D2D: Keypoint Extraction with Describe to Detect Approach

Yurun Tian[1] Vassileios Balntas[2] Tony Ng[1] Axel Barroso-Laguna[1]
Yiannis Demiris[1] Krystian Mikolajczyk[1]
[1] Imperial College London
[2] Scape Technologies
{y.tian, tony.ng14, axel.barroso17, y.demiris, k.mikolajczyk}@imperial.ac.uk
vassileios@scape.io

Abstract

In this paper, we present a novel approach that exploits the information within the descriptor space to propose keypoint locations. Detect then describe, or detect and describe jointly are two typical strategies for extracting local descriptors. In contrast, we propose an approach that inverts this process by first describing and then detecting the keypoint locations. Describe-to-Detect (D2D) leverages successful descriptor models without the need for any additional training. Our method selects keypoints as salient locations with high information content which is defined by the descriptors rather than some independent operators. We perform experiments on multiple benchmarks including image matching, camera localisation, and 3D reconstruction. The results indicate that our method improves the matching performance of various descriptors and that it generalises across methods and tasks.

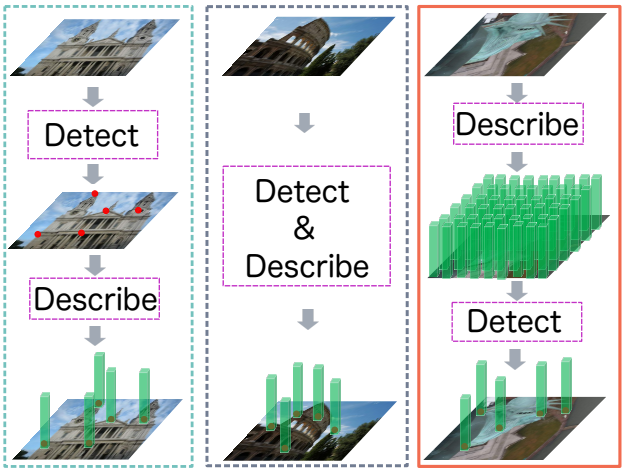

Figure 1: Comparison of our proposed Describe-to-Detect framework (right) to the existing Detect-then-Describe and Detect-and-Describe frameworks.

1. Introduction

One of the main problems in computer vision is concerned with the extraction of 'meaningful' descriptions from images and sequences. These descriptions are then used for the correspondence problem which is critical for applications such as SLAM [9, 32], structure from motion [43, 45], retrieval [35], camera localisation [39], tracking [49], etc. The key issue is how to measure the 'meaningfulness' from the data and which descriptors are the best. Extensive survey of salient region detectors [52] attempts to identify the main properties expected from 'good' features which include repeatability, informativeness, locality, quantity, accuracy, and efficiency. It has also been noted that the detector should be adapted to the needs of the application, *i.e.*, the data.

In contrast to the significant progress on local descriptors achieved with neural networks, keypoint detectors enjoyed little success from using learning methods, with few notable exceptions [10, 11, 17]. As a consequence, keypoint detectors based on handcrafted filters such as Difference-of-Gaussians, Harris, Hessian [52], which all originate from research in 1980-ties are still used in many applications.

In the era of deep learning, there are three main research directions towards improving image matching, namely non-detector-specific description [50, 28, 14, 51], non-descriptor-specific detection [53, 17], as well as jointly learnt detection-description [54, 34, 46, 11, 36]. What underlines the concept of disjoint frameworks is their sub-optimal compatibility between the detection and description. In contrast to the CNN based descriptors [50, 28, 24, 14, 23, 51], the performance of jointly learnt detection-description [54, 34, 11, 36] does not seem to generalise well across datasets and tasks. CNN descriptors perform significantly better if trained and applied in the same data domain. Similarly, different keypoint detectors are suitable for dif-

ferent tasks. In addition, fine-tuning a descriptor for a specific keypoint detector further improves the performance. With all available options finding optimal pair of detector-descriptor for a dataset or a task requires extensive experiments. Therefore, an approach that adapts keypoint detector to a descriptor without training and evaluation is highly sought for various applications.

Our approach is inspired by detectors based on various saliency measures [15, 42] where the saliency was defined in terms of local signal complexity or unpredictability; more specifically the Shannon entropy of local descriptor was suggested. Despite the appealing idea, such methods failed to be widely adopted due to the complexity of the required dense local measurements. However, currently available CNN dense descriptors allow revisiting the idea of using saliency measured by descriptors to define keypoint locations. Top performing learnt descriptors [50, 28, 14, 51] all share the same fully convolutional network (FCN) that adapts to varying image resolution and output dense descriptors. Furthermore, joint methods like SuperPoint [10], D2-Net [11] and R2D2 [36] also provide dense features. The proposed approach can be seen as a combination of the classical saliency-based methods [15, 42] and the modern deep attention mechanisms [33, 35, 11].

In summary, our main contributions are:

- We propose a novel Describe-to-Detect (D2D) framework for keypoint detection that requires no training and adapts to any existing CNN based descriptor.
- We propose a relative and an absolute saliency measure of local deep feature maps along the spatial and depth dimensions to define keypoints.
- We demonstrate on several benchmarks and different tasks that matching performance of various descriptors can be consistently improved by our approach.

2. Related Works

In this section, we briefly introduce some of the most recent learning-based methods for local feature detection and description. There are several survey articles that provide comprehensive reviews of this field [26, 27, 52, 19].

Local Feature Detection. Most of the existing hand-crafted [22, 20, 1, 2] or learned [53, 18, 56, 41, 55, 29, 17] detectors are not descriptor-specific. The main property required from keypoints is their repeatability such that their descriptors can be correctly matched. TILDE [53] trains a piece-wise linear regression model as the detector that is robust to weather and illumination changes. CNN models are trained with feature covariant constraints in [18, 56]. Unsupervised trained QuadNet [41] assumes that the ranking of the keypoint scores should be invariant to image transformations. A similar idea is also explored in [55] to detect keypoint in textured images. AffNet [29] learns to predict the affine parameters of a local feature via the hard negative-constant loss based on the descriptors. Key.Net [17] combines hand-crafted filters with learned ones to extract keypoints at different scale levels. Recently, it has been shown that pre-trained CNNs on standard tasks such as classification can be adapted to keypoint detection [6]. However, the local feature matching pipeline is by nature different from classification. In contrast, our method directly leverage CNNs pre-trained for description to achieve detection.

