AERIAL TRIANGULATION WITH LEARNING-BASED TIE POINTS

F. Remondino[1*], L. Morelli[1,2], E. Stathopoulou[1,3], M. Elhashash[4,6], R. Qin[4,5,6,7]

[1] 3D Optical Metrology (3DOM) unit, Bruno Kessler Foundation (FBK), Trento, Italy
Web: http://3dom.fbk.eu – Email: <remondino><lmorelli><estathopoulou>@fbk.eu

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

[3] Laboratory of Photogrammetry, National Technical University of Athens (NTUA), Athens, Greece

[4] Geospatial Data Analytics Lab, The Ohio State University, Columbus, USA

[5] Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, USA

[6] Department of Electrical and Computer Engineering, The Ohio State University, Columbus, USA

[7] Translational Data Analytics Institute, The Ohio State University, Columbus, USA - Email: <elhashash.3><qin.324>@osu.edu

Commission II – WG II/1

KEY WORDS: aerial triangulation, tie points, deep learning, image matching, accuracy analysis

ABSTRACT

Aerial triangulation (AT) has reached outstanding progress in the last decades, and now fully automated solutions for nadir and oblique images are available. Usually, image correspondences (tie points) are found using hand-crafted methods, such as SIFT or its variants. But in the last years, there were many investigations and developments to promote the use of machine and deep learning solutions within the photogrammetric processing pipeline. The paper explores learning-based methods for the extraction of tie points in aerial image blocks. Image correspondences are used to perform aerial triangulation (AT) and successively generate dense point clouds. Two different datasets are used to compare conventional hand-crafted detector/descriptor methods with respect to learning-based methods. Accuracy analyses are performed using GCPs as well as ground truth LiDAR point clouds. Results confirm the potential of learning-based methods in finding reliable image correspondences in the aerial block, still showing space for improvements due to camera rotations.

1. INTRODUCTION

Photogrammetry is one of the most widely used techniques for the determination of 3D metric information at various scales and from diverse imaging platforms (satellite, aerial, drone, terrestrial and underwater). The typical aerial photogrammetric workflow consists of the identification of image correspondences via sparse image matching, the estimation of the unknown camera parameters and 3D object coordinates (image triangulation) with a bundle adjustment (BA) method, the generation of dense point clouds via dense image matching (or Multi-View Stereo - MVS) and the realization of by-products like mesh models or orthophotos. Photogrammetric methods – since ever – aim to provide practical, reliable, and daily-based routines and solutions for geospatial data generation, geometric processing, and semantic interpretation – even with manual intervention to keep accuracy as high as possible. For two decades the community has provided many automated algorithms, also based on Artificial Intelligence (AI), to speed up geospatial data generation and interpretation, increase efficiency as well as robustness (Hartmann et al., 2015; Zhu et al., 2017; Becker et al., 2018; Gong and Ji, 2018; Yao et al., 2018; Liu et al., 2019; Griffiths and Boehm, 2019; Stathopoulou et al., 2019; Heipke and Rottensteiner, 2020; Huang et al., 2018; Shan et al., 2020; Chen et al., 2020a; Oezdemir et al., 2021; Qin and Gruen, 2021; Remondino et al., 2021). For sure there is still a hype around deep learning in research activities and in the media, but these methods are really a treasure trove for innovation in the geospatial field.

Following this momentum, this work aims to investigate the use of learning-based algorithms for the extraction of tie points in aerial image blocks. The work applies learning-based methods to full-size aerial images and highlights the performances of these methods in performing aerial triangulation (AT) and, successively, generating dense point clouds. For comparison analyses, using two different datasets (Table 1), tie points are automatically extracted using learning-based approaches as well as traditional hand-crafted detector/descriptor methods. Accuracy analyses are performed using GCPs as well as ground truth LiDAR point clouds.

2. RELATED WORK

AT has achieved remarkable improvement in the last decades, and now fully automated solutions for both nadir and oblique images are available (Rupnik et al., 2013; Rupnik et al., 2015; Maset et al., 2021). Automated AT based on a bundle block adjustment moved from point-based to feature-based methods, including also linear features (Habib et al., 2002; Schenk, 2004; Triggs et al., 2000). The identification of image correspondences (tie points – Figure 1) is traditionally performed using hand-crafted keypoint detectors and descriptors (Lowe, 2004; Bay et al., 2006; Alcantarilla et al., 2013; Bellavia et al., 2021). These hand-crafted approaches are based on *a priori* knowledge inspired by professional knowledge and intuitive experience (Yao et al., 2021). Despite their good performance, there are still open issues in case of large perspective or temporal differences as well as scale and illumination changes between the images.

