Co-registering Laser Scanning Point Clouds and Photogrammetric Images with Deep Learning Multi-Modal Matching

Luca Morelli [1,2], Giulio Perda [1], Francesco Ioli [3], Paweł Trybała [1], Andrea Sterpin [4], Simone Rigon [1], Neil Sutherland [5], Marco Medici [6], Fabio Remondino [1], Alfonso Vitti [2]

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

[2] Dept. of Civil, Environmental and Mechanical Engineering (DICAM), University of Trento, Italy - Email: alfonso.vitti@unitn.it

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

[4] Department of Architecture, University of Ferrara, Italy - Email: andrea.sterpin@unife.it

[5] Nottingham Geospatial Institute, University of Nottingham, Nottingham NG7 2TU, UK - Email: neil.sutherland@nottingham.ac.uk

[6] INCEPTION, Spinoff of the University of Ferrara, Italy - Email: marco.medici@inceptionspinoff.com

Keywords: image-to-geometry registration, multi-modal matching, learning-based matching, laser scanning, photogrammetry, fusion

Abstract:
The integration of laser scanning and photogrammetry has become a critical approach in architectural and civil surveying, leveraging the geometric precision of Terrestrial Laser Scanners and the high-quality textures achievable through photogrammetric surveys. Despite the advances, challenges persist in efficiently merging these data sources, particularly due to limitations in sensor integration and varying levels of Ground Sampling Distance. This study presents a novel data fusion methodology, operating at raw and intermediate levels, bypassing the need for data pre-alignment, sensor trajectories or coloured point clouds. The approach employs deep learning-based matchers to achieve automated co-registration of RGB images and TLS data, offering advantages such as global registration, multi-modal matching, direct scaling and referencing, and enhanced sensor fusion during the photogrammetric bundle adjustment. Additionally, the method enables the direct orientation of single images and texture mapping without requiring dense point clouds. The pipeline is validated with an architectural surveying scenario, demonstrating its efficacy in comparison with commercial solutions.

1. INTRODUCTION

Laser scanning and photogrammetric point clouds, under appropriate conditions, can achieve similar results in terms of geometric 3D reconstruction and accuracy (Guarnieri et al., 2006; Charbonnier et al., 2013; Teza et al., 2016). Moreover, when a laser scanner is paired with high-resolution integrated cameras, it has the potential to provide good texture quality, but with wide variability depending on the scanner (Julin et al, 2020). For architectural and civil applications, even high-end Terrestrial Laser Scanners (TLS) often do not achieve the same texture quality as photogrammetric surveys with a full-frame or large-sensor high-resolution camera (Crombez et al., 2015; Carraro et al., 2019; Julin et al, 2020). This discrepancy is primarily due to the limitations in the camera sensor size in the TLS and the fact that images are typically acquired from a limited number of stations, resulting in occlusions and a significantly variable Ground Sampling Distance (GSD). Nevertheless, in some cases, TLS does not have an integrated camera. As a result, the combination of laser scanning and photogrammetry is becoming increasingly common in civil and architectural surveys, leveraging TLS point cloud accuracy and high-quality textures. In certain applications, TLS is also used to have a metric base for the surveying and avoid time-consuming topographic networks for georeferencing and scaling a photogrammetric reconstruction.

Therefore, an accurate and efficient data fusion process should be established to merge the data coming from the two different sources (Corsini et al., 2009; Moussa et al., 2012; Ramos and Remondino, 2015; Luhmann, 2019; Bruno et al., 2022). Data can be integrated at various levels (Medici et al., 2024): at raw level by directly coupling images and individual laser scans, at an intermediate level after some processing of both scans and images, or at a high level by simply co-registering and merging the final point clouds of both techniques. The majority of approaches in literature propose integration at high level (Fiorillo et al., 2012; Suwardhi et al., 2015), with few recent approaches working at raw or intermediate levels (Jonassen et al., 2023; Markiewicz et al., 2023; Medici et al., 2024), generally requiring the sensors trajectories, in particular for aerial LiDAR datasets (Glira et al., 2019).

1.1 Aim of the work

This study proposes a data fusion methodology working at raw and intermediate level based on LiDAR point cloud renderings. The method is not assuming a specific acquisition setup (e.g., aerial or terrestrial) and it does not require sensor trajectories nor coloured point clouds. The proposed approach leverages deep learning-based matchers to automatically co-register RGB and scan data featuring intensity values without any user intervention, with the following advantages:

- Global registration approach: it is a sparse approach based on 2D tie points between RGB images and multiple renders, one for each single TLS stations. TLS stations do not need pre-alignment. The RGB dataset and TLS renders are oriented together in a unique block up to a scale factor.
- Multi-modal matching: exploiting deep learning capabilities, RGB colors in the LiDAR data are not strictly required.
- Directly scale and reference the photogrammetric data: the laser scanning point cloud is used to support scaling and referencing during the bundle adjustment (BA).
- Sensor fusion: the laser scanning observations (3D points) can be used as constraint in the photogrammetric BA, potentially improving the accuracy of the final 3D reconstruction.
- Orient single images: even without a robust camera network, single images can be (geo)referenced since the geometry of the model relies on the laser scanning point cloud.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-2-W8-2024-335-2024 | © Author(s) 2024. CC BY 4.0 License.