Local Feature Description. The emergence of several large scale local patch datasets [7, 3, 30] stimulated the development of deep local descriptors [47, 4, 50, 28, 16, 14, 51] that are independent of the detectors. However, this paper is concerned with keypoint detection. Therefore we refer the reader to [3] for a detailed review and evaluation of recent descriptors. In our experiments we include several recent descriptors such as HardNet [28] and SOSNet [51]. SIFT [22] is the most widely used handcrafted descriptor still considered as a well-performing baseline. HardNet [28] combines triplet loss with a within-batch hard negative mining that has proven to be remarkably effective and SOSNet [51] extends HardNet with and second-order loss.

Joint Detection and Description. Joint training of detection-description has received more attention recently [54, 34, 10, 11, 23, 36, 25, 12]. SuperPoint [10], D2-Net [11], and R2D2 [36] are the three representatives of recent research direction, where patch cropping is replaced by fully convolutional dense descriptors. SuperPoint [10] leverages two separate decoders for detection and description on a shared encoder. Synthetic shapes and image pairs generated from random homographies are used to train the two parts. In D2-Net [11], local-maxima within and across channels of deep feature maps are defined as keypoints, with the same maps used for descriptors. R2D2 [36] aims at learning keypoints that are not only repeatable but also reliable together with robust descriptors. However, the computational cost for current joint frameworks is still high. Besides, the generation of training data is typically laborious and method-specific.

Therefore, a keypoint detection method that is based on a trained descriptor model, thus adapted to the data without requiring any training, can be considered a novel and significant contribution.

3. Describe-to-Detect

In this section, we first define keypoints in terms of the descriptor saliency, then we present our approach to integrate D2D with existing state-of-the-art methods.

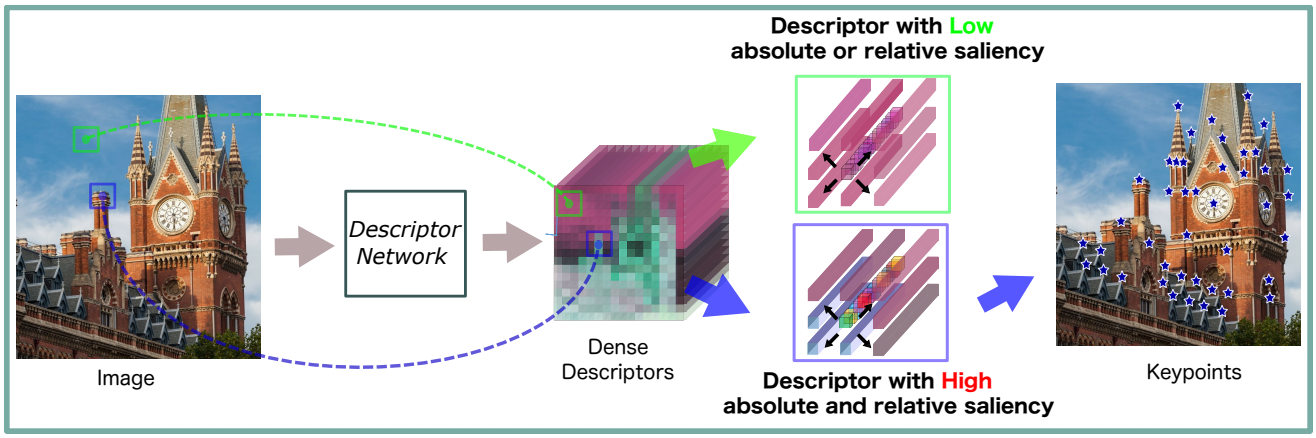

Figure 2: The Describe-to-Detect pipeline. Locations with high variation across channels (high absolute saliency) as well as high saliency w.r.t spatial neighbours (relative saliency) are detected as keypoints.

3.1. What is a keypoint?

Despite the absence of a unified definition, it is widely accepted that keypoints should be image points that have the potential of being repeatably detected under different imaging conditions. As mentioned, according to [52], such points should satisfy several requirements such as repeatability, informativeness, locality, quantity, accuracy and efficiency.

In this work, we argue that the informativeness, which we refer to as saliency, is the property that can lead to satisfying most of the other requirements. We define the saliency in relative terms *i.e.* w.r.t the other descriptors in the neighbourhood, as well as in absolute terms as the information content of the descriptor. Therefore, we depart from the following assumptions:

Assumption 1 *A point in an image has a high absolute saliency if its corresponding descriptor is highly informative.*

The idea of exploiting salient regions in an image has been adopted by many classical [42, 15] methods as well as recent attention-based models [33, 35, 11]. In tasks such as image retrieval, saliency/attention is defined on image regions with rich semantic information [35, 33]. In feature matching, local image structures that exhibit significant variations in shape and texture can be considered salient. However, absolute saliency alone is not sufficient for identifying keypoints. For instance, highly informative but spatially non-discriminative structures should be avoided as they cannot be uniquely and accurately localised. Therefore a relative saliency should also be considered.

Assumption 2 *A point in an image has a high relative saliency if its corresponding descriptor is highly discriminative in its spatial neighbourhood.*

The success of handcrafted detectors that define keypoints according to this criteria [31, 22, 20, 1, 2, 17] validates this assumption. Descriptors on repeated textures can lead to geometrically noisy correspondences, therefore their spatial uniqueness is essential. Similarly to the absolute saliency, the relative saliency alone is not sufficient for detection. For example, corner points of uniform regions can exhibit high relative saliency, whereas their descriptors information content is not high.

