EVALUATING HAND-CRAFTED AND LEARNING-BASED FEATURES FOR PHOTOGRAMMETRIC APPLICATIONS

F. Remondino, F. Menna, L. Morelli

3D Optical Metrology (3DOM) unit, Bruno Kessler Foundation (FBK), Trento, Italy
Web: http://3dom.fbk.eu – Email: <remondino><fmenna>@fbk.eu, lcmorelli.eng@gmail.com

Commission II, WGII/1

KEY WORDS: Keypoints, Detectors, Descriptors, Tie points**,** Deep learning, Accuracy, Point cloud, RMSE.

ABSTRACT:
The image orientation (or Structure from Motion - SfM) process needs well localized, repeatable and stable tie points in order to derive camera poses and a sparse 3D representation of the surveyed scene. The accurate identification of tie points in large image datasets is still an open research topic in the photogrammetric and computer vision communities. Tie points are established by firstly extracting keypoint using a hand-crafted feature detector and descriptor methods. In the last years new solutions, based on convolutional neural network (CNN) methods, were proposed to let a deep network discover which feature extraction process and representation are most suitable for the processed images. In this paper we aim to compare state-of-the-art hand-crafted and learning-based method for the establishment of tie points in various and different image datasets. The investigation highlights the actual challenges for feature matching and evaluates selected methods under different acquisition conditions (network configurations, image overlap, UAV vs terrestrial, strip vs convergent) and scene's characteristics. Remarks and lessons learned constrained to the used datasets and methods are provided.

1. INTRODUCTION

The extraction of accurate, reliable, and well-distributed tie points among images is a prerequisite for the accurate recovery of camera parameters and the generation of 3D geometry. Tie points are traditionally found coupling hand-crafted detectors (Lowe, 2004; Bay et al., 2006; Leutenegger et al., 2011; Rublee et al., 2011; Alcantarilla et al., 2013; Tombari and Di Stefano, 2014) and descriptors (Trzcinski et al., 2013; Calonder et al., 2011; Tola et al., 2010; Alahi et al., 2012) with feature matching comparison methods (brute force, FLANN, etc.) (Gonzalez-Aguilera et al., 2020).
To be correctly coupled, the detected and described keypoints must have a high level of repeatability, be discriminative, geometrically invariant, not very sensitive to changes in the brightness of the scene, and sparse to reduce memory usage (Apollonio et al., 2014).
Since few years, alternative deep learning methods based on convolutional neural networks (CNN) have been proposed and evaluated (Balntas et al., 2017; Schönberger et al., 2017; Fan et al., 2019; Jin et al., 2020; Bojanić et al., 2020). It is well-known that hand-crafted methods are bounded by a priori knowledge. So researchers aimed to let a deep network discover automatically which feature extraction process and representation are most suited to the data (Revaud et al., 2019). Some end-to-end methods for image-based 3D reconstruction purposes under challenging conditions are now available and solutions based on traditional hand-crafted methods are beginning to be outperformed by state-of-the-art learning-based approaches (Yi et al., 2016; DeTone et al., 2018; Ono et al., 2019; Dusmanu et al., 2019; Revaud et al., 2019; Christiansen et al, 2019; Luo et al, 2020). Among the current limitations for these learning-based methods, we can list the lack of shape-awareness and the general invariance to geometric, radiometric and scale changes, low localization accuracy, repeatability, etc. These issues can degrade the extractor performance in very different ways depending on how it has been designed (Revaud et al., 2019; Luo et al, 2020).

The aim of this work is to assess existing learning-based approaches to extract tie points across images for photogrammetric applications, especially those jointly performing detection and description ("end-to-end" methods). These are very interesting and can potentially lead to a better performance in terms of fewer outliers. On the other hand, as reported in Fan et al. (2019), Luo et al. (2020), Jin et al. (2020) and Bojanić et al. (2020), we should expect a very low keypoint localization accuracy or the extraction of a limited number of features (DeTone et al., 2018), which, in both cases, lead to a less accurate 3D reconstruction. Unlike similar investigations, we propose to evaluate learning-based methods with different metrics, using various image blocks (Table 1) and considering bundle adjustment statistics as well as 3D reference points (targets measured with topographic methods) as ground truth.

2. RELATED WORK

2.1 Learning-based methods

CNN-based features and methods can be applied in keypoint detection and description or can simultaneously perform both steps. A detector and a descriptor can be merged into a single architecture or be studied separately, even if some works (Yi et al., 2016; Ono et al., 2019; Revaud et al., 2019) suggest not to separate their training to obtain more reliable keypoints in the matching process. Loss functions include pairwise, triplet or structured loss whereas applications vary from image retrieval to camera pose estimations and dense 3D reconstructions. Network structures are highly variable and depend on the chosen approach (one-braced, siamese, multi-braced, etc.). Methods include:
Learning-based detectors. They train a CNN to detect keypoint, among which: TILDE (Verdie et al., 2015), Quad-Net (Savinov et al., 2017) and Key.Net (Barroso et al., 2019). They focused on identifying repeatable keypoints that could not be reliable for matching, as highlighted in Revaud et al. (2019). End-to-end methods also have their own detectors, such as LIFT (Yi et al.,

2016), LF-Net (Ono et al., 2019), SuperPoint (DeTone et al., 2018), D2-Net (Dusmanu et al., 2019).

Learning-based descriptors. Several stand-alone learning-based descriptors have been proposed, such as L2-Net (Tian et al., 2017), HardNet (Mishchuk et al., 2017), SOS-Net (Tian et al., 2019), LogPolarDesc (Ebel et al., 2019). They are usually trained on cropped patches centred on known keypoint (e.g. from SIFT), with the risk of creating a descriptor with a SIFT-like behaviour.