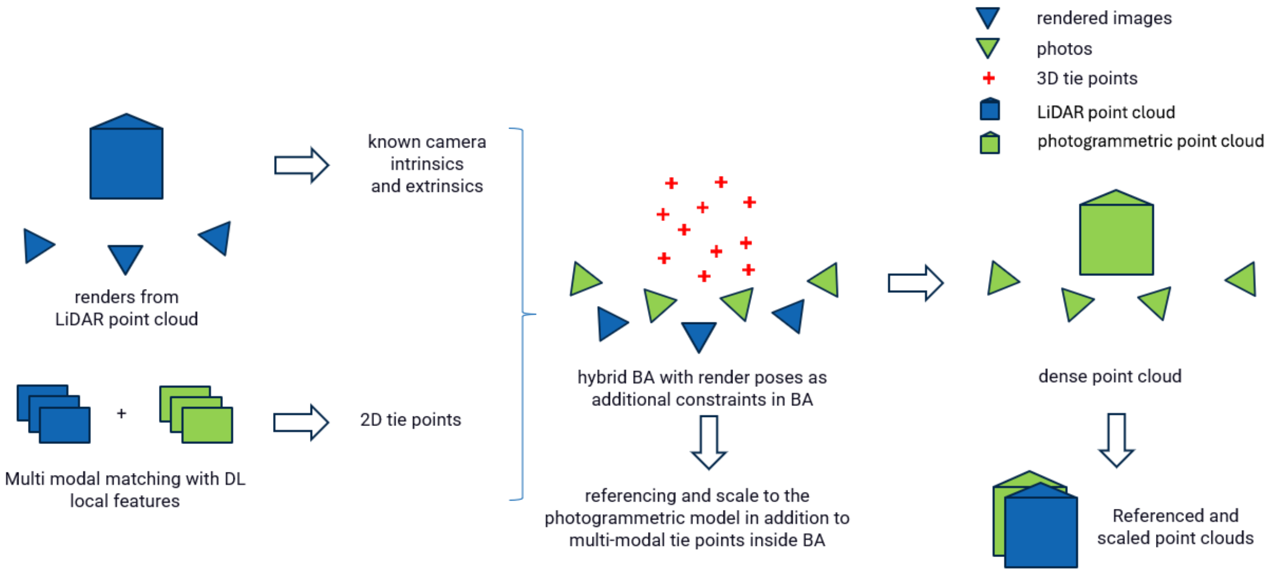

rendered images
photos
3D tie points
LiDAR point cloud
photogrammetric point cloud
known camera intrinsics and extrinsics
renders from LiDAR point cloud
+
2D tie points
Multi modal matching with DL local features
hybrid BA with render poses as additional constraints in BA
referencing and scale to the photogrammetric model in addition to multi-modal tie points inside BA
dense point cloud
Referenced and scaled point clouds

Figure 1: Scheme of the proposed approach to co-register TLS and RGB images with multi-modal image matching.

- Direct texture mapping: registered images can be directly mapped onto the laser scanning point cloud without generating a dense point cloud from the image dataset.

The proposed pipeline, reported in Section 2, was already successfully used to co-register crowdsourcing images and a TLS point cloud (Morelli et al., 2024a), while in this work it is refined and applied to an architectural surveying scenario. It is affine to Markiewicz et al. (2023), which projects single laser scans into equirectangular images and employs deep learning matchers to co-register TLS scans among them. We extend this approach to the multi-imodal RGB-TLS case. The proposed approach is compared with two commercial software solutions, Reality Capture (Capturing Reality, 2024) and Agisoft Metashape (2024), both of which enable the fusion of LiDAR point clouds and RGB data at different levels of integration.

2. DATA AND METHOD

2.1 Methodology

Low-level data fusion approaches involve matching 3D laser scanning data with 2D images, a process that is highly complex due to the need for extracting repeatable and accurate keypoints, as well as describing these areas with descriptors to identify 2D-3D homologous points. Detection, description and matching processes are inherently difficult and often unreliable as they involve working within different domains (Li et al., 2023), often referred to as multi-modal matching (Jiang et al., 2021). For example, 3D descriptors rely on feature extraction based on local geometric properties (e.g., point distribution), while 2D descriptors focus on local radiometric variations, which are not solely related to changes in geometry.

To address these challenges, we propose a 2D-2D multi-modal matching approach, instead of 2D-3D, by rendering synthetic views of the laser scanning point cloud. Its workflow is showcased in Figure 1. A similar method was proposed by Elias et al. (2023) to co-register individual thermal images with LiDAR point clouds. Other studies in the literature have explored co-registration through the rendering of RGB images and LiDAR point clouds. However, they emphasize the challenges of working with LiDAR point clouds lacking colorimetric information (Kehl et al., 2017; Elias et al., 2019), a scenario addressed in this work. To the best of the authors' knowledge, this is the first study to employ deep learning-based local features and matchers to co-register TLS renders with RGB images.

Given one or more TLS scans, each can be rendered into a series of synthetic views, following approaches such as the Reality Capture (six views on a cube) or Agisoft Metashape (one large equirectangular projection). Reality capture leverages the intensity values recorded by the TLS to generate these views, while is not known the approach adopted by Metashape. In our approach, for each single scan, a single view is rendered in Blender (Blender Foundation, 2024), with an orientation and field of view that allow good part of the scan to be visualized in a single render. As TLS scans generally lack RGB information, renders are generated with a rendering engine which uses TLS intensity values as texture. The core idea is to integrate the renders directly into the photogrammetric adjustment: individual TLS scans are co-registered with each other in a single step, along with the co-registration of the scans with the photogrammetric image block. This procedure fuses data at raw level producing a unified block where individual TLS scans and the RGB block are co-registered up to a scale factor. At this stage, a single TLS scan is sufficient to assign the scale and reference the block. In the case study experiment, multiple scans are pre-registered directly by the TLS. Consequently, in this paper the scale and referencing are achieved by treating the poses of the TLS renders—already known within a single reference system due to pre-alignment—as fixed in the bundle adjustment, achieving an intermediate data fusion level. Moreover, the possible known/georeferenced laser scanning poses together with the multi-modal tie points can potentially improve the final reconstruction accuracy adding additional observations in BA respect the photogrammetric model with only tie points from the RGB image block.