Based on Assumptions 1 and 2, our definition for keypoints based on their corresponding descriptors is:

Definition 1 *A point in an image is a keypoint, if its corresponding descriptor's absolute and relative saliencies are both high.*

Definition 1 is a generalization of the keypoints defined for low-level pixel intensities, either by simple operators such as autocorrelation [31] or by early saliency based methods [42, 15], to high-level descriptors. In contrast to existing Detect-then/and-Describe frameworks, in Definition 1, we define the detector by the properties of the descriptor. Thus, the key idea of Describe-to-Detect (D2D) is a description-guided detection. Moreover, we claim that descriptors that are specifically trained to be robust to the changes of imaging conditions can provide data driven discriminativeness and thus, more reliable detections. It is worth noting that our D2D differs from other works that utilize the deep feature map response, but do not exploit the full representation potential of a descriptor. For example, the detection step of D2-Net [11] is performed by considering each feature activation separately, as a score map for keypoint detection, whereas D2D detects keypoints via descriptor similarity in the metric space and therefore makes use of the rich information content across entire depth.

In summary, to identify the keypoints, Definition 1 is

concerned with two properties: Firstly, when evaluating itself, the descriptor should be informative. Secondly, when comparing to others, the descriptor should be discriminative.

3.2. How to detect a keypoint?

Measuring the absolute saliency of a point can be achieved by computing the entropy of a descriptor. It has been shown in the design of binary descriptors [37, 5], that selecting binary tests with high entropy will encourage compact and robust representation. Therefore, we propose to measure the informativeness of a descriptor by its entropy, interpreted as a N-dimensional random variable. Unlike in binary descriptors where discrete entropy can be computed directly, for real-valued descriptors differential entropy is needed. However, computing an accurate differential entropy requires probability density estimation, which is computationally expensive. Thus, similarly to the binary case [37, 5], we employ the standard deviation as a proxy for the entropy:

$$\mathbf{S}_{\text{AS}}(x,y) = \sqrt{\mathbb{E}[\boldsymbol{F}^2(x,y)] - \bar{F}(x,y)^2} \tag{1}$$

where $\bar{F}(x,y)$ is the mean value of descriptor $\boldsymbol{F}(x,y)$ across its dimensions.

Measuring the relative saliency of a point is based on Assumption 2. A function that measures the relationship between a variable's current value and its neighboring values is the autocorrelation. It has been successfully used by the classic Moravec corner detector [31] as well as the well known Harris detector [13]. However, their simple operators rely directly on pixel intensities which suffer from poor robustness to varying imaging conditions. The autocorrelation was implemented as a sum of squared differences (SSD) between the corresponding pixels of two overlapping patches:

$$\boldsymbol{S}_{\text{SSD}}(x,y) = \sum_u \sum_v \boldsymbol{W}(u,v)(\boldsymbol{I}(x,y) - \boldsymbol{I}(x+u, y+v))^2 \tag{2}$$

where $I(x,y)$ indicate pixel intensity at (x,y), (u,v) are window indexes centered at (x,y), and $\boldsymbol{W}(u,v)$ are weights. A high value of $\boldsymbol{S}_{\text{SSD}}(x,y)$ means low similarity. As a natural generalization of SSD for measuring the relative saliency, we replace pixel intensities with dense descriptors :

$$\boldsymbol{S}_{\text{RS}}(x,y) = \sum_u \sum_v \boldsymbol{W}(u,v)||\boldsymbol{F}(x,y) - \boldsymbol{F}(x+u, y+v)||_2 \tag{3}$$

where $\boldsymbol{F}(x,y)$ indicates the descriptor centered at location (x,y), and $||\cdot||_2$ is the L2 distance. A high value of $\boldsymbol{S}_{\text{RS}}(x,y)$ defines points with high relative saliency, *i.e.*, this point stands out from its neighbours according to the description provided by the pre-trained descriptor model. Using Equations (1) and (3), we assign a score to each point by:

$$\mathbf{S}_{\text{D2D}}(x,y) = \mathbf{S}_{\text{AS}}(x,y)\mathbf{S}_{\text{RS}}(x,y). \tag{4}$$

3.3. Dense Descriptors

All existing description methods can extract dense descriptors for a given image. For example, patch-based methods can be used to generate dense descriptors by extracting patches with a sliding window. However, such strategy is infeasible in large scale tasks such as 3D reconstruction, due to its computational cost. Fortunately, most recent state-of-the-art methods adopt the fully convolutional network architecture without fully-connected layers [50, 28, 14, 51, 10, 11]. Dense descriptor maps can be extracted with a single forward pass for images with various resolutions. To guarantee the efficiency, we apply the proposed D2D to fully convolutional network descriptors only. Specifically, in Section 4, we evaluate D2D with two state-of-the-art descriptors, *i.e.*, HardNet [28] and SOSNet [51]. We further validate D2D on joint detection-description methods SuperPoint [10] and D2-Net [11].

3.4. Implementation Details

Computation of $\boldsymbol{S}_{\text{AS}}(x,y)$ is done on descriptors before L2 normalization, since it has an effect of reducing the standard deviation magnitude across the dimensions. It has been shown [3] that the descriptor norm, that also reflects the magnitude of variance, is not helpful in the matching stage, however, we use it during the detection to identify informative points.

Computation of $\boldsymbol{S}_{\text{RS}}(x,y)$. We define the size of the window $\boldsymbol{W}(u,v)$ in Equation (3) as r_{RS}. Considering that the receptive fields of neighbouring descriptors overlap and that the descriptor map resolution is typically lower than the input image, we sample the neighbouring descriptors with a step size of 2 and calculate the relative saliency with respect to the center descriptor. Note that the operation in Equation (3) can be implemented efficiently with a convolution, therefore when the window size r_{RS} is small and the sampling step is 2, the computational cost is negligible.

