

# PRISM: PRogressive dependency maxImization for Scale-invariant image Matching

Xudong Cai  
Renmin University of China  
Beijing, China  
HAOMO.AI  
Beijing, China  
xudongcai@ruc.edu.cn

Yongcai Wang\*  
Renmin University of China  
Beijing, China  
ycw@ruc.edu.cn

Lun Luo  
HAOMO.AI  
Beijing, China  
luolun@haomo.ai

Minhang Wang  
HAOMO.AI  
Beijing, China  
wangminhang@haomo.ai

Deying Li  
Renmin University of China  
Beijing, China  
deyingli@ruc.edu.cn

Jintao Xu  
HAOMO.AI  
Beijing, China  
xujintao@haomo.ai

Weihao Gu  
HAOMO.AI  
Beijing, China  
guweihao@haomo.ai

Rui Ai  
HAOMO.AI  
Beijing, China  
airui@haomo.ai

## Abstract

Image matching aims at identifying corresponding points between a pair of images. Currently, detector-free methods have shown impressive performance in challenging scenarios, thanks to their capability of generating dense matches and global receptive field. However, performing feature interaction and proposing matches across the entire image is unnecessary, because not all image regions contribute to the matching process. Interacting and matching in unmatchable areas can introduce errors, reducing matching accuracy and efficiency. Meanwhile, the scale discrepancy issue still troubles existing methods. To address above issues, we propose PRogressive dependency maxImization for Scale-invariant image Matching (PRISM), which jointly prunes irrelevant patch features and tackles the scale discrepancy. To do this, we firstly present a Multi-scale Pruning Module (MPM) to adaptively prune irrelevant features by maximizing the dependency between the two feature sets. Moreover, we design the Scale-Aware Dynamic Pruning Attention (SADPA) to aggregate information from different scales via a hierarchical design. Our method's superior matching performance and generalization capability are confirmed by leading accuracy across various evaluation benchmarks and downstream tasks. The code is publicly available at <https://github.com/Master-cai/PRISM>.

\*Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

MM '24, October 28–November 1, 2024, Melbourne, VIC, Australia.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 979-8-4007-0686-8/24/10  
<https://doi.org/10.1145/3664647.3681209>

## CCS Concepts

- Computing methodologies → Vision for robotics; Matching

## Keywords

Image Matching, Patch Pruning, Scale-Invariant, Detector-free

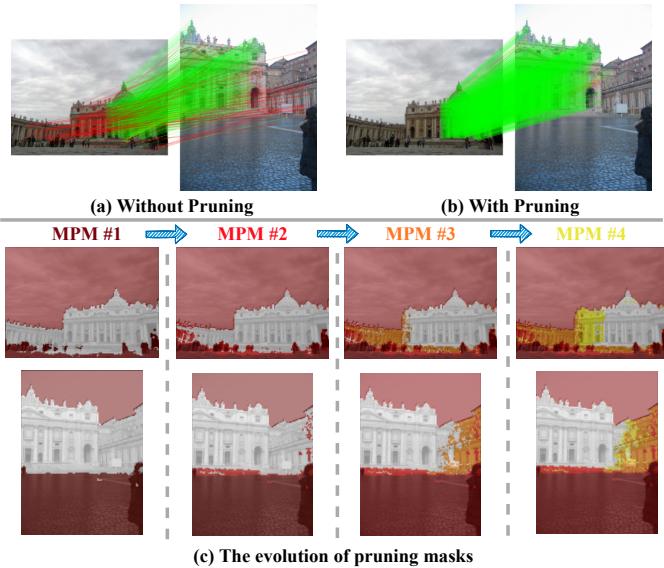
## ACM Reference Format:

Xudong Cai, Yongcai Wang, Lun Luo, Minhang Wang, Deying Li, Jintao Xu, Weihao Gu, and Rui Ai . 2024. PRISM: PRogressive dependency maxImization for Scale-invariant image Matching. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28–November 1, 2024, Melbourne, VIC, Australia*. ACM, New York, NY, USA, 15 pages. <https://doi.org/10.1145/3664647.3681209>

## 1 Introduction

Image matching, a pivotal task in computer vision [26, 27], finds extensive applications in areas as image stitching [35, 47], visual localization [7, 55, 57, 76, 77], 3D reconstruction [60, 61], etc. Traditionally, it involves identifying corresponding points across two images, which is a detector-based paradigm [3, 39, 54]. This paradigm follows a three-step pipeline: (1) keypoint detection, (2) keypoint description, and (3) matching based on descriptor similarity. Although detector-based methods are generally effective and efficient, they falter in complex environments such as poor texture, repetitive patterns and significant viewpoint changes, where keypoint detection may not yield sufficient keypoints.

To address the limitation, researchers propose deep learning-based methods to improve the reliability of traditional pipeline. Superpoint [16], Lift [79] and some other works [17, 41, 50] enhance the repeatability and distinguishability of the keypoints. Super-Glue [58] and its variations [37, 62] propose transformer-based [72] matchers, jointly matching sparse points and rejecting outliers. Although these works gain improvements in more reliably matching, they still share the limitations of the detector-based paradigm. The keypoint detection is generally a bottleneck and the performances



**Figure 1: The basic idea of our proposed methods.** Given two images, not all image patches are helpful to the matching process. Conducting feature interactions and searching matches across the entire image can be detrimental (Without Pruning). We propose to gradually prune the irrelevant patches by maximizing the dependency between two images, resulting in more robust and accurate matches (With Pruning). (c) shows the pruning masks estimated by successive MPM. MPM prunes irrelevant patches from shallow to deep layers successively ■ → ■ → ■ → ■. Feature interactions and match searches are only conducted in the white mask regions.

in some complex environments (e.g., poor texture, repetitive patterns and significant viewpoint changes) are unsatisfactory.

To address these issues, another line of works [33, 51, 52] abandons the keypoint detection, so-called detector-free methods. They first use CNN [36] to extract image features. In order to generate dense matches, they exhaustively search for matches across entire feature maps. Potential patch-level matches are proposed from the whole feature map by constructing correlation cost volume between two feature maps. The coarse matches are then filtered by threshold and Mutual Nearest Neighborhood criterion. Finally, the valid coarse matches are further refined into pixel-level matches. Considering the feature maps generated by CNN have a limited receptive field, LoFTR [65] and its variants [11, 21, 80] utilizes transformer backbone [29] to model long-range dependencies, resulting in better feature matching performance.

While detector-free methods address the limitations of detector-based methods, these techniques still encounter significant challenges. On one hand, detector-free methods suffer from irrelevant feature interference. Feature interactions and match searches in unnecessary areas may lead to erroneous matches, as shown in Figure 1. Notably, for a pair of images divided into  $M$  and  $N$  patches, this matching paradigm proposes  $M \times N$  potential matches. However, the maximum viable matches are only  $\max(M, N)$  in theory. This discrepancy suggests that a significant proportion of these matches are erroneous (such as matches in unmatchable areas like sky and clouds or non-overlapping regions). TopicFM [21] and

AdaMatcher [25] utilize semantic and co-visible information to filter match proposals. But the feature interactions still span the entire image. HCPM [13] keeps a fixed proportion of significant patches but may discard useful ones in overlapping scenarios. Efficient LoFTR [78] uses Max-pooling for pruning but can still include harmful patches. SGAM [83] and MESA [82] propose to matches areas with similar semantics before establishing point matches, but they need prior image segmentation.

On the other hand, detector-free methods fall short in addressing the scale discrepancy. As detector-free methods generate matches in pixel-level, exhaustively searching matches across different scales can bring unbearable computational costs. Although ASTR [80] and AdaMatcher [25] attempt to estimate the scale variation between images via patch-level matching and adjust the patch size in fine matching process based on the estimated scale ratio, the patch-level matching itself can also be erroneous due to the scale problem. PATS [46] introduces an innovative yet time-intensive framework that iteratively rescales and segments images into smaller, scale-consistent patches to mitigate the scale issue, but the time consumption is significantly increased.

To deal with the above challenges, we propose a novel framework, PRogressive dependency maxImization for Scale-invariant image Matching (PRISM), which simultaneously prunes irrelevant patches and tackles scale difference. The basic idea is illustrated in Figure 1. The key innovation is to prune unnecessary image patches adaptively and gradually, and model the scene of various scales simultaneously within the same attention mechanism. To eliminate the interference of the irrelevant features, we propose Multi-scale Pruning Module (MPM) to dynamically prune irrelevant features by gradually maximizing the dependency between the two feature sets, where the dependency is usually measured by Mutual Information. By pruning irrelevant features gradually, the computational resources can be focused on those features that are most informative and relevant for matching. In addition, to solve the scale discrepancy, we propose a novel Scale-Aware Dynamic Pruning Attention (SADPA) mechanism, which injects a scale space analysis into the attention mechanism via a hierarchical design and focuses attention on the selected features. This scheme gives SADPA favorable computational efficiency alongside the ability to model multi-scale features.

Experimental evaluations show that PRISM sets a new state-of-the-art (SOTA), surpassing both detector-based and detector-free baselines across various tasks, such as homography estimation, relative pose estimation, and visual localization. Experiments also showcase PRISM’s robust generalization capabilities. Our ablation studies verify the effectiveness of the proposed MPM and SADPA. The key contributions are as following:

- A novel image matching framework PRISM is proposed, employing a Multi-scale Pruning Module (MPM) to aggregate information in different scales and prune irrelevant features by maximizing the dependency gradually.
- A Scale-Aware Dynamic Pruning Attention (SADPA) is proposed, which dynamically adjusts the attention focus and aggregates information across multiple scales.
- PRISM is demonstrated to achieve state-of-the-art results across a comprehensive set of benchmarks, showcasing its robust generalization capabilities across diverse datasets.

## 2 Related Work

### 2.1 Detector-based image matching

Detector-based image matching has been studied for several decades. It involves identifying keypoints in images and finding their correspondences across different views. The evolution begins with foundational techniques like the Harris Corner Detector [23], advancing to sophisticated systems such as SIFT [39] for improved scale and rotation resilience. Developments in detection speed and reliability followed [3, 53, 54]. With the success of CNN in the field of image processing [24, 63, 66], numerous researchers have begun to employ CNNs for keypoint detection and description, achieving impressive results. SuperPoint [16] jointly trains the detector and descriptor, largely improving the accuracy and robustness of matching. Many works follow this line, further improving the reliability and uniqueness of the keypoints [17, 32, 40, 41, 50, 71, 75].

After detecting and describing keypoints, matching typically involves a nearest neighbor search in descriptor space, complemented by heuristic filters like Lowe's ratio test [39]. However, this approach struggles under challenging conditions. SuperGlue [58] addresses this by employing an Attentional Graph Neural Network to jointly match features and filter outliers. However, it has quadratic complexity with the number of keypoints due to the attention mechanism. Subsequent efforts, such as SGMNet [10], ClusterGNN [62], and LightGlue [37], have aimed to reduce computational load through strategies like initializing with a subset of reliable matches, dividing keypoints into subgraphs, and adapting network depth and width based on image pair difficulty. However, the dependency on keypoint and descriptor repeatability limits the robustness of detector-based methods against extreme variations in viewpoint, repetitive patterns, and texture-less surfaces.