In the last years, driven by rapid developments in deep learning networks, researchers proposed various innovative learning-based solutions aiming to overcome the limitations of hand-crafted methods (Verdie et al., 2015; Jin et al., 2021). Such solutions include *detect-then-describe* approaches where the detector (Verdie et al., 2015; Savinov et al., 2017; Barroso et al., 2019; Truong et al., 2020) and the descriptor (Mishchuk et al., 2017; Tian et al., 2017; Mishkin et al., 2018; Ebel et al., 2019; Pautrat et al., 2020; Pultar, 2020; Parihar et al., 2021) can be both learned methods or a combination of hand-crafted and learning-

* Corresponding author

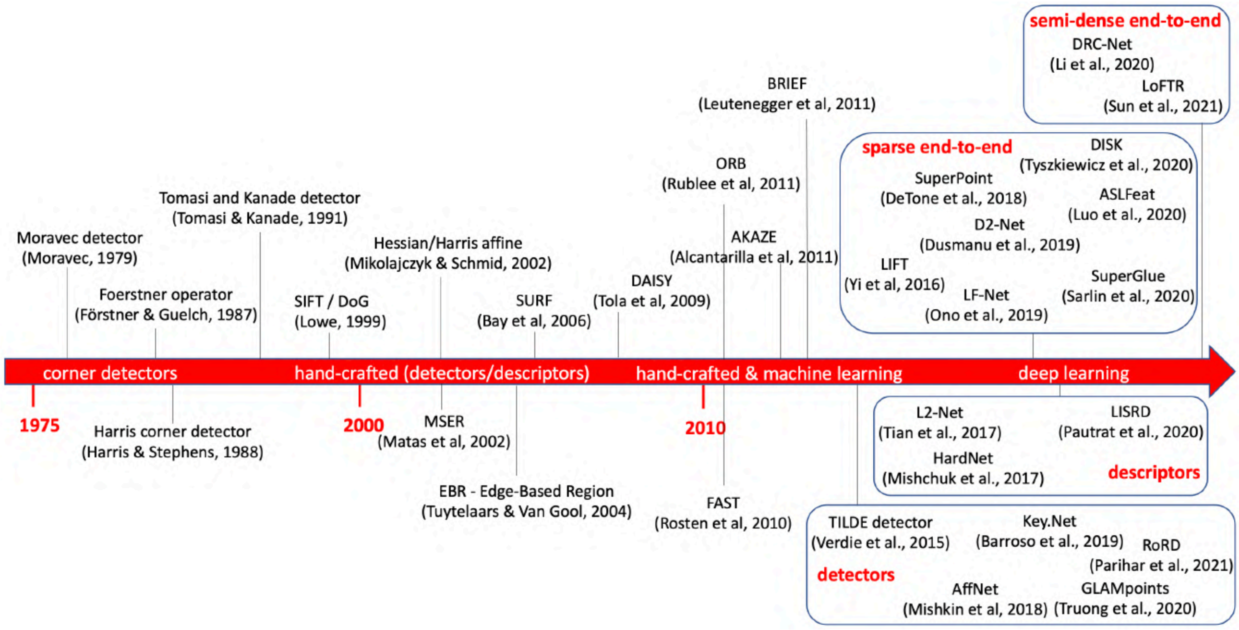


Figure 1: Some milestone methods within the evolution of image matching for tie point extraction.

based (Bellavia and Mishkin, 2021; Bellavia et al., 2022). Other approaches, called *end-to-end*, jointly optimize the entire pipeline to extract sparse image correspondences, e.g., LIFT (Yi et al., 2016), LF-Net (Ono et al., 2019), SuperPoint (DeTone et al., 2018), R2D2 (Revaud, 2019), D2-Net (Dusmanu et al., 2019), ASLFeat (Luo et al., 2020), SuperGlue (Sarlin et al., 2020), DISK (Tyszkiewicz et al., 2020). End-to-end methods were demonstrated to increase both the keypoint repeatability and reliability and, consequently, the image matching success rate and the final pose estimation accuracy (Remondino et al., 2021). More recently, various researchers (Choy et al., 2016; Rocco et al., 2018; Li et al., 2020) proposed end-to-end detector-free local feature matching methods that remove the feature detector phase and directly produce dense descriptors or dense feature matches. Among these, Sun et al. (2021) created the LoFTR approach based on Transformer (Vaswani et al., 2017): instead of performing image feature detection, description, and matching sequentially, it establishes pixel-wise dense matches at a coarse level and later refines the good matches at a fine level.
The use of learning-based methods to automatically orient image blocks is primarily applied to terrestrial datasets (Schonberger et al., 2017; Bojanić et al., 2019; Jin et al., 2021) with very few experiments on UAV datasets (Remondino et al., 2021; Bellavia et al., 2022) and aerial modern (Chen et al., 2020b) and historical (Ressl et al., 2020; Zhang et al., 2021) blocks. This is mainly because most of the existing deep architectures for tie point extraction are not suitable for general-purpose photogrammetric applications, particularly aerial blocks, due to their limitation in handling large image sizes, small scales and camera rotations among strips.

