Structured Epipolar Matcher for Local Feature Matching

Jiahao Chang[*], Jiahuan Yu[*], Tianzhu Zhang[†]

University of Science and Technology of China

Abstract

Local feature matching is challenging due to textureless and repetitive patterns. Existing methods focus on using appearance features and global interaction and matching, while the importance of geometry priors in local feature matching has not been fully exploited. Different from these methods, in this paper, we delve into the importance of geometry prior and propose Structured Epipolar Matcher (SEM) for local feature matching, which can leverage the geometric information in an iterative matching way. The proposed model enjoys several merits. First, our proposed Structured Feature Extractor can model the relative positional relationship between pixels and high-confidence anchor points. Second, our proposed Epipolar Attention and Matching can filter out irrelevant areas by utilizing the epipolar constraint. Extensive experimental results on five standard benchmarks demonstrate the superior performance of our SEM compared to state-of-the-art methods. Project page: https://sem2023.github.io.

1. Introduction

Local feature matching, which aims to find correspondences between a pair of images, is essential for many important tasks in computer vision, such as Structure-from-Motion (SfM) [34], 3D reconstruction [7], visual localization [33, 38], and pose estimation [12, 27]. Given its broad application, local feature matching has received significant attention, leading to the development of many research studies. Despite this, accurate local feature matching remains a challenging task, particularly in scenarios with poor texture, repetitive texture, illumination variations, and scale changes.

Numerous methods [8,19,28,30,35] have been proposed to overcome the challenges mentioned above, which can be divided into two major groups: detector-based methods [2,8,9,24,28,32] and detector-free methods [5,10,15, 19,29,30,35,48]. The detector-based approach consists of three separate stages: first, extracting keypoints from the

---
[*]Equal contribution
[†]Corresponding author

Figure 1. With the help of anchor points (marked in red), we can correctly match point A to point C and remove interference from point B.

images; then, generating descriptors of the keypoints; and finally, finding correspondences between keypoints using a matcher. The performance of the detector-based method highly relies on the quality of keypoints. Unfortunately, reliable keypoint detection in textureless or repetitive texture areas is a highly challenging problem that limits the final performance of detector-based methods. In comparison, detector-free approaches build dense correspondences between pixels instead of extracted keypoints. Recent works show that this method can handle poor textured regions better and demonstrate excellent performance. Some recent works [5, 16, 35, 46] leverage the Transformer architecture and shows impressive performance. However, most of these works use an all-pixel-to-all-pixel matching process, which introduces irrelevant pixels and negatively affects the result in some scenes, such as regions with repetitive textures.

Upon studying the previous matching methods, we have identified two issues that cannot be ignored in obtaining dense correspondences between images: (1) **How to extract more distinguishable features in areas with poor or repetitive textures.** Existing methods focus on extracting features from appearance, which is not distinguishable enough. As shown in motivation Figure 1, points A, B and C are similar in appearance, making it difficult to obtain accurate correspondence. In contrast, textured anchor points (marked in red) can easily match with the correct correspondence through appearance. We have discovered an interesting fact that by using the relative positional relationship between point A and other anchor points, we can easily determine the correct matching position of A. Therefore, it is necessary to consider the relative positional relationship

Figure 2. Using the epipolar geometry prior to calculate the candidate matching area of point D, we can exclude the influence of point F on D.

with the anchor point during matching, which can help us make better judgments. (2) **How to filter out irrelevant regions as much as possible during the matching process.** Existing methods usually use Transformers for global feature update and matching, which would make matching process influenced by unrelated areas. As shown in motivation Figure 2, the correct match of point D is point E. If point D aggregates the features of point F, it will confuse the network to determine the final matching result. Therefore, it is necessary to design a way to filter out the negative influence of irrelevant areas such as F.

To address the aforementioned issues, we propose the **Structured Epipolar Matcher** (SEM) with an **Iterative Epipolar Coarse Matching** stage, which consists of a **Structured Feature Extractor** and **Epipolar Attention/Matching Module**. In the Structured Feature Extractor, we select high-confidence matching pixels as anchor points, and encode the relative position between each point of the image and the anchor points in a scale and rotation invariant manner as a supplement to the appearance features, making the features more discriminating. In the Epipolar Attention/Matching Module, we consider the epipolar geometry prior to filter out candidate matching regions. First we select points with high confidence, which is used to calculate the relative pose of the camera. And then, we use the epipolar area corresponding to each point as the valid area (painted in blue in Figure 2) for feature interaction and matching. The entire Iterative Epipolar Coarse Matching process is iterated to gradually optimize the matching results.

The contributions of our method could be summarized into three-fold: (1) We proposed a novel Structured Epipolar Matcher (SEM), which includes a Structured Feature Extractor and Epipolar Attention/Matching Module in a unified architecture. (2) We design the Structured Feature Extractor, which can generate structured feature to complement the appearance features, making them more discriminating. Also, we design Epipolar Attention/Matching Module, which uses epipolar geometry prior to filter out irrelevant matching regions as much as possible, and Iterative Epipolar Coarse Matching process will gradually optimize

the matching results. (3) Extensive experimental results on four challenging benchmarks show that our proposed method performs favorably against state-of-the-art image matching methods.

2. Related Work

This section provides a brief overview of the methods related to local feature matching.

Local Feature Matching. The popular local feature matching methods can be divided into two paradigms: detector-based methods and detector-free methods. Detector-based methods involve three main stages: feature detection, feature description, and feature matching. Hand-crafted local features such as SIFT [22] and ORB [31] are the most widely used. Recently, learning-based local features like LIFT [47] and SuperPoint [8] have been shown to significantly improve performance compared to classical methods. Some researchers also focus on enhancing the feature matching stage, representative works include D2Net [9], R2D2 [28] and SuperGlue [32]. Nevertheless, detector-based methods rely on keypoint detectors, which limits the performance in challenging scenarios such as repetitive textures, weak textures, and illumination changes. On the contrary, Detector-free approaches directly match dense features between pixels without the local feature detector. While classical methods [14, 23] exist, few of them outperform detector-based methods. Learning-based methods have revolutionized this field, with cost-volume-based methods [19,30,42–44] and Transformer-based methods [5,16,35,46,48] leading the charge. While cost-volume-based methods archive promising performance, the limited perceptual field of CNNs is its inescapable shortcoming. Recently, Transformer-based methods overcome this problem and leads the benchmark. Given that detector-free methods have been shown to perform better in local feature matching, we have chosen this paradigm as our baseline approach, and use our proposed new module to overcome the challenging scenarios.

