Scale-Free Image Keypoints Using Differentiable Persistent Homology

Giovanni Barbarani [1] **Francesco Vaccarino** [1] **Gabriele Trivigno** [2] **Marco Guerra** [3] **Gabriele Berton** [2] **Carlo Masone** [2]

Abstract

In computer vision, keypoint detection is a fundamental task, with applications spanning from robotics to image retrieval; however, existing learning-based methods suffer from scale dependency, and lack flexibility. This paper introduces a novel approach that leverages Morse theory and persistent homology, powerful tools rooted in algebraic topology. We propose a novel loss function based on the recent introduction of a notion of subgradient in persistent homology, paving the way toward topological learning. Our detector, MorseDet, is the first topology-based learning model for feature detection, which achieves competitive performance in keypoint repeatability and introduces a principled and theoretically robust approach to the problem.

1. Introduction

The ability to extract salient points (*keypoints*) and associated features from an image is a cornerstone of computer vision, as it underpins several applications such as visual localization (Sarlin et al., 2019; Sattler et al., 2018; Toft et al., 2020), SLAM (Mur-Artal et al., 2015; Durrant-Whyte & Bailey, 2006; Bailey & Durrant-Whyte, 2006), Structure-from-Motion and 3D reconstruction (Schönberger & Frahm, 2016; Heinly et al., 2015; Schönberger et al., 2016), as well as retrieval and place recognition (Barbarani et al., 2023; Noh et al., 2017). Traditional pipelines relied on handcrafted filters that were engineered to detect salient points such as corners (Harris & Stephens, 1988), blobs (Tuytelaars & Van Gool, 2000; Lowe, 2004; Mikolajczyk & Schmid, 2004) or edges (Bhardwaj & Mittal, 2012). These keypoints would then be associated with a feature vector obtained typically from local derivatives of the image (Calonder et al., 2010; Bay et al., 2006; Lowe, 2004).

Ideally, a good feature detector should provide keypoints with the following desirable properties: high repeatability (*i.e.*, consistent across image pairs) and scale-invariance, while being robust to noise and distortion (Ghahremani et al., 2020; Revaud et al., 2019; Lowe, 2004). Scale-Space theory (Lindeberg, 1994) provides a formulation of the concept of keypoint that guarantees the properties mentioned above (Lindeberg, 1994; Lowe, 2004; Ghahremani et al., 2020), and it operates by building a scale-space feature pyramid from the image, in which keypoints are detected as local extrema. Many classical handcrafted detectors exploit this theoretical framework (Mikolajczyk & Schmid, 2004; Bay et al., 2006), the most popular of which is SIFT (Lowe, 2004).

Recently, several learning-based detectors have been introduced, which, in the spirit of deep learning, propose to forego the formal definition of keypoints and rely on a data-driven approach to teach a neural network how to select salient points (Yi et al., 2016; Savinov et al., 2017; Tian et al., 2017; Revaud et al., 2019). Despite their learning-based approach, these methods are still bound to several design choices: for reliability, keypoints are often defined as locations that are easily matched (DISK) or that are more discriminative (R2D2); and for repeatability, they are usually defined as local maxima, in a local patch of arbitrary size.

This formulation presents some inherent flaws due to the coarse heuristic used to model the maxima (*i.e.*, patch-wise) being inflexible. More specifically it (i) does not guarantee that chosen points are critical points of the feature maps, (ii) leads to keypoints whose density is hard-coded in a hyperparameter, which (iii) makes the keypoints inherently scale-dependent. Whereas the issue of scale dependency is often approached through multi-scale inference (*i.e.*, processing the same image multiple times at different resolutions). We note that a mathematical framework that is able to model and thus locate local maxima with inherent guarantees of scale independence is currently missing in the literature.

To this end, we propose a novel and differentiable formulation of keypoints based on Morse theory (Milnor, 1963) and

[1]Department of Mathematical Sciences "Giuseppe Luigi Lagrange", Politecnico di Torino, Italy [2]Department of Control and Computer Engineering, Politecnico di Torino, Italy [3]Institut Fourier, Université Grenoble Alpes, France. Correspondence to: Giovanni Barbarani <giovanni.barbarani@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

persistent homology (Edelsbrunner et al., 2002; Zomorodian & Carlsson, 2004) from algebraic topology. This formulation leverages the connection between local maxima and differentiable topological invariants (Carriere et al., 2021; Leygonie et al., 2021) offering a more robust and elegant solution, thus enabling a genuinely unsupervised framework for keypoint detection, thus without requiring hardcoded hyperparameters that determine the density or frequency of keypoints. We believe this work can lay a foundation for future research to expand our proposed framework for integrating topology in computer vision. Our implementation and trained models have been publicly released[1].

To summarize, our contributions are the following:

- We show a connection between topological data analysis (TDA) and keypoint detection. Our method, MorseDet, is the first learnable method based entirely on a TDA framework;
- The first loss based on persistent homology for unsupervised learning, capable of modeling a set of features arbitrary in cardinality and shape;
- We demonstrate that our detector, thanks to its strong theoretical foundation, shows promising performance in terms of keypoints repeatability.

2. Related Work

Feature detection. The task of local feature detection and description has witnessed a growing interest in the past years due to its central role in several vision-related applications (Csurka et al., 2018; Salahat & Qasaimeh, 2017). Traditionally, the detection step was performed independently via handcrafted methods based on local derivatives such as (Lowe, 2004; Harris & Stephens, 1988). Early attempts to train learnable detectors also adopted this decoupled approach (Mishkin et al., 2018; Barroso-Laguna et al., 2019). The common denominator of these early works is that they all rely on a supervised notion of keypoint, which is problematic because it is limited by the prior knowledge that the researchers have on what should be considered a keypoint, either via handcrafted detectors or via learnable filters that mimic the handcrafted ones. SuperPoint (DeTone et al., 2018) was the first to implement a semi-supervised approach. Starting from synthetic shapes with user-defined keypoints, it uses homographic adaptation to enhance the equivariance of the detected keypoints to such transformation. This approach is still limited because keypoints are still arbitrarily defined as intersections of known structures.

D2-Net (Dusmanu et al., 2019) and R2D2 (Revaud et al., 2019) proposed a joint detection and description pipeline, arguing that keypoint repeatability cannot be achieved via supervised training, which would result in mimicking a previously available detector. They are directly inspired by Scale-Space theory (Lindeberg, 1994) and model keypoints as local maxima of their output maps. More recent self-supervised approaches, namely SiLK (Gleize et al., 2023) and DISK (Tyszkiewicz et al., 2020), avoid the problem of defining keypoints, and promote the detection of points that are easy to match and are based on probabilistic loss formulations. ALIKE (Zhao et al., 2023b) proposes a patch-wise softmax relaxation of keypoints at training time. ALIKED (Zhao et al., 2023a) further improves this approach by relying on deformable convolutional filters capable of adapting to the keypoints support.

These recent works demonstrated the effectiveness of a learnable approach based on detecting local maxima from the response map of a feature extractor. Nevertheless, they lack an analytical tool to locate these local extrema reliably and rely typically on a softmax-based approximation inside local patches of predefined size, thus bounding the detection frequency to this hyperparameter and fails to achieve scale invariance. We propose a formulation based on the persistent homology theory that fills this gap in the literature.

