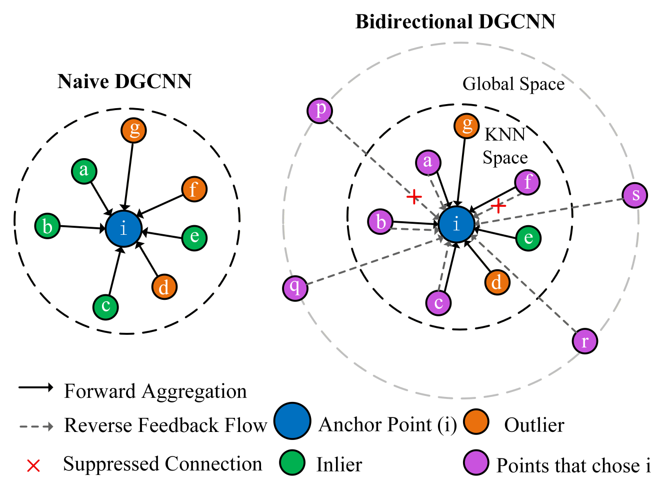

Figure 1: Comparison of Naive DGCNN and Bidirectional DGCNN.

1 Introduction

Establishing robust correspondences between two images stands as a pivotal component in the field of multimedia processing, which can provide a foundation for downstream visual applications, such as visual simultaneous localization and mapping (SLAM) [Huang *et al.*, 2020], structure from motion (SFM) [Wang *et al.*, 2017], point cloud registration [Chen *et al.*, 2019], image retrieval [Sattler *et al.*, 2012], and image fusion [Kaur *et al.*, 2021]. By establishing stable feature correspondences between two images, essential constraints can be provided for geometric modeling and robust localization. However, in real-world scenarios, initial matching results are often heavily contaminated by erroneous correspondences (outliers) due to occlusions, illumination variations, viewpoint changes, and repetitive structures. These outliers not only significantly degrade matching accuracy but also induce error propagation in downstream tasks such as pose estimation and 3D reconstruction. Consequently, effectively distinguishing true correspondences from false ones under high outlier ratios remains a core challenge in two-view correspondence pruning.

Traditional methods, represented by RANSAC [Fischler and Bolles, 1981] and its variants [Barath and Matas, 2018], [Barath *et al.*, 2020], estimate geometric models through random sampling and consistency verification. Although they exhibit certain robustness under low outlier ratios, their performance and efficiency deteriorate markedly in the presence of high outlier rates or complex scene structures. Moreover, their reliance on predefined parameters and geometric assumptions further limits their applicability in challenging real-world conditions. With the advancement of deep learning, learning-based methods have achieved significant progress in corre-

*Corresponding Authors.

spondence pruning. Approaches such as LFGC-Net [Yi *et al.*, 2018] and OANet [Zhang *et al.*, 2019] formulate pruning as a point-wise binary classification problem and incorporate global contextual information to suppress random outliers. Subsequent methods, including ACNe [Sun *et al.*, 2020], enhance feature interactions through attention mechanisms. More recently, CLNet [Zhao *et al.*, 2021] and MS^{2}DGNet [Dai *et al.*, 2022] explicitly model geometric dependencies among correspondences using graph structures, while ConvMatch [Zhang and Ma, 2023], SC-Net [Lin *et al.*, 2026], and Match-Mamba [Wu *et al.*, 2026] improve representational capacity in complex scenes through local convolutions, spatial-channel modeling, or state space mechanisms.

Despite continuous performance improvements, most existing methods share a similar underlying discriminative principle: the reliability of a correspondence is primarily determined by whether it receives sufficient consistency support from other correspondences. Whether based on dynamic graph aggregation, local convolution, or global context modeling, these methods aim to reinforce correspondences supported by the majority while suppressing isolated or inconsistent outliers. However, in complex scenarios with repetitive structures or systematic noise, incorrect correspondences may form highly consistent support clusters and thus be assigned high confidence. In contrast, some key inliers that are critical to the overall geometric structure may be spatially sparse or locally inconspicuous, leading to insufficient support and erroneous removal. This observation indicates that consistency strength alone is insufficient to characterize the true importance of a correspondence within the global structure.

A natural question follows: if consistency support alone fails to expose key inliers, what other structural cue is hidden in the correspondence graph and yet remains unused? Motivated by this insight, we shift the pruning perspective from whether a correspondence is supported to the role it plays in global graph propagation, concerned not only with which correspondences appear consistent, but also with which are widely referenced during aggregation. To this end, we propose the Bidirectional Receptive-Field Graph Network (BAG-Net). At its core, the BiDGC module performs forward aggregation to capture neighborhood support and reverse aggregation to characterize reference patterns. Fig. 1 contrasts this design with naive DGCNN: while DGCNN models only outgoing edges, BiDGC additionally exploits incoming edges from the global space, revealing how each correspondence is invoked across the graph and suppressing misleading connections. Building upon this, SGEC adopts a receptive field expansion-contraction strategy for multi-scale context modeling, combined with cross-stage feature modulation (SSE) and intra-stage deep guidance (DGM) to stabilize feature evolution and improve pruning robustness in complex scenarios. The main contributions of this work are summarized as follows:

- We introduce a bidirectional aggregation perspective that jointly considers the consistency support a correspondence receives from its neighborhood and its reference pattern during graph aggregation, exposing a structural cue overlooked by prior pruning methods. To our knowledge, this is the first attempt to leverage reverse structural dependency in KNN graphs to handle the task of two-view correspondence pruning.
- We propose BiDGC, which performs forward aggregation to capture neighborhood support and reverse aggregation to model structural dependency, providing complementary cues for robust correspondence discrimination in complex scenes.
- We design BAG-Net with an SGEC architecture, which combines multi-scale receptive field evolution with SSE and DGM to stabilize feature propagation and improve pruning robustness, where SSE ensures inter-stage feature continuity while DGM adaptively refines shallow features using deep guidance.

