From Robustness to Efficiency: Deformation-Aware and Efficient Local Feature Extraction for Images

Guilherme Potje*
Dep. de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte – MG, Brazil
Email: guipotje@dcc.ufmg.br

Renato Martins
Lab. Interdisc. Carnot de Bourgogne
Université de Bourgogne
Dijon, France
Email: renato.martins@u-bourgogne.fr

Erickson R. Nascimento
Dep. de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte – MG, Brazil
Email: erickson@dcc.ufmg.br

***Abstract*—Just as animals rely on visual perception and geometric understanding to navigate the 3D world, modern computers emulate this ability through Simultaneous Localization and Mapping (SLAM), image-based 3D reconstruction, and visual place recognition techniques, all relying on image features for obtaining correspondences. However, most feature extraction methods handle only affine transformations, ignoring non-rigid deformations, ubiquitous in the real-world. This work investigates deformation-aware local features, leveraging RGB-D images to compute geodesics, where RGB denotes the visible channels (Red, Green, Blue) and D represents image depth. Then, we generalize the concept to RGB-only images via learned representations. We introduce a novel RGB-D dataset with non-rigid deformations for real-world benchmarking, where experiments showed significant improvements in foundational vision tasks as matching and registration when adopting our proposed strategies. Finally, we present an efficient local feature extractor, balancing accuracy with reduced computational cost, expanding visual perception for mobile computers.**

I. Introduction

For over two decades, keypoint detection and local feature description have been central to image matching, supporting tasks as Visual SLAM [1], Structure-from-Motion (SfM) [2], [3], and image retrieval [4], [5], which rely on pixel-level and image-level correspondences for subsequent processing. Since the Scale-Invariant Feature Transform (SIFT) [6], numerous descriptors have emerged, categorized by input type (intensity/depth) and design (handcrafted/learning-based). Fast processing is crucial, driving research on binary features [7], [8] to efficiently establish correspondences. Feature invariance to geometric and photometric changes is fundamental [9], yet most descriptors achieve approximate affine invariance, struggling with non-affine changes. We show that explicitly modeling invariance significantly enhances robustness in challenging image matching scenarios, as shown in Fig. 1.

Geometric information, such as depth images, has gained popularity for computing distinctive feature descriptors [15], [16], offering robustness to lack of texture and improving rigid surface descriptions. Multimodal approaches combining intensity and depth data enhance keypoint descriptions but struggle under strong affine or non-rigid deformations. Since real-world objects often undergo isometric deformations (where geodesic distances remain unchanged), capturing this invariance is crucial for robust feature extraction. We study isometry by designing local features with explicit isometric sampling using RGB-D images. However, depth sensors suffer from noise, miscalibration, and missing data, while depth data itself is often inaccessible, making robust RGB-based descriptors essential. Deep learning improves invariance to illumination and perspective but remains weak against non-rigid deformations, even with relevant training data. To address invariance, we introduced novel strategies that incorporate explicit deformation-awareness into deep learning frameworks using single RGB images. Additionally, we highlight the overlooked role of keypoint detection in deformation-aware methods and propose a joint detection-description framework optimized for non-rigid surfaces. Finally, we present a lightweight Convolutional Neural Network (CNN) based image matching

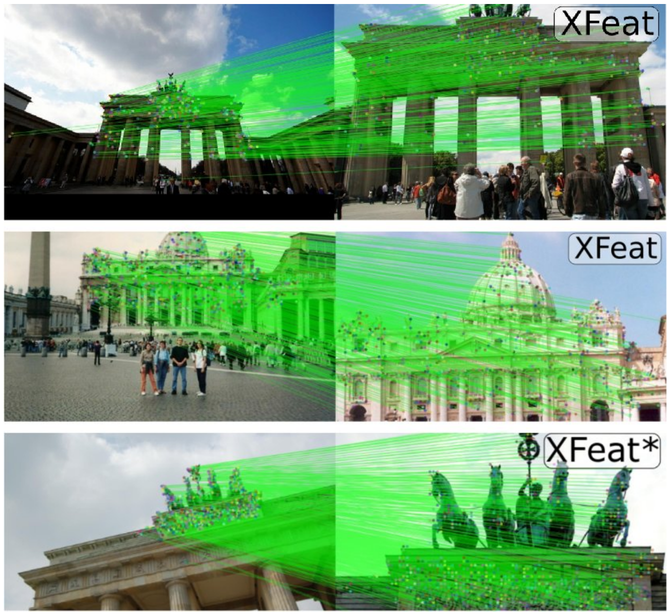

Fig. 1. **Visual matching under strong changes in viewpoint and illumination conditions.** Our lightweight descriptor (XFeat) stands out with its dual ability to perform both sparse and semi-dense matching, providing fast feature extraction for a wide range of applications from visual localization with sparse matches to 3D reconstruction with denser correspondences.

[*] This work is related to the Ph.D. thesis of Guilherme Potje.