3. METHODOLOGY

3.1 Considered methods

Initial analyses on state-of-the-art hand-crafted and deep learning methods were performed to understand rotation and scale invariance issues in the case of aerial views (Figure 2). RootSIFT (Relja and Zisserman, 2012) was chosen to represent the hand-crafted family as it proved to be the most reliable and versatile solution (Schonberger et al., 2017). On the other hand, among the available learning-based solutions, we considered two rotation-invariant frameworks: LF-Net (Ono et al., 2018) as end-to-end architecture and KeyNet (Barroso et al., 2019) coupled with AffNet (Mishkin et al., 2018) and HardNet (Mishchuk et al., 2017) – available in the Kornia library (Riba et al., 2020), as a detect-then-describe approach. Both frameworks showed good performances in previous evaluations (Remondino et al., 2021; Bellavia et al., 2022), accommodating various scenarios and contexts. They also seem to be suitable for retraining processes to include photogrammetric scenarios. Moreover, to the best of authors' knowledge, they are among the very few methods which are partially invariant to camera rotations.

3.2 Image tiling approach

As learning-based methods demand many computational resources and can generally handle only small image sizes, a tiling approach is proposed in order to extract tie points in the full resolution images. Normally keypoints are not detected along the perimeter of the images due to the padding used during convolutions. Therefore, to avoid having no keypoints in areas of adjacent tiles, thus obtaining a not uniform keypoints distribution in the entire image, tiles (2500x2500 pixel) are overlapped vertically and horizontally by some 30 pixels. Features are detected/described on these tiles then tiles are reassembled for the matching and verification steps.

3.3 Datasets

Two different sets of aerial images (Table 1) are employed to test the capabilities of learning-based methods within AT processes and their influence on the generation of dense point clouds: the ISPRS/EuroSDR Dortmund benchmark (Nex et al., 2015) and the Dublin benchmark (Ruano and Smolic, 2021). These urban datasets were chosen due to their complementarity in terms of acquisitions, resolution, and ground truth (GT). They both feature nadir and oblique images, varying GSD (and image scale), picturing complex urban scenarios.

3.4 Processing pipeline

The AT process consists of features detection and description (Section 3.1), features matching, geometric verification, and final bundle adjustment (BA). The number of detected keypoints was set to be around 10,000 per image, while descriptors consist of 128 (rootSIFT and HardNet) and 256 (LF-Net) parameters. The OpenCV Brute-Force method with L2 distance is used, albeit slow, to handle descriptors of variable sizes and ensure a fair comparison between methods. Matches are then imported into

