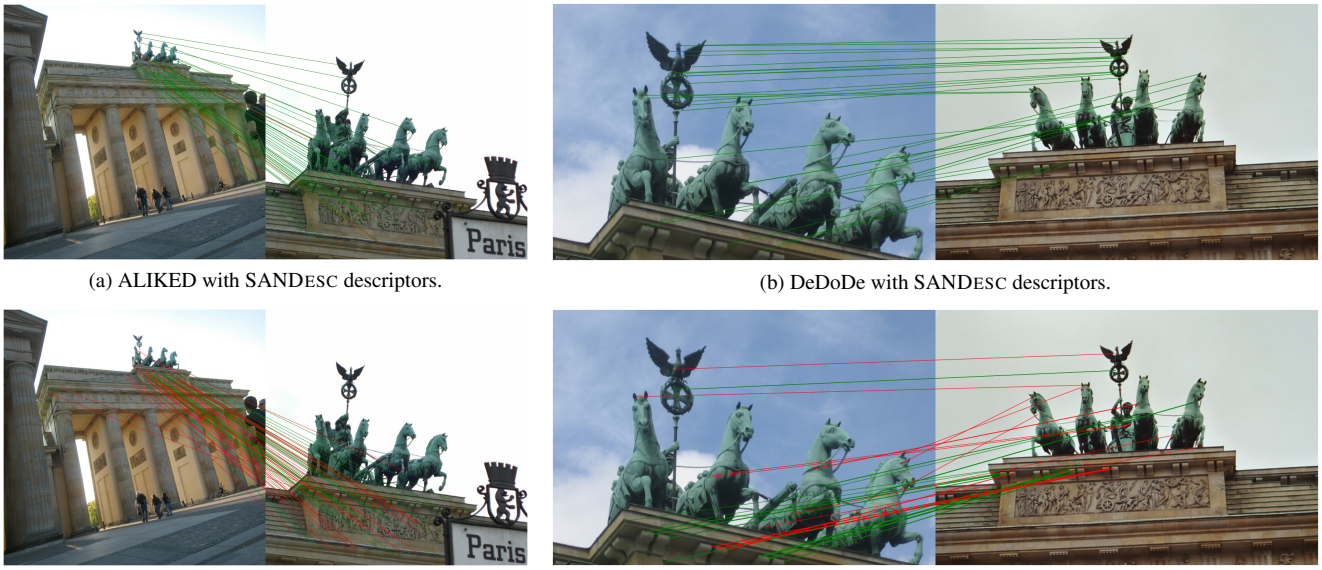
A Streamlined Attention-Based Network for Descriptor Extraction

Mattia D'Urso[1] Emanuele Santellani[1] Christian Sormann[2] Mattia Rossi[2]
Andreas Kuhn[2] Friedrich Fraundorfer[1]
[1]Graz University of Technology {name.surname@tugraz.at}, [2]Sony {name.surname@sony.com}



(a) ALIKED with SANDESC descriptors.

(b) DeDoDe with SANDESC descriptors.

(c) ALIKED with original descriptors.

(d) DeDoDe with original descriptors.

Figure 1. **Qualitative comparison.** Examples of local feature matching between pairs of images exhibiting large scale differences. Inliers are displayed in green; outliers in red.

Abstract

We introduce **SANDESC** *a Streamlined Attention-based Network for Descriptor extraction that aims to improve on existing architectures for keypoint description.*

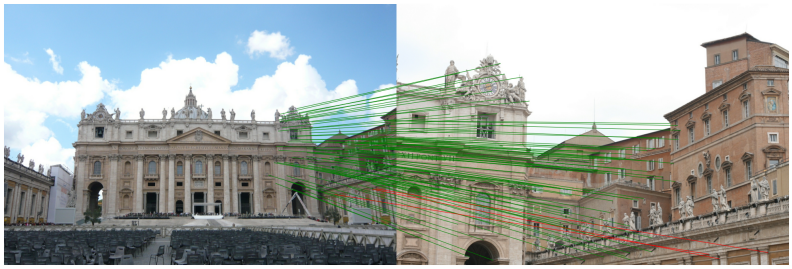
Our descriptor network learns to compute descriptors that improve matching without modifying the underlying keypoint detector. We employ a revised U-Net-like architecture enhanced with Convolutional Block Attention Modules and residual paths, enabling effective local representation while maintaining computational efficiency. We refer to the building blocks of our model as Residual U-Net Blocks with Attention. The model is trained using a modified triplet loss in combination with a curriculum learning–inspired hard negative mining strategy, which improves training stability.

Extensive experiments on HPatches, MegaDepth-1500, and the Image Matching Challenge 2021 show that training SANDESC *on top of existing keypoint detectors leads to improved results on multiple matching tasks compared to the original keypoint descriptors. At the same time,* SANDESC *has a model complexity of just 2.4 million parameters.*

As a further contribution, we introduce a new urban dataset featuring 4K images and pre-calibrated intrinsics, designed to evaluate feature extractors. On this benchmark, SANDESC *achieves substantial performance gains over the existing descriptors while operating with limited computational resources.*

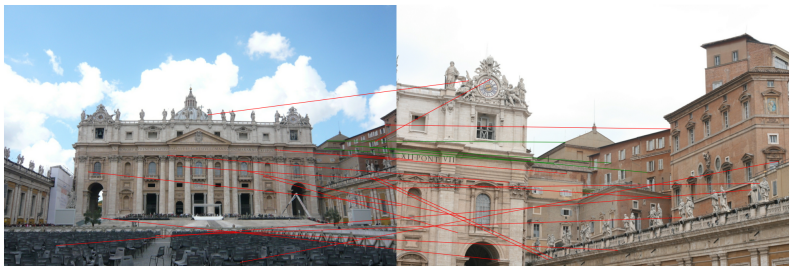
1. Introduction

Finding reliable correspondences between images is a fundamental problem in computer vision, with key applications in Structure-from-Motion (SfM) [32], Visual Localization (VL) [31] and Object Tracking [42]. The classical SfM pipeline involves three main stages: feature extraction (keypoint detection and description), feature matching and geo-



(a) SuperPoint with SANDESC descriptors.