2.2 Multi-modal matching

By shifting from 2D-3D to 2D-2D matching, the search of correspondences remains within the same domain, yet it still encounters challenges arising from significant radiometric differences between the rendered TLS images and those captured by cameras. For this reason, RGB-RGB matching, and especially the multi-modal matching, is performed using learning-based approaches trained to find robust tie points against extreme variations in illumination and perspective distortions (Chen et al., 2021; Jin et al., 2021). Our hypothesis is that these AI models are sufficiently general to be used for multi-modal matching without retraining, even though they were not explicitly designed for multi-modal matching. In this work, the methods available in the DIM toolbox[1] (Morelli et al., 2024b) are used. DIM is a library of deep learning local features and matchers, able to match images at full size with a tiling approach, and that support easy integration in various photogrammetric software. The multi-modal matching is performed using SuperPoint (DeTone et al., 2018) combined with SuperGlue (Sarlin et al., 2020) and LightGlue (Lindenberger et al., 2024) matchers, DISK (Tyszkiewicz et al., 2020) and ALIKED (Zhao et al., 2023) paired with LightGlue, and DeDoDe (Edstedt et al., 2024) matched with the nearest-neighbor strategy. These matchers are compared with traditional methods: RootSIFT (Arandjelović and Zisserman, 2012), as implemented in COLMAP (Schönberger and Frahm, 2016) and Agisoft Metashape, to provide an example of a widely used commercial software in both practical applications and research. In addition, deep learning-based strategies are compared with the results of RIFT (Li et al., 2019), the most widely used multi-modal matching method, particularly used for aerial and satellite imagery datasets.

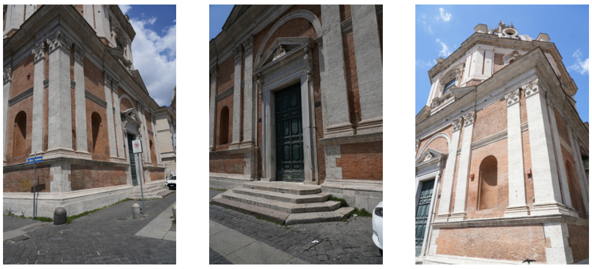


Figure 2: Some images acquired for the photogrammetric processing.

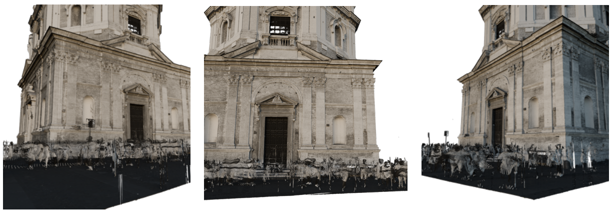


Figure 3: Rendered images from the TLS point clouds using only the intensity values from the TLS data.

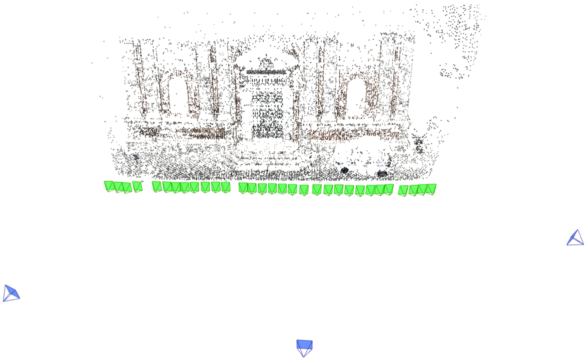


Figure 4: Camera network for the Santa Maria di Loreto dataset. RGB camera poses in green, positions of TLS renders in blue.

2.3 Testing dataset and camera network

To evaluate the proposed approach, a survey of the external facades of Santa Maria di Loreto in Rome (Medici et al., 2024) has been used. Data were acquired with Leica P40 (3D point position accuracy of 3 mm at 50 meters) and a Sony Alpha 7 (33 MP full-frame sensor - Figure 2). To highlight the advantages of the proposed method, a strip of 31 images taken at a constant distance from the building facade was selected, without including oblique images or additional strips that could enhance the self-calibration of the RGB sensor. Figure 3 shows three renders created from the TLS scans of the facade. Although the renders seem to include RGB data, the point cloud is rendered using only the intensity values from the TLS. The RGB images were captured at a distance of approximately 5 meters from the church facade, while the renders were generated from a distance of about 18 meters. The camera network of the RGB images and the poses of the three renders are shown in Figure 4, where the multi-modal block is co-registered using SuperPoint combined with LightGlue. Multi-modal matching is complex not only because of the radiometry differences between TLS intensity renders and camera images, but also because of different image scales and resolution (see Section 3.2). Due to actual limitations of learning-based methods (e.g. scale invariance), the TLS renders and image sizes had to be adjusted to 4094 x 4094 px 2803 x 4205 px for the TLS and RGB images, respectively, in order to maximize the number of extracted matches. To address the scale limitation, we would need to extract SuperPoint features on an image pyramid with varying scales.

