CoMatcher: Multi-View Collaborative Feature Matching

Jintao Zhang[1], Zimin Xia[2]*, Mingyue Dong[1], Shuhan Shen[3], Linwei Yue[4], Xianwei Zheng[1]*
[1]The State Key Lab. LIESMARS, Wuhan University
[2]École Polytechnique Fédérale de Lausanne (EPFL)
[3]Chinese Academy of Sciences [4]China University of Geosciences

Abstract

This paper proposes a multi-view collaborative matching strategy for reliable track construction in complex scenarios. We observe that the pairwise matching paradigms applied to image set matching often result in ambiguous estimation when the selected independent pairs exhibit significant occlusions or extreme viewpoint changes. This challenge primarily stems from the inherent uncertainty in interpreting intricate 3D structures based on limited two-view observations, as the 3D-to-2D projection leads to significant information loss. To address this, we introduce CoMatcher, a deep multi-view matcher to (i) leverage complementary context cues from different views to form a holistic 3D scene understanding and (ii) utilize cross-view projection consistency to infer a reliable global solution. Building on CoMatcher, we develop a groupwise framework that fully exploits cross-view relationships for large-scale matching tasks. Extensive experiments on various complex scenarios demonstrate the superiority of our method over the mainstream two-view matching paradigm.

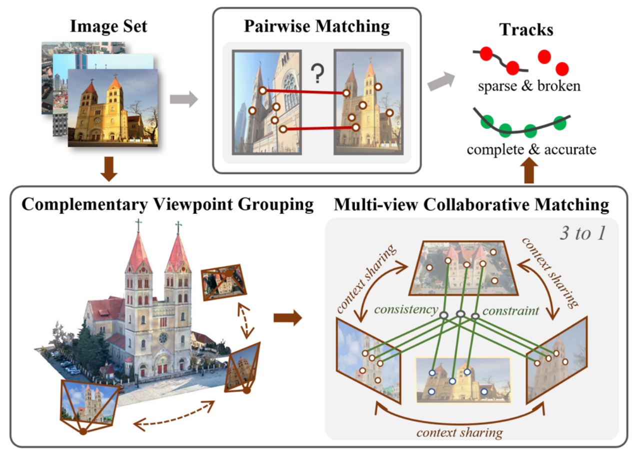

Figure 1. **From two view to multi-view.** Unlike pairwise schemes prone to uncertainty, we first partition the image set into co-visible groups and collaboratively estimate group-wide correspondences using a reliable deep multi-view matcher. Benefiting from holistic scene understanding and consistency constraint, our method generates high-quality tracks in complex scenarios.

1. Introduction

Research on recovering 3D structures and camera poses from multiple views storing scene information has a long history [20, 23, 34]. A core component of this process is feature matching, which aims to estimate point correspondences across the image set [25, 32, 58]. For computational flexibility, existing methods often decompose the set into pairs of co-visible images [1, 18] and apply a two-view matching approach to each pair independently [31, 43]. The pairwise matches are then merged into comprehensive multi-view tracks [1, 45]. Following this pairwise framework to build systems for localization [22, 42] and mapping [29, 45] has become a entrenched research paradigm.

Recent advances in learning-based two-view matchers have greatly improved performance in most cases [15, 30, 43, 50]. Inspired by Transformer [54], these methods employ deep networks to jointly analyze global context from both views. This enables models to learn to infer underlying 3D scenes by integrating spatial and visual cues, which has been proven crucial for accurate estimations [27, 43].

Despite great progress, current two-view matchers still struggle in challenging wide-baseline scenarios, particularly with severe occlusions and repetitive textures [25]. We argue that one potential cause for this difficulty lies in the inherent uncertainty of interpreting complex 3D structures from limited two-view observations— a task that even humans find prone to perceptual ambiguity. Projecting complex 3D geometries onto 2D planes inevitably results in significant loss of scene information. For example, spatially distant regions may appear close in 2D. This makes it highly unreliable to infer the original scene solely from ambiguous two-view context [6, 57]. Moreover, while matchers can learn two-view geometric priors, this alone is also in-

*Corresponding author.

adequate—capturing abrupt depth discontinuities requires stronger constraints [4, 27]. Errors in pairwise correspondences compound during merging, significantly impacting downstream tasks, especially in large-scale datasets.

To address these challenges, we propose a shift in focus: instead of optimizing two-view matchers, a more effective strategy is to directly leverage the rich relationships inherent in raw multi-view observations. Building on this, we introduce a unified formulation that collaboratively establishes correspondences between a complementary view group and a target view. This approach fosters holistic scene understanding by integrating contextual information from multiple views, while multi-view geometric priors provide stronger constraints for reliable reasoning. Moreover, matches across views naturally adhere to cross-view projection consistency. This facilitates a mutual verification process, ensuring confident inference of a global solution.

Inspired by existing two-view methods [43, 50], we develop CoMatcher, a deep-learning architecture for multi-view feature matching. Context sharing is achieved through a GNN with multi-view receptive fields. While a larger feature space enriches context, it also introduces noise, complicating learning. To address this, we propose to constrain the search space for each point by leveraging cross-view projection geometry. Furthermore, we redesign the feature correlation layers in cross-attention to enable consistent multi-view reasoning. This design exploits the property that correlation scores encode both reasoning outcomes and confidence, enabling progressive layer-by-layer integration of multi-view estimates into a globally consistent solution.

To scale CoMatcher for large-scale tasks like SfM [36, 45], we develop a groupwise matching pipeline (see Fig. 1). It introduces a novel set partitioning algorithm that uses co-visibility information to group images, followed by group-level matching using CoMatcher. By preserving cross-view relationships, our framework achieves comprehensive improvements over pairwise methods in extensive experiments on wide-baseline scenarios. The main contributions are summarized as follows: (i) A novel multi-view matching paradigm, comprising a deep collaborative matcher and a scalable groupwise pipeline, designed to overcome the inherent uncertainty of pairwise approaches. (ii) A multi-view representation learning network architecture that enables holistic scene understanding. (iii) A multi-view feature correlation strategy that integrates uncertain individual estimates into a globally consistent solution.