### 2.2 Detector-free image matching

Detector-free image matching methods eschew the keypoint detection step and directly generate dense matches from the image. Early works are cost volume-based, such as NCNet series [33, 51, 52]. They use CNN to extract dense feature maps and construct 4-D cost volume to exhaust all potential matches. Although they have made some progress, the receptive field of CNN is limited, and the image resolution is restricted due to the expensive computational cost. Recently, LoFTR [65] extends the limited receptive field to global consensus with the help of the global receptive field and long-range dependencies of Transformers [72]. However, LoFTR and its successors [8, 19, 78, 86] still encounter significant challenges, particularly regarding the scale disparity problem and the distraction issue of linear attention. AdaMatcher [25] and ASTR [80] estimate the scale variation via coarse matching results and resize the patch size by scale ratio, but they ignore that the coarse level matching itself can be erroneous due to the scale problem. PATS [46] models the scale problem as a patch area transportation problem and designs an iterative framework to find matches from coarse to fine, but the time cost is unacceptable. ASpanFormer [11] uses estimated optical flow to guide the attention span. ASTR [80] proposes a spot-guided attention framework to restrict the feature aggregations. Quadtree attention [69] builds token pyramids and computes attention in a coarse-to-fine manner. However, they neglect the comprehensive understanding of scene semantics and contextual

relationships, achieving limited enhancements. TopicFM [21] constrains matches to regions with identical semantics, but it does not account for completely unmatchable categories (e.g., sky, clouds) or non-overlapping regions. Other works [28, 68] that try to directly decode the coordinates of matching keypoints are also related.

### 2.3 Mutual Information based Feature Selection

Feature selection involves choosing a subset of the available features based on specific criteria to eliminate irrelevant, redundant, or noisy features, which is a crucial aspect of data mining. The use of Mutual Information (MI) to assess the dependency among features for the purpose of feature selection was first introduced in [2], referred to as Mutual Information Maximization. MI assesses the information contribution of variables towards the learning goal. Several methods [5, 42] leveraging MI for feature selection have been proposed to enhance performance across diverse learning tasks. One notable approach is the minimal-redundancy-maximal-relevance (mRMR) criterion [49], which uses average MI as a criterion and selects features with a trade-off between dependency and redundancy of the selected features. Further advancing the field, the Normalized Mutual Information Feature Selection (NMIFS) technique was introduced to mitigate MI's bias towards multi-valued features by normalizing MI (NMI) values to a [0,1] range [18]. This method uses average NMI to select features, improving upon the standard MI approach by addressing its inherent bias. Subsequent methods [20, 22, 74, 74] are proposed to enhance the performance.

MI distinguishes itself from other dependency measures by its ability to quantify any relationship between variables and its stability under space transformations like rotations and translations [18, 30]. This makes MI-based feature selection particularly appealing for patch feature pruning in the context of image matching. It underpins our study's innovative patch pruning technique, aimed at eliminating irrelevant patch features.

## 3 Method

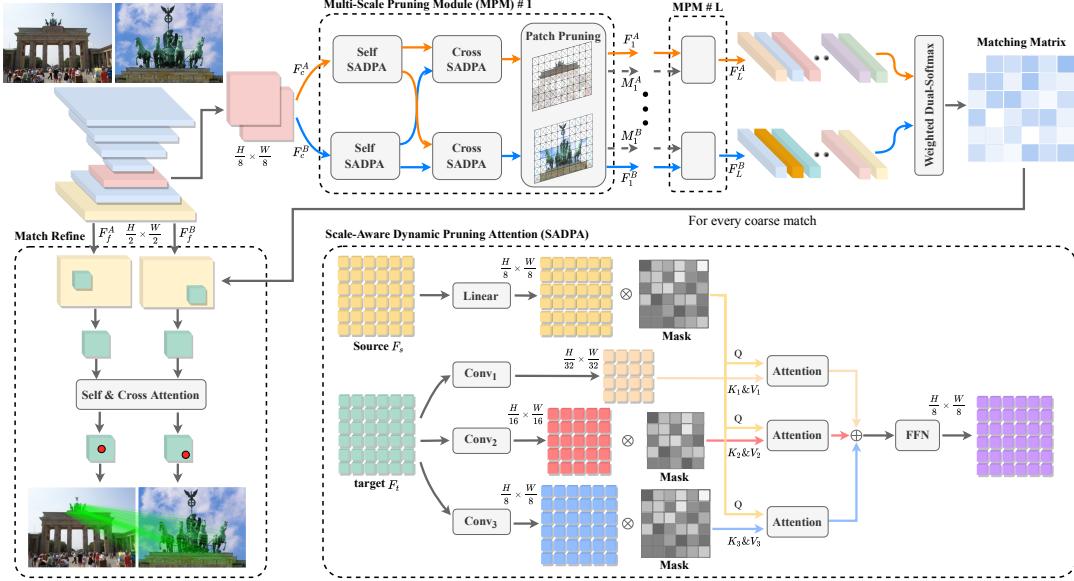
### 3.1 Overview

Given a pair of images, our task is to identify reliable matches across images. To address the scale discrepancy and reduce the interference of irrelevant features, we propose to gradually eliminate redundant image patches through adaptive pruning and simultaneously model scenes of various scales within the same attention framework. The overview of PRISM is shown in Figure 2. Mathematically, for image  $I^A$  and  $I^B$ , PRISM generates matches as in Algorithm 1:

In PRISM, each **for** loop constitutes a Multi-scale Pruning Module (MPM), and there are total  $L$  MPMs. Each MPM takes the last MPM's output, i.e.,  $F_{l-1}^A, F_{l-1}^B, M_{l-1}^A, M_{l-1}^B$  as input and conducts  $f_{\text{transform}}$  and dependency maximization.  $M_l^A$  is the mask to select the best subset  $\hat{F}_l^A$ :  $\hat{F}_l^A = M_l^A \otimes F_l^A$ , that maximizes the dependency  $D(\cdot)$  between the two feature sets.  $M_l^B$  is defined in a similar way.

### 3.2 Multi-scale Pruning Module

Given the coarse-level features maps  $F_c^A$  and  $F_c^B$  at  $\frac{1}{8}$  resolution, the Multi-scale Pruning Module extracts multi-scale features and progressively eliminates irrelevant features. Specifically, in each MPM, the coarse feature maps are first transformed by  $f_{\text{transform}}(\cdot)$ ,



**Figure 2: Overview of PRISM.** PRISM starts from a CNN-based backbone to extract coarse-level  $F_c^A, F_c^B$  and fine-level features  $F_f^A, F_f^B$ .

$F_c^A, F_c^B$  are fed into the proposed iterative Multi-scale Pruning Module (MPM) for updating and pruning (Sec. 3.2). In each MPM layer, the features are first transformed by the self- and cross- SADPA with a hierarchical design (Sec. 3.2.2) to aggregate information from selected features of various scales. Then the Patch Pruning module (Sec. 3.2.1) eliminates irrelevant features to maximize the NMI between the two feature sets. After  $L$  MPM blocks, the final features  $F_L^A$  and  $F_L^B$  are used to acquire the coarse matching Matrix by Weighted Dual-softmax (Sec. 3.3). Finally, we use the mutual nearest neighbor strategy and the threshold  $\theta_c$  to filter the valid coarse matches  $M_c$ . Then  $M_c$  are projected to fine level features maps  $F_f^A, F_f^B$  and refined to sub-pixel precision matches  $M_f$ .

#### Algorithm 1 PRISM: PRogressive dependency maxImization for Scale-invariant image Matching

```

1: Input: a pair of images  $I^A$  and  $I^B$ 
2: Output: Matches  $M_f$ 
3:  $F_c^A, F_c^B, F_f^A, F_f^B = \text{CNN}(I^A, I^B)$ 
4:  $M_0^A, M_0^B \leftarrow$  Masks of all 1 with the same size of  $F_c^A, F_c^B$ 
5:  $F_0^A, F_0^B = F_c^A, F_c^B$ 
6: for  $l = 1$  to  $L$  do
7:    $F_l^A, F_l^B = f_{\text{transform}}(F_{l-1}^A, F_{l-1}^B, M_{l-1}^A, M_{l-1}^B)$ 
8:    $M_l^A = \text{argmax } D(\hat{F}_l^A, F_l^B), M_l^B = \text{argmax } D(\hat{F}_l^B, F_l^A)$ 
     $M_l^A$   $M_l^B$ 
9: end for
10:  $M_c = f_{\text{matching}}(F_L^A, F_L^B)$ 
11:  $M_f = f_{\text{refine}}(M_c, F_f^A, F_f^B)$ 
12: return  $M_f$ 
```

which consists of a self SADPA and a cross SADPA. Then, the Patch Pruning module eliminates irrelevant features by maximizing the dependency between the two feature sets, resulting in two pruning masks for future use. We first introduce the Patch Pruning in MPM.

**3.2.1 Patch Pruning.** As stated in Introduction (Sec. 1), existing methods [8, 11, 25, 65, 80] may search matches and perform feature interactions in unmatchable areas. It harms the matching accuracy and increases the computational cost. A pruning module is needed to exclude irrelevant patch features.

Inspired by the Mutual Information based Feature Selection methods [18, 49], we innovatively propose to identify the most characterizing features by maximizing the dependency between the two feature sets in the context of image matching. Specifically, in  $l$ th MPM layer, the patch pruning module takes the transformed feature maps  $F_l^A$  and  $F_l^B$  as input, where  $|F_l^A| = M$  and  $|F_l^B| = N$ . For  $F_l^A$ , our target is to find a feature subset  $\hat{F}_l^A \subseteq F_l^A$  which has the largest dependency on  $F_l^B$ :

$$M_l^A = \underset{\hat{F}_l^A}{\operatorname{argmax}} D(\hat{F}_l^A, F_l^B) \quad (1)$$

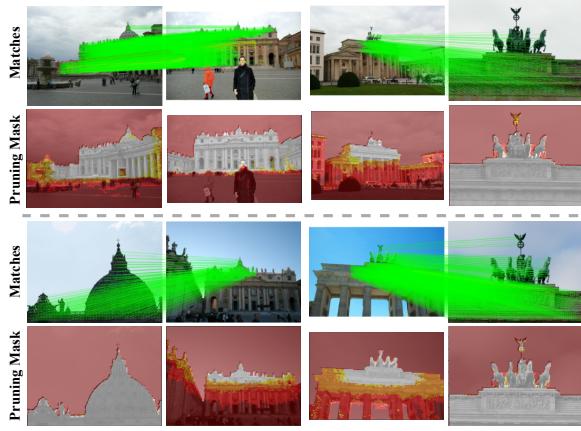
The dependency  $D(\cdot)$  is usually characterized in terms of mutual information (MI) as follows:

$$D(\hat{F}_l^A, F_l^B) = I(\{\hat{F}_l^A(i)|i=1, \dots, m\}; F_l^B), \hat{F}_l^A(i) \in \hat{F}_l^A \quad (2)$$

where  $F_l^B$  can be treated as a multivariate variable and  $I(\cdot)$  denotes the MI. Given two random variables  $X, Y$ , the MI is defined as:  $I(X; Y) = D_{KL}(p(x, y)||p(x)p(y))$ .  $D_{KL}$  is the KL-divergence between  $p(x, y)$  and  $p(x)p(y)$ , which represents the joint distribution and product of the marginal distributions of  $x$  and  $y$ , respectively.