2.4 Image block orientation with DIM features

Using DIM, a maximum of 4096 matches per image pair, filtered using MAGSAC (Barath et al., 2019), are extracted. The feature extraction is performed at medium quality, corresponding to half the original resolution of the images (see Section 3.2). DIM extracts features and matches that are saved in a database with the same format as COLMAP. Using pycolmap[2], the image block is oriented with a pinhole camera model, and the poses and 3D tie points are exported in *Bundler* format. Then, all data are imported into Agisoft Metashape for a final BA, with the poses of the renders fixed as known, and the images undergoing a self-calibration process. The transition from COLMAP to Agisoft Metashape is necessary because COLMAP, specifically pycolmap, does not support the use of additional constraints on camera poses.

2.5 Fusion of TLS and photogrammetric point clouds

As shown at the end of the processing in Figure 1, the proposed pipeline produces a photogrammetric point cloud scaled and referenced from a BA where 2D multi-modal observations are used, without any 3D constraints a part for the known render poses. However, it does not fully utilize all the information available from the TLS point cloud. Specifically, while the multi-modal tie points are included in the bundle adjustment, their 3D coordinates derived from the TLS scan are not.

Under the assumption that the TLS point cloud is more accurate than the photogrammetric one, an additional constraint can be introduced in the BA by treating the 3D coordinates of the multi-modal tie points (available from the TLS ranges) as known values. In the experiments, a more general approach was tested: we employ the methodology described in Figure 1 up to the step where the multi-modal block is oriented. At this stage, the 3D tie points and the TLS point cloud are co-registered. By performing a nearest-neighbor search, each 3D tie point can be associated with a corresponding point in the TLS point cloud, and this coordinate can be treated as a known value in an additional BA run. This method allows not only the multi-modal tie points—which may be limited in number—to be assigned TLS-derived

[1] https://github.com/3DOM-FBK/deep-image-matching

[2] https://github.com/colmap/pycolmap

coordinates but also enables the assignment of TLS coordinates to non-multi-modal tie points.

3. RESULTS AND DISCUSSION

3.1 Performance evaluation of commercial software

Co-registering RGB data with LiDAR data is not a straightforward task, particularly when performed in the early stages of processing, prior to the generation of independent point clouds (Medici et al., 2024). A fusion of raw data at this stage theoretically allows for combining observations to achieve enhanced geometric accuracy. As previously noted, some commercial software solutions enable this type of integration. For instance, Reality Capture supports the alignment of individual TLS scans or groups of pre-aligned scans with an approach similar to that proposed in this study. From the input LiDAR point clouds, six cube-distributed renders are generated around the station point using a pinhole camera model. It is assumed that the images are subsequently matched with these renders to enable multi-modal co-registration. Agisoft Metashape, on the other hand, uses an equirectangular projection of the TLS points and intensity and, presumably, follows a similar matching strategy. In the tested dataset, both Reality Capture and Agisoft Metashape fail to co-register LiDAR and RGB data effectively, whether the scans are provided individually or pre-aligned (Figure 5). This failure is likely due to the lack of RGB information in the TLS point cloud leading to an inability of matching intensity-based images with RGB images. This result aligns with the findings reported in Medici et al. (2024), from which the dataset used in this study is a subset.

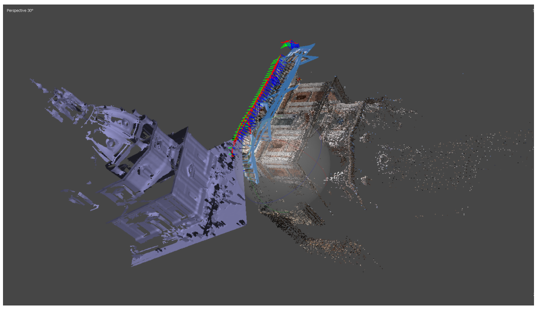


Figure 5: A failed co-registration of the RGB data with the TLS point cloud in Agisoft Metashape.

3.2 Image matching with different local features

Given the radiometric differences arising from real versus artificial illumination conditions in the renders, the data are not only matched using classical methods such as RootSIFT but are also processed using neural network-based matchers. As said, although these matchers are not explicitly trained to align LiDAR data with RGB images, it is hypothesized that their training on highly complex datasets enables them to generalize effectively to multi-modal datasets. Table 1 presents the results of the different matching approaches tested. Despite the variety of deep learning-based approaches available in the literature, only SuperPoint demonstrates the ability to generalize to our case study. All other methods fail to detect the minimum number of tie points required between renders and RGB images to enable the orientation of the renders within the photogrammetric block. Furthermore, the combination of SuperPoint and LightGlue proves to be the fastest matching method between learning-based approaches. The results reported in the table refers to the specific implementation of DIM that works per tiles and apply MAGSAC.

