D^{2}Former: Jointly Learning Hierarchical Detectors and Contextual Descriptors via Agent-based Transformers

Jianfeng He[1,*], Yuan Gao[1,*], Tianzhu Zhang[1,2,†], Zhe Zhang[2], Feng Wu[1]
[1] University of Science and Technology of China [2] Deep Space Exploration Laboratory
{hejf, wazs98}@mail.ustc.edu.cn, {tzzhang, fengwu}@ustc.edu.cn, cnclepzz@126.com

Abstract

Establishing pixel-level matches between image pairs is vital for a variety of computer vision applications. However, achieving robust image matching remains challenging because CNN extracted descriptors usually lack discriminative ability in texture-less regions and keypoint detectors are only good at identifying keypoints with a specific level of structure. To deal with these issues, a novel image matching method is proposed by Jointly Learning Hierarchical Detectors and Contextual Descriptors via Agent-based Transformers (D^{2}Former), including a contextual feature descriptor learning (CFDL) module and a hierarchical keypoint detector learning (HKDL) module. The proposed D^{2}Former enjoys several merits. First, the proposed CFDL module can model long-range contexts efficiently and effectively with the aid of designed descriptor agents. Second, the HKDL module can generate keypoint detectors in a hierarchical way, which is helpful for detecting keypoints with diverse levels of structures. Extensive experimental results on four challenging benchmarks show that our proposed method significantly outperforms state-of-the-art image matching methods.

1. Introduction

Finding pixel-level matches accurately between images depicting the same scene is a fundamental task with a wide range of 3D vision applications, such as 3D reconstruction [35, 53, 55], simultaneous localization and mapping (SLAM) [15, 25, 39], pose estimation [13, 29], and visual localization [35, 43]. Owing to its broad real-world applications, the image matching task has received increasing attention in the past decades [9, 16, 31, 33, 34]. However, realizing robust image matching remains difficult due to various challenges such as illumination changes, viewpoint transformations, poor textures and scale variations.

To conquer the above challenges, tremendous image matching approaches have been proposed [7, 9, 12, 16, 31, 34, 42], among which some dense matching methods [7, 16, 42] are proposed to consider all possible matches adequately and have achieved great success. However, because of the large matching space, these dense matching methods are expensive in computation cost and memory consumption. To achieve high efficiency, we notice that the detector-based matching methods [4, 9, 20, 31] can effectively reduce the matching space by designing keypoint detectors to extract a relatively small keypoint set for matching, thus having high research value. Generally, existing detector-based matching methods can be categorized into two main groups including detect-then-describe approaches [18, 37, 40, 41, 54] and detect-and-describe approaches [12, 20, 31]. Detect-then-describe approaches refer to first detect repeatable keypoints [3, 5, 18], and then keypoint features [19, 23, 28] are represented by describing image patches extracted around these keypoints. In this way, matches can be established by nearest neighbor search according to the Euclidean distance between keypoint features. However, since the keypoint detector and descriptor are usually designed separately in detect-then-describe approaches, keypoint features may not be suitable for detected keypoints, resulting in poor performance under extreme appearance changes. Differently, detect-and-describe approaches [12, 31] are proposed to tightly couple the keypoint detector learning with the descriptor learning. For example, both D2-Net [12] and R2D2 [31] use a single convolutional neural network (CNN) for joint detection and description. These methods

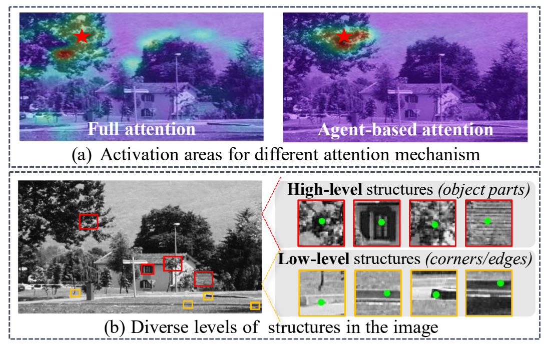

Figure 1. Illustration of our motivation. (a) shows the comparison between our proposed agent-based attention and full attention, where full attention would aggregate features from irrelevant areas. (b) shows the diverse structures contained in the image.

*Equal Contribution
†Corresponding Author

have achieved great performance mainly benefiting from the superiority of joint learning. However, the receptive field of features extracted by CNN is limited, and keypoint detectors are usually learned at a single feature scale, which restricts further progress.

Based on the above discussions, we find that both the descriptor and detector learning are crucial for detector-based matching methods. To make image matching more robust to real-world challenges, the following two issues should be taken into consideration carefully. (1) **How to learn feature descriptors with long-range dependencies.** Current detector-based matching methods [9, 12, 31] usually use CNN to extract image features. Due to the limited receptive field of CNN, the extracted features would lack discriminative ability in texture-less regions. Although several works [11, 48] leverage full attention to capture long-range dependencies, as shown in Figure 1 (a), full attention may aggregate irrelevant noise, which is harmful to learn discriminative features. Besides, the computation cost of full attention is rather expensive. Therefore, an effective and efficient attention mechanism needs to be proposed urgently to capture long-range contexts of features. (2) **How to learn keypoint detectors suitable for various structures.** As shown in Figure 1 (b), there are diverse levels of structures in an image, from simple corner points (low-level structures) to complex object parts (high-level structures). However, existing keypoint detectors are usually good at identifying keypoints with a specific level of structure, such as corners (or edges) [14, 49], and blobs [18, 21]. Thus, it is necessary to learn hierarchical keypoint detectors to detect keypoints with different structures.

Motivated by the above observations, we propose a novel model by Jointly Learning Hierarchical Detectors and Contextual Descriptors via Agent-based Transformers (D^{2}Former) for image matching, which mainly consists of a contextual feature descriptor learning (CFDL) module and a hierarchical keypoint detector learning (HKDL) module. In the **contextual feature descriptor learning module**, it is proposed to capture reliable long-range contexts efficiently. Specifically, original image features are first extracted by a standard CNN. Then, we design a set of descriptor agents to aggregate contextual information by interacting with image features via attention mechanisms. Finally, contextual features are obtained by fusing the updated descriptor agents into original features. In the **hierarchical keypoint detector learning module**, it is proposed to detect keypoints with different structures, which can achieve robust keypoint detection. Specifically, we design a set of detector agents, which can interact with contextual features via attention mechanisms to obtain low-level keypoint detectors. Then, we aggregate these low-level keypoint detectors to form high-level keypoint detectors in a hierarchical way. Finally, the hierarchical keypoint detectors are obtained by gathering keypoint detectors from different levels.