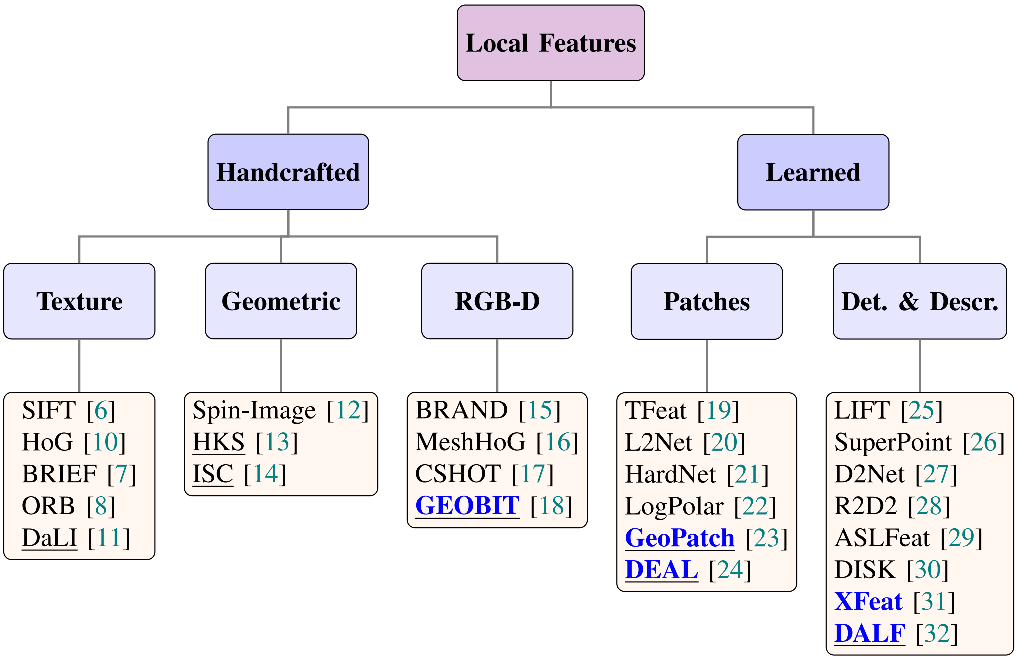

Fig. 2. **Taxonomy of local feature extraction methods.** BLUE BOLD highlights our contributions. Modern approaches are categorized as handcrafted or learning-based, grouped by input modality. Most learned methods rely on RGB, with further division into patch-based and detect & describe paradigms. Underlined denotes methods handling non-rigid deformations. Our work bridges gaps in RGB and RGB-D descriptors by integrating texture and geometric cues, leveraging isometric invariance and explicit deformation modeling in deep networks.

solution that balances accuracy and computational efficiency, achieving real-time performance on standard CPUs.

A. Contributions

This work advances local feature extraction invariant to non-rigid deformations in RGB-D and RGB images in five key contributions: (1) We introduce *GeoBit* (ICCV'19), a geodesic-aware binary descriptor, and *GeoPatch* (CVIU'22), a learning-based descriptor that efficiently samples pixels over geodesic distances, enabling CNN-based training. (2) To overcome depth dependence, we then propose *DEAL* (NeurIPS'21), an end-to-end trainable architecture embedding geometric transformations directly into CNNs, achieving state-of-the-art results on real-world data. (3) Extending DEAL, we develop *DALF* (CVPR'23), a joint keypoint detector and descriptor trained end-to-end for deformation-aware local features. (4) We provide an RGB-D benchmark with 11 real-world objects and a large synthetic dataset, including ground-truth dense flow fields for evaluating non-rigid matching. (5) Finally, we present *XFeat* (CVPR'24), a compact CNN for local feature extraction, offering a 5× speedup over existing methods while maintaining competitive accuracy.

Our works establish a principled framework for learning deformation-invariant representations, enabling computers to perceive the world through robust, structure-aware understanding of non-rigid transformations. In parallel, we introduce efficient deep networks for real-time image matching, advancing the design of compact and discriminative feature representations that *expand the perceptual intelligence of computers* across practical and constrained computing environments. All code and datasets are available at https://www.verlab.dcc.ufmg.br/descriptors.

II. Related Work

Local feature extraction has transitioned from handcrafted techniques to learning-based approaches. Traditional methods such as SIFT [6] rely on image gradients and blob detection to extract stable keypoints. To explicitly introduce deformation invariance, methods such as DaLI [11] leveraged computational geometry to improve the invariance to deformations. More recently, unified learning frameworks have integrated keypoint detection and description [26], [28], [29], enhancing efficiency and robustness compared to decoupled approaches. Despite advances in feature representation, current handcrafted and deep-learning-based methods suffer from high computational costs. Furthermore, modern descriptors lack explicit deformation handling and are expensive to employ in high-resolution images, essential for image matching.

A. Research Contextualization and Relevance

Fig. 2 provides a comprehensive overview of existing literature, with our contributions (highlighted in blue) advancing local feature representations. We introduce three key approaches: (1) *Geodesic-Aware Descriptors*, which exploit RGB-D depth data to ensure isometric invariance while maintaining robustness to rotation and scale; (2) *Deformation-Aware Descriptors*, which learn geometric transformations from RGB images using spatial transformers and reinforcement learning; and (3) *Accelerated Features*, where we devise strategies to push the

