BCLNet: Bilateral Consensus Learning for Two-View Correspondence Pruning

Xiangyang Miao[1, 2], **Guobao Xiao**[1]*, **Shiping Wang**[2], **Jun Yu**[3]

[1]School of Electronics and Information Engineering, Tongji University,
[2]College of Computer and Data Science, Fuzhou University,
[3]School of Computer Science and Technology, Hangzhou Dianzi University
mxyttkx@163.com, x-gb@163.com, shipingwangphd@163.com, yujun@hdu.edu.cn

Abstract

Correspondence pruning aims to establish reliable correspondences between two related images and recover relative camera motion. Existing approaches often employ a progressive strategy to handle the local and global contexts, with a prominent emphasis on transitioning from local to global, resulting in the neglect of interactions between different contexts. To tackle this issue, we propose a parallel context learning strategy that involves acquiring bilateral consensus for the two-view correspondence pruning task. In our approach, we design a distinctive self-attention block to capture global context and parallel process it with the established local context learning module, which enables us to simultaneously capture both local and global consensuses. By combining these local and global consensuses, we derive the required bilateral consensus. We also design a recalibration block, reducing the influence of erroneous consensus information and enhancing the robustness of the model. The culmination of our efforts is the Bilateral Consensus Learning Network (BCLNet), which efficiently estimates camera pose and identifies inliers (true correspondences). Extensive experiments results demonstrate that our network not only surpasses state-of-the-art methods on benchmark datasets but also showcases robust generalization abilities across various feature extraction techniques. Noteworthily, BCLNet obtains significant improvement gains over the second best method on unknown outdoor dataset, and obviously accelerates model training speed.

Introduction

Accurate two-view correspondences play a pivotal role in various computer vision applications, including stereo matching (Yao et al. 2021), simultaneous localization and mapping (SLAM) (Kazerouni et al. 2022), and structure from motion (Xiao et al. 2021), *etc*. Standard sparse matching pipeline relies on off-the-shelf methods (SIFT (Lowe 2004) or Superpoint (DeTone, Malisiewicz, and Rabinovich 2018)) to establish initial correspondences. However, these initial correspondences often harbor a considerable number of outliers (false correspondences) due to the intricate image variations (Jin et al. 2021; Ma et al. 2021), such as illumination changes, repeated textures, image blurs and occlusions. Therefore, to alleviate the repercussions on downstream tasks, the incorporation of a correspondence pruning algorithm emerges as an indispensable stride.

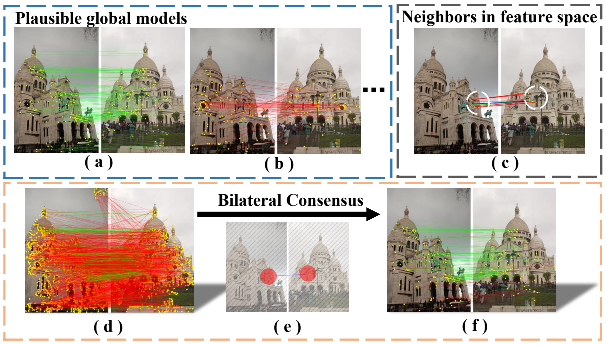


Figure 1: Bilateral consensus acquisition process. Both local and global contexts are inevitably affected by outliers, neglecting the interaction between them tends to exacerbate the propagation of erroneous information. There may be multiple models in the network that satisfy global constraints (a) and (b). Neighbors based on k-nearest neighbor search also contain many outliers (c). Given a set of putative correspondences (d), we adopt existing blocks and the designed BCMA block (e) to extract local and global consensuses, respectively. Subsequently, we facilitate their interaction to achieve bilateral consensus, which ultimately generates the network's prediction (f). The red lines represent outliers, green lines represent inliers, and blue represents selected correspondences.

Existing correspondence pruning algorithms can currently be categorized into two classes: traditional and learning-based algorithms. Among traditional algorithms, RANSAC (Fischler and Bolles 1981) and its diverse adaptations (Chum and Matas 2005; Barath, Matas, and Noskova 2019; Torr and Zisserman 2000; Torr and Bishop 2002) are the most popular ones, which primarily rely on a generate-and-verify strategy. Although they demonstrate promising results on specific tasks, their theoretical execution time tends to experience exponential growth as the proportion of outliers increases (Zheng et al. 2022).

Fortunately, the advancement of deep learning offers a novel solution for the correspondence pruning task. Following the pioneering work LFGC (Yi et al. 2018), learning-

*Corresponding author

based works mostly partition correspondence pruning into a correspondence classification task and a camera pose estimation task. Then, CLNet (Zhao et al. 2021) initially introduces a progressive local-to-global learning strategy and it has gradually become the default standard for subsequent researches. The most recent work, ConvMatch (Zhang and Ma 2023), firstly adopts convolutions to solve this problem by mapping correspondences onto grids, but it also abides by the progressive local-to-global strategy. It's worth mentioning that while this progressive strategy yields favorable results, a crucial aspect has frequently been neglected–the interaction between local and global contexts. As depicted in Figure 1 (a)-(c), challenging scenarios reveal that both local and global contexts are susceptible to significant outlier contamination. Neglecting the precise information shared between these contexts could potentially lead to an amplification of erroneous information, will this have a negative impact on our task?