Detect-then-describe. Firstly, the keypoints are detected and then a patch extracted around each feature is passed to the description step. In this way, sparse local features are obtained with the advantage of lower memory usage and a detection that refers to low-level structures, such as corners and blobs, which allows a precise keypoint localization. The feature detector often considers only small image regions and typically focuses on low-level structures such as corners or blobs. Therefore, the descriptor captures higher level of information in a larger patch around the keypoint (Dusmanu et al., 2019). The detector and the descriptor can be hand-crafted, learned methods or a combination of the two. Since the detector looks for keypoints in an area of the image which is significantly smaller than the one used by the descriptor, noisy and low-resolution images can lead to significant variations of the low-level radiometric values that do not allow the coupling, while the corresponding descriptors would still be able to be coupled (Dusmanu et al., 2019).

End-to-end. These approaches perform a simultaneous detection and description in order to extract recognizable and uniquely describable keypoints (Yi et al., 2016; DeTone et al., 2018; Ono et al., 2019; Dusmanu et al., 2019; Revaud et al., 2019; Christiansen et al, 2019; Luo et al., 2020). The joint training of descriptor and detector avoids extracting non-discriminative keypoints and selects only repeatable interest points to improve the overall feature matching pipeline. These methods have a variable degree of architectural sharing for detection and description. For example, SuperPoint shares a deep representation between detection and description, however they rely on different decoder branches which are trained independently with specific losses (Dusmanu et al., 2019). D2-Net shares all parameters between detection and description and uses a joint formulation simultaneously optimized for both tasks (Dusmanu et al., 2019). LF-Net instead has two different networks for detection and description, trained together in an end-to-end way, but they do not share computations (Christiansen et al, 2019).

Detect-and-describe. This is a subcategory of end-to-end methods where detector and descriptor completely share the same network: ASLFeat (Luo et al., 2020), D2-Net (Dusmanu et al., 2019), and R2D2 (Revaud et al., 2019).

In terms of localization accuracy, Luo et al. (2020) emphasize-the need to recover spatial accuracy in the keypoint localization of end-to-end methods and identify some critical issues: while LF-Net's and D2-Net's lack in accuracy due to low-resolution feature maps, i.e., 1/4 of the original size, for SuperPoint it could be related to its decoder, used to restore spatial resolution. Finally, R2D2 keeps the original resolution using dilated convolution, but performs the detection deeply in CNN's layers.

2.2 Comparisons between hand-crafted and learning-based

While early studies on learning-based tie point extraction methods based their evaluation criteria on repeatability and matching score in isolation (Schönberger et al., 2017), more recent works emphasize instead the importance of performing downstream evaluations of the entire photogrammetric pipeline, going beyond pure descriptor matching (Yi et al., 2016; Schönberger et al., 2017; Ono et al., 2018; Jin et al., 2020). New feature extractors, being hand-crafted or learning-based, are generally tested on benchmarks - such as Balntas et al. (2017) or Bojanić et al. (2020), evaluating keypoint verification, matching and retrieval. Nevertheless, in geomatic applications, it is essential to test them with metrics specifically tailored for the object space, in particular 3D coordinates of the surveyed area.

Interesting evaluations in multiview scenarios are currently available (Schönberger et al., 2017; Jin et al., 2020), using extensive datasets, with different shooting angles, camera and environmental and lighting conditions, but without accuracy evaluations in metric terms in object space. The comparison between different methods assumes choosing several hyperparameters which can significantly influence the outcome of the assessment:

- Number of extracted features: it affects the accuracy of the image orientation process and extracting more and more features leads to a plateau (Jin et al., 2020).
- Detector and descriptor parameters: in our experiments we use default parameters both for hand-crafted and learning-based methods.
- Ratio threshold and features number: the ratio helps in filtering outliers or non-discriminative keypoints. Jin et al. (2020) pointed out that the ratio test is critical for performance analysis and one could arbitrarily select a threshold that favours one method over another, which shows the importance of proper benchmarking. Schönberger et al. (2017) proposed not to enforce the ratio test by pruning descriptors whose top-ranked nearest neighbours are very similar. Therefore, Schönberger et al. (2017) carried out only the cross-check without the ratio test whereas Jin et al. (2020) performed both the cross-check and the ratio test (the latter after looking for the optimal value). In our experiment we chose the second approach in order to use the results of that research both for the features number (8000) and the optimal ratio thresholds.
- Evaluation criteria: Schönberger et al. (2017) focused on evaluating descriptors starting from SIFT features, testing only descriptors, except for LIFT (Yi et al., 2016) which has its own detector. In most datasets an external ground truth is not used and evaluations are based on statistics obtained by COLMAP after the bundle adjustment. Unfortunately, some of the commonly used metrics, such as the reprojection error, are not able to disclose underlying systematic effects in the object space, in particular for imaging networks with poor geometry and redundancy. On the contrary, in some cases, low reprojection errors may even correspond to bad accuracy in object space. This is the case, for example, when the used functional model (pinhole camera) is incomplete, thus introducing systematic errors that can be absorbed by the camera network and therefore resulting in object space polynomial deformation (also called dome effect) (Menna et al., 2020; Nocerino et al., 2014; James and Robson, 2014). In some cases, the ground truth refers to camera location, that may be highly correlated with the interior orientation parameters, also depending on the imaging geometry (interior and exterior orientation parameters) and object shape (planar, three-dimensional). Jin et al. (2020), as ground truth, use SfM results obtained processing a larger number of images and evaluate the performance of small subgroups of images comparing the retrieved camera angles with respect to the ground truth values. We believe that the accuracy analysis of a photogrammetric method should be carried out on the triangulated 3D points using reference 3D data in the form of GCPs measured with an independent measurement technique (e.g. differential GNSS, geodetic surveying, laser tracker). In our experiments, all datasets feature a sufficient number of reference 3D coordinates (Table 1).