(b) DISK with SANDESC descriptors.



(c) SuperPoint with original descriptors.

(d) DISK with original descriptors.

Figure 2. **Qualitative comparison.** Examples of local feature matching between pairs of images exhibiting large scale differences. Inliers are displayed in green; outliers in red.

metric verification, and 3D mapping. While early methods, such as SIFT [21] and SURF [3], use handcrafted features, modern descriptor learning techniques, such as L2-Net [35] and HardNet [23], have significantly improved matching performance.

A major shift occurred with the introduction of all-in-one networks, such as SuperPoint [4], which jointly optimize keypoint detection and description within a single unified architecture. However, coupling detection and description introduces interdependencies that can constrain flexibility and degrade overall performance, while recent works indicate that decoupling the two tasks not only simplifies the training process but also enhances robustness and generalization capabilities [16]. A growing line of methods, including XFeat [24], S-TREK [30], and DeDoDe [6] further highlight the benefits of decoupling keypoint from descriptor learning.

Despite the strong accuracy of DeDoDe's decoupled design [6], its descriptors' compute and memory demand is a bottleneck in long-duration sequences, large-scale SfM reconstructions, and with high-resolution imagery where more pixels inflate keypoints/descriptors and increase the cost of extraction and matching. This overhead limits the deployment on resource-constrained platforms and slows pipelines processing millions of descriptors, underscoring the need for a scalable descriptor architecture that preserves robustness while substantially reducing training and inference costs.

In order to address these issues, we propose SANDESC, a lightweight descriptor network architecture that aims at providing existing keypoint detectors with more discriminative descriptors than their native ones. In particular, SANDESC leverages a compact U-Net-like architecture enhanced with an attention mechanism from the Convolutional Block Attention Module (CBAM) [39], which allows for effective local feature representation while maintaining computational efficiency. The model is trained using a modified triplet-based loss combined with a curriculum-inspired hard negative mining strategy, improving both stability and discriminative power. SANDESC leverages the sequential dependence of description on detection: its descriptor performs best when trained on the keypoint distribution of the target detector, as discussed in Section 5.7.

Extensive evaluations on HPatches, MegaDepth-1500, and the Image Matching Challenge 2021 show that SANDESC consistently provides more robust descriptors across diverse detectors. Moreover, to evaluate performance on high-resolution imagery, we introduce a new high-resolution urban dataset with pre-calibrated intrinsics and poses estimated via COLMAP. This benchmark not only tests a model's ability to extract reliable local features in high-resolution images but also stresses its efficiency under constrained computational budgets.

Figures 1, 2 and 3 compare matches obtained with SANDESC against other method's native descriptors across challenging cases. These figures highlight failure cases and limitations of state-of-the-art methods. They further demonstrate that our proposed descriptors can address these challenging scenarios. Additional examples are provided in the Supplementary Material.

(a) RIPE with SANDESC descriptors.

(b) RIPE with original descriptors.

Figure 3. **Qualitative comparison.** Example of local feature matching between a pair of images exhibiting large scale differences. Inliers are displayed in green; outliers in red.

Our main contributions are as follows:

- We introduce a new attention-based descriptor architecture with only 2.4M parameters, trained using a revised triplet loss combined with a curriculum learning–inspired hard negative mining approach for stable and efficient learning.
- We present a new real-world benchmark consisting of carefully collected, high-resolution urban scenes captured with pre-calibrated intrinsics, extending the typical evaluation scheme with a practically relevant scenario.
- With extensive experiments across several established benchmarks, we demonstrate significant image matching improvements using the proposed SANDESC descriptor.

2. Related Work

Learned Patch Descriptors. In recent years, a significant number of works in the literature have studied the advantages of deep learning-based approaches compared to traditional handcrafted local descriptors such as SIFT [21] and SURF [3]. Early works, such as MatchNet [11] and DeepDesc [33], focused on learning descriptors from image patches extracted at given keypoint locations (often SIFT detections). These initial methods demonstrated the advantage of training deep networks to compare local patches, significantly improving upon classical descriptors. Subsequent works refined descriptor learning through improved objectives and mining strategies. For example, the triplet ranking loss [1] became a standard for training robust descriptors by maximizing the margin between the hardest positive and negative pairs in a batch, leading to methods such as L2-Net [35] and HardNet [23]. Additional regularization techniques, such as second-order similarity in SOS-Net [36], boosted descriptor invariance. By the end of the 2010s, learned patch descriptors achieved superior matching accuracy and robustness compared to handcrafted ones, as demonstrated in benchmarks such as HPatches [2] and the Image Matching Challenge 2021 [14].

Joint Detection and Description. Rather than relying on an external detector, a second generation of methods learned to perform detection and description jointly within a single network. SuperPoint [4] pioneered self-supervised joint learning in this context, inspiring a series of end-to-end models such as D2-Net [5], R2D2 [26], and MD-Net [29]. Notably, DISK [38] integrated a reinforcement learning approach optimizing a probabilistic detector-descriptor network with a cycle-consistency reward, allowing back-propagation through the detection process and achieving state-of-the-art results on the Image Matching Challenge 2021 [14].

The Rationale for Decoupling. An emerging line of research revisits the idea of decoupling detection and description, based on the observation that training keypoint detectors and descriptors together can create harmful interdependencies that reduce overall performance [16]. Li et al. [16] showed that training the detector independently, without forcing it to align with the descriptor during training, produces higher-quality keypoints, since the detector can optimize solely for its own task rather than compromising for descriptor compatibility.

