

RIPE: Reinforcement Learning on Unlabeled Image Pairs for Robust Keypoint Extraction

Johannes Künzel^{1,2} Anna Hilsmann¹ Peter Eisert^{1,2}

¹Fraunhofer Heinrich-Hertz-Institut, HHI, Germany, ²Humboldt University Berlin, Germany,

Abstract

We introduce *RIPE*, an innovative reinforcement learning-based framework for weakly-supervised training of a keypoint extractor that excels in both detection and description tasks. In contrast to conventional training regimes that depend heavily on artificial transformations, pre-generated models, or 3D data, *RIPE* requires only a binary label indicating whether paired images represent the same scene. This minimal supervision significantly expands the pool of training data, enabling the creation of a highly generalized and robust keypoint extractor.

RIPE utilizes the encoder’s intermediate layers for the description of the keypoints with a hyper-column approach to integrate information from different scales. Additionally, we propose an auxiliary loss to enhance the discriminative capability of the learned descriptors.

Comprehensive evaluations on standard benchmarks demonstrate that *RIPE* simplifies data preparation while achieving competitive performance compared to state-of-the-art techniques, marking a significant advancement in robust keypoint extraction and description. To support further research, we have made our code publicly available at <https://github.com/fraunhoferhhi/RIPE>.

1. Introduction

Given two images, how can we determine whether they depict the same scene and precisely identify matching keypoints? This task, as shown in Fig. 1, is intuitive for humans: we can identify distinctive keypoints in one image and look for their counterparts in the other, ignoring distractions such as moving cars, changing foliage, all while remaining unaffected by noise or lighting variations. This natural human capability raises an intriguing question: can neural networks learn robust keypoint extraction exclusively from binary labels indicating whether two images depict the same scene?

Traditional keypoint detection methods such as SIFT [21], ORB [29], and SURF [3] rely on handcrafted feature detectors and descriptors that struggle significantly with long-term registration tasks, particularly when images are taken hours,



Figure 1. Can you tell if these images depict the same scene? While humans naturally ignore noise and lighting variations to solve this effortlessly, teaching a neural network to do the same using only image pairs poses significant challenges. (Images: MegaDepth [18])

days or months apart. Variations in weather, lighting conditions and appearance (most notably from vegetation) pose significant problems for traditional detectors. To overcome these limitations, deep-learning-based methods were introduced to directly learn feature representations from images. Current state of the art (SOTA) approaches, including DeDoDe [9], DISK [38] and ALIKED [47] rely on training datasets like MegaDepth [18], which provide relative pose and depth information. MegaDepth, for example, consists of touristic imagery along with 3D models generated with Structure from Motion using COLMAP [34], itself reliant on classical SIFT keypoints. Other methods, like SuperPoint [6] or SiLK [13] rely on self-supervision through applying artificial augmentations, but face limitations from the available training data and the domain-gap between real and simulated scenarios, which inadequately replicate long-term changes of the real world.

In this paper we propose *RIPE* (**R**einforcement Learning on **I**mage **P**airs for **K**eypoint **E**xtraction), a novel approach that trains keypoint detectors using only image pairs labeled as either showing the same image scene or not, eliminating the need for depth or pose information. This significantly expands the pool of usable datasets to include diverse real-world scenarios, such as large-scale autonomous driving data (ACDC [30]) or place recognition data (Tokyo 24/7 [37]), including challenging weather and illumination changes.

Training on this broader set of datasets helps keypoint detectors become more robust to real-world conditions, especially for challenging long-term and dynamic localization tasks, as we show in our evaluation (Sec. 4.6 and Sec. 6.1).

However, using only binary labels substantially weakens training supervision. To overcome this and address the non-differentiable nature of the keypoint selection process, we introduce a probabilistic formulation for keypoint selection via Reinforcement Learning (RL). Unlike previous RL-based methods [4, 26, 38] that still require depth or pose information, we propose to derive the reward exclusively from labeled image pairs. Our key insight is to leverage the epipolar constraint, a fundamental concept in computer vision, that all *true* keypoint matches in a positive pair must satisfy. This eliminates the need for pre-generated 3D models and further broadens the range of suitable training datasets.

To efficiently associate every keypoint location with its descriptor, we incorporate intermediate multi-scale information by leveraging hyper-column features from intermediate layers of the encoder, rather than relying solely on the final low-resolution output. We further strengthen the descriptiveness of the descriptors by introducing a robust loss function explicitly designed for binary-labeled image pairs.

We evaluate RIPE on MegaDepth 1500 and HPatches, achieving competitive results compared to the state-of-the-art. On Aachen Day-Night and Boreas we demonstrate its increased robustness to adverse weather and illumination changes thanks to our minimal supervision strategy.

In summary, the key contributions of our work include:

- **Innovative Weakly-Supervised Training Framework:** Introduction of RIPE, a novel method based on Reinforcement Learning that trains keypoint detectors using only labeled image pairs, effectively removing the dependency on depth or pose information.
- **Improved Generalizability:** Ability to utilize diverse training datasets, improving keypoint detector performance in varying real-world conditions.
- **Epipolar Geometry-based reward:** Utilization of the epipolar constraint to derive rewards from labeled image pairs, ensuring that the optimization process adheres to fundamental principles of computer vision.
- **Multi-Scale Feature Representation:** Integration of multi-scale hyper-column features to enhance the association between keypoint locations and descriptors, leading to more informative and discriminative representations.
- **Robust Descriptor Loss:** Development of a robust loss function based on labeled image pairs, further strengthening the descriptiveness and reliability of the keypoint descriptors.

2. Related Work

Evolving from classical hand-crafted approaches, current training-based methods for learning keypoint detection and

description either rely on artificial augmentations or the availability of 3D information such as pose or depth, often derived from pre-generated 3D data. We systematically categorize current state-of-the-art methods based on their underlying training principles.

Artificial augmentations Methods in this category generate training data by applying artificial augmentations to existing images, creating image pairs with known transformations. This can be achieved by using photometric and/or homographic data augmentation (SuperPoint [6], DomainFeat [45], SiLK [13]). As these are limited in their representation of illumination changes, style-transfer techniques were introduced to synthesize night images (DomainFeat [45], Melekhov *et al.* [22]).

