An evaluation of recent local image descriptors for real-world applications of image matching

An Evaluation of Recent Local Image Descriptors for Real-World Applications of Image Matching

Fabio Bellavia and Carlo Colombo
Università degli Studi di Firenze, via di S. Marta n.3, 50124, Firenze, Italy
{fabio.bellavia, carlo.colombo}@unifi.it

Abstract

This paper discusses and compares the best and most recent local descriptors, evaluating them on increasingly complex image matching tasks, encompassing planar and non-planar scenarios under severe viewpoint changes. This evaluation, aimed at assessing descriptor suitability for real-world applications, leverages the concept of Approximated Overlap error as a means to naturally extend to non-planar scenes the standard metric used for planar scenes. According to the evaluation results, most descriptors exhibit a gradual performance degradation in the transition from planar to non-planar scenes. The best descriptors are those capable of capturing well not only the local image context, but also the global scene structure. Data-driven approaches are shown to have reached the matching robustness and accuracy of the best hand-crafted descriptors.

1 Introduction

Local image descriptors constitute the basic layer of almost all computer vision applications dealing with point correspondences among several images, encompassing object detection [16], image stitching [7], 3D reconstruction [27] and visual odometry [12]. This has ensured that the topic remained well alive through the years, up to the recent advances on both hand-crafted and data-driven descriptors. The latter, which leverage deep learning progress, availability of big data and modern hardware capabilities, are yielding particularly promising results.

Several factors influencing descriptor performance must be taken into account for developing practical applications. These factors include the nature of scene content, the image transformations involved, the computational constraints, the requirements in matching accuracy and robustness. Concurrently with the evolution of descriptor design, better evaluation benchmarks that can expose both potential strengths and weaknesses of descriptors are called for. In particular, adaptability to non-planar scene content and relevant viewpoint changes are the main aspects to consider when defining an effective descriptor evaluation benchmark, as they reflect the most general real-world environment.

Well-consolidated benchmarks exist for the evaluation of planar scenes, from the standard Oxford benchmark [18, 20] to the more recent HPatches [2]. Here, the overlap error between local descriptor patches and their re-projections is used as the error metric, while ground-truth (GT) information consists just in the homograhpy transformation between the input images, which can be estimated in a very accurate and easy way. Nevertheless, evaluation on planar scenes provides only a limited insight into overall descriptor properties. In order to overcome this limitation, benchmarks exploiting directly or indirectly non-planar environments have been devised. In the former case, GT is directly estimated (a) using stereo matching [14] or Structure-from-Motion [25], (b) through complex sensor-based system setups [10, 28], or (c) according to some approximation scheme [5, 23]. On the other hand, indirect evaluation of local image descriptors is done (d) by checking the correctness of the output for a given specific application task, such as object retrieval [11] or visual odometry [6]. All these solutions have some drawbacks: GT may not be available for some image region (a,b), can be erroneously estimated (a,c,d) or biased towards the considered application (a,d). For example, SIFT was found to give the best results on evaluations based on Structure-from-Motion pipelines, that are usually built and optimized over SIFT itself [25].

In this paper, a comparative evaluation of the best recent local descriptors is carried out, focusing on image matching tasks. Test images include both planar and non-planar scenes, the latter being particularly effective at assessing descriptor suitability for practical applications. Descriptor performance with planar scenes is evaluated in terms of overlap error. For non-planar scenes, the Approximated Overlap error (AO) metric introduced in [5] was chosen for two main reasons. First, AO takes into account the whole local descriptor patch, thus representing a natural extension of the overlap error to the more complex non-planar case. Second, AO was shown to give a very low false positive rate in GT estimation, thereby not affecting descriptor ranking order, and to avoid the bias issues experimented with recent setups.

The rest of the paper is organized as follows. Recent state-of-the-art local image descriptors are reviewed in Sec. 2. The planar and non-planar evaluation setups and datasets are described in Secs. 3 and 4, respec-

tively. Comparative experimental results are discussed in Sec. 5. Conclusions and future work are outlined in Sec. 6.

2 Recent State-of-the-Art Local Descriptors

Nowadays, there are mainly two ways of classifying local image descriptors. The first way is to consider whether the descriptor uses a priori data knowledge and is trained according to some machine learning approach. If this is the case, the descriptor is termed data-driven, otherwise it is termed hand-crafted. The second way is to divide descriptors by the data type used to represent their vector elements. Specifically, if a single bit per element is used, the descriptor is referred to as binary, and non-binary otherwise. Binary descriptors are usually less robust yet faster and more compact than non-binary ones.