The main contributions of this work can be summarized as follows. (1) A novel image matching method is proposed by jointly learning hierarchical detectors and contextual descriptors via agent-based Transformers, which can extract discriminative feature description and realize robust keypoint detection under some extremely challenging scenarios. (2) The proposed CFDL module can model long-range dependencies effectively and efficiently with the aid of designed descriptor agents. And the HKDL module can generate keypoint detectors in a hierarchically aggregated manner, so that keypoints with diverse levels of structures can be detected. (3) Extensive experimental results on four challenging benchmarks show that our proposed method performs favorably against state-of-the-art detector-based image matching methods.

2. Related Work

In this section, we briefly overview detect-then-describe image matching, detect-and-describe image matching and applications of Transformers in vision-related tasks.

Detect-then-describe image matching. Detect-then-describe methods [18, 37, 40, 41, 54] generally consist of three stages: detection, description, and matching. First, a set of salient and repeatable keypoints are first detected by a keypoint detector [3, 14, 46], then keypoint descriptors are computed based on a patch centered around each keypoint [19, 23, 28, 45], and finally, keypoints and feature descriptors are paired together to form a candidate matching space from which matches with high confidence can be retrieved through the mutual nearest neighbour criterion [24]. Traditional methods utilize handcrafted keypoint detectors and descriptors [18], which makes them limited by the priori knowledge. To alleviate the problem, several learning-based methods have been proposed, which can learn the keypoint detector [37, 54] or the feature descriptor [40, 41] in a data-driven manner. For example, LIFT [51] designs three differentiable branches, where keypoints are first detected by a convolutional branch, and cropped regions are then fed to the second branch to estimate the orientation. Finally, the third convolutional branch is used to perform description. However, detect-then-describe methods typically perform poorly under extreme appearance changes because repeatable keypoints are hard to detect. Besides, due to the separate design of keypoint detectors and feature descriptors, keypoint features may not be suitable for detected keypoints. Thus, in our work, we propose to learn keypoint detection and description jointly in a unified framework.

Detect-and-describe image matching. Recently, several methods [9, 12, 20, 26, 31] propose to tightly couple keypoint detection and description. Among these methods, D2-Net [12] proposes to utilize a single CNN for jointly optimizing detection and description, and demonstrates that the describe-and-detect strategy performs significantly better under challenging conditions. Further, R2D2 [31] is proposed to extract feature descriptors from the standard CNN backbone and learn a keypoint detector ($1\times$ 1 convolutional kernel) by constraining the detector to be both repeatable and reliable. However, the detector is learned and output from a fixed feature resolution, which limits the detection of keypoints with diverse levels of structures. Although ASLFeat [20] proposes keypoint detection on image features with different resolutions, the features are directly obtained from the CNN backbone, which may lack discriminative ability in texture-less regions. Besides, keypoint detectors of ASLFeat are learned independently on features at different resolutions without interaction, which limits the detector to perceive various levels of structures. Differently, our proposed contextual feature descriptor learning module can model long-range dependencies effectively and efficiently. And the designed hierarchical keypoint detector learning module can generate keypoint detectors in a hierarchically aggregated manner to identify keypoints with diverse levels of structures, which is vital for detector-based image matching approaches.

Transformers in vision-related tasks. Transformers [48] were initially widely used in the natural language processing field, which has achieved great success [10]. Due to their powerful global interaction capabilities, Transformers have gained increasing attention to a variety of computer vision tasks, such as object detection [6, 22] and image classification [11]. As a representative work, DETR [6] innovatively views object detection as a direct set prediction problem, and adopts an encoder-decoder architecture based on Transformers. Thanks to the attention mechanisms [1] which can model long-range dependencies, DETR [6] has successfully achieved state-of-the-art performance. Recently, attention mechanisms have been also introduced to the image matching task, where LoFTR [42] and ASpanFormer [7] are representative works. As can be seen, the global interaction ability of attention mechanism is useful for vision-based tasks. Thus, in this paper, we introduce the attention mechanisms to the detector-based image matching task, which can help learn discriminative feature descriptors with long-range dependencies. And hierarchical keypoint detectors can be learned by exploiting the global interaction ability of the attention mechanism, which is helpful for detecting keypoints from different structures.

3. Our Approach

In this section, we present our proposed method by Jointly Learning Hierarchical Detectors and Contextual Descriptors via agent-based Transformers for image matching. The overall architecture is illustrated in Figure 2.

3.1. Overview

As shown in Figure 2, our proposed model mainly consists of a contextual feature descriptor learning (CFDL) module and a hierarchical keypoint detector learning (HKDL) module. Given an input image $\mathbf{I}$, we first extract its original image features $\widehat{\mathbf{F}}$ via a feature extractor inspired by R2D2 [31]. Then, the image features $\widehat{\mathbf{F}}$ are flattened to $\mathbb{R}^{d\times hw}$ and are sent into the proposed CFDL module to generate contextual feature descriptors. Specifically, we first define a set of descriptor agents $\widehat{\mathbf{A}} \in \mathbb{R}^{d\times M}$ in the CFDL module, which can interact with flattened features $\widehat{\mathbf{F}}$ via an attention operation to obtain updated descriptor agents $\mathbf{A}$. The similarity $\mathbf{S}$ between features $\widehat{\mathbf{F}}$ and updated descriptor agents $\mathbf{A}$ is then calculated. And the final contextual feature descriptors $\mathbf{F} \in \mathbb{R}^{d\times h\times w}$ are obtained by a weighted sum of $\mathbf{A}$ based on the calculated similarity. After obtaining the contextual feature descriptors, we aim to produce hierarchical keypoint detectors in the HKDL module. Specifically, we first down-sample the contextual features $\mathbf{F}$ with convolutions (Convs) and obtain features $\mathbf{F}^l$ with different resolutions. A set of detector agents $\widehat{\mathbf{D}}^l$ is then defined by leveraging the agent initialization strategy. Next, for each level, the detector agents $\widehat{\mathbf{D}}^l$ are used to interact with the low-level keypoint detectors $\mathbf{D}^{l-1}$ via the detector decoder to produce high-level keypoint detectors $\mathbf{D}^l$. Finally, we can generate hierarchical keypoint detectors $\mathbf{D}$ by concatenating keypoint detectors $\mathbf{D}^l$ at different levels.