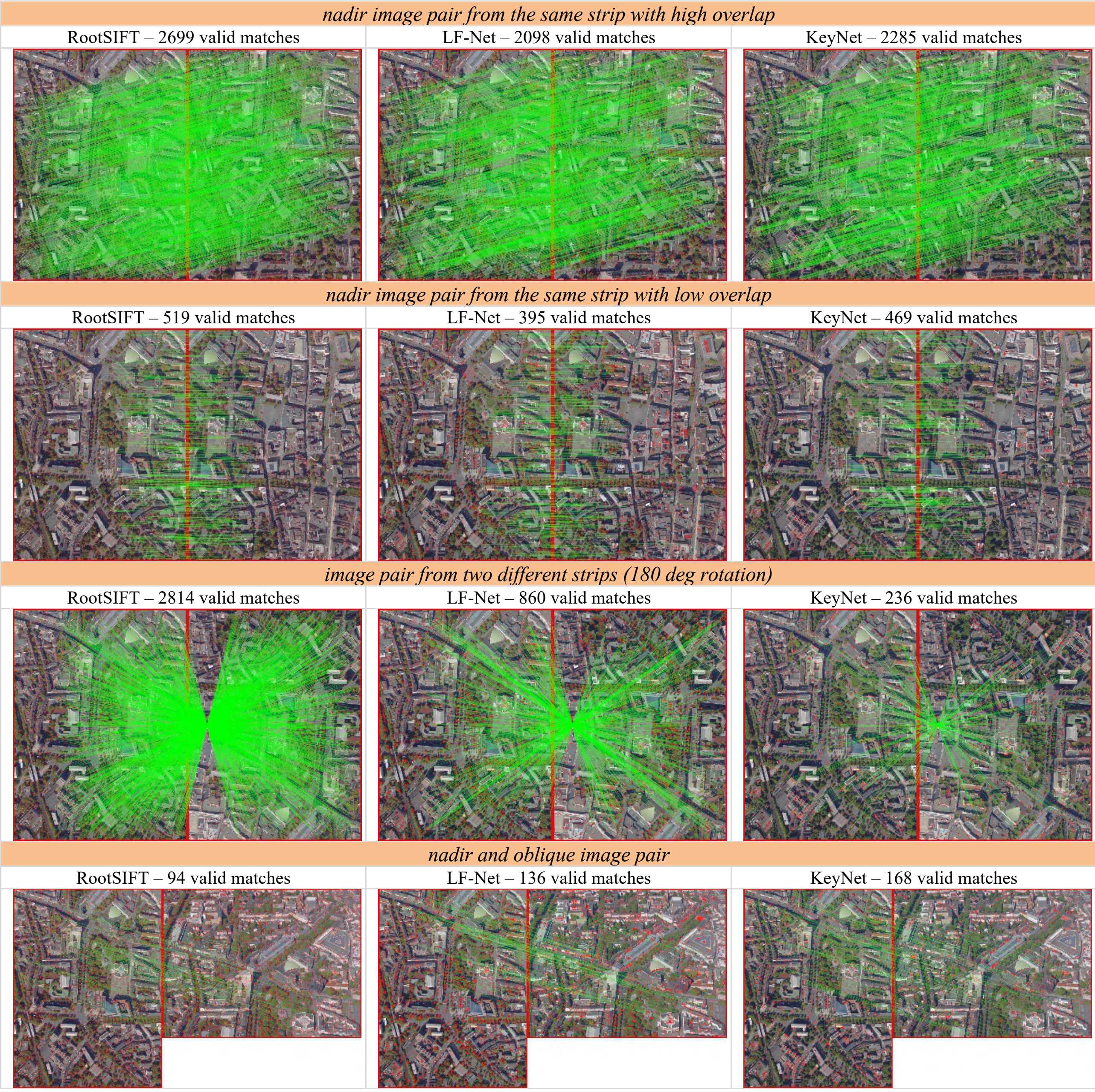


Figure 2: Tie point extraction on image pairs using hand-crafted and learning-based methods.

COLMAP, where first the geometric verification with RANSAC is performed. Ratio test thresholds of 0.80 for RootSIFT, 0.85 for KeyNet+HardNet, and 1.00 for LF-Net are used.
Finally, image observations are employed within the COLMAP incremental BA in free network mode. The available GCP imagecoordinates are used as tie points to triangulate, and the computed 3D coordinates are then used for the accuracy evaluation (RMSEs) based on a Helmert transformation (7-parameters).
On the other hand, to evaluate the influence of AT results on dense point clouds, an SGM-based MVS algorithm developed at OSU[1] is applied. The OSU-MVS is a dense image matching software developed for DSM true ortho-photo generation from aerial (nadir and oblique) frame camera images. In our analyses, the photogrammetric point clouds are registered with an ICP method to the available GT, and then a cloud-to-cloud comparison is performed.

3.5 Evaluation protocol

In geomatic applications, it is essential to test algorithm performance with metrics specifically tailored for the 3D object space. In our evaluations, the accuracy of tie points extraction methods is evaluated based on:
- RMSEs on GCPs/CPs;
- multiplicity/redundancy (Mean Track Length – MTL);
- cloud-to-cloud comparison with respect to LiDAR ground truth;
- point cloud completeness/accuracy.

4. RESULTS AND ANALYSES

4.1 Dortmund dataset

Different sets of interior parameters are used, and the available 12 GCPs (6 targets and 6 natural points) lead to RMSEs shown in Table 2.

[1] https://u.osu.edu/qin.324/msp/

	Dortmund	**Dublin**
number of images	16 nadir (N), 43 oblique (O)	145 nadir (N), 73 oblique (O)
camera	Pentacam IGI	Leica RCD30, Nikon D800E
image resolution	6132 x 8176 px (N), 8176 x 6132 px (O)	9000 x 6732 px (N), 7360 x 4912 px (O)
focal length	50 mm (N), 80 mm (O)	53 mm (N), 50 mm (O)
pixel size	6 µm (N), 6 µm (O)	6 µm (N), 4.8 µm (O)
platform	airborne	helicopter
overlap (N)	75/80	not constant
average GSD	12 cm (N), 10-14 cm (O)	3.4 cm (N)
Ground Truth (GT)	12 GCPs, LiDAR (not simultaneously acquired)	LiDAR (simultaneously acquired)

Table 1: Main characteristics of the two aerial datasets employed in this work.

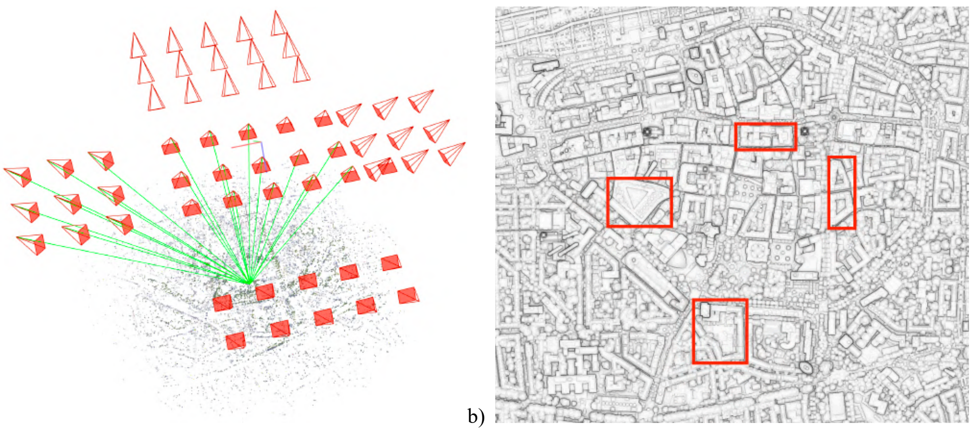
Features / Tie points	BA	Int. Param.	RMSE [m]	Mean reproj. error [px]	MTL	3D points
RootSIFT	COLMAP	f, cx, cy	**0.139**	1.05	2.9	98,575
RootSIFT	COLMAP	f, cx, cy, k1	0.192	1.04	2.9	98,554
RootSIFT	COLMAP	f, cx, cy, k1, k2, p1, p2	0.184	1.04	2.9	98,554
LFNet	COLMAP	f, cx, cy	**0.228**	0.32	2.9	80,082
LFNet	COLMAP	f, cx, cy, k1	0.266	0.32	2.9	79,242
LFNet	COLMAP	f, cx, cy, k1, k2, p1, p2	0.253	0.30	2.9	77,978
KeyNet+AffNet+HardNet	COLMAP	f, cx, cy	0.274	0.37	3.2	75,048
KeyNet+AffNet+HardNet	COLMAP	f, cx, cy, k1	**0.206**	0.37	3.2	75,047
KeyNet+AffNet+HardNet	COLMAP	f, cx, cy, k1, k2, p1, p2	0.536	0.34	3.2	75,010
Metashape	Metashape	f, cx, cy	**0.139**	0.57	3.0	53,994
Metashape	Metashape	f, cx, cy, k1	0.160	0.53	3.0	53,387
Metashape	Metashape	f, cx, cy, k1, k2, p1, p2	0.154	0.53	3.0	53,773