To address this concern, we introduce a novel parallel strategy that concurrently models both local and global contextual information while capturing their interaction. Specifically, we employ the existing Order-Aware (OA) block as the module for capturing local consensus and design a novel Bilateral Consensus Mining Attention (BCMA) block that operates in parallel with it, serving as the module for capturing global consensus. As demonstrated in Figure 1 (e), for each correspondence, the BCMA block not only establishes global dependencies but also highlights local information, akin to the matching process performed by the human central and peripheral vision. This embedding of local context is achieved by employing k-nearest neighbor (KNN) search separately in the query, key and value feature spaces to seek consistent neighbors. Subsequently, we concatenate the learned local consensus by OA block and global consensus by BCMA block in the channel dimension and facilitate their interaction through several additional ResNet blocks, resulting in a more dependable bilateral consensus. As shown in Figure 1 (d)-(f), through the acquisition of bilateral consensus, our network demonstrates remarkable performance in the correspondence pruning task.

In addition, we discover that inliers have higher consistency in the bilateral consensus feature space, a simple interaction operation is inadequate in the face of complex matching scenarios. To enhance the robustness of our network, we present the Bilateral Consensus Recalibrate (BCR) block to revalidate bilateral consensus at both local and global scales. To elaborate further, BCR block entails direct compression of bilateral consensus feature maps to generate a global scalar, along with the utilization of a KNN search search again to acquire a local vector. The fusion of these two tensors facilitates a soft selection process on the bilateral consensus, thereby amplifying the feature map's representational capacity. Finally, with all the proposed blocks, we formulate a Bilateral Consensus Learning Network tailored for the task of two-view correspondence pruning.

We summarize our contributions as follows:

- We propose a novel consensus learning strategy for the two-view correspondence pruning task. In contrary to previous progressive learning strategy, we concurrently learn local and global consensuses in parallel and obtain bilateral consensus by establishing interdependencies between them. To our knowledge, it is the first time leveraging bilateral consensus to handle the task of two-view correspondence pruning.
- We propose a simple yet effective BCMA block as the global consensus learning module in bilateral consensus and a BCR block to rectify bilateral consensus. Through the process of learning and recalibration, our network is equipped to handle intricate matching scenarios.
- We develop an effective BCLNet for correspondence pruning task. Extensive experiments demonstrate the effectiveness of our proposed BCLNet on the correspondence classification task and the camera pose estimation task. Noteworthily, BCLNet obtains 3.98% mAP5° gains over the second best method on unknown outdoor dataset, and obviously accelerates model training speed.

Related Work

Correspondence Pruning

Over the years, correspondence pruning method has undergone extensive development and gradually divided into two branches, i.e., traditional hand-craft methods and learning-based methods. In traditional methods, hypothesize-and-verify strategy is a prevalent paradigm, *e.g.*, RANSAC (Fischler and Bolles 1981), PROSAC (Chum and Matas 2005), MAGSAC (Barath, Matas, and Noskova 2019), LO-RANSAC (Chum, Matas, and Kittler 2003), NG-RANSAC (Brachmann and Rother 2019), *etc.* To be specific, RANSAC (Fischler and Bolles 1981) repeatedly selecting a random subset of the data to fit models until the number of inliers meets a specified threshold. PROSAC (Chum and Matas 2005) exploits LMI constraints to accelerate the model optimization process based on RANSAC, while MAGSAC (Barath, Matas, and Noskova 2019) proposes σ-consensus to eliminate the need of user-defined threshold. But these methods are mostly task-specific, and the computation time adds exponentially with outliers increases, which is not enough to deal with the complex reality scenes.

Fortunately, the advent of deep learning makes it possible. As a pioneering work, LFGC (Yi et al. 2018) first proposes a PointNet-like architecture that treats correspondence pruning as a collection of correspondence classification and camera pose regression tasks. The impressive efficacy of it makes the adoption of permutation invariant pattern to handle correspondences a de facto standard. Subsequent works primarily focus on designing additional modules to capture richer context. Notably, OANet (Zhang et al. 2019) implicitly captures local context by mapping correspondences to clusters through differential pooling operation. CLNet (Zhao et al. 2021) initially proposes a local-to-global learning strategy and incorporates pruning operations. However, they only consider the improvement of a certain context or separately utilize detached contexts, ignoring the potential interactions between them, which limits the improvement of network performance.