However, the Max-Dependency is hard to implement in high-dimensional space, and searching the feature subspaces exhaustively costs  $O(2^M)$  [49]. An alternative is to select features based on Max-Relevance, which approximates the Max-Dependency with the mean value of all mutual information values as in [49]:

$$D(\hat{F}_l^A, F_l^B) \approx \frac{1}{|\hat{F}_l^A|} \sum_{\hat{F}_l^A(i) \in \hat{F}_l^A} I(\hat{F}_l^A(i); F_l^B). \quad (3)$$



**Figure 3: Visualization of pruning masks and the matching results on MegaDepth dataset.** The Patch pruning can identify redundant image patches and exclude them from subsequent feature interactions gradually (from shallow to deep layers ■ → ■ → ■ → ■). It avoids most incorrect matches.

According to this equation, we can increase the Max-Dependency by eliminating patch features with low MI. The MI between a patch feature  $\hat{F}_l^A(i)$  and another feature set  $F_l^B$  can be expressed compactly in terms of multi-information as in [44, 73]. To standardize the measure of shared information between variables, we utilize Normalized MI (NMI) as in [18]:

$$\text{NMI}(\hat{F}_l^A(i); F_l^B) = 2 \frac{I(\hat{F}_l^A(i); F_l^B)}{H(\hat{F}_l^A(i)) + H(F_l^B)} \in [0, 1] \quad (4)$$

$\text{NMI}(\hat{F}_l^A(i); F_l^B)$  quantifies the amount of information that  $\hat{F}_l^A(i)$  shares with  $F_l^B$ . This concept is intrinsically linked to the fundamental principle of matching, where a pair of features is considered to be matched when they describe the same scene, signifying that their information is shared. However, MI and NMI have historically been difficult to compute [48]. Exact computation is only tractable for discrete variables or a limited family of problems where the probability distributions are known. Inspired by MINE [4], we propose to train a neural network to estimate the NMI:

$$\text{NMI}(\hat{F}_l^A(i); F_l^B) \approx \text{Sigmoid}(\Phi_l(f_{\text{transform}}(F_{l-1}^A, F_{l-1}^B, M_{l-1}^A, M_{l-1}^B)(i))) \quad (5)$$

where  $\Phi_l$  represents an MLP at the last of the  $l$ th layer MPM module. To maximize the dependency, we prune the features whose NMI is lower than a threshold  $\theta_p = 0.05$ . The mask for the locations of the removed features is set to 0, resulting in the updated pruning mask  $M_l^A$ .  $M_l^B$  is obtained in a similar way.

The Patch Pruning module plays a critical role by effectively reducing the search space for potential matches, enhancing both accuracy and efficiency. The structure and computational flow of these components ensure that as the matching process evolves, the algorithm becomes increasingly focused on the most promising match candidates. Figure. 3 shows the effect of this module. We can see most of the invaluable regions such as the sky with less mutual information are pruned. This improves both the precision and computational efficiency of the image matching process.

### 3.2.2 Scale-Aware Dynamic Pruning Attention.

**self and cross SADPA.** In each MPM, we use a succession of one self SADPA and one cross SADPA to update the features, as shown in Figure 2. The  $l$ th MPM takes feature maps  $F_{l-1}^A, F_{l-1}^B$  and masks  $M_{l-1}^A, M_{l-1}^B$  as input, where  $F_{l-1}^A, F_{l-1}^B$  are first input to two self-SADPAs respectively. The cross SADPA takes the output of two self SADPAs as input to further interact features across images. For clarity of presentation, we denote the input feature maps of SADPA as  $F_s$  and  $F_t$  respectively. Thanks to this design, MPM not only enhances the model’s ability to capture intra-image information but also broadens the understanding of inter-image relationships.

**Preliminaries.** Before delving into SADPA, let’s briefly introduce the commonly used vanilla attention. The vanilla attention takes three input vectors: Query  $Q$ , Key  $K$  and value  $V$ . The  $Q$  queries information from the  $K - V$  pairs according to the similarity matrix between  $Q$  and  $K$ :  $\text{Attention}(Q, K, V) = \text{softmax}(QK^T)V$ . However, the size of the weight matrix  $\text{softmax}(QK^T)$  increases quadratically with the image size and the computational cost is unacceptable. As a result, directly applying scale-space analysis by conducting attention between features of different scales is infeasible. The absence of the scale-space analysis impedes the capability of the models to capture multi-scale scenes. Although existing methods [11, 21, 65] optimize the quadratic complexity to linear using linear attention [29], it comes at the cost of sacrificing representational capability [6] and matching accuracy [69, 80]. Therefore, we introduce our Scale-Aware Dynamic Pruning Attention, which injects the scale space analysis into the attention mechanism.

**SADPA.** The design of SADPA is shown in the right-down part of Figure. 2, which performs attention at different scales in parallel. SADPA takes the source and target feature maps  $F_s$  and  $F_t$  and their corresponding pruning masks  $M_s$  and  $M_t$  as input. SADPA uses a Linear module to project  $F_s$  into Query and then trims unnecessary source features by the mask  $M_s$ :

$$Q = \text{Linear}(F_s) \otimes M_s \quad (6)$$

where  $\otimes$  denotes element-wise mask operation and  $\text{Linear}(\cdot)$  is a learned linear transformation. SADPA captures multi-scale features by downsampling the target feature map  $F_t$  to construct a 3-level feature pyramid using convolution layers with varying kernel sizes and strides. Specifically,  $F_t$  is reduced to  $\frac{1}{32}$  resolution to be the coarsest Key and Value. No pruning mask is applied to model the long-range dependencies and large scenes:

$$K_i = V_i = \text{Conv}_i(F_t), i = 1 \quad (7)$$

where  $\text{Conv}_i$  means a convolutional operator with kernel size and stride of  $r_i$ . The other two layers in the pyramid are downsampled into  $\frac{1}{16}$  and  $\frac{1}{8}$  resolution separately to encode the local neighborhood constraints and small scenes. In order to reduce the computational cost and avoid disruption caused by irrelevant features, pruning masks is applied to prune irrelevant features. We down-sample the pruning masks to the same size as the feature maps:

$$K_i = V_i = \text{Conv}_i(F_t) \otimes \text{down}_i(M_t), i \in \{2, 3\} \quad (8)$$

where  $\text{down}_i$  denotes nearest-neighbor interpolation by the ratio  $r_i$ . Then, the attention is performed at different scales and the retrieved messages  $m_i$  from different scales are concatenated and fused with an FFN to update the source features:

$$m_i = \text{Attention}(Q, K_i, V_i), i \in \{1, 2, 3\}, \quad (9)$$

$$F_{\text{out}} = \text{FFN}(m_1 \oplus m_2 \oplus m_3, F_s).$$

So SADPA injects the scale space analysis into the attention mechanism by projecting the  $K$  and  $V$  into different scales via convolutions

before computing the attention matrix. It allows attention processing at various scales: fine-level attention retains more local details, whereas coarse-level attention captures broader image contexts. By excluding the irrelevant features during the fine-level attentions, the computational cost and disruption are largely reduced.

**Positional Encoding.** The spatial relationship of features is crucial for matching. But the attention mechanism falls short in recognizing spatial positional relationships. Therefore, a positional encoding is necessary. Previous methods [11, 21, 65] use the 2D extension of the standard sinusoidal encoding following DETR [9]. However, in the context of two-view geometry, it's apparent that the positioning of visual elements changes consistently in relation to camera movements within the image plane. This phenomenon underscores the need for an encoding that prioritizes relative position over absolute position. we adopt the Rotary Position Embedding (RoPE) [64] to remedy this problem. It allows the model to effectively identify the relative positioning of point  $j$  from point  $i$ . We only apply RoPE in self SADPA, because it makes no sense to compute the relative positions across images.

### 3.3 Match Prediction by Weighted Dual-Softmax

After the update by  $L$  MPM blocks, we get the final transformed features  $F_L^A$  and  $F_L^B$  and flatten them for further use. We also obtain their corresponding estimated NMI  $\sigma_L^A$  and  $\sigma_L^B$  at the last MPM layer. We calculate the matching matrix  $\mathcal{P}$  combining both the similarity and the estimated NMI:

$$\mathcal{P}(i, j) = \sigma_L^A(i)\sigma_L^B(j)\text{softmax}(S(i, \cdot))_j \cdot \text{softmax}(S(\cdot, j))_i \quad (10)$$

where  $S$  is the similarity matrix computed by the features:  $S(i, j) = \tau \cdot \langle F_L^A(i), F_L^B(j) \rangle$ .  $\tau$  is the temperature coefficient. The NMI is used to weight the matching matrix, as the valid match points should be both relevant and similar. We selected matches with  $\mathcal{P}(i, j) > \theta_c$  and filtered them using the mutual nearest neighbor strategy, resulting the coarse matches  $\mathcal{M}_c$ .

### 3.4 Supervision

The final loss is composed of three parts: coarse matching loss  $\mathcal{L}_c$ , sub-pixel refinement loss  $\mathcal{L}_f$  and patch pruning loss  $\mathcal{L}_p$ :

$$\mathcal{L} = \mathcal{L}_c + \mathcal{L}_f + \mathcal{L}_p \quad (11)$$

**Coarse Matching Loss.** We use cross entropy loss to supervise the coarse matching matrix  $\mathcal{P}$ :

$$\mathcal{L}_c = -\frac{1}{|\mathcal{M}_c^{gt}|} \sum_{(i,j) \in \mathcal{M}_c^{gt}} \log \mathcal{P}(i, j) \quad (12)$$

The ground-truth coarse matches  $\mathcal{M}_c^{gt}$  is calculated from the camera poses and depth maps at coarse resolution.

**Sub-pixel Refinement Loss.** Following LoFTR [65], we use the L2-distance between each refined coordinate and the ground truth reprojection coordinate and normalize it by the variance  $\phi$ :

$$\mathcal{L}_f = \frac{1}{|\mathcal{M}_f|} \sum_{(\hat{i}, \hat{j}') \in \mathcal{M}_f} \frac{1}{\phi^2(\hat{i})} \left\| \hat{j}' - \hat{j}'_{gt} \right\|_2, \quad (13)$$

We compute the the  $\hat{j}'_{gt}$  by warping  $\hat{i}$  on  $I^A$  to  $I^B$  with the ground-truth pose and depth.