Table 2: AT results for the Dortmund dataset.

Results show that learning-based methods are still slightly worse than RootSIFT. To support the BA metrics provided by COLMAP, Agisoft Metashape[2] is also used, confirming the obtained values. Using the AT results (Figure 3a) with the smallest RMSEs, dense point clouds are derived and compared to the available LiDAR GT (average surface density of ca 10 pts/sqm - Figure 3b).
The cloud-to-cloud analyses (Figure 4 and Table 3) do not reveal significant variations among the dense clouds. Some differences (truncated to 1m) are only present, due to the older LiDAR flight, at the vegetation, and where some new buildings were constructed. Four sub-areas containing only buildings were also chosen, showing differences in the order of 2x the average GSD.

Features	*Mean [m]*	*Std [m]*
RootSIFT	0.368 / 0.270	0.205 / 0.158
LF-Net	0.369/ 0.263	0.204 / 0.151
KeyNet+AffNet+HardNet	0.384 / 0.262	0.206 / 0.156

Table 3: Mean and standard deviation of the cloud-to-cloud differences. The second value is the average of the 4 sub-areas including only buildings (Figure 3b).



a) b)

Figure 3: Dortmund dataset - Recovered image network (a) and the available GT LiDAR (area of ca 1.3x1.3 km) with the four chosen built-up areas (b).

[2] https://www.agisoft.com/

RootSIFT – 30,5 mil. pts *LF-Net – 31 mil. pts* *KeyNet+AffNet+HardNet – 33 mil. pts*

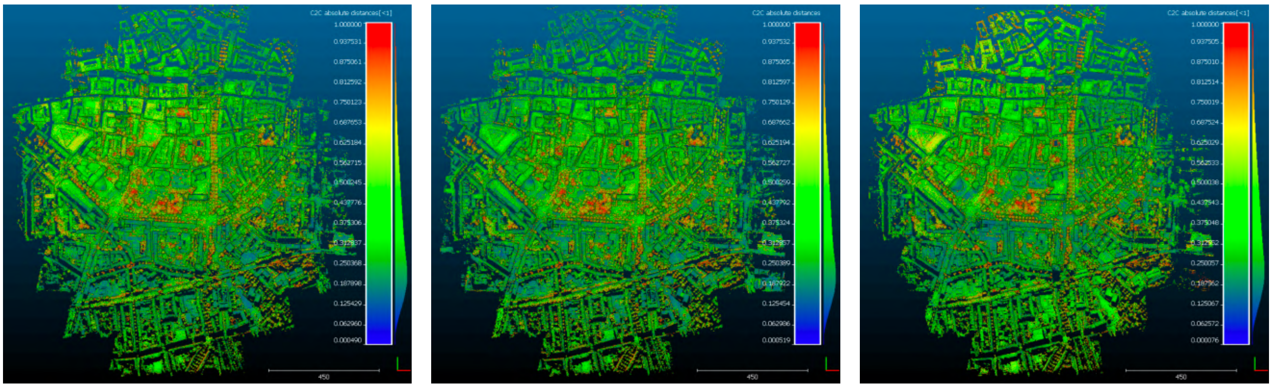


Figure 4: Dortmund dataset - Cloud-to-cloud comparisons of MVS results obtained using the different ATs (RootSIFT, LF-Net, and Key.net). Metrics in Table 3.

4.2 Dublin dataset

AT results for the Dublin dataset are given in Table 4. As no GCPs are provided in the benchmark, metrics are only in image space. Interestingly, LF-Net provides much more points with multiplicity 2 albeit the average MTL is similar for both methods.

Features / Tie points	*Mean reproj. error*	*MTL*	*3D points*
RootSIFT	1.11 px	5.8	204,057
LFNet	0.46 px	5.5	169,992
KeyNet+AffNet+HardNet	0.71 px	6.6	173,697

Table 4: AT results for the Dublin dataset.