Scale-Invariant Feature Transform (**SIFT**) [16] is a quite popular and valid hand-crafted, non-binary descriptor, generally used as baseline for benchmark evaluations. SIFT is obtained as the concatenation of the Gaussian-weighted gradient histograms associated to the regions into which the keypoint patch is divided, after being rotated towards the dominant gradient orientation. In the attempt to improve its robustness, several SIFT extensions has been proposed over the years. Among these, **RootSIFT** [1] and the doubled shifting Gradient Local Orientation Histogram [3] , in both its non-binary (**sGLOH2**) and binary (**BisGLOH2**) versions, are considered in the proposed evaluation. RootSIFT improves SIFT by employing the Hellinger distance instead of the Euclidean distance. sGLOH2 and BisGLOH2 are more effective at handling patch rotations.

Other hand-crafted descriptors considered for the evaluation are Local Intensity Order Pattern (**LIOP**) [31], employing intensity order pooling and histograms computed on the relative order of neighbor pixels to achieve rotation invariance, and the Multiple-Kernel Local-Patch Descriptor (**MKD**) [24], using alternative kernels for defining histograms. For both descriptors, optimized data-driven versions exist, exploiting among others Principal Component Analysis (PCA) to achieve better matching results while reducing the associated vector dimensions. These descriptors, denoted respectively as Mixed Intensity Order Pattern (**MIOP**) [31] and $\mathbf{MKD}_W$, will also be taken into account.

Binary data-driven descriptors have also been proposed. In particular, Receptive Field Descriptor (RFD) [9] thresholds regions of the patch gradient map, where threshold values, positions and sizes of the patch regions are learned from training data. Two different RFDs have been included in the evaluation, namely $\mathbf{RFD}_R$ and $\mathbf{RFD}_G$, making use of rectangular and Gaussian regions, respectively.

Deep descriptors are also considered for the comparative evaluation. This kind of data-driven descriptors is built upon Convolutional Neural Network (CNN) architectures, generally exploiting triplet loss and hard negative mining for optimization at the training stage. **DeepDesc** [26], **L2-Net** [29] and **HardNet++** [21] have been included together with some variants. In particular, **BiL2-Net** and $\mathbf{L2\text{-}Net}_{CS}$ denote respectively the binary and center-symmetric versions of L2-Net [29]. **HardNetPS** [22] will be evaluated too. It employs an alternative massive patch dataset for training, aiming at overcoming the lack of generalization ability, as consequence of data insufficiency, common to all learning-based approaches. The very recent **GeoDesc** [17], also in its quantized form here denoted as $\mathbf{GeoDesc}_Q$, is also included in the comparison. Differently from the previous approaches, this descriptor also exploits geometric information for network training.

3 Planar evaluation setup

The setup follows the guidelines described in [19] with slight changes. More in detail, given two images I_1 and I_2 of the same planar scene, keypoints are extracted using the HarrisZ detector [4]. Descriptors are then computed from the corresponding normalized patches, matched in a pairwise way, and sorted according to the Nearest Neighbor Ratio (NNR) [16]. The GT homography $H_{2\to1}$ relates points $\mathbf{x}_1 \in I_1$ and $\mathbf{x}_2 \in I_2$ so that $\mathbf{x}_1 = H_{2\to1}\mathbf{x}_2$ in homogeneous coordinates. The overlap error between two generic regions A and B of the same image is defined as

$$\epsilon(A, B) = 1 - \frac{A \cap B}{A \cup B} \tag{1}$$

A match is considered correct if $\epsilon(E_1, E_{2\to1}) \leq t$, where $t = 0.5$ is a given threshold, $E_1 \in I_1$ and $E_2 \in I_2$ are the two elliptical patches corresponding respectively to the matching pair elements, and $E_{2\to1}$ is the projection of E_2 onto I_1 through $H_{2\to1}$. Notice that, differently from the original implementation in [19], where the overlap error is computed by finite approximations, the *exact* analytical solution described in [15] is used in this work for the computation of the overlap error.