**Patch Pruning Loss.** We supervise the Patch Pruning module as the negative log-likelihood loss over the estimated NMI for all

features. Patch features derived from  $I^A$  that can find matches in  $I^B$  are defined as  $A_m$ , and the rest patch features can not find matches are defined as  $A_n$ .  $B_m$  and  $B_n$  is defined similarly. For NMI estimated at  $l$ -th MPM layer, the loss  $\mathcal{L}_p^A(l)$  is defined as:

$$\mathcal{L}_p^A(l) = -\left(\frac{1}{|A_m|} \sum_{i \in A_m} \log(\sigma_l^A(i)) + \frac{1}{|A_n|} \sum_{i \in A_n} \log(1 - \sigma_l^A(i))\right) \quad (14)$$

$\mathcal{L}_p^B(l)$  is defined in the same way. The final  $\mathcal{L}_p$  is defined as:

$$\mathcal{L}_p = \frac{1}{L} \sum_l \frac{\mathcal{L}_p^A(l) + \mathcal{L}_p^B(l)}{2} \quad (15)$$

### 3.5 Implementation Details

We adopt the same ResNet-FPN [24, 36] architecture as LoFTR [65] to extract image features. The dimension for coarse features and fine features are 256 and 128, respectively. We use 4 MPM layers for feature updating and pruning. For each SADPA, the convolutional kernel sizes are 4, 2 and 1 from coarse to fine. We use an efficient implementation of attention [15] and each attention unit has 4 heads. We only train the model on the MegaDepth [34] dataset without any data augmentation and test it on all datasets and tasks to demonstrate the generalization ability. We follow the same train-test split as in LoFTR [65]. We use the AdamW [38] optimizer with an initial learning rate of  $8 \times 10^{-4}$ . The model is trained end-to-end with a batch size of 24 on 8 NVIDIA A100, taking 1.5d to converge.

## 4 Experiments

### 4.1 Homography Estimation

Homography is crucial in two-view geometry. It enables the transformation of perspectives between two images of the same scene. We assess homography accuracy by measuring corner correctness. We warp the four corners of a reference image to another using estimated homography and ground truth homography respectively and calculate the corner error between the warped points, as in [58, 65].

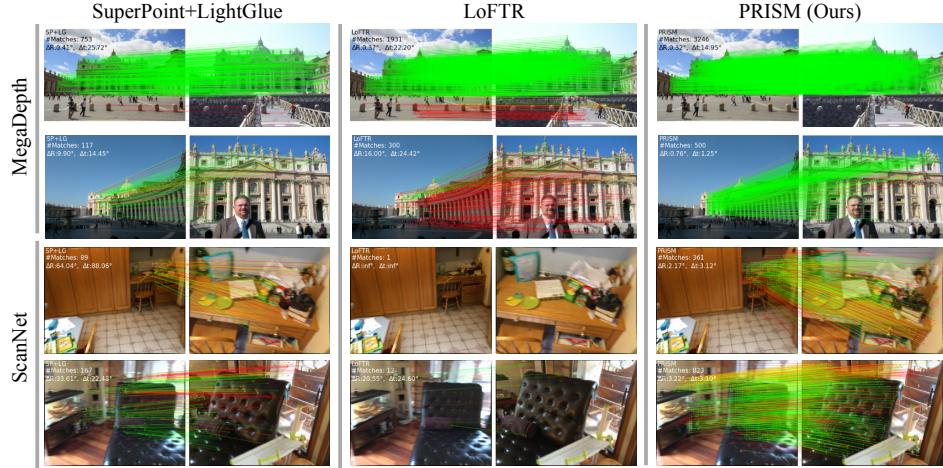
**Setup.** We evaluate on the widely used HPatches dataset [1]. HPatches comprises 52 sequences showing significant changes in illumination and 56 sequences displaying pronounced changes in viewpoint. Each sequence includes 1 reference image alongside 5 query images. We report the Area Under the cumulative Curve (AUC) of the corner error up to thresholds of 3, 5, and 10 pixels. OpenCV's RANSAC algorithm is used for robust estimation. We compare with two categories of methods: dense methods [11, 21, 33, 43, 51, 65, 69, 78, 80] and sparse methods [17, 50, 58, 58, 71, 85].

**Results.** Table 1 shows that PRISM outperforms other baselines under all error thresholds, which strongly proves the effectiveness of our method. We attribute the outstanding performance to the ability to capture multi-scale contexts provided by the SADPA. The iterative pruning paradigm also contributes to the accuracy by greatly reducing the mismatches.

### 4.2 Relative Pose Estimation

Relative pose estimation plays a fundamental role for various applications. We measure the relative pose error by the maximum angular error in rotation and translation. To determine the camera pose, we compute the essential matrix using predicted match points and apply both the RANSAC of OpenCV and LO-RANSAC of Poselib [31] for robust estimation.

**Setup.** We evaluate our PRISM model on MegaDepth [34] and ScanNet [14] dataset. MegaDepth is an extensive outdoor dataset



**Figure 4: Qualitative Results.** We compare PRISM with SP [16]+LG [37] and LoFTR [65] in ScanNet [14] and MegaDepth [34] dataset. As shown in the figure, PRISM can generate more dense matches and avoid most outliers in both indoor and outdoor scenes. The red color indicates epipolar error beyond  $5 \times 10^4$  (in the normalized image coordinates). More visualizations are provided in the Appendix.

Category	Methods	Homography AUC		
		@3px	@5px	@10px
Sparse	SP [16]+SG [58]	53.9	68.3	81.7
	D2Net [17]+NN	23.2	35.9	53.6
	R2D2 [50]+NN	50.6	63.9	76.8
	Patch2Pix [85]	46.4	59.2	73.1
	DISK [71]+NN	52.3	64.9	78.9
Dense	DRC-Net [33]	50.6	56.2	68.3
	Sparse-NCNet [51]	48.9	54.2	67.1
	LoFTR [65]	65.9	75.6	84.6
	Quadtree [69]	66.3	76.2	84.9
	ASpanFormer [11]	67.4	76.9	85.6
	3DG-STFM [43]	64.7	73.1	81.0
	ASTR [80]	71.7	80.3	88.0
	TopicFM [21]	70.9	80.2	<b>88.3</b>
	EfficientLoFTR [78]	66.5	76.4	85.5
	PRISM(Ours)	<b>71.9</b>	<b>80.4</b>	<b>88.3</b>

**Table 1: Homography estimation on Hpatches.**

	Methods	RANSAC AUC			Lo-RANSAC AUC		
		@5°	@10°	@20°	@5°	@10°	@20°
Sparse	SP [16]+SG [58]	42.2	61.2	76.0	65.8	78.7	87.5
	D2-Net [17]+NN	19.6	36.3	54.6	33.4	47.3	60.4
	R2D2 [50]+NN	36.2	54.1	69.4	48.0	62.8	73.8
	SP [16]+LG [37]	49.4	67.2	80.1	66.3	79.0	87.9
Dense	LoFTR [65]	52.8	69.2	81.2	64.3	76.6	85.3
	Quadtree [69]	54.6	70.5	82.2	65.6	78.0	86.5
	ASpanFormer [11]	55.3	71.5	83.1	68.1	80.0	88.3
	3DG-STFM [43]	52.6	68.5	80.0	-	-	-
	AdaMatcher [25]	52.4	69.7	82.1	64.1	76.8	85.6
	HCPM [13]	52.6	69.2	81.4	-	-	-
	ASTR [80]	58.4	73.1	83.8	-	-	-
	TopicFM [21]	58.2	72.8	83.2	64.1	76.7	85.6
	EfficientLoFTR [78]	56.4	72.2	83.5	67.5	79.1	87.0
	MESA_ASpan [82]	58.4	74.1	84.8	-	-	-
	PRISM (ours)	<b>60.0</b>	<b>74.9</b>	<b>85.1</b>	<b>68.8</b>	<b>80.6</b>	<b>88.9</b>

**Table 2: Relative pose estimation on MegaDepth dataset.**

	Methods	RANSAC AUC		
		@5°	@10°	@20°
Sparse	D2-Net [17]+NN	5.5	14.5	28.0
	SP [16]+SG [58]	16.2	33.8	51.8
	SP [16]+OANet [81]	11.8	26.9	43.9
	SP [58]+LG [37]	17.7	34.6	51.2
Dense	DRC-Net [33]	7.7	17.9	30.5
	LoFTR [65]	16.9	40.8	50.6
	Quadtree [69]	19.0	37.3	53.5
	ASpanFormer [11]	19.6	37.7	54.4
	ASTR [80]	19.4	37.6	54.4
	TopicFM [21]	17.3	34.5	50.9
	Patch2Pix [85]	9.6	20.2	32.6
	EfficientLoFTR [78]	19.2	37.0	53.6
	PRISM (Ours)	<b>23.9</b>	<b>41.8</b>	<b>58.9</b>

**Table 3: Relative pose estimation on ScanNet dataset.**

with 1 million images across 196 scenes, reconstructed by COLMAP [60]. For testing, we use the same 1500 pairs as in [65], resizing images to a longer dimension of 1152. ScanNet features monocular sequences with ground truth data and is challenging due to wide baselines and textureless regions. Following [65], we resize images to 640x480. To verify the PRISM’s generalizability, the model is only trained on MegaDepth and tested on both datasets. Following [58, 65], we report the AUC of pose error at thresholds of 5°, 10°, and 20°. Note that we use the official codes, configurations and pre-trained weights to report the AUC under the Lo-RANSAC solver.

**Results.** Table 2 and Table 3 provide the AUC of pose error for MegaDepth and ScanNet, respectively. As we can see, our proposed PRISM achieves new state-of-the-art performance for all evaluation metrics. Thanks to the proposed Multi-scale Pruning Module, PRISM can avoid a large number of incorrect matches and perceive multi-scale information. Figure 4 qualitatively demonstrates our method’s performance against others. For the ScanNet dataset, our method notably improves by 4.3% in AUC@5° and 4.5% in AUC@20° compared to the best model trained on MegaDepth, indicating the impressive generalization capability of our method.

	Methods	DUC1	DUC2	overall
		(0.25m, 10°) / (0.5m, 10°)	(1m, 10°)	
Sparse	SP [16]+SG [58]	49.0/68.7/80.8	53.4/ <b>77.1</b> /82.4	68.6
	ClusterGNN [62]	47.5/69.7/79.8	53.4/ <b>77.1</b> /84.7	68.7
	SP [16]+LG [37]	49.0/68.2/79.3	<b>55.0</b> /74.8/79.4	67.6
Dense	LoFTR [65]	47.5/72.2/84.8	54.2/74.8/ <b>85.5</b>	69.8
	ASpanFormer [11]	51.5/73.7/86.0	55.0/74.0/81.7	70.3
	Patch2Pix [85]	44.4/66.7/78.3	49.6/64.9/72.5	62.7
	ASTR [80]	53.0/73.7/87.4	52.7/76.3/84.0	71.2
	TopicFM [21]	52.0/74.7/87.4	53.4/74.8/83.2	70.9
	CasMTR [8]	<b>53.5</b> /76.8/85.4	51.9/70.2/83.2	70.2
	PRISM (Ours)	53.0/ <b>77.8</b> / <b>87.9</b>	54.2/72.5/83.2	<b>71.4</b>

Table 4: Indoor visual localization on InLoc dataset.

### 4.3 Visual Localization

Visual Localization is essential in computer vision. The percentage of pose errors satisfying both angular and distance thresholds is reported, as in the Long-Term Visual Localization Benchmark [70].

**Setup.** We evaluate PRISM on the InLoc [67] dataset for indoor scenes and the Aachen Day-Night v1.1 [59, 84] dataset for outdoor scenes. The InLoc dataset consists of 9,972 geometrically registered RGBD indoor images and 329 query images with verified poses, posing challenges in textureless or repetitive environments. We use the two scenes named DUC1 and DUC2 for test as in [11, 65]. The Aachen dataset provides 6,697 daytime and 191 nighttime images, highlighting the difficulty of matching under significant illumination changes, especially at night. The metrics of the daytime and nighttime divisions are reported. We follow the benchmark [70] to compute query poses. For both datasets, candidate image pairs are identified using the pre-trained HLoc [56] system following [11, 37, 65]. Camera poses are estimated utilizing our model trained on MegaDepth dataset.