Successively, the images are further processed to generate dense point clouds. Figure 5, Table 5, and Table 6 report color-coded views and accuracy values of the cloud-to-cloud assessments, respectively. These analyses do not reveal significant differences originating from the different AT input data.

Features	*Mean [m]*	*Std [m]*
RootSIFT	0.090	0.068
LF-Net	0.079	0.058
KeyNet+AffNet+HardNet	0.088	0.070

Table 5: Mean and standard deviation of the cloud-to-cloud differences.

Features / Tie points	*Precision*	*Recall*	*F1*
RootSIFT	0.996	0.649	0.786
LFNet	0.997	0.627	0.770
KeyNet+AffNet+HardNet	0.989	0.660	0.792

Table 6: Precision (accuracy), recall (completeness) and F1 scores for tolerance $\tau = 0.5$ m.

5. CONCLUSIONS

The paper presented an investigation of learning-based methods to extract tie points in aerial image blocks. AT and MVS results revealed that deep learning could be also a valuable way to find reliable and accurate image correspondences in aerial datasets. Accuracy values provide a clear message that AT could be performed both by hand-crafted and learning-based methods in common AT survey conditions, even if the real potential of these methods lies in managing aerial datasets with images that are difficult to be correctly co-registered due to strong variations in the appearance of the images, and in particular in multi-temporal datasets (Bellavia et al., 2022b; Farella et al., 2022). Moreover, most of these deep architectures still suffer when high camera rotations are present in the datasets. Researchers so far primarily solved the problem by manually rotating images in order to have the same format (Jin et al., 2020), although new methods were developed to match images under large camera rotations (Parihar et al., 2021; Bellavia et al., 2022a).

We believe that deep learning will offer more valuable solutions for photogrammetry in the near future, inspiring and impacting research in our field through collaboration with colleagues in neighbouring disciplines.

REFERENCES

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.

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

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

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

Bellavia, F., Mishkin, D., 2021. HarrisZ+: Harris corner selection for next-gen image matching pipelines. *arXiv*:2109.12925.

Bellavia, F., Morelli, L., Menna, F., Remondino, F., 2022a. Image orientation with a hybrid pipeline robust to rotations and wide-baselines. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Vol. XLVI-2/W1-2022, pp. 73-80.

Bellavia, F., Colombo, C., Morelli, L., Remondino, F., 2022b. Challenges in image matching for cultural heritage: an overview and perspective. Proc. FAPER2022, Springer LNCS, in press.

Becker, C., Rosinskaya, E., Häni, N., D'Angelo, E., Strecha, C., 2018. Classification of aerial photogrammetric 3D point clouds. *Photogramm. Eng. Remote Sens*., Vol. 84, 287-295.

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.

Chen, L., Rottensteiner, F., Heipke, C., 2020a. Feature detection and description for image matching: from hand-crafted design to deep learning. *Geo-spatial Information Science*, Vol. 24.