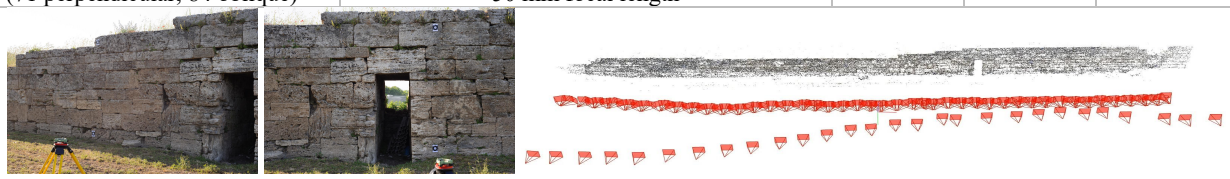
Images	Camera	GSD	Ground Truth	Acquisition
DATASET 1: Neptune Temple in Paestum (Italy)				
17 (12 perpendicular, 5 oblique)	Nikon D3X, 24 MP, full frame sensor, 14 mm focal length	9 mm	6 GCPs	Terrestrial
				
DATASET 2: Neptune Temple in Paestum (Italy)				
11	Canon EOS 550D, 18 MP, APS-C sensor, 25 mm focal length	14 mm	6 GCPs	UAV
				
DATASET 3: Paestum Wall (Italy) - Perpendicular + Convergent images				
155 in 2 strips (71 perpendicular, 84 oblique)	Nikon D3X, 24 MP, full frame sensor, 50 mm focal length	4-9 mm	22 GCPs	Terrestrial
				
DATASET 4: Ventimiglia Roman Theatre (Italy) - Nadiral + Oblique images				
64 in 4 strips (52 nadir, 12 oblique)	Nikon D3X, 24 MP, full frame sensor, 50 mm focal length	11 mm	7 GCPs	UAV
				
DATASET 5: Saranta Kolones (Cyprus)				
176	Nikon D3X, 24 MP, full frame sensor, 28 mm focal length	2-4.5 mm	12 GCPs	Terrestrial
				

Table 1: Summary of the datasets (images and camera networks with sparse point clouds) employed in the presented evaluation. The ground truth is given by reference 3D points measured with topographic methods (GCP-like): these 3D points are not included in the bundle adjustment as constraint but used to estimate an Helmert transformation and derive RMSEs of 3D coordinates. The reported min/max Ground Sampling Distance - GSD values are resampled from the original ones considering the used image resolution (i.e., 1500x1000 px – see Section 3.6) during the evaluations.

3. METHODOLOGY

The working methodology extends past assessments of learning-based methods by focusing on the following aspects:
Accuracy. We want to investigate whether learning-based methods can be a valid alternative to hand-crafted ones, by verifying point localization and accuracy in object (3D) space.
Evaluation criteria. We turn the attention from camera poses (Jin et al. 2020) to object point accuracy, reporting the absolute error of computed 3D coordinates compared to known reference points.
Datasets. Instead of photo-tourism datasets (Schönberger et al., 2017; Jin et al., 2020), we use sets of images acquired for photogrammetric purposes, in order to evaluate learning-based methods in contexts closer to engineering, heritage or architectural practice and surveys (Table 1). The employed datasets feature strong appearance changes due to significant variations in scale and viewing angle.
Camera configuration. We use more controlled scenarios, with image networks and scales typical of topographic and photogrammetric surveys, including single/unique camera, parallel (terrestrial or UAV) strips and loop closures acquisitions.

3.1 Considered methods

3.1.1 Learning-based methods

Among the available solutions, we mainly considered the end-to-end methods, i.e., those that jointly use CNNs for both the detection and description step. This choice was dictated by the following reasons:

- We intend to investigate the contribution of CNNs in the keypoints extraction, without the result being influenced by hand-crafted methods (usually the detector).
- Networks performing a joint learned detection and description show better performances (Yi et al., 2016; Ono et al., 2018).
- Selected networks provide sparse features hence they are preferred with respect to dense approaches in order to avoid excessive computation time on high-resolution images.
- Chosen methods were generated as general-purpose solutions hence they should accommodate various scenarios and contexts.
- These methods were trained for wide-baseline, a very interesting scenario that we tested adding oblique images, with strong appearance variance due to different scale and viewing angles with respect to perpendicular/nadir images.
- Selected methods seem to be suitable for retraining in order to include photogrammetric scenarios not considered so far (e.g. UAV and aerial).

The considered methods represent the state of the art and include:

- **LFNet** (Ono et al., 2018): it first identifies keypoints with its own detector, trained without using an existing one, then it calculates the position, scale and orientation of each feature, cutting an image patch around each keypoint, and finally it passes them to the descriptor.
- **R2D2** (Revaud et al., 2019): it joins detector and descriptor steps, focusing on features not only repeatable but with a good chance to be matched. It extracts a descriptor for each pixel of the image and obtains two confidence maps for repeatability and reliability. Keypoints are chosen where there is a maximum in both maps. The detector was trained in a self-supervised manner avoiding the usage of keypoints extracted with existing detectors. R2D2 requires a large amount of memory usage as it uses dilated convolutions.
- **SuperPoint** (DeTone et al., 2018): it consists of an encoder and two decoders, one for the localization of the features, the others for their description. The method is trained on synthetic images in a self-supervised way in the first phase, then the training is reinforced using real images. This method tends to extract fewer features with respect to other end-to-end methods.
- **ASLFeat** (Luo et al., 2020): it aims to recover shape-awareness and accuracy in keypoint localization with a multi-level detection mechanism. Its structure is based on D2-Net (Dusmanu et al., 2019) and performs both detection and description in a single step.
- **Key.Net+HardNet** (Barroso et al., 2019; Mishchuk et al., 2017)**:** in the first handcrafted and learned filters are combined to detect repeatable keypoints (Revaud et al., 2019), while the latter is a learning-based descriptor. Their combination performed well in Jin et al. (2020) and it was chosen as an example of detect-then-describe method.