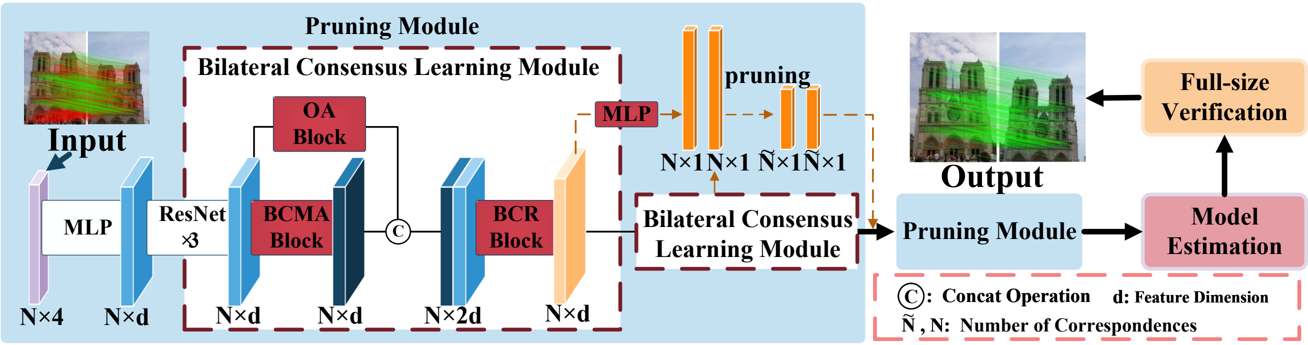


Figure 2: Architecture of BCLNet for correspondence pruning. We take putative correspondences $C \in R^{N\times4}$ as inputs and finally output the inlier probabilities $\omega \in R^{N\times1}$. The entire process involves two pruning modules, systematically refining correspondences into more reliable subsets. Each pruning module consists of our proposed Bilateral Consensus Mining Atttention block, Bilateral Consensus Recalibrate block and existing Order-Aware block.

Consensus in Correspondences

The inliers between two images satisfy epipolar geometry constraint or homography transformation, while outliers are randomly distributed. This correspondence consensus has been studied predates the popularity of learning-based methods, *e.g.*, GMS (Bian et al. 2017) represents consensus as the statistical number of supported correspondences in a small area. LPM (Ma et al. 2019) considers the topological structure within the neighborhood of keypoints is always consistent. Inspired by these methods, learning-based methods also introduce the concept of consensus. CLNet (Zhao et al. 2021), for instance, firstly differentiates between global and local consensus and proposes a local-to-global learning strategy. As an improved version of CLNet (Zhao et al. 2021), NCMNet (Liu and Yang 2023) achieves significant improvements in exploring neighbor consistency in different feature spaces. ConvMatch (Zhang and Ma 2023) maps correspondences onto grids and learns motion consistency using convolutions. In this paper, we also leverage the concept of consensus, but emphasize on interacting with different consensus information.

Method

Problem Formulation

Given a pair of images (I, I'), we first extract keypoints by existing feature detectors (Lowe 2004; DeTone, Malisiewicz, and Rabinovich 2018). Then a nearest neighbor matching algorithm is used to obtain the initial correspondence set C:

$$C = \{c_1, c_2, c_3, \ldots, c_n\} \in R^{N\times4}, c_i = (x_i, y_i, x_i', y_i') \quad (1)$$

where c_i is the i-th correspondence. (x_i, y_i) and (x_i', y_i') represent the coordinates of keypoints that are normalized by camera intrinsics in the image I and I', respectively.

It is worth noting that the initial correspondence set C is dominated by outliers. With C as input, the goal of correspondence pruning is to predict the probability of each correspondence as an inlier and recover the relative camera motion. Specifically, reference to literature (Zhao et al. 2021), BCLNet contains two pruning modules to progressive pruning C into a more reliable candidate subsets $\hat{C}$. As show in Figure 2, the complete architecture can be expressed as:

$$(\omega_1, C_1) = f_\theta'(C), \quad (\omega_2, C_2) = f_\varphi'(C_1), \quad (2)$$

$$\hat{E} = g(C_2, \omega_2), \quad \omega = h(\hat{E}, C) \quad (3)$$

where $f_\theta'(\cdot), f_\varphi'(\cdot)$ represent different network weights of the two pruning modules; $g(\cdot)$ denotes the weight eight-point algorithm; $h(\cdot)$ is full-size verification operation. Particularly, here $C_1 \in R^{N_1\times4}$ and $C_2 \in R^{N_2\times4}$ are pruned correspondence sets, in which $N > N_1 > N_2$; $\omega_1 \in R^{N\times2}$ and $\omega_2 \in R^{N_1\times2}$ are the weights predicted by bilateral consensus learning modules; $\omega \in R^{N\times1}$ is the inlier probabilities of the final prediction.

Bilateral Consensus Mining Attention Block

To achieve bilateral consensus, it is imperative to model local and global consensuses primarily. The existing Order-Aware block has demonstrated commendable accuracy in extracting local consensus. Consequently, an additional module must be fashioned to capture global consensus information. The potency of the self-attention mechanism in establishing global dependencies instills us with optimism. Therefore, we design a distinctive and lightweight Bilateral Consensus Mining Attention (BCMA) block. What's more, acknowledging the significance of local information in the correspondence pruning task, BCMA block embeds local context before acquiring global consensus.

Local Context Embedding In contrast to conventional self-attention (Vaswani et al. 2017), BCMA computes similarity score across channels, rather than in the spatial dimension, which gives it linear complexity while encodes the global context implicitly. We also introduce annular convolution (Zhao et al. 2021) to highlight local context before obtaining global attention map. From input tensor $X \in R^{N\times d}$, where d represents channel dimensions, we first generate query Q, key K and value V via a layer normalization followed by MLP layers. Then, three neighbor embedding graphs are constructed using KNN search for each correspondence on value, query and key projections:

$$\mathcal{G}_i^Q = \left\{\mathcal{V}_i^Q, \mathcal{E}_i^Q\right\}, \mathcal{G}_i^K = \left\{\mathcal{V}_i^K, \mathcal{E}_i^K\right\}, \mathcal{G}_i^V = \left\{\mathcal{V}_i^V, \mathcal{E}_i^V\right\} \tag{4}$$

