A Differentiable Topological Notion of Local Maxima for Keypoint Detection

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 which achieves competitive performance in keypoint repeatability and introduces a principled and theoretically robust approach to the problem.

1. Introduction

The ability to extract points from an image in consistent way across different views (*keypoint detection*) is a fundamental task of computer vision that found applications as a basic step of many complex applications of visual localization [1]–[3], SLAM [4]–[6], Structure-from-Motion and 3D reconstruction [7]–[9], as well as retrieval and place recognition [10], [11].

A theoretical framework for the problem is provided by scale-space theory [12], [13]. In this context, the keypoints of an image $I \in \mathbb{R}^{H\times W}$ are modeled as the set of local extrema (maxima and minima) of a suitable one-parameter image operator $F(I, s) \in \mathbb{R}^{H\times W}$, where the parameter s represents the scale at which we are looking for. The principle used to design the aforementioned scale-space operator is the *non-creation property*: when processing the image at multiple scales, what is noticeable from a distance (on a coarse scale), should have already been visible in the details (on a fine-grained scale). Therefore, an ideal scale-space operator should be coherent across different scales by detecting, as local extrema, the same keypoints at a greater scale s_2 and also at the smaller scale $s_1 < s_2$. For an appropriate relaxation of these requirements, under certain hypotheses, the work in [14] indicated that the convolution with a Gaussian of a variance equal to the scale parameter is the best solution among linear operators.

Many classical handcrafted keypoints detectors exploit the scale-space theoretical framework [15], [16], the most popular of which is SIFT [17]. In particular, the latter one operates by building a feature maps pyramid from the image by repeatedly applying Gaussian convolution and downsample operations. Ultimately, keypoints are detected as local extrema of the features maps.

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 [18]–[25]. Inspired by scale-space theory, at inference time, these approaches model keypoints as local maxima of a scalar map that is the output of a respective learnable convolutional neural network. However, at training time, several differentiable relaxations have been applied. For example, D2Net [22] makes use of a pixel score calculated in a fixed-size patch centered on the pixel. Similarly, R2D2 [21] considers the local maxima within a fixed-size $N \times N$ sliding window. DISK [23] provides a probabilistic formulation where the probability of being a keypoint still depends on a softmax logit calculated only on a neighboring patch.

Despite these recent innovations based on deep learning, classical handcrafted solutions still remain competitive and often outperform their learnable counterparts. We hypothesize that one of the main reasons for this is that the current formulation of keypoints adopted in the deep learning literature is based on a fixed-size patch-wise differentiable relaxation of the concept of local maxima. Indeed, this approach incentivizes models to detect keypoints at a given frequency, introducing a scale dependency that is in direct contrast with the non-creation property that earlier literature has identified as a key requirement. Therefore, a new

[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>.

Published at the 2^{nd} Differentiable Almost Everything Workshop at the 41^{st} International Conference on Machine Learning, Vienna, Austria. July 2024. Copyright 2024 by the author(s).

approach, based on a scale-agnostic and differentiable formulation of local maxima, is needed to develop unbiased learnable methods for keypoint detection.

To this end, we propose a novel scale-independent formulation of keypoints based on Morse theory [26] and persistent homology [27], [28] from algebraic topology. This formulation leverages the connection between local maxima and differentiable topological invariants [29], [30], offering a rigorous and differentiable solution without requiring hardcoded hyperparameters that determine the density or frequency of keypoints. Furthermore, we demonstrate the validity of our approach by achieving promising results on benchmarks for keypoints repeatability. Our implementation and trained models have been publicly released[1].

2. Background

2.1. Morse Theory

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 [26] is the mathematical framework that precisely captures the relationship between critical points and changes in topology. Formally, 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 sublevel 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 sublevel 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. An example of this case can be seen in fig. 1.