Additionally, different losses are used to ensure that feature maps are robust against domain changes. For instance, the triplet loss (DomainFeat [45]) enhances descriptor descriptiveness, while the minimization of margins for corresponding patches (HardNet [24]) or hard-negative sampling of local features (Melekhov *et al.* [22]) is also utilized.

Pose or depth information Methods of the second category depend on known pose or depth information, typically pre-generated by Structure-from-Motion (SfM) techniques. Many methods [7, 9, 12, 40, 41] utilize the work of Li and Snavely [18], leveraging the MegaDepth dataset (see Fig. 1 for two example images), which contains photos of landmark collections, automatically annotated with pixel-wise depth information. This dataset is constructed using COLMAP [34, 35] based on SIFT [21], to generate a 3D model along with pixel-wise depth information for each image.

Known depth information is used to calculate keypoint correspondences to train detection and description (Dusmanu *et al.* [7]) or to calculate reward values (DISK [38]). DeDoDe [9] introduced the direct use of 3D tracks originating from reconstructed points in the 3D model as a supervision signal. ALIKE [40] proposed to ground the learning of descriptors on relative pose information by introducing a differentiable matching layer and translating the relative poses into epipolar constraints. ALIKED [39] additionally introduced attention-weighted local descriptors to include image-level spatial awareness into the descriptor, training on pixel-wise correspondences enriched with random homography transformations. Recently, the dense (RoMa [10], Mast3r [16], Dust3r [42]) and semi-dense (S2DNet [12], LoFTR [36], Efficient LoFTR [43]) matching methods moved into focus. These methods also use the available depth information from the pre-generated 3D models.

Tyszkiewicz *et al.* [38] (DISK) and Bhowmik *et al.* [4] (Reinforced Feature Points) independently introduced Reinforcement Learning for learning keypoint detection and description to overcome the non-differentiability of keypoint detection. Bhowmik *et al.* trained with a complete com-

puter vision pipeline, treating the matching and pose estimation stages as non-differentiable black boxes, requiring known poses to calculate the reward value. In contrast, DISK computes rewards based on the number of correct feature matches (determined by the known relative position), allowing for precise calculation of matching probabilities. Potje *et al.* [26] (DEAL) extended the DISK approach by introducing an additional Warp Module, increasing robustness against non-rigid image deformations.

SOTA Limitations Ultimately, existing methods rely on known pixel-wise correspondences derived from artificial augmentations, depth information (measured or estimated with SfM) or relative pose. Consequently, these methods remain dependent on limited datasets or augmentation techniques. RL also remains underutilized, as depth (DISK [38]), pose (Reinforced Feature Points [4]) or artificial augmentations (DEAL [26]) are still required.

To this end, we introduce a more radical Reinforcement Learning approach, which allows us to train without known poses, without depth information, and without pixel-wise correspondences – just using paired images. This is comparable to semantic keypoint matching works like Rocco *et al.* [28], but differs in the usage of RL, the inclusion of negative pairs and the enforcement of a valid epipolar geometry.

3. Method

In the following, we present our approach to learning keypoint detection and description solely from unlabeled image pairs by using Reinforcement Learning (RL). A visual overview of our method is provided in Fig. 2. For each image in a given pair, a neural network generates a heatmap from which keypoint positions are sampled (Sec. 3.1). Each keypoint is then associated with a descriptor, extracted from the decoder using hyper-column features (Sec. 3.2). We process pairs of images and the resulting keypoints are matched and filtered by estimating the fundamental matrix. The final number of successfully matched keypoints is used as the reward signal, based on the label of the input pair. Using Reinforcement Learning (Eq. (4)), the reward encourages the network to generate a greater number of matchable keypoints – consistent with the epipolar constraint – and to produce fewer keypoints for negative pairs. This formulation effectively leverages the feedback from both geometric consistency and image similarity to guide the learning process.

Consequently, the only required label is binary, indicating whether the paired images depict the same scene (*i.e.* a sufficient number of 2D image keypoints correspond to projections of the same 3D point from the depicted scene) or not. Hence, we assume a dataset $\mathcal{D} = \{(I_\kappa, I'_\kappa, \lambda_\kappa) \mid \kappa = 1, \dots, N\}$ where each tuple contains two images and a

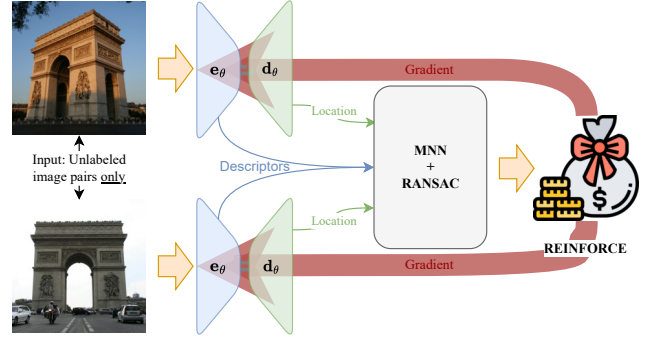


Figure 2. Overview of RIPE, our approach for learning keypoint detection and description from unlabeled image pairs using reinforcement learning. For an image pair, heatmaps are generated for probabilistic keypoint sampling, with descriptors derived from hyper-column features. These keypoints are matched and filtered via the fundamental matrix, with the number of matchable keypoints serving as the reward signal. This encourages the network to produce a large number of keypoints fulfilling the epipolar constraint for positive pairs and fewer keypoints for negative pairs.

binary label λ_κ with

$$\lambda_\kappa = \begin{cases} 1, & \text{if } I_\kappa \text{ and } I'_\kappa \text{ show the same scene} \\ -1, & \text{otherwise} \end{cases} \quad (1)$$

to indicate positive and negative pairs.