2. Related work

Two-view matching is a fundamental problem in computer vision. The most prevalent paradigm involves sparsifying images into keypoints, each represented by a high-dimensional vector encoding its visual context [25, 32]. This transforms matching into a search problem, with match likelihood expressed by the inner product of corresponding vectors [31]. Modern methods leverage deep networks for representation learning to extract feature points [10, 13, 14, 39, 53, 59], outperforming traditional hand-crafted encodings [5, 31, 41] and achieving significant advancements on mainstream benchmarks [3, 25, 46]. In parallel, other works aim to address the limitations of heuristic matching strategies by leveraging networks with global modeling capabilities to directly learn the matching process [7, 24, 30, 43]. Such data-driven methods can effectively capture priors of the underlying 3D scene, which are essential for accurate estimation. Unlike keypoint-based approaches, another typical method retains all pixels for dense inference [8, 15, 27, 50, 52]. While this has led to notable progress in weakly textured regions, such methods require greater computational resources [30, 56] and struggle with multi-view inconsistency [21]. In summary, while these methods perform well in most scenarios, they remain constrained in complex wide-baseline settings due to the limited receptive field inherent to two-view systems.

Multi-view matching focuses on establishing correspondences across a larger-scale image collection. A typical pairwise matching approach decomposes the problem into two-view matching instances and subsequently merges the pairwise results using multi-view projection consistency [1, 42, 44, 45]. During merging, some works explore how consistency assumptions can filter and refine coarse matches [33, 51], while others focus on efficiency, developing better strategies to reduce redundant matching computations [35, 60]. However, as a post-processing step, merging cannot address catastrophic failures that occur during the matching stage. Another line of work treats matching as a point tracking problem [11, 12, 19, 26], using optical flow-like frameworks to reason about the positions of query pixels across multiple images. While this approach is well-suited for video inputs, its effectiveness on unstructured, wide-baseline image collections remains largely unverified. The most relevant method to ours is [40], which performs end-to-end multi-view matching and pose estimation. However, it is limited in scalability, as it can only handle a small image group. In contrast, our framework is highly flexible, with the potential to scale to an arbitrary number of images.

3. Methodology

Formulation. Given a set of N_I images, $\mathcal{I} = \{I_i \mid i = 1, \ldots, N_I\}$, of a scene, the objective of feature matching is to identify the 2D position coordinates $\mathcal{M}_k = \{\mathbf{p}^{I_i}\}$ in $\mathcal{I}$ that correspond to identical 3D scene points $\mathcal{X} = \{\mathbf{x}_k \mid k = 1, \ldots, N_X\}$, where $\mathbf{p}^{I_i} \in \mathbb{R}^2$ are the pixel coordinates in I_i, and each $\mathcal{M}_k$ is called a *track*.

Motivation. Constructing complete tracks requires estimating correspondences between each image $I_t \subset \mathcal{I}$ and its co-visible image set $\mathcal{C}_t$. While the pairwise approach of-

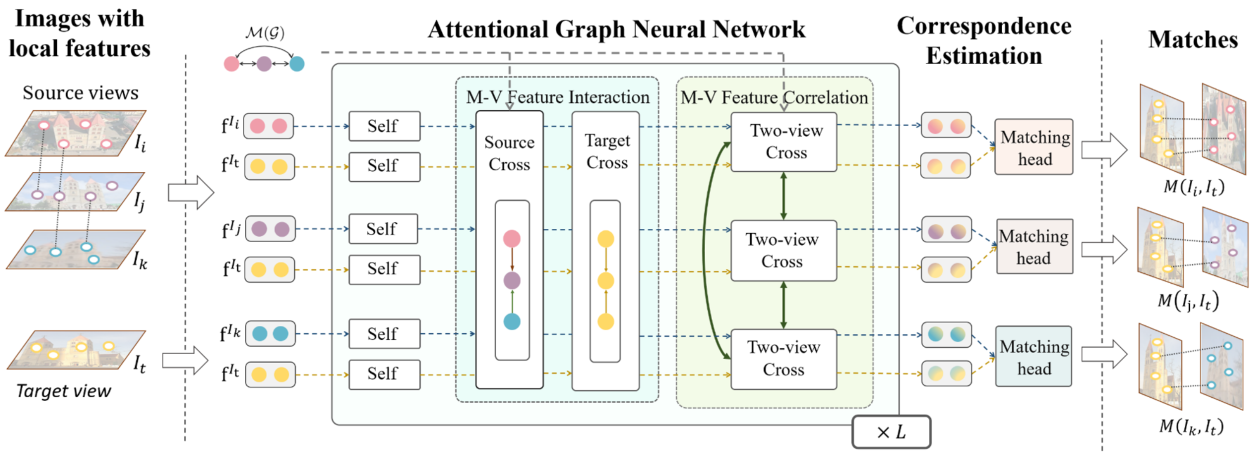

Figure 2. **The CoMatcher architecture** can be viewed as an extension of parallel two-view matchers, enhanced with two core components: the multi-view (M-V) feature interaction module (Sec. 3.2) and the multi-view (M-V) feature correlation strategy (Sec. 3.3).The first component employs multi-view cross-attention to enhance the representation of all source and target view features, while leveraging the cross-view projection geometry from $\mathcal{M}(\mathcal{G})$ to constrain the point search space. The second component aggregates correlation scores from multiple two-view cross-attention modules based on estimated confidence, guiding the network to achieve multi-view consistent reasoning.

fers implementation flexibility for this problem, its reliance on two-view inputs in a single estimation often leads to uncertainty in complex scenes. In practice, images within $\mathcal{C}_t$ are often not independent—most are captured from adjacent positions but with differing viewing angles. This inspires us to group $\mathcal{C}_t$ based on image correlations, where each complementary group offers a mapping from multiple planar views to a 3D sub-scene. In the image matching problem, the correspondence between I_t and a group must adhere to certain 2D-to-3D physical constraints, such as cross-view projection consistency. Rather than independently estimating and then integrating them with constraints, it is more efficient to unify this process and learn underlying priors directly from the data. Thus, we formulate CoMatcher to address the 1-to-N matching problem (see Fig. 2).