**Results.** Table 4 shows the results on the InLoc dataset. PRISM is better than all baselines on DUC1 and on par with state-of-the-art methods on DUC2. Overall, we achieve the best performance on average. On Aachen V1.1 dataset, as shown in table 5, PRISM performs best on the day queries and the results on the night queries are slightly lower than that of LightGlue [37]. We argue that the Dense methods require quantification for the triangulation step of the localization pipeline, which harms the accuracy. Compared to dense methods, PRISM’s performance ranks among the top tier on night queries. In general, PRISM shows promising performances and generalization in visual localization tasks. These evaluations demonstrate our network’s versatility across different task settings.

### 4.4 Understanding PRISM

**Ablation Study.** To fully understand the design decisions in PRISM, we follow the same setting in Sec. 3.5 and conduct ablation experiments on MegaDepth, as shown in Tab. 6. 1) Replacing the RoPE with the absolute positional encoding of LoFTR results in a degraded AUC. 2) Without the Patching Pruning module, there is a significant drop in pose estimation accuracy as expected. This demonstrates the efficacy of the proposed Patch Pruning module. 3) Using LoFTR’s Linear Attention instead of SADPA leads to a noticeably declined result. We attribute this to the ability of SADPA to aggregate multi-scale information. 4) Replaceing SADPA with only single level attention at  $\frac{1}{8}$  resolution will lead to degraded

	Methods	Day	Night
		(0.25m, 2°) / (0.5m, 5°) / (1m, 10°)	
Sparse	SP [16]+SG [58]	88.2/95.5/98.7	86.7/92.9/ <b>100</b>
	SGMNet [10]	86.8/94.2/97.7	83.7/91.8/99.0
	ClusterGNN [62]	89.4/95.5/98.5	81.6/ <b>93.9</b> / <b>100</b>
Dense	SP [16]+LG [37]	89.2/95.4/98.5	<b>87.8</b> / <b>93.9</b> / <b>100</b>
	LoFTR [65]	88.7/95.6/99.0	78.5/90.6/99.0
	ASpanFormer [11]	89.4/95.6/99.0	77.5/91.6/99.5
	AdaMatcher [25]	89.2/95.9/99.2	79.1/92.1/99.5
	PATS [46]	89.6/95.8/ <b>99.3</b>	73.8/92.1/99.5
	ASTR [80]	89.9/95.6/99.2	76.4/92.1/99.5
	TopicFM [21]	<b>90.2</b> /95.9/98.9	77.5/91.1/99.5
EfficientLoFTR [78]	EfficientLoFTR [78]	89.6/96.2/99.0	77.0/91.1/99.5
	PRISM (Ours)	89.4/ <b>96.2</b> / <b>99.3</b>	78.5/91.1/99.5

Table 5: Outdoor visual localization on Aachen Day-Night v1.1 dataset.

Method	Pose estimation AUC		
	@5°	@10°	@20°
1)Replace RoPE to absolute positions	57.3	73.0	83.9
2)Without Patch Pruning	56.6	72.6	83.6
3)Replace SADPA with LoFTR’s Attention	55.3	70.9	82.9
4)Replace SADPA with single level design	57.7	73.2	84.3
5)Without weighted Softmax	58.5	73.4	84.1
<b>Full</b>	<b>60.0</b>	<b>74.9</b>	<b>85.1</b>

Table 6: Ablation study on MegaDepth dataset.

Methods	resolution			
	640 × 640	832 × 832	960 × 960	1152 × 1152
LoFTR [65]	<b>89.6</b> /11.1	<b>107.7</b> /17.3	<b>145.1</b> /22.3	212.5/13.9
AspanFormer [11]	119.3/16.7	173.0/20.4	208.3/22.5	289.2/13.9
PRISM(Ours)	99.4/ <b>5.4</b>	119.7/ <b>7.7</b>	153.3/ <b>9.9</b>	<b>209.1</b> / <b>13.5</b>

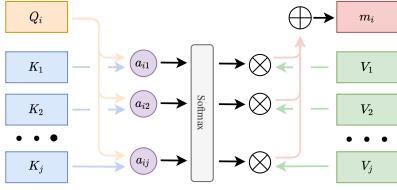
Table 7: Impact of test image resolution on the MegaDepth dataset [34]. We report both the runtime(ms)/VRAM(GiB).

pose accuracy. It further validates the essentiality of the design of SADPA. 5) The absence of the weights (estimated NMI) of Softmax will lead to a certain degree of performance degradation.

**Impact of test image resolutions.** We test the runtime and VRAM of PRISM under different image resolutions and compare with LoFTR [65] and AspanFormer [11], as shown in Table. 7. All results are based on one single A100 GPU. For the runtime, PRISM is slightly slower than LoFTR and largely outperforms AspanFormer. In terms of VRAM, PRISM can save about half of the VRAM usage compared to baselines in most cases.

### 5 Conclusion

In this paper, we propose PRogressive dependency maxImization for Scale-invariant image Matching (PRISM). PRISM gradually prunes irrelevant patch features during the feature interactions. Meanwhile, in order to better handle the scale discrepancy, we propose the Scale-Aware Dynamic Pruning Attention to aggregate information from different scales via a hierarchical design. Extensive experimental results on a wide range of benchmarks demonstrate the effectiveness and generalization of PRISM. With careful engineering optimizations, PRISM’s time efficiency can be enhanced.



**Figure 5: The process of attention mechanism.** The  $\otimes$  means the product and the  $\oplus$  represents the summation. The RoPE considers the relative distance between query and key when computing the attention score  $a_{ij}$ .

## Acknowledgments

Dr. Li is supported in part by the National Natural Science Foundation of China Grant No. 12071478. Dr. Wang is supported in part by the National Natural Science Foundation of China Grant No. 61972404, Public Computing Cloud, Renmin University of China, and the Blockchain Lab, School of Information, Renmin University of China.

## APPENDIX

### A Summary

In this supplementary material, we provide more details and experiment results about our proposed image matching method: PRogressive dependency maxImization for Sclae-invariant image Matching (PRISM). We first elaborate on the RoPE [64] and supervision of the Patch Pruning module mentioned in the main manuscript (Sec. B). Then, we conduct more experiments to validate the effectiveness of our design choices and superior performance (Sec. C). Finally, we give more qualitative results on MegaDepth [34], ScanNet [14] and InLoc [67] datasts (Sec. D).

### B Implementation Details

#### B.1 Details about Positional Encoding

We adopt the Rotary Position Embedding (RoPE) [64] to model the spatial positional relationships between patch features. Given the  $i$ th projected query  $Q_i \in \mathbb{R}^d$  and  $n$  projected K-V pairs where  $K_j \in \mathbb{R}^d$  and  $V_j \in \mathbb{R}^d$ , the attention message  $m_i$  can be defined as:

$$m_i = \sum_{j=1}^n \text{softmax}(a(Q_i, K_j)) V_j \quad (16)$$

The RoPE does not fully model the positional information of each input. Instead, it considers the relative distance between the current position and the position being attended to when computing attention score. In this case, the attention score between the query feature  $Q_i$  and  $j$ th key feature  $K_j \in \mathbb{R}^d$  is defined as:

$$a_{ij} = Q_i^T R(\Delta x, \Delta y) K_j \quad (17)$$

Where  $R(\cdot) \in \mathbb{R}^{d \times d}$  is the rotary encoding and  $\Delta x = x_i - x_j, \Delta y = y_i - y_j$ .  $x_i, y_i$  is the coordinates of  $Q_i$ , defined as the normalized central coordinate of grid corresponding to  $Q_i$  at  $\frac{1}{8}$  resolution.  $x_j, y_j$  is the coordinates of  $K_j$  and defined in the similar way. To compute the rotary encoding, the feature space is first divided into  $\frac{d}{2}$  2D

subspaces and each of them is rotated by an angle:

$$R(\Delta x, \Delta y) = \begin{pmatrix} R_1(\theta_1) & & & \\ & R_2(\theta_2) & & \\ & & \ddots & \\ & & & R_{\frac{d}{2}}(\theta_{\frac{d}{2}}) \end{pmatrix} \quad (18)$$

$$R_k(\theta) = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}, \theta = b_k^T(\Delta x, \Delta y) \quad (19)$$

Where  $R_k(\theta)$  is the rotary matrix for  $k$ th subspace and  $b_k \in \mathbb{R}^2$  is a learned parameter to project the the distance between features into frequencies. The whole process is shown in Figure 5.

### B.2 Patch Pruning Supervision

Since the Patch Pruning is applied on coarse features at  $\frac{1}{8}$  resolution, We use the central coordinates of grids at  $\frac{1}{8}$  resolution to represent the coordinates of the coarse features. We project the central points in left image to the right image using its groundtruth depth map and camera pose, and take its nearest neighbor as a matching candidate. Then the central points in right image are projected to the left images in the same way. Based on the matching relationships in these two directions, we use MNN (Mutual Nearest Neighbors) filtering to obtain reasonable matches. The filtered matches can be used to indicate whether a feature can be matched or not and supervise the training.

## C More Experiments Results

### C.1 Evaluating Pose Estimation Errors on a scale-split Megadepth test set

To further demonstrate the effectiveness of PRISM in handling scale discrepancy, we evaluate the pose estimation errors on a scale-split Megadepth test set. The image pairs are split by their scale ratio ranges in  $[1, 2]$ ,  $[2, 3]$ ,  $[3, 4]$  and  $[4, +\infty)$ , following [12, 25]. All images are resized so that the longest dimension equals 840 (ASpanFormer [11] and our method (PRISM) use 832 due to the need for an image resolution divisible by 16). The pose error AUC at thresholds of  $5^\circ$ ,  $10^\circ$ , and  $20^\circ$  are reported.

We compare with sparse methods [16, 39, 45, 50, 58, 71] and dense methods [11, 25, 65], as shown in table 8. Our method outperforms all baselines under various relative scale ratios substantially. We attribute the top performance to the hierarchical design of SADPA, which can aggregate features across different scales. The Patch Pruning module also contributes to the accuracy by pruning irrelevant features.

### C.2 Impact of test image resolutions

We farther study the impact of test image resolutions to PRISM on the Pose estimation AUC and report the optimized VRAM usage (inference using Half-Precision), as shown in table 9. All results are based on one single A100 GPU. As the resolution is reduced, PRISM can still achieve competitive performance and the inference time and memory efficiency have both been greatly improved.