3.2. Contextual Feature Descriptor Learning

In order to capture long-range contexts efficiently and effectively, we adopt an agent-based attention mechanism in the proposed contextual feature descriptor learning (CFDL) module. Given flattened image features $\widehat{\mathbf{F}} \in \mathbb{R}^{d\times hw}$, we first design M descriptor agents $\widehat{\mathbf{A}} \in \mathbb{R}^{d\times M}$ to interact with $\widehat{\mathbf{F}}$ via the attention operation, where descriptor agents are initialized with a set of learnable parameters [50]. Specifically, keys and values arise from image features $\widehat{\mathbf{F}}$, and queries arise from the descriptor agents $\widehat{\mathbf{A}}$. Formally,

$$\mathbf{Q} = \mathbf{W}^{\mathcal{Q}}\widehat{\mathbf{A}}, \mathbf{K} = \mathbf{W}^{\mathcal{K}}\widehat{\mathbf{F}}, \mathbf{V} = \mathbf{W}^{\mathcal{V}}\widehat{\mathbf{F}}, \tag{1}$$

where $\mathbf{W}^{\mathcal{Q}} \in \mathbb{R}^{d_k\times d}, \mathbf{W}^{\mathcal{K}} \in \mathbb{R}^{d_k\times d}, \mathbf{W}^{\mathcal{V}} \in \mathbb{R}^{d\times d}$ are linear projections. Then, the descriptor agents are updated to obtain $\mathbf{A}$ in the following way,

$$\mathbf{A} = \mathrm{Attention}\left(\mathbf{Q}, \mathbf{K}, \mathbf{V}\right) = \mathbf{V}\cdot \mathrm{Softmax}(\mathbf{K}^\top\mathbf{Q}). \tag{2}$$

Motivated by [48], Eq. (2) is implemented with the multi-head attention. In this way, $\mathbf{A}$ can effectively capture long-range contexts. Thus, we update original features $\widehat{\mathbf{F}}$ by fusing $\mathbf{A}$ to obtain contextual feature descriptors. To this end, we calculate similarity scores $\mathbf{S}$ between $\widehat{\mathbf{F}}$ and updated descriptor agents $\mathbf{A}$. And original features $\widehat{\mathbf{F}}$ are updated as follows,

$$\mathbf{F} = \widehat{\mathbf{F}} + \mathbf{A}\mathbf{S}, \text{ where } \mathbf{S} = \mathbf{A}^\top\widehat{\mathbf{F}}. \tag{3}$$

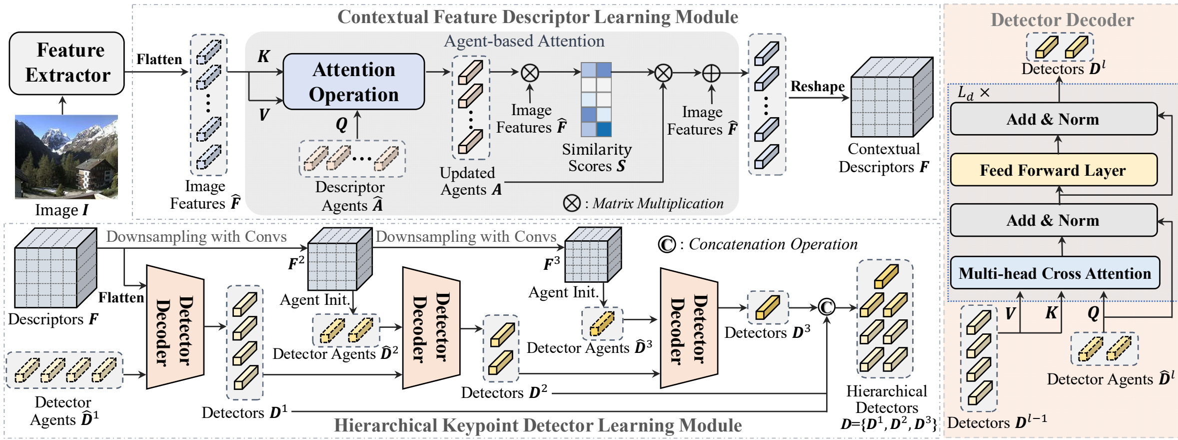

Figure 2. The architecture of our D²Former consists of two major components, including a contextual feature descriptor learning (CFDL) module and a hierarchical keypoint detector learning (HKDL) module. The image $\mathbf{I}$ is first sent into a feature extractor to obtain features $\widehat{\mathbf{F}}$. Then, in the CFDL module, we define a set of descriptor agents $\widehat{\mathbf{A}}$ to interact with flattened features $\widehat{\mathbf{F}}$, and use updated agents $\mathbf{A}$ to produce contextual descriptors $\mathbf{F}$. Next, in the HKDL module, for each level $l \in \{1, 2, 3\}$, we leverage an agent initialization (Agent Init.) strategy to generate detector agents $\widehat{\mathbf{D}}^l$, which are used to interact with $\mathbf{D}^{l-1}$ via the detector decoder to produce detectors $\mathbf{D}^l$. Finally, we can generate hierarchical detectors $\mathbf{D}$ by concatenating detectors $\mathbf{D}^l$ at different levels. For more details, please refer to the text.

The above operations (Eq. (1) to Eq. (3)) constitute the agent-based attention mechanism. And the final contextual descriptors are obtained by reshaping $\mathbf{F}$ to $\mathbb{R}^{d\times h\times w}$.