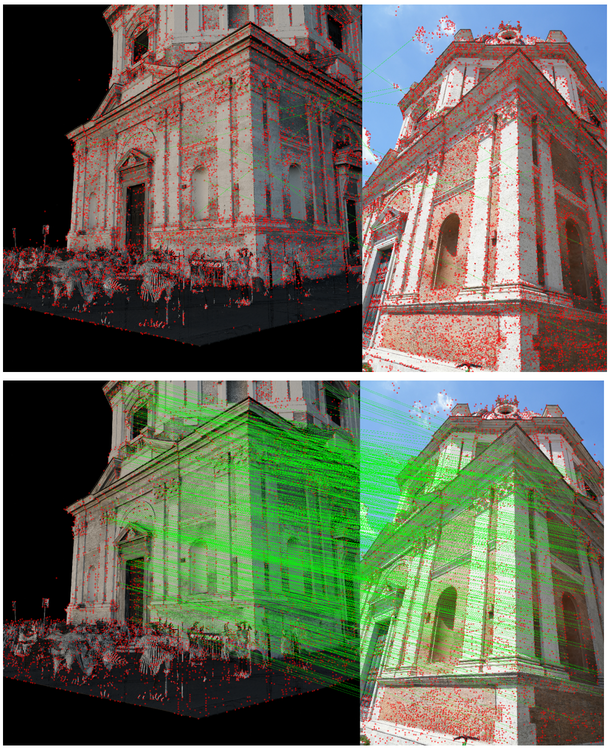


Figure 6: Matches (green lines) between TLS renders and RGB images with RootSIFT (top) & SuperPoint+LightGlue (bottom).

Matching TLS renders with RGB images using SuperPoint is not feasible across all resolutions, as mentioned in 2.4. Consistent with findings in Marelli et al. (2023), the scale invariance of deep learning-based methods is limited. Although DIM supports high-resolution matching, aligning RGB images and renders at full resolution results in the extraction of a limited number of matches. This indicates that the SuperPoint descriptor does not operate effectively across a wide range of scales, unlike RootSIFT. Consequently, it is necessary to manually determine the most appropriate scale for matching. Matching at a lower resolution clearly affects the actual GSD of the scene, leading to some 4 mm for RGB and 1cm for TLS renders.

Similarly to most DL-based descriptors, RootSIFT in the COLMAP implementation also fails, as does Agisoft Metashape, when attempting to orient TLS renders and RGB images together. Similarly, RIFT, even after testing various combinations of image down sampling, fails to find correspondences. Figure 6 presents an example of matching between a TLS render and a RGB image, using RootSIFT and SuperPoint combined with LightGlue.

	Extraction [min:sec]	*Matching [min:sec]*	*Status*
Agisoft Metashape	00:07		Fails
COLMAP (RootSIFT)	01:12	00:05	Fails
SuperPoint + LightGlue	00:45	01:18	Oriented
SuperPoint + SuperGlue	00:45	13:09	Oriented
DISK + LightGlue	02:26	01:01	Fails
ALIKED + LightGlue	01:56	01:12	Fails
DeDoDe + NN	01:20	00:57	Fails
Key.Net + HardNet + NN	02:24	00:43	Fails
RIFT	> 1 h		Fails

Table 1: Matching time and success status for different matching strategies and image triangulation. All tests were conducted using an NVIDIA GeForce RTX 3050 Laptop GPU.

3.3 3D accuracy evaluation

The accuracy assessment was performed using one façades of the church (Figure 7). Pre-aligned TLS scans are used as ground truth, to assess the accuracy of the point cloud generated from RGB images oriented using the renders, that is expected to be less accurate because of the poor camera network. The renders enable the creation of a photogrammetric point cloud that is already scaled and referenced, and it is then compared with the classical methodology, which involves manually co-registering a photogrammetric point cloud roughly, followed by fine co-registration using the iterative closest point algorithm (ICP).