Groupwise matching pipeline. Based on the above, we propose a three-step pipeline for constructing tracks: grouping, connecting, and matching. First, we treat the original set of views $\mathcal{I}$ as a union of smaller image groups $\{\mathcal{G}_s \mid s = 1, \ldots, N_G\}$, each representing a localized scene. This involves a pre-processing step to partition the set. The most critical aspect of this is determining the correlation between two images, i.e., whether they describe the same structure. Building on existing research [2, 38, 48], we employ heuristic techniques to design a grouping algorithm. Next, although CoMatcher can directly learn multi-view geometry, certain fine-grained physical rules are challenging to model. Therefore, we explicitly establish relationships within each group to guide subsequent matching. For each group, we construct tracks using existing frameworks, which provide a set of 3D points along with their multi-view projections and implicitly encode camera poses. These cues are incorporated as inputs to the network to steer the inference process. Finally, each group of images is treated as a whole to collaboratively match with other images in $\mathcal{I}$ using CoMatcher. Additional details on grouping and the overall framework are provided in the supplementary materials.

3.1. CoMatcher overview

As shown in Fig. 2, CoMatcher, as a deep sparse matcher, estimates point correspondences between each *source* view in a group $I_i \in \mathcal{G}$ and a *target* view I_t. The network takes as input the extracted local features $\{(\mathbf{p}_k^{I_i}, \mathbf{d}_k^{I_i}) \mid k = 1, \ldots, N_F\}$ from the M source views and the target view, along with the precomputed tracks $\mathcal{M}(\mathcal{G})$ of $\mathcal{G}$. Each local feature consists of keypoint locations $\mathbf{p}_k$ and their corresponding visual descriptors $\mathbf{d}_k$. Initially, the target view features are broadcast and paired with those of each source view to establish the initial states. These M feature pairs are then refined using a Graph Neural Network (GNN) with alternating self-attention and cross-attention modules, repeated L times. Finally, the enhanced features are fed into matching heads to infer the final correspondences.

The design of CoMatcher revolves around two core questions: (i) how can multi-view context be leveraged to learn better point representations (Sec. 3.2), and (ii) how can the estimations across multiple views be constrained to satisfy cross-view consistency (Sec. 3.3)? After thoroughly discussing these questions, we introduce the prediction heads (Sec. 3.4) and the loss function (Sec. 3.5).

3.2. Multi-view feature interaction

Consider an occluded scene. When observing such structures from a frontal view, points near occlusion boundaries are often contaminated by irrelevant context. This noise significantly compromises the quality of point features, creating substantial challenges for matching. Multi-view learning provides a promising solution to this issue. For these ambiguous point features, it can integrate observations of the same area from multiple other viewpoints to refine their representations. This involves a querying process, which establishes a multi-view receptive field for each point.

Given a set of co-visible source views, CoMatcher leverages a multi-view cross-attention mechanism to aggregate these context (Source Cross). For a point u in source view I_i, we attend to all points in the other source views:

$$\mathbf{m}_u^{I_i \leftarrow \mathcal{W}_j} = \frac{1}{M-1} \sum_{I_j \in \mathcal{G} \setminus I_i} \sum_{v \in \mathcal{W}_j} \mathrm{Softmax}\left(a_{uv}^{I_i I_j}\right) \mathbf{v}_v^{I_j}. \quad (1)$$

Here, $\mathcal{W}_j$ denotes the point set in the source view I_j, while $a_{uv}^{I_i I_j}$ represents the similarity score computed by treating point u in I_i as the query and point v in I_j as the key. $\mathbf{v}_v^{I_j}$ is the value of point v. $\mathbf{m}_u^{I_i}$ represents the message vector from $\mathcal{W}_j$ to u of the GNN [43]. We uniformly aggregate features from each source view, encouraging the model to integrate multi-view information comprehensively rather than overemphasizing images from similar perspectives.

However, for a point near an occlusion boundary in one source view, how does it identify the corresponding points observing the same location in other source views? This implies a matching process. While the network could directly learn this, it would introduce additional complexity, as establishing correspondences between source views is not our primary objective. Moreover, while multi-view setups provide abundant contextual information, much of it is irrelevant, such as regions lacking co-visibility. This redundancy also imposes an additional burden on learning.

To address this, we propose a geometric constraint mechanism to explicitly guide the attention range of each point. By incorporating relative positional encoding into the attention score computation, we enable the scores to depend on two critical factors: feature correlation and the geometric distance $\Delta\mathbf{p}$ between two points:

$$a_{uv}^{I_i I_j} = \left(\mathbf{q}_u^{I_i}\right)^\top \mathbf{R}\left[\Delta\mathbf{p}_{uv}^{I_i I_j}\right] \mathbf{k}_v^{I_j}. \quad (2)$$

Here, $\mathbf{R}[\cdot]$ is the rotation encoding matrix [49].

While the 2D relative distance between two points within a single image is well-defined, across different views, it depends on the projection transformation. We use the precomputed tracks $\mathcal{M}(\mathcal{G})$ to find the projection position of $\mathbf{p}_u^{I_i}$ on I_j, which can be denoted as $\mathbf{p}_w^{I_j}$. When no corresponding point in I_j can be found for $\mathbf{p}_u^{I_i}$, we simply assign the relative position to $\mathbf{0}$ so that the attention score depends solely on the feature similarity. The calculation of the relative position can be formulated as follows:

$$\Delta\mathbf{p}_{uv}^{I_i I_j} = \begin{cases} \mathbf{p}_w^{I_j} - \mathbf{p}_v^{I_j} & (u, w) \in M(I_i, I_j) \\ \mathbf{0} & (u, \emptyset) \in M(I_i, I_j). \end{cases} \quad (3)$$

Additionally, we also interact with all target view features after source cross-attention (Target Cross). For a point t in view I_t paired with source view I_i, we attend to the same point t across the remaining $M-1$ target views. This is analogous to operations in some works that aggregate features along the temporal dimension [12, 26]. Unlike the approach for source views, we compute the attention scores solely based on feature similarity. Despite its low computational complexity, this module complements the Source Cross, mutually enhancing each other to help the network implicitly infer globally optimal correspondences.

3.3. Multi-view feature correlation

For a track in a source view set originating from a 3D point, its corresponding 2D keypoints $\left\{\mathbf{p}_u^{I_i}, \mathbf{p}_v^{I_j}, \ldots\right\}$ should exhibit consistent matches in the target view—either corresponding to the same single point or having no corresponding point at all. This physical constraint is often utilized by existing methods as a post-processing step to filter out erroneous results when merging multiple two-view correspondences. However, as a multi-view matcher, we aim for CoMatcher to inherently satisfy this constraint in its reasoning, thereby enhancing its reliability and confidence.