Discussions. Here, we discuss differences between our proposed agent-based attention mechanism and the full attention mechanism [11,48] to model long-range dependencies. In terms of efficiency, it is well known that the complexity of full attention [11] is $\mathcal{O}((hw)^2)$, where (h, w) is the resolution of features. Differently, by analyzing Eq. (2) and Eq. (3), our agent-based attention has the complexity of $\mathcal{O}((hw)\cdot M)$, where M is the number of descriptor agents. Since M is far smaller than hw, the agent-based attention is more efficient than the full attention. Besides, as shown in Figure 1, the agent-based attention can focus more on valid regions than full attention. Therefore, with the aid of proposed agent-based attention mechanism, we can capture long-range contexts efficiently and effectively to produce contextual descriptors.

3.3. Hierarchical Keypoint Detector Learning

After obtaining the contextual feature descriptors $\mathbf{F}$, we aim to learn hierarchical keypoint detectors, which is suitable for detecting keypoints with various structures. To this end, we aggregate low-level keypoint detectors to form high-level keypoint detectors in a hierarchical way. Specifically, we first leverage an agent initialization strategy to generate detector agents $\widehat{\mathbf{D}}^l$ at the l^{th} level, where $l \in \{1, 2, 3\}$. Then, these detector agents are interacted with the $(l-1)^{th}$ level keypoint detectors $\mathbf{D}^{l-1}$ via the designed detector decoder to produce the l^{th} level keypoint detectors $\mathbf{D}^l$. Finally, we can generate hierarchical keypoint detectors $\mathbf{D}$ by concatenating keypoint detectors $\mathbf{D}^l$ at different levels. Below, we introduce the designs of agent initialization and detector decoder in detail.

Agent initialization. For the first level ($l = 1$), the detector agents $\widehat{\mathbf{D}}^1$ are simply initialized with a set of learnable parameters. For other levels ($l \geq 2$), we generate detector agents by using contextual features $\mathbf{F}$. Specifically, we first use convolutional operations to down-sample $\mathbf{F}$, and obtain $\mathbf{F}^l \in \mathbb{R}^{d\times h_l\times w_l}$. Here, $h_l = h/2^{l-1}$ and $w_l = w/2^{l-1}$. Then, a 1×1 convolutional layer is applied on $\mathbf{F}^l$ to produce $N_l = N/2^{l-1}$ agent masks $\mathbf{M}^l \in \mathbb{R}^{N_l\times h_l\times w_l}$. Finally, $\mathbf{F}^l$ and $\mathbf{M}^l$ are flattened and the detector agents $\widehat{\mathbf{D}}^l$ are initialized as follows,

$$\widehat{\mathbf{D}}^l = \mathbf{F}^l \otimes [\mathbf{M}^l]^\top, \tag{4}$$

where $\otimes$ represents the matrix multiplication operator.

Detector decoder. As shown in the right of Figure 2, we aim to utilize detector agents $\widehat{\mathbf{D}}^l$ to aggregate information from the $(l-1)^{th}$ keypoint detectors $\mathbf{D}^{l-1}$. In this way, we can obtain the l^{th} keypoint detectors $\mathbf{D}^l$, which is formulated as follows,

$$\mathbf{Q} = \mathbf{W}^{\mathcal{Q}}\widehat{\mathbf{D}}^l, \mathbf{K} = \mathbf{W}^{\mathcal{K}}\mathbf{D}^{l-1}, \mathbf{V} = \mathbf{W}^{\mathcal{V}}\mathbf{D}^{l-1}, \tag{5}$$

$$\bar{\mathbf{D}}^l = \mathrm{LN}(\widehat{\mathbf{D}}^l + \mathbf{V}\cdot\mathrm{Softmax}(\mathbf{K}^\top\mathbf{Q})), \tag{6}$$

$$\mathbf{D}^l = \mathrm{LN}(\bar{\mathbf{D}}^l + \mathrm{MLP}(\bar{\mathbf{D}}^l)). \tag{7}$$

Here, LN is layer normalization and MLP denotes the multi-layer perception. For the first level ($l = 1$), there is no definition of $\mathbf{D}^0$. Thus, we simply replace $\mathbf{D}^0$ with the flattened $\mathbf{F}$, which means that keypoint detectors $\mathbf{D}^1$ at the first level are generated according to image features $\mathbf{F}$.

3.4. Objective Function

After obtaining the contextual descriptors $\mathbf{F} \in \mathbb{R}^{d \times h \times w}$ and hierarchical detectors $\mathbf{D} \in \mathbb{R}^{d \times N_a}$, multiple score maps $\mathbf{S}_N \in \mathbb{R}^{N_a \times h \times w}$ is generated by a dot product operation between them, *i.e.* $\mathbf{S}_N = \mathbf{D}^\top \mathbf{F}$. Here, we denote $N_a = N_1 + N_2 + N_3$, which is the number of hierarchical detectors. The final keypoint detection score map $\mathbf{S}_c \in \mathbb{R}^{1 \times h \times w}$ is obtained by averaging $\mathbf{S}_N$ on the first channel. Then three major objective functions are introduced to guide our model learning. For keypoint repeatability, the cosine similarity loss $\mathcal{L}_{cosim}$ is used to enforce detection score maps between two images to have high similarity in corresponding local patches. For the goal of enforcing the proposed detectors to focus on the salient position, we use the peaky loss $\mathcal{L}_{peaky}$ to maximize the local peakiness of the detection score map $\mathbf{S}_c$. Both $\mathcal{L}_{cosim}$ and $\mathcal{L}_{peaky}$ are inspired from R2D2 [31], and the details of these two losses can be referred to R2D2. Additionally, to expand the discrepancy among updated descriptor agents $\mathbf{A}$, we impose the diversity loss as follows,

$$\mathcal{L}_{div} = \frac{1}{M(M-1)} \sum_{j=1}^{M} \sum_{k=1, k \neq j}^{M} \frac{\langle \mathbf{A}_j, \mathbf{A}_k \rangle}{\|\mathbf{A}_j\|_2 \, \|\mathbf{A}_k\|_2}. \quad (8)$$

Finally, we combine these loss functions by a weighted sum to train our model, *i.e.*,

$$\mathcal{L}_{total} = \mathcal{L}_{cosim} + \alpha_1 \mathcal{L}_{peaky} + \alpha_2 \mathcal{L}_{div}, \quad (9)$$

where α_1 and α_2 are weight terms to balance these losses.

