The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-M-2-2023
29th CIPA Symposium "Documenting, Understanding, Preserving Cultural Heritage:
Humanities and Digital Technologies for Shaping the Future", 25–30 June 2023, Florence, Italy

THE EVALUATION OF STRUCTURE-FROM-MOTION WORKFLOW WITH THE TLS SYNTHETIC IMAGES SIMULATOR – THE CULTURAL HERITAGE APPROACH

J. Markiewicz [1*], M. Kowalczyk [1], K. Karwel [1], P. Kot [2] and Ł. Markiewicz[3]

[1] Faculty of Geodesy and Cartography, Warsaw University of Technology, Warsaw, Poland – (jakub.markiewicz, michal.kowalczyk, karol.karwel)@pw.edu.pl
[2] Built Environment and Sustainable Technologies (BEST) Research Institute, Liverpool John Moores University, Liverpool, United Kingdom, p.kot@ljmu.ac.uk
[3] Institute of Micromechanics and Photonics, Warsaw University of Technology, Warsaw, Poland, lukasz.markiewicz.dokt@pw.edu.pl

KEY WORDS: Cultural Heritage, Evaluation, Open-Source Software, SfM/MVS, TLS

ABSTRACT:

Modern measurement technologies such as Terrestrial Laser Scanning or combined Structure-from-Motion with Multi-View Stereo are commonly utilised to monitor, preserve and document cultural heritage objects and sites. For this reason, it is essential to know the capabilities and limitations of the sensor used, the data processing methods, and in particular, the orientation of the images. However, these algorithms tackle different errors and have different effects on the final accuracy of images orientation. For this reason, it is essential to know how the algorithms implemented in the Structure-from-Motion approach work. Due to the impossibility of obtaining this information for commercial solutions, it is necessary to use synthetic data to assess the quality of the SfM process. Therefore, this article aims to present the method of evaluation of SfM approach implemented in commercial Agisoft Metashape and COLLMAP open-source software based on the synthetic data generated from TLS point clouds of three different Cultural Heritage sites. In addition, obtained results were compared with the author's SfM approach based on BRISK, FAST, CenSurE, SIFT and SURF (and its Affine detectors equivalents) detector implemented (Fig. 1) and Learned-based -feature extraction approach SuperGlue and LoFTR. The second aim of this research is to propose an application to automatically generate scalable benchmark based on point clouds or 3D models of cultural heritage objects.

1. INTRODUCTION

Nowadays, cultural heritage is an integral part of modern societies, and preserving tangible and intangible evidence of the past is necessary. Modern non-destructive measurement technologies are commonly used to monitor and preserve cultural heritage.

The development of image and range-based methods for 3D shape reconstruction has contributed to their use in the inventory of cultural heritage objects and sites. The increasing trend in digital technologies led to the development of various 2D and 3D data sets in several scenarios (i.e., indoor, laboratory, outdoor, urban and buildings) that allow for the evaluation of novel computer vision and photogrammetry algorithms based on the different tasks (i.e., image matching, structure-from-motion, image retrieval and SLAM) (Marelli et al., 2023).

The evaluation of the data sets enables the assessment of the quality of sensors, measurement platforms or data processing algorithms used, and in particular, the methods for processing of 2D and 3D data by photogrammetric and machine vision methods, benchmarked against high-quality and accurate ground truth-data (Bakuła et al., 2019; Marelli et al., 2023). Various ready-to-use benchmarks consisting of real (Gabara and Sawicki, 2023; Seitz et al., 2006; Strecha et al., 2008) or synthetic (Aanæs et al., 2012; Marelli et al., 2023) data in the form of point clouds, ground photos, information on interior and exterior orientation parameters, depth maps and more are available in the literature.

The literature analysis led the authors to prepare a new benchmark dedicated to (but not limited to) verifying the performance of image-matching algorithms for cultural heritage sites. In contrast to existing solutions, the Authors decided to develop and make available an application (the synthetic simulator) that allows the generation of any (at the user's discretion) 'virtual images' based on any point cloud or mesh model.

In addition, this article aims to present the method of evaluation of SfM approach implemented in commercial Agisoft Metashape and COLLMAP open-source software based on the synthetic data generated from TLS point clouds at three different Cultural Heritage sites, namely Royal Castle in Warsaw (Fig. 3a) - Benchmark 1. (Fig, 3b, c) Museum of King Jan III's Palace in Wilanów. Obtained results were compared with the author's SfM approach based on BRISK, FAST, CenSurE, SIFT and SURF (and its Affine detectors equivalents) detector implemented and Learned-based -feature extraction approach SuperGlue and LoFTR.

1.1. Structure-from-Motion and image matching

Nowadays, the combined Structure-from-Motion (SfM) and Multi-View Stereo (MVS) methods (in addition to laser scanning) are widely used in the inventory of historical objects and sites. The combined SfM and MVS workflow allows for 3D shape reconstruction based on the collection of images. The typical SfM approach is based on two parts: (1) the corresponding search phase and (2) incremental reconstruction. Each step can use different algorithms. However, these algorithms tackle various errors and have other effects on the final accuracy of images orientation. For this reason, it is essential to know how the algorithms implemented in the Structure-from-Motion approach work. Due to the impossibility