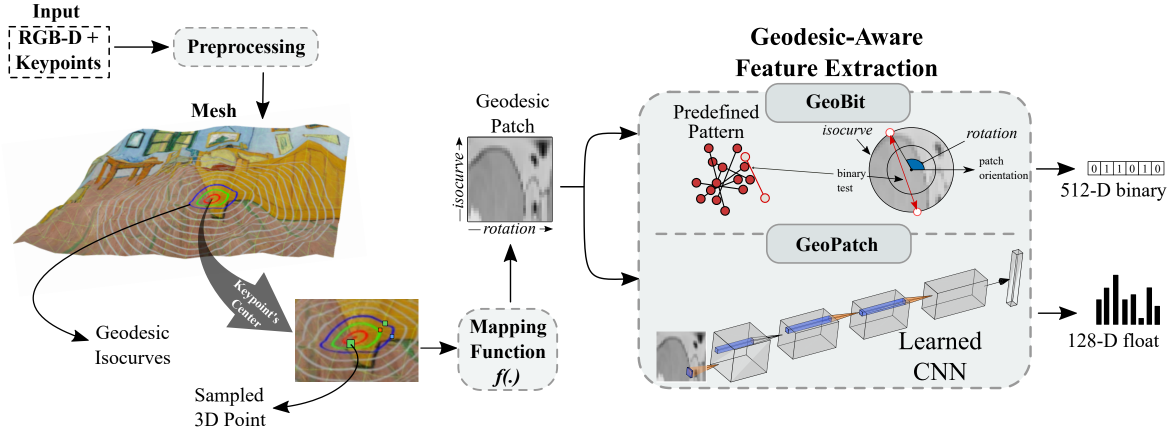

Fig. 3. **Geodesic-awareness in local feature extraction.** Our method processes an RGB-D image through a preprocessing step to reconstruct a consistent mesh, followed by the geodesic mapping function $f(.)$. This mapping function is then used to sample the intensity field of the texture image (geodesic-aware feature extraction), ensuring deformation invariance by design.

efficiency limits of deep learning for feature extraction while preserving competitive accuracy and robustness.

III. Geodesic-Aware Descriptors

The proposed geodesic-aware methods [18], [23] leverage intrinsic surface properties to achieve invariance to non-rigid deformations, making them naturally suited for analyzing deformable surfaces in RGB-D data. By modeling the surface as a smooth 2D manifold, our methods extract isometric-invariant visual features through geodesic isocurves, which inherently preserve the surface's intrinsic geometry. The approach consists of two key steps: first, computing a geodesic mapping function from depth maps, ensuring invariance to isometric deformations; second, extracting distinctive features from pixel intensities. Let $\mathcal{K} = \{\mathbf{k}_i \in \mathbb{R}^2\}_{i=1}^{N}$ denote a set of keypoints in the image domain, and let $I : \Omega \to \mathbb{R}^4$ be an RGB-D image defined over the pixel grid $\Omega \subset \mathbb{R}^2$. We consider a mapping $m : I \to \mathcal{M}$, where $\mathcal{M} = (\mathcal{V}, \mathcal{E}, \mathcal{C})$ is a colored mesh with vertices $\mathcal{V} \subset \mathbb{R}^3$, edges $\mathcal{E}$ defined by the grid neighborhood, and vertex colors $\mathcal{C} : \mathcal{V} \to \mathbb{R}^3$. This formulation enables geodesic computation directly on the reconstructed surface while preserving photometric information.

Mapping function computation. To enhance robustness, we perform depth preprocessing to mitigate noise and fill missing values through denoising and hole-filling, ensuring a reliable surface representation. Then, the geodesic mapping function is computed via heat flow [33] or a more efficient *local geodesic expansion* method proposed in *GeoPatch* [23], preserving intrinsic geometry and allowing invariant feature extraction under surface deformations. In Fig. 3, the method preprocesses an RGB-D image and keypoints to reconstruct a surface mesh. For each keypoint $\mathbf{x}_i$, the isocurve set is defined as the set of mesh points whose geodesic distance to $\mathbf{x}_i$ lies in a discrete set of radii, i.e., $\mathcal{N}_i = \{\mathbf{x}_j \mid d_G(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{R}\}$, where d_G is the geodesic distance and $\mathcal{R} = \{r_k\}$ for k discrete values. The mapping f samples the image intensity at these points and, considering discrete in-plane rotations $\{\theta_l\} \subset \Theta$ for l discrete values, constructs a geodesic patch $\mathbf{P}_i \in \mathbb{R}^{N_r \times N_\theta}$, where N_r and N_θ are the numbers of radii and rotations (e.g., 32 × 32), with the x-axis indexing rotation and the y-axis indexing geodesic radius. These rectified patches are then used as input to geodesic-aware feature extractors, either handcrafted (GeoBit) or learned (GeoPatch), to produce invariant local descriptors.

Descriptor extraction. Two novel descriptors, GeoBit and GeoPatch, employ the geodesic-aware strategy using the proposed mapping function. GeoBit encodes visual information using binary intensity tests on the geodesic coordinates, while GeoPatch utilizes a shallow CNN trained using the geodesic patch representation with a ranking loss on synthetic data, as detailed in Fig. 3.