Combining D2D with descriptors. To evaluate D2D we employ two current state-of-the-art patch-based descriptors, namely HardNet [28] and SOSNet [51]. Given the network architecture [50] and an input image of size $H \times W (H \geq 32, W \geq 32)$, the output feature map size is $(\lfloor H/4 \rfloor - 7) \times (\lfloor W/4 \rfloor - 7)$. The receptive field is of size 51×51. Therefore, each descriptor $\boldsymbol{F}(x,y)$ describes a 51×51 region centered at $(4x + 14, 4y + 14)$. There are two stride-2 convolutional layers in the network, therefore $\boldsymbol{F}$ describes each 51×51 patch with stride of 4. In other words, keypoints are at least 4 pixels away from each

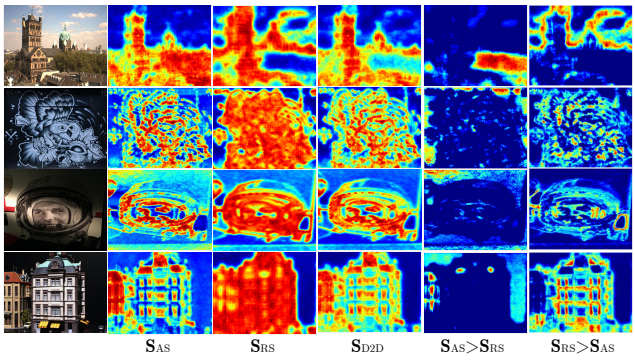

Figure 3: Visualization of the heat maps generated by D2D applied to HardNet [28]. From left to right the columns show images, heat maps of $\mathbf{S}_{\text{AS}}$, $\mathbf{S}_{\text{RS}}$, $\mathbf{S}_{\text{D2D}}$, $\max(0, \mathbf{S}_{\text{AS}} - \mathbf{S}_{\text{RS}})$ and $\max(0, \mathbf{S}_{\text{RS}} - \mathbf{S}_{\text{AS}})$, respectively. $\mathbf{S}_{\text{AS}}$ and $\mathbf{S}_{\text{RS}}$ are normalized so that their values are in $[0, 1]$.

other. Such sparse sampling has also been validated in other works [36, 8]. Finally, with $\mathbf{S}_{\text{D2D}}$ we directly take the top K ranked points as keypoints.

In D2-Net [11], the effect of downsampling layers is mitigated by upsampling the dense descriptors. However, with a large receptive overlap, dense $\boldsymbol{F}$ is redundant. For example, $\boldsymbol{F}(x, y)$ and $\boldsymbol{F}(x + 1, y)$ describe two 51×51 patch with a 47×51 overlap. For networks such as HardNet [28] and SOSNet [51] that are trained to be insensitive to such small changes, additional interpolation of feature maps is unnecessary.

Also, note that the amount of content the network can see in a 51×51 region is defined by the resolution of the image. High resolution and dense sampling can make the neighbouring descriptors indistinguishable. An interesting question is whether a multi-scale strategy to tackle the scale changes is needed. We show in Section 4 that single scale HardNet [28] and SOSNet [51] perform well in different tasks. We claim that there are two reasons for this: First, dramatic scale changes are rare in typical images of the same scenes. Second, scale changes are often global and the ranking of the detected keypoints is not affected by such changes[41].

Furthermore, we give some examples in Figure 3 to show different components of the final keypoint score map and how $\boldsymbol{S}_{\text{AS}}$ and $\boldsymbol{S}_{\text{RS}}$ contribute to $\boldsymbol{S}_{\text{D2D}}$. As shown, $\boldsymbol{S}_{\text{AS}}$ highlights all regions that have high intensity variations, while $\boldsymbol{S}_{\text{RS}}$ has high scores in structured areas. Finally, $\boldsymbol{S}_{\text{D2D}}$ combines the two parts, resulting in a low score for repeated/non-textured areas and edges. Points with $\boldsymbol{S}_{\text{RS}}$ greater than $\boldsymbol{S}_{\text{AS}}$ are informative but not locally discriminative. This includes repeated textures like tree leaves and tiles on building roof, as well as intensity noise in visually homogeneous regions. Otherwise, line structures are less informative but can be discriminative from the adjacent regions, which results in $\boldsymbol{S}_{\text{AS}}$ greater than $\boldsymbol{S}_{\text{RS}}$.

4. Experiments

In this section we present the results for various tasks on different datasets, which include image matching, visual localisation and 3D reconstruction.

4.1. Comparison with the state-of-the-art

We evaluate D2D on three different tasks, *i.e.*, image matching, visual localisation, and 3D reconstruction on three standard benchmarks, *i.e.*, Hpatches [3], Aachen Day-Night [40, 38], and ETH SfM [44], respectively. Each of the tasks tests the compatibility of the detector and the descriptor from a different perspective. We employ HardNet and SOSNet trained on Liberty from UBC dataset [7]. For all experiments in this section, we set r_{RS} to be 5.

We evaluate our method on three different tasks, *i.e.*, image matching, visual localisation, and 3D reconstruction on three standard benchmarks, *i.e.*, Hpatches [3], Aachen Day-Night [40, 38], and ETH SfM [44]. Each of the tasks tests the compatibility of the detector and the descriptor from a different perspective.

Image Matching. Hpatches [3] dataset contains 116 image sequences with ground truth homographies under different viewpoint or illumination changes. Following the evaluation protocol of [11, 26], we report the mean matching accuracy (MMA). In Figure 4, we report MMA for thresholds 1 to 10 pixels averaged over all image pairs. Also, we give the mean number of keypoints, mean number of mutual nearest neighbour matches per image pair, and the ratio between the two numbers.

As shown, combining D2D approach with HardNet and SOSNet can achieve superior or comparable results to other state-of-the-art methods. By comparing the curves of HardNet/SOSNet+D2D with SIFT+HardNet/SOSNet, we can observe that the D2D finds more compatible keypoints for HardNet/SOSNet than SIFT. Also note that when using the SIFT detector, the MMA curves of HardNet and SOSNet almost overlap, however, D2D helps to further reveal their performance difference. This also demonstrates that the detector is a very crucial component of matching, and that optimising descriptor independently from the detector is insufficient. Moreover, we can also see that D2D can detect more keypoints thus leading to a higher number of mutual nearest neighbour matches, which beneficial for various applications. Besides, HardNet+D2D also surpass AffNet+HardNet++, where AffNet is specifically trained with a descriptor loss. This shows that leveraging the absolute and relative saliency of descriptors is an effective approach to detect keypoints.