Topological data analysis. Topological data analysis (TDA) encompasses several data analysis techniques that employ algebraic topology. These include persistent homology (Edelsbrunner et al., 2002), which is often utilized to provide a multi-scale description of point clouds, and discrete Morse theory (Delgado-Friedrichs et al., 2014; Robins et al., 2011), used extensively for processing 2D and 3D images. Persistence-based statistics have proven useful in various domains, such as clustering (Chazal et al., 2013), robust pose estimation (Dey et al., 2010), and as input features for neural networks (Hofer et al., 2017; Giansiracusa et al., 2019). Although in a non-learnable context, the Morse theory and TDA have been applied to image matching (Matas et al., 2004; da Silva et al., 2012; Xu et al., 2014). From a theoretical standpoint, the differentiability of persistence-based functions has been extensively explored (Leygonie et al., 2021; Carriere et al., 2021). Recently, differentiable topological objectives have been integrated into deep learning, serving as regularization tools for shaping the topology of decision boundaries in classification tasks (Chen et al., 2019) and as priors for the latent space in autoencoders (Moor et al., 2020). In supervised learning contexts, persistent homology, along with discrete Morse theory, has been applied to image segmentation tasks to enhance the loss function (Hu et al., 2023; 2021; Clough et al., 2020), often complemented by the binary cross-entropy loss, and to quantify uncertainty (Gupta et al., 2024).

Our work represents a first attempt at introducing completely unsupervised objectives based on persistent homology within the context of deep learning for data-intensive applications.

[1] https://github.com/gbarbarani/MorseDet

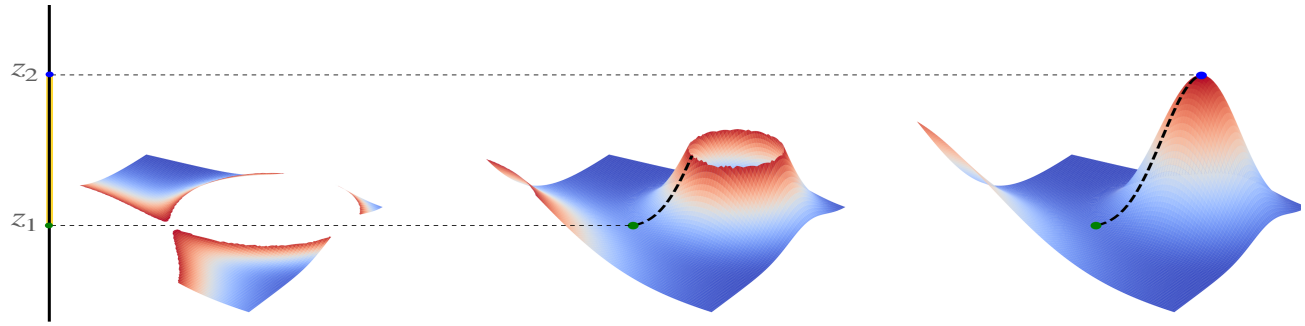

Figure 1: The evolution of the sub-level sets of a surface filtered by height, *i.e.* value on the z axis. As the height crosses z_1, a new loop is born in correspondence with a saddle (green point), then the loop changes smoothly until z hits z_2, the value of a corresponding maximum (blue point), and the loop disappears. z_1 and z_2 are respectively the birth time and the death time of the topological feature.

3. Background

The relationship between critical points of a function (extrema and saddle points) and the evolution of a topology can be intuitively explained using the following analogy: picture the graph of a 2D scalar function as a landscape. When we flood this landscape, we witness a series of transformations: lakes emerge from the lowest valley regions; lakes surround mountains, leaving only their peaks above water, and, ultimately, the lakes blend when they submerge the peaks.

Morse theory (Milnor, 1963) is the mathematical framework that precisely captures the relationship between critical points and changes in topology, while persistent homology (Edelsbrunner et al., 2002) is the algebraic topology tool that defines the computation of the filtration process described above. Finally, discrete Morse theory (Robins et al., 2011) is the tool that is commonly used to implement these ideas in the context of digital 2D and 3D images. We will briefly review these concepts in the context of our scope.

3.1. Morse Theory

A smooth scalar function h defined on a smooth manifold is a Morse function if it has only non-degenerate critical points, i.e., having non-zero Hessian determinants only. This condition is not restrictive: indeed, up to an infinitesimal perturbation, every differentiable function on a compact is Morse. Given a 2D compact surface $\mathcal{X}$ and the choice of a Morse function h, we can study the evolution of the sub-level sets $\mathcal{X}_t = \{x \in \mathcal{X} : h(x) \leq t\}$ for an increasing t. These sets can be considered the union of the bottom of the lakes obtained by pouring water onto our landscape up to level t. When t reaches the value corresponding to a minimum of h, the sub-level changes by adding a new point: a new connected component (lake) is born. When t reaches a saddle point $s = (p, t)$ with $t = h(p)$, two things could happen: (i) the saddle s merges two lakes into one, or (ii) the saddle s creates a single span bridge over a lake, thus producing a new closed path (loop) in the component. Therefore, a saddle either reduces connected components or creates a loop. Finally, when t reaches a maximum value, it corresponds to completely submerging the terrain and its closed paths, and this can be seen as filling the hole surrounded by a closed path.

3.2. Persistent Homology

The homology modules H_0 and H_1 of $\mathcal{X}$ are vector spaces, whose dimensions count the number of connected components and the number of loops of $\mathcal{X}$, respectively. The sub-level filtration $\mathcal{X}_{t_1} \subset \mathcal{X}_{t_2} \subset ... \subset \mathcal{X}_{t_n}$ associated to a function h in an obvious way, induces a sequence of morphisms

$$H_i(\mathcal{X}_{t_1}) \to H_i(\mathcal{X}_{t_2}) \to ... \to H_i(\mathcal{X}_{t_1})$$

The persistent homology module $\mathbb{H}_i(\mathcal{X}, \mathcal{F}) := \bigoplus_t H_i(\mathcal{X}_t)$ of $\mathcal{X}$ keeps track via the morphisms above of the evolution along the sub-level filtration of the topological features associated to the critical points of the function h. Each generator of $e \in \mathbb{H}_i$ (as a graded module over the ring of polynomials, see app. A) is associated with a pair of values $(b(e), d(e))$, $b(e) < d(e)$, the birth time and the death time of e, representing the life span of a topological feature. From sec. 3.1 it should be apparent that, if $e \in \mathbb{H}_0$, $b(e)$ corresponds to a minimum of h and $d(e)$ must be a saddle. On the other hand, if $e \in \mathbb{H}_1$, then $b(e)$ corresponds to a saddle point and $d(e)$ to a maximum; an example of this case is depicted in fig. 1.