Towards graph $\mathcal{G}_i^Q$, nodes $\mathcal{V}_i^Q = \left\{c_{i,1}^Q, c_{i,2}^Q, \cdots, c_{i,k}^Q\right\}$, $c_{i,m}^Q$ denotes the m-th neighbor of c_i^Q according to Euclidean distance between c_i^Q and c_m^Q in the query feature space. Edges $\mathcal{E}_i^Q = \left\{e_{i,1}^Q, e_{i,2}^Q, \cdots, e_{i,k}^Q\right\}$ is formulated as (Wang et al. 2019):

$$e_{i,j}^Q = \left[z_i^Q, z_i^Q - z_j^Q\right], \quad 1 \leq j \leq k \tag{5}$$

where $[\cdot]$ denotes the concatenation operation, $z_i^Q - z_j^Q$ represents residual feature between c_i^Q and c_j^Q.

Whereafter, in order to effectively tapping local consensus through graph $\mathcal{G}^Q \in R^{N\times 2d}$, we categorize the neighbors into g groups based on their affinity to the anchor node c_i^Q, where each group contains k/g nodes. The final feature of c_i^Q is then obtained by a convolution layer with $1 \times k/g$ kernels followed by a convolution layer with $1 \times g$ kernels. By aggregating all correspondences, we acquire a new feature mapping Q_{local}, that encapsulates local information. Let us do the same operation for $\mathcal{G}^K$ and $\mathcal{G}^V$, so that correspondence in each feature space aggregates different local contexts and we get K_{local}, V_{local}.

Attention Mechanism We alter the computation order of Q, K, V in conventional self-attention (Vaswani et al. 2017), so that the similarity map A generated by the dot product of Q and K has only the size of $R^{d\times d}$, instead $R^{N\times N}$. Thus, its computational complexity is linear for the number of correspondences. The attention process is defined as:

$$\hat{X} = W_P \cdot V_{local} \cdot Softmax\left(Q_{local} \cdot K_{local}^T \cdot \alpha\right) + X \tag{6}$$

where X and $\hat{X}$ are the input and output tensor; W_P represents a MLP layer; α is a learnable scaling parameter to adjust the magnitude of the similarity map A.

After completing the above operations, the resulting feature $\hat{X}$ undergoes further refinement to achieve the desired global consensus, which involves an additional layer normalization and a ResNet block. Finally, we concatenate the obtained global consensus with the local consensus learned by the Order-Aware block in the feature dimension. This combined feature undergoes thorough interaction and recalibration through followed Bilateral Consensus Recalibrate block, resulting in the reliable bilateral consensus.

Bilateral Consensus Recalibrate Block

We find that inliers have higher consistency in the feature space after the convergence of bilateral consensus. However, when facing some challenging scenarios, the feature representation simply learned through a interaction operation is deficient. It remains susceptible to inaccuracies in the features when performing correspondence pruning. Hence, we propose to model the interdependence between channels as SENet (Hu, Shen, and Sun 2018) does to adaptively recalibrate the feature response across channels.

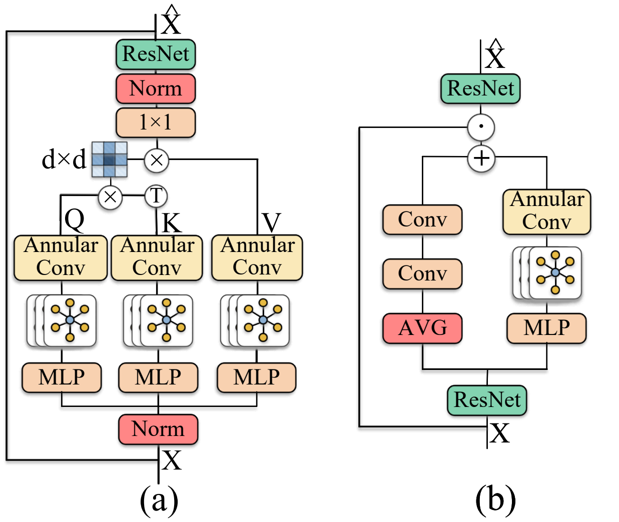


Figure 3: Illustration of the proposed (a) Bilateral Consensus Mining Attention (BCMA) block and (b) Bilateral Consensus Recalibrate (BCR) block.

For each feature map, SENet (Hu, Shen, and Sun 2018) compresses all spatial information into a scalar, which unfortunately neglects the vital local context containing motion consistency that is crucial for our task. To overcome this limitation, literature (Zheng et al. 2022) introduces a multi-scale attention block, but it only addresses this issue by using extra MLP layers to process correspondences individually. Inspired by this concept, we seize the opportunity to devise a Bilateral Consensus Recalibrate (BCR) block capable of capturing both local and global contexts concurrently.

Specifically, given the input feature $X \in R^{N\times d}$, we first apply a global average pooling operation to obtain the global feature map $X_{avg} \in R^{1\times d}$. Thus, point-wise convolutional layers with a bottleneck structure are employed to get the global feature scalar $X_{global} \in R^{1\times d}$. The operations are formulated as:

$$X_{global} = Conv_1\left(Conv_2\left(AVG\left(X\right)\right)\right) \tag{7}$$

where the weights of $Conv_1$ and $Conv_2$ are $d \times d/r \times 1$ and $d/r \times d \times 1$ respectively; r is the channel reduction ratio; X_{global} is the recalibrated global scalar. For simplicity, we omit the Batch Normalization layer and ReLU layer.