Geometry Prior in Local Feature Matching. The significance of geometric priors in local feature matching has been explored in previous works. GeoWrap [3] uses a learned pairwise warping to increase visual overlap between images. Toft et al. [41] propose a method to remove perspective distortion based on single-image depth estimation. RoRD [25] generate rotation-robust descriptors via orthographic viewpoint projection. COTR [16] also estimates scale by finding co-visible region, and uses recursive zooming to enable the matcher to sense geometry scale. Patch2Pix [50] uses the consistent with the epipolar geometry of an input image pair to supervise the training. In our research, we delve deeper into the role of geometric prior in local feature matching and propose a novel and effective method to leverage epipolar constraint and relative position

information in handling challenging scenes.

Iterative Optimization. Iterative optimization is a common paradigm in computer vision and has been explored in many previous work [16, 36, 37, 40, 43, 44]. For example, RAFT [40] iteratively updates a flow field through a recurrent unit that performs lookups on the correlation volumes, ACNe [36] uses an robust iterative optimization to build permutation-equivariant network, and NeuralBF [37] proposes an iterative bilateral filtering with learned kernels for instance segmentation on point clouds. The iterative optimization is also deployed by some works in local feature matching, such as PDC-Net [44], GLU-Net [43] and COTR [16]. In this work, we follow this tried and tested paradigm and combine it with geometry priors.

3. Our Approach

In this section, we introduce our proposed Structured Epipolar Matcher (SEM) for local feature matching. The overall architecture is illustrated in Figure 3. Here we first present the overall scheme of SEM in Section 3.1. In Section 3.2, we describe our Iterative Epipolar Coarse Matching in detail. Then we explain our Structured Feature Extractor in Section 3.3 and our Epipolar Attention/Matching in Section 3.4. At last, Section 3.5 states the loss used in our SEM.

3.1. Overview

As shown in Figure 3, the proposed SEM mainly consists of two stages, including a Iterative Epipolar Coarse Matching stage and refine Matching stage. Here we provide a concise overview of the complete process. For a pair of input images, reference image I_{ref} and source image I_{src}, we first use a Feature Pyramid Network (FPN) [21] to extract multi-level features of each image $\{f_{1/i}^{ref}, f_{1/i}^{src}\}$. Then the feature maps with the size of $1/32$ and $1/8$ are sent into the Iterative Epipolar Coarse Matching stage. With the help of $1/32$ feature maps, coarse matching matrix M are obtained from the feature maps with the size of $1/8$ through a Structured Feature Extractor, Epipolar Attention and Epipolar Matching. Finally, we refine our coarse matching results in refine matching, which is the same as LoFTR [35].

3.2. Iterative Epipolar Coarse Matching

Considering that global feature can help fine feature to locate correct matching area, we use a multi-level strategy [48] in Iterative Epipolar Coarse Matching. First, we apply one self-and-cross attention layer on the feature maps $\{f_{1/32}^{ref}, f_{1/32}^{src}\}$ and one self-and-cross linear attention layer on the feature maps $\{f_{1/8}^{ref}, f_{1/8}^{src}\}$. After that, features at two scales are updated from each other. This progress can

be described as following:

$$\hat{f}_{1/32}^t = f_{1/32}^t + \text{Conv}_{1\times1}(\text{Down}(f_{1/8}^t)), t \in \{ref, src\}$$
$$\hat{f}_{1/8}^t = f_{1/8}^t + \text{Conv}_{1\times1}(\text{Up}(f_{1/32}^t)), t \in \{ref, src\}$$
$$(1)$$

where $\hat{f}_{1/32}^t$ and $\hat{f}_{1/8}^t$ denotes the fused features with the size of $1/32$ and $1/8$. The purpose of the initialization is to establish the global connection in image or cross images. Next, we enter the process of iterative matching. During iterative process, $f_{1/32}^{ref}$ and $f_{1/32}^{src}$ are updated by self-and-cross attention layers to mainatin the global connection. Coarse matching matrix $M \in \mathbb{R}^{HW \times HW}$ is initialized by computing pointwise similarity from flattened $f_{1/8}^{ref}$ and $f_{1/8}^{src}$ and a dural-softmax operator [35]:

$$S(i,j) = \tau \langle f_{1/8}^{ref}(i), f_{1/8}^{src}(j)\rangle,$$
$$M(i,j) = \text{softmax}(S(i,:))(i,j) \cdot \text{softmax}(S(:,j))(i,j),$$
$$(2)$$

where i is a pixel in image I_{ref} and j is a pixel in image I_{src}. τ means the temperature coefficient. $S \in R^{HW \times HW}$ is the similarity matrix. As features on appearance, $f_{1/8}^{ref}$ and $f_{1/8}^{src}$ are supplied by structured feature from Structured Feature Extractor and become more discriminating. And then, in Epipolar Attention, we interact $f_{1/8}^{ref}$ and $f_{1/8}^{src}$ via cross attention within corresponding epipolar banded area, which will be detailed in Section 3.4. After updating features between different scales, in Epipolar Matching, we use $f_{1/8}^{ref}$ and $f_{1/8}^{src}$ to rewrite matching matrix M along epipolar banded areas, which is a sparse similarity calculation with a low computational cost. After several iterations of epipolar matching, we obtain accurate coarse matching results.