IV. Deformation-Aware Descriptors

In contrast to geodesic-aware descriptors that rely on metric depth data, we introduce a deformation-aware local descriptor *DEAL* [24] that uses solely RGB images to extract discrimina-

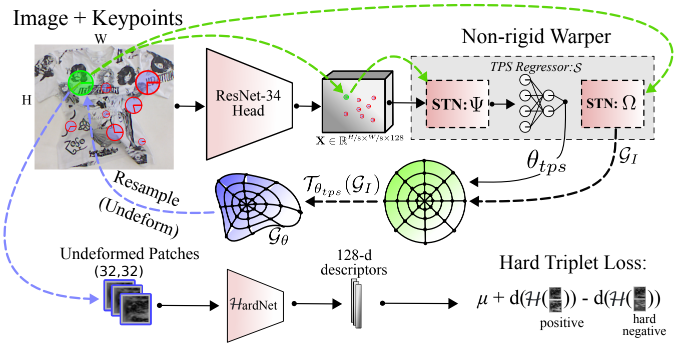

Fig. 4. **Proposed formulation for computing descriptors of deforming objects.** The non-rigid warper undeforms local patches by estimating warp parameters θ_{tps} from global ResNet features $\mathbf{X}$, using two spatial transformers and a TPS regressor. $\mathcal{G}_I$ is an identity polar grid with predefined radius that is warped by the learned transformation into $\mathcal{G}_\theta$. Rectified patches are fed to HardNet to extract discriminant descriptors, and the model is trained end-to-end with a hard triplet loss, whose objective is to pull hard negative samples (sampled in-batch) beyond the margin μ.

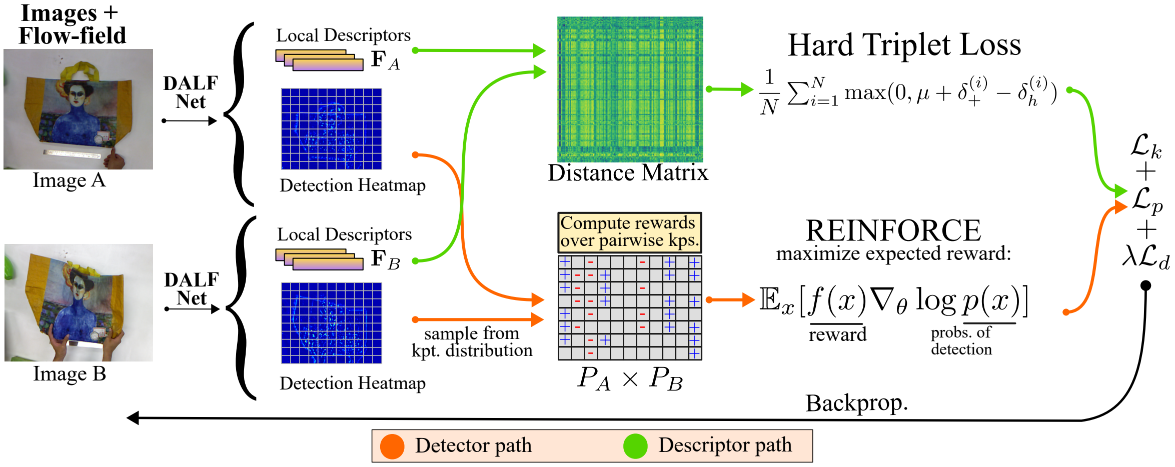

Fig. 5. **Training strategy to learn to detect and describe keypoints robust to deformations.** DALF net, composed of a U-Net CNN and a non-rigid warper, outputs a keypoint heatmap $\mathbf{K} \in \mathbb{R}^{H \times W}$, where H and W denote the image height and width, respectively, and a set of local descriptors $\mathbf{D} \in \mathbb{R}^{128}$. In the detector branch, $\mathbf{K}$ is optimized using the REINFORCE algorithm to promote keypoint repeatability under deformations. In the descriptor branch, $\mathbf{D}$ is learned with a hard triplet loss. The network is trained in a Siamese configuration using image pairs.

tive, deformation-invariant features from surfaces undergoing deformation (e.g., humans, animals, or cloth).

Learning to deform. Due to the ill-posed nature of estimating geometric transformations from single images, our CNN integrates a Spatial Transformer Network (STN) [34] that explicitly learns local surface deformations by estimating differentiable Thin-Plate Spline (TPS) warps to rectify image patches around keypoints. The TPS warps representing 2D coordinate mappings $\mathbb{R}^2 \to \mathbb{R}^2$ are used to model non-rigid deformations in the architecture, allowing differentiable spatial warping via attention mechanism. Unlike existing patch-based (HardNet [21]) or dense methods (SuperPoint [26], R2D2 [28]), *DEAL* encodes deformation cues into mid-level CNN features (e.g., shadows, textures, local shape), explicitly rectifying local regions via the learned non-rigid warp θ_{tps} end-to-end as shown in Fig. 4, thus improving descriptor invariance. Trained end-to-end with our proposed simulated dataset, our differentiable approach achieves accurate correspondences without requiring depth data and human labels, significantly improving matching performance in challenging real-world scenarios.

Learning keypoint detection & description. Considering the main limitation of DEAL, which extracts deformation-invariant features but relies on external keypoint detectors, we subsequently proposed a novel unified framework, DALF (Deformation-Aware Local Feature) [32]. DALF jointly learns keypoint detection and description to robustly address non-rigid image matching challenges. Existing deformation-aware methods typically neglect the keypoint detection phase, significantly limiting their effectiveness under strong deformation conditions. DALF overcomes this limitation by employing a cooperative training strategy that simultaneously optimizes a detection heatmap $\mathbf{K}$ and local descriptors $\mathbf{D}$, both parameterized by a neural network θ. Specifically, DALF leverages reinforcement learning via policy gradient to increase keypoint detection repeatability $\mathbb{E}_x^{\text{reward}}\left[f(x)\,\nabla_\theta \log p(x)^{\text{prob.}}\right]$ under deformation, as shown in Fig. 5, promoting repeatability and reliability of keypoints, while concurrently training the descriptor to rectify and describe image patches invariantly using the differentiable TPS transformations from DEAL's non-rigid warper module. DALF does not require expensive human labels or pseudo-ground-truth data, instead relying solely on synthetically generated deformations as our previous methods. Furthermore, DALF fuses distinctive mid-level CNN features with deformation-invariant features through an attention-based feature fusion layer, achieving a balanced representation that remains robust under significant geometric and photometric changes.