3.1.2 Hand-crafted methods

Following past keypoint analyses (Apollonio et al., 2014; Schönberger et al., 2017; Jin et al., 2020), we employed SIFT (Lowe et al., 2004) - since it has proven to be the most reliable and versatile over time, SURF (Bay et al., 2006) and AKAZE (Alcantarilla et al., 2013). In particular, the latter two have been chosen following the investigations on the optimal ratio threshold presented in Jin et al. (2020).

3.2 Evaluation Pipeline

The SfM / image orientation process includes several steps: features detection and description, keypoints matching, geometric verification and bundle adjustment. The most widely used open-source software, such as VisualSfM (Wu et al, 2013), COLMAP (Schönberger et al., 2016) and OpenMVG (Moulon et al., 2016), can import features only if they are SIFT-like, i.e., consisting of 128 positive integer parameters in the range [0, 255]. There is no implemented method for importing floating-point descriptors or descriptors of arbitrary size. Therefore, using OpenCV-Python libraries, we extracted the features and matched them externally, performing the geometric verification and bundle adjustment in COLMAP. This software has performances comparable to OpenMVG (Stathopoulou et al., 2019), but it offers a more user-friendly graphical interface and provide useful statistics after the bundle adjustment. AliceVision (Moulon et al., 2016; Jancosek et al., 2011) is also an interesting choice since it already integrates AKAZE and SIFT, but currently presents difficulties in registering large datasets (Stathopoulou et al., 2019). Finally, VisualSFM was excluded because it only considers the first radial distortion additional parameter, whereas some of the considered datasets have short focal lengths, thus requiring at least two coefficients to properly model the lens radial distortion.
Features extraction and detection. For the learning-based methods, we used the respective implementations available on GitHub with default parameters. For R2D2 the model *WASF_N8_big.pt* was chosen, designed to extract more keypoints than other models, while for ASLFeat and LF-Net we used *model.ckpt-60000* and *outdoor with rotation augmentation* respectively. For the hand-crafted methods, we used the OpenCV implementation made available within PhotoMatch (González-Aguilera et al., 2020). Note that for the SURF descriptor we have chosen the more discriminative version with 128 parameters instead of 64.
Features matching. This step was performed with the Brute-Force method implemented in OpenCV-Python with distance L2 or Hamming according to the method used. Brute-Force was chosen, albeit slow, to ensure a fair comparison between

methods. The obtained matches were then filtered with the Lowe's ratio test and cross-check.
Bundle adjustment (free network). All matches were imported as *raw matches* to be geometrically checked before the bundle adjustment with the RANSAC method implemented in COLMAP. All intrinsic parameters are shared by all cameras and the first two radial distortion coefficients are used. The initial image pair from which the bundle starts is freely chosen by the software. COLMAP does not allow to include GCPs in the bundle solution, therefore their image coordinates were imported as tie points in order to triangulate their 3D coordinates. These computed values were then used for the accuracy evaluations.
Helmert transformation. The computed 3D coordinates were imported into CloudCompare and roto-translated with a scale factor (7-parameter Helmert transformation) using the available reference coordinates as ground truth (Oniga et al., 2016).

3.3 Keypoint number and ratio threshold

Since the accuracy of each method depends on the number and quality of extracted keypoints and on the way outliers are filtered, our tests were performed as proposed in Jin et al. (2020) in order to use the same optimal ratio thresholds in the Brute-Force matching (SIFT: 0.80; SURF: 0.90; AKAZE: 0.90; ASLFeat: 0.80; R2D2: 0.95; Key.Net+HardNet: 0.85; LF-Net: 0.95; SuperPoint: 0.90). For all methods, we extracted 8000 features except for SuperPoint and LFNet that have been trained to extract fewer keypoints than the other networks. Therefore, for these two methods we extracted 2000 and 8000 features.

3.4 Performance analyses

The performance evaluation is executed using similar metrics and processes presented in (Heinly et al., 2012; Apollonio et al., 2014; Remondino et al., 2017; Stathopoulou et al., 2019). The evaluation includes quantitative analyses for keypoint repeatability, pairwise matching efficiency, root mean square error (RMSE) on reference 3D points and tie point multiplicity. With respect to other evaluations which do not report absolute metric errors (Schönberger et al., 2017; Jin et al., 2020), our work aims to perform assessments also considering known 3D coordinates and comparing these reference values with those achieved triangulating specific points marked in the images.

3.5 Datasets and reference 3D points

The employed datasets (Table 1) have been chosen in order to appropriately represent typical 3D surveying scenarios in civil and cultural heritage applications, in terms of scale, network and camera-object distance. The *Paestum Wall*, which features a predominant dimension over the others, contains orthogonal (*Perpendicular*) as well as convergent (*Oblique*) images in order to reduce possible block deformations (dome effect). In the *Ventimiglia Theatre* dataset, beside two parallel nadiral UAV strips (*Nadir*), we included also two *Oblique* strips with an inclination of ca 45° in order to evaluate highly convergent views, illumination changes and perspective effects. The *Saranta Kolones* is composed of the largest number of images and presents a complex configuration with a strong variation of the camera-object distance and a loop closure.
All datasets contain well-distributed points of known 3D coordinates, materialized with high-contrast photogrammetric circular targets in the scenes. The measurement of their 3D coordinates in a local coordinate system was carried out with geodetic surveying techniques using a total station with 1" angular accuracy. Each point was observed at least two times (face left and face right) in non-prism mode (distance measurement accuracy ca. 3 mm). For the Ventimiglia dataset the targets were fixed on the ground; in this case a prism pole mounted on a tripod was used and collimated from at least two survey stations. The expected accuracy for all the datasets is better than 5mm in the three coordinates.