### C.3 Ablation on the Patch Pruning Threshold

This section analyzes the impact of the Patch Pruning threshold  $\theta_p$ . We train several variants of PRISM using  $\theta_p \in \{0.95, 0.9, 0.85, 0.8\}$

Methods	Scale [1, 2)	Scale [2, 3)	Scale [3, 4)	Scale [4, inf)
	Pose estimation AUC@5° / AUC@10° / AUC@10°			
Sparse	SIFT [39]+HardNet [45]	21.2 / 33.0 / 45.4	10.8 / 18.6 / 28.6	4.6 / 9.3 / 16.2
	DISK [71]	33.7 / 49.8 / 63.3	5.5 / 8.5 / 11.6	0.2 / 0.5 / 0.8
	R2D2 [50]	37.8 / 55.9 / 70.7	22.7 / 36.9 / 51.9	6.6 / 13.0 / 22.0
Dense	SP [16]+SG [58]	50.4 / 67.6 / 80.0	39.4 / 57.78 / 72.3	19.7 / 35.2 / 52.0
	LoFTR [65]	60.2 / 74.7 / 84.5	49.7 / 65.7 / 77.9	24.9 / 39.7 / 55.1
	ASpanFormer [11]	60.9 / 75.3 / 85.0	54.6 / 70.2 / 81.2	33.4 / 51.2 / 66.9
	AdaMatcher [25]	62.4 / 76.0 / 85.4	57.0 / 71.8 / 82.6	41.0 / 58.7 / 73.4
PRISM (Ours)		<b>66.6 / 79.2 / 87.2</b>	<b>61.9 / 75.9 / 85.2</b>	<b>43.5 / 60.7 / 74.9</b>
27.4 / 42.9 / 57.4				

**Table 8: Relative pose estimation on the scale-split Megadepth test set.**

resolution	Pose estimation AUC			VRAM (GiB)	time(ms)
	@5°	@10°	@20°		
1152 × 1152	60.0	74.9	85.1	13.5/11.0	209.1
960 × 960	58.3	73.0	83.6	9.9/6.6	153.3
832 × 832	56.4	71.6	82.7	7.7/4.5	119.7
640 × 640	52.9	68.5	80.2	5.4/3.2	99.4

**Table 9: The Pose estimation AUC, VRAM and inference time under different test image resolutions.** the SP means inference using Single-Precision and HP means Half-Precision.

$\theta_p$	Pose estimation AUC		
	@5°	@10°	@20°
0.8	56.2	<b>71.7</b>	82.8
0.85	56.2	71.6	82.7
0.9	56.1	71.6	<b>82.9</b>
0.95	<b>56.4</b>	71.6	82.7

**Table 10: The impact of Patch Pruning threshold  $\theta_p$  on the MegaDepth dataset.** The test image resolution is 832 × 832.

and test the pose estimation errors at 832 × 832 image resolution. Table 10 shows the results of relative pose estimation results for PRISM on MegaDepth. As shown in the table, the model with  $\theta_p = 0.95$  achieved the best performance. Other models with different  $\theta_p$  perform slightly worse.

#### C.4 Ablation on the gradual design of MPM

We conduct an ablation study to validate the gradual design of MPM. We train a variant of PRISM which only keeps the Patch Pruning module in the last MPM and removes the rest. This design makes patch pruning degenerate into mere co-visible area prediction, only functioning during the Match Prediction phase, similar to existing works [21, 25]. This variant is trained using the same setting as the full model and tested on MegaDepth dataset at 832 × 832 image resolution. The Pose error AUC at thresholds of 5°, 10°, and 20° are reported. As shown in table 11, when only performing patch pruning at the last MPM, the pose estimation AUC declines noticeably. This validates the effectiveness of the gradual design of MPM.

Method	Pose estimation AUC		
	@5°	@10°	@20°
Only pruning at the last MPM	54.6	70.5	81.8
Replace with cosine similarity	54.7	70.3	81.9
gradually pruning at every MPM (Full)	<b>56.4</b>	<b>71.6</b>	<b>82.7</b>

**Table 11: More ablation studies.** The test image resolution is 832 × 832.

#### C.5 Ablation on the NMI estimator

As shown in table 11, we replace the NMI estimator in the patch pruning with cosine similarity and retrain the network in the same setup. The AUC of pose error on MegaDepth is shown. The significantly decreased performance validates the effectiveness and necessity of our choice.

#### C.6 Quality of the predicted pruning mask

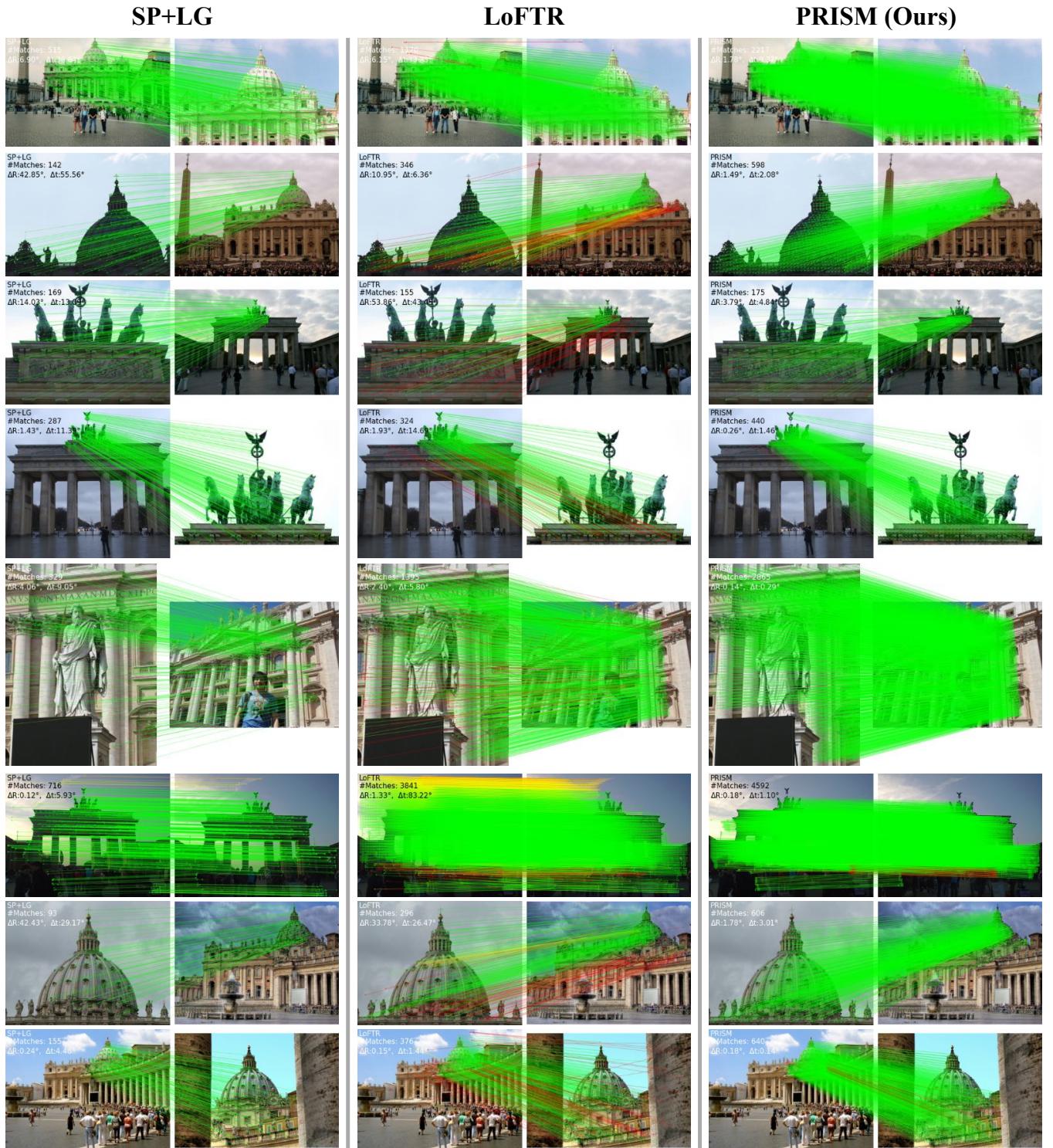
We access the mIoU and AR of the predicted masks, as shown in table 12. The relatively lower mIoU is due to the sparse and discrete nature of ground truth.

Datasets	mIoU (%)	Average Recall (%)
MegaDepth / ScanNet	55.08 / 46.89	99.91 / 92.32

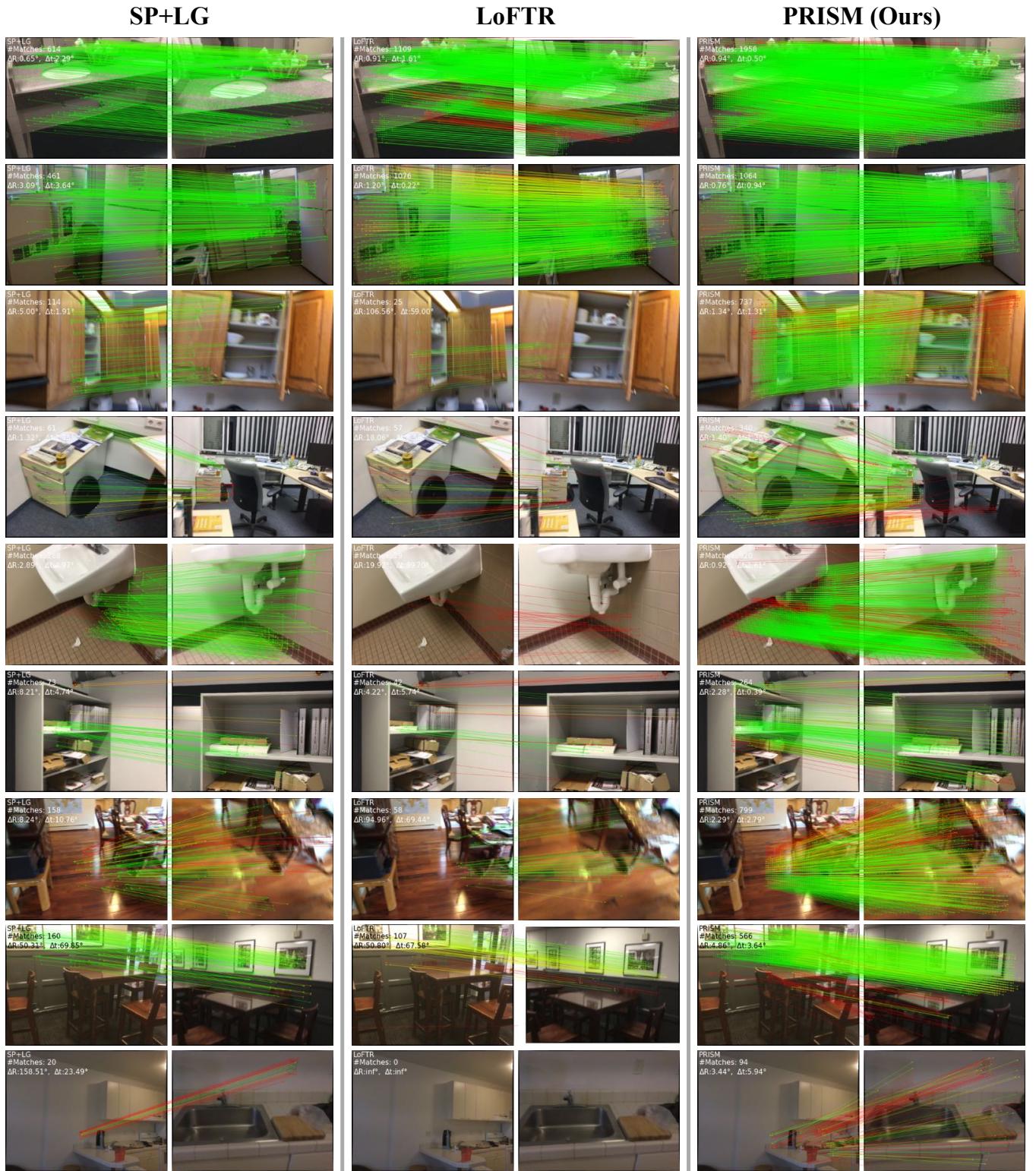
**Table 12: The mIoU and AR of the predicted masks**