3.3. Structured Feature Extractor

As human beings, when we search for correspondences between a pair of images, we do not just compare pixels. Instead, we first look for conspicuous and easily matched anchor points in the images. By utilizing the relative positional relationships between pixels and anchor points, human beings can easily overcome the challenges posed by textureless and repetitive texture areas. This process is also inspiring for learning-based approaches. As shown in Figure 1, even though the appearance features of points B and C are not discriminating enough, suitable anchor points can be selected, and the relative positions can be utilized to build structured features, which can compensate for the shortage of appearance features, resulting in better correspondences in textureless and repetitive texture regions.

We propose the **Structured Feature Extractor** for above purpose. First we define the confidence of correspondence:

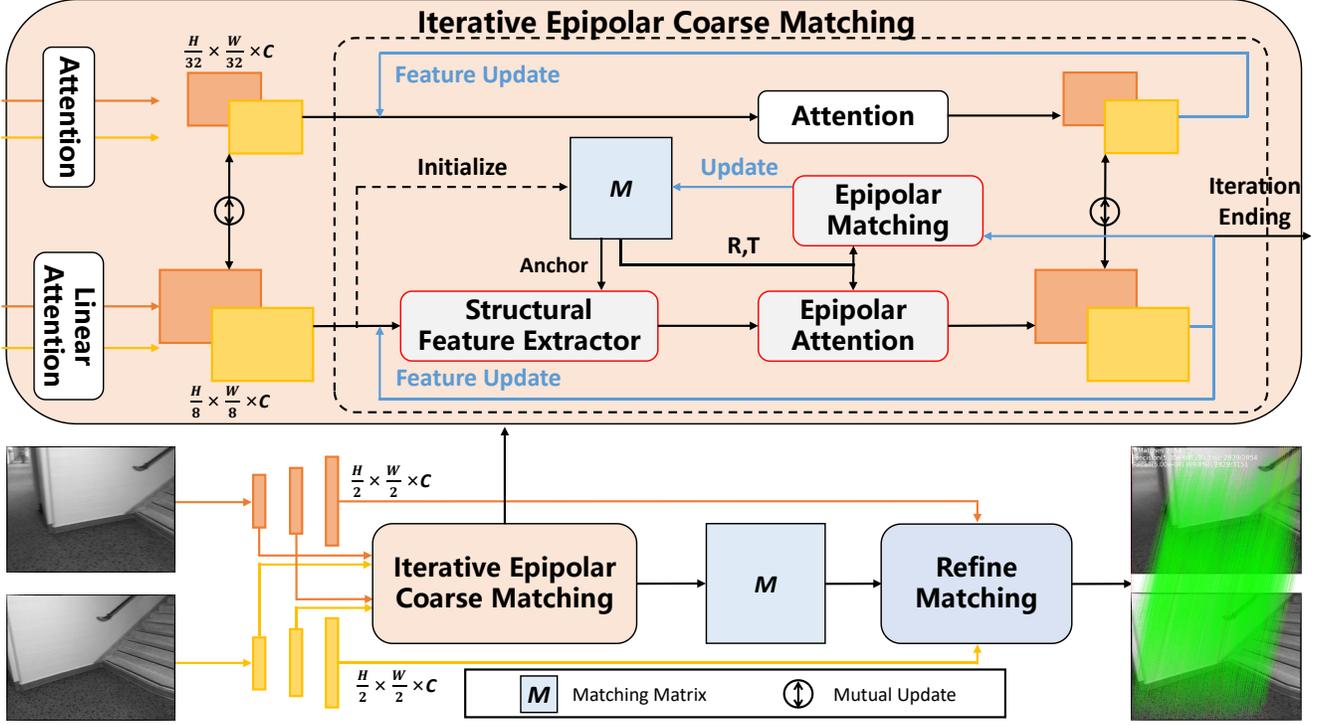$$\text{conf}(p_i) = \max\{M(i,j) \mid j \in \{1, 2, \cdots, HW\}\}, \quad (3)$$

Figure 3. The architecture of our SEM. Our SEM consists of two stages: Iterative Epipolar Coarse Matching stage and refine Matching stage. Here, "Attention" means vanilla self-and-cross attention, "Linear Attention" means linear self-and-cross attention. Please refer to the text for detailed architecture.
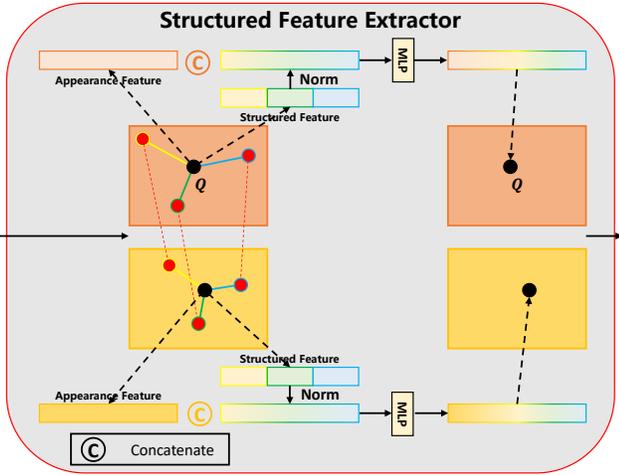

Figure 4. The schematic of our Structural Feature Extractor. We take the black point to examplify the construction of our structural feature, and we show the way to fuse appearance feature and structural feature. The red points means the selected anchor points.

where p_i is the pixel in reference image. As shown in Figure 4, given two image feature maps, f^{ref} and f^{src}, we first refer to matching matrix M to find high-confidence correspondence over a threshold of σ_h. Among these corre-spondences, we randomly select N_A pairs as anchor points (marked in red), denote as

$$\mathcal{A}^{ref} = \left\{ \left(x_i^{ref}, y_i^{ref} \right) \mid i = 1, 2, \cdots, N_A \right\},$$
$$\mathcal{A}^{src} = \left\{ \left(x_i^{src}, y_i^{src} \right) \mid i = 1, 2, \cdots, N_A \right\}. \quad (4)$$

