2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

PG-Match: A Pose-Guided Generalizable Framework for Semi-Dense Feature Matching

Jiayi Pei[1], Peili Song[1], Chenyang Zhao[1], Lei Sun[1], Jingtai Liu[1†]

***Abstract*— Feature matching is a fundamental technique in visual perception, essential for tasks such as 3D reconstruction, SLAM, and visual localization. Existing detector-free methods often struggle to generalize due to their reliance on depth data, which is not available in many datasets. We propose PG-Match, a detector-free feature matching framework that leverages pose supervision instead of depth-based supervision, thereby improving generalization across diverse environments. We further introduce a Differentiable Outlier Rejection Module (DORM) to enhance global consistency and increase the inlier ratio. For efficiency, a coarse-to-fine matching strategy is employed, where specially designed confidence scores are utilized to guide the sampling process. This ensures efficient convergence and avoids local optima. Experiments on the widely used MegaDepth-1500 dataset show that PG-Match consistently outperforms state-of-the-art approaches, highlighting the effectiveness of its pose-guided design. Additionally, experiments on the depth-free PhotoTourism dataset further evaluate generalization of PG-Match, and its performance is also assessed in a downstream Structure from Motion (SfM) task.**

I. INTRODUCTION

Feature matching is a cornerstone of environmental perception, underpinning key tasks such as 3D reconstruction by structure from motion (SfM) [1]–[4], simultaneous localization and mapping (SLAM) and visual localization. The accuracy of feature matching directly influences the performance of these downstream applications, making it a pivotal component in achieving reliable scene understanding. Detector-based methods [5]–[7] face challenges in scenarios such as low-texture regions and scenes with repetitive structures. Even though detector-free methods [8]–[10] have made significant progress in these issues, achieving robustness in challenging scenarios and ensuring generalizability across diverse datasets remain difficult.

As illustrated in Fig. 1 (a), feature matching methods take an image pair as input and outputs the corresponding matches. Detector-based methods first detect sparse keypoints and then compute descriptors for matching. Recent advances, such as SuperPoint [5] and LightGlue [7], have achieved significant breakthroughs in keypoint detection and descriptor representation, respectively. Nevertheless, they still depend on reliable keypoint detection, which remains difficult in low-texture regions. In contrast, detector-free methods, such as AspanFormer [9] and LoFTR series [8], [10], [11], bypass explicit keypoint detection, directly predicting semi-dense correspondences. This approach allows the network to establish matches even in regions where keypoints are sparse or unreliable.

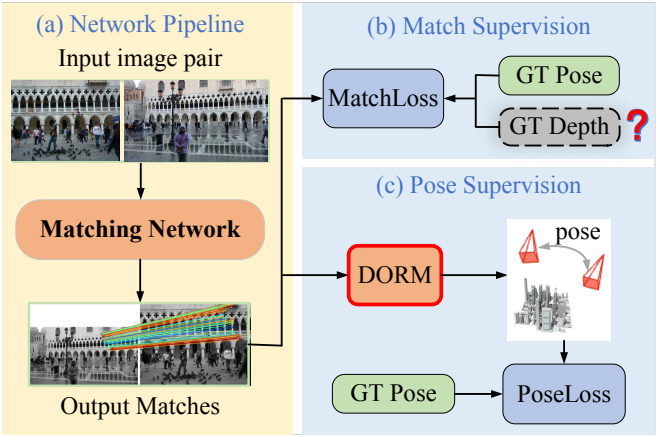


Fig. 1. Overview of the motivation behind our approach. (b) Supervision in existing methods, which requires both ground-truth poses and depth for training. However, datasets with ground truth depth are scarce, limiting the generalization ability of these networks. (c) Supervision in our proposed method. By incorporating the DORM module, we enable end-to-end training with pose supervision, eliminating the dependency on ground truth depth.

However, existing detector-free methods rely on ground truth matches for supervision, which is obtained via reprojections from both accurate pose and depth information, while available datasets with ground truth depth values are limited. Moreover, training with ground truth matches focuses only on local matching errors across correspondences, but does not guarantee global consistency. This limits the inlier ratio and increases the risk of falling into local optima, leading to the performance degradation in downstream applications, such as pose estimation and SfM.

To address these challenges, we propose PG-Match, a detector-free feature matching framework. As illustrated in Fig. 1 (b) and (c), PG-Match leverages ground truth pose as supervision instead of match supervision, which provides global and robust guidance. This also eliminates the reliance on depth information, improving generalization ability. Similar to common approaches, the estimated pose is produced from the predicted correspondences using a RANSAC-based approach. To utilize pose error for backpropagation, we introduce a Differentiable Outlier Rejection Module (DORM), which increases the inlier ratio of the matching results by enhancing global consistency. In addition, a coarse-to-fine matching strategy is used for efficiency, where specially

*This work is supported by the National Natural Science Foundation of China (No. 62573244)

[1]Jiayi Pei, Peili Song, Chenyang Zhao, Lei Sun and Jingtai Liu are with the Institute of Robotics and Automatic Information System, Tianjin Key Laboratory of Intelligent Robotics, and also with TBI center, Nankai University, Tianjin 300350, China.

[†]Corresponding Author. liujt@nankai.edu.cn

designed confidence scores are used to guide the sampling process, preventing convergence to local optima. Pose supervision naturally couples local correspondences with global epipolar geometry, which promotes higher inlier ratios and better pose estimation. Furthermore, by imposing a global geometric constraint, the network is encouraged to refine local matches to be consistent with the overall scene geometry, effectively improving the reliability of local matching.

In summary, the main contributions of this work are as follows.

1) We propose PG-Match, a detector-free feature matching framework that leverages ground truth pose as supervision, which eliminates the reliance on depth information for training and improves generalization ability.
2) We introduce the DORM, which increases the inlier ratio of the matching results by enhancing global consistency. Additionally, we design coarse and fine confidence priors as guidance for the sampling process, which avoids local optima and ensures robust performance across different scenes.
3) We evaluate PG-Match on the MegaDepth-1500 dataset and assess its cross-dataset generalization on PhotoTourism dataset. The results demonstrate that PG-Match outperforms the state-of-the-art methods. Furthermore, we validate its effectiveness in a Structure-from-Motion pipeline, where it improves both the accuracy and completeness of 3D reconstruction.

