DEEP-IMAGE-MATCHING: A TOOLBOX FOR MULTIVIEW IMAGE MATCHING OF COMPLEX SCENARIOS

L. Morelli [a,b], F. Ioli [c], F. Maiwald [d], G. Mazzacca [a,e], F. Menna [a], F. Remondino [a]

[a] 3D Optical Metrology (3DOM) unit, Bruno Kessler Foundation (FBK), Trento, Italy
Web: http://3dom.fbk.eu – Email: <lmorelli><fmenna><remondino>@fbk.eu

[b] Dept. of Civil, Environmental and Mechanical Engineering (DICAM), University of Trento, Italy

[c] Dept. of Civil and Environmental Engineering (DICA), Politecnico di Milano, Milan, Italy – Email: francesco.ioli@polimi.it

[d] Institute of Photogrammetry and Remote Sensing, TU Dresden, Germany – Email: ferdinand.maiwald@tu-dresden.de

[e] Dept. Mathematics, Computer Science and Physics, University of Udine, Italy

Commission II

KEY WORDS: Deep-learning, Multiview image matching, 3D reconstruction, Image retrieval, SuperGlue, LightGlue, LoFTR, SIFT.

ABSTRACT:
Finding corresponding points between images is a fundamental step in photogrammetry and computer vision tasks. Traditionally, image matching has relied on hand-crafted algorithms such as SIFT or ORB. However, these algorithms face challenges when dealing with multi-temporal images, varying radiometry and contents as well as significant viewpoint differences. Recently, the computer vision community has proposed several deep learning-based approaches that are trained for challenging illumination and wide viewing angle scenarios. However, they suffer from certain limitations, such as rotations, and they are not applicable to high resolution images due to computational constraints. In addition, they are not widely used by the photogrammetric community due to limited integration with standard photogrammetric software packages. To overcome these challenges, this paper introduces *Deep-Image-Matching,* an open-source toolbox designed to match images using different matching strategies, ranging from traditional hand-crafted to deep-learning methods (https://github.com/3DOM-FBK/deep-image-matching). The toolbox accommodates high-resolution datasets, e.g. data acquired with full-frame or aerial sensors, and addresses known rotation-related problems of the learned features. The toolbox provides image correspondences outcomes that are directly compatible with commercial and open-source software packages, such as COLMAP and openMVG, for a bundle adjustment. The paper includes also a series of cultural heritage case studies that present challenging conditions where traditional hand-crafted approaches typically fail.

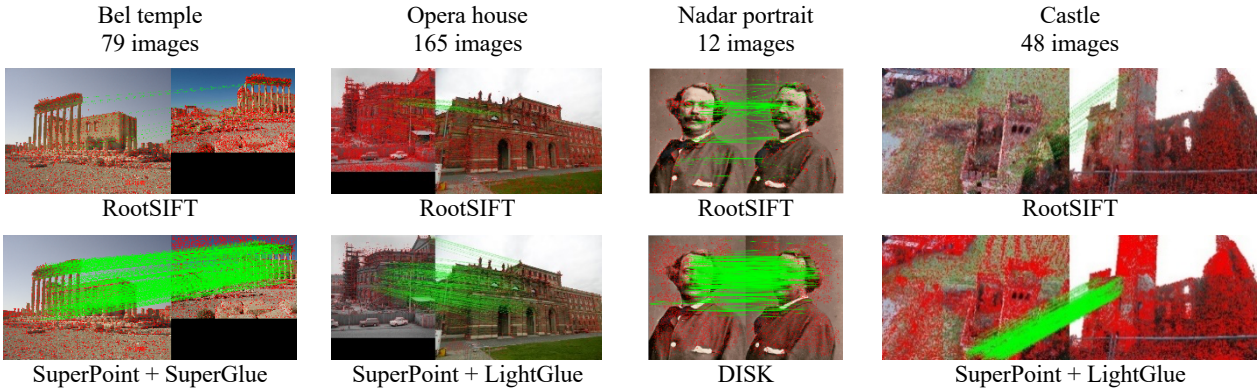


Figure 1: Sample images of the datasets used to test the proposed *Deep-Image-Matching* toolbox.

1. INTRODUCTION

Image matching plays a pivotal role in Structure-from-Motion (SfM), Visual Odometry (VO), simultaneous localization and mapping (SLAM), and various photogrammetric applications. Although traditional hand-crafted local features, such as SIFT (Lowe, 2004) and ORB (Rublee et al., 2011), have facilitated automatic keypoint extraction and matching, these methods have limitations when dealing with significant viewpoint and radiometric variations. These challenging situations can occur in cultural heritage image datasets, for example, when matching historical images with contemporary dataset for valorization projects based on virtual/augmented reality (Maiwald et al., 2021; Morelli et al., 2022) or multitemporal aerial datasets (Zhang et al., 2021; Farella et al., 2022). Typically, in these scenarios the number of historical images is often limited and presents strong variations in viewpoint and radiometric appearance.

Over the last decade, there has been a proliferation of deep learning (DL) approaches for feature extraction and matching (Chen et al., 2021; Jin et al. 2021; Yao et al., 2021) that aim to overcome these limitations and they have demonstrated resilience against varying illumination conditions, multi-temporal datasets, wide baselines, and significantly different view angles. Recently, several works have proved the effectiveness of DL approaches in challenging scenarios, including glacier monitoring with wide camera baselines (Ioli et al., 2023a, Ioli et al., 2023b), multi-temporal image matching (Maiwald et al., 2023), multi-temporal co-registration problems (Maiwald et al., 2021; Morelli et al., 2022), VO and SLAM

(Morelli et al., 2023), aerial triangulation (Remondino et al., 2022) and in terrestrial laser scanning point cloud registration (Markiewicz et al., 2023). However, well known limitations of DL approaches are their computational complexity, limited scale and rotation invariance of the descriptors and their application on high-resolution images.