Without loss of generality, we talk about the reference feature map f^{ref}. For any point Q in reference image whose coordinate is (x, y), we calculate the coordinate differences and Euclidean distances from Q to all anchor points:

$$\Delta X^{ref} = \left(x - x_1^{ref}, x - x_2^{ref}, \cdots, x - x_{N_A}^{ref} \right),$$
$$\Delta Y^{ref} = \left(y - y_1^{ref}, y - y_2^{ref}, \cdots, y - y_{N_A}^{ref} \right),$$
$$D^{ref} = \sqrt{(\Delta X)^2 + (\Delta Y)^2} = \left(d_1^{ref}, d_2^{ref}, \cdots, d_{N_A}^{ref} \right). \quad (5)$$

The structured feature f_{sf}^{ref} is then defined as

$$\Delta X_n^{ref} = \text{norm} \left(\Delta X^{ref} \right),$$
$$\Delta Y_n^{ref} = \text{norm} \left(\Delta Y^{ref} \right),$$
$$D_n^{ref} = \text{norm} \left(\Delta D^{ref} \right),$$
$$f_{sf}^{ref}(x, y) = \Delta X_n^{ref} \| \Delta Y_n^{ref} \| D_n^{ref}, \quad (6)$$

where $\|$ is concatenate operation, norm is the L_1 normalization operation. $f_{sf}^{src}(x, y)$ is calculated by similar process.
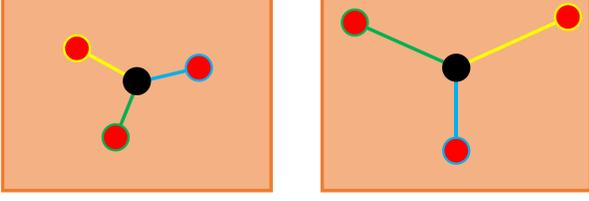
Figure 5. After scaling and rotating the image, the relative positional relationship between the anchor points (marked as red) and the given point (marked as black).
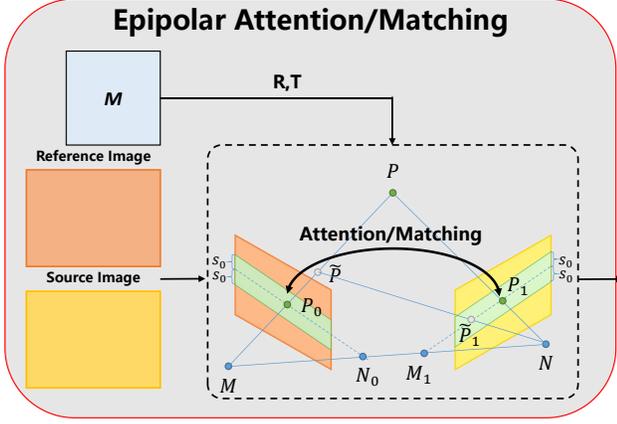

Figure 6. The illustration of our Epipolar Attention and Epipolar Matching.

Why do we need D and L_1 normalization to model the relative position instead of only using ΔX and ΔY? A suitable structured feature should satisfy both **scaling invariance** and **rotational invariance**. As shown in Figure 5, when an image is resized, ΔX, ΔY and D are all scaled proportionally. A L_1 normalization can solve the problems caused by scaling variation and make the structured feature more robust. Also, considering the rotation of the images, ΔX and ΔY will also change despite the normalization. This is the motivation why we introduce D, which is rotational invariant with the normalization.

Finally, the fused feature f_{fused}^{src} can be obtained by applying a MLP to fuse the appearance feature and structural feature:

$$f_{fused}^{src}(x, y) = \text{MLP}\left(f^{src}(x, y) \| f_{sf}^{src}(x, y)\right) \quad (7)$$

3.4. Epipolar Attention and Matching

Existing methods mostly perform global interaction between reference and source images, which will introduce irrelevant noisy points. In Figure 2, the point F has an negative impact on the matching result of point D. Therefore, we attempt to remove irrelevant regions during feature updating and matching as soon as possible. But how?

We think of some "good" points that are easily to match correctly. With the help of epipolar geometry prior, these "good" points can filter out irrelevant areas and reduce candidate regions for other points. Next, we will introduce the preliminaries of attention mechanism and epipolar line in Section 3.4.1. Finally, we states our epipolar attention and matching in Section 3.4.2.

3.4.1 Attention and Epipolar Line

Vanilla attention calculation can be devided into 3 steps. Inputs are converted to query Q, key K and value V by MLPs. To start with, the attention map is computed by dot production from query Q and key K. After the softmax operator, the attention map becomes the weight matrix. Finally, the output is weighted sum of value V with the weight matrix above. The whole process can be expressed as

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T)V. \quad (8)$$

The attention map can produce high computational and memory consumption, which is overwhelming for computer vision tasks. To resolve the problem above, Katharopoulos [17] propose linear attention, which approximates the softmax operator with the product of two kernel functions $\phi(\cdot) = \text{elu}(\cdot) + 1$ and uses the commutative law of matrix multiplication:

$$\text{Linear\_attention}(Q, K, V) = \phi(Q)(\phi(K^T)V), \quad (9)$$

However, some experiments [4, 11] have proved that linear attention will cause a slight decline in model performance.

In Figure 6, P is a point in the real world, P_0 and P_1 is a pair of matching points, which are also the projections of P on the reference and source image. M and N are optical centers of the reference and source camera. N_0 is the projection of N on the reference image. M_1 is the projection of M on the source image. N_0 and M_1 are a pair of epipolar points. $P_0 N_0$ and $P_1 M_1$ are a pair of epipolar lines. According to epipolar constraint, for any point in $P_0 N_0$, its matching point on the source image will definitely fall on $P_1 M_1$.