II. Related Work

Deep neural networks have greatly advanced feature matching in recent years, which can be broadly divided into detector-based and detector-free approaches.

A. Detector-Based Feature Matching

Traditional detector-based methods rely on handcrafted algorithms to detect keypoints, compute descriptors, and establish correspondences. With deep learning, neural networks have significantly improved both keypoint detection [12], [13] and descriptor learning [14]–[17], enhancing robustness and distinctiveness. Some works [18]–[21] jointly optimize detectors and descriptors. A milestone in this line is SuperGlue [6], which introduced Transformers for matching and achieved strong results, albeit with high computational cost when handling many keypoints. Later methods [22], [23] attempted to scale attention more efficiently, often at the expense of accuracy. LightGlue [7] addressed this by adopting adaptive sparse matching, which terminates early for easier pairs, yielding faster performance while remaining competitive with SuperGlue. Nonetheless, reliable keypoint detection remains challenging, especially in texture-poor regions. Our method therefore adopts a detector-free framework, sidestepping these limitations.

B. Detector-Free Feature Matching

Detector-free methods bypass explicit keypoint detection by directly producing semi-dense or dense matches. Early works such as NCNet [24] and Sparse NC-Net [25] employed 4D correlation volumes with varying sparsity, while DRC-Net [26] improved efficiency through a coarse-to-fine design. LoFTR [8] marked a breakthrough by leveraging Transformers for long-range dependencies, though dense global attention introduced heavy computational costs. Efficient LoFTR [10] alleviated this with aggregated attention and adaptive token selection, balancing accuracy and speed. Matchformer [27] and AspanFormer [9] extended attention to multiscale features, introducing flow-guided local windows. QuadTree [11] used hierarchical attention to reduce computation but introduced latency, while TopicFM [28] clustered features into semantic groups for localized attention, which limited long-range modeling. More recently, JamMa [29] adopts a Mamba-based state-space mixer with a joint scan-merge scheme to obtain global context in linear time, yielding a strong accuracy-efficiency trade-off. Dense matching approaches, such as RoMa [30] and DKM [31], further explore fully dense correspondences, achieving high accuracy at the cost of substantially increased computational complexity.

III. Problem Formulation

Nomenclature: Let us define the important notations before introducing our problem.

${}^{A}\mathbf{I},{}^{B}\mathbf{I}$	image pairs (imageA, imageB) to be matched: ${}^{A}\mathbf{I}\in\mathbb{R}^{H_A\times W_A\times 3}$, ${}^{B}\mathbf{I}\in\mathbb{R}^{H_B\times W_B\times 3}$
${}^{A}\mathbf{F}_c,{}^{B}\mathbf{F}_c$	initial coarse feature maps at 1/8 resolution: ${}^{A}\mathbf{F}_c\in\mathbb{R}^{H_A/8\times W_A/8\times D}$, ${}^{B}\mathbf{F}_c\in\mathbb{R}^{H_B/8\times W_B/8\times D}$,
${}^{A}\mathbf{F}_f,{}^{B}\mathbf{F}_f$	initial fine feature maps at 1/s resolution: ${}^{A}\mathbf{F}_f\in\mathbb{R}^{H_A/s\times W_A/s\times D}$, ${}^{B}\mathbf{F}_f\in\mathbb{R}^{H_B/s\times W_B/s\times D}$, where $s\in\{2,4\}$,
${}^{A}\mathbf{F}_c^{tr},{}^{B}\mathbf{F}_c^{tr}$	transformed coarse feature maps: ${}^{A}\mathbf{F}_c^{tr}\in\mathbb{R}^{H_A/8\times W_A/8\times D}$, ${}^{B}\mathbf{F}_c^{tr}\in\mathbb{R}^{H_B/8\times W_B/8\times D}$,
${}^{A}\mathbf{F}_f^{tr},{}^{B}\mathbf{F}_f^{tr}$	transformed fine feature maps: ${}^{A}\mathbf{F}_f^{tr}\in\mathbb{R}^{H_A\times W_A\times D}$, ${}^{B}\mathbf{F}_f^{tr}\in\mathbb{R}^{H_B\times W_B\times D}$,
$M_c,\{m_c\}$	indices of coarse matching point pairs: $M_c=\{m_{c,1},m_{c,2},...,m_{c,n}\}$, where n is number of matches. $m_c=(i_c,j_c)$, where i_c,j_c are feature point indices of ${}^{A}\mathbf{F}_c^{tr}$ and ${}^{B}\mathbf{F}_c^{tr}$, respectively,
P_c	confidences of coarse matching point pairs: $P_c=\{\alpha_{c,k}\mid m_{c,k}\in M_c\}$,
$M_f,\{m_f\}$	indices of fine matching point pairs: $M_f=\{m_{f,1},m_{f,2},...,m_{f,n}\}$. $m_f=(i_f,j_f)$, where i_f,j_f are feature point indices of ${}^{A}\mathbf{F}_f^{tr}$ and ${}^{B}\mathbf{F}_f^{tr}$, respectively, and are equivalent to the feature point indices of ${}^{A}\mathbf{I},{}^{B}\mathbf{I}$,
P_f	confidences of fine matching point pairs: $P_f=\{\alpha_{f,k}\mid m_{f,k}\in M_f\}$.

Our problem is defined as follows.

Definition 1: Given image pairs ${}^{A}\mathbf{I},{}^{B}\mathbf{I}$, generate matching point pairs M_f.

IV. Methods

Fig.2 overviews the proposed framework, which consists of four modules. (1) The RepVGG backbone processes