Modern Decoupled Architectures. Edstedt et al. [6] further support this separation in their recent "Detect, Don't Describe" framework, summarized as DeDoDe. In DeDoDe, the detector is trained independently to identify repeatable, 3D-consistent keypoints, while the descriptor networks are trained at a later stage. Whereas this decoupled design achieves state-of-the-art results on benchmarks such as MegaDepth-1500, the performance gap between the two descriptors proposed implicitly highlights the central role of high-quality descriptors in feature matching and illustrates that performance is not solely determined by keypoint repeatability. Another example of decoupled design is S-TREK [30], a reinforcement learning–inspired feature extractor with a rotation-equivariant detector that excels at

detecting repeatable keypoints across views. S-TREK integrates translation- and rotation-equivariant layers into the detector and employs a reinforcement-learning–inspired sequential strategy to maximize repeatability under large in-plane rotations, while maintaining a lightweight descriptor head. Similarly, XFeat [24] targets runtime scalability: it combines a compute-efficient backbone with a dedicated keypoint branch and a semi-dense match-refinement module, achieving state-of-the-art accuracy while running up to $9\times$ faster than competing methods on low-resources device.

Building on these insights, we focus on descriptor robustness and scalability, addressing DeDoDe's descriptors computational demand with a lightweight architecture that yields strong descriptors at lower compute and memory cost. Our model can be trained on top of any keypoint detector, thus enabling flexible integration into existing pipelines. As we demonstrate in the following sections, this design choice leads to competitive or superior results on multiple challenging benchmarks.

3. Method

For our descriptor model, we adopt a fully convolutional U-Net-like architecture [27], as it provides a good trade-off between local detail preservation and computational efficiency. The proposed architecture begins with an initial $K \times K$ convolution projecting the input $I \in \mathbb{R}^{3\times H\times W}$ into a higher-dimensional feature space. This initial step is followed by four consecutive down- and up-sampling blocks, each implemented through Residual U-Net Blocks with Attention (RUBA). The model computes a final L2-normalized descriptor volume $V \in \mathbb{R}^{128\times H\times W}$.

3.1. Residual U-Net Block with Attention

Each Residual U-Net Block with Attention (RUBA) consists of a main (top) and a residual (bottom) paths, as illustrated in Figure 4. The main path resizes the input either by downsampling via average pooling or by upsampling via bilinear interpolation, followed by feature concatenation. A 1×1 convolution without bias is then used to align the main path with the residual path. The residual path adopts the *pre-activation layer* from He et al. [12], applying batch normalization, activation, and convolution three times in sequence, without weight sharing. The resulting residual features are refined using a Convolutional Block Attention Module (CBAM) [39]. CBAM applies two lightweight attention mechanisms: the *channel attention* and the *spatial attention*. The former mechanism, denoted as $M_c(\cdot)$, prioritizes the most informative feature channels by recalibrating their activations. The latter mechanism, $M_s(\cdot)$, enhances the most significant spatial regions within each feature map. Given a feature map $F \in \mathbb{R}^{C\times H\times W}$, CBAM first applies $M_c(\cdot)$ to enhance relevant channels and then $M_s(\cdot)$ to refine key spatial locations, thus enhancing the overall representational power of the feature map. The process can be expressed as defined in [39]: $F' = M_c(F) \odot F$ and $F'' = M_s(F') \odot F'$, where F' and F'' are the intermediate and final output feature maps, respectively, and $\odot$ denotes element-wise multiplication.

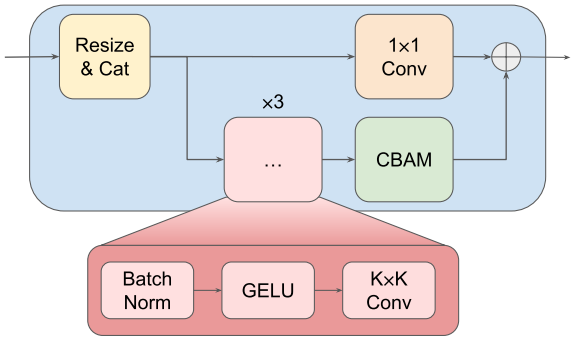


Figure 4. **Residual U-Net Block with Attention (RUBA).** This block serves as the fundamental building unit of the four-layer U-Net. We use four blocks in the encoder and four in the decoder, configured to resize the features through pooling and upsampling, respectively.

This sequential refinement allows CBAM to capture both channel-wise and spatial dependencies effectively. For further details on these attention mechanisms, we refer the reader to [39]. The refined residual is then added back to the main path.

All convolutions use kernel size K and stride 1, with zero-padding $\lfloor K/2 \rfloor$ to preserve spatial resolution. We fix $K = 5$ for convolutions in the residual paths, unless stated otherwise. Within CBAM, the channel-attention module $M_c(\cdot)$ is a two-layer MLP with reduction ratio $r = 16$ to model inter-channel dependencies, whereas the spatial-attention module $M_s(\cdot)$ uses a 7×7 convolution (independent of K) to capture broader spatial context. GELU [13] is used as the activation function throughout the network, including within CBAM, owing to its smooth nonlinearity and strong empirical performance in deep architectures.

3.2. Loss

We employ a variation of the triplet loss [23]. Local features are first matched using the Mutual Nearest Neighbor (MNN) criterion. When a match is established, we form a triplet of descriptors as follows: (i) the **anchor** A, taken from the first image; (ii) the **positive** P, the corresponding matched descriptor in the second image; and (iii) the **negative** N, the descriptor in the second image that achieves the second-best match score with A, selected according to the hardest-negative mining strategy of [23].