3.4.2 Epipolar Attention and Matching

In Figure 6, given Mating Matrix M, we first pick up "good" matching pairs whose confidence is above a threshold σ_h. With these matching pairs, we can obtain the relative pose of two images $\{R, T\}$. Without loss of generality, we take the point P_0 as an example to show the way to find the candidate matching area of P_0. The coordinate of P_0 on the reference image is $(x, y)^T$. Then we find a point $\widetilde{P}$ on MP whose reference camera coordinate is $K_{ref}^{-1} \cdot (x, y, 1)^T$, where K_{ref} denotes the intrinsics of the

reference camera. And the source camera coordinate of $\widetilde{P}$ is $R \cdot K_{ref}^{-1} \cdot (x, y, 1)^T + T$. Project $\widetilde{P}$ to the source image plane and we get $\widetilde{P}_1$. The coordinate of $\widetilde{P}_1$ on the source image $c_{\widetilde{P}_1}$ is

$$
\begin{aligned}
p_{\widetilde{P}_1} &= K_{src} \cdot (R \cdot K_{ref}^{-1} \cdot (x, y, 1)^T + T), \\
c_{\widetilde{P}_1} &= (\frac{p_{\widetilde{P}_1}(1)}{p_{\widetilde{P}_1}(2)}, \frac{p_{\widetilde{P}_1}(0)}{p_{\widetilde{P}_1}(2)}),
\end{aligned}
\tag{10}
$$

where K_{src} is the intrinsics of the source camera. Meanwhile, the coordinate of M_1 on the source image c_{M_1} is

$$
\begin{aligned}
p_{M_1} &= K_{src} \cdot T, \\
c_{M_1} &= (\frac{p_{M_1}(1)}{p_{M_1}(2)}, \frac{p_{M_1}(0)}{p_{M_1}(2)}).
\end{aligned}
\tag{11}
$$

Now we can easily get the slope k and intercept b of $M_1\widetilde{P}_1$:

$$
\begin{aligned}
k &= \frac{p_{\widetilde{P}_1}(1) - p_{M_1}(1)}{p_{\widetilde{P}_1}(0) - p_{M_1}(0)}, \\
b &= p_{M_1}(1) - k \cdot p_{M_1}(0).
\end{aligned}
\tag{12}
$$

Considering that the initial relative pose $\{R, T\}$ may be inaccurate, we expand the epipolar region to a banded region with tolerance s_0, whose boundaries can be represented as $\{k, b - s_0\}$ and $\{k, b + s_0\}$. At this point, we finally locate the candidate matching area of P_0. The candidate matching area of points on the source image can be obtained in the similar way. Each point performs cross attention with points in its corresponding candidate region. For later updating matching matrix M, each point is also matched against points in its corresponding candidate region.

3.5. Loss Function

Our loss function mainly includes two parts, iterative coarse matching loss and refine matching loss. Iterative coarse matching loss L_i is mainly used to supervise the matching matrix M updated by each iteration:

$$
L_i = \sum_k -\frac{1}{|M_c^{gt}|} \sum_{(i,j) \in M_c^{gt}} \log M^k(i, j), \tag{13}
$$

where M_c^{gt} is the ground truth matches at coarse resolution. $M^k(i, j)$ denotes the predicted matching matrix in the k-th iteration. Fine matching loss L_f is a weighted L_2 loss same as LoFTR [35]:

$$
L_f = \frac{1}{|M_f^{gt}|} \sum_{(i,j) \in M_f^{gt}} \frac{1}{\sigma^2(i)} \|j - j_{gt}\|_2, \tag{14}
$$

where M_f^{gt} is the ground truth matches at fine resolution and $\sigma^2(i)$ means the total variance of the heatmap. Our final loss is:

$$
L_{total} = L_i + L_f. \tag{15}
$$

4. Experiments

This section presents the comprehensive evaluation of our SEM through a series of experiments. To begin with, we describe the implementation details, and then we conduct experiments on four widely-used standard benchmarks for image matching. Moreover, a range of ablation studies are conducted to validate the effectiveness of each individual component.

4.1. Implementation Details

Our SEM was implemented using PyTorch [26] and was trained on the MegaDepth dataset [20]. During the training phase, we input images with a size of 832×832. Our CNN extractor is a deepened ResNet-18 [13] with features at $1/32$ resolution. We set the band width s_0 in Epipolar Attention to 10, and the number of anchor point N_A in Structured Feature Extractor is set to 32. The matching threshold σ_h in Iterable Epipolar Coarse Matching equals to 0.5, and the threshold in the matching module before refinement is set to 0.2. We use 4 iterartions in Iterative Epipolar Coarse Matching module. Our network is trained for 15 epochs with a batch size of 8, using the Adam [18] optimizer with an initial learning rate of 1×10^{-3}.

Table 1. Evaluation on HPatches [1] for homography estimation.

| Category | Method | Homography est. AUC | | | matches |
		@3px	@5px	@10px	
Detector-based	D2Net [9]+NN	23.2	35.9	53.6	0.2K
	R2D2 [28]+NN	50.6	63.9	76.8	0.5K
	DISK [45]+NN	52.3	64.9	78.9	1.1K
	SP [8]+SuperGlue [32]	53.9	68.3	81.7	0.6K
	Patch2Pix [50]	46.4	59.2	73.1	1.0K
Detector-free	Sparse-NCNet [29]	48.9	54.2	67.1	1.0K
	COTR [16]	41.9	57.7	74.0	1.0K
	DRC-Net [19]	50.6	56.2	68.3	1.0K
	LoFTR [35]	65.9	75.6	84.6	1.0K
	PDC-Net+ [44]	66.7	76.8	85.8	1.0k
	SEM(ours)	**69.6**	**79.0**	**87.1**	1.0K

4.2. Homography Estimation

Dataset and Metric. HPatches [1] is a widely used benchmark for evaluating image matching algorithms. In line with the methodology presented in [9], we have selected 56 sequences featuring significant viewpoint changes and 52 sequences with substantial illumination variation to assess the performance of our SEM, which was trained on the MegaDepth [20] dataset. We have adopted the same evaluation protocol used in the LoFTR approach [35]. In terms of metrics, we report the area under the cumulative curve (AUC) of the corner error distance up to 3, 5, and 10 pixels. In addition, we have limited the maximum number of output matches to 1,000 as LoFTR [35].