3.3. Discrete Morse Theory

To study the topology of the local extreme values of a digital image I, the cubical complex $\mathcal{K}$ and the filtration associated

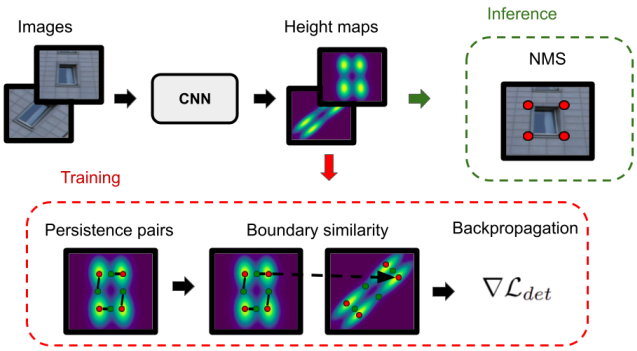

Figure 2: **Pipeline overview**: a convolutional neural network (CNN) is employed to generate height maps from input images. During inference, keypoints are efficiently detected as local maxima of these maps, utilizing non-maximum suppression. For training, our detector loss computation involves the application of discrete Morse theory algorithms to compute persistence pairs, then the maps at the corresponding positions are compared through the boundary similarity. The resulting gradients are subsequently backpropagated.

to I is used (see app. A.1). Discrete Morse theory is an efficient framework to compute the birth and death times of generators of the associated modules $\mathbb{H}_0$ and $\mathbb{H}_1$ through simulated differentiation. Also in this case the critical times are, up to a perturbation, unequivocally associated with the local extreme values of I.

4. Methodology

Overview. We propose a novel keypoint detector, which we call **MorseDet**, whose goal is to detect a set of sparse keypoints that should be repeatable across different transformations of an image as well as scale-invariant. The core novelty of our method is to utilize elements of topological data analysis, namely the persistent homology framework, to model keypoints as local maxima of the feature maps in a CNN in a differentiable way. Our approach differentiates from previous works (Dusmanu et al., 2019; Revaud et al., 2019; Zhao et al., 2023a) where local maxima are computed heuristically within a fixed $N \times N$ patch, imposing a strong prior on the frequency of the detected keypoints, which ultimately makes the detection scale-dependent. In MorseDet, we have re-envisioned these concepts with a scale-independent topological characterization of local maxima. We show that thanks to the theoretical guarantees that underpin our method, we outperform previous sparse detectors in terms of repeatability and scale invariance. Given the novelty of our topology-based approach, in this work, we focus only on the detection step, leaving for future works the integration of a topological descriptor loss.

Problem setting. We aim to learn a feature extractor F_θ that, given an input image $I \in \mathbb{R}^{H \times W \times C}$, outputs a set of discrete pixel locations $\{k_i\} \in \mathbb{R}^2$. Our backbone of choice is a simple fully convolutional network; in particular, we adopt the L2Net (Tian et al., 2017) with the modification proposed in (Revaud et al., 2019) that employs smaller kernels in the last layers, in order to reduce its computational cost. Since we want to obtain a scalar map of the image, in order to exploit Morse theory, we modify the last layer of the backbone used in R2D2 (Revaud et al., 2019) to output a single channel. Thus, forwarding an image through the adopted network, we obtain a response map $F_\theta(I) = \mathfrak{H} \in \mathbb{R}^{H \times W}$. Essentially, the last layer distills the feature volume into a single-channel unified spatial representation, which we call *height map*, in an analogy with the terminology used in Morse theory.

4.1. Topological Detector Loss

During the training process, we model keypoints bijectively with the local maxima of the feature map. The main novelty of our method is implementing a framework that guarantees to find all the extrema. We *refer* to a local maximum via the associated topological feature, i.e., the loop that spawns around the critical point and that gets closed at its peak. Formally, we consider the set $\mathcal{G}(\mathfrak{H})$ of generators of $\mathbb{H}_1$, the 1-dimensional persistent homology module of the cubical complex, with the filtration given by the height map $\mathfrak{H}$ (see app. A.1). The set $\mathcal{G}(\mathfrak{H})$ is in bijection with the set of local maxima of the height map, and its elements should be understood as the loops described above. Every element $e \in \mathcal{G}(\mathfrak{H})$ can be associated to the coordinates of a (creator) saddle $s(e) \in \mathbb{R}^2$ and a (destructor) maximum $m(e) \in \mathbb{R}^2$. The birth time of e is the value attained by $\mathfrak{H}$ at its creator saddle, *i.e.*, $b(e) = \mathfrak{H}[s(e)]$, in the same way, the death time of e is the value of a local maximum $d(e) = \mathfrak{H}[m(e)]$. We make use of the **persistence** of e, a common way to measure the magnitude of a topological feature that is defined as

$$Pers(e) = d(e) - b(e) \tag{1}$$

Notice that this quantity does not depend on the shape or extension of the region filled by the maximum, *i.e.*, the scale, but only on how prominent the peak is. Given a map U of (possibly sparse) correspondences between two height maps, $\mathfrak{H}_1$ and $\mathfrak{H}_2$, for convenience, we define the error matrix as

$$E[i,j] = \mathfrak{H}_1[i,j] - \mathfrak{H}_2[U[i,j]] \tag{2}$$

if U is defined on (i,j), otherwise $E[i,j] = 0$. At this point, we introduce a new term that takes into account how the maps $\mathfrak{H}_1$ and $\mathfrak{H}_2$ differ at the topologically relevant (correspondent through U) positions, namely the **boundary similarity**:

$$Sim(e) = E[s(e)]^2 + E[m(e)]^2 \tag{3}$$

Given a positive constant α, our **detector loss** is finally defined as

$$\mathcal{L}_{det}(\mathfrak{H}_1, \mathfrak{H}_2) = -\sum_{e \in \mathcal{G}(\mathfrak{H}_1)} Pers(e)\left[Pers(e) - \alpha Sim(e)\right] \tag{4}$$

Let us now illustrate this object: the persistence can be thought of as a *peaky term*, as indeed maximizing the persistence can be achieved by increasing the prominence of the peaks or their number. The boundary similarity can be considered a *penalty term*, as it penalizes those topological features that are not reproducible across similar images. The hyperparameter α controls the trade-off of this regularization. Notice that, in contrast to patch-wise methods, we address the genuine set of local maxima found in the height map, which shows a high degree of variability and freedom. In this case, the persistence also takes on the role of a *weight term* that multiplies the total contribution of the feature to the loss, and this prevents the optimization process from getting overwhelmed by thousands of noisy, low-intensity peaks, instead focusing on the refinement of the promising features. Formally, the loss is a quadratic function in the persistence and boundary similarity variables, where the interaction term introduces a convenient relation between their gradients.

The differentiability of our loss function follows from well-established results (Leygonie et al., 2021; Carriere et al., 2021), see app. B for some details. A straightforward implementation within common deep learning automatic differentiation frameworks provides a valid gradient.

4.2. Training and Inference

For training our detector, we adopt WASF, the dataset released in (Revaud et al., 2019) to train R2D2, which provides homographic correspondences between pairs of images. A training instance is composed of two images I_1, I_2 and a ground-truth correspondences map between them $U \in \mathbb{R}^{H \times W \times 2}$, more explicitly $U[i,j] = (i', j')$ if and only if the pixel (i', j') of the second image corresponds to the pixel (i,j) in the first image; notice that U is defined only on covisible regions.