Images from the Oxford [18] and Viewpoint [32] datasets were used. GT homographies are provided as part of the datasets. The Oxford dataset [18] contains 13 sequences of 6 images. Each sequence shows a planar scene undergoing one specific transformation among the following: Scale plus rotation, image blur, illumination, JPEG compression and viewpoint changes. The Viewpoint dataset contains only 6 images for 5 different scenes with various viewpoint changes. Image pairs are generated by setting the first image of each sequence as reference I_1, and using one of the remaining 5 images as I_2, for a total of $(13 + 5) \times 5 = 90$ image pairs.

Results are compared in terms of mean Average Precision (mAP), computed as in [8]. The average

Figure 1. Sample image pairs from the (*top row*) planar, (*middle row)* viewpoint only, and (*bottom row*) non-planar datasets used in the evaluation.

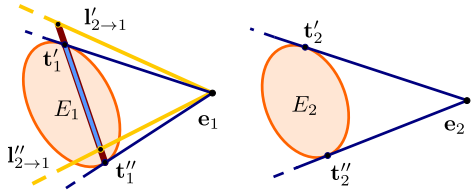

Figure 2. Linear overlap error construction (best viewed in color).

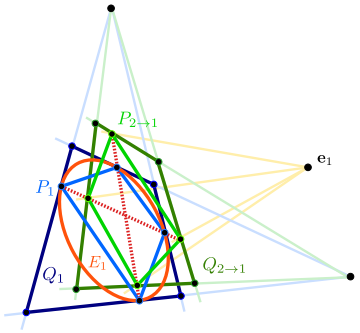

Figure 3. AO construction (best viewed in color).

mAP over all the planar image pairs is considered. In addition, results with the subset obtained by selecting only the image pairs subjected to viewpoint changes (referred to as "viewpoint only," containing $7 \times 5 = 35$ pairs) are reported. This subset represents the most relevant and challenging kinds of image distortion. Two sample image pairs for the planar setup are shown in the first two rows of Fig. 1.

4 Non-Planar evaluation setup

Non-planar evaluation follows the same approach adopted for the planar case, only replacing the overlap error with the Approximated Overlap error (AO) [5]. AO extends to surfaces the linear overlap error introduced in [13], defined hereafter for completeness. As shown in Fig. 2, any tangency relation between the epipole and a given ellipse is preserved under perspective projection (blue). The linear overlap error is defined as the ratio between the small (light blue) and the wider (red) segments, both lying on the line through the points $\mathbf{t}'_1$ and $\mathbf{t}''_1$ where the tangents from the epipole $\mathbf{e}_1$ meet the ellipse E_1. The epipolar lines $\mathbf{l}'_{2\to1}$ and $\mathbf{l}''_{2\to1}$ in I_1 (yellow) correspond to the points $\mathbf{t}'_2$ and $\mathbf{t}''_2$ in I_2 where the tangents from the epipole $\mathbf{e}_2$ meet the ellipse E_2.

AO extends the linear overlap error by observing that, in addition to epipoles, the correspondences employed for computing the GT fundamental matrix are usually available. Given two of such correspondences (see Fig. 3), the associated four tangent points on the ellipse E_1 (orange) define the inscribed P_1 (light blue) and circumscribed Q_1 (dark blue) quadrilaterals. Analogously to the case of the linear overlap error, the quadrilaterals $P_{2\to1}$ (light green) and $Q_{2\to1}$ (dark green) can be constructed through the epipolar mapping from points in I_2 to lines in I_1. AO is defined as

$$\varepsilon = \frac{\epsilon(P_1, P_{2\to1}) + \epsilon(Q_1, Q_{2\to1})}{2} \tag{2}$$

under the assumption that the scene can be approximated by piecewise planar patches. For further computational details see [5].