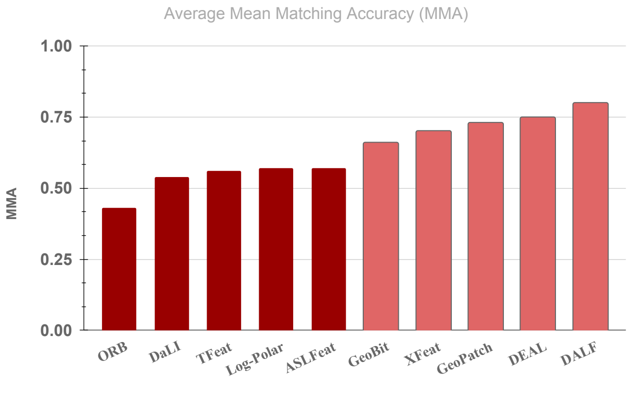

Fig. 6. **Average MMA across all four non-rigid deformation datasets.** Our methods are highlighted in light red. On average, our proposed descriptors achieve significantly better scores for matching non-rigid surfaces.

V. Experiments with Real Non-Rigid Deformations

We conducted experiments on our proposed non-rigid image matching benchmark [23], consisting of images of non-rigid objects with real deformations. All of our proposed descriptors

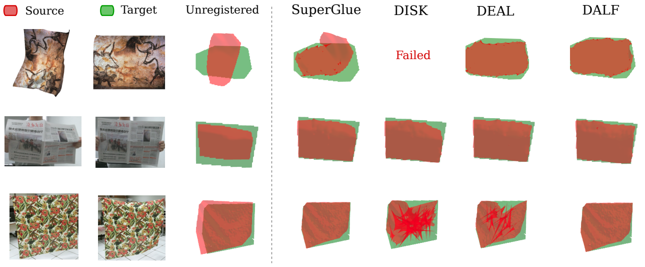

Fig. 7. **Non-rigid registration under challenging scenarios.** Our best method, DALF, can achieve accurate non-rigid registration under large rotations, illumination changes caused by deformations, and highly repetitive patterns. In contrast, state-of-the art matching techniques produce low-quality results in at least one of the challenging scenarios.

were *trained on synthetic data and tested on real images*. We compare our proposed descriptors' performance and that of other detect-and-describe methods against state-of-the-art approaches having available open-source implementation. To that end, we introduced to the community a dataset[1] of 11 deformable objects and 833 image pairs captured with Kinect v1 (manual annotation) and Kinect v2 (automatic annotation via OptiTrack). Dense correspondences are also provided via an intensity-based registration approach employing the sparse pixel-level ground-truth as anchors. In addition, we enhanced the deformable surface tracking (DeSurT) dataset [35] with dense correspondences.

Results & discussion. The performance of all methods is assessed using the top 2,048 keypoints. Figure 6 summarizes the main results, comparing the proposed descriptors with current state-of-the-art methods for keypoint detection and matching. We use the widely recognized metric *mean matching accuracy* (MMA), defined as $MMA = |\mathcal{S}_{gt}|/|\mathcal{K}_{gt}|$, which measures accuracy over successfully detected keypoints $\mathcal{K}_{gt}$ under a pixel threshold. In the bar plot, it can be seen that our proposed strategies significantly improve the quality of the matches in the presence of non-rigid deformations. Worth mentioning is the performance of DEAL and DALF, which only use RGB images and are able to surpass multi-modal approaches such as GeoPatch and GeoBit. This is because depth sensors come with significant noise that is hard to mitigate in practice. Fig. 7 shows registration results on three challenging sequences.

VI. Accelerated Features

In the previous sections, we addressed invariance and distinctiveness of local features under challenging transformations, introducing geodesic-aware and deformation-aware descriptors with competitive computational performance. However, real-time tasks like robot perception, autonomous driving, and applications on embedded devices (e.g., lightweight drones, IoT, augmented reality glasses) remain challenging due to the computational demands of deep-learning-based local feature methods.

Lightweight network. Motivated by the increasing availability of mobile hardware capable of running lightweight neural networks, we introduce XFeat [31], a lightweight CNN architecture offering sparse and semi-dense matching in a single versatile framework. XFeat combines a minimalist learnable keypoint detection branch with a novel match refinement module for efficient pixel-level correspondences without requiring high-resolution feature maps. Our architecture significantly improves the trade-off between computational efficiency and matching accuracy, running up to 5x faster than comparable lightweight methods while maintaining accuracy competitive with heavier models, enabling real-time deployment without specialized optimizations. An overview of the architecture is shown in Fig. 8.