Despite the growing interest in the topic, the practical use of local features and matchers for photogrammetric applications remains limited. This can be attributed to the effectiveness and reliability of SIFT-like approaches under optimal photogrammetric conditions, but also to the lack of an open-source library that easily integrates these new DL approaches into common open-source SfM pipelines such as COLMAP (Schonberger and Frahm, 2016), openMVG (Moulon et al., 2017), or commercial software packages such as Agisoft Metashape and Pix4D Mapper.

The aim of this paper is to introduce *Deep-Image-Matching,* an open-source toolbox for multi-camera image matching using DL approaches. *Deep-Image-Matching* aims to be a flexible toolbox for extracting corresponding points that are ready to be used for a photogrammetric reconstruction and to provide an easy-to-use interface to a wide range of state-of-the-art algorithms that have been recently developed by the computer vision community. Additionally, this paper presents some qualitative and quantitative case studies in cultural heritage, including challenging scenarios for traditional working pipelines.

The key features of *Deep-Image-Matching* are the following:

- availability of both traditional and DL-based local features and matchers in a single toolbox;
- ability to output multi-camera matches ready to be processed e.g., in COLMAP, openMVG or Agisoft Metashape;
- image pair selection with various strategy, including brute-force, low-resolution guided, sequential, image retrieval with global descriptors and custom pairs;
- support for large image formats with a tiling approach;
- support for camera/image rotations;
- support for global descriptors for effectively selecting image pairs in wide scale or complex scenarios;
- support of command line interface (CLI) and graphical user interface (GUI).

To the best of our knowledge, the most similar existing tools are HLOC (Sarlin et al., 2019) and *Image Matching WebUI*[1]. However, they do not support image rotations, large image formats and export for various software, notwithstanding the fact that the latter tool is designed only for image pairs.

2. *DEEP-IMAGE-MATCHING* TOOLBOX

Given a set of unordered images, *Deep-Image-Matching* can perform the matching operations and return the corresponding points between images. It is developed in Python and publicly available on GitHub (https://github.com/3DOM-FBK/deep-image-matching), it supports both CLI and GUI as well as a wide range of local features and matching algorithms, spanning from the traditional ones to recent state-of-the-art learning approaches. Available local features include ORB, SIFT, SuperPoint (DeTone et al., 2020), ALIKE (Zhao et al., 2022), ALIKED (Zhao et al., 2023), DISK (Tyszkiewicz et al., 2020), Key.Net (Barroso-Laguna et al., 2019) + HardNet8 (Pultar, 2020), DeDoDe (Edstedt et al., 2023b). SuperGlue (Sarlin et al., 2020), LightGlue (Lindenberger et al., 2023), LoFTR (Sun et al., 2021), SE2-LoFTR (Bökman and Kahl, 2022), and RoMA (Edstedt et al., 2023a) are implemented as matchers. Additionally, KORNIA python library (Riba et al., 2020) can be used for nearest neighbour matching.

Image pairs to be matched can be chosen by the user (*custom_pair* option), or they can be automatically selected by other strategies, including all possible pairs (*brute_force*), sequential matching (*sequential*), or image retrieval using global descriptors (*retrieval*). Image pairs can also be chosen by running a brute force on low-resolution images to limit computational time (option *matching_lowres*).

For high resolution images (e.g. one size larger than 5000 px), feature extraction and matching are carried out by tiling the images on a regular grid to fit into GPU memory, while the selection of the tiles to be matched is guided by a first matching on low-resolution images. Features matched on each image pair are verified by using PyDegensac (Mishkin et al., 2015) to reject outliers. Geometrically verified tie points are then stored in a SQLite3 database to be imported in COLMAP, or in the openMVG format, ready for the bundle adjustment in the respective software. To import the solution in other photogrammetric software (e.g. Metashape), image orientation is performed with pycolmap library and 3D tie points are exported in the Bundler format (Snavely et al., 2006).

3. DATASETS AND METHOD

3.1 Datasets

To show the potentiality of *Deep-Image-Matching,* four challenging heritage datasets have been selected (Figures 1-5), ranging from architectural heritage to reconstructions of historical figures.

Dataset A – *Bel temple* – is a collection of crowdsourced tourist photos - taken from the online repository REKREI (Vincent et al., 2015, 2016) - of the temple of Bel in Palmyra in Syria (Figure 2) destroyed in 2015. The dataset is composed of 79 images of different formats, quality and resolution, with significant changes in baselines as well as scales and illumination conditions. Dataset A included images acquired from all the four sides of the temple, although primarily on the main facade. Being tourist acquisitions, images do not share common calibration parameters.

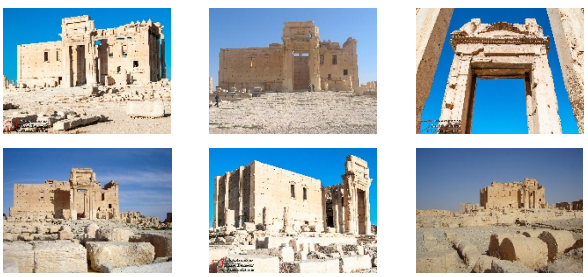


Figure 2: Sample images Dataset A - *Temple of Bel*.

Dataset B – *Semperoper* – shows the famous opera house in Dresden, Germany (Figure 3) and has been introduced in Maiwald et al., 2021 by using an automatic image retrieval approach within a large photo library (https://www.deutschefotothek.de/). For a better comparison of the methods provided by *Deep-Image-Matching,* the dataset has been reduced manually to 165 images so that exclusively the newly reconstructed opera house after 1869 is shown. The majority of images is provided with a maximum edge length of 1600 pixels by the library. The capture dates of the photographs span a period from 1880 to 2020 although most of the images have been taken from 1946 to 1960. Consequently, the

[1] https://github.com/Vincentqyw/image-matching-webui

photographs show vast radiometric and geometric changes. Additionally, the dataset features various image resolutions, varying image rotations and zooms.