2 Related Work

2.1 Learning-Based Methods for Robust Correspondences

Learning-based methods reformulate correspondence pruning as a per-point classification task, and have evolved by progressively enriching the contextual information available to each correspondence. LFGC-Net [Yi *et al.*, 2018] is one of the pioneering works in this area, proposing an end-to-end correspondence classification network that transforms the feature matching problem into a binary classification task of inliers vs. outliers. However, LFGC-Net relies solely on Context Normalization for global statistics, lacking local neighborhood modeling and thus underperforming under high outlier ratios. To overcome this limitation, subsequent research has introduced more complex network structures and contextual modeling mechanisms. OANet [Zhang *et al.*, 2019] models local and global contextual relationships using modules such as DiffPool and DiffUnpool; ACNe [Sun *et al.*, 2020] employs an attention-based contextual normalization mechanism, combining local and global attention operations to enhance feature representation; PGFNet [Liu *et al.*, 2023] uses group attention to compute global preference scores, achieving more effective outlier removal. Despite improvements in contextual information modeling, these methods largely overlook the potential graph-structured context between correspondences, and thus, the performance gains remain limited.

2.2 Graph-Based Methods for Robust Correspondences

In recent years, exploiting graph neural networks (GNNs) to mine geometric consistency among correspondences has emerged as a prominent research direction. GNN-based methods represent correspondences and their relationships as graph structures, capturing consistency patterns through message passing and aggregation among nodes. NM-Net [Zhao *et al.*, 2019] proposes a compatibility-specific local graph context mining strategy, hierarchically extracting and aggregating local correspondences to discover consistent neighborhoods. MS^{2}DGNet [Dai *et al.*, 2022] designs sparse semantic dynamic graphs to capture local topological information among correspondences. CLNet [Zhao *et al.*, 2021] and MGNet [Dai *et al.*, 2024] construct neighborhood graphs by combining KNN with GCN layers, enabling the exploration

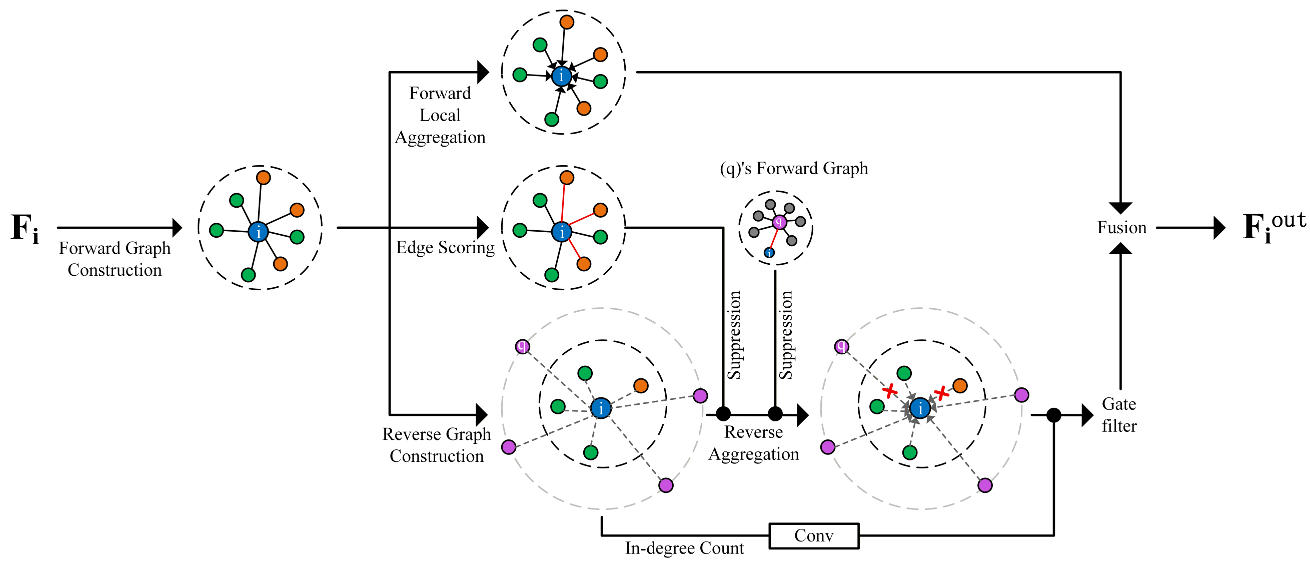

Figure 2: Pipeline of BiDGC. Forward Local Aggregation, Edge Scoring, Conv, and the Gate filter are all MLP-based. Red edges denote unreliable connections identified by Edge Scoring, and red "×" marks reverse edges suppressed during reverse aggregation.

of neighborhood relationships from local to global scales. Going beyond a single neighborhood, NCMNet [Liu and Yang, 2023] explores graph context across coordinate, feature, and global neighborhoods, and DDFNet [Yang *et al.*, 2026] further constructs dual neighborhoods in feature and motion-structure spaces with a dynamic fusion strategy, jointly enriching neighborhood consistency under complex scenes. In addition, GCT-Net [Guo *et al.*, 2024] captures richer contextual dependencies via a graph-context-enhanced transformation module.