Next, considering the strong consistency of inliers at the bilateral consensus level, we adopt the approach that mentioned in the previous local context embedding section to get the local feature vector X_{local}. At last, we combine the features X_{global} and X_{local} and apply a sigmoid function to obtain the final weight for softly selecting X by an elementwise multiplication operator. The entire process is represented as:

$$X_{out} = Sigmod\left(X_{local} + X_{global}\right) \odot X \tag{8}$$

What's more, to effectively manage the bilateral consensus features post recalibration, we also add three Resnet blocks at the end to adequately handle the features.

Network Framework

The overall architecture of the BCLNet proposed by us is shown in Figure 2. BCLNet primarily consists of two pruning modules, a camera motion estimation module, and a full-size verification module. The camera motion estimation module is responsible for computing essential matrix from the last pruned reliable correspondence subset and subsequently passes it to the full-size verification module for classification of all correspondences.

Principally, each pruning module is composed of a MLP layer, three ResNet blocks and two Bilateral Consensus Learning Modules. We use MLP layer to elevate features to higher dimensional spaces d and ResNet block serving as the basic module for extracting correspondence features, which incorporates several MLP layers and normalization layers. An Order-Aware block is incorporated into the Bilateral Consensus Learning module alongside the proposed BCMA block, allowing simultaneous extraction of both local and global contextual information. Subsequently, these two forms of contextual information are combined to form bilateral consensus, which is then revalidated using the BCR block. Combining the above modules, we get a progressive correspondence pruning network.

Loss Function

Follow (Yi et al. 2018), to balance the correspondence classification task and the camera pose estimation task, we develop a hybrid loss function to optimize the proposed B-CLNet:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{ess} \tag{9}$$

where $\mathcal{L}_{cls}$ represents correspondence classification loss, $\mathcal{L}_{ess}$ denotes essential matrix estimation loss, λ is the weight factor used to balance the two losses.

In $\mathcal{L}_{cls}$, to mitigate the effects of label ambiguity, we add adjustable temperature τ (Zhao et al. 2021). For inliers, τ_i is negatively correlated with the epipolar distance d_i, while for outliers $\tau_i = 1$. The loss $\mathcal{L}_{cls}$ is given as:

$$\mathcal{L}_{cls}(\omega, y) = \sum_{i=1}^{N_k} \mathcal{H}(\tau_i \odot \omega_i, y_i) + \mathcal{H}(\hat{\tau} \odot \hat{\omega}, \hat{y}) \tag{10}$$

where ω_i is the predicted logit vector of the $i-th$ bilateral consensus learning module; $\hat{\omega}$ indicates the output of the last MLP layer; y_i and $\hat{y}$ are ground-truth labels; $\mathcal{H}$ indicates the binary cross entropy loss function; $\odot$ denotes the Hadamard product and N_k represents the num of bilateral consensus learning modules.

$\mathcal{L}_{ess}$ can be write as a geometry loss (Ranftl and Koltun 2018):

$$\mathcal{L}_{ess}(\hat{E}, E) = \frac{(p\prime^T \hat{E} p)^2}{\| Ep \|_{[1]}^2 + \| Ep \|_{[2]}^2 + \| E^T p\prime \|_{[1]}^2 + \| E^T p\prime \|_{[2]}^2} \tag{11}$$

where $\hat{E}$ and E denote the predicted essential matrix and ground-truth essential matrix respectively; $\| \cdot \|_{[i]}$ represents the i-th element of this vector; p and $p\prime$ are the virtual correspondence coordinates generated by using E.

Methods	Known Scene			Unknown Scene		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
RANSAC	47.44	52.64	49.90	43.51	50.68	46.82
PointNet++	49.84	86.41	63.22	46.60	84.17	59.99
LFGC	56.64	86.30	68.39	54.67	84.76	66.47
OANet++	60.03	89.31	71.80	55.78	85.93	67.65
MSA-Net	61.98	90.53	73.58	58.70	87.99	70.42
CLNet	76.04	79.27	77.62	75.05	76.41	75.72
MS^{2}DG-Net	63.17	90.98	74.57	59.11	88.40	70.85
ConvMatch	63.07	**91.55**	74.69	58.77	**89.39**	70.92
NCMNet	**78.49**	81.72	79.69	77.07	78.27	77.41
Ours	**78.49**	82.56	**80.10**	**77.39**	79.77	**78.31**

Table 1: Performance comparisons of our network and other models about the Precision, Recall and F-score on the YFCC100M dataset in the correspondence classification task.

Experiments

We first introduce the datasets and evaluation metrics used in the experiment. Subsequently, by comparing with some state-of-the-art methods, we validate the effectiveness of B-CLNet in both correspondence classification task and camera pose estimation task. At last, we conduct some ablation experiments to discuss the setting of hyperparameters and evaluate the role of each module.