Figure 3: Sample images Dataset B - *Semperoper*.

Dataset C – *Nadar* – is a collection of 12 self-portrait images from the opera *Revolving* (1865) of Gaspard-Félix Tournachon, known as Nadar[2] (Figure 4). The images have been downloaded from Wikipedia (resolution of 524 x 671 px), and they show significantly different viewpoints, low radiometric quality and many time-related artefacts. In addition, the dataset is partially ill-posed for photogrammetric purposes since Nadar sometimes changed both facial expressions and the relative position between head and shoulders. All these characteristics make the reconstruction difficult with classical approaches.

Figure 4: Sample images Dataset C - *Nadar*.

Dataset D – *Castle* – is composed of 48 images of the half-destroyed historical Castle of Casalbagliano, Alessandria (Italy). It is a traditional photogrammetric dataset of a cultural heritage site including 25 nadiral UAV images, 11 oblique UAV images and 12 terrestrial images (Fig. 5), with a camera network like a classical photogrammetric survey. Dataset D is the only one with an available ground truth. All the images of Dataset D were acquired by a single Canon Eos M with a fixed focal length of 22 mm and have a size of 5184 x 3456 px. The UAV nadiral images exhibit a modest overlap, approximately 60% in the longitudinal direction and 40% in the transversal direction. The UAV oblique images consist of four convergent shots acquired at each corner of the block, along with five additional images positioned along the exterior perimeter. The terrestrial images are acquired along a circle all around the castle. The main challenge of this dataset is linking together the nadiral UAV images with the terrestrial ones, as they have a strongly different point of view. Moreover, some of the terrestrial images are underexposed and characterized by large dark areas or acquired against the sun and therefore they show strong sunlight reflections.

Dataset D was extracted from a larger and more robust dataset, named dataset D_GT (Figure 6), and which was used as ground truth reference to validate the results obtained with *Deep-Image-Matching*. This dataset is composed of 172 images (83 nadiral, 61 oblique and 28 terrestrial) with an average overlap between the images between 70% and 80% and an average GSD of approximately 9 mm (Gagliolo et al., 2017, Gagliolo et al., 2018). The full dataset also included 19 targets deployed on the ground around the castle and measured by a total station with sub-centimetric accuracy. Dataset D_GT was processed with Metashape, by using 10 targets as Ground Control Points (GCPs) and performing a self-calibration of the camera. The quality of the photogrammetric block was evaluated on the remaining 9 targets, used as Check Points (CP), resulting in an overall RMSE of 1.9 cm in the three directions.

Figure 5: Sample images Dataset D - *Castle*. Images in the first row are samples of the nadiral and oblique UAV images, those in the second row are sample of the terrestrial images, including some of the challenging underexposed or overexposed images.

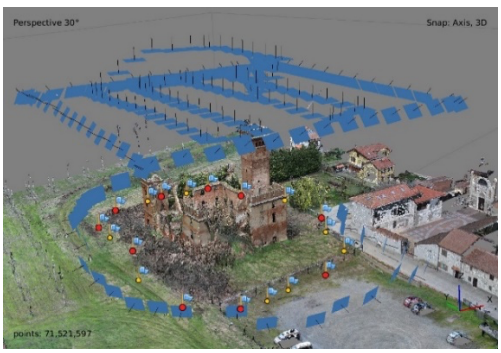


Figure 6: Dataset A_GT, used as the ground truth to evaluate the results on Dataset D. The red flags are the targets used as GCPs while the yellow ones those used as CPs.

3.2 Processing

Datasets were processed by *Deep-Image-Matching* using different DL local features and matchers. In particular, Dataset A was processed using the combination of SuperPoint and SuperGlue, as they have been widely used in challenging viewpoint and illumination scenarios (Ioli et al. 2023a). Similarly, also Dataset B and D were also processed by using SuperPoint features and LightGlue matcher, which is an optimized evolution of SuperGlue. On the other hand, Dataset C, which consists of the challenging set of Nadar's self portrait pictures, was tested with a combination of different DL local features. Except for Dataset D, all the other datasets were processed by using a brute-force pair selection approach and

[2] https://en.wikipedia.org/wiki/Nadar

features were extracted from full-resolution images to avoid losing keypoint detection accuracy.
Dataset D, which included high resolution images acquired by a mirrorless camera, was processed by upsampling the images by a factor of two using a bicubic interpolation technique before extracting features. This was motivated by the fact that SuperPoint lacks subpixel refinement capability in keypoint detection. Therefore, upsampling the images allowed for subpixel accuracy at the half-pixel level. In addition, since the upsampled images of dataset D had a resolution of 10368 x 6912 px and they could not fit into the memory of a consumer-grade GPU, the images were processed by subdividing them into regular tiles with a dimension of 3600 x 2400 px. The tile size was chosen as a compromise to limit the total number of tiles, but at the same time to be able to perform the processing using an NVIDIA RTX A2000 GPU with 12 GB of memory. To reduce the computational time, image pairs of Dataset D were selected by a low-resolution guided approach, which consisted of performing first a matching on all the possible pairs of images downsampled with the longest edge of 2000 px, and then selecting all the possible pairs with at least 30 valid matches.
For all the datasets, the tie points extracted with *Deep-Image-Matching* were imported into COLMAP to orient the image block by performing an incremental bundle block adjustment to estimate the camera poses and the 3D coordinates of the tie points. The results are compared with those obtained using (i) COLMAP with its native feature extraction (RootSIFT) and matching techniques and (ii) Agisoft Metashape, which implements proprietary algorithms for feature extraction and matching. The number of features per image extracted by COLMAP and Metashape was tuned by trial and error. In the end, the default values (8196 features per image for COLMAP and 40000 for Metashape) were used because they provided a good compromise between the number of valid matches extracted, the processing time, and the computational resources required by each software. In particular, the high number of features extracted by Metashape was needed to overcome its known limitations in extracting reliable matches in difficult scenarios. For Deep-Image-Matching, the maximum number of features per image was adjusted on a case-by-case basis to obtain a good number of local features (8000 for the Bel and Nadar datasets, 40000 for the Castle dataset, while no feature limit was set for the historical images of the Semperoper dataset). DL-based matching algorithms, such as SuperGlue and LightGlue, are effective in discriminating the valid matches even with a very high number of features, thanks to the attention mechanisms implemented inside the networks (Sarlin et al., 2020, Lindenberger et al., 2023).
As no ground-truth is available, except for dataset D, the comparison on the quality of the image orientation with different local features was made by comparing the number of oriented images to the total number available images (Table 1) and visually verifying the good orientation and consistency of the sparse reconstruction of tie points. This metric is considered sufficient for the purpose of illustrating the potential of learned local features for cultural heritage datasets, since classical methods (e.g. SIFT and ORB) usually fail almost completely on this kind of datasets.