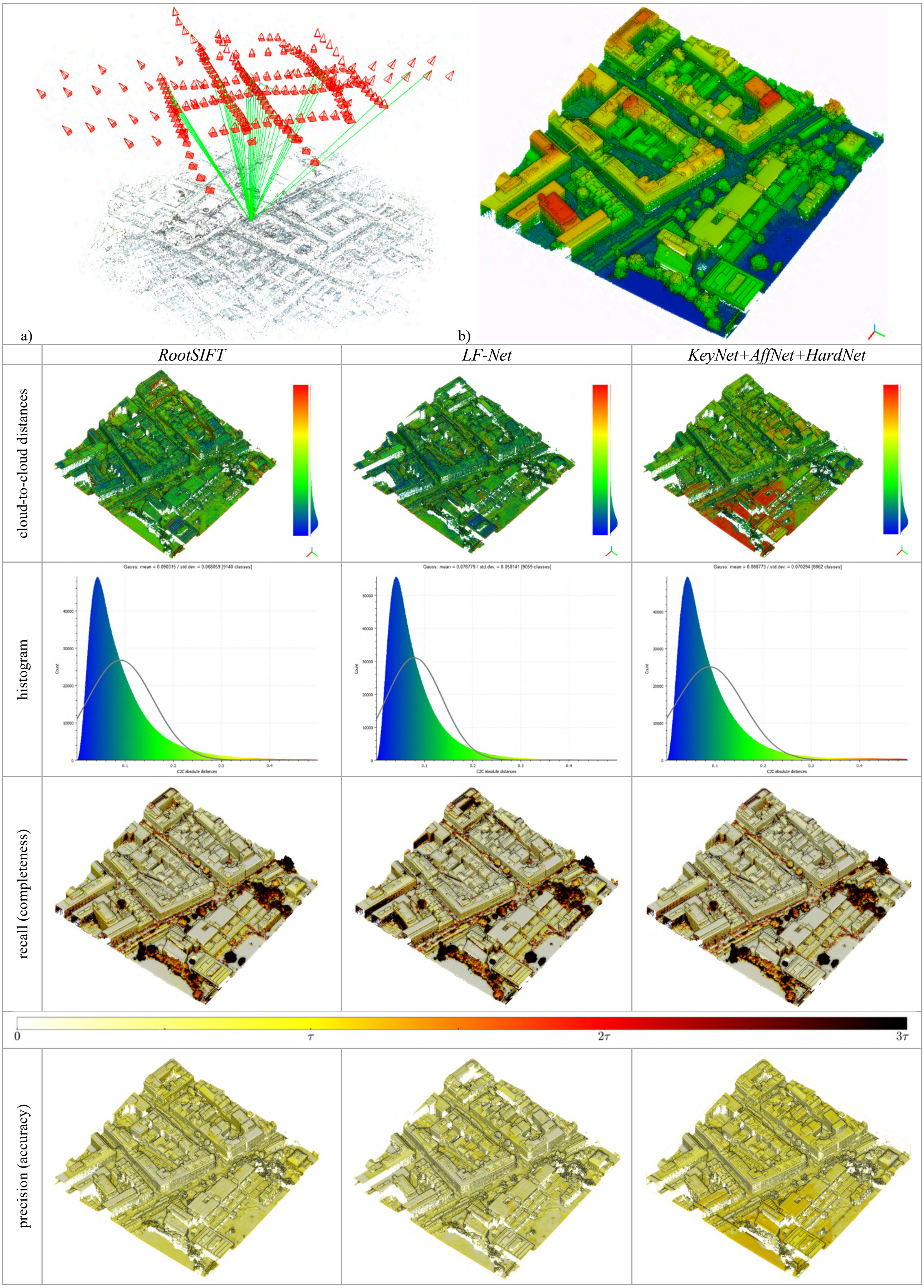


Figure 5: Dublin dataset - Recovered image network (a) and the available LiDAR GT covering an area of ca 250x250m (b). Cloud-to-cloud distances, histogram of point distances, recall and precision for the three MVS results obtained using the different ATs (rootSIFT, LF-Net and Key.net). Metrics in Table 5 and 6.

Chen, L., Rottensteiner, F., and Heipke, C., 2020b. Deep learning based feature matching and its application in image orientation, *ISPRS Annals Photogramm. Remote Sens. Spatial Inf. Sci.*, Vol.2-2020, pp. 25-33.

Choy, C.B., Gwak, J.Y., Savarese, S., Chandraker, M., 2016. Universal correspondence network. Proc. *NIPS*.

DeTone, D., Malisiewicz, T., Rabinovich, A., 2018. SuperPoint: self-supervised interest point detection and description. Proc. *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. Proc. *ICCV*.

Farella, E.M., Morelli, L., Remondino, F., Mills, J. P., Haala, N., Crompvoets, J., 2022. The EuroSDR TIME benchmark for historical aerial images. *ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XXIV ISPRS Congress, in press.*

Gong, J., Ji, S., 2018. Photogrammetry and deep learning. *J. of Geodesy and Geoinformation Science*, Vol. 1(1), pp. 1-15.

Griffiths, D., Boehm, J., 2019. A review on deep learning techniques for 3d sensed data classification. *Remote Sensing*, Vol. 11, 1499.

Habib, A.F., Morgan, M., Lee, Y.-R., 2002. Bundle adjustment with self-calibration using straight lines. *Photogrammetric Record*, Vol. 17(100), pp. 635-650.

Hartmann, W., Havlena, M., Schindler, K., 2015. Recent developments in large-scale tie-point matching. *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 155, pp. 47-62.

Heipke, C., Rottensteiner, F., 2020. Deep learning for geometric and semantic tasks in photogrammetry and remote sensing. *Geo-spatial Information Science*, Vol. 23.

Huang, P.H., Matzen, K., Kopf, J., Ahuja, N., Huang, J.B., 2018. Deepmvs: Learning multi-view stereopsis. Proc. *IEEE CVPR*, pp. 2821-2830.

Li, X., Han, K., Li, S., Prisacariu, V., 2020. Dual resolution correspondence networks. Proc. *NISP*.

Liu, W., Sun, J., Li, W., Hu, T., Wang, P., 2019. Deep learning on point clouds and its application: a survey. *Sensors*, Vol. 19, 4188.

Lowe, D.G., 2004. Distinctive image features from scale invariant keypoints. *Int. J. of Computer Vision*, 60(2), 91-1.

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. *IEEE CVPR*.

Maset, E., Rupnik, E., Pierrot-Deseilligny, M., Remondino, F., and Fusiello, A., 2021. Exploiting multi-camera constraints within bundle block adjustment: an experimental comparison. *ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.,* XLIII-B2-2021, pp. 33-38.

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

Mishkin, D., Radenovic, F., Matas, J., 2018. Repeatability is not enough: Learning affine regions via discriminability. Proc. *ECCV*.