Most existing methods use KNN-based graphs and GCNs to score correspondences, but rely on unidirectional aggregation that only captures "who supports me". This design fails to distinguish locally clustered yet globally isolated outliers from truly important inliers. To overcome this limitation, we propose BAG-Net, which introduces reverse aggregation to model "who needs me" alongside forward support. This bidirectional aggregation jointly captures neighborhood support and reverse structural dependency, suppressing clustered outliers while strengthening critical inliers.

3 Methodology

3.1 Problem Formulation

Given a pair of images (I, I'), keypoints and descriptors are first extracted using standard feature detection and description methods, such as SIFT [Lowe, 2004] or SuperPoint [DeTone *et al.*, 2018]. An initial correspondence set $\mathbf{C} = \{c_i\} \in \mathbb{R}^{N\times 4}, i = 1, 2, \cdots N$, is then established based on nearest-neighbor matching of descriptors, where N denotes the total number of initial correspondences. Here, $c_i = (x_i, y_i, x'_i, y'_i)$ represents the i-th correspondence with coordinates normalized by camera intrinsics.

Since $\mathbf{C}$ typically contains a large proportion of outliers, we propose BAG-Net, which adopts an iterative pruning strategy to select reliable inliers. Specifically, two cascaded pruning modules, f_Φ and f_Ψ, are used to iteratively process the initial correspondence set $\mathbf{C}$ and produce two pruned subsets $\mathbf{C}_1 = f_\Phi(\mathbf{C})$, $\mathbf{C}_2 = f_\Psi(\mathbf{C}_1)$, in which $\mathbf{C}_1 \in \mathbb{R}^{N_1\times 4}$ and $\mathbf{C}_2 \in \mathbb{R}^{N_2\times 4}$ $(N_2 < N_1 < N)$ denote the retained correspondences based on the predicted inlier probabilities. In addition, inlier weights $\mathbf{w}$ are computed for the pruned correspondences to support subsequent weighted essential matrix estimation and full-size verification. Formally, these steps can be written as:

$$\hat{\mathbf{E}} = g(\mathbf{C}_2, \mathbf{w}) \tag{1}$$

$$\hat{\mathbf{y}} = h(\hat{\mathbf{E}}, \mathbf{C}), \tag{2}$$

where $g(\cdot,\cdot)$ is the weighted eight-point algorithm, $h(\cdot,\cdot)$ computes epipolar distances for full-size verification, and $\hat{\mathbf{y}}$ denotes the epipolar distance set used to assess the correctness of each correspondence in $\mathbf{C}$.

3.2 Bidirectional Dynamic Graph Convolution

BiDGC updates correspondence features bidirectionally via forward consistency aggregation and backward structural dependency modeling (Fig. 2). Forward aggregation measures neighborhood support, while backward aggregation captures global structural contribution, enabling reliable identification of critical correspondences even under sparse structures.

Forward Aggregation

Given a correspondence c_i, it is first mapped to a feature vector f_i through several residual blocks. Subsequently, the k-nearest neighbors of f_i are searched in the feature space to construct a directed k-nearest-neighbor graph $G_i^N = (V_i^N, E_i^N)$, where $V_i^N = \{f_{ij}\}_{j=1}^k$ denotes the k-nearest neighbors of f_i and $\varepsilon_i = (e_{i1}, \ldots, e_{ik})$ represents the directed edges to these neighbors. The edge feature e_{ij} is defined as:

$$e_{ij} = [f_i || f_i - f_{ij}], \tag{3}$$

where $f_i - f_{ij}$ is the residual feature and $[\cdot||\cdot]$ represents a concatenation operation along the channel dimension. The edge features are then processed by an MLP, followed by max pooling to aggregate the most salient neighbor responses, thereby emphasizing credible neighborhood cues and updating the forward feature representation:

$$f_i^{fwd} = Maxpooling(MLP(\varepsilon_i)) \tag{4}$$

The forward aggregation propagates information from neighboring nodes to the center node, effectively capturing local graph context information and enhancing neighbor consistency features.

Reverse Aggregation

After the forward aggregation, we build a reverse graph from the forward graph. In this graph, the neighbors of each node are defined as the set of nodes that selected it as a neighbor during the forward aggregation phase. We then aggregate the forward features of these neighbors to the node itself, producing its reverse feature. Specifically, for any correspondence j, we gather all correspondences i that selected j as a neighbor in the forward phase and combine their forward features to obtain node j's initial reverse feature:

$$\hat{f}_j^{rev} = \sum_{i \in N_{\text{rev}}(j)} f_i^{fwd}, \tag{5}$$

where $N_{\text{rev}}(j)$ denotes the set of neighbors pointing to correspondence j in the reverse graph, i.e., all correspondences i that chose j as a nearest neighbor in the forward phase. This aggregation shows how much j is referenced by other correspondences in the global structure. However, unreliable edges may introduce noise. To address this, we introduce a learnable edge scoring mechanism for each directed edge to perform weighted aggregation. The score is predicted by an MLP and converted to a weight using a sigmoid function:

$$\boldsymbol{s}_{ij} = \sigma(MLP(e_{ij})), \tag{6}$$

where e_{ij} is the edge feature constructed in the forward phase, and s_{ij} measures the reliability of i's contribution to j in the reverse aggregation. Based on these edge scores, the reverse aggregation can be rewritten as:

$$\hat{f}_j^{rev} = \sum_{i \in N_{\text{rev}}(j)} s_{ij} \cdot f_i^{fwd} \tag{7}$$