3.1. Keypoint detection

Our proposed method RIPE starts with detecting possible keypoint positions in each image, employing an hourglass network composed of an encoder $e_\theta(\cdot)$ and a decoder $d_\theta(\cdot)$ connected via skip connections with learnable parameters θ . A graphical overview is given in the upper part of Fig. 3. For an input image $\mathbf{I} \in \mathbb{R}^{h \times w \times c}$, the network generates a detection heatmap $\mathbf{H} \in \mathbb{R}^{h \times w}$ with $\mathbf{H} = d_\theta(e_\theta(\mathbf{I}))$, indicating potential keypoint locations. The heatmap \mathbf{H} is divided into a regular grid, with each cell \mathbf{c}_i having a size $m \times m$. The total number of cells is denoted by $C = \lfloor \frac{h}{m} \rfloor \times \lfloor \frac{w}{m} \rfloor$, with $\lfloor \cdot \rfloor$ denoting the floor function. Each cell $\mathbf{c}_i \in \mathbb{R}^{m \times m}$ holds the corresponding logit values. The logit values in each cell constitute a categorical probability distribution from which exactly one keypoint location per cell is sampled, resulting in a keypoint position s_i , with an initial probability \hat{p}_i and logit l_i . As a result, the network learns to define the keypoint locations by shaping the logit values accordingly. Acknowledging that some image regions (such as overexposed sections, sky, etc.) may not be optimal for reliable keypoint detection, a sigmoid function is applied to the keypoint logit forming an acceptance indicator $a_i = \sigma(l_i)$ for each cell. This enables the network to probabilistically discard unsuitable keypoints. Consequently, the final probability $p_i = \sigma(l_i) \cdot \hat{p}_i$ captures both the initial sampling chance and the likelihood of keypoint retention.

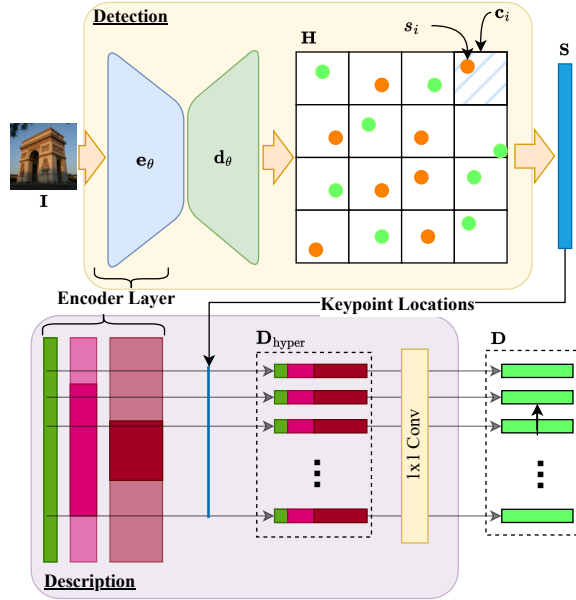


Figure 3. Keypoint detection (top) and description (bottom) for a single input image \mathbf{I} . The network outputs a logit heatmap \mathbf{H} for each input image. Potential keypoint locations \mathbf{S} are sampled from equally sized patches based on logit values, with rejected keypoints marked in orange and accepted ones in green. At the keypoint locations, descriptors \mathbf{D} are generated from the intermediate encoder layers, linking each location to a descriptor integrating context at different scales from the intermediate layers.

The keypoint detection results in a list $\mathbf{S}_{\mathbf{I}} \in \mathbb{R}_{2 \times C}$ of keypoint locations, acceptance indicators $\mathbf{a}_{\mathbf{I}} \in \mathbb{R}_{1 \times C}$, and associated probability values $\mathbf{p}_{\mathbf{I}} \in \mathbb{R}_{1 \times C}$ gathered from the individual cells of each keypoint in the image \mathbf{I} . The total number of keypoints is equivalent to the number of cells C .

3.2. Keypoint description

Once the keypoints are detected, the next step is to assign a descriptor to each location in order to enable the matching between images. This process is visualized in the lower part of Fig. 3. Potje *et al.* [26] demonstrated that features derived from the encoder and bilinearly upsampled to the input resolution retain a high distinctiveness under small to moderate photometric and geometric changes. However, this approach proved insufficient for our requirements, as it does not facilitate a reliable matching, as discussed in Sec. 6.2. To enhance (with minimal computational overhead) the quality of the features created simply by upsampling the final feature layer, we employ hyper-column features [14], a concept also proposed by Germain *et al.* [11] and tested in diverse settings (e.g. Li *et al.* [17], Min *et al.* [23]) to combine multi-scale information for increased descriptiveness.

In each layer l of the encoder, feature vectors \mathbf{e}_i^l at the detected keypoint locations $s_i \in \mathbf{S}$ are bilinearly interpolated (please refer to Fig. 3 for a visualization). This results in a list

($\mathbf{e}_i^1, \mathbf{e}_i^2, \dots, \mathbf{e}_i^L$) of feature vectors that encode information at different image scales. These feature vectors are then concatenated to form the intermediate hyper-column features $\mathbf{D}_{\text{hyper}} \in \mathbb{R}_{\hat{d} \times C}$, with \hat{d} being the overall sum of channels for the encoder layers. As the final dimensionality \hat{d} is usually large (960 for a VGG-19), a final 1×1 convolution is applied to reduce the dimensionality to d , resulting in compact descriptors $\mathbf{D} \in \mathbb{R}_{d \times C}$.

For a single image \mathbf{I} , the detection and description results in: keypoint locations $\mathbf{S} \in \mathbb{R}_{2 \times C}$ (in image coordinates), acceptance indicators $\mathbf{a} \in \mathbb{R}_{1 \times C}$, selection probabilities $\mathbf{p} \in \mathbb{R}_{1 \times C}$, and the descriptor map $\mathbf{D} \in \mathbb{R}_{d \times C}$. During training, we repeat the entire process to obtain \mathbf{S}' , \mathbf{a}' , \mathbf{p}' , and \mathbf{D}' for the second image \mathbf{I}' of a pair.

3.3. Reinforcement of matchable keypoints

To address the discrete nature of the keypoint detection process and to allow training from unlabeled image pairs only, we formulate it as a reinforcement problem. Starting from a brief general introduction, we develop the methodical foundation of RIPE.