Detection & description In our network, we design an efficient parallel branch for keypoint detection that leverages low-level features, first representing the image as 8×8 local grids of 64-dimensional embeddings and regressing keypoint positions via efficient 1×1 convolutions, yielding a keypoint embedding $\mathbf{K} \in \mathbb{R}^{H/8 \times W/8 \times (64+1)}$ for fine-grained, fast, and robust keypoint localization. Each local grid is responsible for localizing one keypoint inside it. For the description branch, a multi-scale feature pyramid merges representations at $\{1/8, 1/16, 1/32\}$ resolution to compute a dense but coarse descriptor map $\mathbf{F} \in \mathbb{R}^{H/8 \times W/8 \times 64}$ and a reliability map $\mathbf{R} \in \mathbb{R}^{H/8 \times W/8}$ via convolutional fusion, enhancing local feature robustness and feature matchability.

Dense matching. A lightweight dense matching module that selects top-K reliable regions by $\mathbf{R}_{i,j}$ and uses a multi-layer perceptron (MLP) for coarse-to-fine matching on $\mathbf{F}$ is proposed. Let $\mathbf{f}_a \in \mathbf{F_1}$ and $\mathbf{f}_b \in \mathbf{F_2}$ be two coarsely matched features obtained by traditional nearest neighbor matching from an image pair $(\mathbf{I_1}, \mathbf{I_2})$. We predict offsets $\mathbf{o} = \text{MLP}(\text{concat}(\mathbf{f}_a, \mathbf{f}_b))$, classifying the offset (x, y) that leads to the correct pixel-level match at original image resolution. The proposed stragegy significantly improves efficiency compared to current semi-dense matchers [36], [37].

Experiments. We consider sparse (XFeat) and semi-dense (XFeat*) setting using the same backbone; in the sparse case, up to 4,096 keypoints are selected based on $score = \mathbf{K}_{i,j} \cdot \mathbf{R}_{i,j}$, with local descriptors bicubically interpolated from $\mathbf{F}$, while XFeat* leverages most regions in $\mathbf{F}$ using the proposed dense matching approach. As shown in Fig. 9, XFeat is $5\times$ faster than the fastest alternative (ALIKE [38]), with competitive sparse matching on MegaDepth [39] and state-of-the-art performance on ScanNet [40] (please check some qualitative matching examples in Fig. 1), demonstrating superior efficiency and generalization with compact descriptors.

VII. Conclusion

In this work, we addressed the challenge of learning local image features that are robust to non-rigid deformations and viewpoint changes, with a focus on two key aspects: invariance and efficiency. Although large-scale foundation models have gained prominence, domain-specific applications remain constrained by limited computational resources and insufficient data. To mitigate these limitations, we introduced inductive biases mechanisms for deformation invariance using RGB-D

[1] Available at https://www.verlab.dcc.ufmg.br/descriptors

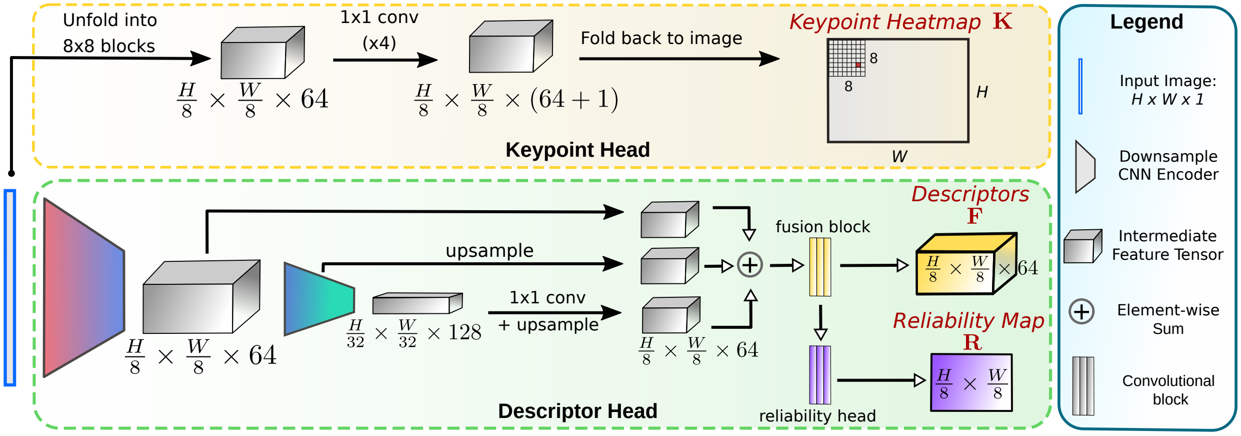
Unfold into
8x8 blocks
$\frac{H}{8} \times \frac{W}{8} \times 64$
1x1 conv
(x4)
$\frac{H}{8} \times \frac{W}{8} \times (64+1)$
Fold back to image
Keypoint Heatmap K
8
8
H
W
Keypoint Head
Legend
Input Image:
H x W x 1
Downsample
CNN Encoder
Intermediate
Feature Tensor
Element-wise
Sum
Convolutional
block
$\frac{H}{8} \times \frac{W}{8} \times 64$
upsample
$\frac{H}{32} \times \frac{W}{32} \times 128$
1x1 conv
+ upsample
$\frac{H}{8} \times \frac{W}{8} \times 64$
fusion block
Descriptors
F
$\frac{H}{8} \times \frac{W}{8} \times 64$
Reliability Map
R
$\frac{H}{8} \times \frac{W}{8}$
reliability head
Descriptor Head

Fig. 8. **Accelerated feature extraction network architecture.** XFeat extracts a keypoint heatmap $\mathbf{K}$, a compact 64-D dense descriptor map $\mathbf{F}$, and a reliability heatmap $\mathbf{R}$. It achieves unparalleled speed via early downsampling and shallow convolutions, followed by deeper convolutions in later encoders for robustness. Contrary to typical methods, it separates keypoint detection into a distinct branch, using 1×1 convolutions on an 8×8 tensor-block-transformed image for fast processing.