2.2. Discrete Morse Theory

In the context of digital images, we rely on a discretized version of the former theoretical framework, namely discrete Morse theory [31], and the concept of 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 $I_Q := \{l_1, \dots, l_d\}$ 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. 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.

In this setting a $2D-$greyscale digital image is represented as cubical complexes as follows: $0-$cubes are the image 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*. In layman's terms, we can think of the filtration of a cubical complex as the sublevel sets of a step function, where vertices, edges, and faces gradually appear. As shown in fig. 2, this process results in loops forming and disappearing at certain critical times, that are in correspondence with the respective saddle edge or maximum face.

2.3. Persistent Homology

Every topological feature e, that appears in the evolution of the sublevel sets, is associated with a pair of values $(b(e), d(e))$, $b(e) < d(e)$, the birth time and the death time of e. If e is a connected component, $b(e)$ corresponds to a minimum of h, and $d(e)$ must be a saddle. On the other hand, if e is a loop, then $b(e)$ corresponds to a saddle point

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

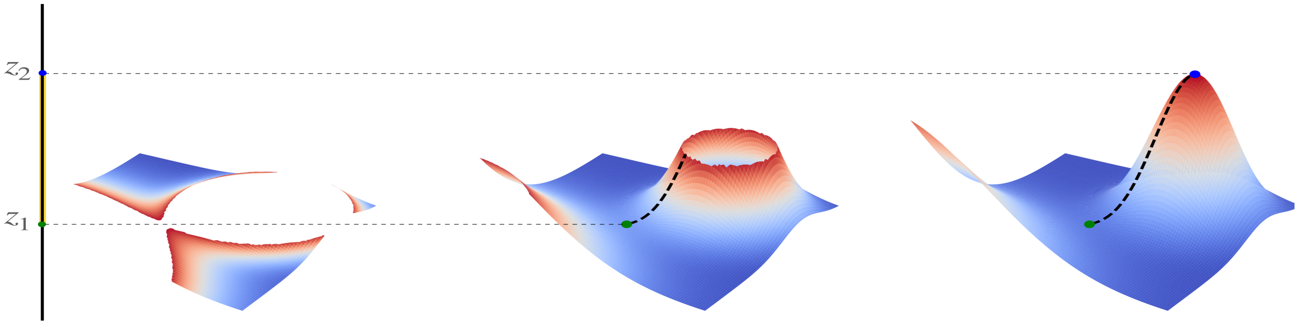


Figure 1: The evolution of the sublevel sets of a surface filtered by height, *i.e.* value on the z axis. As the height crosses z_1, a new loop is born in correspondence with a saddle (green point), then the loop changes smoothly until z hits z_2, the value of a corresponding maximum (blue point), and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

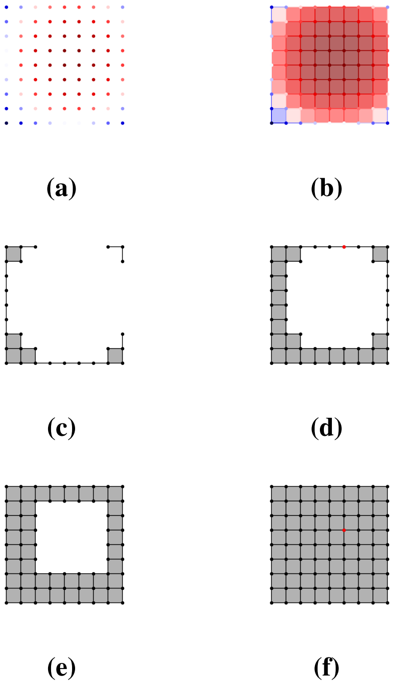


Figure 2: (a) A scalar image can be associated with (b) a cubical complex made of vertices, edges, and squares. In both examples, the color represents the pixel value and the filtration value of the cube, i.e., the maximum value among its vertices. Adding the cubes according to the filtration time, we observe the following sequence: (c) still no loop exists, (d) a loop is born as a saddle edge is added, (e) the loop *persists*, and (f) the loop gets closed by a maximum square. The pixels responsible for the birth and death of the loop are marked in red.

and $d(e)$ to a maximum. The *persistence* is defined as $Pers(e) = d(e) - b(e)$ and it represents the life span of a topological feature [27], [28].