If ellipse E_1 correctly matches with E_2, but not with E'_2, a false positive may nevertheless arise when E_2 and E'_2 share, either exactly or approximately, the same tangent lines through the epipole $\mathbf{e}_1$. Even if AO has been shown to give a very low false positive rate (less than 5%), that does not affect descriptor ranking in unsupervised evaluations [5], hereafter a heuristic is introduced that further decreases false positive matches. Let I_1 and I_2 be $m \times n$ pixel images, and $\mathbf{c}_1$ and $\mathbf{c}_2$ the centers of E_1 and E_2, with flow length $\| \mathbf{c}_1 - \mathbf{c}_2 \|$. Consider the set F of flow lengths relative to all the ellipses in I_1 whose centers are inside a radius of $\min(m, n)/15$ from $\mathbf{c}_1$. A match is discarded if

$$\| \mathbf{c}_1 - \mathbf{c}_2 \| > \mu + 2.5\sigma \tag{3}$$

where μ and σ are the median and Median Absolute Deviation (MAD) over F, respectively. The same process is repeated by working on I_2 and switching the

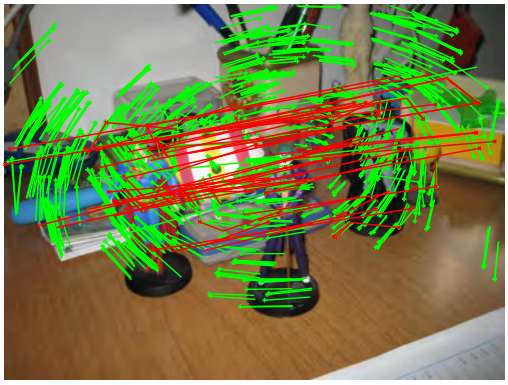

Figure 4. Flow vectors of the retained (green) and discarded (red) matches using the proposed heuristic for the sample non-planar image pair of Fig. 1 (best viewed in color).

roles of E_1 and E_2. An example of wrong matches discarded by the proposed heuristic is shown in Fig. 4.

For the non-planar evaluation, experimental results were obtained with the dataset introduced in [3]. This dataset is made up of 42 different image pairs of non-planar scenes exhibiting various degrees of viewpoint changes (a sample image pair is shown in the last row of Fig. 1). GT fundamental matrices for epipolar transfer and correspondences for constructing approximated quadrilaterals are provided by the authors. As for the planar case, the AO threshold is set to $t = 0.5$ and results are reported in terms of average mAP.

5 Results

Results for all of the state-of-the-art descriptors referred to in Sec. 2, some of which are quite recent, are reported in Table 1. For each descriptor under test, the table also lists the following characteristics: (1) matching distance (L_1, L_2, Hamming, or dot product), (2) class (hand-crafted or data-driven), (3) rotational invariance, (4) vector dimension and data type, (5) bibliographic reference. The choice of the L_1 distance for SIFT, RootSIFT and LIOP may appear unusual, as these descriptors are typically matched according to the L_2 distance. Nevertheless, our experiments confirmed the result found in [3] that these hand-crafted descriptors perform better with L_1 than with L_2. For descriptors that are not rotationally invariant , local image patches were rotated according to the SIFT dominant gradient orientation using the VLFeat [30] implementation. The freely available code from [3] was used for the computation of the overlap error and AO. For all descriptors, with the exception of SIFT and RootSIFT employing the VLFeat implementation, the code from their respective authors is used. Notice also that for the sake of clarity, Table 1 refers to sGOr2h* and BisGOr2h* matching strategies [3] as sGLOH2 and BisGLOH2, respectively.

According to the results, mAP decreases in the transition from planar through viewpoint to non-planar scenes, and are well aligned with those reported for the HPatches dataset, with respectively the easy, hard and tough setups [17].

GeoDesc and GeoDesc$_Q$ achieve the best results for any setup, closely followed by sGLOH2 and its binarized counterpart BisGLOH2, with HardNet++ and HardNetPS ranked after them. Comparing GeoDesc against Geodesc$_Q$, quantization does not seem to affect the matching robustness, while it provides a faster and compact descriptor. HardNet++ performs better than HardNetPS for non-planar scenes, while the opposite happens in the case of planar scenes, underlining the strict and critical dependency of deep descriptors from training data. Among the evaluated descriptors, only GeoDesc, Geodesc$_Q$, sGLOH2 and BisGLOH2 exploit the spatial geometric structure in the image. Being data-driven, GeoDesc is learned a priori according to this kind of information, while sGLOH2 and BisGLOH2 use it explicitly at runtime time thanks to their matching strategies, that behave like statistical filters on the data. Negative mining techniques, employed by deep descriptors, also seem to be able to implicitly extract the image statistical context.