Evaluation Protocols

Dataset Towards an effort to demonstrate the effectiveness of BCLNet, we conduct experiments on both indoor and outdoor scenes. For the outdoor dataset, we utilized Yahoo's YFCC100M (Thomee et al. 2016), a vast collection containing 100 million pieces of multimedia data. As for the indoor setting, we relied on SUN3D (Xiao, Owens, and Torralba 2013), which is an RGBD video dataset encompassing entire rooms. Followed the data division approach outlined in (Zhang et al. 2019), we train all models at the same training setting to ensure an equitable comparison.

Evaluation Metrics Our proposed network addresses two key tasks: correspondence classification and camera pose estimation. For classification task, we employ precision (P), recall (R), and F-score (F) as evaluation metrics to measure the performance of classification task. Precision is the ratio of correctly identified correspondences to preserved correspondences, while Recall is the ratio of correctly identified correspondences to all correct correspondences in the initial set. The F-score combines Precision and Recall for a comprehensive measure, defined as $2 * P * R/(P + R)$. For camera pose estimation task, we use mean average precision (mAP) of angular differences under various error thresholds as evaluation metric. The angular difference is determined by comparing the rotation and translation vectors predicted by our model with the corresponding ground truth values.

Implementation Details

In our experiments, feature detectors(SIFT (Lowe 2004), Superpoint (DeTone, Malisiewicz, and Rabinovich 2018))

Methods	YFCC100M(%)		SUN3D(%)	
	Known	Unknown	Known	Unknown
RANSAC	5.81	9.07	4.52	2.84
PointNet++	10.49	16.48	10.58	8.10
LFGC	17.45	25.95	11.55	9.30
OA-Net++	32.57	38.95	20.86	16.18
MSA-Net	39.53	50.65	18.64	16.86
CLNet	38.27	51.80	19.20	15.83
MS^{2}DG-Net	38.36	49.13	22.20	17.84
ConvMatch	43.48	54.62	**25.36**	**21.71**
NCMNet	49.10	62.10	24.91	19.57
BCLNet(Ours)	**52.62**	**66.08**	24.59	19.96

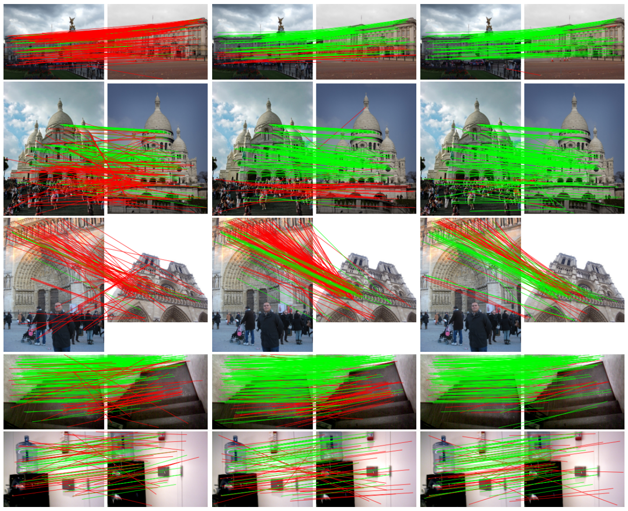
Table 2: Comparative results about our network and other networks when using SIFT as feature extractor on the YFCC100M and SUN3D datasets, with mAP5° (%) is reported.

and nearest neighbor matching method are adopted to establish $N = 2000$ initial correspondences and then our network. We go through two pruning modules with a pruning ratio of 0.5 and end up with a reliable set of correspondences of $N/4$. Throughout both pruning modules, we increase the feature dimension d to 128. For the first pruning module, we use the initial set of correspondences as input and the number of k neighbors is empirically set to 9. For the second pruning module, we set k to 6 and use the weights predicted in the previous module along with the pruned correspondences set as input. Within the BCMA layer and BCR layer of the two pruning modules, we set the number of groups g to 3 and 2, respectively. Additionally, in the Order-Aware block, we set the cluster number to 150 for outdoor scenes and 250 for indoor scenes. All networks are implemented in PyTorch (Paszke et al. 2019) and trained using the Adam optimizer (Kingma and Ba 2014) with an initial learning rate of 10^{-3} and a batch size of 32. The training process consists of 500k iterations. In Equation 9, the weight λ is initialized to 0, and then fixed at 0.5 after the first 20k iterations.

Correspondence Classification

We compare BCLNet with RANSAC (Fischler and Bolles 1981) and recent state-of-the-art learning-based methods for both known and unknown outdoor scenes, which included PointNet++ (Qi et al. 2017), LFGC (Yi et al. 2018), OANet (Zhang et al. 2019), MSANet (Zheng et al. 2022), CLNet (Zhao et al. 2021), MS2DNet (Dai et al. 2022), ConvMatch (Zhang and Ma 2023), and NCMNet (Liu and Yang 2023). Table 1 presents the comparative results in the classification task, measured using Precision, Recall, and F-score metrics. For all correspondences, we consider them as inliers if their epipolar distance is less than a certain threshold (10^{-4}).