4. RESULTS AND DISCUSSION

4.1 Temple of Bel

As reported in Table 1, the RootSIFT features implemented in COLMAP oriented 105 out of 149 images of dataset A, covering only three sides of the temple (Figure 7a). Metashape oriented almost all the images, but inconsistently: two coherent sub-blocks of images were correctly oriented, but wrongly connected because of wrong matches. Moreover, as shown in Figure 7b, the plan of the temple is duplicated. This could be the reason of the small reprojector error.
The combination of SuperPoint local features and SuperGlue matcher consistently oriented 141/149 images (Figure 7c). Even if a ground-truth is not available, the advantages of using deep-learning local features is clearly visible in the completeness of the dense cloud obtained from SuperGlue (Figure 7f) with respect to standard COLMAP (Figure 7d). For the processing, the simple-radial camera model has been used (focal length, principal point, and one parameter for radial distortion) with image variant calibration parameters, since images were taken by different tourists and different camera sensors.

Table 1: Summary of the results obtained with *Deep-Image-Matching* (DIM), compared to the results obtained with COLMAP (with RootSIFT features) and Agisoft Metashape with its own proprietary feature extractor and matchers. (*) Dataset C1 has been oriented combining different local features: SIFT, KeyNey + HardNet, ALIKED, SuperPoint, and DISK. (**) As COLMAP produced two non-linked reconstructions, the reported results refer to the reconstruction with the highest number of oriented images.

Label	Dataset	Local feature extractor and matcher	Oriented / total images	Mean reprojection error [px]	Mean track length	3D tie points
A1	Bel	DIM: SuperPoint + SuperGlue	141/149	1.33	5.6	118494
A2	Bel	COLMAP (RootSIFT)	105/149	0.59	4.3	17970
A3	Bel	Metashape (proprietary)	134/149	0.46	2.9	48141
B1	Semperoper	DIM: SuperPoint + LightGlue	161/165	1.47	7.0	17197
B2	Semperoper	COLMAP (RootSIFT)	147/165	0.75	4.7	20080
B3	Semperoper	Metashape (proprietary)	119/165	1.03	2.6	24970
C1	Nadar	DIM: Combination of local features (*)	12/12	1.09	2.6	2791
C2	Nadar	COLMAP (RootSIFT)	0/12	NA	NA	NA
C3	Nadar	Metashape (proprietary)	3/12	0.36	2.0	294
D1	Castle	DIM: SuperPoint + LightGlue	48/48	0.90	3.3	75274
D2	Castle	COLMAP (RootSIFT) (**)	31/48	0.94	3.7	11367
D3	Castle	Metashape (proprietary)	48/48	0.50	2.5	59679