3.6 Extraction time, image resolution and hardware

Due to the actual limitations of learning-based methods, all images have been downsampled to 1500x1000 px to optimize the computer resources used. For reference, in the pipeline of Jin et al. (2020) all images are downsampled to a maximum size of 1024 pixels. Table 2 shows the average time needed to extract 8000 features. Each method uses a different deep learning framework (TensorFlow, PyTorch) and CUDA version.The technical specifications of the employed hardware are as follows: Processor Intel Core i7-4510U CPU @ 2.00GHz 2.60 GHz, RAM 16.0 GB, System 64 bit, GPU GeForce GTX 850M (5 Cores @ 901 MHz, 4096 MB).
On the other hand, using the same hardware, all hand-crafted methods were able to extract keypoints on full-resolution images, indicating that currently the learning methods require a greater commitment of IT resources. To perform a fair comparison, in our experiments all hand-crafted methods worked on the same downsampled resolution like learning-based methods.

METHOD	8000 keypoints per image [mm:ss]
SIFT PhotoMatch	00:03
SURF PhotoMatch	00:05
AKAZE PhotoMatch	00:02
ASLFeat	00:07
R2D2 GPU/CPU	00:26 / 02:36
Key.Net+HN	00:19
LF-Net	00:04
SuperPoint	00:04

Table 2: Average time required to extract 8000 keypoints per image (1500x1000 px) with the available implementations. For learning-based methods, a variable extraction time of 3-7 minutes must be added for loading the pre-trained model, regardless of the number of keypoints extracted.

4. RESULTS

4.1 Rotation invariance

Most of the published studies have tested learning-based methods on datasets with only roughly "upright" images, i.e., all images always present the sky in the upper part of the image and the object is always oriented in the same way with respect to the sensor. This is due to the fact that most of the learning-based methods (except SuperPoint and LF-Net) have no invariance to rotation or it is limited to small angles (about 20-30° for ASLFeat). The absence of rotation invariance in many of the tested learning-based methods is a very limiting deficiency for accurate photogrammetric applications where a single sensor with variable orientation is used, for example:

- In UAV surveys, between one strip and the next one, there is an inversion of the orientation of the aircraft's bow with consequent rotation of the sensor by about 180° (see *Ventimiglia Roman Theatre* dataset).
- In terrestrial surveys, it often happens that the sensor is rotated of 90° by the operator to better picture the scene or acquire an image network suitable for self-calibration.

ASLFeat, R2D2 and Key.Net+HardNet have not a sufficient rotation invariance, therefore in all our datasets presenting a

sensor rotation of 90° or 180°, the following approach is chosen: first the keypoints are extracted on the original rotated images, then the image coordinates are rotated accordingly in order to use only a single sensor in COLMAP elaborations. This procedure has been applied to perform a fair comparison among the methods, without affecting the accuracy in interest points localization. In this way, all image datasets could be properly registered, except for *Saranta Kolones*, where it was not always possible to identify a position to be considered "upright" (see 3rd image in the last row of Table 1), preventing implementation of the above-mentioned procedure and causing incomplete image orientation for ASLFeat, R2D2, and Key.Net. For these methods, specific retraining is required to achieve a much greater rotational invariance.

4.2 Mean Reprojecting Error (MRE)

Comparing the obtained Mean Reprojection Error (Figure 3) and the RMSEs (Figure 6-8), it is clear that the MRE is not a sufficient evaluation parameter to compare the quality of the derived 3D information. There is no similar trend between the two parameters, so it may not be really discriminatory to use only the MRE to evaluate a 3D reconstruction pipeline (Schönberger et al., 2017). In fact, the MRE can vary significantly from point to point for many reasons: it could be lower for a point with low Mean Track Length (MTL) and bad camera configuration (small base between cameras) or higher for a point with high MTL but good camera configuration.

4.3 Observations per image and Mean Track Length (MTL)

In terms of Mean Track Length (Figure 4), ASLFeat, R2D2 and Key.Net+HardNet achieve significantly higher values than the other methods. The difference is particularly pronounced in the *Paestum Wall Perpendicular+Oblique*, indicating an excellent ability of these methods to recognize the same point even on images with a very different aspect, i.e., under large perspective distortions. The ability to identify ("track") the same point on many images also emerges from the number of computed 3D points: for the same number of extracted features, a definitely less dense cloud is obtained (Figure 5).

4.4 Root Mean Square Error (RMSE) on 3D coordinates

The photogrammetric targets were manually marked on all possible images and then triangulated within the bundle adjustment. From the accuracy analyses (Figure 6-8) with the reference coordinates (Helmert/similarity transformation), it appears:

- The performances of hand-crafted and learning-based methods are comparable when the camera configuration is good in terms of base / object distance ratio, image overlap and above all the inclusion of oblique/convergent images in the scene, as the bundle adjustment succeed to compensate for the lack in keypoint localization accuracy.
- When the images are only nadiral, like in UAV surveys, or more generally, a quite flat scene is surveyed without the inclusion of oblique images (*Roman Theatre Nadiral* and *Paestum Wall Perpendicular*), the situation is more variable both for learning-based and hand-crafted methods. Only SIFT and ASLFeat perform well for both datasets.