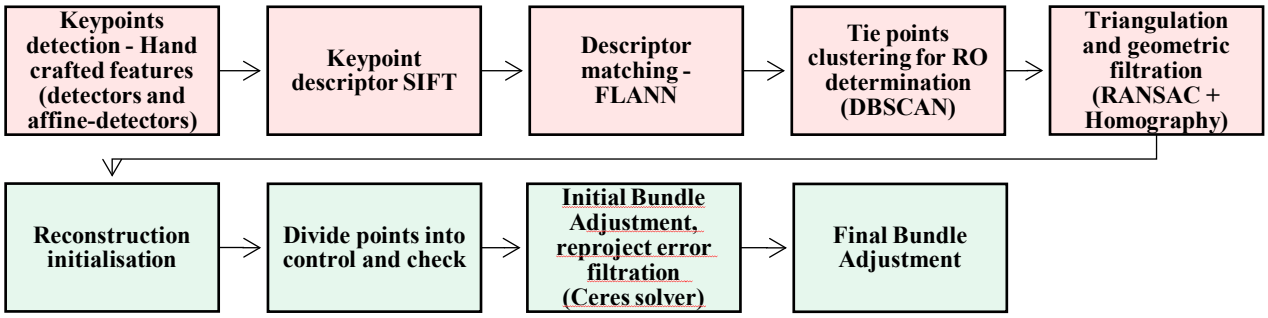

Figure. 1 The diagram of the proposed Structure-from-Motion approach; red rectangles – correspond to search part; green rectangles – incremental reconstruction

of obtaining this information for commercial solutions, it is necessary to use synthetic data to assess the quality of the SfM process.

There are currently two approaches to keypoints detection based on a group of Hand-crafted algorithms, i.e., SIFT (Lowe, 2004) or SURF (Bay and Ess, 2008), and a Learned-based -feature extraction approach i.e., SuperGlue or LoFTR. When using hand-crafted detectors, keypoints are detected based on the grayscale gradients values in the nearest neighbourhood (blob detectors, i.e., SIFT , SURF or CenSurE) or by comparing the differences between grayscale compared to the analysed pixel (corner detectors, i.e., FAST (Rosten and Drummond, 2006), BRISK (Leutenegger et al., 2011)).
In recent years, various novel learning-based solutions were developed to overcome the existing limitations of hand-crafted methods (Verdie et al., 2015). Various solutions, namely (1) detect-then-describe wherethe detector (Barroso-Laguna et al., 2019; Verdie et al., 2015) and the descriptor (Ebel et al., 2019; Mishchuk et al., 2017) can be both learned methods or a combination of hand-crafted and learning- based. Other approaches, called (2) end-to-end, jointly optimise the entire pipeline to extract sparse image correspondences, e.g., SuperPoint (DeTone et al., 2017), SuperGlue (Sarlin et al., 2019), DISK (Tyszkiewicz et al., 2020). End-to-end methods were used 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, researchers like Choy et al., 2016; Rocco et al., 2018; Li et al., 2020 proposed (3) end-to-end detector-free local feature matching methods that eliminate 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.

2. MATERIALS AND METHODS

2.1. The structure-from-motion evaluation

An essential step in the orientation of the images using the Struture-from-Motion method is the detection of keypoints, which are then used to determine the elements of the relative orientation and to determine the internal orientation parameters in the self-calibration process. For this purpose, it was decided to analyse the impact of the selection of hand-crafted points (BRISK, FAST) and blob detectors (SIFT, SURF and CenSurE) together with their affine counterparts and to compare the results obtained with the learned-bassed approach based on the LoFTR and SuperGlue algorithms. In addition, the results were compared with those from ready solutions implemented in Colmap and Agisfot software. Figure 1 shows the proposed schematic of the author's SfM approach.
The evaluation process of the various SfM approaches were a multi-stage process consisting of the analysis of (1) the number of tie points, (2) the reprojection error values of the automatically detected and matched tie points, (3) the distribution of tie points in space and on the image; (4) the analysis of the accuracy of the determination of the exterior orientation parameters; (5) the analysis of the correctness of the determination of the internal orientation elements; (6) the quality of the point clouds generated from the oriented images. For this purpose, 'virtual images' were generated based on what it was possible to simulate image distortion parameters.

2.2 The synthetic images simulator

The aim of this paper is the evaluation of the Structure-from-Motion workflow with the TLS synthetic images simulator. For this purpose, benchmarks were generated based on point clouds of historic interiors located at the Royal Palace in Warsaw and the Museum of King Jan III's Palace in Wilanów. In preparing benchmarks based on point clouds, it was decided to prepare an application that allowed the preparation of (1) any configuration of 'virtual images', (2) any geometric distortion and (3) external orientation elements. The Synthetic Images Simulator (Fig. 2) generated "virtual images" based on the OpenGL graphics environment.