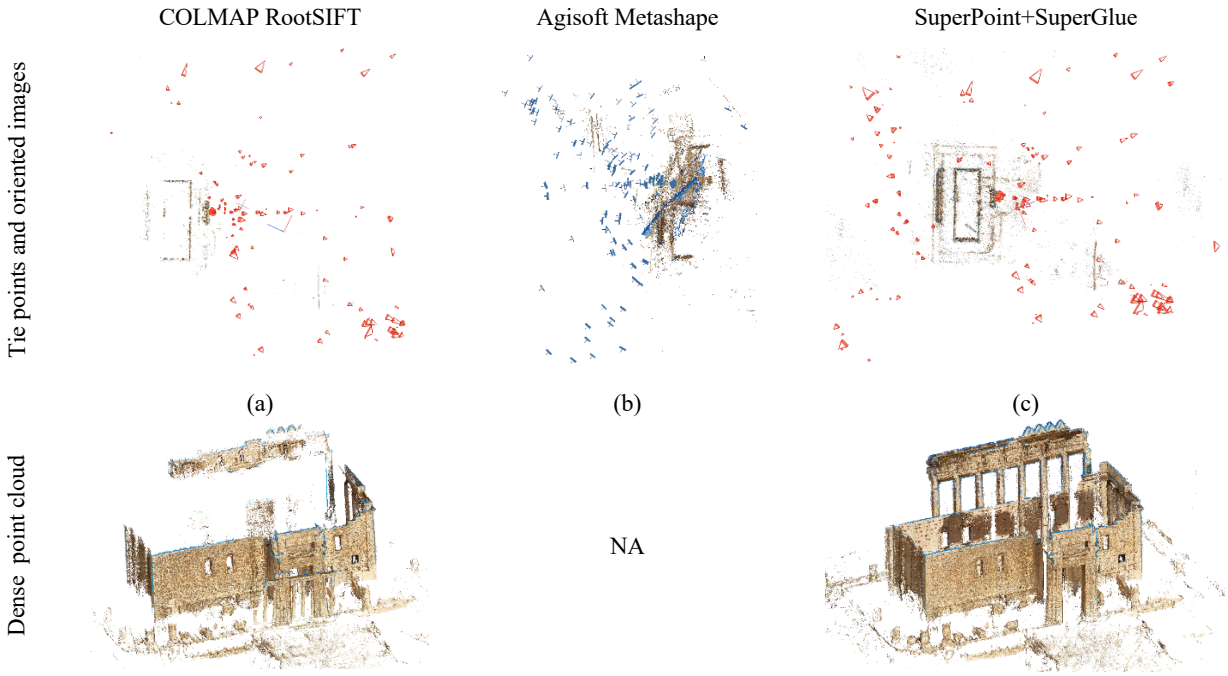
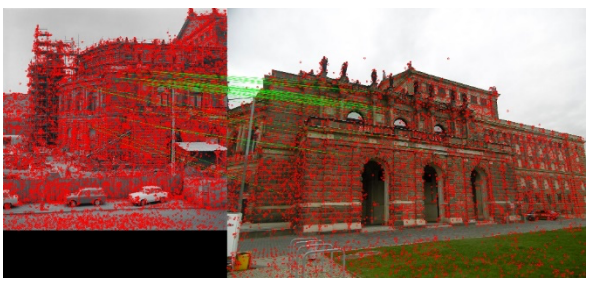
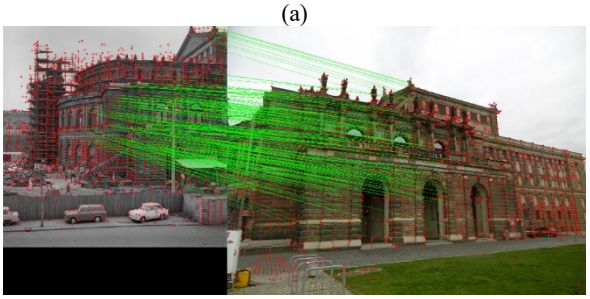


Figure 7: Results for the *Bel temple* dataset. (a-c) Tie points and camera orientation and (d-f) dense reconstruction for COLMAP RootSIFT, Agisoft Metashape, and SuperPoint+SuperGlue approaches, respectively. Metashape dense reconstrunction has not been run because of the incosistent image orientation output.



(a)



(b)

Figure 8: *Semperoper* image pair example with matches represented as green lines, while the red dots are the non-matched keypoints. (a) Features correctly matched by RootSIFT; (b) feature matched by SuperPoint + LightGlue. On this image pair Metashape could not find any correct match.

4.2 Semperoper

For dataset B, Metashape was able to correctly orient 119/165 images with a mean reprojection error of 1.03 pixels and with a relatively small mean track length (2.6 images), i.e., the number of images on which the same keypoint was found. However, Metashape was not able to orient the images with wide baselines such as those with sideviews of the building, even if parts of the front view are visible. Additionally, the images taken at night cannot be properly oriented. COLMAP was able to orient 147/165 with the smallest mean reprojection error of all experiments (0.75 pixels). COLMAP also finds only few correct feature points for the sideviews (Figure 8a). On the other hand, the DL-based approach (SuperPoint + LightGlue) enabled features to be matched even for image pairs with strong radiometric differences and wide baselines (Fig. 8b), with a reasonable large number for the mean track length. The total number of registered images is 161/165 with a mean reprojection error of 1.47 pixels. The larger reprojection error is probably related to the fact that SuperPoint lacks subpixel accuracy in keypoint detection, because it operates at the pixel level. With almost all images registered, the DL-based feature matching results is the best solution.

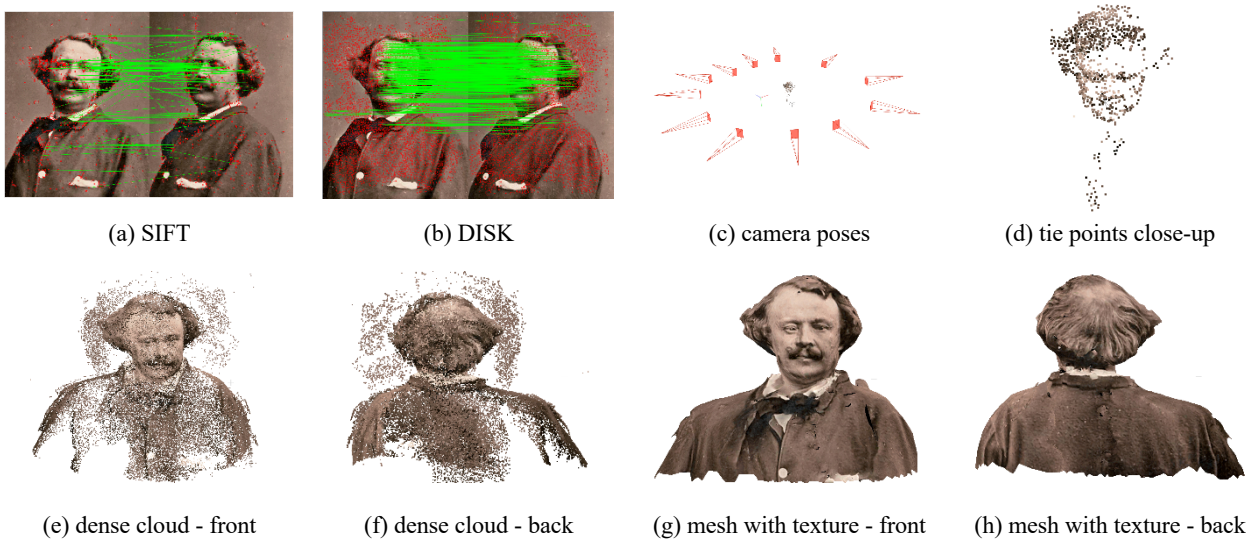
4.3 Nadar dataset

For the *Nadar* dataset, different local features have been tested: SIFT, KeyNey + HardNet, ALIKED, SuperPoint, and DISK. SuperPoint and DISK are matched with LightGlue, while the others are matched with a nearest neighbor approach. None of these approaches managed to orient more than three images, except DISK that found significantly more tie points and oriented seven images. In Figure 9a, a matching pair example is reported for SIFT (a) and DISK (b). Metashape completely failed to orient the dataset.