The policy is defined as a probability distribution over actions \mathcal{A} , conditioned on the current state \mathcal{S} and parameterized by θ with

$$\pi_{\theta}(\mathcal{S}) = \mathbb{P}[\mathcal{A}|\mathcal{S}, \theta]. \quad (2)$$

This constitutes a probability distribution, from which an action \mathcal{A} is sampled. Based on the sampled action, the agent receives a reward signal that indicates a good or bad action. The learning objective is then formulated as maximizing the expected cumulative reward over a trajectory τ (a sequence of state, action, reward tuples) scaled by the reward R :

$$\max_{\theta} J(\theta) = \mathbb{E}_{x \sim \pi_{\theta}} [R(\tau)], \quad (3)$$

However, directly calculating the derivative of J is not feasible, as it would require differentiating through all possible trajectories and the state distribution. Nonetheless, REINFORCE [44] provides an approximation for the derivative:

$$\nabla_{\theta} J(\theta) \approx \hat{g} = \sum_{t=0} \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) R(\tau). \quad (4)$$

In RIPE, the encoder-decoder network acts as a trainable policy, with the input image \mathbf{I} representing the state. This leads to the following formulation for the policy:

$$\begin{aligned} \pi_{\theta}(\mathbf{s}) &= d_{\theta}(e_{\theta}(\mathbf{I})) = \mathbf{p} \\ &= \left[\mathbb{P}_1[a_1 | I, \theta], \mathbb{P}_2[a_2 | I, \theta], \dots, \mathbb{P}_c[a_c | I, \theta] \right], \end{aligned} \quad (5)$$

where \mathbf{p} is a list of distributions for each cell \mathbf{c} in the heatmap \mathbf{H} . As described in Sec. 3.1, keypoint locations are sampled from these distributions (i.e. the keypoint localization corresponds to an action).

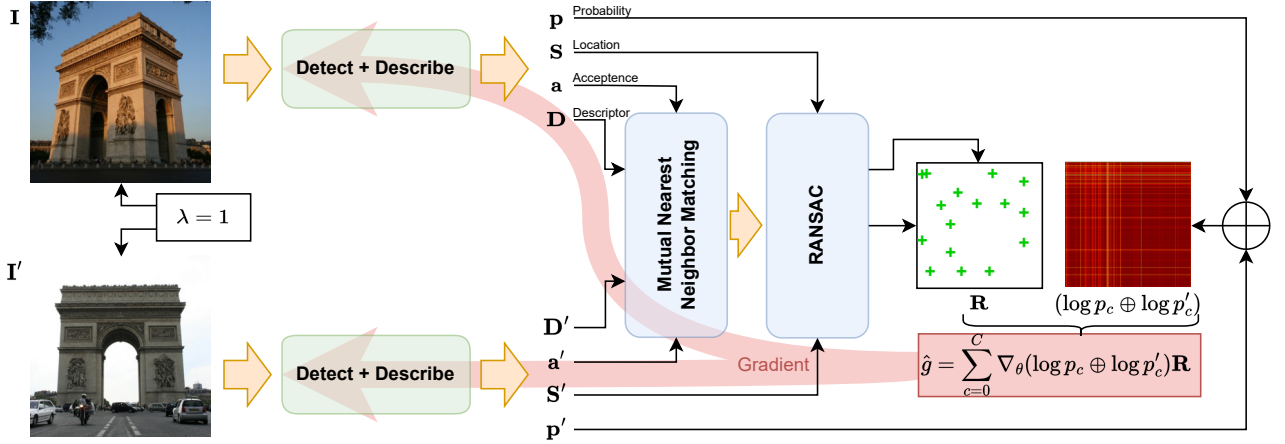


Figure 4. Overview of the proposed Reinforcement Learning formulation for learning keypoint detection and description from unlabeled image pairs. A network/ agent generates probability distributions over potential keypoint locations. Actions (keypoint locations \mathbf{S}) are sampled from these distributions and associated with their respective probability \mathbf{p} . The agent receives rewards based on the number of mutual nearest neighbors further filtered by estimating the fundamental matrix. This encourages the detection of matching keypoints in positive pairs ($\lambda = 1$) while penalizing incorrect detections in negative pairs ($\lambda = -1$). Using REINFORCE [44], gradients are derived and utilized to update the network parameters accordingly.

Working with image pairs, we introduce a second list of probabilities \mathbf{p}' and compute the joint probability across all combinations of cells. This allows us to approximate the gradient, as given by:

$$\hat{g} = \sum_{c=0}^C \nabla_{\theta} (\log p_c \oplus \log p'_c) R, \quad (6)$$

with \oplus denoting the outer sum¹.

The only ingredient missing is the reward R . As motivated in our introductory example, the goal is to produce many matching keypoints for positive pairs and few for a negative pairs. Consequently, the network should be rewarded for detecting keypoints in positive image pairs that are both matchable and pass the geometric filtering based on the epipolar constraint. Conversely, a negative reward (penalty) should be applied when the network incorrectly detects keypoints that appear matchable and filterable, yet originate from different scenes. The reward is computed using mutual nearest-neighbor estimation and RANSAC filtering, which, despite being non-differentiable, are used here solely for the reward calculation, and thus do not require gradient computation.

RIPE identifies mutual nearest neighbors between the keypoint descriptors in \mathbf{D} and \mathbf{D}' by computing their L2-distance, yielding a list of index pairs \mathbf{C} . This list is again filtered by estimating the fundamental matrix \mathbf{F} using the 8-point algorithm in combination with RANSAC. The resulting mask \mathbf{M} indicates all mutual nearest neighbors that can be explained by a common epipolar geometry.

¹With $x \in \mathcal{R}_{m \times 1}$ and $y \in \mathcal{R}_{m \times 1}$ the outer sum gets calculated as $x \oplus y \in \mathcal{R}_{m \times n}$ with $(x \oplus y)_{i,j} = x_i + y_j$

The reward matrix $\mathbf{R} \in \mathbb{R}^{C \times C}$ with the individual reward values $r_{i,j} \in \mathbf{R}$ is created by

$$r_{p,q} = \begin{cases} \text{sign}(\lambda_{\kappa})\rho, & (p,q) \in \mathbf{C} \text{ and } \mathbf{M}_{p,q} \text{ is True} \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

with the reward ρ and the signum function $\text{sign}(\cdot)$. Simply inverting the reward for negative pairs produces in a symmetric training signal, which we found to be beneficial during training.