The final training loss given a pair of images is the following:

$$\mathcal{L}(I_1, I_2) = \mathcal{L}_{det}(\mathfrak{H}_1, \mathfrak{H}_2) \tag{5}$$

Where $\mathcal{L}_{det}$ is our novel detector loss and $\mathfrak{H}_1, \mathfrak{H}_2$ are the height maps obtained by forwarding the pair of images through our backbone.

At inference time, the keypoints are simply obtained by performing a fast non-maximum suppression algorithm that selects the locations corresponding to a local maximum of the height map with a value above a given threshold γ. Fig. 2 summarizes the training and inference steps.

5. Experiments

5.1. Dataset and Metrics

We assessed the capability of our method to predict repeatable keypoints using the well-established HPatches benchmark (Balntas et al., 2017). This dataset comprises 116 scenes, split into 696 images, with the first 57 scenes emphasizing variations in illumination and the subsequent 59 containing changes in viewpoint. Each sequence in the dataset comprises image pairs of increasing difficulty. We focus on this dataset, given that it represents a classical, longstanding benchmark for the task of keypoint detection, to assess the validity of our framework.

Regarding evaluation, our main concern is comparing the quality of the extracted keypoints for different detectors. Thus, as a metric, we use the formulation of repeatability proposed in (Lenc & Vedaldi, 2018), which evaluates the consistency of keypoints across different images while overcoming the issues of previous definitions of repeatability, which can bias toward detecting clusters of keypoints (Rey-Otero et al., 2015). This adaptation is detailed in C and aims to assess the unique association of keypoints by preventing a single keypoint from matching multiple counterparts, thus quantifying the proportion of keypoints that are each other's nearest neighbors in the corresponding images and are closer than a predefined distance threshold.

Method	Illumination					Viewpoint				
	250	500	1000	2000	4000	250	500	1000	2000	4000
D2-Net	21.1	22.0	23.6	26.4	28.7	12.1	13.6	19.5	18.6	22.1
R2D2	27.3	28.6	29.8	30.5	30.7	24.3	25.5	26.5	27.6	28.3
SIFT	34.9	37.2	38.8	40.4	41.2	37.8	38.9	39.9	40.7	40.4
SuperPoint	42.4	**47.7**	49.8	49.5	49.4	27.5	36.0	43.6	**46.8**	46.4
DISK	42.2	45.9	49.8	**54.2**	**57.4**	30.6	35.0	39.3	44.0	**47.6**
ALIKED	14.8	24.4	37.3	47.0	51.9	6.5	10.6	18.1	29.7	43.1
MorseDet (ours)	**44.3**	47.3	**50.3**	53.4	55.2	**40.6**	**42.8**	**44.6**	46.1	47.2

Table 1: Repeatability for illumination and viewpoint splits of HPatches, computed using various values for the maximum number of keypoints allowed. The **best** and second-best results are indicated in each column.

Method	Avg	75%	50%	25%
D2-Net	24.6	31.9	19.2	22.8
R2D2	48.5	55.7	56.2	33.7
SIFT	**63.6**	75.9	**64.8**	**50.2**
SuperPoint	60.6	73.3	63.0	45.6
DISK	56.0	71.8	57.4	38.8
ALIKED	18.7	24.2	16.5	15.4
MorseDet (ours)	62.2	**82.2**	63.0	41.3

Table 2: Repeatability of the detector on resized HPatches images as the scale factor progressively reduces. The **best** and second-best results are indicated in each column.

5.2. Baselines

In our study, we benchmark our model against a range of established detectors and state-of-the-art models to ensure a comprehensive evaluation:

- **SIFT**: a handcrafted detector designed to be robust against scale changes.
- **D2-Net**: employs a multi-scale inference approach, detecting local maxima across multiple output maps.
- **R2D2**: an unsupervised detector that uses multi-scale inference.
- **SuperPoint**: a semi-supervised detector trained to generalize to real images from labeled synthetic shapes.
- **DISK**: utilizes a probabilistic formulation to jointly model detection and matching.
- **ALIKED**: features deformable convolution, adapting receptor fields to the keypoints' support.

5.3. Implementation Details

For training, we employed AdamW (Loshchilov & Hutter, 2019) as an optimizer, with batches of 8 pairs of images with resolution 208×208. To generate image pairs, we follow the protocol from R2D2, using the same training datasets (WASF). Hyperparameter search and early stopping are performed on the validation split of MegaDepth (Li & Snavely, 2018) used in (DeTone et al., 2018). The final hyperparameters configuration included $\alpha = 10$, weight decay equal to 0.005, and repeatability threshold $\gamma = 0.7$ for inference. The training process on a single TITAN X GPU with 12GB of VRAM concluded in approximately 10 hours until convergence.

5.4. HPatches Benchmarks

Detector Repeatability

In this experiment, we evaluate the detector repeatability across changes in point of view and illumination on the common benchmark HPatches. Following (Revaud et al., 2019), we provide results across different values for the maximum number of detected keypoints allowed. The results are shown in the tab. 1, where the metrics are averaged across all thresholds up to 5px.

We can see that MorseDet's keypoints achieve consistently good performances, regardless of the number of keypoints or settings (*i.e.* illumination and viewpoint changes), being either best or second best across the table. Some other methods perform competitively with MorseDet under specific settings, although none is competitive in all cases. Notably, DISK has strong results with a high number of keypoints, and SIFT is second best with fewer keypoints under viewpoint changes but performs poorly under illumination changes. On average, SuperPoint is second-best.

Scale Repeatability

We posit that models employing a fixed-size window approach for keypoint modeling during training learn to predict keypoints at a specific frequency. Building on this premise, such models may struggle to consistently replicate keypoints under rescaling transformations. To study this idea in isolation, we designed the following experiment using the images of HPatches. We evaluated for every method the repeatability metric between every image resized to 1000×1000, and the image resized to smaller sizes to have approximately 75%, 50%, and 25% the pixel area of the original image. As the number of keypoints deeply influences repeatability, we