Only combining all the tie points from the previous approaches, excluded Metashape, it was possible to orient the whole dataset (Figure 9c-d). Tie points with multiplicity equals to two were excluded because considered not sufficiently robust and prone to outliers. In addition, no ratio threshold has been used to retain more matches. Because of the camera network and the scarcity of tie points, images were first oriented using a rough nominal focal length, then focal length and one radial distortion parameter were updated in a final bundle adjustment with self-calibration.

With regard to 3D model reconstruction, the poor radiometry of

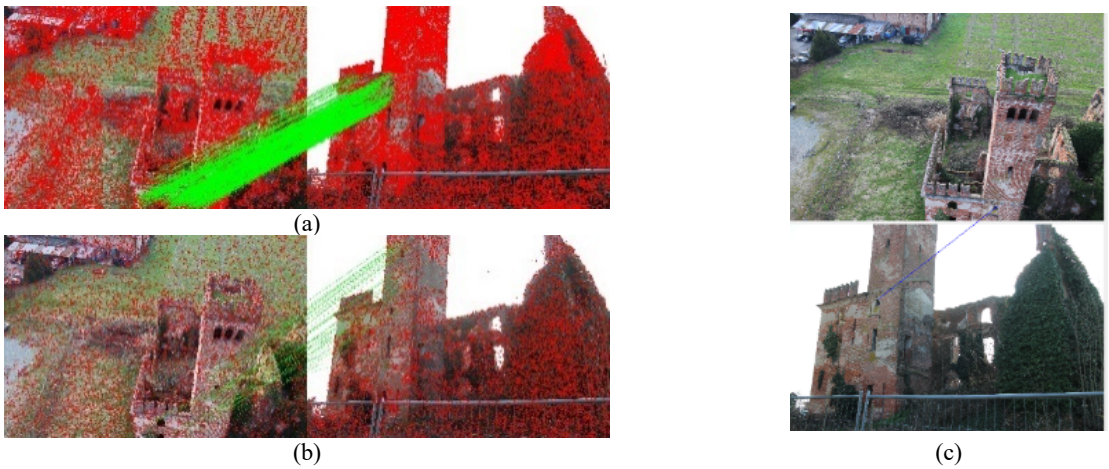
the images caused the dense matching of COLMAP and Metashape to fail. Therefore, to obtain a point cloud dense enough to build a meshed textured 3D model, the deep learning-based semi-dense matcher RoMA available in *Deep-Image-Matching* is applied (Figure 9e and 9f). Finally, using Metashape functionalities, a textureized model is created (Figure 9g and 9h).



(a) SIFT (b) DISK (c) camera poses (d) tie points close-up

(e) dense cloud - front (f) dense cloud - back (g) mesh with texture - front (h) mesh with texture - back

Figure 9: Results for Nadar dataset. (a) SIFT matches and (b) DISK matches on an image pair; (c) camera poses and (d) 3D tie points; (e-f) semi-dense point cloud generated from RoMA tie points; (g-h) textured mesh 3D model.



(a) (b) (c)

Figure 10: Example of matched features with the different approaches on a challenging image pair (green or blue lines are the valid matches, while the red dots are the rejected keypoints): (a) SuperPoint + LightGlue (658 valid matches); (b) COLMAP RootSIFT (49 valid matches); (c) Agisoft Metashape (1 valid match).

4.4 Castle dataset

The results obtained with SuperPoint + LightGlue (Tab. 1 - D1) were significantly better only compared to those obtained with a traditional COLMAP processing pipeline (D2), while they were similar to the outcomes obtained with Metashape (D3). With both LightGlue and Metashape all 48 images were oriented. On the other hand, COLMAP failed to orient all the images together, but it created two different not-linked models. The largest model consisted of only 31 oriented images, as detailed in Table 1. The smallest average reprojection error of 0.5 px was obtained by processing the dataset using Metashape with its proprietary local feature implementation, while a slightly higher reprojection error of 0.9 px was obtained by SuperPoint + LightGlue (Table 1), as SuperPoint did not have subpixel accuracy in keypoint detection. On the other hand, the mean track length of 3.3 obtained with SuperPoint + LightGlue was larger than 2.5 obtained with Metashape, guaranteeing higher redundancy of the observations in the bundle adjustment.

Figure 10 shows the matched keypoints for a challenging pair composed of a UAV oblique image and a terrestrial image, with a wide baseline and rather bad lighting conditions for the terrestrial image. The combination of SuperPoint + LightGlue, which works better under strong viewpoints conditions (Ioli et al., 2023b), allowed for extracting more than 600 valid matches, while Metashape was able to find only a single valid match. Surprisingly for this pair, COLMAP with RootSIFT was able to detect more matches than Metashape, probably thanks to its ability to estimate affine descriptors (Lindeberg et al., 1997). However, for other challenging pairs the results of SuperPoint + LightGlue were comparable to those of Metashape or COLMAP, without providing any relevant improvement. The matches obtained with SuperPoint + LightGlue were finally imported into COLMAP for the bundle adjustment and reconstruction (Figure 11a).