3. Topological Detector Loss

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 $\{k_i\} \in \mathbb{R}^2$. We propose a novel keypoint detector, which we call **MorseDet**, based on a CNN backbone, that outputs a scalar feature map which we call *height map*, in an analogy with the terminology used in Morse theory. Thus, forwarding an image through the adopted network, we obtain a response map $F_\theta(I) = \mathfrak{H} \in \mathbb{R}^{H\times W}$.

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.

As for the previous approaches, 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.

In contrast with previous heuristical methods, during the training process, we model keypoints bijectively with the local maxima of the feature map. 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. Let $\mathcal{G}(\mathfrak{H})$ be the set of such topological features of the cubical complex associated with $\mathfrak{H}$. Every element $e \in \mathcal{G}(\mathfrak{H})$ can be associated with 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, as defined in sec. 2.3, to measure the magnitude of a topological feature. 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.

For convenience, we define the error matrix between two height maps, $\mathfrak{H}_1$ and $\mathfrak{H}_2$, based on the correspondences map U:

$$E[i,j] = \mathfrak{H}_1[i,j] - \mathfrak{H}_2[U[i,j]] \tag{1}$$

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*:

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

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

$$\mathcal{L}_{det}(\mathfrak{H}_1, \mathfrak{H}_2) = -\sum_{e \in \mathcal{G}(\mathfrak{H}_1)} Pers(e)\left[Pers(e) - \alpha Sim(e)\right] \tag{3}$$

In practice, minimizing the loss function aims to increase the number and prominence of local maxima in the height map, as long as they are reproducible across similar images.

3.1. Differentiability

Consider a vectorized scalar image $\mathfrak{H} \in \mathbb{R}^{HW}$ where all entries are distinct, and let d be the minimum distance between these values. When moving within a neighborhood defined by the open ball $\mathcal{B}(\mathfrak{H}, d)$, the order in which the cubes spawn along the associated filtration remains unchanged, as do the critical times locations. Therefore, within a region where the entries follow a strict order $S = \{x \in \mathbb{R}^{HW} \mid x_{\sigma(1)} < \cdots < x_{\sigma(HW)}\}$, the function P_S that selects the relevant entries for the m critical time pairs, $P_S(\mathfrak{H}) = (b_1, d_1, \ldots, b_m, d_m)$, acts as a fixed linear projection. The persistence and boundary similarity terms of our loss can thus be expressed as compositions of P_S with the height maps $\mathfrak{H}_1$ and $\mathfrak{H}_2$. Consequently, the loss function is differentiable almost everywhere.

A problem arises when multiple entries in $\mathfrak{H}$ have the same value. To address this, we can introduce an arbitrary perturbation that ensures all values are distinct while preserving any other order relations. 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 distinct entries. For $\epsilon \in (0, d)$, all $\mathfrak{H}_\epsilon$ belong to the same region S corresponding to a fixed projection P_S. This allows us to compute the loss and its gradient by extending their values continuously along the arbitrary direction.

For many functions of the persistence terms, such as their sum, stronger results hold [29], [30]. These functions are locally Lipschitz, and the process described above provides a subgradient along the trajectory V. In these cases, we have a guarantee for the convergence to a local minimum of the gradient methods. In contrast, our loss function is not continuous at the boundaries of the regions S and its value strictly depends on the choice of the arbitrary perturbation, but we still obtain a directional derivative.