Achieving this involves two key processes: (i) identifying points that may lead to errors in the *early* layers of the network, and (ii) promptly guiding these points toward correct estimation using inference information from other views. To accomplish this, we propose a two-step multi-view feature correlation strategy (see Fig. 3).

First, to identify ambiguous points in the source views, we utilize a lightweight head at the end of each layer to predict the confidence of each point:

$$c_u^{I_i} = \mathrm{Sigmoid}\left(\mathrm{MLP}\left(\mathbf{f}_u^{I_i}\right)\right) \in [0, 1]. \quad (4)$$

Here, MLP represents a multi-layer perceptron and $\mathbf{f}_u^{I_i}$ is the feature of u. This confidence score reflects the assessment of each point in its current state: higher values indicate that, at this layer, the point is either converging toward a reliable match or confidently identified as having no match.

In each layer, we define a distinct hyperparameter threshold θ. Points with confidence scores below this threshold are identified as potentially ambiguous and likely to lead to incorrect inference. We progressively increase the value of θ in later layers, as the confidence of overall inference generally improves with each successive layer.

Next, we redesign the parallel two-view cross-attention modules to correct the erroneous estimations of ambiguous

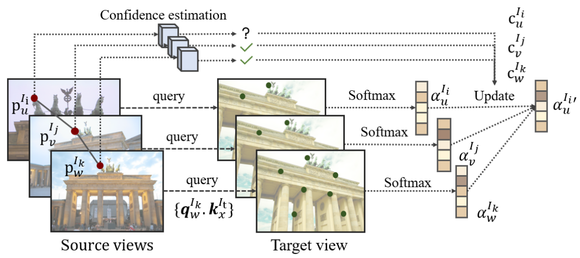

Figure 3. **The multi-view feature correlation strategy** consists of two steps: we first identify potentially erroneous points using a confidence estimator, and then correct their estimation by aggregating attention correlation vectors from other views.

points. In these modules, each point in the source view computes an attention distribution vector to query [54]:

$$\boldsymbol{\alpha}_u^{I_i} = \underset{x \in \mathcal{W}_t}{\mathrm{Softmax}}(a_{ux}^{I_i I_t}). \tag{5}$$

This vector represents the degree to which the point attends to in view I_t, with a peak indicating potential matching regions of interest. For ambiguous points, this similarity is not reliable, often leading to incorrect feature aggregation. Assuming that point u in the source view I_i is an ambiguous point, we update its original attention distribution by applying a weighted average to the distributions of its corresponding points across multi-view:

$$\boldsymbol{\alpha}_u^{I_i\prime} = c_u^{I_i} \boldsymbol{\alpha}_u^{I_i} + (1 - c_u^{I_i}) \frac{\sum_{v \in \mathcal{D}_m^u} c_v^{I_j} \boldsymbol{\alpha}_v^{I_j}}{\sum_{v \in \mathcal{D}_m^u} c_v^{I_j}} \quad \text{s.t. } c_u^{I_i} < \theta. \tag{6}$$

Here, $\mathcal{D}_m^u$ represents other points in the track. We embed this strategy into the model inference process, which significantly enhances the reliability of estimation under challenging conditions despite its simple implementation.

3.4. Correspondence prediction

As shown in Fig. 2, a lightweight matching head operates in parallel for each pair of network output features to predict correspondences between two views [30, 43]. In each head, we begin by calculating a score matrix from feature correlations between two views: $\mathbf{S}(u, x) = \mathrm{L}(\mathbf{f}_u^{I_i}) \cdot \mathrm{L}(\mathbf{f}_x^{I_t})$, where L is the linear transform. Then we apply dual-softmax [30, 50, 53] to derive the matching probabilities of two points: $\mathbf{S}'(u, x) = \mathrm{Softmax}(\mathbf{S}(u, \cdot))_x \cdot \mathrm{Softmax}(\mathbf{S}(\cdot, x))_u$. Additionally, we employ the method proposed in [30] to calculate the matching probability for each point: $\sigma_i^{I_i} = \mathrm{Sigmoid}(\mathrm{L}(\mathbf{f}_i^{I_i}))$. The final assignment matrix is a combination of the two probabilities:

$$\mathbf{P}(u, x) = \mathbf{S}'(u, x) \sigma_u^{I_i} \sigma_x^{I_t}. \tag{7}$$

Using mutual nearest neighbors and a threshold, we filter the matching results from the assignment matrix [43, 50].

3.5. Loss

CoMatcher supervises each source-target image pair individually, with the loss for each pairs composed of two components: correspondence and confidence:

$$\mathcal{L}_{\text{total}} = \frac{1}{M} \sum_{I_i \in \mathcal{G}} (\mathcal{L}_{\text{corr}}(I_i, I_t) + \alpha \mathcal{L}_{\text{conf}}(I_i)). \tag{8}$$

Correspondence supervision. For each pair I_i and I_t, we perform two-view transformations using relative poses or homography to compute ground truth labels [30, 43, 50]. We minimize the negative log-likelihood of the assignment matrix, following [30]. More details are provided in the supplementary materials.

Confidence supervision. To train the confidence estimators at each layer, we define the point confidence as the consistency probability between its correspondence estimated at the current layer and the final estimation [16, 30, 47]. At each layer, predictions are obtained via dual-softmax on current features, without introducing additional matching probabilities or parameters. The ground truth label y indicates whether the two estimations are consistent. This is supervised using a binary cross-entropy loss:

$$\mathcal{L}_{\text{conf}}(I_i) = \frac{1}{L-1} \sum_{\ell} \sum_{u \in \mathcal{W}_i} \mathrm{CE}({}^{\ell}c_u^{I_i}, {}^{\ell}y_u^{I_i}), \tag{9}$$

where $\ell \in \{1, \ldots, L-1\}$.

4. Experiments

In this section, we first introduce the datasets used, followed by our implementation details. Then, our CoMatcher network is compared to previous state-of-the-art baselines for homography and camera pose estimation. Next, we integrate CoMatcher into groupwise matching framework and evaluate it against pairwise method on a large-scale benchmark. Finally, an extensive ablation study is provided.