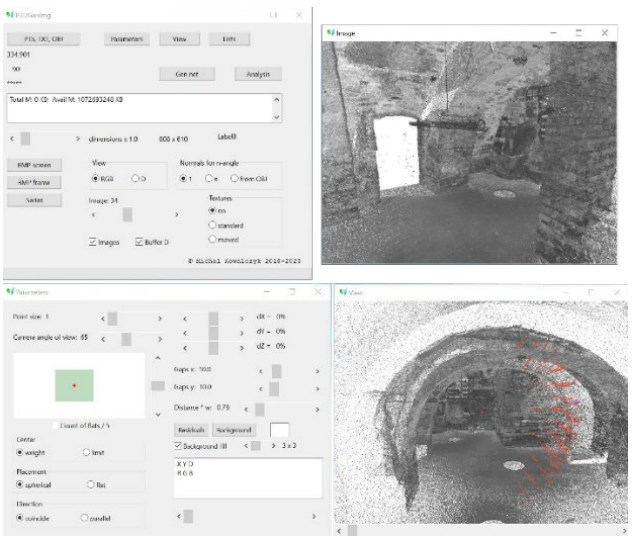

Figure. 2 The Graphical User Interference of The Synthetic Images Simulator

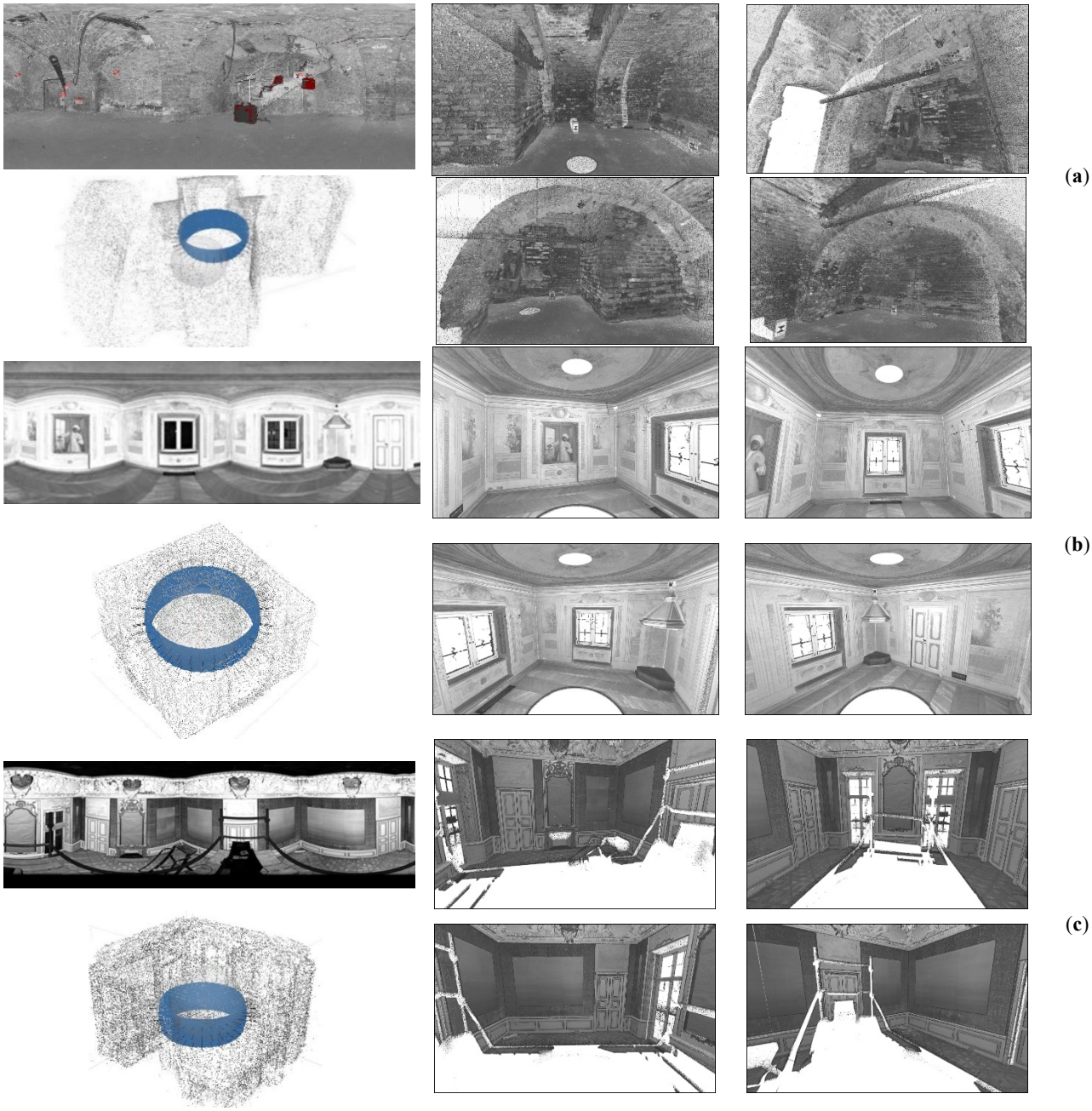

Figure. 3 Examples of images distribution: (a) Benchmark 1 (b) Benchmark 2 and (c) Benchmark 3

Individual image is designed with user-defined parameters such as (1) image size and resolution, (2) the focal length, (3) camera position and (4) interior orientation parameters, namely radial and tangential distortion. In the first stage, the projection centre and camera angle are defined. The point clouds are reprojected onto the reference plane, considering the model transformation matrix, projection and observation range. The texture was based on the TLS intensity considering the appropriate setting of the image depth buffer. The second data set that can generate virtual images, which are used to create 3D models saved as a regular or irregular mesh in ply or obj format. The way the virtual images are generated, and the parameters determination are set is the same as for point clouds.

2.3. Test site description

In preparation for the ground-truth data that are part of the Benchmark based on the innovative *The synthetic images simulator*, it was decided to use 3 groups of point clouds representing historic interiors characterised by different numbers of ornaments, decorations and geometric complexity. For this purpose, point clouds of the basement rooms located on the lowest floor of the Tin-Roofed Palace, which is a part of the architectural complex of the Royal Castle in Warsaw (Fig. 3a) - Benchmark 1. (Fig, 3b, c) the Museum of King Jan III's Palace in Wilanów were used for Benchmark 2 and 3.

The point clouds represent Benchmark 1 are constructed of bricks filled with mortar. The historical basement has an irregular shape with a ceiling in the shape of arches, with a maximum height of approximately 3.2 m and a minimum of about 2.1 m.