The resulting gradients of Eq. (6) adjust the network parameters θ to increase the likelihood of selecting keypoints (i.e., actions taken) that result in a positive reward. Since the expectation is approximated by the mean, a substantial number of samples is required for a reliable estimate. In our approach, the sampling of C keypoint locations for every image acts as an equivalent of sampling multiple trajectories (as in classical reinforcement learning), and allows a sufficient approximation of the expectation.

3.4. Descriptor loss

To enhance the descriptiveness and robustness of the descriptors, we integrate a second loss

$$L_{\text{desc}} = \begin{cases} \frac{1}{N} \sum_{n=1}^N \max(0, \mu + \delta_+^n - \delta_h^n), & \lambda_{\kappa} = 1 \\ \frac{1}{N} \sum_{n=1}^N \max(0, \mu - \delta_+^n), & \text{otherwise} \end{cases} \quad (8)$$

where N is the number of inliers after the RANSAC filtering, μ is a positive threshold, and δ the L2-distance between

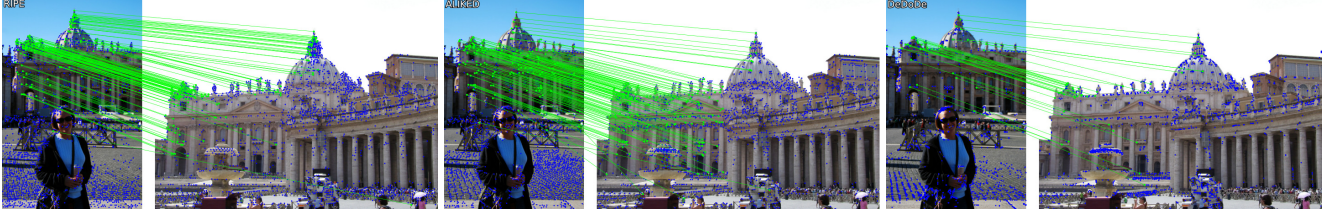


Figure 5. Example results for RIPE (ours), ALIKED [47], and DeDoDe [9] from the MegaDepth 1500 benchmark. RIPE demonstrates its ability to effectively utilize the unlabeled paired images to disregard irrelevant regions and identify highly discriminative keypoints. DeDoDe tends to cluster its keypoints in specific areas, whereas the keypoints of ALIKED and RIPE are more evenly distributed across the image.

two descriptors. This loss behaves differently depending on the type of image pair, with δ_{\pm}^n being the L2-distance between two keypoint descriptors, which are matched *and* were validated (i.e., identified as inliers) by geometric filtering, while δ_h^n is the distance to the second closest neighbor. In the positive case, this loss forces the descriptors of putative matches (putative, as we have no ground truth verification for belonging to the same 3D point) to become more similar while repelling all other descriptors. In contrast, in the negative case, where no matches should exist, the loss drives the descriptors of incorrectly matched keypoints further apart, thus increasing robustness by minimizing false positives.

3.5. Final loss

We define the final loss for the descriptor as $\mathcal{L}_{\text{desc}} = -\mathbb{E}_{\mathbf{K}}[\mathbf{R}]$. Inspired by the work of Potje *et al.* [26], we add a regularization term

$$\mathcal{L}_{\text{low}} = -\sum_p \sum_q \log p_i \cdot \epsilon, \quad (9)$$

with ϵ being a small negative constant to prevent the network from generating low probability keypoints. The final loss is calculated by

$$\mathcal{L} = \mathcal{L}_{\text{dect}} + \mathcal{L}_{\text{low}} + \psi \mathcal{L}_{\text{desc}}, \quad (10)$$

where ψ is a small positive constant, to balance detection and description loss.

By framing keypoint detection as an RL problem, RIPE can backpropagate the final loss to optimize both keypoint locations and descriptors, effectively tackling the challenge of learning without relying on extensive labeled datasets. The use of RL and formulating the detection process as a probabilistic decision also overcomes the inherent discreteness of the keypoint selection, enabling direct optimization. With this approach, we demonstrate how keypoints can be learned effectively using only unlabeled image pairs – mirroring the way humans would perform the task.

4. Experiments

4.1. Implementation details

We used the TORCHVISION implementation of VGG-19 with strides of [1, 2, 4, 8] and pretrained on ImageNet

Variable	Value	Purpose
h, w	560	image size
m	8	cell size
ρ	1.0	reward Eq. (7)
ϵ	-7e-8	constant regularization Eq. (9)
ψ	5.0	descriptor loss weight Eq. (10)

Table 1. Overview of our hyperparameter configuration. as our encoder. The feature maps on each scale have [64, 128, 256, 512] channels, respectively, resulting in $d' = 960$ (see Sec. 3.2), reduced to $d = 256$, by the 1x1 convolutional layer. For the decoder, we follow DeDoDe [9] and use the depthwise convolutional refiners proposed by Edstedt *et al.* in DKM [8], with 8 blocks per scale and internal dimension [64, 128, 256, 512]. We use the GPU-accelerated mutual nearest neighbor implementation from the Kornia library [27] and the fundamental matrix estimation from PoseLib [15]. The network is trained with AdamW [20] with a learning rate starting from 0.001 and linearly decaying to 1e-6. We train for 80,000 steps with a batch size of 6 and a gradient accumulation over 4 batches, taking three days on a single A100 GPU. Input images are normalized, resized and padded to meet the desired input resolution, without altering the aspect ratio. In addition, we set the hyperparameters according to Tab. 1. To facilitate the initial stages of training, we linearly increase the influence of ϵ during the first third of the training process.

4.2. Training data

We use the same Megadepth subset as [38], reducing the dataset to image pairs which show a sufficient number of covisible 3D points. However, our innovative training regime enables the use of an extended data basis. In Sec. 4.6 and Sec. 6.1 we therefore demonstrate how this can improve the robustness of our method.

