[Supplementary Material] Enhancing Deformable Local Features by Jointly Learning to Detect and Describe Keypoints

Guilherme Potje[1] Felipe Cadar[1] André Araujo[2]
Renato Martins[3,4] Erickson R. Nascimento[1,5]

[1]Universidade Federal de Minas Gerais [2]Google Research
[3]Université de Bourgogne [4]Université de Lorraine, LORIA, Inria [5]Microsoft
{guipotje,cadar,erickson}@dcc.ufmg.br, renato.martins@u-bourgogne.fr, andrearaujo@google.com

In this supplementary material to our paper, we provide additional details about the retrieval and non-rigid 3D surface tracking applications, as well as further qualitative, quantitative results and visualization of the results presented in the paper. Please also have a look in our video showing detailed views of the non-rigid 3D surface registration for different approaches.

1. Deformable object retrieval

The nonrigid dataset contains various sequences of different objects being deformed over time. We selected one frame for each sequence to serve as a query image. The other frames of all sequences compose the search database, from where the application must retrieve the results.

For each method, we detect and describe a maximum of 1,024 keypoints inside a mask delimiting the object pixels. Next, we sample an equal amount of descriptors for each image to collect about 10,000 descriptors. Then we use the sampled descriptors to compute 300 centroids using the K-Means algorithm. The centroids are then used to calculate one global representation for each image using the Bag-of-Visual-Words approach to aggregate all the described keypoints. Given a query, we use the global descriptor to retrieve the closest K images using K-Nearest Neighbors. We evaluate each method with the mean retrieval accuracy for each value of K from 1 to 20.

We compare our method against the best-performing description methods, in addition to DELF [10], a state-of-the-art descriptor designed and trained specifically for image retrieval. DALF achieved the best performance in the retrieval task, as shown in Figure 1. Note that at $K = 10$, all methods achieve similar scores because they can correctly retrieve the easy images. However, note that the task becomes hard when $K > 10$, where all methods but DALF degrade as they cannot reliably retrieve the images of the heavily deformed objects, while DALF exhibits superior performance. The full retrieval result for each query seen in Figure 2. The code for the retrieval task will be publicly available; its objective is to be an easy-to-run benchmark for detectors and descriptors.

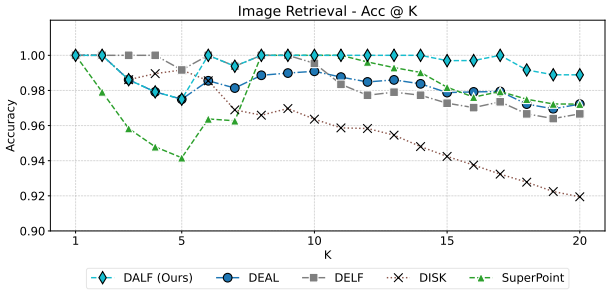

Figure 1. Accuracy@K metric for the nonrigid object retrieval task. The normalized area-under-the-curve for each method is the following DISK: 96.12%, SuperPoint: 97.92%, DEAL: 98.34%. DELF: 98, 57%, and DALF (Ours): 99, 49%.

2. Additional quantitative results

In this section, we report additional metrics beyond the matching scores and mean matching accuracy, and present a more detailed analysis of the ablation study.

Keypoint Repeatability. Recent methods [5, 11], do not report keypoint repeatability because it often does not correlate well with downstream performance. Nevertheless, we computed repeatability across all the datasets, and our approach obtains the best average repeatability across all datasets. The scores are the following. DALF: 0.58, DISK: 0.57, non-rigid detector [7]: 0.47, SIFT: 0.41, and R2D2: 0.35.