Results. Table 1 shows that our SEM sets a new state-of-the-art and notably performs better on HPatches [1] under all error threshold. This is a strong demenstration of the effectiveness of our method. SEM outperforms the previ-

Table 2. Evaluation on MegaDepth [20] for outdoor relative position estimation.

Category	Method	Pose estimation AUC		
		@5°	@10°	@20°
Detector-based	SP [8]+SuperGlue [32]	42.2	59.0	73.6
	SP [8]+SGMNet [4]	40.5	59.0	73.6
Detector-free	DRC-Net [19]	27.0	42.9	58.3
	PDC-Net+(H) [44]	43.1	61.9	76.1
	LoFTR [35]	52.8	69.2	81.2
	MatchFormer [46]	53.3	69.7	81.8
	QuadTree [39]	54.6	70.5	82.2
	ASpanFormer [5]	55.3	71.5	83.1
	SEM(ours)	**58.0**	**72.9**	**83.7**

Table 3. Evaluation on ScanNet [6] for indoor relative position estimation. * indicates models trained on MegaDepth [20].

Category	Method	Pose estimation AUC		
		@5°	@10°	@20°
Detector-based	D2-Net [9]+NN	5.3	14.5	28.0
	SP [8]+OANet [49]	11.8	26.9	43.9
	SP [8]+SuperGlue [32]	16.2	33.8	51.8
Detector-free	DRC-Net [19]*	7.7	17.9	30.5
	MatchFormer [46]*	15.8	32.0	48.0
	LoFTR-OT [35]*	16.9	33.6	50.6
	SEM(ours)*	**18.7**	**36.6**	**52.9**

ous state-of-the-art method by a margin of **2.9%**, **2.2%** and **1.3%** under 3, 5, 10 pixels respectively. This shows that our proposed Structured Feature Extractor can effectively model the geometric prior of the image content under illumination changes, while Epipolar Attention/Matching can effectively filter out the effect of irrelevant regions due to the viewpoint variations.

4.3. Relative Pose Estimation

Dataset and Metric. To evaluate the performance of our SEM in relative pose estimation, we conduct experiments on two datasets, MegaDepth [20] and ScanNet [6]. MegaDepth [20] is a large-scale outdoor dataset consisting 1 million internet images of 196 different outdoor scenes, reconstructed by COLMAP [34]. Depth maps as intermediate results can be used to obtain ground truth matches. We follow the same testing pairs and use the same 1500 image pairs as LoFTR [35]. All test images are resized to 1216 in their longer dimension. On the other hand, ScanNet [6] is usually used to validate the performance of indoor pose estimation, which is composed of monocular sequences with ground truth poses and depth maps. Wide baselines and extensive textureless regions in image pairs make ScanNet benchmark challenging. All test images are resized to 640×480. It is worth noting that we evaluate our SEM trained on MegaDepth [20] on ScanNet [6]. We report the AUC of pose error at thresholds $(5°, 10°, 20°)$ in both benchmark as LoFTR [35]. Our SEM with proposed modules achieves state-of-the-art performance on both datasets.

Results. As Table 2 illustrated, our SEM show an outstanding performance on MegaDepth [20] compared to other methods. Respectively, our method outperformances the previous best method with **2.7%** and **1.4%** in AUC@5°
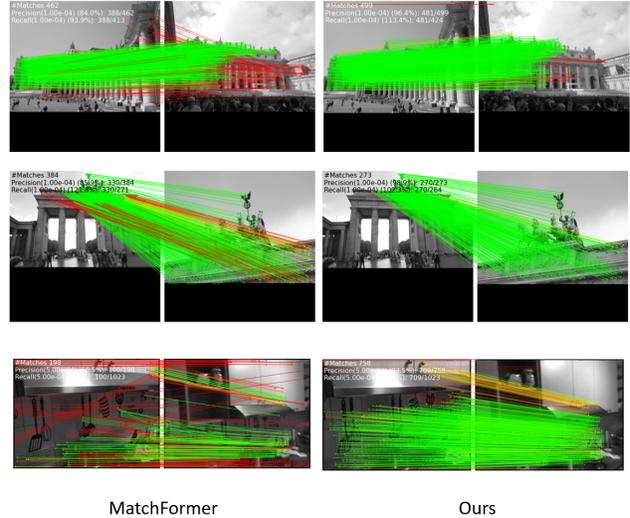

Figure 7. Visual qualitative comparison of MatchFormer [46] and our SEM on MegaDepth [20] dataset.

and AUC@10°. This result shows the outstanding performance of our method on outdoor scenes. Furthermore, Table 3 compares the performance of our proposed SEM with other state-of-the-art models on ScanNet [6] dataset. Similar with MegaDepth [20], our method ranks the first although the model is not trained on ScanNet [6] dataset, which demonstrates the strong generalization ability of our proposed SEM. The main challenge of indoor dataset lies in the widespread presence of textureless and repetitive texture regions, where structured information is critical. Our method generalizes well in indoor dataset thanks to the utilization of geometry prior in proposed Structured Feature Extractor and Epipolar Attention and Matching. In order to more strongly demonstrate the effectiveness of our proposed SEM, Figure 7 provides a qualitative result compared with other methods. Our method has an outstanding performance on textureless and reppetitive texture regions.

4.4. Visual Localization

Dataset and Metric. To evaluate the performance of our SEM in visual localization, we use the InLoc [38] dataset, which consists of 9972 RGBD images. Among them, 329 RGB images are selected as queries for visual localization. InLoc [38] presents challenges such as textureless regions and repetitive patterns under large viewpoint changes. We evaluate the performance of our SEM trained on MegaDepth [20] in the same way as LoFTR [35]. The metric used in InLoc [38] measures the percentage of images registered within given error thresholds.