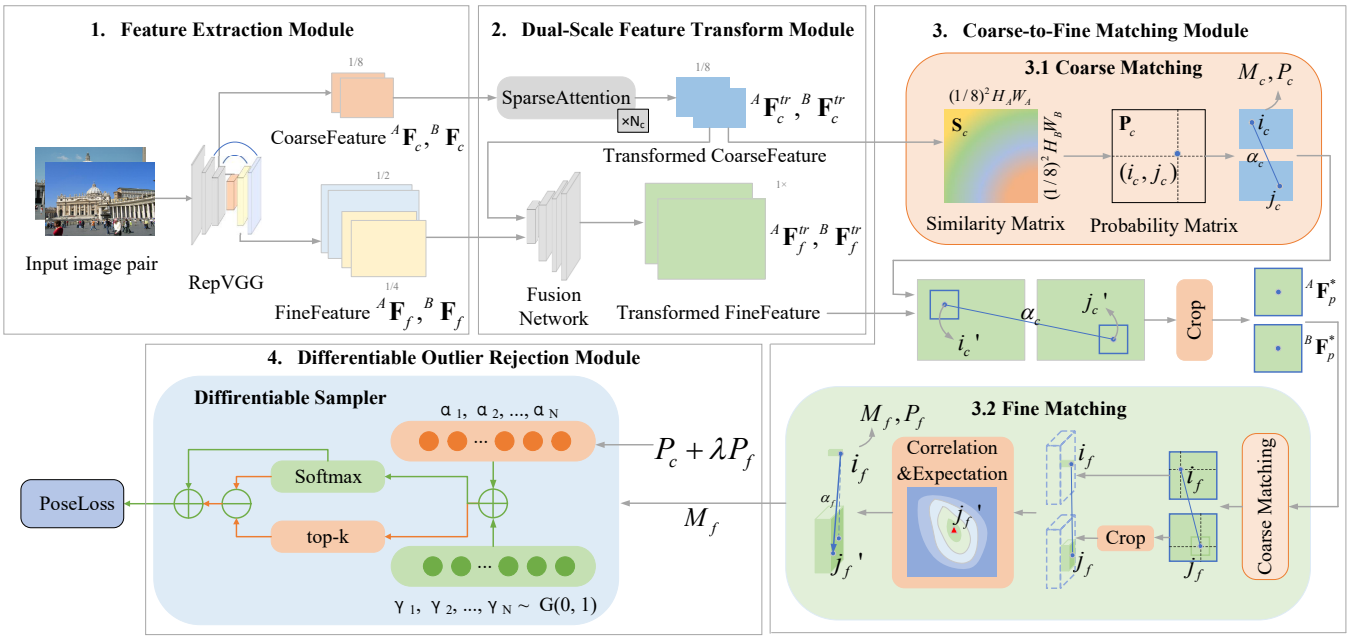


Fig. 2. Pipeline Overview of PG-Match. (1) The feature extraction module employs a convolutional neural network (CNN) to initially generate multi-scale feature maps. (2) The dual-scale feature transform module enhances initial feature maps using attention to produce transformed feature maps at both low resolution and original resolution. (3) The coarse-to-fine matching module leverages the low resolution feature maps for coarse matching, which are then mapped back to the original resolution feature maps and cropped for fine matching. (4) The differentiable outlier rejection module utilizes a weighted combination of coarse and fine confidence scores as priors to guide the sampling of fine matches for pose estimation, enabling the network supervised using only ground truth pose data.

input image pairs ${}^{A}\mathbf{I},{}^{B}\mathbf{I}$, yielding initial coarse feature maps ${}^{A}\mathbf{F}_c,{}^{B}\mathbf{F}_c$ at 1/8 resolution and fine feature maps ${}^{A}\mathbf{F}_f,{}^{B}\mathbf{F}_f$ at 1/4 and 1/2 resolutions. (2) The dual-scale feature transform module enhances initial feature maps using attention to produce transformed coarse feature maps ${}^{A}\mathbf{F}_c^{tr},{}^{B}\mathbf{F}_c^{tr}$ and fine feature maps ${}^{A}\mathbf{F}_f^{tr},{}^{B}\mathbf{F}_f^{tr}$. (3) The coarse-to-fine matching module employs a hierarchical matching strategy that first establishes coarse correspondences M_c with confidences P_c using coarse-scale feature maps to constrain the search space. Then these correspondences are refined using fine-scale feature maps to yield sub-pixel matches M_f with confidences P_f. (4) The differentiable outlier rejection module samples fine matches M_f using coarse and fine matching confidences P_c, P_f as priors to guide a differentiable RANSAC process, optimizing robust relative pose estimation.

A. Feature Extraction Module

The feature extraction module employs a RepVGG-based CNN to process input image pairs ${}^{A}\mathbf{I},{}^{B}\mathbf{I}$, initially generating multi-scale feature maps. RepVGG utilizes a multi-branch architecture during training to enhance feature expressiveness, reparameterized into a single convolution path at inference for computational efficiency. The module outputs coarse feature maps ${}^{A}\mathbf{F}_c,{}^{B}\mathbf{F}_c$ at 1/8 resolution to capture global contextual relationships, alongside fine feature maps ${}^{A}\mathbf{F}_f,{}^{B}\mathbf{F}_f$ at 1/4 and 1/2 resolutions for precise local correspondence estimation.

B. Dual-Scale Feature Transform Module

The dual-scale feature transform module is designed to enhance the robustness and expressiveness of feature maps for reliable semi-dense matching, by processing coarse feature maps ${}^{A}\mathbf{F}_c,{}^{B}\mathbf{F}_c$ and fine feature maps ${}^{A}\mathbf{F}_f,{}^{B}\mathbf{F}_f$ derived from the feature extraction module. These initial feature maps, generated by CNN, often lack sufficient capability of representing features in challenging scenarios such as low-texture or repetitive regions. This limitation stems from the reliance of CNN on local neighborhood aggregation, which results in insufficient feature representation in low-texture areas due to the lack of surrounding information. CNN also struggle to differentiate features in repetitive scenes, where similar local contexts can lead to inaccurate or failed matches.

To overcome these limitations, the dual-scale feature transform module utilizes attention mechanisms to enhance feature representations. Attention allows each feature point to weigh and aggregate information from the entire image, capturing long-range dependencies that CNNs overlook. Self-attention integrates intra-image relationships, enabling features in low-texture regions to draw from broader contextual cues, thus improving distinctiveness. Cross-attention incorporates inter-image relationships, distinguishing repetitive patterns by leveraging complementary information from paired images. This global aggregation addresses the representational deficiencies of CNN-based features, resulting in more robust and discriminative transformed maps for accurate matching.

The dual-scale feature transform module outputs enhanced coarse feature maps ${}^{A}\mathbf{F}_c^{tr}, {}^{B}\mathbf{F}_c^{tr}$ and fine feature maps ${}^{A}\mathbf{F}_f^{tr}, {}^{B}\mathbf{F}_f^{tr}$. The transformed coarse feature maps at low resolution are utilized for initial coarse matching, followed by mapping these coarse matches back to the original image resolution. Subsequently, the transformed fine feature maps at original resolution are employed to further refine and correct the matches with enhanced precision.