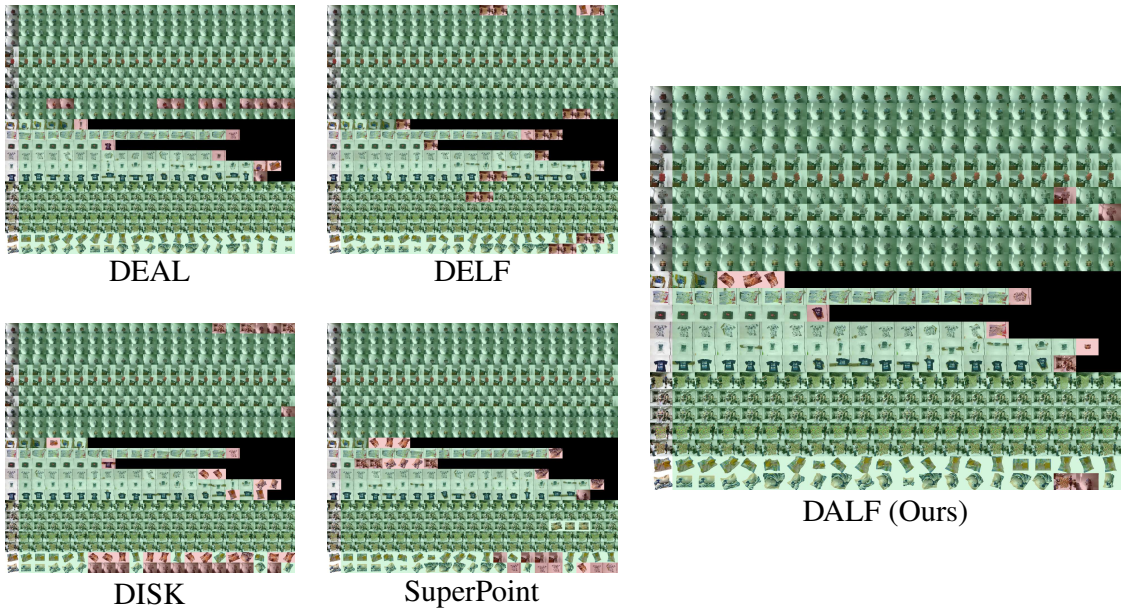

Figure 2. Our method has the best result in retrieving images of real and simulated deformed objects. The first column of each image shows the object queries, and the rows show the results from different queries. Green images correspond to the same object as the query, and red images do not correspond. Some objects are smaller and difficult to deform, so they may have less them 20 occurrences in the dataset. In that case, we lower the value of K to the exact number of occurrences of the object. For this reason, we can see some empty squares in the qualitative results. The black squares indicate no correspondent objects available.

Table 1. **Extended ablation.** Matching score @ 3 pixels for each configuration C following the order of Tab. 2 of the main paper, *e.g.*, C1 corresponds to distinct-only, C2 to invariant-only, etc. Best result in red, second best in green, third best in blue.

Dataset	C1	C2	C3	C4	C5
Kinect1	0.58	0.52	0.53	0.55	0.54
Kinect2	0.54	0.56	0.60	0.63	0.62
DeSurT	0.48	0.43	0.46	0.50	0.49
Simulation	0.27	0.52	0.50	0.34	0.42

Extended ablation analysis. Although the two-stage training is not mandatory in our learning pipeline, it offers a better trade-off between invariance and distinctiveness, as shown in the top 3 performances on every dataset according to Tab. 1 (which presents the scores per dataset from Tab. 2 of the main paper), thus we opt for C5 as the final design choice. Note that the fusion of the invariant and distinct features (C3–5), one of our novel contributions, achieves much better rankings on average across all datasets.

3. Non-rigid 3D surface registration

In this section, we describe in detail the implementation of the surface registration application using the as-rigid-as-possible (ARAP) [9] optimization, and also show qualitative results derived from the experiments of Tab. 3 (3D surface registration) of the paper.

Non-rigid 3D surface registration aims to accurately align two RGB-D frames of the same surface, viewed from different viewpoints at the same time that the object is affected by non-rigid deformations. Figure 3 shows an overview of the registration pipeline. Surface alignment is a crucial step used by non-rigid reconstruction frameworks [1,3] that allow complete 3D reconstruction of deforming objects. Improvements in registration accuracy can significantly increase the quality of the reconstruction, enabling the use of such systems in critical, challenging applications, such as the live reconstruction of human organs [6].

3.1. Implementation details

Our application considers the most difficult scenario: wide-baseline registration, where the object can be in an arbitrary viewpoint and deformed shape. Thus, it is challenging to filter outlier matches, in contrast with rigid registration, where it is possible to fit a homography or fundamental matrix using a minimal correspondence sample and perform RANSAC to remove the outlier correspondences with high confidence.