We compute the similarity scores s as:

$$s_p = A \cdot P, \quad s_n = A \cdot N, \tag{1}$$

where $\cdot$ denotes the dot product. Only triplets $t \in T$ that violate the margin constraint $s_p - s_n < m$, with margin m, are retained. This procedure is applied bidirectionally for each image pair (image 1 $\rightarrow$ image 2 and image 2 $\rightarrow$ image 1). The final triplet loss is defined as:

$$\mathcal{L}_{\text{Triplet}} = \frac{1}{|T|} \sum_{t \in T} \left(s_n^t - s_p^t \right). \tag{2}$$

To stabilize training and avoid premature plateauing due to overly difficult triplets, we incorporate a curriculum learning strategy: the hardest negative is used with probability $1 - \gamma$, and a random negative is chosen otherwise. γ is decayed after each training step, gradually introducing harder examples.

4. Training Setup

4.1. Dataset

We train our network on the MegaDepth [17] dataset following the protocols described in previous works [6, 7, 30, 38]. The dataset includes images, camera poses, intrinsic parameters, and depth maps obtained from 3D-reconstructed scenes. We adopt the dataset split proposed by DISK [38], excluding scenes overlapping with the Image Matching Benchmark [14]. For each scene, we randomly select image pairs from a predefined list of 10 000 triplets. Each image is resized such that its shortest side measures 512 pixels, then the longer dimension is cropped to produce square images. Additionally, images undergo random rotations with angles uniformly sampled from the range $[-30°, +30°]$.

4.2. Details

We follow the approach of training the descriptor on top of prior keypoint detections, as proposed in [6, 30]. For each method evaluated, we *separately* train a dedicated descriptor network, ensuring that the resulting models are specifically optimized for the corresponding keypoint detectors.

We use the AdamW [20] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 0.01. At the beginning of the training, we linearly increase the learning rate from the minimum value η_{min} to the maximum value η_{max} over the first w warm-up steps to carefully build the gradient moments. After the warm-up phase, the learning rate is decayed exponentially by a factor d at every iteration, until η_{min} is reached. We set $\eta_{\text{max}} = 0.005$, $\eta_{\text{min}} = 0.0001$, $w = 2048$, and $d = 0.99996$.

For the loss function, we initialize parameter γ to 1 and apply an exponential decay with a factor of 0.9993 at each training step. The loss margin α is fixed at 0.5.

We adopt automatic mixed-precision training to increase memory efficiency and enable larger batch sizes. Specifically, we use a batch size of 14 on a single NVIDIA RTX 4090 GPU. The model is trained on approximately 250 000 image pairs. The total training time depends on the speed of the chosen detector.

5. Experiments

We evaluate our descriptors on several benchmarks, comparing them against the following state-of-the-art methods.

- **SuperPoint** [4] is an all-in-one model that jointly trains keypoints and descriptors, first on synthetic images for corner detection and then using homographic adaptation for fine-tuning, all within a self-supervised training framework. The authors employ a hinge-loss-based function that minimizes the L2-distance between descriptor pairs at corresponding locations, while maximizing distances for non-corresponding pairs. It is trained on Synthetic Shapes [4] and MS-COCO 2014 [18].
- **DISK** [38] is an end-to-end trainable U-Net-based model that learns keypoint detection and description jointly using a policy gradient reinforcement learning framework. It leverages ground-truth geometry (including depth) to assign positive rewards to correct feature matches and vice versa. It is trained on Megadepth [17].
- **RIPE** [15] is a reinforcement learning–based framework that trains a VGG-19 model pretrained on ImageNet as an all-in-one detector–descriptor. It maximizes an epipolar geometry–based reward to guide learning, requiring only labels indicating whether two images belong to the same scene. It extracts descriptors from multi-scale hypercolumn features. RIPE is trained on the MegaDepth [17], Tokyo 24/7 [37], and ACDC [28] datasets.
- **ALIKED** [41] builds upon ALIKE [40], incorporating a Sparse Deformable Descriptor Head (SDDH) that extracts descriptors exclusively at keypoint locations. Descriptors are trained with a sparse variation of the Neural Reprojection Error loss [9, 40]. It is trained on Megadepth [17], R2D2 [26], Oxford and Paris [25], and the Aachen [31] datasets.
- **DeDoDe** [6] separately trains detector and descriptor models starting from the same pre-trained VGG-19 model. The authors propose *two descriptors* denoted as -B and -G, paired with the *same detector*. The -G version incorporates DINOv2 [22] features, resulting in a significantly larger model. Both descriptors are trained via a MNN negative log-likelihood loss on Megadepth [17].

To ensure a fair comparison, we re-evaluated all methods under the same experimental conditions. We compare them using both their original descriptors and SANDESC across three tasks: *homography estimation* on HPatches, *stereo matching* on the Phototourism test set from the Image Matching Challenge 2021, and finally *stereo pose recovery* on MegaDepth-1500 and on our newly introduced Graz4K dataset.