4. Experiments

Our experiments assessed the ability of MorseDet to predict repeatable keypoints that are robust to changes in scale, viewpoint, or illumination. We adopted the HPatches benchmark and repeatability metric to compare our model against the best current deep learning detectors, namely D2-Net, R2D2, SuperPoint, DISK, and ALIKED, as well as the classical handcrafted solution SIFT. Due to space constraints, the details of the implementation, experimental settings, and results are covered in app. A.

To summarize our findings, we observed that MorseDet performs better than the other learned detectors against scale shifts, ranking second on average in this scenario after SIFT. Moreover, we found that, in viewpoint and illumination tests, our model achieves the most consistent performance. Notably, while SIFT demonstrates strong scale invariance properties, it fails to generalize as effectively in noisy settings, such as changes in illumination.

5. Conclusion and Limitations

In this paper, we introduce an algebraic topology technique to address a ubiquitous problem in the image matching literature: the scale dependency of keypoints due to patchwise differentiable relaxation of local maxima. By utilizing the concept of sublevel set filtration and its connection with Morse theory, we model local maxima in a scale-independent manner that is, thanks to persistence homology, suitable for gradient methods. Experimental results demonstrate the validity of our approach, showing strong results in terms of keypoints repeatability against changes in scale, viewpoint, and illumination.

Nevertheless, challenges remain in terms of differentiability (see sec. 3.1). The regularization term we introduced meets only minimal requirements and lacks a guarantee of convergence for gradient methods due to the discontinuous nature of the loss function. Therefore, we believe there is significant potential for further improvements by extending our framework to achieve complete differentiability.

References

[1] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 716–12 725.

[2] T. Sattler, W. Maddern, C. Toft, *et al.*, "Benchmarking 6DOF outdoor visual localization in changing conditions," in *CVPR*, 2018.

[3] C. Toft, W. Maddern, A. Torii, *et al.*, "Long-Term Visual Localization Revisited," *TPAMI*, pp. 1–1, 2020.

[4] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "Orb-slam: A versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[5] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: Part i," *IEEE robotics & automation magazine*, vol. 13, no. 2, pp. 99–110, 2006.

[6] T. Bailey and H. Durrant-Whyte, "Simultaneous localization and mapping (slam): Part ii," *IEEE robotics & automation magazine*, vol. 13, no. 3, pp. 108–117, 2006.

[7] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *CVPR*, 2016.

[8] J. Heinly, J. L. Schonberger, E. Dunn, and J.-M. Frahm, "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)*, Jun. 2015.

[9] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, "Pixelwise view selection for unstructured multi-view stereo," in *ECCV*, 2016.

[10] G. Barbarani, M. Mostafa, H. Bayramov, *et al.*, "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*, 2023, pp. 6154–6164.

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

[12] T. Lindeberg, "Scale-space theory: A basic tool for analyzing structures at different scales," *Journal of applied statistics*, vol. 21, no. 1-2, pp. 225–270, 1994.

[13] T. Lindeberg, "Scale invariant feature transform," 2012.

[14] J. J. Koenderink, "The structure of images," *Biological cybernetics*, vol. 50, no. 5, pp. 363–370, 1984.

[15] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *International Journal of Computer Vision*, vol. 60, pp. 63–86, Oct. 2004.

[16] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *ECCV*, 2006.

[17] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91–110, 2004.

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

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

[20] Y. Tian, B. Fan, and F. Wu, "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.

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

[22] M. Dusmanu, I. Rocco, T. Pajdla, *et al.*, "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.

[23] M. Tyszkiewicz, P. Fua, and E. Trulls, "Disk: Learning local features with policy gradient," *Advances in Neural Information Processing Systems*, vol. 33, pp. 14 254–14 265, 2020.

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

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

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

[27] Edelsbrunner, Letscher, and Zomorodian, "Topological persistence and simplification," *Discrete & Computational Geometry*, vol. 28, pp. 511–533, 2002.

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

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

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

[31] V. Robins, P. J. Wood, and A. P. Sheppard, "Theory and algorithms for constructing discrete morse complexes from grayscale digital images," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 33, no. 8, pp. 1646–1658, 2011.