Our solution to this problem was to tune the AdaLAM [2] filtering method to perform outlier detection in the presence of image deformations. AdaLAM checks the affine consistency of local point clusters and filters the correspondences that are inconsistent with their neighboring matches. As we have observed empirically, the assumption of localized affine consistency is a reasonable approximation for non-rigid cor-

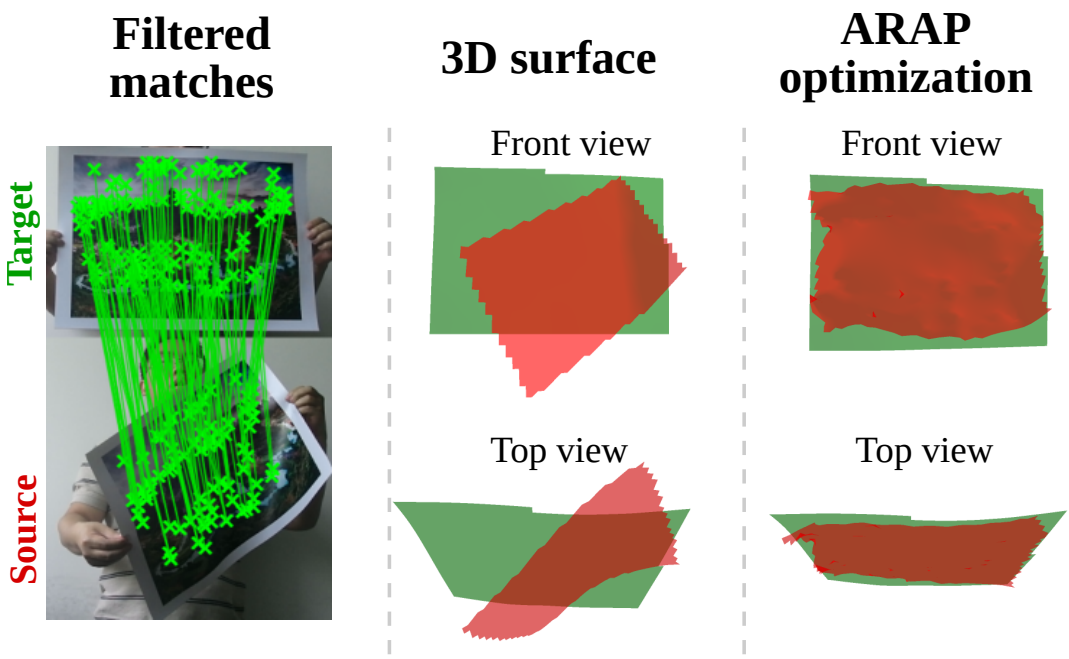

Figure 3. **Non-rigid 3D surface registration overview**. We use the filtered correspondences (left) to align two meshes of the same surface obtained from their respective RGB-D frames (middle) to the same reference pose and deformation (right), using as-rigid-as-possible (ARAP) refinement.