SparseAttention is applied to initial coarse feature maps ${}^{A}\mathbf{F}_c, {}^{B}\mathbf{F}_c$, producing transformed coarse feature maps ${}^{A}\mathbf{F}_c^{tr}, {}^{B}\mathbf{F}_c^{tr}$. To make the process faster and more efficient, the module performs sparse processing. First, it groups together the attention areas of nearby query tokens that usually look very similar, which avoids the same calculations. Second, it picks out the most important key tokens by focusing on the ones that carry the most weight in the attention process, since only a few of them really matter. By doing this, it simplifies the attention step and cuts down on unnecessary work.

As is shown in Fig. 3(a), for a query feature map ${}^{i}\mathbf{F}_c$, a convolutional layer downsamples it. For the other feature map ${}^{j}\mathbf{F}_c$, max pooling is used. This aggregation preserves essential local information while significantly lowering the computational cost of the subsequent attention mechanism.

The fusion network obtains fine feature maps ${}^{A}\mathbf{F}_f^{tr}, {}^{B}\mathbf{F}_f^{tr}$ by fusing upsampled transformed coarse feature maps ${}^{A}\mathbf{F}_c^{tr}, {}^{B}\mathbf{F}_c^{tr}$ with initial fine feature maps ${}^{A}\mathbf{F}_f, {}^{B}\mathbf{F}_f$. As is shown in Fig. 3(b), the fusion process employs convolution and upsampling to integrate the high expressiveness of low-resolution coarse features with the detailed spatial information of high-resolution fine features, ensuring a balance between computational efficiency and feature quality while mitigating the limitations of CNN-based local aggregation.

C. Coarse-to-Fine Matching Module

The coarse-to-fine matching module is designed to generate precise feature correspondences between two input images by leveraging the transformed coarse feature maps ${}^{A}\mathbf{F}_c^{tr}, {}^{B}\mathbf{F}_c^{tr}$ and fine feature maps ${}^{A}\mathbf{F}_f^{tr}, {}^{B}\mathbf{F}_f^{tr}$ from the dual-scale feature transform module. This module employs a hierarchical matching strategy that first establishes coarse correspondences using coarse-scale feature maps to constrain the search space and then refines these correspondences using fine-scale feature maps.

1) Coarse Matching: To establish initial correspondences, the coarse-to-fine matching module first performs coarse matching using the transformed coarse feature maps ${}^{A}\mathbf{F}_c^{tr}, {}^{B}\mathbf{F}_c^{tr}$. The feature maps are flattened into sequences of tokens,

$$ {}^{A}\mathbf{F}_c^{tr}, {}^{B}\mathbf{F}_c^{tr} = \text{Flatten}({}^{A}\mathbf{F}_c^{tr}, {}^{B}\mathbf{F}_c^{tr}) \in \mathbb{R}^{N_c \times D}, \tag{1} $$

where $N_c = H/8 \times W/8$. A similarity matrix is computed to measure the similarity between feature points:

$$ \mathbf{S}_c = \frac{{}^{A}\mathbf{F}_c^{tr} \cdot ({}^{B}\mathbf{F}_c^{tr})^T}{\sqrt{D}} \in \mathbb{R}^{N_c \times N_c}, \tag{2} $$

where the scaling factor $\sqrt{D}$ normalizes the dot product. To ensure robust matching, a mutual matching probability matrix is constructed:

$$ \mathbf{P}_c(i,j) = \text{softmax}(\mathbf{S}_c[i,:])[j] \cdot \text{softmax}(\mathbf{S}_c[:,j])[i], \tag{3} $$

where $\text{softmax}(\mathbf{S}[i,:])[j]$ represents the probability of feature point i in image A matching feature point j in image B, and $\text{softmax}\mathbf{S}[:,j])[i]$ ensures the reverse consistency. Coarse correspondences M_c are extracted by selecting high-probability mutual matches:

$$ \begin{aligned} M_c &= \{(i_c, j_c) \mid \mathbf{P}_c(i_c, j_c) > \theta\}, \\ P_c &= \{\alpha_c \mid \alpha_c = \mathbf{P}_c(i_c, j_c)\}, \end{aligned} \tag{4} $$

where i_c and j_c represent feature point indices in transformed coarse feature maps ${}^{A}\mathbf{F}_c^{tr}$ and ${}^{B}\mathbf{F}_c^{tr}$, respectively, and θ is a confidence threshold to filter reliable matches. The indices (i_c, j_c) are mapped to their corresponding coordinates to form point pairs in $\mathbb{R}^2 \times \mathbb{R}^2$. $\mathbf{P}_c(i_c, j_c)$ serves as the confidence score for the matching pair (i_c, j_c), reflecting its reliability.

2) Fine Matching: To generate precise correspondences, the coarse-to-fine matching module performs fine matching using the coarse correspondences M_c and the transformed fine feature maps ${}^{A}\mathbf{F}_f^{tr}, {}^{B}\mathbf{F}_f^{tr}$. Each coarse match $(i_c, j_c) \in M_c$ is mapped to the fine resolution grid by upsampling (scaling by 8), denoted as (i'_c, j'_c) and a local window of size $w \times w$ is cropped around the corresponding points in ${}^{A}\mathbf{F}_f^{tr}$ and ${}^{B}\mathbf{F}_f^{tr}$, denoted as ${}^{A}\mathbf{F}_p^{*}, {}^{B}\mathbf{F}_p^{*}$. The fine matching module refines each coarse match $(i_c, j_c) \in M_c$ within the corresponding ranges of ${}^{A}\mathbf{F}_p^{*}$ and ${}^{B}\mathbf{F}_p^{*}$, yielding fine matches $(i_f, j'_f) \in M_f$

Using the same method as coarse matching, a local similarity matrix $\mathbf{S}_f$ and mutual matching probability matrix $\mathbf{P}_f$ are computed to obtain the fine matching pair indices