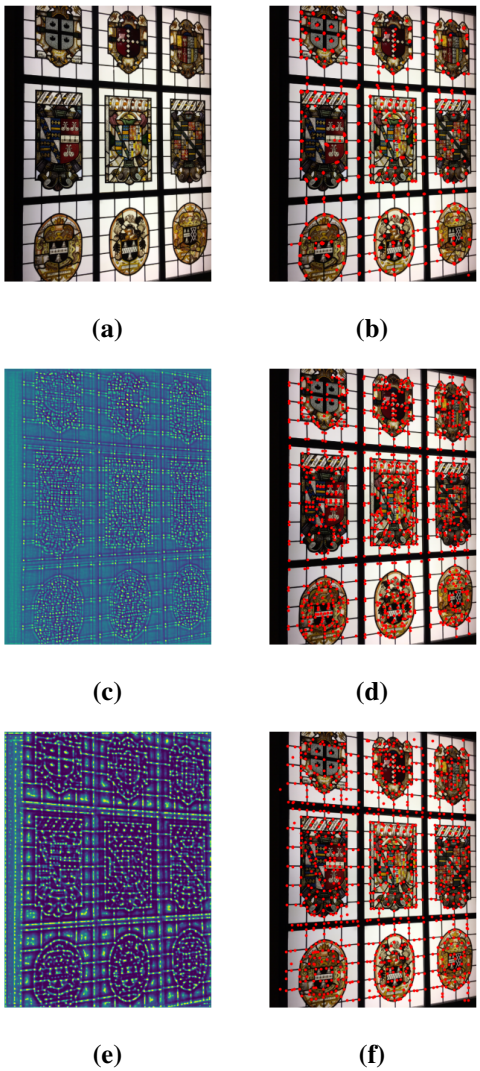

Figure 3: (a) HPatches image. (b) R2D2 multi-scale inference keypoints. (c) MorseDet height map. (d) MorseDet keypoints. (e) R2D2 repeatability map. (f) R2D2 single-scale inference keypoints. R2D2 tends to detect keypoints at a fixed resolution, creating artifacts along edges and in untextured areas. In contrast, MorseDet better adapts its keypoints to the scale of the image content. While multi-scale inference can improve R2D2, it often results in redundant predictions.

limit keypoints to 500, to ensure that every method uses the same number of keypoints at every scale for fair comparisons, thus also measuring how the methods can prioritize their most robust keypoints. The metrics are summarized in the tab. 2 by their average above all the thresholds till 5px.

The results show that MorseDet obtains second-best results on average after SIFT. In particular, MorseDet shines with 75% image resize (*i.e.* to images of 750×750), outperforming the second best method, SIFT, by 6.3 points. For extreme scale changes (*i.e.*, 25% of the original resolution), the best model is SIFT, which is a handcrafted detector built to be scale-invariant, followed by SuperPoint and MorseDet. Overall, the only learnable model competitive with MorseDet is SuperPoint, which benefits from a human-informed prior on keypoints. Notably, despite SIFT being proposed nearly two decades ago, it still outperforms modern detectors in this setup; MorseDet performs significantly better than every other learnable method in this task. This is a direct consequence of the fact that previous learnable methods lack a principled framework for modeling local maxima, which is our method's core contribution.

5.5. Qualitative Results

In fig. 3, we compare the height map generated by MorseDet and the repeatability map produced by R2D2, using a HPatches image. The figure highlights that MorseDet exhibits superior adaptability to the context compared to the constraints posed by a patch-wise unsupervised detector loss. The repeatability map of R2D2 shows a bias towards detecting keypoints at a fixed resolution, potentially resulting in the exclusion of some features and the generation of artifacts, especially along edges and in untextured areas. In contrast, MorseDet can adjust its keypoints according to the image content, effectively detecting features at both large and fine-grained scales. Although R2D2 partially mitigates some of these issues through multi-scale inference, it tends to produce redundant predictions by clustering keypoints around the same semantic feature while potentially missing other points of interest.

In fig. 4, we compare the outcomes of MorseDet with other detectors based on scalar maps, namely SuperPoint, DISK, and ALIKED, using an image from the Megadepth dataset. The figure illustrates inherent limitations in all previous approaches. Specifically, DISK exhibits poor localization, densely detecting keypoints on buildings. SuperPoint demonstrates a limited understanding of keypoints, correctly identifying the most evident features (e.g. sharp corners in clear areas). Still, it favors noisy keypoints in untextured areas rather than finding additional keypoints in relevant structures. ALIKED performs better, but a hard-coded resolution bias persists in its scalar map. Conversely, our method does not suffer from these limitations. Moreover, it exhibits an attractive emergent property by effectively capturing much of the input structure in its height map.

5.6. Ablation Study

To better understand the design of the loss function and the respective effects of the boundary similarity and the persistence terms, we conducted an ablation study on the effect of training with different values of α. The results regarding the repeatability of the HPatches viewpoint data

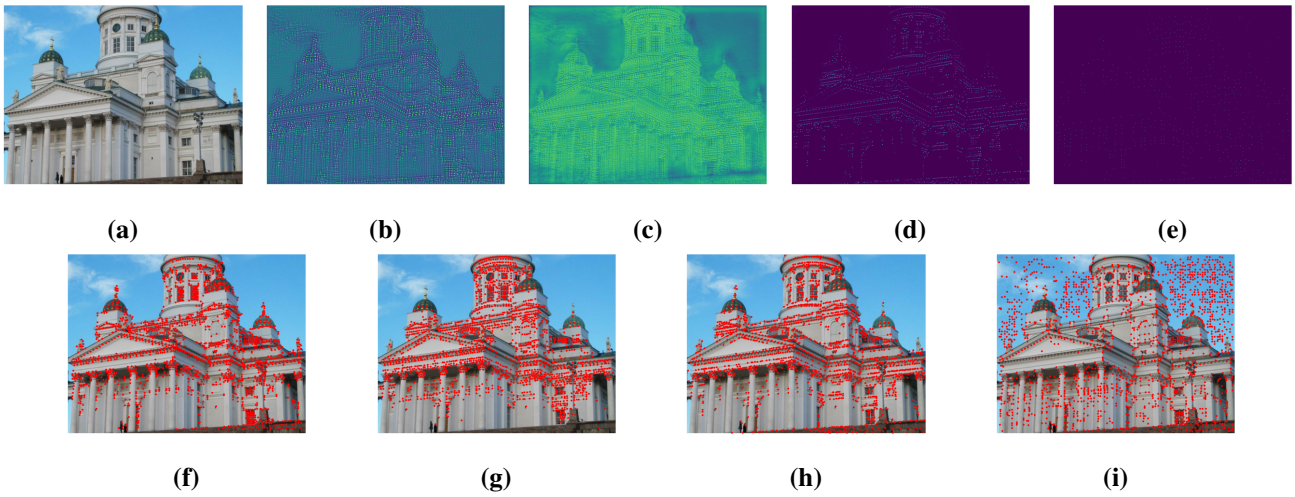

Figure 4: (a) Input image. (b) MorseDet, (c) DISK, (d) ALIKED, (e) SuperPoint scalar map. (f) MorseDet, (g) DISK, (h) ALIKED, (i) SuperPoint keypoints. DISK tends to generate equally-spaced keypoints, even in untextured areas. ALIKED predominantly produces keypoints at a fixed resolution. SuperPoint identifies the most clear features while overlooking others (to the extent of favoring the sky). MorseDet detects keypoints across various scales, and its height map reveals structural information from the input.

split are reported in the tab. 3.

The results show that the model's performance is poor in the absence of the Sim loss term (*i.e.* $\alpha = 0$). Conversely, increasing α increases performance across all considered keypoint quantities. At $\alpha = 20$ the repeatability continues to rise when the number of keypoints is limited but does not gain additional benefits from allowing a more significant number of keypoints, suggesting a trade-off imposed by the parameter's value, balancing the quantity and quality of detected keypoints.

Indeed, without the Sim term (*i.e.* $\alpha = 0$), the loss simplifies to the maximization of squared persistence, thus amounting to seeking an output feature map with as many local maxima as possible, leading to a trivial and uninformative solution, a grid of 1s surrounded by 0s in every 3×3 image patch, disregarding the input image value. The limitation of the model and training process impedes achieving an ideal grid pattern exactly, as shown in fig. 5. Nonetheless, training without the Sim component yields an output that resembles the trivial solution. On the contrary, setting a sensible value for α enables learning repeatability from the data.