5. CONCLUSIONS AND FUTURE WORKS

In this investigation we compared hand-crafted and learning-based feature extraction methods, focusing on the accuracy achieved in object space. From our experiences and from the achieved results, we can conclude:

- In terms of RMSE of 3D coordinates, learning-based achieve comparable results to hand-crafted methods when the image network is well designed. In the *Saranta Kolones* dataset, RMSE oscillate from a min of 0.011mm for Key.Net+HardNet to a max of 0.016mm for SuperPoint. For *Paestum Wall Perpendicular+Oblique* (Figure 8), the best learning-based is R2D2 which achieves 2.8 cm RMSE.
- When datasets have a redundant image configuration and strong geometry (*Neptune Temple, Paestum Wall Perpendicular+Oblique and Ventimiglia Nadiral+Oblique*), the bundle adjustment results have higher reliability and sensitivity to disclose gross observation errors. For this reason, all the methods perform very similarly in terms of RMSE on the reference points when the camera network geometry is redundant and well designed. On the contrary, with only nadir and perpendicular imaging networks, the RMSE show significant differences most likely due to the presence of gross observation errors that remain undetected as they are absorbed by the exterior and interior orientation parameters due to perspective coupling. This effect results in an inaccurate dome shape, especially for the *Paestum* dataset.
- None of the learning-based method could jointly tie and process Neptune terrestrial and UAV images, most probably due to large scale changes.
- ASLFeat, R2D2, Key.Net+HardNet have a greater ability to recognize repeatable keypoints (higher MTL – Figure 4) even when images have a really different viewing angles.
- LF-Net and Key.Net+HardNet achieved mediocre results in the most challenging datasets (*Paestum Wall Perpendicular and Ventimiglia Theatre Nadiral)*, with RMSE more than double of SIFT and ASLFeat for the *Paestum Wall Perpendicular* dataset.
- Only SIFT and ASLFeat performed well across all datasets, notwithstanding the lack of rotation invariance for the latter.
- ASLFeat, R2D2, Key.Net+HardNet are not invariant to rotation, an essential property in photogrammetric applications, requiring a specific retraining of the CNNs in order to stand such variation of the camera configuration.
- R2D2 exhibits controversial behaviour in the most challenging datasets, with a very high RMSE in *Paestum Wall Perpendicular*, probably linked to a non-optimal localization of the features. On the other hand, its keypoints are the best in terms of repeatability and reliability (always high MTL values) and excellent results in *Roman Theatre Nadiral*, which has little overlap between strips, but a good base to object distance ratio.
- Due to low number of extracted features, SuperPoint seems to be the less suitable method for UAV photogrammetry, in particular when there is less overlap between strips (despite good accuracy results in terrestrial datasets).

For sure more tests and datasets are needed to draw further lessons learnt but we believe that learning-based methods, being only at the dawn of their developments, are becoming a valuable and powerful alternative to traditional hand-crafted methods.
In the near future, the following aspects will be investigated: (i) increase the image resolution where learning-based methods can work and features can be extracted, (ii) generate dense point clouds from the retrieved image orientation and perform a cloud-to-cloud analyses, (iii) expand the analysis to learned descriptors, coupling them to hand-crafted detectors, (iv) consider processing time and memory usage, (v) re-train some methods in order to better accommodate high-resolution images, camera rotations, scale and illumination changes, etc.

	Neptune Terrestrial		*Theatre Nadir*		*Wall Perpendicular*
	Neptune UAV		*Theatre Nadir + Oblique*		*Wall Perpendicular +Oblique*

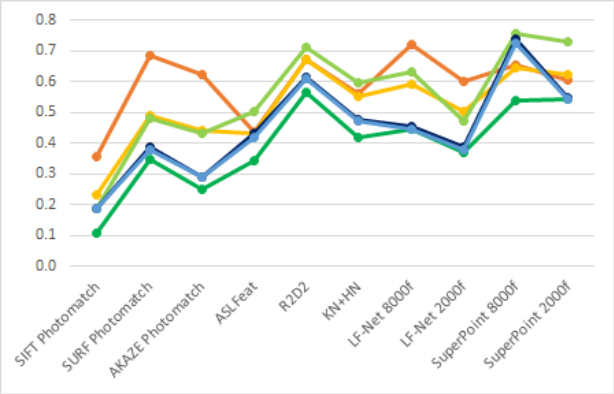


Figure 3: Mean reprojection error (MRE) [pixel] for the different datasets (see legend above) and tested methods.

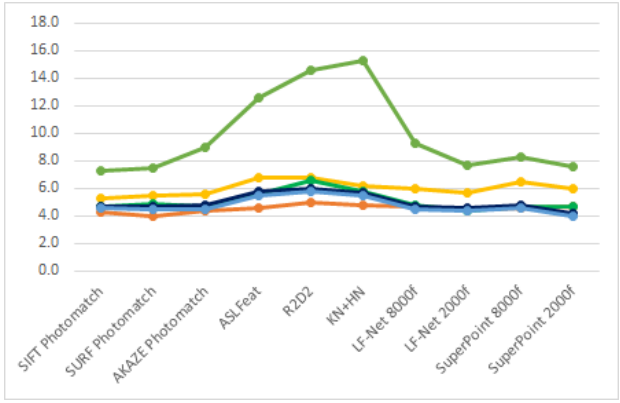


Figure 4: Average MTL for extracted tie points. ASLFeat, R2D2 and Key.Net+HardNet always show a higher MTL value with respect to the other learning-based methods.

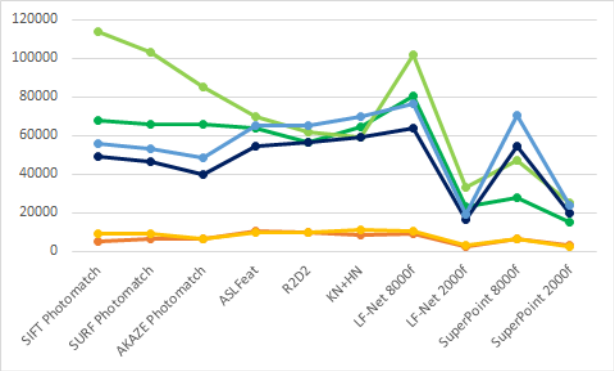


Figure 5: Computed 3D points (sparse cloud) within the bundle adjustment. In *Paestum Wall Perpendicular + Oblique*, ASLFeat, R2D2 and Key.Net+HardNet methods, having a higher MTL, create significantly fewer 3D points.