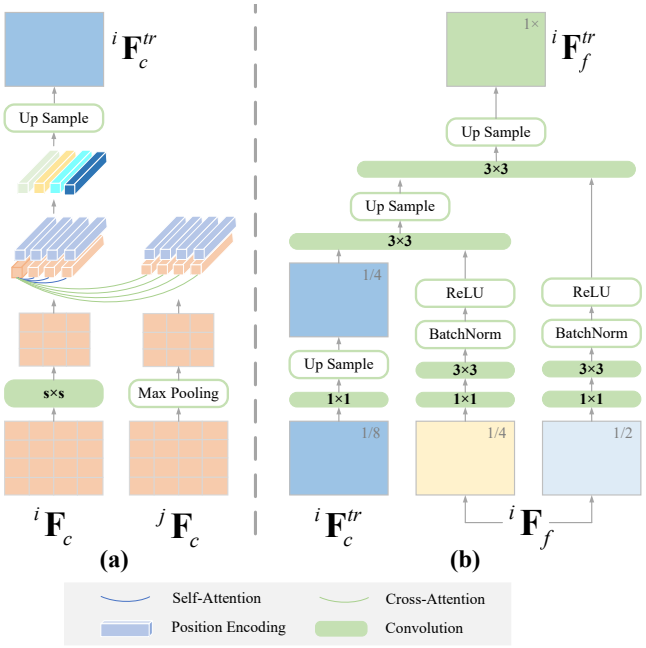


Fig. 3. (a) SparseAttention: Architecture processing coarse feature maps ${}^{A}\mathbf{F}_c, {}^{B}\mathbf{F}_c$ to yield transformed coarse feature maps ${}^{A}\mathbf{F}_c^{tr}, {}^{B}\mathbf{F}_c^{tr}$, enhancing distinctiveness via global information. (b) Fusion Network: Structure fusing upsampled transformed coarse features ${}^{A}\mathbf{F}_c^{tr}, {}^{B}\mathbf{F}_c^{tr}$ with fine feature maps ${}^{A}\mathbf{F}_f, {}^{B}\mathbf{F}_f$ to produce transformed fine feature maps ${}^{A}\mathbf{F}_f^{tr}, {}^{B}\mathbf{F}_f^{tr}$.

(i_f, j_f). To achieve sub-pixel accuracy, the point i_f in image A is fixed, and a heatmap $\mathbf{H}$ is computed by correlating the feature vectors in the $w \times w$ window around the point j_f in image B:

$$\mathbf{H} = \text{softmax}(\mathbf{S}_f[i_f, :]/T), \tag{5}$$

where T is a temperature parameter. The sub-pixel coordinate j'_f is obtained by computing the spatial expectation over the heatmap. Fine correspondences M_f and matching confidence are defined as:

$$\begin{aligned} M_f &= \{(i_f, j'_f)\}, \\ P_f &= \{\alpha_f \mid \alpha_f = \max(\mathbf{H})\}. \end{aligned} \tag{6}$$

D. Differentiable Outlier Rejection Module

To ensure robust feature correspondences for downstream pose estimation, the DORM filters out unreliable matches from the fine correspondence set M_f using the associated confidence scores P_c and P_f. A key challenge in this process is that discrete sampling of matches based on confidence thresholds is non-differentiable, preventing end-to-end training of the network:

$$M'_f = \{m_k \in M_f \mid w_k > \theta \text{ or } w_k \in \text{top-}k(w)\}, \tag{7}$$

where θ is a confidence threshold, w_k is the confidence score associated with the k-th sampled correspondence, and top-k selects the k highest-scoring matches. This process relies on non-differentiable operations like thresholding or argmax, which yield zero gradients, preventing gradient-based optimization of the network.

To address this, we employ the Gumbel-Softmax relaxation method [32], [33] to approximate discrete match selection with a differentiable process. For each correspondence $m_k \in \mathbf{M}_f$ with confidence $\alpha_{c,k}$ and $\alpha_{f,k}$, the module computes differentiable weights using the Gumbel-Softmax distribution, enabling robust outlier rejection. The implementation proceeds as follows:

A weight w_k is computed for the k-th correspondence by augmenting its confidence score with Gumbel noise:

$$\begin{aligned} \hat{y}_k &= \log(\pi_k) + \gamma_k, \\ \pi_k &= \lambda\, \alpha_{f,k} + \alpha_{c,k}, \end{aligned} \tag{8}$$

where $\gamma_k \sim \text{Gumbel}(0, 1)$. These scores are transformed into a continuous probability distribution via a softmax operation with temperature τ:

$$w_k = \text{Gumbel-Softmax}(\pi_k, \tau) = \frac{\exp(\hat{y}_k/\tau)}{\sum_{m=1}^{M} \exp(\hat{y}_m/\tau)}. \tag{9}$$

As illustrated in differentiable outlier rejection module of Fig. 2, the forward pass uses discrete thresholding as equation 7, this discrete selection corresponds to connections marked by red arrows, indicating no gradient flow due to the non-differentiable nature of thresholding and top-k operations. To ensure differentiability, the gradients in the backward pass flow through the continuous w_k as equation 9, bypassing the non-differentiable thresholding, represented by green arrows.

TABLE I
EVALUATION ON MEGADEPTH-1500 WITH POSE ACCURACY METRICS

Category	Method	Pose Accuracy Metrics		
		AUC@5°	AUC@10°	AUC@20°
Sparse	SP + NN$_{\text{(CVPRW 18)}}$	31.7	46.8	60.1
	SP + SG$_{\text{(CVPR 20)}}$	49.7	67.1	80.6
	SP + LG$_{\text{(ICCV 23)}}$	49.9	67.0	80.1
Semi-Dense	DRC-Net$_{\text{(NeurIPS 20)}}$	27.0	42.9	58.3
	LoFTR$_{\text{(CVPR 21)}}$	52.8	69.2	81.2
	QuadTree$_{\text{(ICLR 22)}}$	54.6	70.5	82.2
	MatchFormer$_{\text{(ACCV 22)}}$	53.3	69.7	81.8
	TopicFM$_{\text{(AAAI 23)}}$	54.1	70.1	81.6
	AspanFormer$_{\text{(ECCV 22)}}$	55.3	71.5	83.1
	eLoFTR$_{\text{(CVPR 24)}}$	<u>56.4</u>	<u>72.2</u>	<u>83.5</u>
	JamMa$_{\text{(CVPR 25)}}$	53.7	69.8	81.6
	Ours	**57.8**	**73.0**	**84.0**

E. Supervision

To optimize the network, we employ a loss function based on the average pose error, which quantifies the discrepancy between estimated and ground truth relative camera poses.