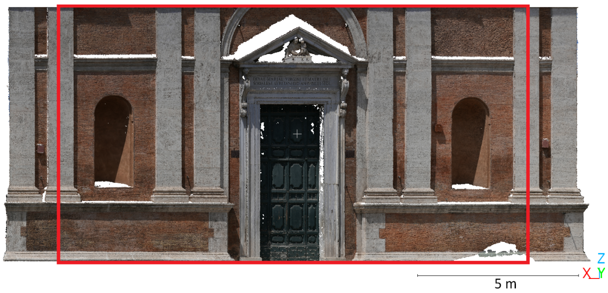


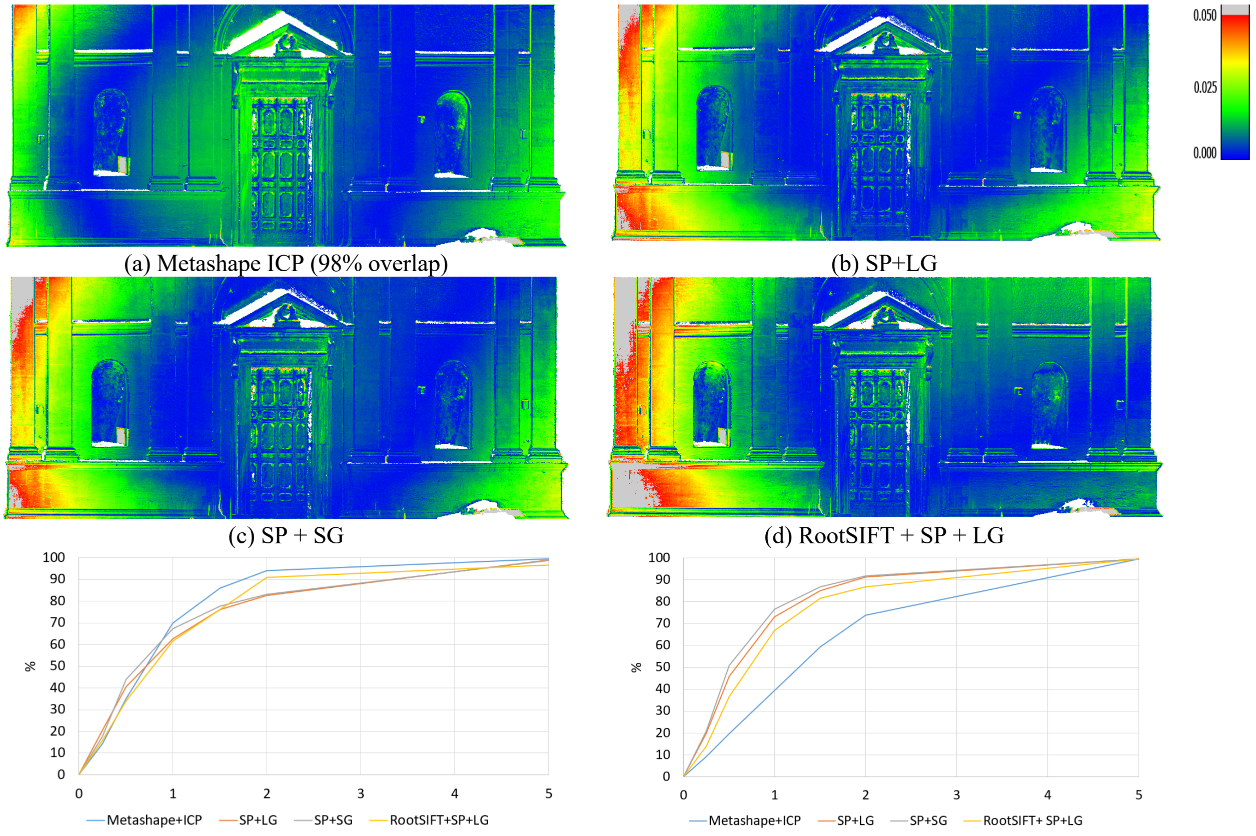
Figure 7: The area of the façade used for the accuracy evaluation. Results are reported for the entire areas shown in the figure and for just the area inside the red rectangle.

RGB images were oriented at the *highest quality* option in Metashape, and a dense point cloud was generated at high quality with moderate depth filtering. The same settings were used to produce the dense point cloud via the proposed method using SuperPoint, where no manual co-registration or ICP has been required.

Figures 8a–8d illustrate the C2C distance maps with scale bars in centimetres between each evaluated approach and the TLS point cloud taken as reference. Figures 8e and 8f present the percentage of points with C2C distances below increasing thresholds from 0 to 5 cm. While 8e shows the curve for the entire generated point clouds, 8f reports the C2C distances excluding the peripheral regions, since they are more instable and reconstructed with higher uncertainty due to the configuration of the camera network (Figure 7).

The first key finding is that the proposed method effectively enables the automatic referencing and scaling of the photogrammetric block using LiDAR data. Figures 8b–8d demonstrate absolute C2C errors of less than 5 cm for over 95% of the points. This metric highlights the potential of multi-modal matching as a method for global co-registration. In relative terms, the four generated point clouds are equivalent. Performing a final ICP on the point clouds (Figure 8b-8d) yields an RMSE identical to that obtained in Figure 8a, specifically 4.4 cm. However, working with ICP presents drawbacks, particularly in poorly reconstructed areas, such as the peripheral regions in this case. These areas heavily influence the ICP process, although they minimize the overall error, they worsen accuracy in regions with strong camera network coverage, such as the central areas.

Figure 8f underscores the advantages of the proposed approach: in the central area, over 75% of the points achieve an accuracy of less than 1 cm, and 90% fall below 2 cm. Conversely, the same chart reveals that approach in Figure 8a penalizes the central area.



(a) Metashape ICP (98% overlap)

(b) SP+LG

(c) SP + SG

(d) RootSIFT + SP + LG

(e)

(f)

Figure 8: Cloud-to-cloud distances between photogrammetric and TLS point clouds after applying classical co-registering approach with ICP (a) and co-registration via multi-modal matching using different combinations of descriptors (b–d). Accuracy comparisons for the full façade (e) and the center part only (f) in terms of percentage of points with C2C distances below increasing thresholds from 0 to 5 cm. Common color scale for plots (a–d) is shown. Both the scale bar and the accuracy thresholds are in centimeters.