Results. Our method's performance on the InLoc [38] benchmark is summarized in Table 4, where it achieves the best performance on DUC1 and performs comparably to state-of-the-art methods on DUC2. This demonstrates our

Table 4. Visual localization evaluation on the InLoc [38] benchmark.

Method	DUC1	DUC2
	$(0.25m, 10°)$ / $(0.5m, 10°)$ / $(1m, 10°)$	
LoFTR [35]	47.5 / 72.2 / 84.8	54.2 / 74.8 / **85.5**
MatchFormer [46]	46.5 / 73.2 / 85.9	**55.7** / 71.8 / 81.7
ASpanFormer [5]	51.5 / 73.7 / 86.4	55.0 / 74.0 / 81.7
SEM(ours)	**52.0 / 74.2 / 87.4**	50.4 / **76.3** / 83.2

Table 5. Ablation Study of each component on MegaDepth [20]. SF indicates Structured Feature extractor, and EAM refers to Epipolar Attention and Matching.

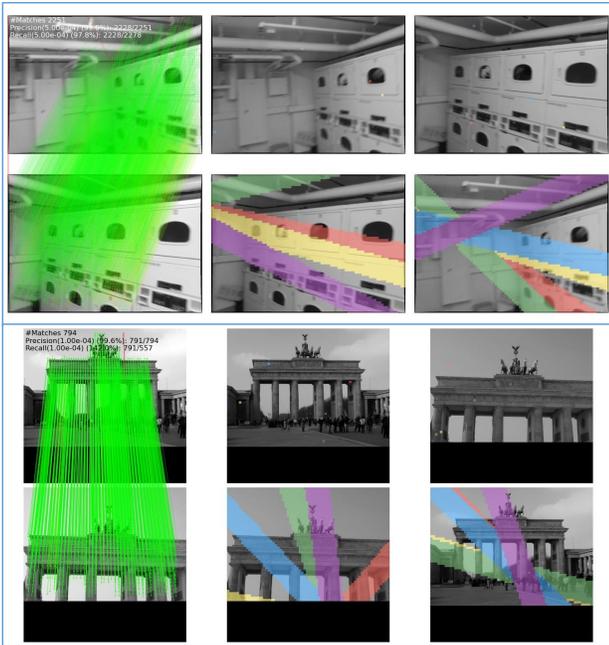| Index	Multi-Level	SF	EAM	Pose estimation AUC		
				@5°	@10°	@20°
1				45.6	62.2	75.3
2	✓			46.7	63.1	76.3
3	✓	✓		47.3	64.3	76.8
4	✓	✓	✓	**48.1**	**64.7**	**77.4**

Figure 8. Visualization of the epipolar banded areas corresponding to some points and matching results on MegaDepth [20] and ScanNet [6]. Points and corresponding epipolar regions are identified by the same color.

method's strong generalization ability in visual localization, even in challenging indoor scenes.

4.5. Ablation Study

To deeply analyze and evaluate the effectiveness of each component in SEM, we conducted detailed ablation studies on MegaDepth [20]. Here, we use images with a size of 544 for both training and evaluation. In Table 5, we gradually added the proposed components to the baseline to analyze their impact.

Effectiveness of Structured Feature Extractor. According to [5], we learn that the multi-level strategy is

Table 6. Ablation Study with different s_0 in Epipolar Attention on MegaDepth [20].

s_0	Pose estimation AUC		
	@5°	@10°	@20°
5	45.6	62.7	76.2
10	**48.1**	**64.7**	**77.4**
15	47.5	64.3	77.2
20	46.7	62.4	76.4

useful in local matching, which is verfied by the results of Index-2 and Index-1. Comparing the results of Index-2 and Index-1, we find that the proposed Structured Feature Extractor helps during the inference phase, increasing performance with an improvement of +0.6 AUC@5°, +1.2 AUC@10° and +0.5 AUC@20°.

Effectiveness of Epipolar Attention. As shown in Table 5, We replace the original linear attention and global matching with our Epipolar Attention and Epipolar Matching in Index-4. Comparing the results of Index-4 and Index-3, we can see that the performance is improved, which indicates the effectiveness of our proposed Epipolar Attention and Epipolar Matching. In Figure 8, we visualize the matching results for several pair of images. Specifically, we randomly select points in the reference image and identify their corresponding epipolar band regions (marked by same color) in the source image. For the points in the source image, the corresponding area is also found in the reference image. The attention is computed within these points and their corresponding band regions. The visualization results show that our module can effectively utilize the relative camera poses to filter out irrelevant regions. To further explore the effect of eliminating irrelevant regions, we explored the effect of different s_0 on MegaDepth dataset. As Table 6 shows, an appropriate value for s_0 is crucial as it can balance the filtering of unwanted regions and tolerance to camera pose estimation errors. If s_0 is too small, it may lead to the exclusion of correct matches when the initial camera pose estimation is imprecise. On the other hand, if s_0 is too large, it may not effectively filter out irrelevant regions. This shows that the modules we designed play the role we expected.

5. Conclusion

In this paper, we propose a novel Structured Epipolar Matcher (SEM) for local feature matching utilizing the geometry priors. To make features more discriminating, we design a Structured Feature Extractor to complement the appearance features. To exclude the regions of interference as much as possible, we propose Epipolar Attention and Matching in the iterable coarse matching stage. Extensive experimental results on five benchmarks demonstrate the effectiveness of the proposed method.