Method	MMA↑ @1px	MMA↑ @2px	MMA↑ @3px	MS↑ @1px	MS↑ @2px	MS↑ @3px	Hom. Acc.↑ @1px	Hom. Acc.↑ @2px	Hom. Acc.↑ @3px
SuperPoint	30.3	50.6	62.2	19.0	31.5	38.4	37.6	63.7	75.0
↳ w/ SANDESC	**31.7**	**53.7**	**65.8**	**20.9**	**35.0**	**42.6**	**41.1**	**66.3**	**78.3**
DISK	**45.4**	**67.1**	**76.3**	27.7	39.9	44.8	39.8	61.9	71.3
↳ w/ SANDESC	40.9	59.9	68.1	**27.8**	**40.1**	**45.3**	**42.0**	**64.8**	**75.7**
RIPE	**38.3**	**58.2**	67.3	25.4	38.0	43.6	44.2	67.8	**78.9**
↳ w/ SANDESC	38.0	**58.2**	**67.8**	**25.9**	**39.1**	**45.2**	**44.3**	**70.6**	78.5
ALIKED	**41.8**	**64.3**	**73.6**	**28.6**	**43.0**	**48.8**	40.4	66.7	**77.6**
↳ w/ SANDESC	41.1	62.5	70.9	**28.6**	42.6	48.0	**42.6**	**68.1**	77.0
DeDoDe-B	44.7	62.0	69.2	22.4	30.6	33.9	50.6	71.3	79.3
DeDoDe-G	**45.6**	**63.9**	**71.7**	**22.7**	**31.3**	**34.8**	50.4	71.3	78.7
↳ w/ SANDESC	44.2	60.5	67.0	22.4	30.2	33.2	**52.0**	**71.5**	**79.4**

Table 1. HPatches results with budget of 2048 keypoints. DeDoDe-B and -G share the same detector. DeDoDe+SANDESC overall achieves the highest Hom. Acc.

Method	BM	FLC	LM	LB	MC	MR	PSM	SF	SPC	AVG
SuperPoint	14.1	33.6	37.9	30.4	13.7	13.0	7.1	24.9	17.6	21.3
↳ w/ SANDESC	**25.6**	**50.0**	**40.5**	**45.9**	**23.9**	**17.9**	**10.2**	**36.1**	**34.8**	**31.6**
DISK	**24.1**	54.7	**52.4**	**56.5**	35.4	26.5	**13.7**	**44.3**	42.3	**38.8**
↳ w/ SANDESC	24.0	**56.0**	50.6	54.7	**36.5**	**26.8**	12.7	44.0	**44.0**	**38.8**
RIPE	18.7	53.6	**48.9**	**54.0**	27.4	19.9	10.2	39.1	41.1	34.7
↳ w/ SANDESC	**25.5**	**55.0**	44.5	53.3	**31.8**	**25.1**	**12.3**	**42.1**	**44.0**	**37.0**
ALIKED	**27.7**	61.0	**44.2**	54.2	37.7	25.1	**17.2**	**46.5**	47.9	40.1
↳ w/ SANDESC	19.5	**61.2**	41.7	**58.8**	**41.9**	**30.9**	15.6	46.3	**50.7**	**40.7**
DeDoDe-B	26.8	52.6	34.5	45.2	32.4	19.9	10.8	37.7	44.4	33.8
DeDoDe-G	**28.3**	**55.2**	**36.8**	**56.4**	**34.3**	19.1	**12.8**	**43.4**	**46.6**	**36.9**
↳ w/ SANDESC	24.7	53.4	34.2	45.1	30.3	**24.7**	10.4	41.2	44.2	34.2

Table 2. IMC21 results in terms of AUC@5. The keypoints budget is set to 2048. DeDoDe-B and -G share the same detector. ALIKED+SANDESC overall achieves the highest average score.

5.1. HPatches

HPatches [2] is a widely used benchmark consisting of image sequences exhibiting either viewpoint changes or illumination variations. Following the protocol of [5, 30], we evaluate on a subset of 108 scenes, each comprising one reference image and five target images with corresponding ground-truth homographies. We fix the keypoints budget to 2048 and perform feature matching using MNN. Performance metrics are computed pairwise between the reference and each target. We report the **Mean Matching Accuracy (MMA)** as the percentage of correct matches within a threshold of ϵ pixels; the **Matching Score (MS)** as the number of correct matches (up to a pixel threshold) divided by the average number of keypoints in the overlapping area between the two images; and the **Homography Accuracy (Hom. Acc.)** as the Area Under the Curve (AUC) of the percentage of estimated homographies with corner error below ϵ [4, 30]. The corner error is computed as the average distance between the four corners of the reference image and the corresponding corners of the target image warped using the estimated homography. Following [30], we compute the relative homography using the OpenCV `findHomography` function with multiple RANSAC thresholds and report the highest Hom. Acc. score for each method. This last metric is often considered the most important for this benchmark, as it measures the effectiveness on a typical downstream SfM task.

We do not include the repeatability metric in our experiments, neither here nor later, as it relates only to the keypoints, which are not modified by SANDESC.

As shown in Table 1, SANDESC trades a small drop in MMA for higher geometric reliability, with MS preserved or improved and Hom. Acc. frequently increasing. The gains are pronounced for SuperPoint, where Hom. Acc. rises by 3.5 at 1 px and MS improves by up to 4.2. With DISK, MS improves slightly and Hom. Acc. increases consistently. RIPE and ALIKED show mixed changes in MMA yet still gain in Hom. Acc. at tight thresholds. Overall, paired with DeDoDe, SANDESC achieves the best Hom. Acc. across all methods: 52.0 at 1 px, 71.5 at 2 px, and 79.4 at 3 px.

5.2. The Image Matching Challenge 2021

The Image Matching Challenge 2021 (IMC) [14] evaluates local feature matching methods in complex real-world settings. We adopt the IMC 2021 Phototourism test set, which consists of nine scenes, each comprising 100 tourist photos captured with diverse cameras, viewpoints, and lighting conditions. Each scene is identified by its capitals. Images within each scene are exhaustively compared. Following the authors' protocol, we evaluate estimated poses using the AUC of relative pose accuracy with a 5 degrees error threshold. Pose error is defined as the larger of the rotation and translation errors, with failures assigned when this error exceeds 10 degrees. For each method, we first extract keypoints and descriptors with its original pipeline, then replace the descriptors with SANDESC for a fair comparison. The results are summarized in Table 2.