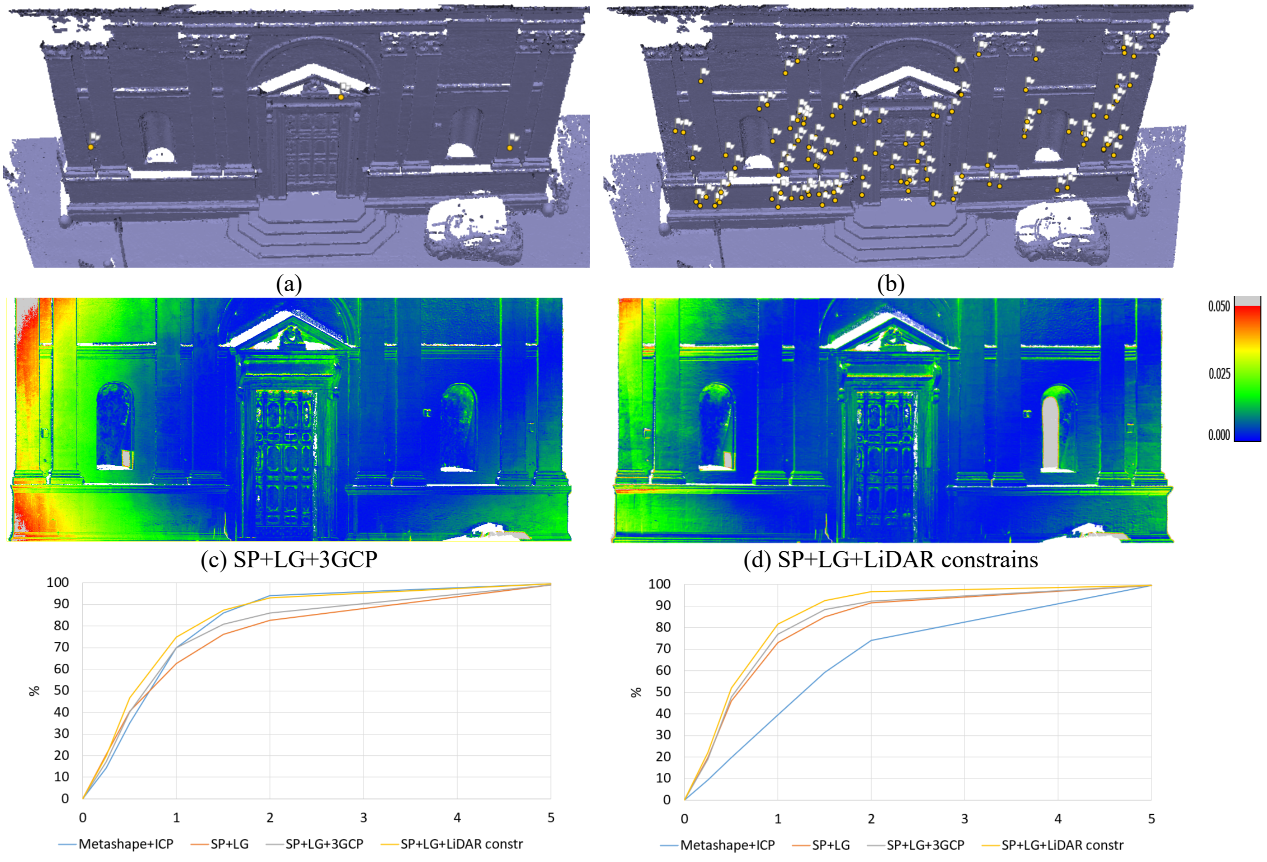
In addition, RootSIFT was run on full resolution images to leverage the potential accuracy gains from using such images and combined with SuperPoint features, increasing the number of keypoints from 4,096 kpts/image to 9,269 kpts/image. However, the lack of improvement is likely tied to the available GSD, which may not sufficiently support the expected gains in accuracy.

3.4 Influence of 3D constraints on tie points

The proposed approach is not merely a method for global co-registration, but also offers the capability for deep integration of LiDAR and photogrammetric data. In the case study, it is assumed that the TLS point cloud can serve as a reference due to its higher accuracy compared to the photogrammetric point cloud, which is particularly affected by a dome effect caused by a suboptimal camera network (see C2C distribution in Figure 8a). The TLS data can thus be used as an anchor for the photogrammetric point cloud to improve the estimation of both extrinsic and intrinsic parameters, as well as calibration parameters, thereby reducing the dome effect. In the case study, the constraints on the tie points are introduced as actual ground control points with one millimeter accuracy. If the point-by-point TLS point cloud covariances are known, the same approach can be applied by appropriately weighting the additional constraints using the known a priori standard deviation.

Two possible solutions are demonstrated (Section 2.4). The first involves manually selecting a few natural points to include as constraints in the BA, referred in Figure 9 as *SP+LG+3GCP*. Figure 9a shows the distribution of three natural points chosen to cover a significant portion of the cloud. They are clearly identifiable in the TLS point cloud and on the RGB images, and they are treated as ground control points, e.g., adding 2D projections on the images and the 3D LiDAR coordinates as constraints in the BA. Figure 9b illustrates an alternative approach. The results obtained with the global co-registration approach evaluated in the previous section can be treated as an initial orientation. At this stage, each 3D tie point can be associated with a corresponding TLS point using a nearest-neighbor approach. These points, as 2D projections and their 3D LiDAR point correspondences, are then used as constraints in the BA. Empirically, 150 tie points were randomly sampled for this procedure. Although the points are not evenly distributed due to random sampling, they still cover the entire facade, unlike in Figure 9a. This approach is referred in Figure 9 as SP+LG+LiDAR constrains.

As shown in Figure 9e and 9f, the introduction of constraints using three natural points or 150 3D tie points leads to a significant improvement in accuracy compared to the baseline method of SuperPoint + LightGlue. In the second case, 80% of the points achieve an accuracy of less than or equal to 1 cm in the non-peripheral area. This approach results in a substantial enhancement in the final co-registration accuracy between the RGB and TLS point clouds.