Day-Night Visual Localisation. In this section, we further evaluate our method on the task of long-term visual localization using the Aachen Day-Night dataset [38, 40].

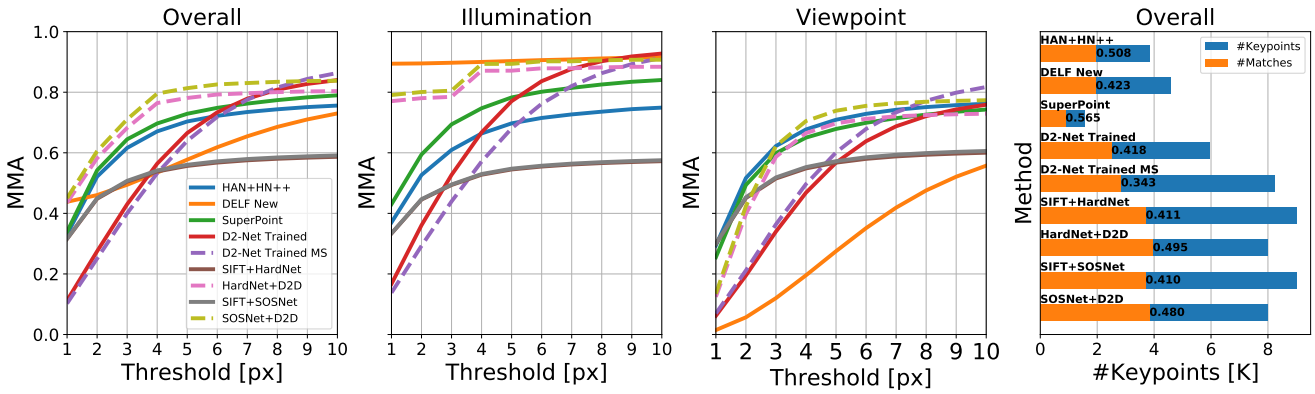

Figure 4: Experimental results for the HPatches [3] dataset. The results are reported with Mean Matching Accuracy. We observe that the proposed D2D method significantly outperforms other approaches, especially in the crucial high-accuracy area of $< 5px$

Method	#Dim	#Kp	0.5m, 2°	1m, 5°	5m, 10°
SIFT	128	11K	33.7	52.0	65.3
DELF(New) [33]	1024	11K	39.8	61.2	85.7
HAN+HN++ [29, 28]	128	11K	39.8	61.2	77.6
SuperPoint [10]	128	3.7K	42.8	57.1	75.5
D2-Net SS [11]	512	12K	44.9	66.3	88.8
D2-Net MS [11]	512	12K	44.9	64.3	88.8
R2D2 (N=8) [33]	128	10K	**45.9**	**66.3**	**88.8**
SIFT+HardNet [28]	128	11K	34.7	52.0	69.4
HardNet+D2D	128	16K	41.8	61.2	84.7
SIFT+SOSNet [51]	128	11K	36.7	53.1	70.4
SOSNet+D2D	128	16K	42.9	64.3	85.7

Table 1: Comparison to the state of the art on the Aachen Day-Night dataset. We report the percentages of successfully localized images within 3 error thresholds as in [11, 36].

This task evaluates the performance of local features under challenging conditions including day-night and viewpoint changes. Our evaluation is performed via a localisation pipeline[1] based on COLMAP [43] and The Visual Localization Benchmark[2].

In Table 1, we report the percentages of successfully localized images within three error thresholds. As can be seen, D2D significantly boost the performance of HardNet and SOSNet. Even though D2-Net and R2D2 are still the best performers on this dataset, their advantage may come from the training data or network architecture, *i.e.*, D2-Net uses VGG16 network [48] pre-trained on ImageNet and then trained on MegaDepth [21] while R2D2 is also trained on Aachen Day-Night dataset. However, HardNet and SOSNet are only trained on 450K 32×32 patches from Liberty dataset [7]. We will show in the next experiments that, these two models trained on patches labeled by an SfM pipeline are especially effective for 3D reconstruction tasks.

3D Reconstruction. We test our method on the ETH SfM benchmark [44] in the task of 3D reconstruction. We compare the reconstruction quality by comparing the number of registered images, reconstructed sparse and dense points, mean track length, and the reprojection error. Following [44], no nearest neighbour ratio test is conducted to better expose the matching performance of the descriptors. The reconstruction results are listed in Table 2. As shown, HardNet/SOSNet+D2D shows consistent performance increase in terms of the number of registered images, the number of sparse points, and the track length, which are important indicators of the reconstruction quality. This observation is expected as in this experiment, both HardNet and SOSNet are trained on local patches that are extracted and labeled via the SfM pipeline, and therefore are more suitable for this task.

Efficiency. In this experiment, we compare the feature extraction speed of several methods. Specifically, we record the extraction time over 108 image sequences in Hpatches [3], where there are 648 images with various resolutions (the average resolution is 775×978). All methods are tested on a RTX 2080 GPU, and the results are shown in Figure 5. SuperPoint and D2-Net has 1.3M and 15M parameters, respectively, whereas HardNet/SOSNet+D2D only relies 0.3M. Worth noting that R2D2 also uses similar architecture to HardNet/SOSNet, however it has no downsampling layers, thus the computational cost increases linearly with the depth. HardNet/SOSNet+D2D is slightly slower than SuperPoint, due to the extra time that is mostly spend on ranking the $\mathbf{S}_{\text{D2D}}$ score of keypoints, whereas SuperPoint takes a thresholding operation.