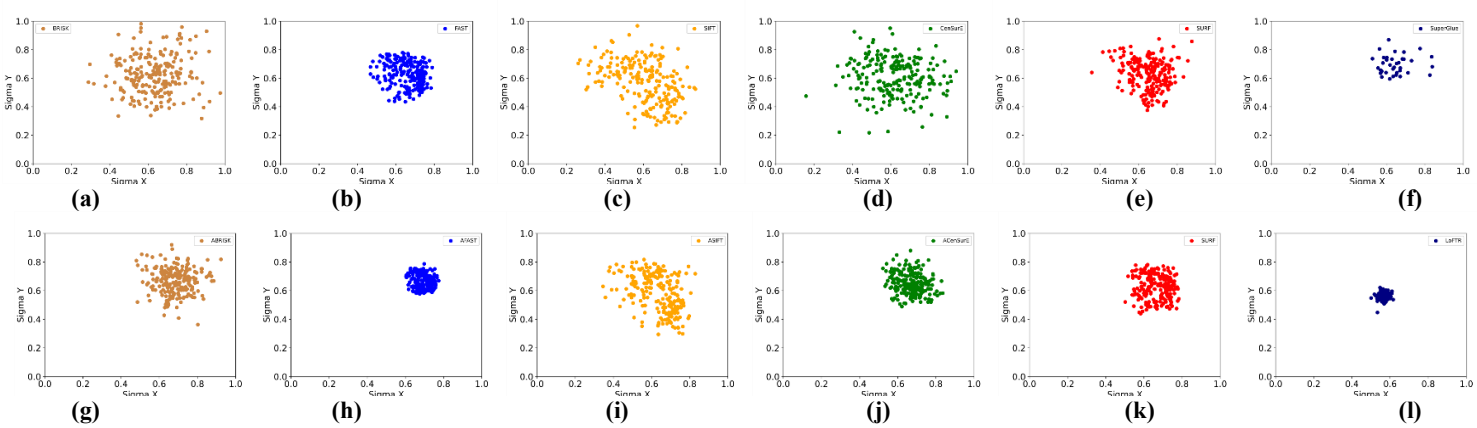

Figure. 4 The diagram of the relationship between reprojection error for X and Y coordinates for (a) BRISK, (b) FAST, (c) SIFT, (d) CenSurE, (e) SURF, (f) SuperGlue, (g) ABRISK, (h) AFAST, (i) ASIFT, (j) ACenSurE, (k) ASURF and (l) LoFTR.

Due to its historical character and the prevailing humidity conditions, part of the room has damp walls and fragments of bricks are crumbling. TLS data used in this Benchmark was acquired by phase-shift scanners Z + F 5006h from different positions and heights with angular resolution and point resolution 6.3 mm/10 m. Benchmark 2 "The Chamber with a Parrot"- Museum of King Jan III's Palace at Wilanów is characterised by the small number of ornaments and the lack of bas-reliefs, facets, or fabrics on the walls. In this Test Site, the walls were painted with patterns, which imitated spatial effects. Data was acquired by the Z+F 5006h scanner with angular resolution and point resolution 6.3 mm/10 m. Due to the 360°/320° prohibition to place marked points on historical surfaces, automatically detected points defined as check points, were used for the accuracy analysis. The dimensions of "The Chamber with a Parrot" are approximately 4.2 m x 4.2 m x 2.6 m. The Benchmark 3 called "The Queen's Bedroom"-Museum of King Jan III's Palace at Wilanów Test site III was characterised by geometric complexity in the form of rich ornaments, bas-reliefs, and facets. Moreover, there were mirrors in golden frames, decorative fireplaces, fabrics, etc., hanging on the walls."The Queen's Bedroom" is approximately 6.4 m x 7.3 m x 5.3 m. Similar to the benchmark 2 , the Z+F 5006h laser scanner was used for point cloud acquisition.

3. RESULTS

3.1. Tie points determination on pairs of images.

The first step of SfM workflow evaluation involves pair keypoints detection and matching on pairs of images. To do this, reprojection error values were assessed for pairs of matched keypoints in a descriptor-matching process (hand-crafted methods) or using a learning-based approach. Thanks to the knowledge of the external orientation elements of the images, it was possible to determine the values of these errors and to perform outlier filtering for error values greater than 1 pixel.

When assessing the mean values of the number of points detected in the image pairs (Tab. 1), the following relationships can be observed: **(1)** The smallest number of points was obtained for the BRISK detector (about 36 points on average – Benchmark 1, 12 – Benchmark 2 and 44 – Benchmark 3) and the highest for AFAST (9028 points on average – Benchmark 1, 8259 – Benchmark 2, 7047 – Benchmark 3). **(2)** The same relationship was observed for the standard deviation number of points - for the BRISK detector it was 13.5 points for the AFAST 4053. **(3)** The ratio between the affine-detectors and detectors averages are for ABRISK: 4.5-times, AFAST – 4.4-times, ASIFT – 11.4 -times, ACenSurE – 8.9 – times and ASUR – 7.4 – times. **(4)** For the detectors, an average of 3.5 times fewer points were obtained for the ABRISK detector, 4.2 times for AFAST, 13 times for ASIFT, 7 times for ACenSurE and 8.3 times for ASURF than for the affine detectors. **(5)** Comparing the number of points detected for the Hand-crafted and the Learning approach, it can be seen that an average of 287 and 320 points were detected for all 3 benchmarks using the LoFTR and SuperGlue methods. These methods have approximate standard deviations, which are LoFTR: 130 - Benchmark 1, 141 - Benchmark 2 and 158 - Benchmark 3 and SuperGlue: 127 - Benchmark 1, 150 - Benchmark 2, 66 - Benchmark 3, respectively.