Concerning the remaining descriptors, LIOP and MIOP boost their performance in the planar case, while results become comparable to those provided by RFD$_R$, RFD$_G$, SIFT, RootSIFT, L2-Net, MKD and MKD$_W$ in the non-planar case. L2-Net$_{CS}$ and BiL2-Net$_{CS}$ exhibit the opposite behavior. Analogously, DeepDesc behaves nearly as the worst in the planar case, save when only viewpoint transformations are considered or in the non-planar case, for which results are well aligned with the others. This can be again due to the training dataset and approach employed by DeepDesc. Comparing LIOP and MKD against MIOP and MKD$_W$, respectively, PCA provides only little improvements in terms of matching, but can greatly reduce descriptor dimensions, thus improving efficiency. BiL2-Net is the one with the worst performance in this evaluation. However, considering its strictly limited descriptor vector length, BiL2-Net$_{CS}$ can be useful for applications dealing with non-complex images and requiring fast matching.

The proposed evaluation does not report any analysis about running times, since these are quite dependent from the hardware and software implementations (e.g. CPU, SIMD, GPU). However, descriptor vector total byte length, that can be derived by descriptor dimension and data type, is in general sufficient to outline computational requirements at matching time. According to this assumption, binary descriptors are faster than the others, while `float` type descriptors are the slowest on optimized implementations. Notice, however, that this discussion does not hold for the

sGLOH2 and BisGLOH2 descriptors, whose matching strategies are different from the others and very time consuming [3].

6 Conclusion and Future Work

This paper compared recent state-of-the-art local image descriptors for real-world matching applications, thanks to the concept of Approximated Overlap error as a means to naturally extend the analysis from planar to non-planar scenarios, without introducing biases as it happened with other recent evaluations.

Overall, most descriptors exhibit a gradual performance degradation in the transition from planar, through viewpoint, to non-planar scenes. The best descriptors are those capable of capturing well not only the local image context, but also the global scene structure. Indeed, it seems that descriptors are now very close to reach their allowable discriminability power when considered alone. Injecting global scene knowledge, either a priori or at matching time, and either implicitly or explicitly, can be the next step to look for better solutions in the field. According to the evaluation results, data-driven approaches are so matured as to have reached and even surpassed the matching robustness and accuracy of the best hand-crafted descriptors. Nevertheless, training data still remain a crucial aspect to be considered, in particular when descriptors have to be designed for very specific application domains.

Future work will include an expansion of the non-planar dataset, and the analysis of further descriptor properties omitted in this paper, such as implementation issues and computational times.

Acknowledgment

The Titan Xp used for this research was donated by the NVIDIA Corporation.

Table 1. Evaluation results

		hand-crafted	rotationally invariant	mAP (%) planar	mAP (%) viewpoint only	mAP (%) non-planar	dim	type	refs
L_1	SIFT	✓		63.93	47.48	37.58	128	`uchar`	[16]
	RootSIFT	✓		63.71	49.09	38.88	128	`float`	[1]
	LIOP	✓	✓	74.11	55.22	39.52	144	`uchar`	[31]
	sGLOH2	✓	✓	75.64	63.51	50.68	256	`uchar`	[3]
L_2	MIOP		✓	76.36	57.02	40.54	128	`float`	[31]
	DeepDesc			55.38	47.84	38.35	128	`float`	[26]
	L2-Net			59.91	48.62	43.00	128	`float`	[29]
	L2-Net$_{CS}$			67.00	54.64	48.12	256	`float`	[29]
	HardNet++			70.73	58.37	47.54	128	`float`	[21]
	HardNetPS			73.94	59.86	45.77	128	`float`	[22]
	GeoDesc			78.75	65.10	51.51	128	`float`	[17]
	GeoDesc$_Q$			78.78	65.03	51.53	128	`uchar`	[17]
H	RFD$_R$			68.26	54.13	38.48	293	bit	[9]
	RFD$_G$			68.77	55.63	40.25	406	bit	[9]
	BiL2-Net			48.70	36.58	34.33	128	bit	[29]
	BiL2-Net$_{CS}$			61.42	49.35	43.31	256	bit	[29]
	BisGLOH2	✓	✓	74.26	61.49	49.31	1152	bit	[3]
$*$	MKD	✓		62.84	48.64	40.10	128	`float`	[24]
	MKD$_W$			62.65	48.89	40.67	238	`float`	[24]

hand-crafted · rotationally invariant · refs · $*$ dot product · planar · viewpoint only · non-planar · H Hamming distance

References

[1] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2911–2918, 2012.