## D More Qualitative Results

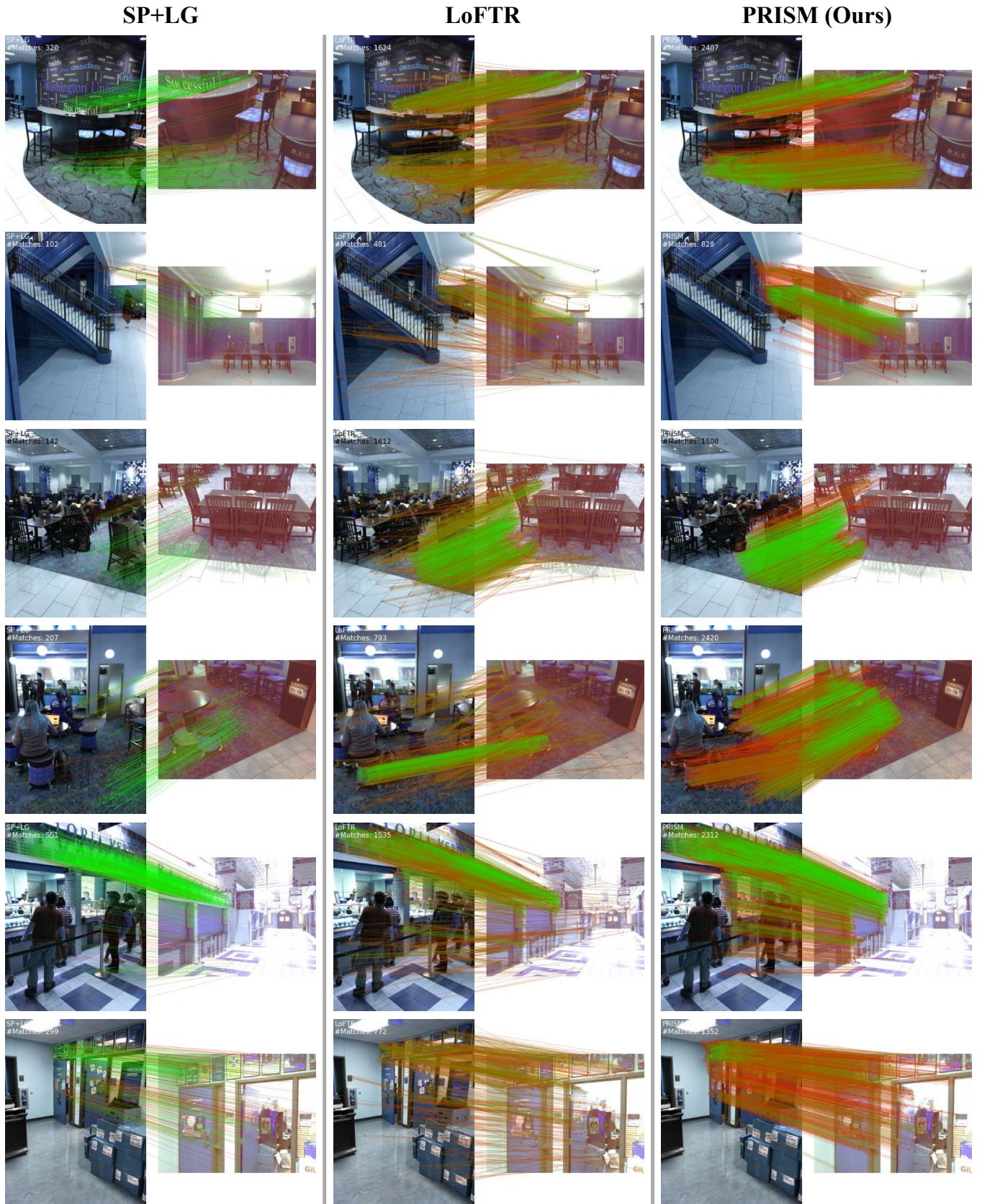
More qualitative comparisons between PRISM and LoFTR [65] and SP [16]+LG [37] are provided. Figure 6 shows the qualitative results on MegaDepth dataset and Figure 7 shows the qualitative results on ScanNet dataset. More qualitative results on the InLoc benchmark are shown in Figure 8.



**Figure 6: Qualitative Results on MegaDepth dataset.** We compare with the sparse method SP [16]+LG [37] and dense method LoFTR [65]. The red color indicates the epipolar error beyond  $5 \times 10^{-4}$  (in the normalized image coordinates). We report the pose error in the upper left corner in each pair of image, please zoom in for details.



**Figure 7: Qualitative Results on ScanNet dataset.** We compare with the sparse method SP [16]+LG [37] and dense method LoFTR [65]. The red color indicates the epipolar error beyond  $5 \times 10^{-4}$  (in the normalized image coordinates). We report the pose error in the upper left corner in each pair of image, please zoom in for details.



**Figure 8: Qualitative Results on InLoc dataset.** We compare with the sparse method SP [16]+LG [37] and dense method LoFTR [65]. We color the matches with matching scores since there is no ground-truth pose available. The green indicates higher score and red for the opposite.

## References

- [1] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. 2017. HPatches: A Benchmark and Evaluation of Handcrafted and Learned Local Descriptors. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017*. IEEE Computer Society, 3852–3861. <https://doi.org/10.1109/CVPR.2017.410>
- [2] Roberto Battiti. 1994. Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. Neural Networks* 5, 4 (1994), 537–550. <https://doi.org/10.1109/72.298224>
- [3] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. 2008. Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding* 110, 3 (2008), 346–359. <https://doi.org/10.1016/j.cviu.2007.09.014> Similarity Matching in Computer Vision and Multimedia.
- [4] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. 2018. Mutual information neural estimation. In *International conference on machine learning*. PMLR, 531–540.
- [5] Gavin Brown, Adam Craig Pocock, Ming-Jie Zhao, and Mikel Luján. 2012. Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection. *J. Mach. Learn. Res.* 13 (2012), 27–66. <https://doi.org/10.5555/2503308.2188387>
- [6] Han Cai, Chuang Gan, and Song Han. 2022. EfficientViT: Enhanced Linear Attention for High-Resolution Low-Computation Visual Recognition. *CoRR* abs/2205.14756 (2022). [https://doi.org/10.48550/ARXIV.2205.14756 arXiv:2205.14756](https://doi.org/10.48550/ARXIV.2205.14756)
- [7] Xudong Cai, Yongcai Wang, Zhe Huang, Yu Shao, and Deying Li. 2024. VOLoc: Visual Place Recognition by Querying Compressed Lidar Map. *arXiv preprint arXiv:2402.15961* (2024).
- [8] Chenjie Cao and Yanwei Fu. 2023. Improving Transformer-based Image Matching by Cascaded Capturing Spatially Informative Keypoints. *ArXiv* abs/2303.02885 (2023). <https://api.semanticscholar.org/CorpusID:257365749>
- [9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-End Object Detection with Transformers. *ArXiv* abs/2005.12872 (2020). <https://api.semanticscholar.org/CorpusID:218889832>
- [10] Hongkai Chen, Zixin Luo, Jiahui Zhang, Lei Zhou, Xuyang Bai, Zeyu Hu, Chiew-Lan Tai, and Long Quan. 2021. Learning to Match Features with Seeded Graph Matching Network. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV) (2021)*, 6281–6290. <https://api.semanticscholar.org/CorpusID:237213601>
- [11] Hongkai Chen, Zixin Luo, Lei Zhou, Yurun Tian, Mingmin Zhen, Tian Fang, David N. R. McKinnon, Yanghai Tsin, and Long Quan. 2022. ASpanFormer: Detector-Free Image Matching with Adaptive Span Transformer. In *European Conference on Computer Vision*.
- [12] Ying Chen, Di He Huang, Shang Xu, Jianlin Liu, and Yong Liu. 2022. Guide Local Feature Matching by Overlap Estimation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelfth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 365–373. <https://doi.org/10.1609/AAAI.V36I1.19913>
- [13] Ying Chen, Yong Liu, Kai Wu, Qiang Nie, Shang Xu, Huifang Ma, Bing Wang, and Chengjie Wang. 2024. HCPM: Hierarchical Candidates Pruning for Efficient Detector-Free Matching. *arXiv preprint arXiv:2403.12543* (2024).
- [14] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. 2017. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017*. IEEE Computer Society, 2432–2443. <https://doi.org/10.1109/CVPR.2017.261>
- [15] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In *Advances in Neural Information Processing Systems*.
- [16] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2018. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 224–236.
- [17] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. 2019. 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*. 8092–8101.
- [18] Pablo A. Estévez, M. Tesmer, Claudio A. Perez, and Jacek M. Zurada. 2009. Normalized Mutual Information Feature Selection. *IEEE Trans. Neural Networks* 20, 2 (2009), 189–201. <https://doi.org/10.1109/TNN.2008.2005601>
- [19] Miao Fan, Ming lei Chen, Chen Hu, and Shuchang Zhou. 2023. Occ2Net: Robust Image Matching Based on 3D Occupancy Estimation for Occluded Regions. *ArXiv* abs/2308.16160 (2023). <https://api.semanticscholar.org/CorpusID:261339755>
- [20] Wanfu Gao, Liang Hu, Ping Zhang, and Jialong He. 2018. Feature selection considering the composition of feature relevancy. *Pattern Recognit. Lett.* 112 (2018), 70–74. <https://doi.org/10.1016/J.PATREC.2018.06.005>
- [21] Khang Truong Giang, Soohwan Song, and Sung-Guk Jo. 2022. TopicFM: Robust and Interpretable Feature Matching with Topic-assisted. *ArXiv* abs/2207.00328 (2022). <https://api.semanticscholar.org/CorpusID:250243816>
- [22] Xiangyuan Gu, Jichang Guo, Lijun Xiao, Tao Ming, and Chongyi Li. 2020. A Feature Selection Algorithm Based on Equal Interval Division and Minimal-Redundancy-Maximal-Relevance. *Neural Process. Lett.* 51, 2 (2020), 1237–1263. <https://doi.org/10.1007/S11063-019-10144-3>
- [23] Christopher G. Harris and M. J. Stephens. 1988. A Combined Corner and Edge Detector. In *Alvey Vision Conference*. <https://api.semanticscholar.org/CorpusID:1694378>
- [24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [25] Dihe Huang, Ying Chen, Shang Xu, Yong Liu, Wen-Qi Wu, Yikang Ding, Chengjie Wang, and Fan Tang. 2022. Adaptive Assignment for Geometry Aware Local Feature Matching. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)*, 5425–5434.
- [26] Zhe Huang, Yongcai Wang, Xingu Tang, and Hongyu Sun. 2023. Boundary-aware set abstraction for 3D object detection. In *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 01–07.
- [27] Zhe Huang, Yongcai Wang, Jie Wen, Peng Wang, and Xudong Cai. 2023. An object detection algorithm combining semantic and geometric information of the 3D point cloud. *Advanced Engineering Informatics* 56 (2023), 101971.
- [28] Wei Jiang, Eduard Trulls, Jan Hendrik Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. 2021. COTR: Correspondence Transformer for Matching Across Images. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV) (2021)*, 6187–6197. <https://api.semanticscholar.org/CorpusID:232380104>
- [29] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and Francois Fleuret. 2020. Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. In *International Conference on Machine Learning*.
- [30] Solomon Kullback. 1997. *Information theory and statistics*. Courier Corporation.
- [31] Viktor Larsson and contributors. 2020. PoseLib - Minimal Solvers for Camera Pose Estimation. <https://github.com/vlarsson/PoseLib>
- [32] Kunhong Li, Longguang Wang, Li Liu, Qing Ran, Kai Xu, and Yulan Guo. 2022. Decoupling makes weakly supervised local feature better. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15838–15848.
- [33] Xinghui Li, K. Han, Shuda Li, and Victor Adrian Prisacariu. 2020. Dual-Resolution Correspondence Networks. *ArXiv* abs/2006.08844 (2020). <https://api.semanticscholar.org/CorpusID:219708544>
- [34] Zhengqi Li and Noah Snavely. 2018. MegaDepth: Learning Single-View Depth Prediction From Internet Photos. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2041–2050. <https://doi.org/10.1109/CVPR.2018.00218>
- [35] Kang Liao, Lang Nie, Chunyu Lin, Zishuo Zheng, and Yao Zhao. 2023. RecRecNet: Rectangling rectified wide-angle images by thin-plate spline model and DoF-based curriculum learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10800–10809.
- [36] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. 2016. Feature Pyramid Networks for Object Detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)*, 936–944. <https://api.semanticscholar.org/CorpusID:10716717>
- [37] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. 2023. LightGlue: Local Feature Matching at Light Speed. In *ICCV*.
- [38] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019*. OpenReview.net. <https://openreview.net/forum?id=Bkg6RiCqY7>
- [39] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60 (2004), 91–110.
- [40] Zixin Luo, Tianwei Shen, Lei Zhou, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. 2019. Contextdesc: Local descriptor augmentation with cross-modality context. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2527–2536.
- [41] Zixin Luo, Lei Zhou, Xuyang Bai, Hongkai Chen, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. 2020. Aslfeat: Learning local features of accurate shape and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6589–6598.
- [42] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511809071>
- [43] Runyu Mao, Chen Bai, Yatong An, Fengqing Zhu, and Cheng Lu. 2022. 3DG-STFM: 3D Geometric Guided Student-Teacher Feature Matching. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII (Lecture Notes in Computer Science, Vol. 13688)*, Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer, 125–142. [https://doi.org/10.1007/978-3-031-19815-1\\_8](https://doi.org/10.1007/978-3-031-19815-1_8)
- [44] William McGill. 1954. Multivariate information transmission. *Transactions of the IRE Professional Group on Information Theory* 4, 4 (1954), 93–111.