From Table 1, we can observe that BCLNet achieves the best results on Precision and F-score metrics across all scenarios. However, it is worth noting that both our method and CLNet (Zhao et al. 2021), NCMNet (Liu and Yang 2023) show a significant decrease on Recall compared to other learning-based networks. This reduction is attributed to the adoption of pruning strategy, leading to the removal of inliers during the pruning process. Although full-size verification restores some inliers, the predicted geometric model still implicitly inherits this attribute. Specifically, NCMNet (Liu and Yang 2023) represents an improved version of CLNet (Zhao et al. 2021), capturing richer local consensus and being the most effective model prior to our work. But our network demonstrates improvements over NCMNet (Liu and Yang 2023) in all performance indicators, proving that learning bilateral consensus effectively suppresses the spread of erroneous context. Furthermore, with comparable parameters, the training time of BCLNet has been reduced by nearly 25 hours compared to NCMNet (Liu and Yang 2023), thanks to our key improvements in the attention mechanism. Additionally, we also visualized some visualization results of RANSAC, CLNet and BCLNet in Figure 4. As can be seen, BCLNet achieves the best results.



(a) RANSAC (b) CLNet (c) BCLNet(Ours)

Figure 4: Visualization results of two-view correspondence pruning on the unknown outdoor scenes and unknown indoor scenes. From left to right are the results of RANSAC, CLNet and BCLNet, respectively. Inliers(green lines) and outliers(red lines) retained by algorithms are exhibited.

Camera Pose Estimation

The camera pose estimation aims to calculate the positional changes (translation and rotation) between two cameras when capturing an image pair. We compare BCLNet with other methods described in the previous section on both indoor and outdoor scenes to verify the generalization of our network. In addition, since the quality of the initial correspondences will inevitably affect the effect of camera pose estimation, we also utilize traditional (SIFT) and learn-based methods (SuperPoint) as feature extractors respectively to prove that the network has strong robustness.

As shown in Table 2, the comparison results for indoor and outdoor scenes are presented. We report mAP5° (%) for both known and unknown scenes. It is evident that BCLNet achieves state-of-the-art performance in nearly all scenar-

	Matcher	Known Scene		Unknown Scene	
		5°	20°	5°	20°
SIFT	RANSAC	5.81	16.88	9.07	22.92
	PointNet++	10.49	31.17	16.48	42.09
	LFGC	13.81	35.20	23.95	52.44
	OA-Net++	32.57	56.89	38.95	66.85
	MSA-Net	39.53	61.75	50.65	77.99
	CLNet	38.27	62.48	51.80	75.76
	MS^{2}DG-Net	38.36	64.04	49.13	76.04
	ConvMatch	43.48	66.14	54.62	77.24
	NCMNet	49.10	70.80	62.10	81.67
	BCLNet(Ours)	**52.62**	**72.89**	**66.08**	**83.38**
SuperPoint	RANSAC	12.85	31.22	17.47	38.83
	PointNet++	11.87	33.35	17.95	49.32
	LFGC	12.18	34.75	24.25	52.70
	OA-Net++	29.52	53.76	35.27	66.81
	MSA-Net	30.63	53.74	38.53	68.56
	CLNet	27.56	50.82	39.19	67.37
	MS^{2}DG-Net	31.15	55.16	39.19	70.36
	ConvMatch	38.34	60.25	**48.80**	74.59
	NCMNet	40.33	62.37	48.60	75.38
	BCLNet(Ours)	**40.56**	**62.71**	48.07	**75.84**

Table 3: Comparative results about our network and other networks when using SIFT and SuperPoint as feature extraction methods on the YFCC100M dataset. mAP5° (%) and mAP20° (%) are reported.

IPS	BCMA	BCR	OA	MAP5°(%)	MAP20°(%)
√				51.80	75.76
√	√			59.98	80.72
√	√	√		61.03	81.18
√	√	√	√	66.08	83.38

Table 4: Ablation studies regarding performance gains of the key components in each pruning module on outdoor dataset. mAP5° (%) and mAP20° (%) are reported. IPS: the iterative pruning strategy. BCMA: the Bilateral Consensus Mining Attention block. BCR: the Bilateral Consensus Recalibrate block. OA: the Order-Aware block.

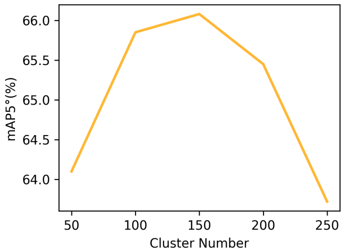


Figure 5: Impact of cluster number on BCLNet performance.

ios, particularly with significant improvements of 3.52% and 3.98% in outdoor known and unknown scenarios compared to the second best method respectively. In indoor scenarios, our performance slightly lags behind that of ConvMatch, albeit with the advantage of our model size being almost one-third smaller.

In addition, Table 3 presents comparison results obtained using learning-based feature extractor and hand-craft feature extractor on the YFCC100M dataset, with mAP5° (%) and mAP20° (%) reported. BCLNet continues to demonstrate superior performance across all scenarios, affirming the robustness of our network to the quality of the initial correspondence set.

Ablation

In this section, we conduct ablation studies on the unknown scene of YFCC100M to verify the role of each key component in our network. We also perform some parameter analysis to balance performance and efficiency.

Main Components The proposed BCLNet comprises three main modules *i.e.* BCMA block, BCR block, and OA block. Utilizing the pruning strategy from the reference (Zhao et al. 2021) as a baseline, we evaluate the performance gains of these key components. The BCMA block is employed to run in parallel with the OA block to separately obtain local and global consensus, while the BCR block is used to re-verify bilateral consensus. As depicted in Table 4, with the integration of each module, the network's performance shows gradual improvement. When all modules are combined, we achieve the best result.