In summary, from the results on three different tasks with three different datasets we observe that with D2D,

[1] https://github.com/tsattler/visuallocalizationbenchmark/tree/master/local_feature_evaluation

[2] https://www.visuallocalization.net/

		# Image	# Registered	# Sparse Points	# Dense Points	Track Length	Reproj. Error
Fountain	SIFT	11	11	14K	292K	4.79	0.39px
	SuperPoint		11	7K	304K	4.93	0.81px
	D2-Net		11	19K	301K	3.03	1.40px
	HardNet+D2D		11	20K	304K	6.27	1.34px
	SOSNet+D2D		11	20K	305K	6.41	1.36px
Herzjesu	SIFT	8	8	7K	241K	4.22	0.43px
	SuperPoint		8	5K	244K	4.47	0.79px
	D2-Net		8	13K	221K	2.87	1.37px
	HardNet+D2D		8	13K	242K	5.73	1.29px
	SOSNet+D2D		8	13K	237K	6.06	1.34px
South Building	SIFT	128	128	108K	2.14M	6.04	0.54px
	SuperPoint		128	125k	2.13M	7.10	0.83px
	D2-Net		128	178K	2.06M	3.11	1.36px
	HardNet+D2D		128	193K	2.02M	8.71	1.33px
	SOSNet+D2D		128	184K	1.94M	8.99	1.36px
Madrid Metropolis	SIFT	1344	500	116K	1.82M	6.32	0.60px
	SuperPoint		702	125K	1.14M	4.43	1.05px
	D2-Net		787	229K	0.96M	5.50	1.27px
	HardNet+D2D		899	710K	1.13M	5.31	1.08px
	SOSNet+D2D		865	626K	1.15M	6.00	1.14px
Gendarmenmarkt	SIFT	1463	1035	338K	4.22M	5.52	0.69px
	SuperPoint		1112	236K	2.49M	4.74	1.10px
	D2-Net		1225	541K	2.60M	5.21	1.30px
	HardNet+D2D		1250	1716K	2.64M	5.32	1.16px
	SOSNet+D2D		1255	1562K	2.71M	5.95	1.20px

Table 2: Evaluation results on ETH dataset [44] for SfM. We can observe that with our proposed D2D, the shallow networks trained on local patches can significantly surpass deeper ones trained on larger datasets with full resolution images.

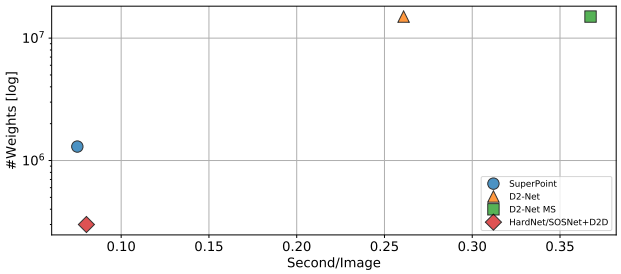

Figure 5: Comparison of efficiency.

patch-based descriptors HardNet and SOSNet can achieve competitive performance compared to joint detection-description methods such as D2-Net and SuperPoint. With significantly less parameters and faster speed, HardNet and SOSNet can achieve comparable/superior results to/than the state-of-the-art methods. These results validate our hypothesis that the networks trained for descriptors can be also used for detection.

4.2. Ablation Study

Combining D2D with joint methods. In order to further validate the effectiveness of the proposed D2D, we test it in combination with detect-and-describe methods[3] namely

[3]By the time of submission, there was no released code or model available for R2D2 [36], therefore we omitted it.

AS	RS	SuperPoint	D2-Net	HardNet	SOSNet
√		67.51	61.20	71.38	72.66
	√	67.58	60.07	69.32	72.77
√	√	**67.64**	**61.42**	**72.40**	**75.40**

Table 3: Ablative study in terms of Absolute Saliency(AS) and Relative Saliency(RS). Numbers are in terms of the average MMA on Hpatches [3] across pixel error threshold 1 to 10.

D2-Net[11] and SuperPoint [10]. Each of the two methods has its unique detection strategy: SuperPoint detects via thresholding of deep score maps while D2-Net selects local maxima. We adapt D2D in the following way: For SuperPoint, we generate a new threshhold α^* by:

$$\alpha^* = \frac{\mathbb{E}[\boldsymbol{S}_{\text{D2D}}\boldsymbol{S}_{\text{O}}]}{\mathbb{E}[\boldsymbol{S}_{\text{O}}]}\alpha, \tag{5}$$

where α and $\boldsymbol{S}_{\text{O}}$ are the original threshold and score map, respectively. For D2-Net, we choose local maxima that also have high $\mathbf{S}_{\text{D2D}}$. Specifically, if (x, y) is a keypoint than it should be detected by the non-maxima-suppression as well as have:

$$\boldsymbol{S}_{\text{D2D}}(x, y) > \mathbb{E}[\boldsymbol{S}_{\text{D2D}}] \tag{6}$$

In Figure 6, D2D improves the MMA score and the ratio of mutual nearest neighbour matches on Hpatches [3].

Method	#Dim	#Kp	0.5m, 2°	1m, 5°	5m, 10°
SuperPoint	256	3.7K	42.8	57.1	75.5
SuperPoint+D2D	256	3.7K	41.8	59.2	78.6
D2-Net SS	512	12K	44.9	66.3	88.8
D2-Net SS+D2D	512	8.3K	44.9	66.3	88.8

Table 4: Performance of combining D2D with SuperPoint [10] and D2-Net [11] on Aachen Day-Night [40, 38]

Moreover, in Table 4, SuperPoint+D2D achieves remarkably better localisation accuracy. D2-Net+D2D can maintain the same accuracy with much fewer detections indicating that keypoints not contributing to the localisation are filtered out by D2D. These results demonstrates that D2D can also improve the jointly trained detection-description methods.

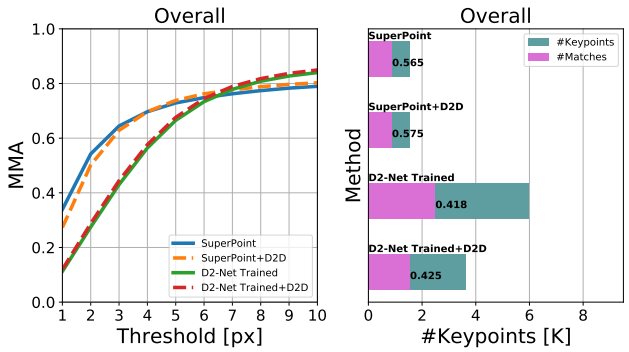

Figure 6: Performance of combining D2D with SuperPoint [10] and D2-Net [11] on Hpatches [3].