4.1. Datasets

The **HPatches** dataset is used for homography estimation [3]. It consists of 116 strictly planar scenes, each containing 6 images with variations in viewpoint and illumination. For camera pose estimation tasks, we utilize the **MegaDepth** dataset, an outdoor dataset that exhibits strong occlusions and significant structural changes [28]. We selected two scenes, “Sacre Coeur” and “St. Peter’s Square”, from which 1500 co-visible quadruplets were sampled in a way that balances difficulty based on visual overlap [30, 50, 56]. The **Image Matching Challenge 2020** benchmark provides a comprehensive evaluation protocol, including datasets that cover multiple challenging outdoor wide-baseline scenes. From the phototourism dataset, we selected 3 validated scenes, each consisting of about 100

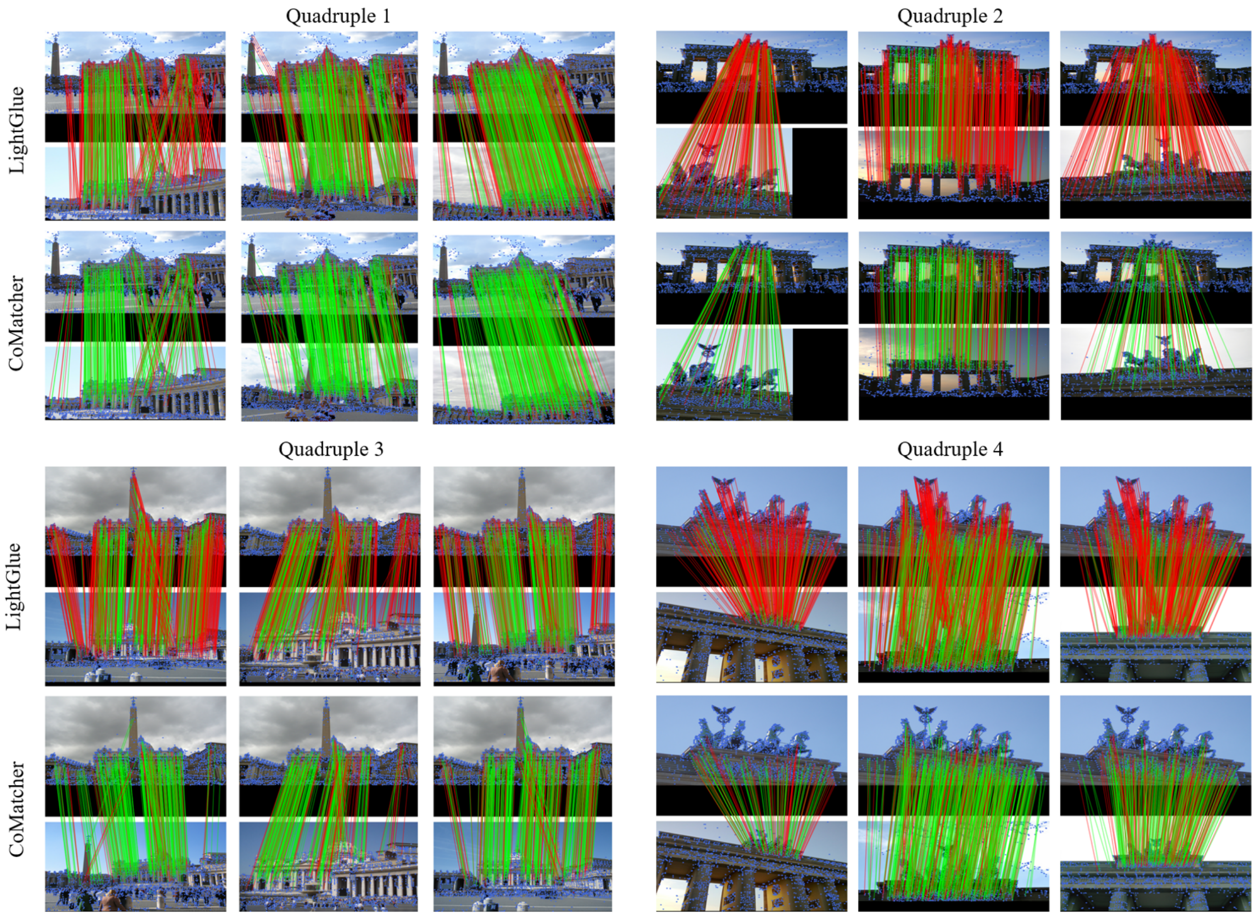

Figure 4. **Qualitative comparison on MegaDepth.** For each quadruple with a target view (top), correspondences (green for correct, red for incorrect) predicted by CoMatcher and LightGlue are shown. Using identical local features as input, CoMatcher achieves significantly more *reliable* estimations, even at challenging semantic edges with depth discontinuities. This stems from its holistic scene understanding through multi-view cues to effectively address occlusions (Q3), large-scale variations (Q2, 4), and repetitive textures (Q1, 2, 3, 4).

images. Additionally, we conducted an ablation study on a synthetic homography dataset [37] used for training.

4.2. Implementation details

CoMatcher is trained in two steps, following [30, 43]. We first pre-train the model on a large-scale synthetic homography dataset [37], leveraging noise-free homography for ground truth. Next, fine-tuning is performed on the MegaDepth dataset [28] using the camera poses recovered by SfM as ground truth. We sample 200 co-visible multi-view groups per scene and randomly select one image as the target view. The size of source views is set to $M = 4$ during training. The training process is carried out on two NVIDIA GeForce 4090 GPUs, taking about 6 days in total. More details can be found in the supplementary materials.

4.3. Homography estimation

CoMatcher is compared with three types of baselines. First, we include state-of-the-art (SOTA) sparse two-view matchers: nearest-neighbor with mutual check [31], SuperGlue [43], SGMNet [7], and LightGlue [30], each paired with different feature extractors [10, 31, 53]. Second, we evaluate the multi-view matcher End2End [40]. For reference, we also compare against typical dense matchers LoFTR [50] and PDC-Net [52]. For CoMatcher and End2End, we use a single forward pass to obtain matches across all five image pairs, while others follow a pairwise approach. To ensure a fair comparison, we adopt the setups of [30, 50] for the number of features and image resizing. Homography accuracy is evaluated using both robust (RANSAC [17]) and non-robust (weighted DLT [20]) solvers. For each pair, we compute the mean reprojection error of the four image corners and report the area under the cumulative error curve (AUC) up to 1px, 3px, and 5px.