Nex, F., Gerke, M., Remondino, F., Przybilla, H.-J., Bäumker, M., Zurhorst, A., 2015. ISPRS benchmark for multi-platform photogrammetry. *ISPRS Annals Photogramm. Remote Sens. Spatial Inf. Sci.*, Vol. II-3/W4, pp. 135-142

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

Özdemir E, Remondino F, Golkar A., 2021. An efficient and general framework for aerial point cloud classification in urban scenarios. *Remote Sensing*, Vol. 13(10):1985.

Parihar, U.S., Gujarathi, A., Mehta, K., Tourani, S., Garg, S., Milford, M., Krishna, K.M., 2021. RoRD: Rotation-robust descriptors and orthographic views for local feature matching. Proc. *IEEE CVPR*.

Pautrat, R., Larsson, V., Oswald, M. R. and Pollefeys, M., 2020. Online invariance selection for local feature descriptors. Proc. *ECCV,* pp. 707-724.

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

Qin, R., Tian, J., Reinartz, P., 2016. 3D change detection - Approaches and applications. *ISPRS Journal of Photogrammetry & Remote Sensing*, Vol. 122, pp. 41-56.

Qin, R., Gruen, A., 2021. The role of machine intelligence in photogrammetric 3D modelling – an overview and perspectives. *Int. Journal of Digital Earth*, Vol. 14(1), pp. 15-31.

Relja, A., Zisserman, A., 2012. Three things everyone should know to improve object retrieval. Proc. *IEEE CVPR*.

Remondino, F., Menna, F. Morelli, L., 2021. Evaluating hand-crafted and learning-based features for photogrammetric applications. *ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, *43*, 549-556.

Ressl, C., Karel, W., Piermattei, L., Puercher, G., Hollaus, M. and Pfeifer, N., 2020. Multi-epoch bundle block adjustment for processing large dataset of historical aerial images. In *EGU General Assembly Conference Abstracts* (p. 22544).

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

Riba, E., Mishkin, D., Ponsa, D., Rublee, E., Bradski, G., 2020. Kornia: an open source differentiable computer vision library for pytorch. Proc. IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3674-3683.

Rocco, I., Cimpoi, M., Arandjelovic, R., Torii, A., Pajdla, T., Sivic, J., 2018. *Neighbourhood consensus networks*. Proc. *NIPS*.

Ruano, S., Smolic, A., 2021. A Benchmark for 3D Reconstruction from Aerial Imagery in an Urban Environment. Proc. *VISAPP*, pp. 732-741.

Rupnik, E., Nex, F., Remondino, F., 2013. Automatic orientation of large blocks of oblique images. *ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Vol. 40(1/W1), pp. 299-304.

Rupnik, E., Nex, F., Toschi, I., Remondino, F., 2015. Aerial multi-camera systems: Accuracy and block triangulation issues. *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 101, pp. 233-246.

Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A., 2020. SuperGlue: Learning feature matching with graph neural networks. Proc. *IEEE CVPR*.

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

Shan, J., Hu, Z., Tao, P., Wang, L., Zhang, S., Ji, S., 2020. Toward a unified theoretical framework for photogrammetry. *Geo-spatial Information Science*, Vol. 23 (1), pp. 75-86.

Schenk, T., 2004. From point-based to feature-based aerial triangulation. *ISPRS Journal of Photogrammetry & Remote Sensing*, Vol. 58, pp. 315-329.

Schonberger, J. L., Hardmeier, H., Sattler, T., Pollefeys, M., 2017. Comparative evaluation of hand-crafted and learned local features. Proc. *CVPR*, pp. 1482-1491.

Stathopoulou, E.K., Welponer, M., Remondino, F., 2019. Open-source image-based 3D reconstruction pipeline: review, comparison and evaluation. *ISPRS Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Vol. XLII-2/W17, pp. 331-338.

Sun, J., Shen, Z., Wang, Y., Bao, Y., Zhou, X., 2021. LoFTR: Detector-free local feature matching with Transformers. Proc. *IEEE CVPR*.

Tyszkiewicz, M.J., Fua, P., Trulls, E., 2020. DISK: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, Vol. 33.

Tian, Y., Fan, B., Wu, F., 2017. L2-net: Deep learning of discriminative patch descriptor in euclidean space. Proc. *IEEE CVPR*, pp. 661-669.

Triggs, B., McLauchlan, P.F., Hartley, R.I., Fitzgibbon, A.W., 2000. Bundle Adjustment — A Modern Synthesis.

Truong, P., Apostolopoulos, S., Mosinska, A., Stucky, S., Ciller, C., De Zanet, S., 2020. GLAMpoints: Greedily Learning Accurate Match points. *Arvix*: 1908.06812v3.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. Proc. *NeurIPS*.

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

Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L., 2018. MVSnet: Depth inference for unstructured multi-view stereo. Proc. *ECCV*.

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), 3247.

Yi, K. M., Trulls, E., Lepetit, V., Fua, P., 2016. LIFT: Learned invariant feature transform. Proc. *ECCV,* pp. 467-483.

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.

Zhu, X., Tuia, D., Mou, L., Xia, G., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: a comprehensive review and list of resources. *IEEE GRSS Magazine*, Vol. 5(4), pp. 8-36.