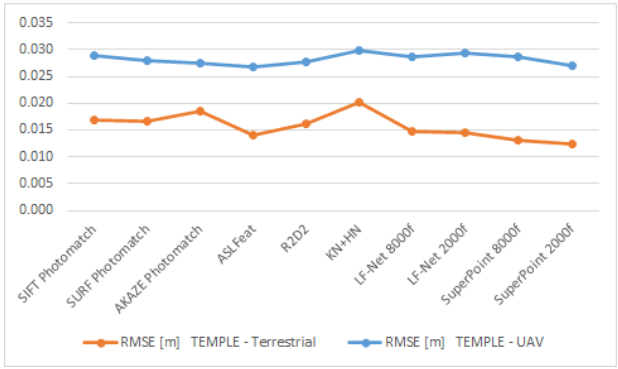


Figure 6: RMSE [m] for the *Neptune Temple* datasets. In both cases (terrestrial and UAV), the RMSE variation is very limited, the different methods behave in a similar way, especially in the UAV dataset where the sensor-object distance is very high.

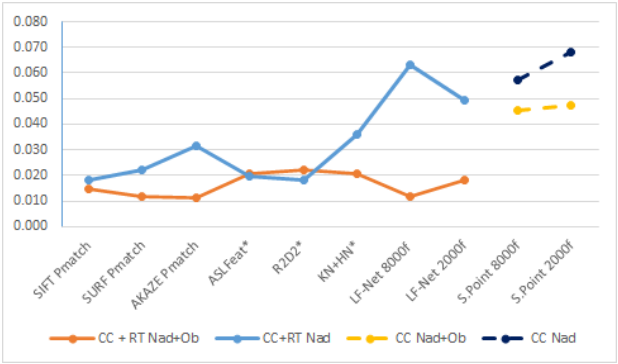


Figure 7: RMSE [m] for the *Ventimiglia Theatre* dataset, with only nadir images (*Nad*) and with the inclusion of oblique images (*Nad+Ob*). The SuperPoint tests are performed with only cross-check (CC) instead of cross-check + ratio test (CC+RT).

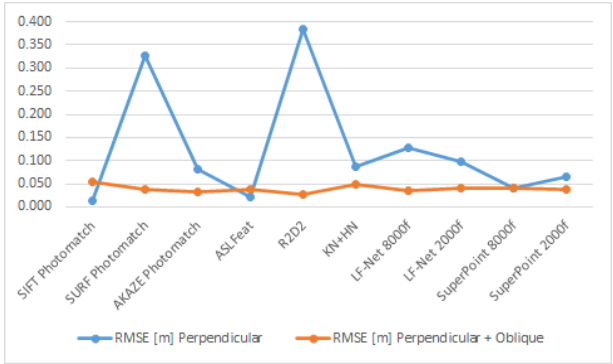


Figure 8: RMSE [m] for the *Paestum Wall*. SURF and R2D2 produced a pronounced curvature/deformation - not visible from the MRE analysis – in case of only perpendicular images. The use of oblique images reveals a minimal RMSE differences between the various methods.

REFERENCES

Alahi, A., Ortiz, R., Vandergheynst, P., 2012. Freak: Fast retina keypoint. Proc. *CVPR*, pp. 510-517.

Alcantarilla, P.F., Nuevo, J., Bartoli, A., 2013. Fast explicit diffusion for accelerated features in nonlinear scale spaces. Proc. *BCMV*, Vol. 34(7), 1281-1298.

Apollonio, F., Ballabeni, A., Gaiani, M., Remondino, F., 2014. Evaluation of feature-based methods for automated network

orientation. *ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Vol. XL-5, pp. 47-54.

Balntas, V., Lenc, K., Vedaldi, A., Mikolajczyk, K., 2017. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. Proc. *CVPR.*

Bay, H., Tuytelaars, T., Gool, L.V., 2006. SURF: Speeded-Up Robust Features. Proc. *ECCV*, pp. 404-417.

Bojanić, D., Bartol, K., Pribanić, T., Petković, T., Donoso, Y. D., Mas, J. S., 2019. On the comparison of classic and deep keypoint detector and descriptor methods. Proc. *ISPA*, pp. 64-69.

Barroso-Laguna, A., Riba, E., Ponsa, D., Mikolajczyk, K., 2019 Key.Net: Keypoint Detection by Handcrafted and Learned CNN Filters. Proc. *ICCV*.

Calonder, M., Lepetit, V., Ozuysal, M., Trzcinski, T., Strecha, C., Fua, P., 2011. BRIEF: Computing a local binary descriptor very fast. *IEEE Trans. PAMI*, Vol. 34(7), 1281-1298.

Christiansen, P. H., Kragh, M. F., Brodskiy, Y., Karstoft, H., 2019. Unsuperpoint: End-to-end unsupervised interest point detector and descriptor. *arXiv preprint arXiv:1907.04011*.

CloudCompare, 2021: http://www.cloudcompare.org/

DeTone, D., Malisiewicz, T., Rabinovich, A., 2018. SuperPoint: Self-Supervised Interest Point Detection and Description. *CVPR*.

Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T., 2019. D2-net: A trainable CNN for joint detection and description of local features. Proc. *CVPR.*

Ebel, P., Mishchuk, A., Yi, K. M., Fua, P., Trulls, E., 2019. Beyond Cartesian Representations for Local Descriptors. *ICCV*.