The pose loss computes the average rotation and translation errors for a batch of estimated models. Given keypoint correspondences M_f and ground truth poses consisting of rotation $R \in SO(3)$ and translation $t \in \mathbb{R}^3$, the loss is calculated as follows:

*1) **Essential Matrix Estimation**:* The keypoint correspondences are used to estimate a set of models $\hat{\mathbf{E}} = \{\hat{\mathbf{E}}_i\}_{i=1}^{N}$, where $\hat{\mathbf{E}}_i$ is essential matrix.

The keypoint correspondences $(\mathbf{p}_1, \mathbf{p}_2)$ are in pixel coordinates, which are first converted into 3D homogeneous coordinates $\mathbf{p}'_1$ and $\mathbf{p}'_2$. A minimal subset of M_f is sampled to directly estimate $\hat{\mathbf{E}}_i$ using the five-point algorithm [34], which enforces the essential matrix constraint:

$$(K_2^{-1}\mathbf{p}'_2)^\top \hat{\mathbf{E}}_i (K_1^{-1}\mathbf{p}'_1) = 0, \tag{10}$$

where K_1, K_2 are the camera intrinsic matrices for the two views.

*2) **Pose Error Calculation**:* For each estimated model $\hat{\mathbf{E}}_i$, the rotation error L_R and translation error L_t are computed relative to the ground truth R and t using a pose recovery function that decomposes $\hat{\mathbf{E}}_i$ into estimated rotation $\hat{R}_i$ and translation $\hat{t}_i$. The error for the i-th model is defined as:

$$L_i = \frac{L_R(\hat{R}_i, R) + L_t(\hat{t}_i, t)}{2}, \tag{11}$$

where L_R and L_t measure the angular discrepancies in rotation and translation, respectively.

The pose loss ensures that the network optimizes both the feature correspondences and the geometric models by minimizing pose errors.

V. EXPERIMENTS

We conduct comprehensive experiments to evaluate our PG-Match framework on the MegaDepth-1500 dataset [35] and CVPR IMW 2020 PhotoTourism dataset [36]. MegaDepth-1500 is widely used for semi-dense feature matching, comprising 1500 image pairs with ground truth relative poses and depth maps. The PhotoTourism

TABLE II

EVALUATION ON MEGADEPTH-1500 DATASET WITH GEOMETRIC AND RUNTIME METRICS

Method/Metrics	Inliers(%)↑	Mean Error(px)↓	Median Error(px)↓	Time(ms)↓
LoFTR	73.99	5.84	1.71	360.70
QuadTree	76.46	3.92	1.75	628.93
eLoFTR	77.91	4.67	1.76	**266.49**
Ours	**79.67**	**3.26**	**1.65**	270.75

TABLE III

CROSS-DATASET EVALUATION ON PHOTOTOURISM DATASET WITH POSE ACCURACY METRICS

Metrics/Method	Grand_place_brussels			Trevi_fountain			Westminster_abbey		
	auc@5°	auc@10°	auc@20°	auc@5°	auc@10°	auc@20°	auc@5°	auc@10°	auc@20°
LoFTR	79.4	89.7	94.8	18.2	55.1	77.6	12.3	34.1	65.4
QuadTree	75.5	87.8	93.9	18.1	45.8	71.8	11.9	38.2	69.1
eLoFTR	77.1	88.5	94.3	25.7	54.1	76.7	11.5	38.4	68.8
Ours	**81.9**	**91.0**	**95.5**	**26.5**	**56.0**	**78.2**	**22.2**	**51.1**	**75.5**

TABLE IV

ABLATION STUDIES ON MEGADEPTH-1500 WITH POSE ACCURACY METRICS

Variants			Pose Accuracy Metrics		
Coarse conf	Fine conf	DORM	auc@5°	auc@10°	auc@20°
×	×	×	56.4	72.2	83.5
×	×	✓	56.9	72.4	83.5
✓	×	×	56.4	72.5	84.0
×	✓	×	57.2	72.4	83.5
✓	✓	×	57.6	72.9	83.8
✓	✓	✓	**57.8**	**73.0**	**84.0**

TABLE V

DOWNSTREAM APPLICATION TO STRUCTURE FROM MOTION ON TEXTURE-POOR SFM DATASET WITH POSE ACCURACY METRICS

Category	Method	auc@5°	auc@10°	auc@20°
Detector-Based	COLMAP (SIFT+NN)	2.87	3.85	4.95
	SIFT+NN+PixSfM	3.13	4.08	5.09
	D2Net+NN+PixSfM	1.54	2.63	4.54
	R2D2+NN+PixSfM	3.79	5.51	7.84
	SP+SG+PixSfM	14.00	19.23	24.55
Detector-Free	LoFTR+PixSfM	20.66	30.49	42.01
	Ours	**24.30**	**34.93**	**46.75**

dataset contains diverse real-world image pairs captured from various tourist sites, which provides ground truth poses but does not include depth information. In addition, Texture-Poor SfM dataset [2] is used for downstream SfM experiments, showing its effectiveness and applicability in practical 3D reconstruction tasks.

Following previous works [8], [10], [11], pose AUC metrics are used to evaluate the quality of correspondences, which provides a comprehensive measure related to both local accuracy and global consistency. In addition, we use geometric metrics to separately assess both aspects, where the inlier ratio indicates global consistency and epipolar errors measure local matching accuracy.

A. Evaluation on MegaDepth-1500 Dataset

Comparison experiments and ablation studies are conducted on the MegaDepth-1500 dataset.

Comparison Experiments. We compare PG-Match against state-of-the-art sparse and semi-dense matching methods to demonstrate its effectiveness. Table I shows pose accuracy under different thresholds, PG-Match consistently achieves the highest accuracy, as expected due to direct supervision of ground truth poses. Table II reports geometric and runtime metrics. PG-Match achieves the best overall results while maintaining comparable computational efficiency. The higher inlier ratio indicates the improvements of global consistency, and the lower epipolar errors demonstrate more precise local correspondences.

Qualitative experiments were conducted to compare our method with both detector-based and detector-free approaches, as illustrated in Fig 4 and Fig 5. Compared to the detector-based SP+LG, PG-Match produces denser and more reliable correspondences in challenging regions. Compared to the detector-free eLoFTR, PG-Match yields more inlier correspondences with wider spatial distribution. This more uniform distribution provides stronger global geometric constraints for pose estimation, consistent with the higher pose metrics.