4. Experiments

In this section, we first introduce implementation details. Then, we show experimental results and some visualizations on four public benchmarks. Finally, we conduct a series of ablation studies to verify the effectiveness of each component. Please refer to the **Supplementary Material** for some discussions and more visulization results.

4.1. Implementation Details

In this work, we implement the proposed model in Pytorch [27]. We adopt the same backbone as [31] to extract original image features. In the contextual feature descriptor learning module, the number of descriptor agents M is set to 32. The dimension of image features $d = 128$. And d_k (the dimension of $\mathbf{Q}$ and $\mathbf{K}$) in Eq. (1) is equal to d. In the hierarchical keypoint detector learning module, the number of detector agents N for the first level is set to 16. The detector decoder is composed of $L_d = 4$ layers, and cross-attention heads are set to 8. The weight terms α_1 and α_2 in the objective function are set to 0.6 and 0.8. After obtaining the keypoint detection score map $\mathbf{S}_c$, keypoints can be obtained by applying the local maxima filtering and the threshold constraint [31] on the score map $\mathbf{S}_c$. The model runs about $0.32s$ for a 1600×1200 image pair on an RTX 3090 GPU. For training, we adopt the same outdoor training dataset [30, 35, 36] as R2D2, and the indoor training dataset [8]. All parameters in the feature extractor backbone, the contextual feature descriptor learning module and the hierarchical keypoint detector learning module are randomly initialized, and trained from scratch. We train our model using the Adam optimizer. The learning rate is set to 10^{-4}, and the weight decay is 3×10^{-4}. It converges after 24 hours of training on a single RTX 3090 GPU.

4.2. Datasets and Evaluation Metrics

HPatches. The HPatches [2] dataset is a widely adopted matching benchmark containing 116 image sequences under significant illumination and viewpoint changes. Here, each sequence includes a reference image and five query images, and the ground-truth homography is provided for each image pair. We follow the evaluation procedure of [12, 31, 42] to exclude 8 high-resolution sequences, leaving 108 image sequences, where 52 sequences are under strong illumination changes and 56 sequences are under extreme viewpoint variations. As for the *evaluation metric*, we use the same definition as in [42], and report the area under the cumulative curve (AUC) of the corner error.

ScanNet. The ScanNet [8] is a large-scale indoor dataset with ground truth poses and depth maps, which is used to target the task of indoor pose estimation. This dataset is challenging since it contains image pairs with wide baselines and extensive texture-less regions. We follow the same procedure as [34, 42] and use 1500 image pairs from [34] to evaluate our method. And the *evaluation metric* follows previous work [42], where the AUC of the indoor pose error at thresholds $(5°, 10°, 20°)$ is reported.

YFCC100M. The YFCC100M dataset [44] is usually used to validate the performance of outdoor pose estimation, including a diverse collection of complex real-world scenes ranging from 200,000 street-life-blogged photos to snapshots of daily life, holidays, and events. The main challenging factors for YFCC100M are extreme scale and illumination variations. We adopt the same test pairs as [34, 53] to evaluate, i.e. on 4 scenes of this dataset, where each scene is composed of 1000 image pairs. As for the *evaluation metric*, we report the AUC of the pose error at thresholds $(5°, 10°, 20°)$, similar to [34, 52, 53]. Here, the pose error is defined as the maximum of angular error in rotation and translation, which is computed between the ground truth pose and the predicted pose vectors.

MegaDepth. The MegaDepth [17] is composed of 1M internet images of 196 scenes. In addition, the sparse 3D point clouds of these images constructed by COLMAP [38] and depth maps are also provided. The main challenge of

Table 1. Evaluation results for homography estimation on the HPatches. We report the AUC of the corner error in percentage.

Methods	AUC@3px	AUC@5px	AUC@10px	#matches
Sparse-NCNet [32]	48.9	54.2	67.1	1.0K
DRC-Net [16]	50.6	56.2	68.3	1.0K
LoFTR [42]	65 .9	75.6	84.6	1.0K
D2-Net [12] + NN	23.2	35.9	53.6	0.2K
R2D2 [31] + NN	50.6	63.9	76.8	0.5K
DISK [47] + NN	52.3	64.9	78.9	1.1K
SuperPoint [9] + SuperGlue [34]	53.9	68.3	81.7	0.6K
D^{2}Former + NN (ours)	**71.6**	**81.3**	**89.7**	0.6K

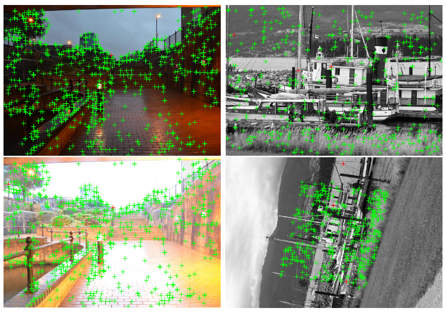

Figure 3. Qualitative results on HPatches. The images in each column form a pair for image matching. Green and red dots denote correct and incorrect matches respectively. (The threshold is 3px).

MegaDepth is strong viewpoint changes and repetitive patterns. We take the same 1500 image pairs as [42] to evaluate the proposed model. Here, the *evaluation metric* we adopt is the same as [42], where the AUC of the pose error at thresholds $(5^\circ, 10^\circ, 20^\circ)$ is reported.

4.3. Comparison with State-of-the-art Methods

Results on HPatches dataset. We compare our model with previous state-of-the-art image matching methods [9, 12, 16, 31, 32, 34, 42, 47]. As shown in Table 1, our method achieves 71.6% in AUC@3px, 81.3 % in AUC@5px and 89.7% in AUC@10px, outperforming all other methods significantly. Compared with LoFTR [42], our D^{2}Former improves by 5.7% in AUC@3px, 5.7% in AUC@5px and 5.1% in AUC@10px, respectively. Finally, we show some qualitative results in Figure 3. Our model can achieve robust keypoint detection and establish accurate matches when facing challenges like extreme illumination (the first column) and viewpoint changes (the last column), which fully proves the effectiveness of our proposed contextual feature descriptor learning module and hierarchical keypoint detector learning module.