Parameter Analysis In ours experiments, we observe a significant enhancement in the network's performance upon incorporating the Order-Aware block. This is attributed to the importance of local context in the correspondence pruning task. As bilateral consensus seems to weaken the local consensus, therefore, we adjust cluster number in the Order-Aware block to compensate for this deficiency. Intuitively, higher values of the cluster in the Order-Aware block are expected to improve network performance. However, this also bring in increased network parameters and computational burden. To strike a balance between model size and efficiency, we conduct tests to analyze the impact of different cluster values on outdoor dataset, as shown in Figure 5. Ultimately, we set it to 150 on outdoor scenes.

Conclusion

In this paper, we propose an effective Bilateral Consensus Learning Network (BCLNet) to cope with correspondence pruning problem. Considering the complementarity between local and global contexts, we interact these two types of information to get a bilateral consensus for accurately identify inliers. Simultaneously, to handle complex matching scenarios and mitigate the impact of incorrect information, we further design a Bilateral Consensus Recalibrate block that enhances the feature representation capability. Numerous experiments conducted on public benchmarks consistently demonstrate the superiority of our method over the current state-of-the-art technologies.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants 62072223, 62125201 and 62020106007.

References

Barath, D.; Matas, J.; and Noskova, J. 2019. MAGSAC: marginalizing sample consensus. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10197–10205.

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

Brachmann, E.; and Rother, C. 2019. Neural-guided RANSAC: Learning where to sample model hypotheses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4322–4331.

Chum, O.; and Matas, J. 2005. Matching with PROSAC-progressive sample consensus. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, 220–226. IEEE.

Chum, O.; Matas, J.; and Kittler, J. 2003. Locally optimized RANSAC. In *Pattern Recognition: 25th DAGM Symposium, Magdeburg, Germany, September 10-12, 2003. Proceedings 25*, 236–243. Springer.

Dai, L.; Liu, Y.; Ma, J.; Wei, L.; Lai, T.; Yang, C.; and Chen, R. 2022. MS2DG-Net: Progressive correspondence learning via multiple sparse semantics dynamic graph. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8973–8982.

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

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

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141.

Jin, Y.; Mishkin, D.; Mishchuk, A.; Matas, J.; Fua, P.; Yi, K. M.; and Trulls, E. 2021. Image matching across wide baselines: From paper to practice. *International Journal of Computer Vision*, 129(2): 517–547.

Kazerouni, I. A.; Fitzgerald, L.; Dooly, G.; and Toal, D. 2022. A survey of state-of-the-art on visual SLAM. *Expert Systems with Applications*, 205: 117734.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Liu, X.; and Yang, J. 2023. Progressive Neighbor Consistency Mining for Correspondence Pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9527–9537.

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

Ma, J.; Jiang, X.; Fan, A.; Jiang, J.; and Yan, J. 2021. Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision*, 129: 23–79.

Ma, J.; Zhao, J.; Jiang, J.; Zhou, H.; and Guo, X. 2019. Locality preserving matching. *International Journal of Computer Vision*, 127: 512–531.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems*, 30.

Ranftl, R.; and Koltun, V. 2018. Deep fundamental matrix estimation. In *Proceedings of the European Conference on Computer Vision*, 284–299.

Thomee, B.; Shamma, D. A.; Friedland, G.; Elizalde, B.; Ni, K.; Poland, D.; Borth, D.; and Li, L.-J. 2016. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2): 64–73.

Torr, N. S. J., Philip Hilaire; and Bishop, J. M. 2002. Napsac: High noise, high dimensional robust estimation-its in the bag. In *British Machine Vision Conference*, volume 2, 3.

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

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M. M.; and Solomon, J. M. 2019. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5): 1–12.

Xiao, G.; Wang, H.; Ma, J.; and Suter, D. 2021. Segmentation by continuous latent semantic analysis for multi-structure model fitting. *International Journal of Computer Vision*, 129(7): 2034–2056.

Xiao, J.; Owens, A.; and Torralba, A. 2013. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proceedings of the IEEE International Conference on Computer Vision*, 1625–1632.

Yao, C.; Jia, Y.; Di, H.; Li, P.; and Wu, Y. 2021. A decomposition model for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6091–6100.

Yi, K. M.; Trulls, E.; Ono, Y.; Lepetit, V.; Salzmann, M.; and Fua, P. 2018. Learning to find good correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2666–2674.

Zhang, J.; Sun, D.; Luo, Z.; Yao, A.; Zhou, L.; Shen, T.; Chen, Y.; Quan, L.; and Liao, H. 2019. Learning two-view correspondences and geometry using order-aware network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5845–5854.

Zhang, S.; and Ma, J. 2023. ConvMatch: Rethinking Network Design for Two-View Correspondence Learning. In *Proc. AAAI Conf. Artif. Intell*, 1–12.

Zhao, C.; Ge, Y.; Zhu, F.; Zhao, R.; Li, H.; and Salzmann, M. 2021. Progressive correspondence pruning by consensus learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6464–6473.

Zheng, L.; Xiao, G.; Shi, Z.; Wang, S.; and Ma, J. 2022. MSA-Net: Establishing reliable correspondences by multiscale attention network. *IEEE Transactions on Image Processing*, 31: 4598–4608.