Ablation Studies. Table IV reports ablation results on the MegaDepth-1500 dataset. We evaluate the impact of three key components: coarse confidence priors, fine confidence priors, and the DORM. The results demonstrate that combining all components yields the best performance. Specifically, incorporating the DORM not only enables the network to be supervised with pose but also yields modest performance improvements. Moreover, incorporating coarse confidence priors leads to improvements in coarse-threshold metrics (AUC@10°, AUC@20°), while fine confidence priors enhance performance under stricter thresholds (AUC@5°, AUC@10°), which is consistent with intuitive expectations. Applying both priors jointly as weighted guidance results in consistent gains across all AUC metrics.

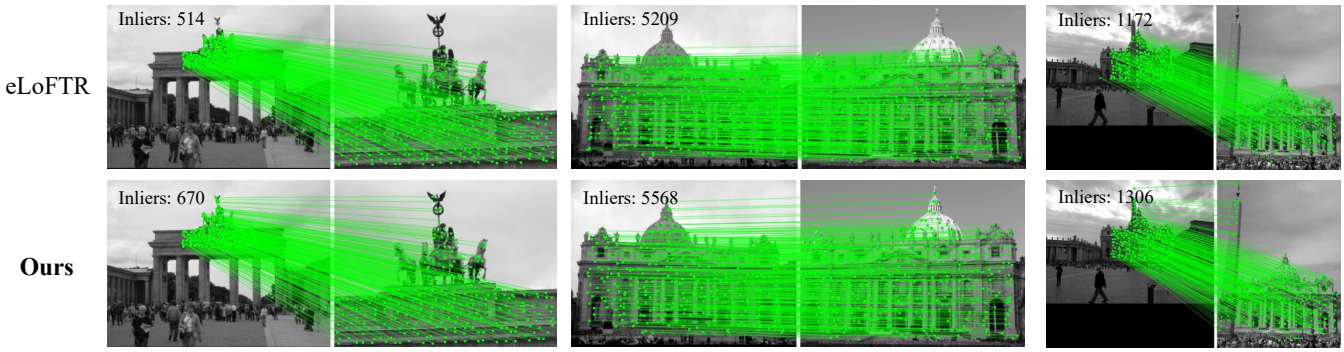
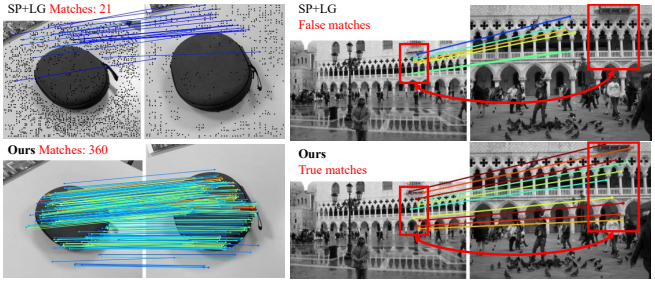


Fig. 4. **Qualitative Results.** Comparison of inlier matches between our method and the detector-free baseline eLoFTR. Our method produces more inliers with a wider spatial distribution.



(a) Low-texture areas (b) Scenes with repetitive structure

Fig. 5. **Qualitative Results.** Our method is qualitatively compared with the detector-based baseline SP+LG. (a) Our method produces more matches in low-texture regions. (b) The red boxes indicate the corresponding areas in the image pair, demonstrating that our method ensures correct matches even in repetitive scenarios. Matching lines with colors approaching red indicate higher confidence, while those approaching blue indicate lower confidence.

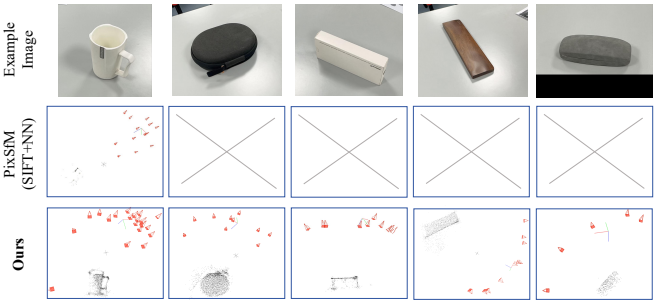


Fig. 6. **Qualitative Results.** SfM with our method is qualitatively compared with the detector-based baseline PixSfM on Texture-Poor SfM Dataset.

B. Cross-Dataset Evaluation on PhotoTourism Dataset

To demonstrate the cross-dataset generalization of PG-Match, we evaluate its performance on the CVPR IMW 2020 PhotoTourism dataset. Existing semi-dense matching methods rely on ground truth both pose and depth for supervision during training. The lack of depth information of the dataset limits their generalization, resulting in significant performance degradation in unseen scenes. In contrast, PG-Match requires no depth information for training, enabling the method to be applied to diverse datasets.

We train PG-Match on the Colosseum_exterior scene, consisting of 4950 image pairs split 3:1 into training and validation sets. The model is evaluated on multiple test scenes from Grand_place_brussels, Trevi_fountain, and Westminster_abbey sequences using pose accuracy metrics, as reported in Table III. PG-Match consistently outperforms the LoFTR series across all sequences, demonstrating the generalization across diverse scenes.

C. Downstream Application to Structure from Motion

To evaluate the practical utility of PG-Match, we assess its performance in the downstream task of SfM [2], a critical application for 3D reconstruction from a collection of images. Experiments are performed on Texture-Poor SfM dataset, where traditional feature-based methods often struggle due to the lack of distinctive visual patterns. The results, presented in Table V, demonstrate the effectiveness of PG-Match when integrated into a SfM pipeline, compared against both detector-based and detector-free methods. The qualitative experimental results are illustrated in Fig. 6.

D. Limitations

Although PG-Match maintains competitive efficiency during inference, the training process remains relatively resource-intensive. This is primarily due to the large network size and the need to process high-dimensional correlation volumes for semi-dense matching, which results in substantial memory consumption and longer training times. Recent works, such as Jamma [29], address similar challenges by employing parameter-efficient network architectures and optimized attention mechanisms, significantly reducing model size and improving training efficiency. Inspired by these strategies, future work could explore lightweight design and efficient training schemes for PG-Match, aiming to further lower computational demands while preserving both local precision and global consistency of feature correspondences.