Results on ScanNet dataset. Here, we present the performance comparison of the indoor pose estimation between our method and other state-of-the-art methods. As shown in Table 2, our proposed method outperforms other state-of-the-art methods favorably at all 3 thresholds. Specifically, compared with ASpanFormer [42], our method improves by 5.43% in AUC@5°, 5.69% in AUC@10° and 5.87% in AUC@20°, which demonstrate that our model is able to

Table 2. Evaluation results on the ScanNet dataset. We report the AUC of the pose error at thresholds $(5^\circ, 10^\circ, 20^\circ)$.

Methods	AUC@5°	AUC@10°	AUC@20°
DRC-Net [16]	7.69	17.93	30.49
LoFTR [42]	22.06	40.80	57.62
ASpanFormer [7]	25.60	46.00	63.30
D2-Net [12] + NN	5.25	14.53	27.96
R2D2 [31] + NN	7.43	17.45	28.64
SuperPoint [9] + NN	9.43	21.53	36.40
SuperPoint [9] + PointCN [52]	11.40	25.47	41.41
SuperPoint [9] + OANet [53]	11.76	26.90	43.85
SuperPoint [9] + SuperGlue [34]	16.16	33.81	51.84
D^{2}Former + NN (ours)	**31.03**	**51.69**	**69.17**

Figure 4. Qualitative results on the ScanNet dataset. The images in each column form a pair for image matching. Green and red dots denote correct and incorrect matches respectively. (The epipolar error threshold is 5×10^{-4}).

Table 3. Evaluation results on the YFCC100M dataset. We report the AUC of the pose error at thresholds $(5^\circ, 10^\circ, 20^\circ)$.

Methods	AUC@5°	AUC@10°	AUC@20°
LoFTR [42]	40.28	61.17	77.80
SIFT [18] + SuperGlue [34]	30.49	51.29	69.72
R2D2 [31] + NN	33.85	52.44	68.53
SuperPoint [9] + NN	16.94	30.39	45.72
SuperPoint [9] + OANet [53]	26.82	45.04	62.17
SuperPoint [9] + SuperGlue [34]	38.72	59.13	75.81
D^{2}Former + NN (ours)	**56.78**	**73.71**	**85.37**

establish accurate correspondences for indoor pose estimation. Finally, we show some qualitative results in Figure 4. It can be seen that our proposed method can realize image matching robust to the existence of texture-less regions in the ScanNet. The reason may be that our designed contextual feature descriptor learning module can generate discriminative descriptors with a large receptive field. Moreover, the designed hierarchical keypoint detector learning module can perceive high-level structures, which is helpful for detecting repeatable keypoints in texture-less regions and achieving robust matching.

Results on YFCC100M dataset. As shown in Table 3, we attempt to compare our method with previous state-of-the-art approaches to validate the effectiveness of our D^{2}Former for outdoor pose estimation. The results show that our proposed method can surpass the other image matching methods by a large margin, gaining by 16.50% in AUC@5°, 12.54% in AUC@10° and 7.57% in AUC@20° compared to LoFTR [42]. Furthermore, as shown in Figure 5, our pro-

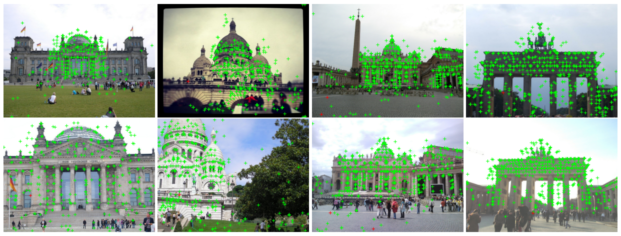

Figure 5. Qualitative results on the YFCC100M (the first two columns) and MegaDepth datasets (the last two columns). The images in each column form a pair for image matching. Green and red dots denote correct and incorrect matches respectively. (The epipolar error threshold is 5×10^{-4}).

Table 4. Evaluation results on the MegaDepth dataset. We report the AUC of the pose error at thresholds $(5^\circ, 10^\circ, 20^\circ)$.

Methods	AUC@5°	AUC@10°	AUC@20°
DRC-Net [16]	27.01	42.96	58.31
LoFTR [42]	52.80	69.19	81.18
ASpanFormer [7]	55.30	71.50	83.10
R2D2 [31] + NN	37.14	55.09	69.65
SuperPoint [9] + SuperGlue [34]	42.18	61.16	75.96
D^{2}Former + NN (ours)	**66.27**	**78.44**	**86.81**

posed D^{2}Former can establish reliable matches when facing extreme scale and viewpoint variations, demonstrating the effectiveness of our designed two modules.

Results on MegaDepth dataset. Here, we attempt to compare our proposed method with other state-of-the-art image matching methods on the Megadepth dataset. As shown in Table 4, our proposed method obtains the best performance in pose accuracy among all image matching methods. As for the comparison with ASpanFormer [7] which performs the best on this dataset, our model improves by 10.97% in AUC@5°, 6.94% in AUC@10° and 3.71% in AUC@20°. The visualization results are also shown in Figure 5, and our model can establish accurate correspondences when facing extreme viewpoint changes and repetitive patterns.

4.4. Ablation Studies

To analyze the effects of each component in D^{2}Former, we perform detailed ablation studies on the ScanNet dataset. In Table 5, the model [A] is the same as R2D2. We first extract original features $\widehat{\mathbf{F}}$ using the same backbone as R2D2. Then, for model [B], $\widehat{\mathbf{F}}$ are processed by CFDL to obtain contextual descriptors, while the keypoint detector is implemented with a 1×1 convolutional kernel. For model [C], $\widehat{\mathbf{F}}$ are not processed by CFDL, and hierarchical detectors are learnt by sending $\widehat{\mathbf{F}}$ into the HKDL. The model [D] is the full model of our D^{2}Former.

Effects of the contextual feature descriptor learning (CFDL) module. As shown in Table 5, with the proposed CFDL module, the performance on the ScanNet is improved notably. In specific, the performance of model [B] is improved by 11.25% in AUC@5°, 19.04% in AUC@10° and 26.53% in AUC@20°, compared to the model [A]. And the model [D] also performs better than the model [C]. The main reason is that our CFDL module can model long-range dependencies for feature descriptors, which is beneficial to handling texture-less regions for robust image matching.