Impact of absolute and relative descriptor saliency. In Table 3, we show how $\boldsymbol{S}_{\text{AS}}$ and $\boldsymbol{S}_{\text{RS}}$ impact the matching performance. We observe that each of the two terms enables the detection, and the performance is further boosted when they are combined. This indicates that the absolute and relative saliency, *i.e.*, informativeness and distinctiveness of a point are two effective and complementary factors.

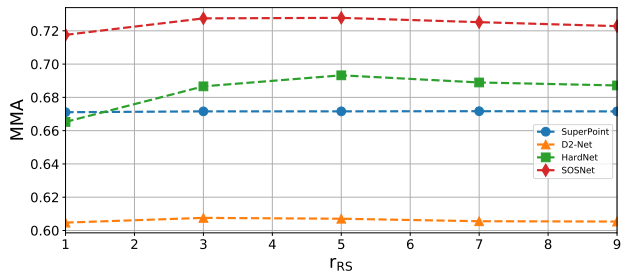

Figure 7: Performance in terms of MMA with different choice of r_{RS}.

Impact of r_{RS}. Matching performance in terms of different window size r_{RS} for computing relative saliency is shown in Figure 7, where the experiment is done using only $\boldsymbol{S}_{\text{RS}}$ as the keypoint score. For HardNet and SOSNet, the best r_{RS} is 5, which means that it is better to compare patches that are 20 pixels (stride 4 times 5) away from the center, which is approximately half of the receptive field size. Descriptors that are too close are indistinguishable.

Keypoint complementarity. Table 5 shows the results of a repeatability test across different descriptors combined with D2D. This is to demonstrate the complementarity of keypoints detected with different methods. The off diagonal scores are normalised with the diagonal scores for example, keypoints from HardNet+D2D are compared to those detected by SOSNet+D2D. Low normalised repeatability score indicates that the keypoints are mostly different *i.e.* different locations, thus the methods are complementary. Similarly HardNet and SOSNet give high score. This may be expected as both share the same architecture and similar training process. However, high repeatability between SuperPoint and D2-Net which indicates that the two descriptors are not complementary *i.e.* measure the same type of information that D2D uses for detecting keypoints.

	SuperPoint	D2-Net	HardNet	SOSNet
SuperPoint	1	1.0154	0.745	0.765
D2-Net	1.136	1	0.675	0.690
HardNet	0.849	0.729	1	0.952
SOSNet	0.868	0.738	0.950	1

Table 5: Keypoint repeatbality on Hpatches [3] with different detectors. Column:detector used on source image. Row:detector used on destination image. Numbers are the percentage of repeatbality change in terms of the original repeatbality (diagonal).

5. Conclusion

We proposed a new Describe-to-Detect (D2D) framework for the task of keypoint detection given dense descriptors. We have demonstrated that CNN models trained to describe can also be used to detect. D2D is simple, does not require training, is efficient and can be combined with any existing descriptor. We defined the descriptor saliency as the most important property and proposed an absolute and relative saliency measure to select keypoints that are highly informative in descriptor space and discriminative in their local spacial neighbourhood.

Our experimental evaluation on three different tasks and different datastes show that D2D offers a significant boost to the matching performance of various descriptors. It also improves results for camera localisation and 3D reconstruction.

Acknowledgements. This work was supported by the UK EPSRC research grant EP/S032398/1.

References

[1] P. F. Alcantarilla, A. Bartoli, and A. J. Davison. Kaze features. In *European Conference on Computer Vision*, pages 214–227. Springer, 2012. 2, 3

[2] J. Aldana-Iuit, D. Mishkin, O. Chum, and J. Matas. In the saddle: chasing fast and repeatable features. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 675–680. IEEE, 2016. 2, 3

[3] V. Balntas, K. Lenc, A. Vedaldi, and K. Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 4, page 6, 2017. 2, 4, 5, 6, 7, 8

[4] V. Balntas, E. Riba, D. Ponsa, and K. Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *British Machine Vision Conference (BMVC)*, volume 1, page 3, 2016. 2

[5] V. Balntas, L. Tang, and K. Mikolajczyk. Bold - binary online learned descriptor for efficient image matching. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 4

[6] A. Benbihi, M. Geist, and C. Pradalier. Elf: Embedded localisation of features in pre-trained cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7940–7949, 2019. 2

[7] M. Brown, G. Hua, and S. Winder. Discriminative learning of local image descriptors. *IEEE PAMI*, 33(1):43–57, 2011. 2, 5, 6

[8] P. H. Christiansen, M. F. Kragh, Y. Brodskiy, and H. Karstoft. Unsuperpoint: End-to-end unsupervised interest point detector and descriptor. *arXiv preprint arXiv:1907.04011*, 2019. 4

[9] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse. Monoslam: Real-time single camera slam. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):1052–1067, 2007. 1

[10] D. DeTone, T. Malisiewicz, and A. Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 224–236, 2018. 1, 2, 4, 6, 7, 8

[11] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler. D2-net: A trainable cnn for joint detection and description of local features. *arXiv preprint arXiv:1905.03561*, 2019. 1, 2, 3, 4, 5, 6, 7, 8

[12] H. Germain, G. Bourmaud, and V. Lepetit. S2dnet: Learning accurate correspondences for sparse-to-dense feature matching. *arXiv preprint arXiv:2004.01673*, 2020. 2

[13] C. G. Harris, M. Stephens, et al. A combined corner and edge detector. In *Alvey vision conference*, volume 15, pages 10–5244. Citeseer, 1988. 4

[14] K. He, Y. Lu, and S. Sclaroff. Local descriptors optimized for average precision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 596–605, 2018. 1, 2, 4

[15] T. Kadir and M. Brady. Saliency, scale and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001. 2, 3

[16] M. Keller, Z. Chen, F. Maffra, P. Schmuck, and M. Chli. Learning deep descriptors with scale-aware triplet networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018. 2