Fan, B., Kong, Q., Wang, X., Wang, Z., Xiang, S., Pan, C., Fua, P. (2019). A performance evaluation of local features for image-based 3D reconstruction. *IEEE Transactions on Image Processing*, *28*(10), 4774-4789.

González-Aguilera, D., Ruiz de Oña, E., López-Fernandez, L., Farella, E. M., Stathopoulou, E. K., Toschi, I., Remondino, F., Rodríguez-Gonzálvez, P., Hernández-López, D., Fusiello, A., and Nex, F..2020. Photomatch: an open-source multi-view and multi-modal feature matching tool for photogrammetric applications. *ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.,* Vol. 43(B5-2020), pp. 213–219. Code available at: https://github.com/TIDOP-USAL/photomatch

Heinly, J., Dunn, E., Frahm, J.M., 2012. Comparative Evaluation of Binary Features. Proc. *ECCV*

James, M.R., Robson, S., 2014. Mitigating systematic error in topographic models derived from UAV and ground-based image networks. *Earth Surface Processes and Landforms,* Vol. 39(10), pp.1413-1420.

Jancosek, M., Pajdla, T., 2011. Multi-view reconstruction preserving weakly-supported surfaces. *CVPR*, pp. 3121-3128.

Jin, Y., Mishkin, D., Mishchuk, A. et al., 2020. Image Matching Across Wide Baselines: From Paper to Practice. *Int Journal of Computer Vision*, Vol. 129, pp. 517-547

Lowe, D.G., 2004. Distinctive image features from scale invariant keypoints. *Int. Journal of Computer Vision*, Vol. 60(2), 91-110.

Luo, Z., Zhou, L., Bai, X., Chen, H., Zhang, J., Yao, Y., Li, S., et al., 2020. ASLFeat: Learning Local Features of Accurate Shape and Localization. Proc. *CVPR*.

Menna, F., Nocerino, E., Ural, S., Gruen, A., 2020. Mitigating image residuals systematic patterns in underwater photogrammetry. *ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Vol. 43, 977–984.

Nocerino, E., Menna, F. and Remondino, F., 2014. Accuracy of typical photogrammetric networks in cultural heritage 3D modeling projects. *ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Vol. 45.

Mishchuk, A., Mishkin, D., Radenovic, F., Matas, J., 2017. Working Hard to Know Your Neighbor's Margins: Local Descriptor Learning Loss. Proc. *NIPS.*

Moulon, P., Monasse, P., Perrot, R. Marlet, R., 2016. OpenMVG: Open multiple view geometry. In *International Workshop on Reproducible Research in Pattern Recognition*, pp. 60-74.

Oniga, E., Savu, A., Negrilă, A., 2016. The evaluation of CloudCompare software in the process of TLS point clouds registration. *RevCAD J. Geodesy Cadastre*, *21*, 117-124.

Ono, Y., Trulls, E., Fua, P., Yi, K.M., 2019. LF-Net: Learning local features from images. Proc. *NIPS.*

Remondino, F., Nocerino, E., Toschi, I., Menna, F., 2017. A critical review of automated photogrammetric processing of large datasets. *ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Vol. 42(2/W5), pp. 591-599.

Revaud, J., Weinzaepfel, P., de Souza, C.R., Humenberger, M., 2019. R2D2: Repeatable and Reliable Detector and Descriptor. Proc. *NIPS.*

Rublee, E., Rabaud, V., Konolige, K., Bradski, G.R., 2011. ORB: An efficient alternative to SIFT or SURF. Proc. *ICCV*.

Savinov, N., Seki, A., Ladicky, L., Sattler, T., Pollefeys, M., 2017. Quad-networks: unsupervised learning to rank for interest point detection. Proc. *CVPR.*

Shen, T., Luo, Z., Zhou, L., Zhang, R., Zhu, S., Fang, T., Quan, L., 2018. Matchable Image Retrieval by Learning from Surface Reconstruction. Proc. *ACCV.*

Schönberger, J., Frahm, J., 2016. Structure-From-Motion Revisited. Proc. *CVPR*.

Schönberger, J.L., Hardmeier, H., Sattler, T., Pollefeys, M., 2017. Comparative Evaluation of Hand-Crafted and Learned Local Features. Proc. *CVPR*.

Stathopoulou, E.-K., Welponer, M., and Remondino, F., 2019: open-source image-based 3d reconstruction pipelines: review, comparison and evaluation. *ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Vol. 42(2/W17), 331–338.

Tian, Y., Fan, B., Wu, F., 2017. L2-Net: Deep Learning of Discriminative Patch Descriptor in Euclidean Space. *CVPR*.

Tian, Y., Yu, X., Fan, B.,Wu, F., Heijnen, H., Balntas, V., 2019. SOSNet: Second Order Similarity Regularization for Local Descriptor Learning. Proc. *CVPR*.

Tombari, F., Di Stefano, L., 2014. Interest Points via Maximal Self-Dissimilarities. Proc. *ACCV*, pp. 586-600.

Tola, E., V. Lepetit, P. Fua, 2010. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE Trans. PAMI*, Vol. 32(5), 815-830.

Trzcinski, T. M. Christoudias, V. Lepetit., 2013. Learning Image Descriptors with Boosting. *IEEE Trans. PAMI*, Vol. 37(3), 597-610.

Verdie, Y., Yi, K. M., Fua, P., Lepetit, V., 2015 TILDE: A Temporally Invariant Learned DEtector. Proc. *CVPR*.

Wu, C., 2013. Towards Linear-Time Incremental Structure from Motion. Proc. *3DV*.

Yi, K.M., Trulls, E., Lepetit, V., Fua, P., 2016. LIFT: Learned invariant feature transform. Proc. *ECCV.*