Tab. 1 shows that CoMatcher yields more accurate estimates than all sparse two-view matchers, highlighting the advantage of inference in multi-view feature space. How-

	Method	AUC - DLT	AUC - RANSAC
		@1px / @3px / @5px	
dense	LoFTR	**38.5** / 66.0 / 71.4	**40.7** / 68.3 / 78.5
	PDC-Net	36.0 / 65.3 / 73.0	37.9 / 67.6 / 77.4
sparse 2-view	SIFT+NN+mutual	0.0 / 0.0 / 0.0	35.9 / 65.0 / 75.6
	SP+NN+mutual	0.0 / 1.9 / 3.4	34.8 / 64.1 / 74.8
	SP+SuperGlue	32.2 / 65.1 / 75.7	37.2 / 68.0 / 78.7
	SP+SGMNet	31.7 / 64.9 / 76.0	37.7 / 66.4 / 77.5
	SP+LightGlue	35.4 / 67.5 / 77.7	37.2 / 67.8 / 78.1
	DISK+NN+mutual	1.8 / 5.2 / 7.8	37.9 / 58.0 / 68.3
	DISK+LightGlue	34.4 / 64.5 / 74.4	38.1 / 65.2 / 77.2
multi-view	SP+End2End	34.3 / 66.9 / 75.5	37.0 / 67.2 / 77.5
	SP+CoMatcher	37.1 / **69.0 / 78.8**	38.4 / **68.9 / 79.0**
	DISK+CoMatcher	36.3 / 66.2 / 75.9	38.7 / 68.2 / 78.4

Table 1. **Homography estimation on HPatches.** We report the area under the cumulative error curve (AUC) up to values of 1px, 3px and 5px, using DLT and RANSAC [17] as homography solver.

	Method	AUC RANSAC	AUC LO-RANSAC	Time
		5° / 10° / 20°		(ms)
dense	LoFTR	58.0 / 73.1 / 84.4	67.4 / 81.7 / 89.3	182
	PDCNet	54.7 / 73.1 / 83.5	67.1 / 80.2 / 87.0	231
	DUSt3R	42.4 / 56.7 / 64.2	58.1 / 70.4 / 79.6	264
	MASt3R	51.5 / 65.3 / 75.6	63.5 / 76.3 / 85.2	317
SuperPoint	NN+mutual	35.3 / 58.3 / 53.7	51.4 / 67.3 / 75.9	9
	SuperGlue	55.8 / 72.8 / 84.1	65.1 / 77.2 / **89.2**	87
	LightGlue	56.2 / 72.7 / 83.5	67.2 / 80.1 / 88.0	51
	End2End	55.3 / 71.4 / 81.2	67.4 / 81.5 / 87.0	152
	CoMatcher	**57.2 / 73.9 / 84.8**	**68.3 / 82.2** / 89.1	69
DISK	NN+mutual	50.9 / 66.7 / 77.7	64.0 / 79.5 / 87.6	9
	LightGlue	53.2 / 69.2 / 80.2	**68.6** / 80.4 / 87.2	54
	CoMatcher	**54.9 / 71.2 / 82.0**	68.5 / **82.1 / 88.4**	73

Table 2. **Relative pose estimation on MegaDepth.** We report the AUC at 5°, 10°, and 20° using different robust estimator, and the average runtime of matching each quadruplet.

ever, despite End2End also jointly reasoning over multiple views like ours, its performance is notably inferior. This underscores the superiority of a N-to-1 architecture over N-to-N for multi-view representation learning. When compared to LoFTR, we observe slightly lower performance under low thresholds. We attribute this mainly to the limitations of the feature extractor in keypoint non-repeatability and localization errors [29]. Additionally, DLT achieves accuracy close to RANSAC on most metrics, reflecting the high-quality correspondences from CoMatcher.

4.4. Relative pose estimation

Next, we evaluate CoMatcher on challenging wide-baseline scenes, using the same baselines as in Sec. 4.3. For each quadruplet, including a selected target view, we still use a single forward pass for multi-view methods. Additionally, we test two newer dense methods, DUSt3R [55] and MASt3R [27], following their evaluation setups for image resizing. All other settings remain consistent with [30]. The essential matrix is estimated from the correspondences using both vanilla RANSAC [17] and LO-RANSAC [9], following [30]. The pose error is then computed as the maximum angular error of rotation and translation, derived from the decomposition of the essential matrix. We report the AUC at 5°, 10°, and 20°, and record the average runtime of matching each quadruplet on a single 4090 GPU.

Tab. 2 shows that CoMatcher, as a sparse method, comprehensively enhances the estimation performance on different local features [10] and [53]. Importantly, compared to other sparse two-view matchers like LightGlue, a key strength of CoMatcher is its reliability, as illustrated in Fig. 3. This highlights the advantage of multi-view collaborative reasoning in understanding complex 3D structures, such as occlusions, even with suboptimal input local features in these areas. Additionally, compared to End2End, CoMatcher achieves more accurate relative poses without the need for cumbersome end-to-end training. While some dense methods may achieve better accuracy on certain metrics, their efficiency is generally much lower. By employing a space-for-time strategy to minimize inference steps, CoMatcher runs significantly faster than most methods.

4.5. Evaluation on the IMC 2020 benchmark

Next, we evaluate our groupwise framework against pairwise baselines [30, 31, 43] on the Image Matching Challenge (IMC) 2020 benchmark [25]. Given a large-scale unordered image collection, the benchmark requires providing matching results for all image pairs, which are then assessed on two downstream tasks: stereo and multi-view reconstruction. For the stereo task, the accuracy of relative camera poses for each pair is evaluated from correspondences using RANSAC, as in Sec. 4.4. For the multi-view task, all correspondences are fed into COLMAP [45] for SfM, with the final accuracy evaluated based on the estimated multi-view camera poses. We report the AUC at 5° and 10° across both tasks. Additionally, the average runtime is reported, which is the total matching time for the entire image set divided by the number of pairs.