Meanwhile, we define the weighted in-degree W_j of correspondence j as the sum of the weights of all edges pointing to j in the reverse graph:

$$W_j = \sum_{i \in N_{\text{rev}}(j)} s_{ij} \tag{8}$$

This in-degree counts how many correspondences chose j as a neighbor. To eliminate the influence of varying in-degree scales, we normalize the reverse-aggregated feature to obtain the final reverse feature:

$$f_j^{rev} = \hat{f}_j^{rev} / W_j \tag{9}$$

Adaptive Gated Fusion

To fuse the forward and reverse features, we use an adaptive gating mechanism based on in-degree. Specifically, for correspondence j, we first apply a logarithmic transformation to its weighted in-degree W_j, and then map it to a 16-dimensional in-degree embedding through an MLP:

$$h_j = MLP(\log(W_j + 1)) \tag{10}$$

The gating weight g_j is computed from both the reverse feature and the in-degree embedding:

$$g_j = \sigma(MLP([f_j^{rev} \| h_j])) \tag{11}$$

Finally, the forward feature and the gated reverse feature are concatenated, processed by an MLP, and combined with the input through a residual connection:

$$f_j^{out} = MLP([f_j^{fwd} \| g_j \cdot f_j^{rev}]) + f_j \tag{12}$$

3.3 Stage-wise Graph Expansion-Contraction Module

Architecture Design

To stabilize feature evolution during multi-stage pruning, we adopt a breathing receptive field architecture with expansion and contraction, combined with lightweight cross-stage feature modulation. The overall architecture is shown in Fig. 3.

Inspired by U-shaped networks [Ronneberger *et al.*, 2015], the breathing receptive field architecture adopts a symmetric expansion-contraction design within each pruning stage. The receptive field expands and then contracts, enabling progressive feature evolution across multiple scales. Each stage consists of three hierarchical graph context embeddings: Neighbor-level Graph Context Embedding (BiDGC) for local structural modeling, Cluster-level Graph Context Embedding (OABlock) [Zhang *et al.*, 2019] for mid-scale clustering consistency, and Global Graph Context Embedding (GCN) [Zhao *et al.*, 2021] for global geometric aggregation.

The feature evolution within a single stage starts from the initial feature F of the network and can be written as:

$$\begin{aligned} F_1 &= \text{BiDGC}(F) \\ F_2 &= \text{OABlock}(F_1) \\ F_3 &= \text{GCN}(F_2) \\ F_4 &= \text{OABlock}(F_3) \\ F_{\text{out}} &= \text{BiDGC}(F_4), \end{aligned} \tag{13}$$

where BiDGC$(\cdot)$ represents the Bidirectional Dynamic Graph Convolution module, OABlock$(\cdot)$ represents the Cluster-level Graph Context Embedding module, and GCN$(\cdot)$ represents the Global Graph Convolution module. The receptive field progressively expands during the first half (F_1 to F_3) and contracts during the second half (F_3 to F_{out}), enabling fine-grained modeling of local structural relationships under global consistency constraints. We also introduce a Deep-Guided Modulation (DGM) Module to modulate the intermediate features; its design is detailed in the next subsection.

Meanwhile, to propagate feature information across adjacent pruning stages, we introduce a lightweight cross-stage

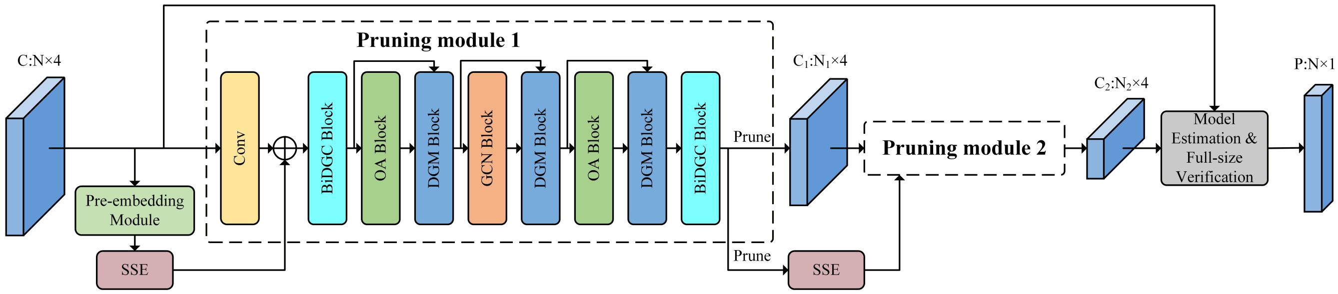

Figure 3: Overall architecture of the proposed BAG-Net.

feature modulation module, called Stage-wise Squeeze-and-Excitation (SSE). SSE uses a channel-spatial dual attention mechanism [Hu *et al.*, 2018] to adaptively modulate the output features of the current stage:

$$S^m = \text{SSE}(F_{out}^m) \tag{14}$$

The modulation signal is then added to the initial feature map of the next stage in a residual manner:

$$F^{m+1} = \text{Conv}(C^{m+1}) + S^m, \tag{15}$$

where F_{out}^m is the output of the m-th pruning stage, C^{m+1} represents the input coordinates of the $(m+1)$-th stage, Conv$(\cdot)$ maps coordinates to initial features, S^m is the cross-stage modulation signal generated by SSE, and F^{m+1} is the initial feature of the next stage after residual injection.