4.3. Inference

During inference, we pass a single image through the network and sample the top K most likely keypoints and their positions from \mathbf{H} , accompanied with non-maximum suppression in a 3x3 window. Each keypoint is associated with its hyper-column descriptor, as described in Sec. 3.2. The mean inference time (measure for the setting of Sec. 4.4 was

Category	Method	MegaDepth 1500			Label	HPatches		
		AUC@5°	AUC@10°	AUC@20°		AUC@1px	AUC@3px	AUC@5px
(Semi-)Dense	Eff. LoFTR[43] _{CVPR'24}	64.63	78.03	87.21	D	42.38	63.26	72.75
	Mast3r[16] _{ECCV'24}	22.84	35.29	47.62	D	29.63	54.04	66.86
	RoMa[10] _{CVPR'24}	68.28	80.53	88.77	D	38.38	63.75	73.37
Sparse	SIFT[21] _{IJCV'04}	37.29	50.61	61.89	—	29.98	53.93	65.01
	SuperPoint[6] _{CVPRW'18}	49.5	62.51	72.12	H	36.61	56.92	67.28
	DISK[38] _{NeurIPS'20}	51.83	64.49	74.3	P/D	37.54	57.68	68.08
	ALIKED[47] _{TIM'23}	56.71	69.86	79.54	P+H	31.95	57.65	<u>69.35</u>
	SiLK[13] _{ICCV'23}	29.83	40.59	49.98	H	<u>39.05</u>	57.96	67.64
	DeDoDe-B[9] _{3DV'24}	55.02	67.99	77.31	P	39.88	60.59	70.36
	RIPE	<u>55.11</u>	<u>68.34</u>	<u>78.03</u>	Pair	37.93	<u>58.93</u>	69.20

Table 2. Results for the evaluation of relative pose estimation on MegaDepth-1500 (left) and homography estimation on HPatches (right). We specify the training data for each approach, with D indicating depth, P position and H homography/ artificial augmentations. Crucially, our method is the only one that requires only unlabeled image pairs, yet achieves competitive performance to the sparse SotA methods. The **best** and second-best performances for the sparse methods are highlighted.

0.47 s for RIPE (for comparison: DeDoDe 0.45 s, ALIKED 0.08 s, DISK 0.17 s).

4.4. Relative pose estimation

Dataset We evaluated the relative pose estimation performance using the MegaDepth-1500 subset. It contains two (*Brandenburger Tor* and *St. Peters Square*) out of the 196 scenes from the original MegaDepth dataset and was introduced in LoFTR [36]. The main challenges are large viewpoint, illumination changes and repetitive patterns. Following recent evaluation protocols ([43], [19]), we resized the longer side of the input images to 1600 for the dense methods and to 1200 for the (semi-)sparse methods. To assess the quality of the extracted keypoints, we use mutual-nearest-neighbor matching for all sparse methods.

Baselines We compare RIPE against SotA methods for sparse keypoint detection and description. Due to their recent success, we also integrate dense (RoMA [10], Mast3r [16]) and semi-dense matching methods for comparison (Efficient LoFTR [43]). For the sparse methods, we use mutual-nearest-neighbor matching to establish correspondences.

Metrics Building on preceding approaches, the accuracy of matches is assessed by evaluating the relative poses they yield. The pose error is characterized as the greatest of the angular discrepancies in both rotation and translation. We provide the Area Under the Curve (AUC) of the pose error at the thresholds of 5°, 10°, and 20°. We used the glue-factory library, kindly provided by the authors of GlueStick [25] and LightGlue [19]. We use the poselib [15] for the robust pose estimation and select the top 2048 keypoints.

Results The left half of Tab. 2 and demonstrate that our method, RIPE, ranks a close second to the current state-of-the-art sparse method, ALIKED, while not requiring images with known poses during training and leveraging a significantly weaker training signal. Remarkably, RIPE even out-

performs DeDoDe, despite the latter employing two distinct networks for detection and description, which nearly doubles the number of required parameters. Furthermore, our method is the first to achieve this performance without relying on artificial homographies (as seen in methods like SuperPoint and SiLK) or requiring pose or depth information from a pre-registered 3D model. Qualitative results can be found in Fig. 5 and in the supplementary (Fig. 6).

4.5. Homography Estimation

Dataset We evaluate on the HPatches dataset [2], which contains sequences of planar scenes, taken with viewpoint or illumination changes. We resized the input smaller side of the input images to 480 pixels.

Metrics To assess the quality of the homography estimation we calculate the mean reprojection error of the corner points and report the AUC for the thresholds of 1, 3 and 5 pixels. For all methods, we use mutual-nearest-neighbor matching and the implementation of poselib [15] for the robust homography estimation and restricted the number of keypoint to 1024.

Results As illustrated on the right side of Tab. 2, RIPE once again performs on par with the state-of-the-art methods SiLK and DeDoDe, despite not utilizing artificial homographies like SiLK or pose information as employed by DeDoDe. The strong performance of SiLK highlights the advantages of its training regimen, which relies solely on artificial augmentations and therefore closely resembles the testing data. Conversely, the impressive performance of ALIKED diminishes on the HPatches benchmark, where it only narrowly surpasses SIFT.

4.6. Outdoor localization day-night

Dataset To evaluate our approach in the context of visual localization (*i.e.* the estimation of the 6-DoF pose for

Method	Day			Night		
	.25m/2°	.5m/5°	5m/10°	.25m/2°	.5m/5°	5m/10°
ALIKED	87.3	93.9	97.3	73.3	88.0	96.9
DeDoDe	82.2	89.0	92.6	47.1	56.5	64.4
SIFT	82.5	88.5	91.9	30.9	38.2	46.1
RIPE	<u>81.6</u>	<u>89.2</u>	<u>93.1</u>	52.9	67.5	79.1
↓ + Tokyo	+0.0	-0.2	-0.7	+7.3	+5.3	+4.7
RIPE	<u>81.6</u>	89.0	92.4	<u>60.2</u>	<u>72.8</u>	<u>83.8</u>

Table 3. Outdoor visual localization on the Aachen Day-Night v1.1. The results emphasize the significance of available training data: ALIKED outperforms DeDoDe by incorporating images from the Aachen dataset. Additionally, RIPE demonstrates substantial improvements by including day-night training pairs from Tokyo 24/7. The **best** and second-best performances are highlighted.

a query image relative to a 3D scene model), we use the Aachen v1.1 [32, 33, 46] dataset. This dataset is especially challenging because of its large viewpoint and illumination (day-night) changes. We used the HLoc localization framework [31] for the evaluation and first triangulate a 3D model from the 6,697 reference images. For each of the 1015 (824 daytime, 191 nighttime) queries we retrieve 50 images using NetVLAD [1] and match them. To evaluate the influence of additional training data, we replaced 20% of the training samples with images from the Tokyo 24/7 *et al.* [37] query dataset, originally intended for place recognition. This dataset contains images from 125 distinct locations. Images were captured at each location from three different viewing directions, across day, dusk, and night, resulting in a total of 1,125 images. The images are paired based on their geo-position only. No 3D positions or depth maps are available.