References

[1] 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 Conference on Computer Vision and Pattern Recognition*, pages 5173–5182, 2017. 6

[2] Axel Barroso-Laguna, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Key. net: Keypoint detection by handcrafted and learned cnn filters. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5836–5844, 2019. 1

[3] Gabriele Berton, Carlo Masone, Valerio Paolicelli, and Barbara Caputo. Viewpoint invariant dense matching for visual geolocalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12169–12178, 2021. 2

[4] 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. 5, 7

[5] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 20–36. Springer, 2022. 1, 2, 7, 8

[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 7, 8

[7] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics (ToG)*, 36(4):1, 2017. 1

[8] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 224–236, 2018. 1, 2, 6, 7

[9] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint detection and description of local features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2, 6, 7

[10] Johan Edstedt, Mårten Wadenbäck, and Michael Felsberg. Deep kernelized dense geometric matching. *arXiv preprint arXiv:2202.00667*, 2022. 1

[11] Hugo Germain, Vincent Lepetit, and Guillaume Bourmaud. Visual correspondence hallucination. *arXiv preprint arXiv:2106.09711*, 2021. 5

[12] Alexander Grabner, Peter M Roth, and Vincent Lepetit. 3d pose estimation and 3d model retrieval for objects in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3022–3031, 2018. 1

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6

[14] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981. 2

[15] Shuaiyi Huang, Qiuyue Wang, Songyang Zhang, Shipeng Yan, and Xuming He. Dynamic context correspondence network for semantic alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2010–2019, 2019. 1

[16] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. Cotr: Correspondence transformer for matching across images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6207–6217, 2021. 1, 2, 3, 6

[17] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pages 5156–5165. PMLR, 2020. 5

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[19] Xinghui Li, Kai Han, Shuda Li, and Victor Prisacariu. Dual-resolution correspondence networks. *Advances in Neural Information Processing Systems*, 33, 2020. 1, 2, 6, 7

[20] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2041–2050, 2018. 6, 7, 8

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 3

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

[23] Bruce D Lucas, Takeo Kanade, et al. *An iterative image registration technique with an application to stereo vision*, volume 81. Vancouver, 1981. 2

[24] Yuki Ono, Eduard Trulls, Pascal Fua, and Kwang Moo Yi. Lf-net: learning local features from images. In *Advances in Neural Information Processing Systems*, pages 6237–6247, 2018. 1

[25] Udit Singh Parihar, Aniket Gujarathi, Kinal Mehta, Satyajit Tourani, Sourav Garg, Michael Milford, and K Madhava Krishna. Rord: Rotation-robust descriptors and orthographic views for local feature matching. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1593–1600. IEEE, 2021. 2

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming

Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[27] Mikael Persson and Klas Nordberg. Lambda twist: An accurate fast robust perspective three point (p3p) solver. In *Proceedings of the European Conference on Computer Vision*, pages 318–332, 2018. 1

[28] Jerome Revaud, Philippe Weinzaepfel, César Roberto de Souza, and Martin Humenberger. R2D2: repeatable and reliable detector and descriptor. In *Advances in Neural Information Processing Systems*, 2019. 1, 2, 6

[29] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. Efficient neighbourhood consensus networks via submanifold sparse convolutions. In *Proceedings of the European Conference on Computer Vision*, pages 605–621, 2020. 1, 6

[30] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. In *Advances in Neural Information Processing Systems*, pages 1658–1669, 2018. 1, 2

[31] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 2

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

[33] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al. Benchmarking 6dof outdoor visual localization in changing conditions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8601–8610, 2018. 1

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

[35] 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, 3, 6, 7, 8

[36] Weiwei Sun, Wei Jiang, Eduard Trulls, Andrea Tagliasacchi, and Kwang Moo Yi. Acne: Attentive context normalization for robust permutation-equivariant learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11286–11295, 2020. 3

[37] Weiwei Sun, Daniel Rebain, Renjie Liao, Vladimir Tankovich, Soroosh Yazdani, Kwang Moo Yi, and Andrea Tagliasacchi. Neuralbf: Neural bilateral filtering for top-down instance segmentation on point clouds. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 551–560, 2023. 3

[38] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. Inloc: Indoor visual localization with dense

matching and view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7199–7209, 2018. 1, 7, 8

[39] Shitao Tang, Jiahui Zhang, Siyu Zhu, and Ping Tan. Quadtree attention for vision transformers. *arXiv preprint arXiv:2201.02767*, 2022. 7

[40] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 3

[41] Carl Toft, Daniyar Turmukhambetov, Torsten Sattler, Fredrik Kahl, and Gabriel J Brostow. Single-image depth prediction makes feature matching easier. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 473–492. Springer, 2020. 2

[42] Prune Truong, Martin Danelljan, Luc V Gool, and Radu Timofte. Gocor: Bringing globally optimized correspondence volumes into your neural network. *Advances in Neural Information Processing Systems*, 33:14278–14290, 2020. 2

[43] Prune Truong, Martin Danelljan, and Radu Timofte. Glunet: Global-local universal network for dense flow and correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6258–6268, 2020. 2, 3

[44] Prune Truong, Martin Danelljan, Radu Timofte, and Luc Van Gool. Pdc-net+: Enhanced probabilistic dense correspondence network. *arXiv preprint arXiv:2109.13912*, 2021. 2, 3, 6, 7

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

[46] Qing Wang, Jiaming Zhang, Kailun Yang, Kunyu Peng, and Rainer Stiefelhagen. Matchformer: Interleaving attention in transformers for feature matching. In *Proceedings of the Asian Conference on Computer Vision*, pages 2746–2762, 2022. 1, 2, 7, 8

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

[48] Jiahuan Yu, Jiahao Chang, Jianfeng He, Tianzhu Zhang, and Feng Wu. Adaptive spot-guided transformer for consistent local feature matching. *arXiv preprint arXiv:2303.16624*, 2023. 1, 2, 3

[49] Jiahui Zhang, Dawei Sun, Zixin Luo, Anbang Yao, Lei Zhou, Tianwei Shen, Yurong Chen, Long Quan, and Hongen Liao. Learning two-view correspondences and geometry using order-aware network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5845–5854, 2019. 7

[50] Qunjie Zhou, Torsten Sattler, and Laura Leal-Taixe. Patch2pix: Epipolar-guided pixel-level correspondences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4669–4678, 2021. 2, 6