For the first pruning stage, we introduce a Pre-embedding Module to generate the initial modulation signal from the raw input C. This module consists of an initial feature mapping followed by two BiDGC layers:

$$\begin{aligned} S^0 &= \text{SSE}(\text{Pre-embedding Module}(C)) \\ &= \text{SSE}(\text{BiDGC}(\text{BiDGC}(\text{Conv}(C)))) \end{aligned} \tag{16}$$

The Pre-embedding Module maps the raw input C into high-dimensional graph context features, providing a stable initial modulation signal for SSE.

Overall, the breathing receptive field refines structural relations within each stage, while lightweight cross-stage modulation propagates features across stages, ensuring stable structural representations throughout multi-stage evolution.

Structural Property Analysis

Correspondence pruning is a progressive process rather than a single-step operation. Local neighborhoods reveal structural breaks and inlier-outlier relations first; thus, pruning should start from local interactions and gradually incorporate global constraints. Based on this insight, we adopt hierarchical graph embeddings within each stage. The receptive field expands from local to global to form stable geometric structures, and then symmetrically contracts to propagate global consistency back to local neighborhoods while avoiding over-smoothing. This expansion-contraction process stabilizes feature representations and preserves local discrimination, enabling reliable separation of inliers from outliers.

3.4 Deep-Guided Modulation Module

To fuse features across layers, we introduce the DGM module, which uses deep graph-enhanced features to generate adaptive weights for modulating shallow features.

Specifically, deep features F_H are mapped to channel and point attention. Two lightweight branches are used: a local branch for position-aware responses and a global branch using average pooling. Their outputs are fused and passed through a Sigmoid to get modulation weights W.

$$A_l = \text{Conv}(\text{Conv}(F_H)) \tag{17}$$

$$A_g = \text{Conv}(\text{Conv}(\text{Avgpool}(F_H))) \tag{18}$$

$$W = \sigma(A_l + A_g), \tag{19}$$

where Conv$(\cdot)$ is point-wise convolution, Avgpool$(\cdot)$ is global average pooling, and $\sigma(\cdot)$ is the Sigmoid activation. $A_l \in \mathbb{R}^{n\times d}$ and $A_g \in \mathbb{R}^{n\times d}$ denote local attention and global statistics, respectively. The resulting weights $W \in \mathbb{R}^{n\times d}$ encode global consistency and reflect the relative reliability of correspondences.

The modulation weights W scale the shallow features F_L point-wise and are combined with deep features F_H via a residual connection to produce the final output:

$$F_{out} = W \odot F_L + F_H \tag{20}$$

Here, $\odot$ denotes element-wise multiplication. DGM refines the stage's graph-context output, providing adaptive inputs for subsequent graph embeddings and enabling effective feature fusion and robust propagation.

3.5 Loss Function

As in benchmarks [Zhao *et al.*, 2021], we adopt a hybrid loss function to optimize BAG-Net, which includes a binary cross-entropy loss for correspondence classification and a loss function L_{ess} for regressing the essential matrix:

$$L = L_{cls} + \beta L_{ess}(\mathbf{E}, \widehat{\mathbf{E}}), \tag{21}$$

where $\mathbf{E}$ and $\hat{\mathbf{E}}$ are the ground truth essential matrix and the predicted essential matrix, respectively. β is a weight parameter, which is used to balance these two losses. L_{cls} is formulated as:

$$\begin{aligned} L_{cls} = &\sum_{m=1}^{\text{M}} \ell_{bce}(\sigma(\tau_m \odot \mathbf{p}_m), \mathbf{G}) \\ &+ \ell_{bce}\left(\sigma\left(\hat{\tau} \odot \hat{\mathbf{y}}\right), \mathbf{G}\right), \end{aligned} \tag{22}$$

Method	YFCC100M			SUN3D		
	P (%)	R (%)	F1 (%)	P (%)	R (%)	F1 (%)
RANSAC	41.83	57.08	48.28	44.11	46.42	45.24
LFGC	53.12	85.51	65.53	47.24	83.45	60.32
OANet++	55.65	85.80	67.51	46.54	83.43	59.74
CLNet	74.89	76.79	75.83	59.97	68.35	63.89
MS^{2}DGNet	59.11	88.40	70.85	46.95	84.55	60.37
U-Match	60.30	90.63	72.42	47.59	85.59	61.17
NCMNet	77.26	78.57	77.91	61.00	69.05	64.78
DeMatch	58.74	91.04	68.91	48.27	85.24	58.08
BCLNet	77.90	80.07	78.73	61.14	68.33	63.92
DeMo	61.41	**91.28**	71.22	47.87	**85.74**	57.75
BAG-Net(ours)	**80.32**	81.80	**81.05**	**62.68**	69.49	**65.91**

Table 1: Quantitative comparative results of outlier removal on the YFCC100M and SUN3D datasets. The best performance is highlighted in bold.