α	250	500	1000	2000	4000
0	1.0	1.6	1.9	3.0	5.9
5	34.8	37.5	40.3	42.9	<u>45.1</u>
10	<u>40.6</u>	<u>42.8</u>	<u>44.6</u>	**46.1**	**47.2**
20	**44.0**	**44.8**	**45.3**	<u>45.1</u>	44.9

Table 3: Repeatability on HPatches viewpoint across various value for the maximum number of keypoints allowed, for models trained with different values of α. The **best** and <u>second-best</u> results are indicated in each column.

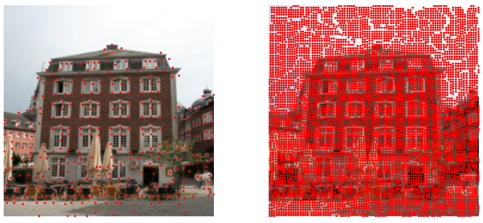

Figure 5: The images display all detected keypoints for a 208×208 training image, comparing outcomes from (a) the actual model with $\alpha = 10$ and (b) a model trained with $\alpha = 0$.

6. Conclusion and Limitations

In this study, we introduced the first training pipeline driven entirely by a loss function based on persistent homology. We based our approach on the observation that Morse theory and persistent homology are naturally suited for modeling the local extrema of a scalar map in a scale-independent way. Through the discrete Morse theory framework, we showed that these concepts can be applied to model the salient points of a digital image in a differentiable manner suitable for keypoint detection. Our experiments on the classical HPatches benchmark demonstrated promising re-

sults that validate our methods. We believe this work can serve as a stepping stone for future research in extending the applications of topological data analysis to the fields of computer vision and image matching.

However, this novelty also presents certain limitations. Due to the limited popularity of discrete Morse theory within the deep learning literature, more open-source software must be needed for efficient implementation. Our current implementation of the loss function, aligned with the methodologies of (Robins et al., 2011), operates on a CPU. Despite this, as detailed in sec. 5.3, training takes roughly 10 hours.

Acknowledgements

This study was carried out within the project FAIR - Future Artificial Intelligence Research - and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions; neither the European Union nor the European Commission can be considered responsible for them.

Impact Statement

This paper presents MorseDet, a work that aims to advance the field of Machine Learning. Our work has many potential societal consequences, none of which we feel must be specifically highlighted here.

References

Bailey, T. and Durrant-Whyte, H. Simultaneous localization and mapping (slam): Part ii. *IEEE robotics & automation magazine*, 13(3):108–117, 2006.

Balntas, V., Lenc, K., Vedaldi, A., and Mikolajczyk, K. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5173–5182, 2017.

Barbarani, G., Mostafa, M., Bayramov, H., Trivigno, G., Berton, G., Masone, C., and Caputo, B. Are local features all you need for cross-domain visual place recognition? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6154–6164, 2023.

Barroso-Laguna, A., Riba, E., Ponsa, D., and Mikolajczyk, K. Key. net: Keypoint detection by handcrafted and learned cnn filters. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5836–5844, 2019.

Bay, H., Tuytelaars, T., and Van Gool, L. SURF: Speeded up robust features. In *ECCV*, 2006.

Bhardwaj, S. and Mittal, A. A survey on various edge detector techniques. *Procedia Technology*, 4:220–226, 2012. ISSN 2212-0173. 2nd International Conference on Computer, Communication, Control and Information Technology(C3IT-2012) on February 25 - 26, 2012.

Calonder, M., Lepetit, V., Strecha, C., and Fua, P. Brief: Binary robust independent elementary features. In Daniilidis, K., Maragos, P., and Paragios, N. (eds.), *Computer Vision – ECCV 2010*, 2010.

Carriere, M., Chazal, F., Glisse, M., Ike, Y., Kannan, H., and Umeda, Y. Optimizing persistent homology based functions. In *International conference on machine learning*, pp. 1294–1303. PMLR, 2021.

Chazal, F., Guibas, L. J., Oudot, S. Y., and Skraba, P. Persistence-based clustering in riemannian manifolds. *Journal of the ACM (JACM)*, 60(6):1–38, 2013.

Chen, C., Ni, X., Bai, Q., and Wang, Y. A topological regularizer for classifiers via persistent homology. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2573–2582. PMLR, 2019.

Clough, J. R., Byrne, N., Oksuz, I., Zimmer, V. A., Schnabel, J. A., and King, A. P. A topological loss function for deep-learning based image segmentation using persistent homology. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8766–8778, 2020.

Csurka, G., Dance, C. R., and Humenberger, M. From handcrafted to deep local features. *arXiv preprint arXiv:1807.10254*, 2018.

da Silva, R. D., Schwartz, W. R., and Pedrini, H. Scalar image interest point detection and description based on discrete morse theory and geometric descriptors. In *2012 19th IEEE International Conference on Image Processing*, pp. 1877–1880, 2012.

Delgado-Friedrichs, O., Robins, V., and Sheppard, A. Skeletonization and partitioning of digital images using discrete morse theory. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):654–666, 2014.

DeTone, D., Malisiewicz, T., and Rabinovich, A. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 224–236, 2018.

Dey, T. K., Li, K., Luo, C., Ranjan, P., Safa, I., and Wang, Y. Persistent heat signature for pose-oblivious matching of incomplete models. In *Computer Graphics Forum*, volume 29, pp. 1545–1554. Wiley Online Library, 2010.

Durrant-Whyte, H. and Bailey, T. Simultaneous localization and mapping: part i. *IEEE robotics & automation magazine*, 13(2):99–110, 2006.

Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., and Sattler, T. D2-Net: A Trainable CNN for Joint Detection and Description of Local Features. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Edelsbrunner, Letscher, and Zomorodian. Topological persistence and simplification. *Discrete & Computational Geometry*, 28:511–533, 2002.

Edelsbrunner, H. and Harer, J. L. *Computational topology: an introduction*. American Mathematical Society, 2022.

Ghahremani, M., Liu, Y., and Tiddeman, B. Ffd: Fast feature detector. *IEEE Transactions on Image Processing*, 30:1153–1168, 2020.

Giansiracusa, N., Giansiracusa, R., and Moon, C. Persistent homology machine learning for fingerprint classification. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 1219–1226. IEEE, 2019.

Gleize, P., Wang, W., and Feiszli, M. Silk: Simple learned keypoints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22499–22508, October 2023.

Gupta, S., Zhang, Y., Hu, X., Prasanna, P., and Chen, C. Topology-aware uncertainty for image segmentation. *Advances in Neural Information Processing Systems*, 36, 2024.

Harris, C. G. and Stephens, M. J. A combined corner and edge detector. In *Alvey Vision Conference*, 1988.

Heinly, J., Schonberger, J. L., Dunn, E., and Frahm, J.-M. Reconstructing the world* in six days *(as captured by the yahoo 100 million image dataset). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Hofer, C., Kwitt, R., Niethammer, M., and Uhl, A. Deep learning with topological signatures. *Advances in neural information processing systems*, 30, 2017.