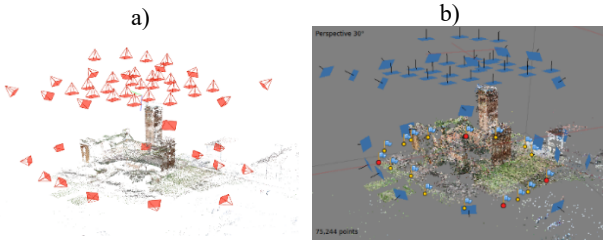


Figure 11: (a) Reconstructed sparse point cloud and oriented camera with SuperPoint + LightGlue after the bundle in COLMAP; (b) the same solution from Metashape, georeferenced with 4 GCPs (red flags), while all the other points are used as CPs (yellow flags).

The two complete solutions D1 (SuperPoint + LightGlue) and D3 (pure Metashape), in which all the 48 images were oriented, were validated by importing the estimated reconstructions into a new Metashape project and adding 4 GCPs at the block corner as a minimum constraint, while leaving the other 15 targets as CPs (Figure 11b). This made it possible to compare the solutions D1 and D3 with the ground truth dataset D_GT and to evaluate the on-ground reconstruction accuracy based on the CPs and the camera pose error by comparing the camera exterior orientation parameters. Comparable results were obtained for both D1 and D3, with a centimetric error on the CP and an average error of less than 5 cm on the camera location (Table 2). This highlights that the dataset D was correctly oriented using both SuperPoint + LightGlue and Metashape processing, without showing a clear superiority of any approach.

Dataset	Approach	RMSE X/Y/Z on CPs [m]	RMSE X/Y/Z on camera location [m]	RMSE Yaw/Pitch/Roll on camera attitude [°]
D1	SuperPoint + LightGlue	0.017/0.010/0.010	0.036/0.043/0.044	0.13/0.04/0.07
D3	Metashape	0.014/0.011/0.012	0.043/0.040/0.041	0.08/0.11/0.11

Table 2: Summary of accuracy evaluation for the solutions obtained with SuperPoint + LightGlue (D1) and Metashape (D3) with respect to the ground truth (D_GT). RMSEs on CPs were computed as the RMS of differences between the 3D coordinates of the targets measured on-the-field and those estimated in D1 and D3. The cameras RMSEs were computed as the RMS of the differences between the estimated 3D coordinates and attitude angles of the cameras in D_GT and those estimated in D1 and D3.

5. CONCLUSIONS

The paper presented *Deep-Image-Matching*, a tool to facilitate the usage of DL-based local features in the photogrammetric community. Compared to other existing tools, *Deep-Image-Matching* implement most of the essential options needed for photogrammetric applications, such as managing high resolution images, being robust to rotations and an *out-of-the-box* implementation that allows convenient interaction with various software packages (COLMAP, openMVG, and Metashape, easily extendable to further photogrammetric software). In addition, we presented some results on challenging datasets where the contribution of DL local features is clearly visible, being them trained to deal with wide camera baselines, significantly different viewpoints and radiometric differences.
A known limitation of several DL local features, including SuperPoint, is the lack of subpixel refinement in keypoint detection, which can easily result in better reprojection error after bundle adjustment. As future work, we plan to add a subpixel refinement routine into *Deep-Image-Matching* using traditional cross-correlation at the location of the matched features. Recent approaches for jointly refine all 2D keypoints that are matched together, such as Pixel-Perfect Structure-from-Motion (Lindenberger et al., 2021), will be also considered. Additionally, we plan to improve *Deep-Image-Matching* processing by exploiting parallelization or batch processing on the GPU.

REFERENCES

Barroso-Laguna, A., Riba, E., Ponsa, D. and Mikolajczyk, K., 2019. Key. net: Keypoint detection by handcrafted and learned CNN filters. *Proc. ICCV,* pp. 5836-5844.

Bökman, G. and Kahl, F., 2022. A case for using rotation invariant features in state of the art feature matchers. *Proc. CVPR*, pp. 5110-5119.

Chen, L., Rottensteiner, F. and Heipke, C., 2021. Feature detection and description for image matching: from hand-crafted design to deep learning. *Geo-spatial Information Science*, 24(1), pp.58-74.

DeTone, D., Malisiewicz, T. and Rabinovich, A., 2018. Superpoint: Self-supervised interest point detection and description. *Proc. CVPR workshops,* pp. 224-236.

Edstedt, J., Bökman, G., Wadenbäck, M. and Felsberg, M., 2023a. DeDoDe: Detect, Don't Describe--Describe, Don't Detect for Local Feature Matching. *arXiv preprint arXiv:2308.08479*.

Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M. and Felsberg, M., 2023b. RoMa: Revisiting Robust Losses for Dense Feature Matching. *arXiv preprint arXiv:2305.15404*.

Farella, E.M., Morelli, L., Remondino, F., Mills, J.P., Haala, N. and Crompvoets, J., 2022. The EuroSDR TIME benchmark for historical aerial images. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, 43, pp.1175-1182.

Gagliolo, S., Fagandini, R., Federici, B., Ferrando, I., Passoni, D., Pagliari, D., Pinto, L. and Sguerso, D., 2017. Use of UAS for the conservation of historical buildings in case of emergencies. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, *42*, pp.81-88.

Gagliolo, S., Fagandini, R., Passoni, D., Federici, B., Ferrando, I., Pagliari, D., Pinto, L. and Sguerso, D., 2018. Parameter optimization for creating reliable photogrammetric models in emergency scenarios. *Applied Geomatics*, *10*, pp.501-514.

Ioli, F., Bruno, E., Calzolari, D., Galbiati, M., Mannocchi, A., Manzoni, P., Martini, M., Bianchi, A., Cina, A., De Michele, C. and Pinto, L., 2023. a Replicable Open-Source Multi-Camera System for Low-Cost 4d Glacier Monitoring. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, 48, pp.137-144.