The evaluation shows that combining ALIKED with SANDESC attains the best average performance (AVG) on this benchmark, demonstrating the effectiveness of our descriptor. Overall, all methods except DeDoDe-G benefit from pairing with SANDESC. In particular, SuperPoint and RIPE exhibit substantial improvements with SANDESC, whereas ALIKED, DISK, and DeDoDe-B achieve results comparable to their native descriptors.

5.3. Megadepth-1500

MegaDepth-1500 (MD1500), originally introduced in LoFTR [34], is a curated subset of the MegaDepth dataset and has since been adopted in several follow-up works [6–8, 10, 19]. The pairs are chosen to maintain a uniform covisibility ratio across the dataset, in contrast to the IMC, where the covisibility distribution follows a Gaussian-like shape [14]. We use the same evaluation protocol used by [6, 7] and a score of 90 degrees is given when the fundamental matrix recovery fails. Table 3 presents the results for all the

Keypoints budget→	2048		30 000	
Method	**AUC@5↑**	**AUC@10↑**	**AUC@5↑**	**AUC@10↑**
SuperPoint	30.1	44.6	15.6	29.2
↳ w/ SANDESC	**42.0**	**59.2**	**35.1**	**52.4**
DISK	35.3	52.0	43.2	58.1
↳ w/ SANDESC	**37.0**	**54.5**	**46.0**	**61.4**
RIPE	43.1	59.1	36.3	52.3
↳ w/ SANDESC	**43.6**	**59.4**	**42.1**	**57.7**
ALIKED	42.0	57.5	41.5	56.5
↳ w/ SANDESC	**43.4**	**60.1**	**48.3**	**63.0**
DeDoDe-B	43.2	59.8	51.1	65.6
DeDoDe-G	46.1	**63.5**	**55.3**	**71.2**
↳ w/ SANDESC	**46.2**	62.5	52.2	67.1

Table 3. MD1500 results with 2048 and 30 000 keypoints budgets. DeDoDe-B and -G share the same detector. DeDoDe+SANDESC scores the highest AUC@5 with 2048 keypoints.

Resolution →	FHD	QHD	4K	FHD	QHD	4K
Method	**AUC@5↑**			**AUC@10↑**		
SuperPoint	39.9	38.4	32.4	54.4	52.4	44.7
↳ w/ SANDESC	**63.0**	**62.2**	**57.7**	**74.8**	**73.9**	**69.8**
DISK	40.8	38.2	33.5	53.4	50.1	44.8
↳ w/ SANDESC	**55.5**	**54.0**	**48.3**	**69.1**	**66.9**	**60.9**
RIPE	54.5	46.5	34.0	67.0	59.0	45.1
↳ w/ SANDESC	**64.0**	**62.8**	**56.7**	**75.5**	**74.4**	**68.2**
ALIKED	55.0	51.7	44.4	67.7	64.9	57.7
↳ w/ SANDESC	**64.0**	**64.5**	**61.8**	**75.9**	**75.6**	**73.4**
DeDoDe–B	56.9	OOM		70.3	OOM	
DeDoDe–G	52.5	OOM		66.5	OOM	
↳ w/ SANDESC	**57.4**	**54.4**	**52.2**	**70.7**	**67.4**	**65.3**

Table 4. Graz4K results with a 2048 keypoints budget for three resolutions. DeDoDe-B and -G share the detector. ALIKED+SANDESC achieves the highest scores in all cases.

evaluated methods, both with original and SANDESC descriptors, at two keypoints budgets set to 2048 and 30 000.

At the 5 degrees (2048 keypoints), all detectors benefit from the use of SANDESC descriptors with the combination DeDoDe+SANDESC achieving the highest accuracy. In all the other columns, SANDESC outperforms all the other descriptors except DeDoDe-G.

Notably, similar to the results in Table 1, SANDESC descriptors allow SuperPoint to significantly improve from 30.1 to 42.0 at 2048 keypoints, a relative gain of 40%, thus outperforming more recent methods like DISK. Interestingly, SuperPoint, RIPE, and ALIKED lose accuracy when the keypoints budget is significantly increased. SANDESC mitigates this issue, enabling stable performance in high-keypoint regimes and delivering substantial gains.

5.4. Graz4K

High-resolution imagery is often essential for capturing fine structure and achieving high accuracy. However, this regime stresses the compute and memory budgets of feature extractors, exposing scalability limits. To evaluate models under these conditions, we curate a new dataset, called Graz4K, comprising six urban scenes captured with three cameras. Each camera was calibrated with OpenCV's ArUco workflow; these intrinsics were then supplied as priors for sparse reconstruction in COLMAP [32] using the default settings. Across all scenes, the resulting sparse models achieved a mean reprojection error of 0.97 ± 0.54 px over a total of 1 331 640 3D points. More details can be found in Supplementary Material. Footage was recorded in 4K at 30 fps and sampled at 1 fps. Then, we export the view graphs from COLMAP databases and prune them keeping only every tenth image pair. We further discard pairs that are too easy or too hard, retaining a pair only if it yields between 100 and 1 000 matches. After filtering, the benchmark comprises 1 866 images and 4 413 image pairs. We evaluate at three resolutions: native 4K (3840×2160) and downscaled QHD (2560×1440) and FHD (1920×1080). Results were computed following the MD1500 protocol on an NVIDIA RTX 4090 with 24GB using mixed precision.

As reported in Table 4, SANDESC improves over the original descriptors in all scenes and at each resolution. Notably, as image resolution increases, SANDESC delivers greater performance gains and is less sensitive to these increases, as evidenced by the large drops in the original models. Furthermore, in this benchmark, SuperPoint exhibits a remarkable improvement when coupled with SANDESC, especially at 4K, achieving the highest gain of +78%. The combination of ALIKED and SANDESC is the top-performing method in this evaluation. DeDoDe-B and -G run out of memory (OOM) when dealing with inputs larger than FHD on a 24GB VRAM GPU. Despite DeDoDe's impressive performance, the substantial memory footprint of both its descriptors makes it unsuitable for scenarios characterized by high resolution images or limited computational resources, thus limiting its general applicability.