		Benchmark 1				Benchmark 2				Benchmark 3			
	Detector	Avg.	Min.	Max	Std.	Avg.	Min.	Max	Std.	Avg.	Min.	Max	Std.
Hand-crafted	BRISK	35.8	8	75	13.5	12.0	3	34	5.7	43.9	6	113	21.0
	ABRISK	124.2	16	344	60.1	79.6	14	173	33.4	148.9	17	386	82.1
	FAST	2155.9	342	6004	1194	2304.0	356	4719	1062.6	1315.8	163	2834	572.7
	AFAST	9027.6	1412	21784	4053	8258.6	1485	18070	4496.3	7047.2	32	15228	3332.0
	SIFT	271.8	46	749	126.6	812.0	196	1339	288.4	307.1	100	580	111.5
	ASIFT	3573.7	583	8495	1631.2	6398.7	998	13456	3264.6	4066.3	496	9290	1856.9
	CenSureE	40.1	5	110	20.1	33.1	8	68	14.2	85.2	7	199	45.2
	ACenSurE	285.6	26	750	143.9	356.6	64	6952	141.6	666.3	36	201	405.1
	SURF	495.9	71	1517	304.9	1113.9	14	2089	445.3	511.5	101	1047	206.2
	ASURF	4155.4	416	12469	2784.6	7042.2	1340	16856	3991.7	4353.2	656	8537	1894.2
Learning approach	SuperGlue	176.6	65	439	97.3	451.8	158	697	149.6	231.1	93	358	66.3
	LoFTR	283.7	51	947	129.6	383.4	89	832	141.1	293.7	32	1062	158.3

Table. 1 The statistic of automatically matched tie points on stereo-pairs of images for all benchmarks.