Metrics The camera pose is estimated with a Perspective-n-Point solver in conjunction with RANSAC and the AUC is reported for thresholds 0.25m/2°, 0.5m/5° and 1.0m/10°.

Results Our results in Tab. 3 highlight the importance of diverse training data for the robustness of trained keypoint extractors. As ALIKED was not only trained on MegaDepth, but also on synthetic and training images from the Aachen[33] dataset, it shows strong performance for the localization of night-time queries. The results from DeDeDo demonstrate the influence of the limited training data (MegaDepth only). RIPE clearly outperforms DeDoDe, even if only trained on MegaDepth. When day-to-nighttime images are added to training data, the margin further increases. This improvement is made possible by our innovative training regime, which facilitates the addition of diverse training data.

Additional experiments presented in the supplementary material (Sec. 6.1) further support these findings, with results obtained from the Boreas [5] dataset, which encompasses challenging weather conditions.

4.7. Evaluation dataset Composition

MegaDepth Tokyo	Day			Night			
	0.25m/2	0.5m/5	5m/10	0.25m/2	0.5m/5	5m/10	
1.0	0.0	81.6	89.2	93.1	52.9	67.5	79.1
0.9	0.1	81.3	87.5	93.2	57.1	68.1	82.7
0.8	0.2	81.6	89.0	92.4	60.2	72.8	83.8
0.7	0.3	80.6	88.1	93	56.5	71.7	82.2
0.6	0.4	79.4	86.0	92.1	58.1	70.7	85.3
0.5	0.5	78.0	84.2	89.4	56.5	68.6	83.2
0.4	0.6	71.8	81.2	87	54.5	70.2	83.8

Table 4. Evaluation on how training data from the Tokyo 24/7 dataset improves the ability of RIPE to handle day to night illumination changes. The **best** performances are highlighted.

We further investigated which data mix between MegaDepth and Tokyo 24/7 yields the best improvement on the Aachen Day-Night benchmark. We trained RIPE with different compositions, as presented in Tab. 4, and found that 80% MegaDepth data and 20% Tokyo data were the most advantageous. A further increase in the proportion of Tokyo data begins to degrade the results, likely due to the lack of viewpoint variability in this dataset, as the images primarily differ in illumination. Consequently, RIPE struggles to learn to cope with the strong viewpoint variations that accompany the day-to-night changes in the Aachen benchmark.

Additional experiments regarding the influence of our hyperparameters can be found in the supplementary (Sec. 6.2).

5. Conclusion

We present a fundamentally new approach to learning keypoint detection and description. By integrating Reinforcement Learning with essential computer vision principles, we successfully train RIPE using only unlabeled image pairs, expanding the pool of available training data. This eliminates the constraints imposed by traditional approaches that depend on precise geometric annotations, making our method scalable and adaptable to diverse real-world scenarios.

Despite leveraging a significantly weaker training signal, RIPE achieves performance on par with state-of-the-art sparse keypoint extractors. Furthermore, our results demonstrate that RIPE effectively benefits from the inclusion of diverse training data, improving its generalization capabilities and robustness to challenging conditions. This highlights the potential of our approach to redefine keypoint learning, enabling broader applicability across various domains.

Acknowledgments We thank our colleague, Wieland Morgenstern, for his valuable feedback on the manuscript. This work was partly funded by the Federal Ministry of Education and Research (RE-FRAME, grant no. 01IS2407A) and the German Federal Ministry for Economic Affairs and Climate Action (DeepTrain, grant no. 19S23005D).