Table 5. Effectiveness of each component on the ScanNet. We report the AUC of the pose error at thresholds $(5^\circ, 10^\circ, 20^\circ)$.

Models	HKDL	CFDL	AUC@5°	AUC@10°	AUC@20°
[A]	✗	✗	7.43	17.45	28.64
[B]	✗	✓	18.68	36.49	55.17
[C]	✓	✗	27.64	48.34	67.05
[D]	✓	✓	**31.03**	**51.69**	**69.17**

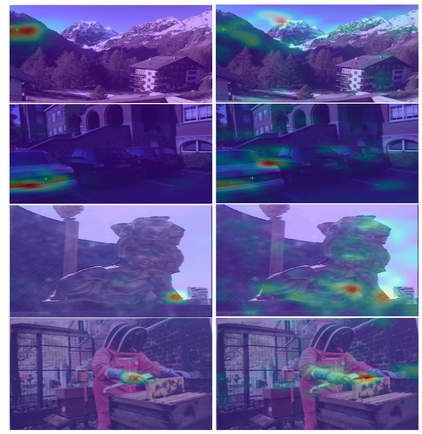

Figure 6. Qualitative comparisons between our proposed agent-based attention mechanism (the first column) and the standard full attention (the second column).

Furthermore, we show some qualitative comparisons between the agent-based attention mechanism in the CFDL module and the standard full attention [1], which can validate the effects of the designed agent-based attention mechanism. As shown in Figure 6, we can find that standard full attention introduces extra noise when conducting global interactions. For example, in the first row, when selecting a pixel from the mountain to conduct interaction with other pixels, pixels from backgrounds also have a high attention score for the standard full attention. By contrast, our proposed agent-based attention mechanism has a clear attention score map as shown in Figure 6, which can effectively reduce noise and generate robust contextual descriptors.

Impacts about the number of descriptor agents in the CFDL module. Here, we study the performance with different numbers of descriptor agents (M) in the CFDL module. M is picked from the set $\{2, 4, 8, 16, 32, 64\}$, and we evaluate the performance on the ScanNet. As shown in Table 6, we find that the overall performance of the model improves with the increase of M, and the model can get the best performance when $M = 32$. There is no performance gain when M continues to increase. The reason may be that the setting $M = 32$ is able to adequately capture different contexts in the input images, and more descriptor agents

Table 6. Impacts of the number of descriptor agents on ScanNet.

Models	AUC@5°	AUC@10°	AUC@20°
M=2	27.56	48.08	66.81
M=4	28.12	48.91	67.39
M=8	28.98	49.33	67.92
M=16	29.83	50.49	68.67
M=32	**31.03**	**51.69**	**69.17**
M=64	30.87	51.31	69.07

Table 7. Impacts of the number of detector agents on ScanNet.

Models	AUC@5°	AUC@10°	AUC@20°
N=4	25.59	45.49	64.66
N=8	29.19	50.11	68.29
N=12	30.24	50.75	68.74
N=16	**31.03**	**51.69**	**69.17**
N=20	30.75	50.99	68.65

may have an adverse influence on the model training due to lack of sufficient explicit constraints.

Effects of the hierarchical keypoint detector learning (HKDL) module. As shown in Table 5, when adding our proposed HKDL module, the performance on the ScanNet can achieve great improvement. Specifically, the performance of model [C] is gained by 20.21% in AUC@5°, 30.89% in AUC@10° and 38.41% in AUC@20°, compared to the model [A]. Besides, the model [D] also performs much better than the model [B]. The main reason is that our proposed HKDL module can generate keypoint detectors in a hierarchically aggregated manner, so that keypoints with diverse levels of structures can be detected, which is vital for robust image matching.

Impacts about the number of detector agents in the HKDL module. To investigate the influences of the number of detector agents (N) in the HKDL module, we pick N from the set $\{4, 8, 12, 16, 20\}$ and evaluate the performance on the ScanNet dataset. As shown in Table 7, we find that when N is set to 16, the model can get the best performance. When N continues to increase from 16, the performance is no longer improved, which reflects that the model with $N = 16$ is sufficient to perceive different levels of structures on the ScanNet dataset.

Visualization about keypoint detection results for different levels. To further validate the effects of our proposed HKDL module, as shown in Figure 7, we show keypoint detection results for different levels. And the detection results from the first row to the third row are obtained by leveraging detectors $\mathbf{D}^1$, $\mathbf{D}^2$, and $\mathbf{D}^3$, respectively. For the first row, keypoints with low-level structures like standard corners or edges are commonly extracted, such as the edge of a refrigerator door in the first column, and the corner of wall in the second column (marked with a red circle). In the second and the third row, we find that detected keypoints are usually far away from simple structures like corners and edges. And keypoints can be detected in texture-less regions. We think

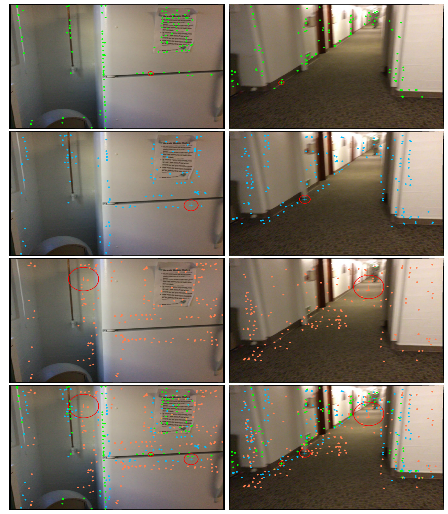

Figure 7. Keypoint detection results for an image (each column). From the first row to the third row, we sequentially show the detection results from low-level detectors to high-level detectors. And the fourth row shows the combined detection results.

the reason is that high-level detectors usually have larger perceived radii, and can perceive some high-level semantics of detected keypoints. For example, given a detected keypoint (the center of red circle), our generated high-level detectors may understand this keypoint is from a refrigerator and can sense how far this keypoint is from the edge of a refrigerator door in the first column. In conclusion, with our well-designed HKDL module, the generated hierarchical detectors can capture keypoints with diverse levels of structures, making our model robust to various challenges such as viewpoint transformations and poor textures, which greatly improves the performance of our method.