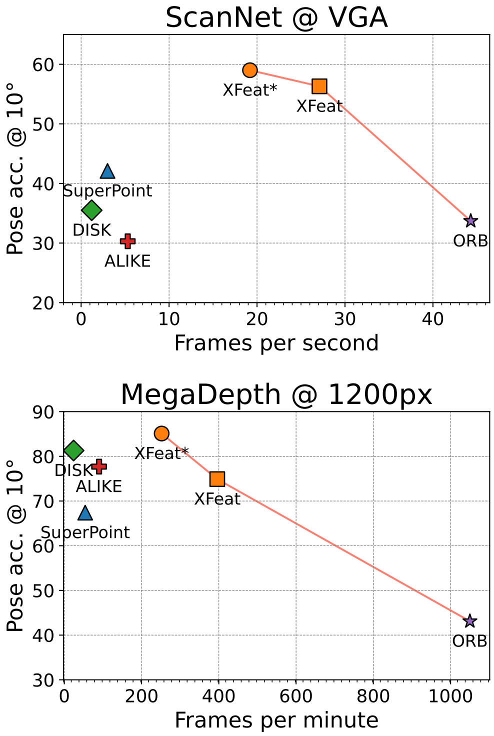

Fig. 9. **In XFeat, accuracy meets efficiency.** XFeat delivers great trade-off between speed and relative pose estimation accuracy on both the Megadepth and ScanNet datasets, as evidenced by the Pareto-frontier curve in orange. Its lightweight architecture enables real-time feature extraction on GPU-free settings and resource-constrained devices without hardware-specific optimizations. Inference speed on a budget-friendly laptop (`Intel(R) i5-1135G7 @ 2.40GHz CPU`). * denotes semi-dense extraction.

inputs, and further extended this approach to RGB-only settings via a deformation-aware module that learns spatial priors from data. We show that explicitly incorporating deformation awareness results in compact representations that remain robust across diverse tasks, even under synthetic supervision.

To further improve inference efficiency, we proposed *XFeat*, a lightweight CNN-based architecture that achieves a Pareto-optimal trade-off between accuracy and computational cost for local feature matching in several benchmarks. Its design facilitates deployment on mobile computers, enabling real-world deployment. Future research directions include geodesic-aware training without requiring depth at inference, learning from synthetic 3D datasets with enhanced realism, and weakly supervised training using image-level labels. Additional routes involve dynamic feature selection to balance distinctiveness and invariance, the development of more efficient learned matchers, and quantization strategies for real-time applications. We hope that our research can guide future advances toward robust, efficient, and scalable local feature extraction in scenarios where large models remain impractical.

Acknowledgments. We thank CAPES, CNPq, FAPEMIG, Google, and ANR (ANR-23-CE23-0003-01) for funding different portions of this work.

VIII. Awards & Publications

The results of this dissertation were published in the *Computer Vision and Image Understanding* (**CVIU**), and in four top-tier international computer vision and machine learning conferences (**ICCV'19**, **NeurIPS'21** and **CVPR'23&24**): *GeoBit* [18] (**ICCV 2019**), the geodesic-aware binary descriptor; *GeoPatch* [23] (**CVIU 2022**), the learning-based descriptor for geodesic sampling that also introduced an RGB-D benchmark with real and synthetic data for non-rigid matching evaluation.; *DEAL* [24] (**NeurIPS 2021**), an end-to-end architecture for non-rigid description; *DALF* [32] (**CVPR 2023**), a deformation-aware keypoint detector and descriptor; and *XFeat* [31] (**CVPR 2024**), a compact CNN for local feature extraction with significant speedup. This work received the Google Latin America Research Award. We also highlight that this PhD work has been selected as the best PhD work of the Department of Computer Science of UFMG and was appointed to the CAPES and UFMG thesis award in 2025.

References

[1] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, “Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam,” *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[2] G. Potje, G. Resende, M. Campos, and E. R. Nascimento, “Towards an efficient 3d model estimation methodology for aerial and ground images,” *Machine Vision and Applications*, vol. 28, pp. 937–952, 2017.

[3] J. L. Schonberger and J.-M. Frahm, “Structure-from-motion revisited,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[4] M. Teichmann, A. Araujo, M. Zhu, and J. Sim, “Detect-to-retrieve: Efficient regional aggregation for image search,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5109–5118.

[5] ——, “Detect-to-retrieve: Efficient regional aggregation for image search,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5109–5118.

[6] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *IJCV*, pp. 91–110, 2004.

[7] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua, “Brief: Computing a local binary descriptor very fast,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1281–1298, 2012.

[8] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “ORB: an efficient alternative to SIFT or SURF,” in *ICCV*, Barcelona, 2011.