References

- [1] Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5297–5307, 2016. 8
- [2] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. HPatches: A Benchmark and Evaluation of Handcrafted and Learned Local Descriptors. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3852–3861, Los Alamitos, CA, USA, 2017. IEEE Computer Society. 7
- [3] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded Up Robust Features. *ECCV 2006, 9th European Conference on Computer Vision*, pages 404–417, 2006. 1
- [4] Aritra Bhowmik, Stefan Gumhold, Carsten Rother, and Eric Brachmann. Reinforced Feature Points: Optimizing Feature Detection and Description for a High-Level Task. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 00:4947–4956, 2020. 2, 3
- [5] Keenan Burnett, David J Yoon, Yuchen Wu, Andrew Z Li, Haowei Zhang, Shichen Lu, Jingxing Qian, Wei-Kang Tseng, Andrew Lambert, Keith YK Leung, Angela P Schoellig, and Timothy D Barfoot. Boreas: A multi-season autonomous driving dataset. *The International Journal of Robotics Research*, 42(1-2):33–42, 2023. 8
- [6] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-Supervised Interest Point Detection and Description. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 337–349, 2018. 1, 2
- [7] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-Net: A Trainable CNN for Joint Description and Detection of Local Features. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8084–8093, 2019. 2
- [8] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. DKM: Dense kernelized feature matching for geometry estimation. *arXiv*, 2022. 6
- [9] Johan Edstedt, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. DeDoDe: Detect, Don’t Describe — Describe, Don’t Detect for Local Feature Matching. *2024 International Conference on 3D Vision (3DV)*, pages 148–157, 2024. 1, 2, 6
- [10] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust Dense Feature Matching. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 00:19790–19800, 2024. 2, 7
- [11] Hugo Germain, Guillaume Bourmaud, and Vincent Lepetit. Sparse-to-Dense Hypercolumn Matching for Long-Term Visual Localization. *2019 International Conference on 3D Vision (3DV)*, 00:513–523, 2019. 4
- [12] Hugo Germain, Guillaume Bourmaud, and Vincent Lepetit. S2dnet: Learning image features for accurate sparse-to-dense matching. In *Computer Vision – ECCV 2020*, pages 626–643, Cham, 2020. Springer International Publishing. 2
- [13] Pierre Gleize, Weiyao Wang, and Matt Feiszli. SiLK: Simple Learned Keypoints. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22442–22451, 2023. 1, 2
- [14] Bharath Hariharan, Pablo Arbelaez, Ross Girshick, and Jitendra Malik. Hypercolumns for Object Segmentation and Fine-Grained Localization. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 447–456, 2015. 4
- [15] Viktor Larsson and contributors. PoseLib - Minimal Solvers for Camera Pose Estimation, 2020. 6, 7
- [16] Vincent Leroy, Johann Cabon, and Jerome Revaud. Grounding Image Matching in 3D with MAST3R. In *European Conference on Computer Vision*, pages 71–91, 2024. 2, 7
- [17] Wanhua Li, Yueqi Duan, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Learning to Compose Hypercolumns for Visual Correspondence. *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV*, pages 346–363, 2020. 4
- [18] Zhengqi Li and Noah Snavely. MegaDepth: Learning Single-View Depth Prediction from Internet Photos. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2041–2050, 2018. 1, 2
- [19] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. In *International Conference on Computer Vision (ICCV)*, 2023. 7
- [20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. 6
- [21] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 1, 2
- [22] Iaroslav Melekhov, Zakaria Laskar, Xiaotian Li, Shuzhe Wang, and Juho Kannala. Digging Into Self-Supervised Learning of Feature Descriptors. In *2021 International Conference on 3D Vision (3DV)*, pages 1144–1155, Los Alamitos, CA, USA, 2021. IEEE Computer Society. 2
- [23] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Hyperpixel Flow: Semantic Correspondence with Multi-layer Neural Features. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 00:3394–3403, 2019. 4
- [24] Anastasiya Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor’s margins: Local descriptor learning loss. In *Neural Information Processing Systems*, 2017. 2
- [25] Rémi Pautrat*, Iago Suárez*, Yifan Yu, Marc Pollefeys, and Viktor Larsson. GlueStick: Robust Image Matching by Sticking Points and Lines Together. In *International Conference on Computer Vision (ICCV)*, 2023. 7
- [26] Guilherme Potje, Felipe Cadar, Andre Araujo, Renato Martins, and Erickson R. Nascimento. Enhancing Deformable Local Features by Jointly Learning to Detect and Describe Keypoints. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1306–1315, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 2, 3, 4, 6

- [27] E. Riba, D. Mishkin, D. Ponsa, E. Rublee, and G. Bradski. Kornia: an open source differentiable computer vision library for pytorch. In *Winter Conference on Applications of Computer Vision*, 2020. 6
- [28] Ignacio Rocco, Relja Arandjelović, and Josef Sivic. End-to-end weakly-supervised semantic alignment. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6917–6925, 2018. 3
- [29] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: an efficient alternative to SIFT or SURF. *2011 International Conference on Computer Vision*, pages 2564–2571, 2011. 1
- [30] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: The Adverse Conditions Dataset with Correspondences for Semantic Driving Scene Understanding. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 00:10745–10755, 2021. 1
- [31] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12708–12717, Los Alamitos, CA, USA, 2019. IEEE Computer Society. 8
- [32] Torsten Sattler, Tobias Weyand, Bastian Leibe, and Leif Kobbelt. Image Retrieval for Image-Based Localization Revisited. In *British Machine Vision Conference (BMVC)*, 2012. 8
- [33] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Fredrik Kahl, and Tomas Pajdla. Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 8
- [34] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2
- [35] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 2
- [36] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiao-wei Zhou. LoFTR: Detector-Free Local Feature Matching with Transformers. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 00:8918–8927, 2021. 2, 7
- [37] Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 Place Recognition by View Synthesis. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1808–1817, 2015. 1, 8
- [38] Michal Tyszkiewicz, Pascal Fua, and Eduard Trulls. DISK: Learning local features with policy gradient. In *Advances in Neural Information Processing Systems*, pages 14254–14265. Curran Associates, Inc., 2020. 1, 2, 3, 6
- [39] Changwei Wang, Rongtao Xu, Ke Lu, Shibiao Xu, Weiliang Meng, Yuyang Zhang, Bin Fan, and Xiaopeng Zhang. Attention Weighted Local Descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10632–10649, 2023. 2
- [40] Qianqian Wang, Zhou Xiaowei, Bharath Hariharan, and Noah Snavely. Learning Feature Descriptors Using Camera Pose Supervision. In *Computer Vision - ECCV 2020*, pages 757–774. Springer International Publishing, 2020. 2
- [41] Qianqian Wang, Xiaowei Zhou, Bharath Hariharan, and Noah Snavely. Learning feature descriptors using camera pose supervision. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, page 757–774, Berlin, Heidelberg, 2020. Springer-Verlag. 2
- [42] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUS3R: Geometric 3D Vision Made Easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697–20709, Los Alamitos, CA, USA, 2024. IEEE Computer Society. 2
- [43] Yifan Wang, Xingyi He, Sida Peng, Dongli Tan, and Xiaowei Zhou. Efficient LoFTR: Semi-Dense Local Feature Matching with Sparse-Like Speed. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 00:21666–21675, 2024. 2, 7
- [44] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256, 1992. 4, 5
- [45] Rongtao Xu, Changwei Wang, Shibiao Xu, Weiliang Meng, Yuyang Zhang, Bin Fan, and Xiaopeng Zhang. DomainFeat: Learning Local Features With Domain Adaptation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(1):46–59, 2024. 2
- [46] Zichao Zhang, Torsten Sattler, and Davide Scaramuzza. Reference pose generation for long-term visual localization via learned features and view synthesis. *Int. J. Comput. Vision*, 129(4):821–844, 2021. 8
- [47] 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. 1, 6