5.5. Speed Comparison

Table 5 compares per image processing time in milliseconds for keypoint detection and description under a budget of 2048 keypoints, and reports the model size, in millions of parameters, in the corresponding column. All images were processed in FHD on an NVIDIA RTX 4090 with 24GB.

Overall, methods such as DISK, SuperPoint, ALIKED, and RIPE run slower when paired with SANDESC, since all-in-one pipelines reuse intermediate features to compute descriptors, whereas SANDESC operates directly on raw images. By contrast, with the decoupled DeDoDe methods, SANDESC remains competitive: its accuracy matches DeDoDe-G and exceeds DeDoDe-B, while its runtime is close to DeDoDe-B and faster than DeDoDe-G. SANDESC alone requires approximately 87 ms on our hardware.

Method	Size	Speed (ms) ↓ Orig.	Ours	VRAM (GB) ↓ Orig.	Ours
DISK	0.26	62.1±0.1	135.2±0.2	6.96	7.01
SuperPoint	0.30	18.3±0.2	114.8±0.2	3.50	7.18
ALIKED	0.32	19.3±0.3	114.2±0.2	3.89	5.76
RIPE	0.24	175.9±0.2	275.0±0.2	6.55	8.42
DeDoDe-B	15.1	181.5±0.1	189.0±0.2	8.11	5.78
DeDoDe-G	323.2	316.8±0.4	189.0±0.2	9.31	5.78

Table 5. Columns report the number of parameters in millions (M), the speed (ms) and the VRAM usage (GB) for original methods and with our descriptor, respectively. Our method has 2.4M parameters. Tests use FHD images on a NVIDIA RTX 4090.

5.6. Ablation

In this section we focus on the impact that different architectural choices and training strategies have on SANDESC results and runtime. Specifically, in Table 6, we evaluate SANDESC descriptor with DeDoDe and RIPE detectors on the MD1500 benchmark. We report AUC@5 and our descriptor inference time in milliseconds while incrementally adding the architectural components and training strategies under examination.

For both detectors, the largest improvement comes from the *random negative decay* strategy, which increases performance by 7.4 and 7.3 points, respectively. This approach stabilizes training by gradually shifting from easier to harder negative samples. Adding residual paths to the RUBA blocks yields further gains of 1.6 and 1.4 points by improving gradient flow during back-propagation. Finally, incorporating the CBAM attention mechanism provides an additional boost of 1.4 and 0.3 points.

Method	DeDoDe	RIPE	ms
Plain U-Net	35.8	34.6	30.1±0.1
+ Random Negative Decay	43.2	41.9	30.1±0.1
+ Residual Paths	44.8	43.3	37.1±0.1
+ Attention (Full)	**46.2**	**43.6**	86.7±0.1

Table 6. Ablation study on MD1500 in terms of AUC@5 with 2,048 keypoints.

5.7. The Need To Retrain on Each Detector

The training framework we propose in this paper trains a **SANDESC** model *separately* for each specific detector. Nevertheless, it is possible to train a SANDESC variant on randomly generated keypoints to remain *detector-agnostic* (DA). Specifically, projecting the image grid back and forth between the two images, retaining only points whose reprojection error is below a small pixel threshold, and then sampling uniformly. We call this version **SANDESC DA**.

To asses descriptors compatibility and the sensitivity of descriptors to detector shift, we run a full cross detector-descriptor evaluation with the methods listed in Section 5. Specifically, we pair each detector with every descriptor under a fixed matching pipeline and test on the MD1500 dataset.

Descriptors	Detectors ALIKED	DeDoDe	DISK	RIPE	SuperPoint	AVG
ALIKED	42.0	38.5	10.3	12.7	37.8	28.3
DeDoDe-B	42.0	43.2	35.7	41.7	40.7	40.7
DeDoDe-G	**45.3**	46.1	36.9	**44.5**	**43.9**	**43.3**
DISK	38.5	38.3	35.3	37.6	35.1	37.0
RIPE	42.5	39.0	34.6	43.1	37.1	39.3
SuperPoint	33.3	29.3	26.7	30.8	30.1	30.0
SANDESC DA	41.7	44.6	36.7	42.9	41.9	41.9
SANDESC	43.4	**46.2**	**37.0**	43.6	42.0	42.4

Table 7. Cross detector-descriptor evaluation on MD1500 in terms of AUC@5 with 2,048 keypoints budget. Boxed entries highlight native detector–descriptor pairs. Last column reports per-descriptor average score.

Table 7's last column reports the average score across detectors for each descriptor. DeDoDe-G emerges as the most flexible and effective descriptor, thanks to its higher parameter count and extensive training. Nevertheless, both SANDESC and SANDESC DA follow closely falling short by only 0.9 (2%) and 1.4 (3.2%) AUC@5 points, respectively. The detector-specific training scheme consistently yields superior performance compared with the DA model, thereby justifying our methodological choice.

6. Conclusion

In this work, we introduced **SANDESC**, an efficient and robust descriptor architecture that can be trained on top of existing keypoint detectors to improve over their original descriptors. SANDESC employs a lightweight U-Net architecture that leverages channel and spatial attention through the proposed **RUBA** block to produce more robust descriptors. We trained SANDESC using a modified triplet loss combined with a curriculum-learning-inspired hard-negative-mining strategy to stabilize the training.