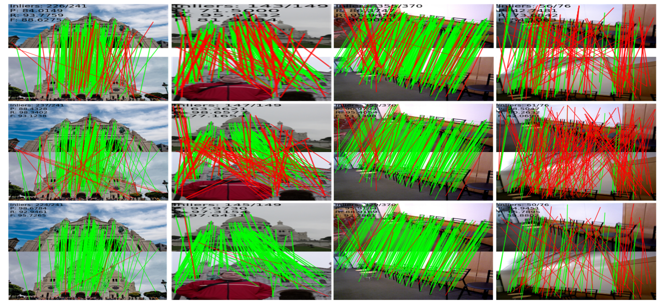

Figure 4: Partial typical visualization results of outlier removal on YFCC100M and SUN3D datasets. From top to bottom: results of CLNet, DeMatch, and BAG-Net. The green lines and red lines represent inliers and outliers, respectively.

where ℓ_{bce} is a binary cross-entropy loss; $\mathbf{G}$ represents weakly supervised ground truth under the epipolar error threshold of 10^{-4}. $\mathbf{p}_m$ denotes the logit value of the m-th pruning module. $\hat{\mathbf{y}}$ is the epipolar distance estimated by our BAG-Net; $\hat{\tau}$ is an adaptive temperature to alleviate the effect of label ambiguity; σ indicates the sigmoid function; $\odot$ represents the Hadamard product; M is the number of pruning blocks. The essential matrix loss is a geometric loss between $\mathbf{E}$ and $\hat{\mathbf{E}}$, and $L_{ess}(\mathbf{E}, \hat{\mathbf{E}})$ is formulated as:

$$L_{ess}(\mathbf{E}, \hat{\mathbf{E}}) = \frac{\left(\mathbf{p_2^T}\hat{\mathbf{E}}\mathbf{p_1}\right)^2}{\|\mathbf{E}\mathbf{p_1}\|_{[1]}^2 + \|\mathbf{E}\mathbf{p_1}\|_{[2]}^2 + \left\|\mathbf{E^T}\mathbf{p_2}\right\|_{[1]}^2 + \left\|\mathbf{E^T}\mathbf{p_2}\right\|_{[2]}^2}, \tag{23}$$

where $\mathbf{p_1}$ and $\mathbf{p_2}$ are virtual correspondences generated by the ground truth $\mathbf{E}$ matrix. $\mathbf{t}_{[i]}$ is the i-th element of vector $\mathbf{t}$.

4 Experiments

4.1 Implementation Details

The number of initial correspondences N for BAG-Net is set to 2000, with a network dimension of 128. In the Pre-embedding Module, the number of neighbors k for the forward neighborhood graph is set to 8, and the number of clusters for the cluster-level graph embedding is set to 64. In the two pruning modules, the first module uses $k = 18$ neighbors and 256 clusters, while the second module uses $k = 12$ neighbors and 128 clusters.

For the iterative pruning strategy, 50% of the correspondences are sampled as candidates in each pruning module, i.e., $N_2 = \frac{1}{2}N_1 = \frac{1}{4}N$. BAG-Net follows the training strategies of prior benchmarks [Zhao *et al.*, 2021] and CGR-Net [Yang *et al.*, 2024], and is trained for $500k$ iterations on Ubuntu 18.04 with an NVIDIA RTX 3090.

4.2 Datasets and Evaluation Metrics

In this section, we first report the experimental performance of our method on the YFCC100M dataset (outdoor scenes) and the SUN3D dataset (indoor scenes), focusing on camera pose estimation and outlier rejection. Finally, we present ablation studies on the YFCC100M dataset to analyze the contributions of the individual components of our method.

4.3 Outlier Rejection

The outlier rejection task aims to identify and remove incorrect correspondences between image pairs, thereby providing reliable geometric constraints for subsequent pose estimation. As shown in Table 1, we compare BAG-Net with traditional methods such as RANSAC, as well as representative learning-based approaches proposed in recent years, including CLNet [Zhao *et al.*, 2021], NCMNet [Liu and Yang, 2023], and BCLNet [Miao *et al.*, 2024], together with recent matching and pruning frameworks such as U-Match [Li *et al.*, 2023], DeMatch [Zhang *et al.*, 2024], and DeMo [Lu *et al.*, 2025].

Experimental results show that BAG-Net achieves the best precision and F1-score on both YFCC100M and SUN3D, while maintaining competitive recall. On YFCC100M, BAG-Net obtains 80.32% precision and 81.05% F1-score, outperforming all compared methods in these two metrics. On SUN3D, it also achieves the best precision and F1-score, demonstrating its effectiveness in suppressing outliers while preserving reliable correspondences.

Our method employs an iterative pruning strategy that progressively reduces the correspondence set across stages to suppress noise and enforce geometric consistency. Since only about 50% of high-confidence correspondences are retained at each stage, some ambiguous true inliers may be removed early, which can limit recall. Nevertheless, the method achieves the best precision and F1-score, indicating a favorable balance between reliability and completeness. As shown in Fig. 4, BAG-Net also produces fewer incorrect matches and better overall matching in both indoor and outdoor scenes.