Tab. 3 shows that our framework significantly outperforms existing pairwise approaches in both tasks. This advantage primarily stems from our framework's ability to better leverage the relationships within the original image

	Method	Stereo AUC 5°/10°	Multi-View AUC 5°/10°	Time (ms)
two-view	SIFT+NN+mutual	31.5 / 44.2	57.2 / 68.5	3
	SP+NN+mutual	28.6 / 40.3	52.9 / 63.4	3
	SP+SuperGlue	36.5 / 50.1	62.3 / 74.8	32
	SP+LightGlue	36.8 / 49.4	64.6 / 75.4	17
	DISK+NN+mutual	35.4 / 47.4	59.3 / 70.2	3
	DISK+LightGlue	42.1 / 55.6	65.2 / 76.2	19
m-v	SP+**CoMatcher**	39.1 / 52.4	66.2 / 77.5	25
	DISK+**CoMatcher**	**44.9 / 57.3**	**67.1 / 78.4**	29

Table 3. **Evaluation on the IMC 2020 benchmark.** We report the pose AUC at 5°, 10° for two subtasks, stereo and multi-view reconstruction, along with the average matching runtime.

collection, enabling the generation of higher-quality tracks in such large-scale tasks. Additional comparisons on track quality for SfM tasks are provided in the supplementary material. In terms of efficiency, our approach is faster than SuperGlue, with the viewpoint grouping process accounting for only about 4% of the total matching time.

4.6. Ablation study

Understanding CoMatcher. The CoMatcher network introduces three key components: multi-view cross-attention modules, a geometric constraint mechanism, and a multi-view feature interaction strategy. We validate these design choices by evaluating on two benchmarks: the challenging synthetic homography dataset [37], where precision and recall metrics are employed, and HPatches [3], where the AUC of homography estimation via weighted DLT is measured. For each image, 512 keypoints are extracted.

Tab. 4 shows the ablation results. Excluding the two cross-attention modules for multi-view interaction reduces the model to a series of parallel two-view matchers [30, 43]. This simplification limits the model's ability to integrate multi-view context, leading to a significant performance decline. Without the geometric constraint mechanism, the model performs even worse than in the previous case. This suggests that direct feature aggregation from multi-view is challenging, as the network requires additional learning to distinguish meaningful context effectively. Without multi-view feature correlation, the model loses its ability to guide inference for ambiguous points by leveraging consistency constraints. This highlights the importance of guiding two-view feature correlation in the attention search process, a factor overlooked in previous work.

Impact of the group size. We evaluated the matching performance on the synthetic homography dataset [37] by varying the size of the source view sets used as network input. As shown in Fig. 5, performance initially improves significantly as more source views are integrated into the collaborative inference process. However, when the number of views exceeds 5, the matching capability begins to decline. We hypothesize that this degradation is attributed to the expanded search space and the excessive redundant information introduced by multi-view attention, which collectively increase the learning complexity of the network.

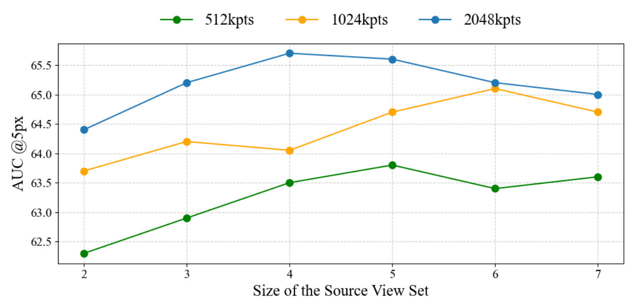

Figure 5. **Impact of the group size.** Under different numbers of keypoints, the AUC initially increases as the source view set expands, then tends to stabilize or shows a downward trend.

CoMatcher	Synthetic precision	Synthetic recall	Hpatches-AUC @1px / @5px
w/o M-V feat. interaction	89.7	96.6	32.4 / 75.1
w/o atten. propagation	87.7	94.2	31.5 / 73.7
w/o M-V feat. correlation	90.5	95.3	33.1 / 74.9
full	**92.7**	**98.9**	**34.7 / 77.1**

Table 4. **Ablation study on synthetic homography dataset and HPatches.** The "full" model is the default model.

5. Conclusion

We proposed a multi-view collaborative matching strategy to address the challenge of reliable matching uncontrolled image sets in complex scenes. Our method, which consists of a deep collaborative matcher (CoMatcher) and a scalable groupwise pipeline, enables a holistic 3D scene understanding while inherently satisfying the cross-view projection consistency constraint. Extensive experiments have demonstrated that exploiting inter-view connections significantly enhances matching certainty, yielding substantial benefits for downstream tasks such as pose estimation and SfM. We hope this work encourages the research community to expand beyond pairwise matching and further explore the understanding and utilization of multi-view information.

Acknowledgement. This research is supported by NSFC projects under Grant 42471447, Development Program of China under Grant 2024YFC3811000, and the Fundamental Research Funds for the Central Universities of China under Grant 2042022dx0001.

References

[1] Sameer Agarwal, Noah Snavely, Ian Simon, Steven M Seitz, and Richard Szeliski. Building Rome in a day. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 72–79. IEEE, 2009. 1, 2

[2] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5297–5307, 2016. 3

[3] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5173–5182, 2017. 2, 5, 8

[4] Axel Barroso-Laguna, Sowmya Munukutla, Victor Adrian Prisacariu, and Eric Brachmann. Matching 2D images in 3D: Metric relative pose from metric correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4852–4863, 2024. 2

[5] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up robust features. In *European Conference on Computer Vision*, pages 404–417. Springer, 2006. 2

[6] Ruojin Cai, Joseph Tung, Qianqian Wang, Hadar Averbuch-Elor, Bharath Hariharan, and Noah Snavely. Doppelgangers: Learning to disambiguate images of similar structures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 34–44, 2023. 1

[7] Hongkai Chen, Zixin Luo, Jiahui Zhang, Lei Zhou, Xuyang Bai, Zeyu Hu, Chiew-Lan Tai, and Long Quan. Learning to match features with seeded graph matching network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6301–6310, 2021. 2, 6

[8] Hongkai Chen, Zixin Luo, Lei Zhou, Yurun Tian, Mingmin Zhen, Tian Fang, David Mckinnon, Yanghai Tsin, and Long Quan. ASpanFormer: Detector-free image matching with adaptive span transformer. In *European Conference on Computer Vision*, pages 20–36. Springer, 2022. 2

[9] Ondřej Chum, Jiří Matas, and Josef Kittler. Locally optimized RANSAC. In *Pattern Recognition*, pages 236–243. Springer, 2003. 7

[10] 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. 2, 6, 7

[11] Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. TAP-Vid: A benchmark for tracking any point in a video. *Advances in Neural Information Processing Systems*, 35:13610–13626, 2022. 2

[12] Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. TAPIR: Tracking any point with per-frame initialization and temporal refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10061–10072, 2023. 2, 4

[13] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten 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*, pages 8092–8101, 2019. 2