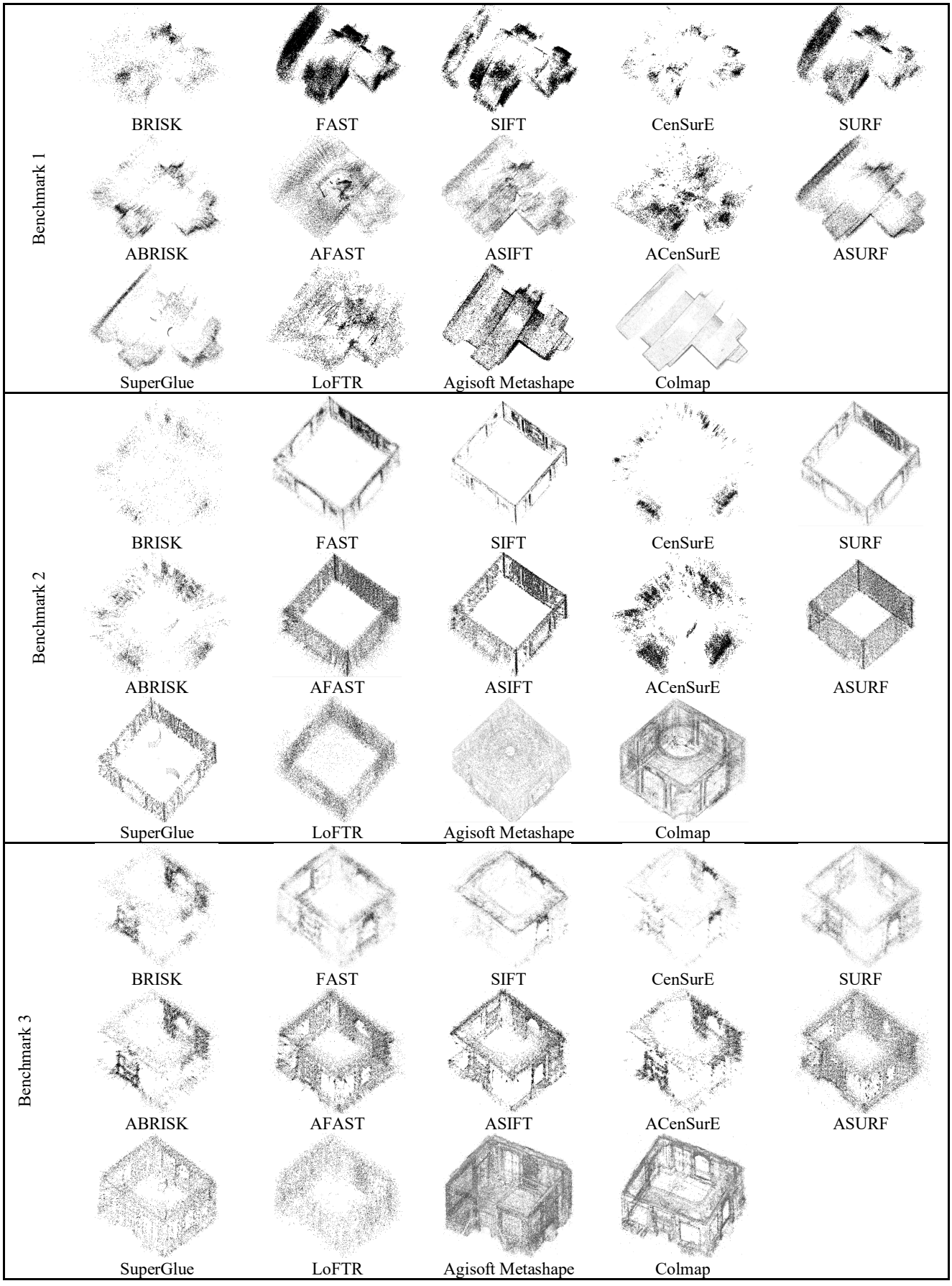

Figure. 5 Examples for tie point distributions for all Benchmarks

Software/ Detector	Benchmark 1					Benchmark 2					Benchmark 3				
	No. of tie point	Mean Reprojection Error [pix]	Geolocation RMSE [mm]			No. of tie point	Mean Reprojec-tion Error [pix]	Geolocation RMSE [mm]			No. of tie point	Mean Reprojec-tion Error [pix]	Geolocation RMSE [mm]		
			X	Y	Z			X	Y	Z			X	Y	Z
Agisoft Metashape	43,449	0.70	0.4	0.4	0.2	28,422	0.60	0.2	0.2	0.3	50,892	0.90	0.2	0.2	0.4
Colmap	56,280	0.39	0.5	0.7	0.6	98,516	0.41	0.4	0.3	0.2	74,920	0.45	0.3	0.3	0.5
BRISK,	8,879	0.32	0.4	0.5	0.2	3,018	0.37	8.2	6.3	4.8	10,538	0.34	0.3	0.4	0.2
ABRISK	68,801	0.37	1.7	1.3	1.5	68,801	0.43	0,7	0,5	0,7	68,801	0.47	1.2	1.9	1.1
FAST	68,801	0.76	0.1	0.2	0.5	58,163	0.21	0.2	0.2	0.1	41,554	0.25	0.1	0.1	0.1
AFAST	68,801	0.40	1.4	1.6	1.1	68,801	0.32	0.5	0.5	0.6	68,801	0.48	0.9	0.7	1.1
SIFT	68,074	0.16	0.6	0.6	0.5	68,290	0.14	0.2	0.1	0.1	32,860	0.18	5.7	5.1	3.1
ASIFT	68,801	0.38	0.1	0.1	0.1	68,738	0.21	0.3	0.3	0.4	68,801	0.35	0.6	0.7	0.9
CenSurE	10,088	0.26	0.1	0.2	0.1	8,303	0.24	4.7	3.6	2.2	18,277	0.27	0.1	0.1	0.1
ACen-SurE	68,801	0.40	0.1	0.1	0.1	68,801	0.52	0.6	0.6	1.1	68,801	0.43	0.7	0.9	0.9
SURF	41,573	0.24	0.5	0.4	0.3	56,235	0.21	0.2	0.2	0.1	42,783	0.25	0.1	0.1	0.1
ASURF	68,801	0.39	0.5	0.5	0.7	68,801	0.37	0.6	0.5	0.4	68,801	0.44	1.0	1.2	1.1
LoFTR	42,881	0.54	0.1	0.1	0.1	42,881	0.57	0.5	0.5	0.9	34,241	0.47	0.5	0.6	0.7
Super-Glue	34,240	0.32	0.4	0.4	0.3	34,240	0.27	0.3	0.3	0.2	34,240	0.3	0.4	0.4	0.4

Table. 2 The quality assessment of tie point detection and bundle adjustment on synthetic images without distortion and self-calibration (focal length and principal point) for all Benchmarks

(6) The analyses show that a higher number of tie points can be obtained using hand-crafted and affine-detectors and that the results obtained approximate those obtained using the SIFT detector.

To assess the impact of the choice of tie points detection methods, the effect of RMSE rejection error values was also analysed on 250 image pairs. Figure 4 shows an example of scatter points for which the X-axis value corresponds to the reprojection error X value and the Y-axis value to the Y reprojection error value. The reprojection error dispersion results show that: (1) The smaller dispersion of values was obtained for the Learning Approach than for the Hand-crafted methods. This demonstrates the greater reproducibility of the detection of tie points in different image pairs. (2) A lower dispersion reprojection error was obtained for hand-crafted detectors, particularly for the ABRISK detector. The lowest dispersion values were obtained for the AFAST, ACenSurE and ASURF detectors, respectively. (3) It is important to emphasise that the repeatability of the distribution of error values for all benchmarks for all methods except the BRISK detector are similar and repetetable. Surprisingly, points were detected worse on Benchmark 2, characterised by good texture and many corners on the wall paintings.

3.2. Tie points determination on the whole photogrammetric block based on images without distortion.

To assess the impact analysis of the tie point detection methods used, the number of tie points detected was analysed, as well as the reprojection error values and the error values of determining the linear values of the external orientation elements (Table 2).

To assess the number, distribution and value of reprojection error, it was decided to use two scenarios to determine these points. Firstly, it was decided to use all the binding points detected using Agisoft Metashape and Colmap software. For the analysis of the data detected using the proposed SfM approach, the points were filtered according to the following scheme: (1) dividing the image into 64 equal sub-areas and detecting 5 points located according to the following scheme - 1 point close to the centre of gravity and 4 points at a distance of 2/3 from the centre of gravity - detection done on the image pair, (2) searching for the remaining corresponding points on other images to increase the number of bundles. This approach allowed a revascularisation of the number of tie points distribution, which is crucial in determining the camera calibration parameters in the self-calibration process.

From the analysis of the distribution of tie points shown in Figure 5, it is apparent that the best distribution was obtained for the Agisoft Metashape and Collmap software and the Hand-crafted detectors (AFAST, ASIFT and ASURF), Learning approach SuperGlue and LoFTR, respectively. For normal-case detectors and ABRISK and ACenSurE, no uniform distribution of tie points was possible to obtain, and the number of them is significantly lower (Fig. 5 and Tab. 2). As mentioned earlier, this may affect the determination of internal orientation elements with lower accuracy. In the case of the tie points detected by the Lear-based approach, the number of points is much smaller (compared to the others), but the geometric distribution of the points nevertheless allows the correct orientation of the images and self-calibration.