- [45] Anastasya Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. 2017. Working hard to know your neighbor's margins: Local descriptor learning loss. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.), 4826–4837. <https://proceedings.neurips.cc/paper/2017/hash/831caa1b600f852b7844499430ecac17-Abstract.html>
- [46] Junjie Ni, Yijin Li, Zhaoyang Huang, Hongsheng Li, Hujun Bao, Zhaopeng Cui, and Guofeng Zhang. 2023. PATS: Patch Area Transportation with Subdivision for Local Feature Matching. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 17776–17786.
- [47] Lang Nie, Chunyu Lin, Kang Liao, Shuaicheng Liu, and Yao Zhao. 2021. Unsupervised Deep Image Stitching: Reconstructing Stitched Features to Images. *IEEE Trans. Image Process.* 30 (2021), 6184–6197. <https://doi.org/10.1109/TIP.2021.3092828>
- [48] Liam Paninski. 2003. Estimation of entropy and mutual information. *Neural computation* 15, 6 (2003), 1191–1253.
- [49] Hanchuan Peng, Fuhui Long, and Chris H. Q. Ding. 2005. Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 8 (2005), 1226–1238. <https://doi.org/10.1109/TPAMI.2005.159>
- [50] Jérôme Revaud, Philippe Weinzaepfel, César Roberto de Souza, No'e Pion, Gabriela Csurka, Johann Cabon, and M. Humenberger. 2019. R2D2: Repeatable and Reliable Detector and Descriptor. *ArXiv abs/1906.06195* (2019). <https://api.semanticscholar.org/CorpusID:189927786>
- [51] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. 2020. Efficient Neighbourhood Consensus Networks via Submanifold Sparse Convolutions. In *European Conference on Computer Vision*. <https://api.semanticscholar.org/CorpusID:216056155>
- [52] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. 2018. Neighbourhood consensus networks. *Advances in neural information processing systems* 31 (2018).
- [53] Edward Rosten and Tom Drummond. 2006. Machine Learning for High-Speed Corner Detection. In *Computer Vision – ECCV 2006*, Aleš Leonardis, Horst Bischof, and Axel Pinz (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 430–443.
- [54] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. 2011. ORB: An efficient alternative to SIFT or SURF. In *2011 International conference on computer vision*. Ieee, 2564–2571.
- [55] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. 2019. From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 12716–12725. <https://doi.org/10.1109/CVPR.2019.01300>
- [56] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. 2019. From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 12716–12725. <https://doi.org/10.1109/CVPR.2019.01300>
- [57] Paul-Edouard Sarlin, Ajaykumar Unagar, Måns Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, and Torsten Sattler. 2021. Back to the Feature: Learning Robust Camera Localization From Pixels To Pose. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 3247–3257. <https://doi.org/10.1109/CVPR46437.2021.00326>
- [58] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2019. SuperGlue: Learning Feature Matching With Graph Neural Networks. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 4937–4946. <https://api.semanticscholar.org/CorpusID:208291327>
- [59] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Fredrik Kahl, and Tomas Pajdla. 2018. Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 8601–8610.
- [60] Johannes Lutz Schönberger and Jan-Michael Frahm. 2016. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [61] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. 2016. Pixelwise View Selection for Unstructured Multi-View Stereo. In *European Conference on Computer Vision (ECCV)*.
- [62] Yanxing Shi, Junxiang Cai, Yoli Shavit, Tai-Jiang Mu, Wensen Feng, and Kai Zhang. 2022. ClusterGNN: Cluster-based Coarse-to-Fine Graph Neural Network for Efficient Feature Matching. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022).
- [63] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [64] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* 568 (2024), 127063.
- [65] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. 2021. LoFTR: Detector-Free Local Feature Matching with Transformers. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 8918–8927.
- [66] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.
- [67] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomás Pajdla, and Akihiko Torii. 2021. InLoc: Indoor Visual Localization with Dense Matching and View Synthesis. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 4 (2021), 1293–1307. <https://doi.org/10.1109/TPAMI.2019.2952114>
- [68] Donglin Tan, Jiangjiang Liu, Xingyu Chen, Chao Chen, Ruixin Zhang, Yunhang Shen, Shouhong Ding, and Rongrong Ji. 2022. ECO-TR: Efficient Correspondences Finding Via Coarse-to-Fine Refinement. *ArXiv abs/2209.12213* (2022). <https://api.semanticscholar.org/CorpusID:252531344>
- [69] Shitao Tang, Jiahui Zhang, Siyu Zhu, and Ping Tan. 2022. QuadTree Attention for Vision Transformers. *ArXiv abs/2201.02767* (2022).
- [70] Carl Toft, Will Maddern, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Tomás Pajdla, Fredrik Kahl, and Torsten Sattler. 2022. Long-Term Visual Localization Revisited. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 4 (2022), 2074–2088. <https://doi.org/10.1109/TPAMI.2020.3032010>
- [71] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. 2020. DISK: Learning local features with policy gradient. *Advances in Neural Information Processing Systems* 33 (2020), 14254–14265.
- [72] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Neural Information Processing Systems*.
- [73] Jorge R Vergara and Pablo A Estévez. 2014. A review of feature selection methods based on mutual information. *Neural computing and applications* 24 (2014), 175–186.
- [74] Jun Wang, Jin-Mao Wei, Zhenglu Yang, and Shu-Qin Wang. 2017. Feature Selection by Maximizing Independent Classification Information. *IEEE Trans. Knowl. Data Eng.* 29, 4 (2017), 828–841. <https://doi.org/10.1109/TKDE.2017.2650906>
- [75] Qianqian Wang, Xiaowei Zhou, Bharath Hariharan, and Noah Snavely. 2020. Learning feature descriptors using camera pose supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I* 16. Springer, 757–774.
- [76] Shuo Wang, Yongcai Wang, Xuewei Bai, and Deying Li. 2023. Communication Efficient, Distributed Relative State Estimation in UAV Networks. *IEEE Journal on Selected Areas in Communications* 41, 4 (2023), 1151–1166. <https://doi.org/10.1109/JSC.2023.3242708>
- [77] Shuo Wang, Yongcai Wang, Deying Li, and Qianchuan Zhao. 2023. Distributed Relative Localization Algorithms for Multi-Robot Networks: A Survey. *Sensors* 23, 5 (2023). <https://doi.org/10.3390/s23052399>
- [78] Yifan Wang, Xingyi He, Sida Peng, Dongli Tan, and Xiaowei Zhou. 2024. Efficient LoFTR: Semi-Dense Local Feature Matching with Sparse-Like Speed. In *CVPR*.
- [79] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. 2016. Lift: Learned invariant feature transform. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VI* 14. Springer, 467–483.
- [80] Jiahuan Yu, Jiahao Chang, Jianfeng He, Tianzhu Zhang, and Feng Wu. 2023. Adaptive Spot-Guided Transformer for Consistent Local Feature Matching. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 21898–21908.
- [81] Jiahui Zhang, Dawei Sun, Zixin Luo, Anbang Yao, Lei Zhou, Tianwei Shen, Yurong Chen, Long Quan, and Hongen Liao. 2019. Learning two-view correspondences and geometry using order-aware network. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5845–5854.
- [82] Yesheng Zhang and Xu Zhao. 2024. MESA: Matching Everything by Segmenting Anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20217–20226.
- [83] Yesheng Zhang, Xu Zhao, and Dahong Qian. 2023. Searching from area to point: A hierarchical framework for semantic-geometric combined feature matching. *arXiv preprint arXiv:2305.00194* (2023).
- [84] Zichao Zhang, Torsten Sattler, and Davide Scaramuzza. 2021. Reference Pose Generation for Long-term Visual Localization via Learned Features and View Synthesis. *Int. J. Comput. Vis.* 129, 4 (2021), 821–844. <https://doi.org/10.1007/S11263-020-01399-8>
- [85] Qunjie Zhou, Torsten Sattler, and Laura Leal-Taixe. 2021. Patch2pix: Epipolar-guided pixel-level correspondences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4669–4678.
- [86] Shengjie Zhu and Xiaoming Liu. 2023. PMatch: Paired Masked Image Modeling for Dense Geometric Matching. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 21909–21918. <https://api.semanticscholar.org/CorpusID:257833673>