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

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

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

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

[36] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *International Conference on Learning Representations*, 2019.

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

A. Experiments Details

A.1. Dataset and Metrics

We assessed the capability of our method to predict repeatable keypoints using the well-established HPatches benchmark [32]. 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 [33], 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 [34]. This adaptation aims to assess the unique association of keypoints by preventing a single keypoint from matching multiple counterparts, in detail, it quantifies the proportion of keypoints that are each other's nearest neighbors in the corresponding images and are closer than a predefined distance threshold, considering both the coordinate system of both the images.

A.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** [17]: a handcrafted detector designed to be robust against scale changes.
- **D2-Net** [22]: employs a multi-scale inference approach, detecting local maxima across multiple output maps.
- **R2D2** [21]: an unsupervised detector that uses multi-scale inference.
- **SuperPoint** [35]: a semi-supervised detector trained to generalize to real images from labeled synthetic shapes.
- **DISK** [23]: utilizes a probabilistic formulation to jointly model detection and matching.
- **ALIKED** [25]: features deformable convolution, adapting receptor fields to the supports of keypoints.

A.3. Implementation Details

We adopted L2Net [20] as backbone with the last convolutional layer modified as in R2D2. For training, we employed AdamW [36] as 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 [37] used in [35]. 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.

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.

A.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 [21], 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.

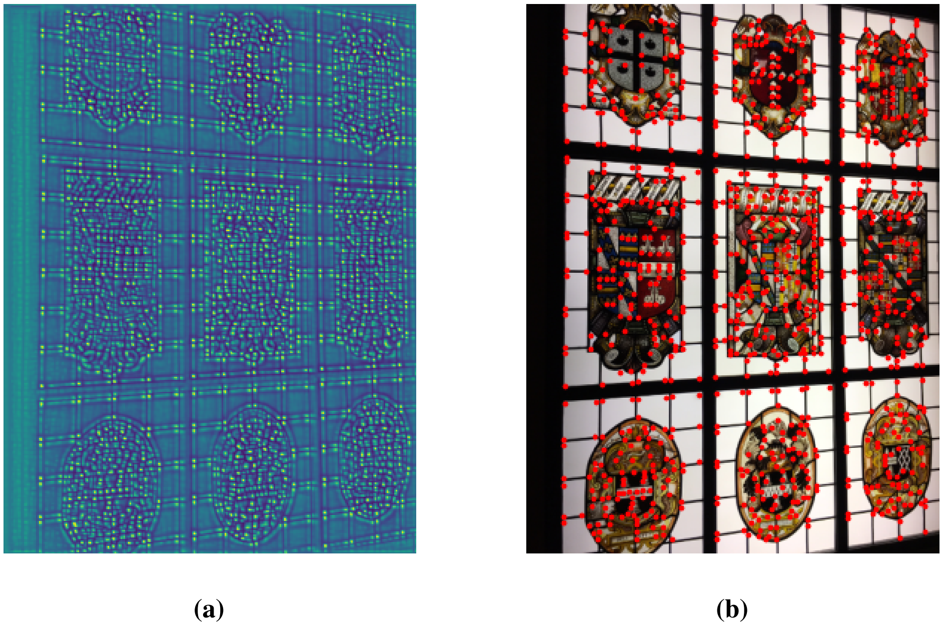
Scale Repeatability

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 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.

A.5. Qualitative Results

Fig. 3 shows an example of our model's height map and detected keypoints. It demonstrates how the model adapts the frequency of keypoints to the image content, effectively detecting both large-scale corners and fine-grained details without



(a) **(b)**

Figure 3: An example of MorseDet's (a) height map and (b) detected keypoints on a HPatches image. Our model adapts the frequency of its keypoints to the scale of the image content.

creating artifacts in low-textured regions.