Comparing the quality of the detected tie points (Table 2) based on Mean Reprojection Error analysis shows that the worst results were obtained for hand-crafted detectors, FAST on Benchamrk1, ACenSurE om Benchmark 2 and and ABRISK\ACenSureE on Benchmark 3.

It was decided to use known exterior orientation parameters (EOP) to assess geolocation accuracy. In the bundle adjustment step, 1/3 EOP were used for images orientation and 2/3 for independent quality assessment. From the results in Table 2, it can be seen that similar results of less than 1 mm were obtained for all solutions (except AFAST and ABRISK for Benchmark 1, BRISK and CenSurE for Benchmark 2 and ABRISK, AFAST, SIFT and ASURF for Benchamrk 3).

Software/ Detector	Benchmark 1				Benchmark 2				Benchmark 3			
	Interior orientation error [pix]				Interior orientation error [pix]				Interior orientation error [pix]			
	f	cx	cy	K1	f	cx	cy	K1	f	cx	cy	K1
Agisoft Metashape	-0.65	-0.03	-0.57	0.012	-0.47	0.05	-0.52	0.015	-0.47	0.51	0.90	0.013
Colmap	-0.31	-0.16	0.36	0.021	-2.14	0.02	-1.03	0.013	-0.33	-0.07	-0.36	0.012
BRISK	10.1	-5.4	-0.8	0.011	8.67	4.28	-1.27	0.012	2.28	3.97	-4.37	0.015
ABRISK	14.35	-0.31	-4.65	0.009	-2.14	-0.01	-1.58	0.010	-0.13	-0.17	-1.86	0.011
FAST	12.3	-1.6	-0.7	0.014	1.16	-0.89	-0.97	0.012	5.72	-1.82	1.45	0.012
AFAST	13.16	0.62	-4.8	0.013	-0,92	0,02	-1.79	0.014	0.01	0.30	2.58	0.011
SIFT	2.64	-2.1	-0.5	0.009	-0.37	0.04	-0.43	0.014	7.51	-1.64	-12.69	0.007
ASIFT	-0.54	-0.23	-1.30	0.006	-1.19	0,04	-1,47	0.008	-0.10	-0.02	2.28	0.006
CenSurE	5.60	-3.1	-0.5	0.006	7.55	-1.79	0.63	0.005	3.03	-0.21	-0.44	0.008
ACenSurE	-0.11	-0.13	0.02	0.006	-1.11	1.03	-1.48	0.009	-0.37	-0.06	0.20	0.006
SURF	5.52	-1.4	0.7	0.011	1.35	-1.48	-0.91	0.013	3.70	-0.13	-0.60	0.015
ASURF	-0.16	-0.1	-0.15	0.009	-0.70	0.01	-1.64	0.007	0.42	-0.04	-1.30	0.011
LoFTR	-0.22	-0.1	0.22	0.013	-0.80	0.04	-1.18	0.012	-0.47	0.10	0.38	0.013
SuperGlue	-0.53	-0.06	0.39	0.011	-1.23	0.03	-0.98	0.012	-0.42	0.01	0.52	0.014

Table. 3 The quality assessment of tie point detection and bundle adjustment on synthetic images with distortion self-calibration (focal length and principal point) for all Benchmarks

Convergent results in the accuracy of the external orientation and the values of the reprojection error obtained using the proprietary SfM approach and point filtration prove the effectiveness of the proposed approach. It is recommended to use hand-crafted affine-detectors blob, in particular ASIFT or ASURF or learning-based approach based on SuperGlue and LoFTR algorithms.

3.3. The analysis of the self-calibration correctness

The correctness of the self-calibration process was divided into the 3 parts: (1) focal length, (2) principal point and (3) radial distortion determination. For this purpose, it was decided to vary the focal length value with a value between 10 and 50 in increments of 10, the principal point with a value between 2 and 10 in increments of 2 pixels, and the radial distortion value k1 with a value between -6E-2 and 6E-2 in increments of 1E-2. Due to the insignificant changes in the focal length and principal point values, it was decided to include the average values for the aforementioned parameters in Table 3.

From the values shown in Table 3 of the differences between the calibrated focal value and the reference value, it can be seen that the best values were obtained for Benchmark 1 for the LoFTR and SuperGlue methods, respectively, using the Colmap and Agisoft Metashape software and the hand-crafted algorithms ASIFT, ACenSurE or ASURF. It should be noted that errors did not exceed 1 pixel. For the corner detectors, i.e., FAST/AFAST and BRISK/ABRISK, these values ranged from 10.1 to 14.35 pixels. For the blob detectors SIFT, SURF, and CenSurE these values were smaller at 2.64, 5.52 and 5.60 pixels, respectively.

For Benchmark 2, worse focal lenght determination values were obtained, for all cases except the results obtained for Agisoft Metashape. This was due to the fact that the test field used was flat and there were no significant changes in depth. It should be noted, however, that the use of affine-crafted detectors has improved the accuracy of focal length determination for ABRISK, ACenSurE and ASURF due to the detection of more points. The use of a learnt-based approach (i.e., LoFTR and SuperGlue) allowed the focal length value to be determined close to that determined by ASURF. As in the case of focal length, improvements can be seen in the accuracy of the determination of the position of the centre of projection in the self-calibration process using affine-detectors.

For Benchmark 3, which is characterized by a greater change in depth, as in the case of Benchmark 2, the use of affine-detectors significantly increases the accuracy of determination of the focal length and the principal point. It can also be stated that similar error values were obtained for affine-detectors, learn-based approach and using Agisoft and Colmap software.

The simulated radial distortion correction tests show that for the Hand-crafted and the Learning-based approach and Agisoft and Colmap, similar constant values were obtained for all Benchmarks. This demonstrates the stability of the proposed SfM solution based on a proprietary point filtering approach, and the results obtained are close to those obtained using commercial software.