Ioli, F., Barbieri, F., Gaspari, F., Nex, F., and Pinto, L., 2023. ICEpy4D: a Python Toolkit for Advanced Multi-Epoch Glacier Monitoring with Deep-Learning Photogrammetry. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLVIII-1/W2-2023, 1037-1044.

Jin, Y., Mishkin, D., Mishchuk, A., Matas, J., Fua, P., Yi, K.M. and Trulls, E., 2021. Image matching across wide baselines: From paper to practice. *International Journal of Computer Vision*, 129(2), pp.517-547.

Lindeberg, T., Gårding, J., 1997. Shape-adapted smoothing in estimation of 3-D shape cues from affine deformations of local 2-D brightness structure. *Image Vision Computing*, *15*(6), 415-434.

Lindenberger, P., Sarlin, P.E., Larsson V. and Pollefeys, M., 2021. Pixel-Perfect Structure-from-Motion with Featuremetric Refinement. *Proc. CVPR.*

Lindenberger, P., Sarlin, P.E. and Pollefeys, M., 2023. LightGlue: Local Feature Matching at Light Speed. *arXiv preprint arXiv:2306.13643*.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60, pp.91-110.

Maiwald, F., Lehmann, C. and Lazariv, T., 2021. Fully automated pose estimation of historical images in the context of 4D geographic information systems utilizing machine learning methods. *ISPRS International Journal of Geo-Information*, 10(11), p.748.

Maiwald, F., Feurer, D. and Eltner, A., 2023. Solving photogrammetric cold cases using AI-based image matching: New potential for monitoring the past with historical aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 206, pp.184-200.

Markiewicz, J., Kot, P., Markiewicz, Ł. and Muradov, M., 2023. The evaluation of hand-crafted and learned-based features in Terrestrial Laser Scanning-Structure-from-Motion (TLS-SfM) indoor point cloud registration: the case study of cultural heritage objects and public interiors. *Heritage Science*, 11(1), p.254.

Mishkin, D., Jiri M., Michal, P.,2015. MODS: Fast and robust method for two-view matching. *Computer vision and image understanding*, 141. 81-93.

Morelli, L., Bellavia, F., Menna, F. and Remondino, F., 2022. Photogrammetry Now and Then - From hand-crafted to deep-learning tie points. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*., 48, pp.163-170.

Morelli, L., Menna, F., Vitti, A., Remondino, F. and Toth, C., 2023, September. Performance Evaluation of Image-Aided Navigation with Deep-Learning Features. *Proc. ION GNSS+ 2023,* pp. 2048-2056.

Moulon, P., Monasse, P., Perrot, R. and Marlet, R., 2017. Openmvg: Open multiple view geometry. In *Reproducible Research in Pattern Recognition: First International Workshop, RRPR 2016,* pp. 60-74, Springer International Publishing.

Pultar, M., 2020. Improving the HardNet descriptor. *arXiv preprint arXiv:2007.09699*.

Remondino, F., Morelli, L., Stathopoulou, E., Elhashash, M., Qin, R., 2022. Aerial triangulation with learning-based tie points. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*., XLIII-B2-2022, 77–84.

Riba, E., Mishkin, D., Ponsa, D., Rublee, E. and Bradski, G., 2020. Kornia: an open source differentiable computer vision library for pytorch. *Proc. WCACV*, pp. 3674-3683.

Rublee, E., Rabaud, V., Konolige, K. and Bradski, G., 2011, November. ORB: An efficient alternative to SIFT or SURF. *Proc. ICCV,* pp. 2564-2571.

Sarlin, P.E., Cadena, C., Siegwart, R., Dymczyk, M., 2019. From Coarse to Fine: Robust Hierarchical Localization at Large Scale. *Proc. CVPR,* pp. 12716-12725.

Sarlin, P.E., DeTone, D., Malisiewicz, T. and Rabinovich, A., 2020. Superglue: Learning feature matching with graph neural networks. *Proc. CVPR,* pp. 4938-4947.

Schonberger, J.L. and Frahm, J.M., 2016. Structure-from-motion revisited. *Proc. CVPR,* pp. 4104-4113.

Snavely, N., Seitz, S.M. and Szeliski, R., 2006. Photo tourism: exploring photo collections in 3D. Proc. *ACM SIGGRAPH,* pp. 835-846.

Sun, J., Shen, Z., Wang, Y., Bao, H. and Zhou, X., 2021. LoFTR: Detector-free local feature matching with transformers. *Proc. CVPR,* pp. 8922-8931.

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

Vincent, M. L., Coughenour, C., Flores Gutierrez, M., LopezMenchero Bendicho, V. M., Remondino, F., Frtisch, D., 2015. Crowd-sourcing the 3D digital reconstructions of lost cultural heritage. *Proc. IEEE Digital Heritage*, Vol. 1.

Vincent, M., Coughenour, C., Remondino, F., Gutierrez, M.F., Lopez-Menchero Bendicho, V.M., Fritsch, D., 2016. Rekrei: A public platform for digitally preserving lost heritage. *Proc. 44th CAA Conference*.

Yao, G., Yilmaz, A., Meng, F. and Zhang, L., 2021. Review of wide-baseline stereo image matching based on deep learning. *Remote Sensing*, 13(16), p.3247.

Zhang, L., Rupnik, E. and Pierrot-Deseilligny, M., 2021. Feature matching for multi-epoch historical aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 182, pp.176-189.

Zhao, X., Wu, X., Miao, J., Chen, W., Chen, P.C. and Li, Z., 2022. Alike: Accurate and lightweight keypoint detection and descriptor extraction. *IEEE Transactions on Multimedia*.

Zhao, X., Wu, X., Chen, W., Chen, P.C., Xu, Q., Li, Z., 2023. ALIKED: A Lighter Keypoint and Descriptor Extraction Network via Deformable Transformation. *IEEE Transactions on Instrumentation & Measurement*, 72, pp.1-16. 10.1109/TIM.2023.3271000.