VI. CONCLUSIONS

In this work, we proposed PG-Match, a detector-free feature matching framework that leverages ground truth poses for supervision, eliminating the need for depth information. Our approach demonstrates improvements in both global consistency and local matching precision, as shown by superior pose accuracy and geometric metrics

on benchmark datasets. The introduction of DORM allows the network to maintain differentiability while optimizing pose accuracy, further enhancing performance. A coarse-to-fine matching strategy is employed for efficiency, where specially designed confidence priors ensure efficient convergence. Through extensive experiments on the MegaDepth-1500 and PhotoTourism datasets, the effectiveness of PG-Match is validated in challenging scenarios, highlighting its generalization ability and applicability to unseen scenes. The results confirm that PG-Match outperforms state-of-the-art methods, providing a robust solution for semi-dense feature matching and downstream tasks, such as pose estimation and 3D reconstruction.

REFERENCES

[1] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski, "Building rome in a day," *Communications of the ACM*, vol. 54, no. 10, pp. 105–112, 2011.

[2] X. He, J. Sun, Y. Wang, S. Peng, Q. Huang, H. Bao, and X. Zhou, "Detector-free structure from motion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 21 594–21 603.

[3] P. Lindenberger, P.-E. Sarlin, V. Larsson, and M. Pollefeys, "Pixel-perfect structure-from-motion with featuremetric refinement," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 5987–5997.

[4] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[5] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 224–236.

[6] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4938–4947.

[7] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "Lightglue: Local feature matching at light speed," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 17 627–17 638.

[8] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8922–8931.

[9] H. Chen, Z. Luo, L. Zhou, Y. Tian, M. Zhen, T. Fang, D. Mckinnon, Y. Tsin, and L. Quan, "Aspanformer: Detector-free image matching with adaptive span transformer," in *European conference on computer vision*. Springer, 2022, pp. 20–36.

[10] Y. Wang, X. He, S. Peng, D. Tan, and X. Zhou, "Efficient loftr: Semi-dense local feature matching with sparse-like speed," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 21 666–21 675.

[11] S. Tang, J. Zhang, S. Zhu, and P. Tan, "Quadtree attention for vision transformers," *arXiv preprint arXiv:2201.02767*, 2022.

[12] N. Savinov, A. Seki, L. Ladicky, T. Sattler, and M. Pollefeys, "Quad-networks: unsupervised learning to rank for interest point detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1822–1830.

[13] H. Yao, N. Hao, C. Xie, and F. He, "Edgepoint: Efficient point detection and compact description via distillation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 766–772.

[14] P. Ebel, A. Mishchuk, K. M. Yi, P. Fua, and E. Trulls, "Beyond cartesian representations for local descriptors," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 253–262.

[15] A. Mishchuk, D. Mishkin, F. Radenovic, and J. Matas, "Working hard to know your neighbor's margins: Local descriptor learning loss," *Advances in neural information processing systems*, vol. 30, 2017.

[16] Y. Tian, B. Fan, and F. Wu, "L2-net: Deep learning of discriminative patch descriptor in euclidean space," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 661–669.

[17] Y. Tian, X. Yu, B. Fan, F. Wu, H. Heijnen, and V. Balntas, "Sosnet: Second order similarity regularization for local descriptor learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 11 016–11 025.

[18] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, "D2-net: A trainable cnn for joint description and detection of local features," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2019, pp. 8092–8101.

[19] Z. Luo, L. Zhou, X. Bai, H. Chen, J. Zhang, Y. Yao, S. Li, T. Fang, and L. Quan, "Aslfeat: Learning local features of accurate shape and localization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6589–6598.

[20] J. Revaud, C. De Souza, M. Humenberger, and P. Weinzaepfel, "R2d2: Reliable and repeatable detector and descriptor," *Advances in neural information processing systems*, vol. 32, 2019.

[21] Y. Tian, V. Balntas, T. Ng, A. Barroso-Laguna, Y. Demiris, and K. Mikolajczyk, "D2d: Keypoint extraction with describe to detect approach," in *Proceedings of the Asian conference on computer vision*, 2020.

[22] H. Chen, Z. Luo, J. Zhang, L. Zhou, X. Bai, Z. Hu, C.-L. Tai, and L. Quan, "Learning to match features with seeded graph matching network," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6301–6310.

[23] Y. Shi, J.-X. Cai, Y. Shavit, T.-J. Mu, W. Feng, and K. Zhang, "Clustergnn: Cluster-based coarse-to-fine graph neural network for efficient feature matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 517–12 526.

[24] I. Rocco, M. Cimpoi, R. Arandjelović, A. Torii, T. Pajdla, and J. Sivic, "Neighbourhood consensus networks," *Advances in neural information processing systems*, vol. 31, 2018.

[25] I. Rocco, R. Arandjelović, and J. Sivic, "Efficient neighbourhood consensus networks via submanifold sparse convolutions," in *European conference on computer vision*. Springer, 2020, pp. 605–621.

[26] X. Li, K. Han, S. Li, and V. Prisacariu, "Dual-resolution correspondence networks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 346–17 357, 2020.

[27] Q. Wang, J. Zhang, K. Yang, K. Peng, and R. Stiefelhagen, "Matchformer: Interleaving attention in transformers for feature matching," in *Proceedings of the Asian conference on computer vision*, 2022, pp. 2746–2762.

[28] K. T. Giang, S. Song, and S. Jo, "Topicfm: Robust and interpretable topic-assisted feature matching," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 2, 2023, pp. 2447–2455.

[29] X. Lu and S. Du, "Jamma: Ultra-lightweight local feature matching with joint mamba," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 14 934–14 943.

[30] J. Edstedt, Q. Sun, G. Bökman, M. Wadenbäck, and M. Felsberg, "Roma: Robust dense feature matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 790–19 800.

[31] J. Edstedt, I. Athanasiadis, M. Wadenbäck, and M. Felsberg, "Dkm: Dense kernelized feature matching for geometry estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 765–17 775.

[32] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.

[33] T. Wei, Y. Patel, A. Shekhovtsov, J. Matas, and D. Barath, "Generalized differentiable ransac," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 649–17 660.

[34] H. Li and R. Hartley, "Five-point motion estimation made easy," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 1. IEEE, 2006, pp. 630–633.

[35] Z. Li and N. Snavely, "Megadepth: Learning single-view depth prediction from internet photos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2041–2050.

[36] J. L. Schonberger and J.-M. Frahm, "Cvpr 2020 workshop on image matching challenge," in *CVPR Workshops*, 2020.