4.4 Camera Pose Estimation

We evaluate the camera pose estimation performance of BAG-Net on the YFCC100M and SUN3D datasets using mAP5°(%) and mAP20°(%) as metrics. As reported in Table 2, BAG-Net achieves the best performance on YFCC100M under both known and unknown scene settings with the SIFT descriptors. In particular, it obtains mAP5°(%) scores of 66.10% and 78.88%, outperforming the second-best methods by 12.89% and 11.03%, respectively. BAG-Net also achieves the best performance on SUN3D under both settings. To further assess robustness, we evaluate BAG-Net on YFCC100M using SuperPoint correspondences. As shown in Table 3, our method again

Dataset	YFCC100M				SUN3D			
Method	Known		Unknown		Known		Unknown	
	5°(%)	20°(%)	5°(%)	20°(%)	5°(%)	20°(%)	5°(%)	20°(%)
RANSAC	5.81	16.88	9.07	22.92	4.73	16.88	3.26	13.48
LFGC	13.81	35.20	23.95	52.44	13.78	38.68	11.40	36.00
OANet++	32.57	56.89	38.95	66.85	20.86	48.77	16.18	42.22
CLNet	39.00	62.48	54.05	75.76	20.62	45.86	16.95	41.45
MS^2DGNet	38.36	76.04	49.13	76.04	22.20	50.00	17.84	43.28
U-Match	46.78	68.75	60.22	80.26	24.99	51.87	21.38	47.12
NCMNet	52.39	72.54	63.52	82.54	26.22	52.47	20.46	45.66
DeMatch	47.53	69.08	59.98	79.97	28.50	55.61	23.51	49.84
BCLNet	53.21	73.48	67.85	84.57	24.32	51.24	20.06	45.83
DeMo	49.78	71.33	63.45	82.91	30.16	57.33	24.00	50.44
BAG-Net(ours)	**66.10**	**82.40**	**78.88**	**90.21**	**31.85**	**58.15**	**25.64**	**51.59**

Table 2: Performance comparison of camera pose estimation on YFCC100M and SUN3D datasets in both known and unknown scenes with SIFT descriptors. The mAP5°/mAP20°(%) is reported. The best performance is highlighted in bold.

Method	Known		Unknown	
	5°(%)	20°(%)	5°(%)	20°(%)
RANSAC	12.85	31.22	17.47	38.83
LFGC	12.18	34.75	24.25	52.70
OANet++	29.52	53.76	35.27	66.81
CLNet	27.56	50.82	39.19	67.37
MS^2DGNet	31.15	55.16	39.19	70.36
U-Match	35.01	56.80	44.29	70.90
NCMNet	38.92	61.28	48.20	74.71
DeMatch	39.73	60.97	48.55	74.36
BCLNet	40.99	63.51	50.30	76.56
BAG-Net(ours)	**51.96**	**71.54**	**61.49**	**83.13**

Table 3: Performance comparison of camera pose estimation on YFCC100M dataset in both known and unknown scenes with SuperPoint descriptors. The mAP5°/mAP20°(%) is reported. The best performance is highlighted in bold.

outperforms all baselines, demonstrating strong robustness to different feature extractors.

4.5 Ablation Experiments

To validate the contribution of each component, we conduct ablation studies on YFCC100M under unknown scene settings. As summarized in Table 4, we progressively add modules to the baseline and report mAP5°(%) and mAP20°(%).

Baseline

We first evaluate the baseline with an expansion-contraction graph network. Multi-scale graph convolution across receptive field scales progressively imposes geometric constraints from local to global. This ordering better captures structural relations than sequential stacking, achieving 69.85% and 85.94% in mAP5°(%) and mAP20°(%) under unknown scenes. Structural information across stages is not explicitly reinforced, leaving room for improvement.

Baseline + DGM

Adding DGM improves the performance to 71.80% and 86.45%. This gain shows that deep-guided modulation effectively uses high-level structural cues to adaptively refine shallow features.

Baseline + BiDGC

Introducing BiDGC further boosts performance to 72.75% and 87.10%. This confirms that modeling bidirectional dependencies enhances robustness by better capturing mutual support among correspondences.

Baseline	SSE	DGM	BiDGC	mAP@5°	mAP@20°
✓				69.85	85.94
✓		✓		71.80	86.45
✓			✓	72.75	87.10
✓	✓			74.92	88.72
✓	✓	✓		76.48	89.04
✓	✓	✓	✓	**78.88**	**90.21**

Table 4: Ablation study results.

Model	mAP@5°	mAP@20°
Unidirectional (forward-only)	76.48	89.04
Bidirectional w/o. gate	77.53	89.68
Bidirectional w/o. score	77.89	89.71
Full BiDGC	**78.88**	**90.21**

Table 5: Ablation Study of BiDGC Components.

Baseline + SSE

Incorporating SSE while keeping the expansion-contraction framework unchanged raises performance to 74.92% and 88.72%. This demonstrates that stage-wise structural enhancement stabilizes information propagation across receptive-field stages.

Baseline + SSE + DGM

Combining SSE and DGM further improves the results to 76.48% and 89.04%, indicating strong complementarity between structural reinforcement and feature modulation.

Full Model

Finally, integrating BiDGC, SSE, and DGM achieves the best performance of 78.88% and 90.21%. These results demonstrate that jointly modeling bidirectional aggregation, stage-wise structural consistency, and deep-guided modulation significantly improves two-view correspondence pruning.