[9] K. Mikolajczyk and C. Schmid, “Scale & affine invariant interest point detectors,” *International journal of computer vision*, vol. 60, pp. 63–86, 2004.

[10] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05)*, vol. 1. Ieee, 2005, pp. 886–893.

[11] E. Simo-Serra, C. Torras, and F. Moreno-Noguer, “DaLI: deformation and light invariant descriptor,” *International Journal of Computer Vision*, vol. 115, no. 2, 2015.

[12] A. E. Johnson and M. Hebert, “Efficient multiple model recognition in cluttered 3-d scenes,” in *Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No. 98CB36231)*. IEEE, 1998, pp. 671–677.

[13] M. M. Bronstein and I. Kokkinos, “Scale-invariant heat kernel signatures for non-rigid shape recognition,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 1704–1711.

[14] I. Kokkinos, M. M. Bronstein, R. Litman, and A. M. Bronstein, “Intrinsic shape context descriptors for deformable shapes,” in *CVPR*, June 2012, pp. 159–166.

[15] E. R. Nascimento, G. L. Oliveira, M. F. M. Campos, A. W. Vieira, and W. R. Schwartz, “BRAND: A Robust Appearance and Depth Descriptor for RGB-D Images,” in *Proc. IROS*, 2012.

[16] A. Zaharescu, E. Boyer, K. Varanasi, and R. P. Horaud, “Surface Feature Detection and Description with Applications to Mesh Matching,” in *CVPR*, Miami Beach, Florida, June 2009.

[17] F. Tombari, S. Salti, and L. D. Stefano, “A combined texture-shape descriptor for enhanced 3D feature matching,” in *ICIP*, 2011.

[18] E. R. Nascimento, G. Potje, R. Martins, F. Cadar, M. F. Campos, and R. Bajcsy, “Geobit: A geodesic-based binary descriptor invariant to non-rigid deformations for rgb-d images,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 004–10 012.

[19] D. P. Vassileios Balntas, Edgar Riba and K. Mikolajczyk, “Learning local feature descriptors with triplets and shallow convolutional neural networks,” in *Proceedings of the British Machine Vision Conference (BMVC)*, 2016.

[20] Y. Tian, B. Fan, and F. Wu, “L2-net: Deep learning of discriminative patch descriptor in euclidean space,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 661–669.

[21] A. Mishchuk, D. Mishkin, F. Radenovic, and J. Matas, “Working hard to know your neighbor’s margins: Local descriptor learning loss,” in *Advances in Neural Information Processing Systems*, 2017, pp. 4826–4837.

[22] P. Ebel, A. Mishchuk, K. M. Yi, P. Fua, and E. Trulls, “Beyond cartesian representations for local descriptors,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 253–262.

[23] G. Potje, R. Martins, F. Cadar, and E. R. Nascimento, “Learning geodesic-aware local features from rgb-d images,” *Computer Vision and Image Understanding*, vol. 219, p. 103409, 2022.

[24] G. Potje, R. Martins, F. Chamone, and E. Nascimento, “Extracting deformation-aware local features by learning to deform,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 10 759–10 771, 2021.

[25] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, “LIFT: Learned invariant feature transform,” in *ECCV*, 2016.

[26] 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.

[27] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, “D2-net: A trainable cnn for joint detection and description of local features,” *arXiv preprint arXiv:1905.03561*, 2019.

[28] J. Revaud, C. De Souza, M. Humenberger, and P. Weinzaepfel, “R2D2: Reliable and repeatable detector and descriptor,” in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 12 405–12 415.

[29] Z. Luo, L. Zhou, X. Bai, H. Chen, J. Zhang, Y. Yao, S. Li, T. Fang, and L. Quan, “Aslfeat: Learning local features of accurate shape and localization,” in *CVPR*, 2020, pp. 6589–6598.

[30] 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.

[31] G. Potje, F. Cadar, A. Araujo, R. Martins, and E. R. Nascimento, “Xfeat: Accelerated features for lightweight image matching,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2682–2691.

[32] ——, “Enhancing deformable local features by jointly learning to detect and describe keypoints,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1306–1315.

[33] K. Crane, C. Weischedel, and M. Wardetzky, “Geodesics in heat: A new approach to computing distance based on heat flow,” *ACM Transactions on Graphics (TOG)*, vol. 32, no. 5, pp. 1–11, 2013.

[34] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, “Spatial transformer networks,” *arXiv preprint arXiv:1506.02025*, 2015.

[35] T. Wang, H. Ling, C. Lang, S. Feng, and X. Hou, “Deformable surface tracking by graph matching,” in *IEEE International Conference on Computer Vision*, 2019.

[36] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, “Loftr: Detector-free local feature matching with transformers,” in *CVPR*, 2021, pp. 8922–8931.

[37] Y. Wang, X. He, S. Peng, D. Tan, and X. Zhou, “Efficient loftr: Semi-dense local feature matching with sparse-like speed,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 21 666–21 675.

[38] X. Zhao, X. Wu, J. Miao, W. Chen, P. C. Chen, and Z. Li, “Alike: Accurate and lightweight keypoint detection and descriptor extraction,” *IEEE TMM*, 2022.

[39] 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.

[40] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, “Scannet: Richly-annotated 3d reconstructions of indoor scenes,” in *CVPR*, 2017, pp. 5828–5839.