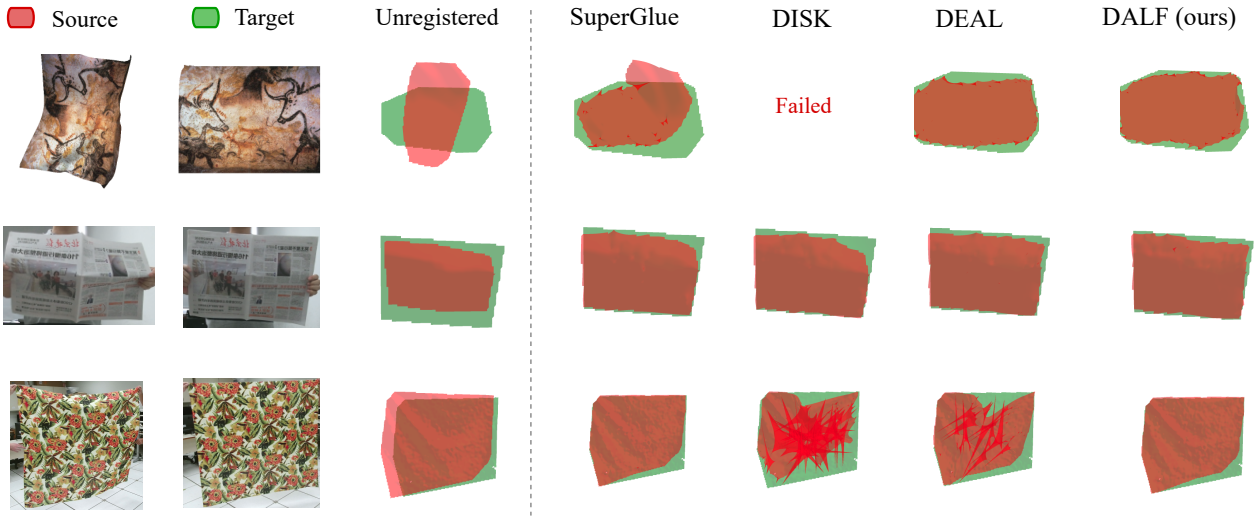

Figure 4. **Non-rigid registration under challenging scenarios**. Our method can achieve accurate non-rigid registration under large rotations, illumination changes caused by deformations, and highly repetitive patterns. In contrast, all other techniques produce low-quality results in at least one of the challenging scenarios. The sharp line artifacts in two registrations from DEAL and DISK indicates that the method produced inconsistent matches even after the filtering step, and the ARAP optimization failed due to local minima. Please check the supplementary video to visualize the registration results in 3D with the depicted image pairs and additional samples.

respondences. We adjusted the sensitivity of the local affine RANSAC of AdaLAM to tolerate more deviation from the base affine transformation, which usually happens in the presence of significant deformations.

AdaLAM tends to provide erroneous consistent affine matches when the scene has repetitive patterns, which is inevitable in practice. Those inconsistent matches introduce large errors in the ARAP optimization, and the method fails to return a meaningful result. Thus, to improve the robustness of the registration, for all methods, we use the best

200 matches according to the Lowe's ratio test [4], which drastically reduces artifacts caused by repetitive patterns, and also accelerates the convergence of the ARAP optimization. The non-rigid registration application source-code will be released alongside the reference implementation of our proposed method.

3.2. Qualitative results

Fig. 4 shows reconstruction results of challenging samples from the non-rigid datasets, where our method obtains robust registration. Aside from this PDF document, we made available a video (please check *registration_visual_results.mp4*) displaying the rendered registered surfaces in 3D from our approach and the competing methods, where it is possible to visualize the registration quality better. It is worth mentioning that SuperGlue [8], the best competing method, requires inputs in the form of image pairs, and employs global self and cross attention across local features when matching them , *i.e.*, the matching problem is conditioned to the input image pair, which significantly improves its robustness, especially in ambiguous regions. In contrast, our method independently detects the features, and a simple nearest neighbor search is used to perform matching. Our strategy can empower SuperGlue with deformation awareness by simply using our descriptors during training. In turn, SuperGlue's global self and cross-attention mechanisms can help our approach become much more robust to matching in challenging scenarios.

References

[1] Aljaž Božič, Michael Zollhöfer, Christian Theobalt, and Matthias Nießner. DeepDeform: Learning non-rigid rgb-d reconstruction with semi-supervised data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[2] Luca Cavalli, Viktor Larsson, Martin Ralf Oswald, Torsten Sattler, and Marc Pollefeys. Adalam: Revisiting handcrafted outlier detection. *arXiv preprint arXiv:2006.04250*, 2020. 2

[3] Matthias Innmann, Michael Zollhöfer, Matthias Nießner, Christian Theobalt, and Marc Stamminger. Volumedeform: Real-time volumetric non-rigid reconstruction. In *European conference on computer vision*, pages 362–379. Springer, 2016. 2

[4] David G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, pages 91–110, 2004. 4

[5] Zixin Luo, Lei Zhou, Xuyang Bai, Hongkai Chen, Jiahui Zhang, Yao Yao, Shiwei Li, Tian Fang, and Long Quan. Aslfeat: Learning local features of accurate shape and localization. In *CVPR*, pages 6589–6598, 2020. 1

[6] Lena Maier-Hein, Peter Mountney, Adrien Bartoli, Haytham Elhawary, D Elson, Anja Groch, Andreas Kolb, Marcos Rodrigues, J Sorger, Stefanie Speidel, et al. Optical techniques for 3d surface reconstruction in computer-assisted laparoscopic surgery. *Medical image analysis*, 17(8):974–996, 2013. 2

[7] Welerson Melo, Guilherme Potje, Felipe Cadar, Renato Martins, and Erickson R Nascimento. Learning to detect good keypoints to match non-rigid objects in rgb images. In *SIBGRAPI*, volume 1, pages 61–66. IEEE, 2022. 1

[8] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, June 2020. 4

[9] Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In *Symposium on Geometry processing*, volume 4, pages 109–116, 2007. 2

[10] Marvin Teichmann, Andre Araujo, Menglong Zhu, and Jack Sim. Detect-to-retrieve: Efficient regional aggregation for image search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5109–5118, 2019. 1

[11] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33:14254–14265, 2020. 1