We demonstrated that **SANDESC** improves multiple existing detectors on several popular benchmarks, including HPatches, the Image Matching Challenge 2021, and MegaDepth-1500. In addition, we introduced the Graz4K datasets showing that SANDESC performance gracefully scales with high-resolution images, which is crucial where accuracy is critical. We further show that our framework remains competitive with the very large DeDoDe-G model, while requiring only a fraction of its cost.

Acknowledgements This work has been supported by the FFG under Contract No. 881844 within the project "Pro²Future".

References

[1] Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, page 3. BMVA, 2016. 3

[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 (CVPR)*, pages 5173–5182. IEEE, 2017. 3, 6

[3] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346 – 359, 2008. 2, 3

[4] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self–supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR W)*, pages 224–236. IEEE, 2018. 2, 3, 5, 6

[5] Mihai Dusmanu, Ignacio Rocco, Tomáš 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8092–8101. IEEE, 2019. 3, 6

[6] Johan Edstedt, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. DeDoDe: Detect, don't describe—describe, don't detect for local feature matching. In *Proceedings of the 2024 International Conference on 3D Vision (3DV)*, pages 148–157. IEEE, 2024. 2, 3, 5, 6

[7] Johan Edstedt, Georg Bökman, and Zhenjun Zhao. DeDoDe v2: Analyzing and improving the dedode keypoint detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4245–4253. IEEE, 2024. 5, 6

[8] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19790–19800. IEEE, 2024. 6

[9] Hugo Germain, Vincent Lepetit, and Guillaume Bourmaud. Neural reprojection error: Merging feature learning and camera pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 414–423, 2021. 5

[10] Pierre Gleize, Weiyao Wang, and Matt Feiszli. SILK: Simple learned keypoints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22499–22508. IEEE, 2023. 6

[11] Xufeng Han, Thomas Leung, Yangqing Jia, Rahul Sukthankar, and Alexander C. Berg. Matchnet: Unifying feature and metric learning for patch–based matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3279–3286. IEEE, 2015. 3

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 630–645. Springer, 2016. 4

[13] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016. 4

[14] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image matching across wide baselines: From paper to practice. *International Journal of Computer Vision*, 129(2):517–547, 2021. 3, 5, 6

[15] Johannes Künzel, Anna Hilsmann, and Peter Eisert. Ripe: Reinforcement learning on unlabeled image pairs for robust keypoint extraction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4868–4877, 2025. 5

[16] Kunhong Li, Longguang Wang, Li Liu, Qing Ran, Kai Xu, and Yulan Guo. Decoupling makes weakly supervised local feature better. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15838–15848. IEEE, 2022. 2, 3

[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 (CVPR)*, pages 2041–2050. IEEE, 2018. 5

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[19] Philipp Lindenberger, Paul–Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17627–17638. IEEE, 2023. 6

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 5

[21] David G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision (ICCV)*, pages 1150–1157. IEEE, 1999. 2, 3

[22] Oquab Maxime, Darcet Timothée, Moutakanni Théo, Vo Huy, Marc Szafraniec, Khalidov Vasil, Fernandez Pierre, Haziza Daniel, Massa Francisco, El-Nouby Alaaeldin, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, pages 1–31, 2024. 5

[23] Anastasiia Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. *Advances in Neural Information Processing Systems*, 30:640–651, 2017. 2, 3, 4

[24] Guilherme Potje, Felipe Cadar, André Araujo, Renato Martins, and Erickson R Nascimento. Xfeat: Accelerated features for lightweight image matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2682–2691, 2024. 2, 4

[25] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting oxford and paris:

Large-scale image retrieval benchmarking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5706–5715, 2018. 5

[26] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2D2: Reliable and repeatable detector and descriptor. *Advances in Neural Information Processing Systems*, 32:124–135, 2019. 3, 5

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U–Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015. 4

[28] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10765–10775, 2021. 5

[29] Emanuele Santellani, Christian Sormann, Mattia Rossi, Andreas Kuhn, and Friedrich Fraundorfer. MD–Net: Multi–detector for local feature extraction. In *Proceedings of the 26th International Conference on Pattern Recognition (ICPR)*, pages 3944–3951. IEEE, 2022. 3

[30] Emanuele Santellani, Christian Sormann, Mattia Rossi, Andreas Kuhn, and Friedrich Fraundorfer. S–Trek: Sequential translation and rotation equivariant keypoints for local feature extraction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9728–9737. IEEE, 2023. 2, 3, 5, 6

[31] 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

[32] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 7

[33] 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 (ICCV)*, pages 118–126. IEEE, 2015. 3

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

[35] Yurun Tian, Bin Fan, and Fuchao Wu. L2–Net: Deep learning of discriminative patch descriptor in euclidean space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 661–669. IEEE, 2017. 2, 3

[36] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11016–11025. IEEE, 2019. 3

[37] Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1808–1817, 2015. 5

[38] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. DISK: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33:14254–14265, 2020. 3, 5

[39] Sanghyun Woo, Jongchan Park, Joon–Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19. Springer, 2018. 2, 4

[40] Xiaoming Zhao, Xingming Wu, Jinyu Miao, Weihai Chen, Peter C. Y. Chen, and Zhengguo Li. Alike: Accurate and lightweight keypoint detection and descriptor extraction. *IEEE Transactions on Multimedia*, 25:3101–3112, 2022. 5

[41] Xiaoming Zhao, Xingming Wu, Weihai Chen, Peter C. Y. Chen, Qingsong Xu, and Zhengguo Li. ALIKED: A lighter keypoint and descriptor extraction network via deformable transformation. *IEEE Transactions on Instrumentation and Measurement*, 72:1–16, 2023. 5

[42] Huiyu Zhou, Yuan Yuan, and Chunmei Shi. Object tracking using sift features and mean shift. *Computer vision and image understanding*, 113(3):345–352, 2009. 1