[14] Johan Edstedt, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. DeDoDe: Detect, don't describe—Describe, don't detect for local feature matching. In *International Conference on 3D Vision*, pages 148–157. IEEE, 2024. 2

[15] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19790–19800, 2024. 1, 2

[16] Maha Elbayad, Jiatao Gu, Edouard Grave, and Michael Auli. Depth-adaptive Transformer. In *International Conference on Learning Representations*, pages 1–14, 2020. 5

[17] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *In Communications of the ACM*, 24(6):381–395, 1981. 6, 7

[18] Jan-Michael Frahm, Pierre Fite-Georgel, David Gallup, Tim Johnson, Rahul Raguram, Changchang Wu, Yi-Hung Jen, Enrique Dunn, Brian Clipp, Svetlana Lazebnik, et al. Building Rome on a cloudless day. In *European Conference on Computer Vision*, pages 368–381. Springer, 2010. 1

[19] Adam W Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle video revisited: Tracking through occlusions using point trajectories. In *European Conference on Computer Vision*, pages 59–75. Springer, 2022. 2

[20] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 1, 6

[21] Xingyi He, Jiaming Sun, Yifan Wang, Sida Peng, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Detector-free structure-from-motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21594–21603, 2024. 2

[22] Martin Humenberger, Yohann Cabon, Nicolas Guerin, Julien Morat, Vincent Leroy, Jérôme Revaud, Philippe Rerole, Noé Pion, Cesar de Souza, and Gabriela Csurka. Robust image retrieval-based visual localization using kapture. *arXiv preprint arXiv:2007.13867*, 2020. 1

[23] Heinly Jared, Johannes L Schonberger, Enrique Dunn, and Jan-Michael Frahm. Reconstructing the world in six days. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015. 1

[24] Hanwen Jiang, Arjun Karpur, Bingyi Cao, Qixing Huang, and André Araujo. OmniGlue: Generalizable feature matching with foundation model guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19865–19875, 2024. 2

[25] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image matching across wide baselines: From paper to practice. *International Journal of Computer Vision*, 129(2):517–547, 2021. 1, 2, 7

[26] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Co-Tracker: It is better to track together. In *European Conference on Computer Vision*, pages 18–35. Springer, 2024. 2, 4

[27] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024. 1, 2, 7

[28] Zhengqi Li and Noah Snavely. MegaDepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2041–2050, 2018. 5, 6

[29] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-perfect structure-from-motion with featuremetric refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5987–5997, 2021. 1, 7

[30] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17627–17638, 2023. 1, 2, 5, 6, 7, 8

[31] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004. 1, 2, 6, 7

[32] Jiayi Ma, Xingyu Jiang, Aoxiang Fan, Junjun Jiang, and Junchi Yan. Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision*, 129(1):23–79, 2021. 1, 2

[33] Eleonora Maset, Federica Arrigoni, and Andrea Fusiello. Practical and efficient multi-view matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4568–4576, 2017. 2

[34] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. ORB-SLAM: A versatile and accurate monocular SLAM system. *In IEEE Transactions on Robotics*, 31(5): 1147–1163, 2015. 1

[35] Carl Olsson and Olof Enqvist. Stable structure-from-motion for unordered image collections. In *Image Analysis*, pages 524–535. Springer, 2011. 2

[36] Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes L Schönberger. Global structure-from-motion revisited. In *European Conference on Computer Vision*, 2024. 2

[37] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting Oxford and Paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5706–5715, 2018. 6, 8

[38] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning CNN image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1655–1668, 2018. 3

[39] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2D2: Reliable and repeatable detector and descriptor. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[40] Barbara Roessle and Matthias Nießner. End2end multi-view feature matching with differentiable pose optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 477–487, 2023. 2, 6

[41] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An efficient alternative to SIFT or SURF. In *Proceeding of the IEEE/CVF International Conference on Computer Vision*, pages 2564–2571. Ieee, 2011. 2

[42] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12716–12725, 2019. 1, 2

[43] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4938–4947, 2020. 1, 2, 4, 5, 6, 7, 8

[44] Torsten Sattler, Akihiko Torii, Josef Sivic, Marc Pollefeys, Hajime Taira, Masatoshi Okutomi, and Tomas Pajdla. Are large-scale 3d models really necessary for accurate visual localization? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1637–1646, 2017. 2

[45] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-Motion revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016. 1, 2, 7

[46] Johannes L Schonberger, Hans Hardmeier, Torsten Sattler, and Marc Pollefeys. Comparative evaluation of hand-crafted and learned local features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1482–1491, 2017. 2

[47] Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Tran, Yi Tay, and Donald Metzler. Confident adaptive language modeling. *Advances in Neural Information Processing Systems*, 35:17456–17472, 2022. 5

[48] Sivic and Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proceedings in the IEEE/CVF International Conference on Computer Vision*, pages 1470–1477. IEEE, 2003. 3

[49] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 4

[50] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8922–8931, 2021. 1, 2, 5, 6

[51] Kun Sun, Jinhong Yu, Wenbing Tao, Xin Li, Chang Tang, and Yuhua Qian. A unified feature-spatial cycle consistency fusion framework for robust image matching. *Information Fusion*, 97:101810, 2023. 2

[52] Prune Truong, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning accurate dense correspondences and when to trust them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5714–5724, 2021. 2, 6

[53] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. DISK: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33:14254–14265, 2020. 2, 5, 6, 7

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 1, 5

[55] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 7

[56] Yifan Wang, Xingyi He, Sida Peng, Dongli Tan, and Xiaowei 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*, pages 21666–21675, 2024. 2, 5

[57] Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. SpatialTracker: Tracking any 2D pixels in 3D space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20406–20417, 2024. 1

[58] Shibiao Xu, Shunpeng Chen, Rongtao Xu, Changwei Wang, Peng Lu, and Li Guo. Local feature matching using deep learning: A survey. *Information Fusion*, 107:102344, 2024. 1

[59] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. LIFT: Learned invariant feature transform. In *European Conference on Computer Vision*, pages 467–483. Springer, 2016. 2

[60] Guofeng Zhang, Zilong Dong, Jiaya Jia, Tien-Tsin Wong, and Hujun Bao. Efficient non-consecutive feature tracking for structure-from-Motion. In *European Conference on Computer Vision*, pages 422–435. Springer, 2010. 2