[17] A. B. Laguna, E. Riba, D. Ponsa, and K. Mikolajczyk. Key.net: Keypoint detection by handcrafted and learned cnn filters. *arXiv preprint arXiv:1904.00889*, 2019. 1, 2, 3

[18] K. Lenc and A. Vedaldi. Learning covariant feature detectors. In *European Conference on Computer Vision*, pages 100–117. Springer, 2016. 2

[19] K. Lenc and A. Vedaldi. Large scale evaluation of local image feature detectors on homography datasets. *arXiv preprint arXiv:1807.07939*, 2018. 2

[20] S. Leutenegger, M. Chli, and R. Siegwart. Brisk: Binary robust invariant scalable keypoints. In *2011 IEEE international conference on computer vision (ICCV)*, pages 2548–2555. Ieee, 2011. 2, 3

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

[22] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 60(2):91–110, 2004. 2, 3

[23] Z. Luo, T. Shen, L. Zhou, J. Zhang, Y. Yao, S. Li, T. Fang, and L. 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

[24] Z. Luo, T. Shen, L. Zhou, S. Zhu, R. Zhang, Y. Yao, T. Fang, and L. Quan. Geodesc: Learning local descriptors by integrating geometry constraints. In *European Conference on Computer Vision (ECCV)*, pages 170–185. Springer, 2018. 1

[25] Z. Luo, L. Zhou, X. Bai, H. Chen, J. Zhang, Y. Yao, S. Li, T. Fang, and L. Quan. Aslfeat: Learning local features of accurate shape and localization. *arXiv preprint arXiv:2003.10071*, 2020. 2

[26] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE PAMI*, 27(10):1615–1630, 2005. 2, 5

[27] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International journal of computer vision*, 65(1-2):43–72, 2005. 2

[28] A. Mishchuk, D. Mishkin, F. Radenovic, and J. Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4826–4837, 2017. 1, 2, 4, 5, 6

[29] D. Mishkin, F. Radenovic, and J. Matas. Repeatability is not enough: Learning affine regions via discriminability. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2, 6

[30] R. Mitra, N. Doiphode, U. Gautam, S. Narayan, S. Ahmed, S. Chandran, and A. Jain. A large dataset for improving patch matching. *arXiv preprint arXiv:1801.01466*, 2018. 2

[31] H. P. Moravec. Obstacle avoidance and navigation in the real world by a seeing robot rover. Technical report, Stanford Univ CA Dept of Computer Science, 1980. 3, 4

[32] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015. 1

[33] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3456–3465, 2017. 2, 3, 6

[34] Y. Ono, E. Trulls, P. Fua, and K. M. Yi. Lf-net: learning local features from images. In *Advances in Neural Information Processing Systems*, pages 6234–6244, 2018. 1, 2

[35] F. Radenović, G. Tolias, and O. Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018. 1, 2, 3

[36] J. Revaud, P. Weinzaepfel, C. De Souza, N. Pion, G. Csurka, Y. Cabon, and M. Humenberger. R2d2: Repeatable and reliable detector and descriptor. *arXiv preprint arXiv:1906.06195*, 2019. 1, 2, 4, 6, 7

[37] E. Rublee, V. Rabaud, K. Konolige, and G. R. Bradski. Orb: An efficient alternative to sift or surf. In *ICCV*, volume 11, page 2. Citeseer, 2011. 4

[38] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. 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. 5, 8

[39] T. Sattler, W. Maddern, A. Torii, J. Sivic, T. Pajdla, M. Pollefeys, and M. Okutomi. Benchmarking 6dof urban visual localization in changing conditions. *CoRR*, abs/1707.09092, 2017. 1

[40] T. Sattler, T. Weyand, B. Leibe, and L. Kobbelt. Image retrieval for image-based localization revisited. In *BMVC*, volume 1, page 4, 2012. 5, 8

[41] N. Savinov, A. Seki, L. Ladicky, T. Sattler, and M. Pollefeys. Quad-networks: unsupervised learning to rank for interest point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1822–1830, 2017. 2, 5

[42] B. Schiele and J. L. Crowley. Recognition without correspondence using multidimensional receptive field histograms. *International Journal of Computer Vision*, 36(1):31–50, 2000. 2, 3

[43] J. L. Schonberger and J.-M. Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016. 1, 6

[44] J. L. Schönberger, H. Hardmeier, T. Sattler, and M. Pollefeys. Comparative evaluation of hand-crafted and learned local features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5, 6, 7

[45] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, pages 501–518. Springer, 2016. 1

[46] X. Shen, C. Wang, X. Li, Z. Yu, J. Li, C. Wen, M. Cheng, and Z. He. Rf-net: An end-to-end image matching network based on receptive field. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8132–8140, 2019. 1

[47] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2

[48] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[49] A. W. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah. Visual tracking: An experimental survey. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1442–1468, 2013. 1

[50] Y. Tian, B. Fan, F. Wu, et al. L2-net: Deep learning of discriminative patch descriptor in euclidean space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 6, 2017. 1, 2, 4

[51] Y. Tian, X. Yu, B. Fan, F. Wu, H. Heijnen, and V. Balntas. Sosnet: Second order similarity regularization for local descriptor learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11016–11025, 2019. 1, 2, 4, 5, 6

[52] T. Tuytelaars, K. Mikolajczyk, et al. Local invariant feature detectors: a survey. *Foundations and trends® in computer graphics and vision*, 3(3):177–280, 2008. 1, 2, 3

[53] Y. Verdie, K. Yi, P. Fua, and V. Lepetit. Tilde: a temporally invariant learned detector. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5279–5288, 2015. 1, 2

[54] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua. Lift: Learned invariant feature transform. In *European Conference on Computer Vision*, pages 467–483. Springer, 2016. 1, 2

[55] L. Zhang and S. Rusinkiewicz. Learning to detect features in texture images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6325–6333, 2018. 2

[56] X. Zhang, F. X. Yu, S. Karaman, and S.-F. Chang. Learning discriminative and transformation covariant local feature detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6818–6826, 2017. 2