Hu, X., Wang, Y., Fuxin, L., Samaras, D., and Chen, C. Topology-aware segmentation using discrete morse theory. In *International Conference on Learning Representations*, 2021.

Hu, X., Samaras, D., and Chen, C. Learning probabilistic topological representations using discrete morse theory. In *International Conference on Learning Representations*, 2023.

Lenc, K. and Vedaldi, A. Large scale evaluation of local image feature detectors on homography datasets. *BMVC*, 2018.

Leygonie, J., Oudot, S., and Tillmann, U. A framework for differential calculus on persistence barcodes. *Foundations of Computational Mathematics*, pp. 1–63, 2021.

Li, Z. and Snavely, N. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2041–2050, 2018.

Lindeberg, T. Scale-space theory: A basic tool for analyzing structures at different scales. *Journal of applied statistics*, 21(1-2):225–270, 1994.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Lowe, D. G. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60: 91–110, 2004.

Matas, J., Chum, O., Urban, M., and Pajdla, T. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, 2004. ISSN 0262-8856. British Machine Vision Computing 2002.

Mikolajczyk, K. and Schmid, C. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60:63–86, 10 2004.

Mikolajczyk, K. and Schmid, C. A performance evaluation of local descriptors. *IEEE transactions on pattern analysis and machine intelligence*, 27(10):1615–1630, 2005.

Milnor, J. W. *Morse theory*. Princeton university press, 1963.

Mishkin, D., Radenovic, F., and Matas, J. Repeatability is not enough: Learning affine regions via discriminability. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 284–300, 2018.

Moor, M., Horn, M., Rieck, B., and Borgwardt, K. Topological autoencoders. In *International conference on machine learning*, pp. 7045–7054. PMLR, 2020.

Mur-Artal, R., Montiel, J. M. M., and Tardós, J. D. Orb-slam: A versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015.

Noh, H., Araujo, A., Sim, J., Weyand, T., and Han, B. Large-scale image retrieval with attentive deep local features. In *CVPR*, 2017.

Revaud, J., De Souza, C., Humenberger, M., and Weinzaepfel, P. R2d2: Reliable and repeatable detector and descriptor. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Rey-Otero, I., Delbracio, M., and Morel, J.-M. Comparing feature detectors: A bias in the repeatability criteria. In *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 3024–3028. IEEE, 2015.

Robins, V., Wood, P. J., and Sheppard, A. P. Theory and algorithms for constructing discrete morse complexes from grayscale digital images. *IEEE Transactions on pattern analysis and machine intelligence*, 33(8):1646–1658, 2011.

Salahat, E. and Qasaimeh, M. Recent advances in features extraction and description algorithms: A comprehensive survey. In *2017 IEEE international conference on industrial technology (ICIT)*, pp. 1059–1063. IEEE, 2017.

Sarlin, P.-E., Cadena, C., Siegwart, R., and Dymczyk, M. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12716–12725, 2019.

Sattler, T., Maddern, W., Toft, C., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., Okutomi, M., Pollefeys, M., Sivic, J., Kahl, F., and Pajdla, T. Benchmarking 6DOF outdoor visual localization in changing conditions. In *CVPR*, 2018.

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

Schönberger, J. L. and Frahm, J.-M. Structure-from-motion revisited. In *CVPR*, 2016.

Schönberger, J. L., Zheng, E., Pollefeys, M., and Frahm, J.-M. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016.

Tian, Y., Fan, B., and Wu, F. L2-net: Deep learning of discriminative patch descriptor in euclidean space. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6128–6136, 2017.

Toft, C., Maddern, W., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., Okutomi, M., Pollefeys, M., Sivic, J., Pajdla, T., Kahl, F., and Sattler, T. Long-Term Visual Localization Revisited. *TPAMI*, pp. 1–1, 2020.

Tuytelaars, T. and Van Gool, L. J. Wide baseline stereo matching based on local, affinely invariant regions. In *BMVC*, 2000.

Tyszkiewicz, M., Fua, P., and Trulls, E. Disk: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33:14254–14265, 2020.

Xu, Y., Monasse, P., Géraud, T., and Najman, L. Tree-based morse regions: A topological approach to local feature detection. *IEEE Transactions on Image Processing*, 23 (12):5612–5625, 2014.

Yi, K. M., Trulls, E., Lepetit, V., and Fua, P. Lift: Learned invariant feature transform. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pp. 467–483. Springer, 2016.

Zhao, X., Wu, X., Chen, W., Chen, P. C., Xu, Q., and Li, Z. Aliked: A lighter keypoint and descriptor extraction network via deformable transformation. *IEEE Transactions on Instrumentation and Measurement*, 2023a.

Zhao, X., Wu, X., Miao, J., Chen, W., Chen, P. C. Y., and Li, Z. Alike: Accurate and lightweight keypoint detection and descriptor extraction. *IEEE Transactions on Multimedia*, 25:3101–3112, 2023b.

Zomorodian, A. and Carlsson, G. Computing persistent homology. In *Proceedings of the twentieth annual symposium on Computational geometry*, pp. 347–356, 2004.

A. Algebraic Topology

A.1. Cubical Complex

A cubical complex is a *finite* family $\mathcal{K}$ of objects $Q \subset \mathbb{R}^d$, such that: for all $Q \in \mathcal{K}$, there is a set of integers $I_Q := \{l_1, \ldots, l_d\}$ such that $Q = I_1 \times \cdots \times I_d$, with $I_j = [l_j, l_j + 1]$ or $I_j = [l_j, l_j]$; and, if $P, Q \in \mathcal{K}$ then either $P \cap Q = \emptyset$ or $P \cap Q \in \mathcal{K}$. When $I = [l, l]$, it is called *degenerate*, and the number of non-degenerate intervals in Q is its dimension, while d is usually called its embedding number. The $0-$dimensional cubes are points, $1-$dimensional cubes are edges, $2-$dimensional cubes are squares, and so on. We denote by $\mathcal{H}_k$ the subset of $k-$dimensional cube in $\mathcal{K}$, thus, for example $\mathcal{K}_0$ are the $0-$cubes, or vertices, or points. Let $P \subset Q \in \mathcal{K}$, then P is called a *face* of Q, and, if the inclusion is proper, $\dim P \leq \dim Q - 1$. A cubical complex $\mathcal{K}$ is a partially ordered set (poset) via inclusion and, if $(\mathcal{P}, \leq_{\mathcal{P}})$ is another poset with $f : \mathcal{K} \to \mathcal{P}$ a monotonically not decreasing map, that is $P \subset Q$ implies $f(P) \leq_{\mathcal{P}} f(Q)$, then it is possible to create a sublevel filtration of $\mathcal{K}$ as follows: $\emptyset \subset \mathcal{K}_1 \subset \mathcal{K}_2 \cdots \subset \mathcal{K}_t = \mathcal{K}$ where $\mathcal{K}_s := \{Q \in \mathcal{K} : f(Q) \leq_{\mathcal{P}} p_s \in \mathcal{P}\}$. Topological invariants of these sublevel sets and their behavior along the filtration are some of the main topics in topological data analysis. In particular, they have been studied in Morse theory, discrete Morse theory, and persistent homology.