[2] V. Balntas, K. Lenc, A. Vedaldi, and K. Mikolajczyk. HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3852–3861, 2017.

[3] F. Bellavia and C. Colombo. Rethinking the sGLOH descriptor. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):931–944, 2018.

[4] F. Bellavia, D. Tegolo, and C. Valenti. Improving Harris corner selection strategy. *IET Computer Vision*, 5(2):86–96, 2011.

[5] F. Bellavia, C. Valenti, C. A. Lupascu, and D. Tegolo. Approximated overlap error for the evaluation of feature descriptors on 3D scenes. In *Proceedings of the International Conference on Image Analysis and Processing (ICIAP)*, pages 270–279, 2013.

[6] J. Bian, L. Zhang, Y. Liu, W. Y. Lin, M. M. Cheng, and I. D. Reid. MatchBench: An evaluation of feature matchers. In *arXiv*, 2018.

[7] M. Brown and D. G. Lowe. Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision*, 74(1):59–73, Aug 2007.

[8] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. volume 88, pages 303–338, 2010.

[9] B. Fan, Q. Kong, T. Trzcinski, Z. Wang, C. Pan, and P. Fua. Receptive fields selection for binary feature description. *IEEE Transactions on Image Processing*, 26(6):2583–2595, 2014.

[10] B. Fan, Q. Kong, X. Wang, Z. Wang, S. Xiang, C. Pan, and P. Fua. A performance evaluation of local features for image based 3D reconstruction. In *arXiv*, 2018.

[11] B. Fan, F. Wu, and Z. Hu. Rotationally invariant descriptors using intensity order pooling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10):2031–2045, 2012.

[12] M. Fanfani, F. Bellavia, and C. Colombo. Accurate keyframe selection and keypoint tracking for robust visual odometry. *Machine Vision and Applications*, 27(6):833–844, 2016.

[13] P. Forssén and D.G. Lowe. Shape descriptors for maximally stable extremal regions. In *Proceedings of the International Conference on Computer Vision*, pages 1–8, 2007.

[14] F. Fraundorfer and H. Bischof. A novel performance evaluation method of local detectors on non-planar scenes. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 33–33, 2005.

[15] G. B. Hughes and M Chraibi. Calculating ellipse overlap areas. *Computing and Visualization in Science*, 15(5):291–301, 2012.

[16] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[17] Z. Luo, T. Shen, L. Zhou, S. Zhu, R. Zhang, Y. Yao, T. Fang, and L. Quan. Geodesc: Learning local descriptors by integrating geometry constraints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[18] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.

[19] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1-2):43–72, 2005.

[20] O. Miksik and K. Mikolajczyk. Evaluation of local detectors and descriptors for fast feature matching. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 2681–2684, 2012.

[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 30: Annual Conference on Neural Information Processing Systems (NIPS)*, pages 4829–4840, 2017.

[22] R. Mitra, N. Doiphode, U. Gautam, S. Narayan, S. Ahmed, S. Chandran, and A. Jain. A large dataset for improving patch matching. In *arXiv*, 2018.

[23] P. Moreels and P. Perona. Evaluation of features detectors and descriptors based on 3D objects. *International Journal of Computer Vision*, 73(3):263–284, 2007.

[24] A. Mukundan, G. Tolias, and O. Chum. Multiple-kernel local-patch descriptor. In *British Machine Vision Conference (BMVC)*, 2017.

[25] J. L. Schönberger, H. Hardmeier, T. Sattler, and M. Pollefeys. Comparative evaluation of hand-crafted and learned local features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[26] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[27] N. Snavely, S.M. Seitz, and R. Szeliski. Modeling the world from internet photo collections. *International Journal of Computer Vision*, 80(2):189–210, 2008.

[28] C. Strecha, W. von Hansen, L. J. Van Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[29] Y. Tian, B. Fan, and F. Wu. L2-Net: deep learning of discriminative patch descriptor in euclidean space. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6128–6136, 2017.

[30] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/`, 2008.

[31] Z. Wang, B. Fan, G. Wang, and F. Wu. Exploring local and overall ordinal information for robust feature description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11):2198–2211, 2016.

[32] K.M. Yi, Y. Verdie, P. Fua, and V. Lepetit. Learning to assign orientations to feature points. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2016.