5. Conclusion

In this work, we propose a novel image matching model by Jointly Learning Hierarchical Detectors and Contextual Descriptors via Agent-based Transformers (D^{2}Former), including a CFDL module and a HKDL module. With these two well-designed modules, our proposed method can learn more discriminative feature description and realize repeatable keypoint detection under some extremely challenging factors, which is vital for robust image matching. Extensive experimental results on four challenging benchmarks demonstrate the superiority of our proposed method.

6. Acknowledgement

This work was partially supported by the National Nature Science Foundation of China (Grant 62022078, 12150007), and National Defense Basic Scientific Research Program of China (Grant JCKY2021601B013).

References

[1] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*, 2015. 3, 7

[2] 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. 5

[3] 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, 2

[4] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up robust features. In *Proceedings of the European Conference on Computer Vision*, pages 404–417, 2006. 1

[5] Assia Benbihi, Matthieu Geist, and Cedric Pradalier. Elf: Embedded localisation of features in pre-trained cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7940–7949, 2019. 1

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*, volume 12346, pages 213–229, 2020. 3

[7] 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. *Proceedings of the European Conference on Computer Vision*, 2022. 1, 3, 6, 7

[8] 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. 5

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

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, 2019. 3

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*, 2021. 2, 3, 4

[12] 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, 5, 6

[13] 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

[14] Christopher G Harris, Mike Stephens, et al. A combined corner and edge detector. In *Alvey vision conference*, volume 15, pages 10–5244, 1988. 2

[15] Georg Klein and David William Murray. Parallel tracking and mapping for small AR workspaces. In *IEEE and ACM International Symposium on Mixed and Augmented Reality*, pages 225–234, 2007. 1

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

[17] 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. 5

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

[19] Zixin Luo, Tianwei Shen, Lei Zhou, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. ContextDesc: Local descriptor augmentation with cross-modality context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2527–2536, 2019. 1, 2

[20] Zixin Luo, Lei Zhou, Xuyang Bai, Hongkai Chen, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. ASLFeat: Learning local features of accurate shape and localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6589–6598, 2020. 1, 2, 3

[21] Jiri Matas, Ondrej Chum, Martin Urban, and Tomás Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, 2004. 2

[22] Meng Meng, Tianzhu Zhang, Zhe Zhang, Yongdong Zhang, and Feng Wu. Task-aware weakly supervised object localization with transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[23] Anastasya Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. In *Advances in Neural Information Processing Systems*, 2017. 1, 2

[24] Marius Muja and David G Lowe. Scalable nearest neighbor algorithms for high dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2227–2240, 2014. 2

[25] Raul Mur-Artal, J. M. M. Montiel, and Juan D. Tardós. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015. 1

[26] Yuki Ono, Eduard Trulls, Pascal Fua, and Kwang Moo Yi. LF-Net: Learning local features from images. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[27] 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. 5

[28] Rémi Pautrat, Viktor Larsson, Martin R Oswald, and Marc Pollefeys. Online invariance selection for local feature descriptors. In *Proceedings of the European Conference on Computer Vision*, pages 707–724. Springer, 2020. 1, 2

[29] 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

[30] Filip Radenovic, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Revisiting Oxford and Paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5706–5715, 2018. 5

[31] Jerome Revaud, Philippe Weinzaepfel, César Roberto de Souza, and Martin Humenberger. R2D2: Repeatable and reliable detector and descriptor. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2, 3, 5, 6, 7

[32] 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. 6

[33] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. *Advances in Neural Information Processing Systems*, 31, 2018. 1

[34] 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, 5, 6, 7

[35] 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, 5

[36] Torsten Sattler, Tobias Weyand, Bastian Leibe, and Leif Kobbelt. Image retrieval for image-based localization revisited. In *Proceedings of the British Machine Vision Conference*, pages 1–12, 2012. 5

[37] Nikolay Savinov, Akihito Seki, Lubor Ladicky, Torsten Sattler, and Marc Pollefeys. Quad-networks: unsupervised learning to rank for interest point detection. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1822–1830, 2017. 1, 2

[38] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016. 5

[39] Weizhao Shao, Srinivasan Vijayarangan, Cong Li, and George Kantor. Stereo visual inertial lidar simultaneous localization and mapping. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, pages 370–377, 2019. 1

[40] Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, and Francesc Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 118–126, 2015. 1, 2

[41] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Learning local feature descriptors using convex optimisation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1573–1585, 2014. 1, 2

[42] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8922–8931, 2021. 1, 3, 5, 6, 7

[43] 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

[44] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 5

[45] Yurun Tian, Xin Yu, Bin Fan, Fuchao Wu, Huub Heijnen, and Vassileios Balntas. SOSNet: Second order similarity regularization for local descriptor learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11016–11025, 2019. 2

[46] Prune Truong, Stefanos Apostolopoulos, Agata Mosinska, Samuel Stucky, Carlos Ciller, and Sandro De Zanet. Glampoints: Greedily learned accurate match points. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10732–10741, 2019. 2

[47] Michal J. Tyszkiewicz, Pascal Fua, and Eduard Trulls. DISK: learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33:14254–14265, 2020. 6

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 2, 3, 4

[49] Deepak Geetha Viswanathan. Features from accelerated segment test (fast). In *Proceedings of the Image Analysis for Multimedia Interactive Services*, pages 6–8, 2009. 2

[50] Philippe Weinzaepfel, Thomas Lucas, Diane Larlus, and Yannis Kalantidis. Learning super-features for image retrieval. In *International Conference on Learning Representations*, 2022. 3

[51] 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

[52] Kwang Moo Yi, Eduard Trulls, Yuki Ono, Vincent Lepetit, Mathieu Salzmann, and Pascal Fua. Learning to find good

correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2666–2674, 2018. 5, 6

[53] 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. 1, 5, 6

[54] Linguang Zhang and Szymon Rusinkiewicz. Learning to detect features in texture images. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6325–6333, 2018. 1, 2

[55] Runze Zhang, Siyu Zhu, Tian Fang, and Long Quan. Distributed very large scale bundle adjustment by global camera consensus. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 29–38, 2017. 1