It is customary to represent $2D-$greyscale digital images as cubical complexes as follows: $0-$cubes are the images pixels laying on the vertices of an integral rectangular lattice in $\mathbb{R}^2$; $1-$dimensional cubes are the edges connecting pixel that differ by 1 in precisely one coordinate; $2-$dimensional cubes, $i.e.$ squares, are the obvious ones. Let $I : \{0 - cubes\} \to [0, 1]$ be the function assigning to each pixel its values. Then, we can associate I with a function $f : \mathcal{K}_I \to [0, 1]$ by $f(Q) = max_{P \in \mathcal{K}_0 \,:\, P \subseteq Q} I(P)$. The complex $\mathcal{K}_I$, with f_I and the corresponding filtration will be called the *cubical complex associated to I* or, more simply, the *I-complex*. We will see how to take advantage of this representation in a moment, but we need some further mathematical devices before proceeding.

A.2. Homology and Persistence

Homology is a topological invariant whose story goes back to the seminal work of H.Poincarè *Analysis Situs* published in 1895. For our setting, discussing homology for dimensions equal to zero and one only would suffice. In these cases, homology will count the number of connected components of a complex (dimension zero) and the number of closed one-dimensional paths that cannot be shortened (cycles or one-dimensional hole in the complex. Homology is defined via two objects: first, for $k = 0, 1, 2$, we consider the vector spaces $C_k(\mathcal{K})$ with bases in bijection with $\mathcal{K}_k$, notice that there are some subtleties regarding the orientation of the cube here, see (Edelsbrunner & Harer, 2022). Then, we introduce the boundary maps $\partial_k : C_k \to C_{k-1}$, which is the algebraic representation of the operation of taking the boundary of a cube. For example $\partial_1([l, l+1]) = [l+1] - [l]$ or $\partial_2([l, l+1] \times [l, l+1]) = [l, l+1] \times [l] + [l+1] \times [l, l+1] - [l, l+1] \times [l+1] - [l] \times [l, l+1]$. It is not difficult to check that $\partial_k \partial_{k+1} = 0$ for all suitable k, and this implies that the image of ∂_k, called B_k for *boundaries*, is a subspace of the kernel of ∂_{k-1}, denoted Z_k for cycles (zyklen). Their quotient $H_k := Z_k / B_k$ is called the $k-$th homology group (of $\mathcal{K}$). When one has a map of complexes $\mathcal{K} \to \mathcal{H}$, it induces a linear mapping $H_k(\mathcal{K}) \to H_k(\mathcal{H})$ for all k and one can check what happens to connected component and cycle of $\mathcal{K}$ when mapped to $\mathcal{H}$. Of paramount importance in our setting is that the sublevel filtration of the $I - complex$ induces a sequence of linear mappings on the homology groups. The space $\mathbb{H}_k(\mathcal{K}_I, f_I) = \bigoplus_{f_I([a,b]) \,:\, [a,b] \in \mathcal{K}_I} H_1(\mathcal{K}_I(f[a, b]))$is called the $k-$th *persistent homology group* of the filtered space $(\mathcal{K}_I, f_I)$. It is generated by persistent cycles as a module over the ring of polynomial in one variable, and, in the case of the cubical complex associated with an image I, it holds that the generators of its first persistent homology group are in one to one correspondence with the local maxima of the image, see (Edelsbrunner & Harer, 2022).

B. Differentiability

From the more general results of (Leygonie et al., 2021; Carriere et al., 2021), given a cubical complex K, there is a stratified function F, denoting with $\Omega(\mathbb{R}^{HW})$ the set of its strata, such that:

- on each stratum $S \in \Omega(\mathbb{R}^{HW})$ the function restricted to S behaves as a permutation P_S.
- if $\mathfrak{H} \in \mathbb{R}^{HW}$ has all distinct values, there is a stratum S such that $\mathfrak{H} \in S$, where P_S returns the m pairs of critical times associated with its local maxima in increasing order of birth time, followed by all the other n unpaired values in increasing order $F(\mathfrak{H}) = (b_1, d_1, ..., b_m, d_m, u_1, ..., u_n)$.

Intuitively, if $\mathfrak{H}$ has all distinct values and d is the minimum distance between them, in a neighborhood constituted by the open ball $\mathcal{B}(\mathfrak{H}, d)$ the location of its critical times cannot change. Consequently, our loss's persistence and boundary similarity terms can be expressed through composition with a fixed permutation P_S on a neighborhood of $\mathfrak{H}$, making them differentiable almost everywhere. Suppose multiple $\mathfrak{H}$ entries have the same value. In that case, we can remove the problem by considering a perturbation that makes all the values distinct while maintaining any other order relation. For instance, $\mathfrak{H}_\epsilon = \mathfrak{H} + \epsilon V$ where $V[i, j] = \frac{i+Ij}{2IJ}$ and ϵ is smaller than d the minimum positive difference between different entries. For $\epsilon \in (0, d)$, all H_ϵ belong to the same

stratum that corresponds to a fixed permutation P_S, allowing us to compute a directional derivative by arbitrarily choosing a trajectory V. In practice, we can arbitrarily break ties if needed.

C. Revisited Repeatability

Following the protocol established by (Mikolajczyk & Schmid, 2005), repeatability is a key metric in literature for assessing the detection of reproducible keypoints. We adopt the version from (DeTone et al., 2018), which aligns with current point-based prediction methods. For clarity, we detail this metric: given two sets of predicted keypoints A, B from a pair of images I_1, I_2 related by a homography U, a keypoint $x \in A$ is positively referenced in B if $\min_{y \in B} ||x - U^{-1}(y)||$ is less than a threshold ϵ , where $U^{-1}(y)$ is the projection of y through the ground truth homography. The repeatability score is the average number of keypoints with a positive reference, typically assessed within covisible areas, acknowledging the detectors' limitation of not knowing a priori which regions will be matched.

However, repeatability is influenced by the number of extracted keypoints. For instance, a uniformly distributed grid of keypoints can artificially inflate the score. To mitigate this, we varied the maximum number of keypoints in our experiments as in (Revaud et al., 2019). Despite this, the metric could still favor detectors producing clustered keypoints, as noted by (Rey-Otero et al., 2015; Lenc & Vedaldi, 2018). The work in (Lenc & Vedaldi, 2018) suggested a method that allows keypoints to match at most one time: computing repeatability based on matches from an optimally constructed bipartite graph, minimizing the sum of a cost function based on distance, with a proposed greedy approximation for the optimization problem.

Our revisited repeatability is a simplification that further requires keypoints to be mutually nearest neighbors, *i.e.*

$$x = \arg\min_{x' \in A} ||y - U(x')|| \tag{6}$$

and

$$y = \arg\min_{y' \in B} ||x - U^{-1}(y')|| \tag{7}$$

This addition helps unequivocally associate underlying features and penalizes redundant detections, favoring detectors more prone to cover image features with a limited number of keypoints comprehensively.