BiDGC Internal Ablation

As shown in Table 5, replacing BiDGC with unidirectional forward aggregation reduces mAP5°(%) and mAP20°(%) by 2.40% and 1.17%, confirming the necessity of backward aggregation. Removing the gating or edge scoring mechanisms also decreases performance, demonstrating their contribution to robust bidirectional feature fusion.

5 Conclusion

In this paper, we propose BAG-Net, a bidirectional receptive-field graph network for two-view correspondence pruning that shifts the pruning perspective from whether a correspondence is supported to the role it plays in global graph propagation. BiDGC jointly models forward consistency aggregation and reverse structural dependency, effectively distinguishing locally clustered outliers from globally critical inliers. Experiments on YFCC100M and SUN3D demonstrate that BAG-Net significantly outperforms state-of-the-art methods. In future work, we will explore extending the bidirectional aggregation paradigm to end-to-end networks and other geometric vision tasks.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62571127, in part by the China White Tea Research Institute's Open Research Project under Grant BCY2022K04, and in part by the Transportation Science and Technology Program Project of Fujian Provincial Department of Transportation under Grant YB202419.

References

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

[Barath *et al.*, 2020] Daniel Barath, Jana Noskova, Maksym Ivashechkin, and Jiri Matas. Magsac++, a fast, reliable and accurate robust estimator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1304–1312, 2020.

[Chen *et al.*, 2019] Hongchen Chen, Xie Zhang, Shaoyi Du, Zongze Wu, and Nanning Zheng. A correntropy-based affine iterative closest point algorithm for robust point set registration. *IEEE/CAA Journal of Automatica Sinica*, 6:981–991, 2019.

[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/CVF Conference on Computer Vision and Pattern Recognition*, pages 8973–8982, 2022.

[Dai *et al.*, 2024] Luanyuan Dai, Xiaoyu Du, Hanwang Zhang, and Jinhui Tang. Mgnet: Learning correspondences via multiple graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3945–3953, 2024.

[DeTone *et al.*, 2018] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 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*, 24:381–395, 1981.

[Guo *et al.*, 2024] Junwen Guo, Guobao Xiao, Shiping Wang, and Jun Yu. Graph context transformation learning for progressive correspondence pruning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1968–1975, 2024.

[Hu *et al.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[Huang *et al.*, 2020] Weijie Huang, Guoshan Zhang, and Xiaowei Han. Dense mapping from an accurate tracking slam. *IEEE/CAA Journal of Automatica Sinica*, 7:1565–1574, 2020.

[Kaur *et al.*, 2021] Harpreet Kaur, Deepika Koundal, and Virender Kadyan. Image fusion techniques: A survey. *Archives of Computational Methods in Engineering*, 28:4425–4447, 2021.

[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.

[Lin *et al.*, 2026] Shuyuan Lin, Hailiang Liao, Qiang Qi, Junjie Huang, Taotao Lai, and Jian Weng. Sc-net: Robust correspondence learning via spatial and cross-channel context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 6979–6987, 2026.

[Liu and Yang, 2023] Xin Liu and Jufeng Yang. Progressive neighbor consistency mining for correspondence pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9527–9537, 2023.

[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.

[Lowe, 2004] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[Lu *et al.*, 2025] Yifan Lu, Jiajun Le, Zizhuo Li, Yixuan Yuan, and Jiayi Ma. Deep motion field consensus with learnable kernels for two-view correspondence learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5829–5837, 2025.

[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.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[Sattler *et al.*, 2012] Torsten Sattler, Tobias Weyand, Bastian Leibe, and Leif Kobbelt. Image retrieval for image-based localization revisited. In *British Machine Vision Conference*, volume 1, page 4, 2012.

[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/CVF Conference on Computer Vision and Pattern Recognition*, pages 11286–11295, 2020.

[Wang *et al.*, 2017] Feng Wang, Feng Xu, and Ya-Qiu Jin. Three-dimensional reconstruction from a multiview sequence of sparse isar imaging of a space target. *IEEE Transactions on Geoscience and Remote Sensing*, 56:611–620, 2017.

[Wu *et al.*, 2026] Yubin Wu, Xiaojie Li, Hao Chen, Changcai Yang, Lifang Wei, and Riqing Chen. Matchmamba: Correspondence pruning via selective state space model. *IEEE Transactions on Circuits and Systems for Video Technology*, 36:161–174, 2026.

[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*, 34:12450–12465, 2024.

[Yang *et al.*, 2026] Changcai Yang, Fengyuan Zhuang, Xiaojie Li, Lifang Wei, Jiayi Ma, and Riqing Chen. Ddfnet: Dual-neighborhoods dynamic fusion network for image feature matching. *IEEE Transactions on Multimedia*, pages 1–14, 2026.

[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/CVF Conference on Computer Vision and Pattern Recognition*, pages 2666–2674, 2018.

[Zhang and Ma, 2023] Shihua Zhang and Jiayi Ma. Convmatch: Rethinking network design for two-view correspondence learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:2920–2935, 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/CVF 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 20278–20287, 2024.

[Zhao *et al.*, 2019] Chen Zhao, Zhiguo Cao, Chi Li, Xin Li, and Jiaqi Yang. Nm-net: Mining reliable neighbors for robust feature correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 215–224, 2019.

[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/CVF International Conference on Computer Vision*, pages 6464–6473, 2021.