4. CONCLUSIONS

The rapid development of new technologies and their rapid entry into the world of cultural heritage has resulted in a need of verifying and testing its accuracy and reliability in the context of the need for the quality and accuracy of architectural documentation. For this reason, it is advisable to control the methods, schemes and algorithms used to generate architectural documentation. To this end, using benchmarks is recommended and should be considered for ground-truth data.

This paper presents a novel approach for using and preparing a benchmark to validate image processing methods for generating architectural documentation. For this purpose, it is possible to use *The synthetic images simulator* (link available in acknowledgement), which, through an intuitive graphical interface, allows the preparation of a scalable benchmark tailored to a specific task.

The paper also compares the SfM method implemented in Colmap and Agisoft Metashape software together with the author's hand-crafted affine and 'classical' detectors and learning-based approach. The following relationships emerge from the analyses:

1. The use of afine-detectors and especially blob detectors (ASIFT, ASURF) allows the highest data orientation accuracy while detecting a large number of evenly spaced tie points for a group of hand-crafted detectors. The reprojection error values were obtained on average 2 times lower than for Agisoft Metashape software and similar for Colmap. The accuracy of the external orientation of the images using the

above-mentioned detectors makes it possible to obtain accuracy similar to that obtained using the commercial software Agisoft Metashape and open-source Colmap.

2. Using a learning-based approach to detect binding points allowed orientation results similar to those of the previously mentioned affine-detectors and the tested software to be obtained. It should be emphasised that the aim of this article was also to analyse available solutions, so pre-trained learning-based solutions were used. This resulted in fewer binding points being detected than in the case of hand-crafted detectors. Despite this, it was possible to find sufficient points to correctly determine the internal orientation elements during the self-calibration process.

3. The proposed SfM approach to image orientation, based on hand-crafted affine-detectors, on an extended method of filtering and selection of tie points, allowed the correct and stable determination of internal and external orientation elements for the benchmark-prepared images. For this reason, it seems reasonable to ask whether it is necessary to use learning-based approaches in the orientation of photographs used for the interior inventory of cultural heritage objects when similar results are obtained for approaches based on hand-crafted solutions.

4. The use of affine-detectors instead of detectors is recommended, as they can detect more stable, reliable and robust tie points used in the Structure-from-Motion workflow.

ACKNOWLEDGEMENTS

The synthetic images simulator is available at the following link: https://wutwaw-my.sharepoint.com/:f:/g/personal/jakub_markiewicz_pw_edu_pl/EkJZ4ertcuhEnvX1LsYuXFwBoyMop4YKG6eUk2azmRhDJA?e=4Bdqao). If you have any comments or suggestions regarding the software, please contact Dr Michał Kowalczyk by e-mail: michal.kowalczyk@pw.edu.pl.

This paper was co-financed under the research grant of the Warsaw University of Technology supporting the scientific activity in the discipline of Civil Engineering and Transport.

REFERENCES

Aanæs, H., Dahl, A.L., Steenstrup Pedersen, K., 2012. Interesting Interest Points. *Int. J. Comput. Vis.* 97, 18–35..

Bakuła, K., Mills, J.P., Remondino, F., 2019. A REVIEW OF BENCHMARKING IN PHOTOGRAMMETRY AND REMOTE SENSING. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci*. XLII-1/W2, 1–8..

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.

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*.

Ebel, P., Mishchuk, A., Yi, K. M., Fua, P., Trulls, E., 2019. Beyond cartesian representations for local descriptors. Proc. *ICCV*.

Gabara, G., Sawicki, P., 2023. CRBeDaSet: A Benchmark Dataset for High Accuracy Close Range 3D Object Reconstruction. *Remote Sensing*. 15, 1116.

Leutenegger, S., Chli, M., Siegwart, R.Y., 2011. BRISK: Binary Robust invariant scalable keypoints. Proc. *IEEE Int. Conf. Comput. Vis*. 2548–2555.

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

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Int. *J. Comput. Vis*. 60, 91–110.

Marelli, D., Morelli, L., Farella, E.M., Bianco, S., Ciocca, G., Remondino, F., 2023. ENRICH: Multi-purposE dataset for beNchmaRking In Computer vision and pHotogrammetry. *ISPRS J. Photogramm. Remote Sens*. 198, 84–98.

Mishchuk, A., Mishkin, D., Radenovic, F., Matas, J., 2017. Working hard to know your neighbor's margins: Local descriptor learning loss. Proc. *NIPS*

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.

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

Rosten, E., Drummond, T., 2006. Machine Learning for High Speed Corner Detection. *Comput. Vis.* -- *ECCV 2006* 1, 430–443.

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

Sarlin, P.-E., DeTone, D., Malisiewicz, T., Rabinovich, A., 2019. SuperGlue: Learning Feature Matching with Graph Neural Networks. Proc. *IEEE CVPR*.

Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R., 2006. A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms, in*: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1 (CVPR'06)*. IEEE, pp. 519–528.

Strecha, C., von Hansen, W., Van Gool, L., Fua, P., Thoennessen, U., 2008. On benchmarking camera calibration and multi-view stereo for high resolution imagery, in: *2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE*, pp. 1–8.

Tyszkiewicz, M.J., Fua, P., Trulls, E., 2020. DISK: Learning local features with policy gradient. *Adv. Neural Inf. Process. Syst*. 2020-December, 1–15.

Verdie, Y., Kwang Moo Yi, Fua, P., Lepetit, V., 2015. TILDE: A Temporally Invariant Learned DEtector, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 5279–5288.

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*.