(a)

(b)

(c) SP+LG+3GCP

(d) SP+LG+LiDAR constrains

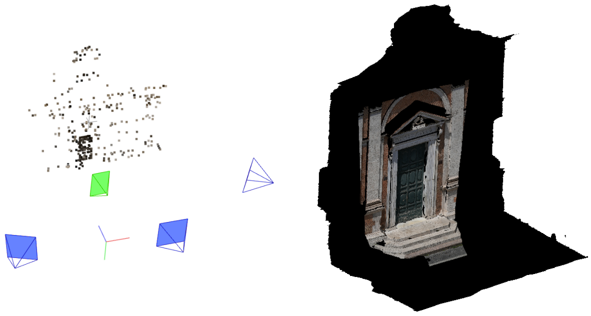
(e)

(f)

Figure 9: Accuracy evaluation of SuperPoint + LightGlue and additional BA constraints with 3 GCPs (a, c and e) and with 150 constraints on 3D tie points (b, d and f). Common color scale for plots (c–d) is shown. Both the scale bar and the accuracy thresholds are in centimeters. Accuracy comparisons for the full façade (e) and the center part only (f) is reported in terms of percentage of points with C2C distances below increasing thresholds from 0 to 5 cm.

3.5 Orientation of single frames on TLS point cloud

The same approach can be applied to align single or a small number of DSLR images with a LiDAR point cloud. In this case, a purely a-posteriori approach on the final point clouds is not feasible, as it is not possible to generate a dense cloud from a single RGB image. Figure 10a illustrates the alignment of a single image from the same dataset used previously, alongside the three renders. The three renders were modelled using a pinhole camera model with a known focal length and their poses have been used to reference and scale the model. Given the limited number of images, the photo was assigned a camera model with a single radial distortion parameter. To verify the accuracy, three natural points were selected as checkpoints, yielding an RMSE of 1.5 cm. Figure 10b shows the projection of the RGB image onto the LiDAR point cloud.



(a) (b)

Figure 10: Orientation of a single photo (in green) to the three TLS renders (in blue) (a) and projection of the single image on the TLS point cloud (b).

4. CONCLUSIONS

The combined use of photogrammetry and TLS point clouds is becoming increasingly common. However, these datasets may have differing resolutions and accuracies, leading to inconsistencies when co-registering point clouds generated by the individual technologies. Several fusion approaches are available in the literature, predominantly for aerial datasets, while terrestrial cases are typically addressed by commercial solutions that can perform poorly when the LiDAR point cloud lacks colorimetric data. This method proposed in this paper aims to advance the integration of terrestrial laser scanning and photogrammetry for precise and efficient 3D reconstruction purposes. Leveraging novel matching approaches based on neural networks robust to significant radiometric variations in the image set, a new approach for RGB to LiDAR co-registration is presented. Specifically, we demonstrated the capability of deep learning matching algorithms to co-register individual TLS scans with a photogrammetric set of images that is thus referenced and scaled to the TLS coordinate system. Furthermore, we presented a simple method to incorporate LiDAR observations as additional constraints within the photogrammetric adjustment, enhancing the accuracy of the final photogrammetric point cloud.

Therefore, multi-modal matching can rely on neural network-based matchers trained to be robust to strong perspective distortions and significant lighting variations. Tests have shown that only SuperPoint local features combined with LightGlue and SuperGlue exhibits sufficient invariance to handle multi-modal matching between RGB images and TLS renders. The only non-automatic step in the described procedure is the selection of the TLS renders resolution and RGB images to maximize the number of tie points in the multi-modal pairs. In future work, this could be automatized by extracting keypoints on an image pyramid with multiple scales.

REFERENCES

Agisoft Metashape, 2024. Agisoft Metashape Professional (Metashape) Software, Version 2.1. Accessed online: (15.11.2024): https://www.agisoft.com/

Arandjelović, R. and Zisserman, A., 2012. Three things everyone should know to improve object retrieval. *Proc. CVPR*, pp. 2911-2918.

Barath, D., Matas, J. and Noskova, J., 2019. MAGSAC: marginalizing sample consensus. *Proc. CVPR,* pp. 10197-10205).

Blender Foundation, 2024. Blender Software, Version 4.2. Accessed online: (15.11.2024): https://www.blender.org/

Bruno, N., Mikolajewska, S., Roncella, R., and Zerbi, A., 2022. Integrated Processing of Photogrammetric and Laser Scanning Data for Frescoes Restoration, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLVI-2/W1-2022, 105–112.

Capturing Reality, 2024. Reality Capture Software, Version 1.4. Accessed online (15.11.2024): https://www.capturingreality.com/realitycapture

Carraro, F., Monego, M., Callegaro, C., Mazzariol, A., Perticarini, M., Menin, A., Achilli, V., Bonetto, J., and Giordano, A., 2019. The 3D survey of the roman bridge of San Lorenzo in Padova (Italy): a comparison between SfM and TLS methodologies applied to the arch structure. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-2/W15, 255–262,

Charbonnier, P., Chavant, P., Foucher, P., Muzet, V., Prybyla, D., Perrin, T., Grussenmeyer, P., Guillemin, S., 2013. Accuracy assessment of a canal-tunnel 3D model by comparing photogrammetry and laserscanning recording techniques. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XL-5/W2, 171-176.

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

Corsini, M., Dellepiane, M., Ponchio, F., Scopigno, R., 2009. Image-to-geometry registration: a mutual information method exploiting illumination-related geometric properties. *Computer Graphics Forum*, Vol. 28, No. 7, pp. 1755-1764.

Crombez, N., Caron, G., and Mouaddib, E., 2015. 3D point cloud model colorization by dense registration of digital images. Int. Arch. Photogramm. *Remote Sens. Spatial Inf. Sci.*, XL-5/W4,123–130.

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

Edstedt, J., Bökman, G., Wadenbäck, M. and Felsberg, M., 2024, March. DeDoDe: Detect, Don't Describe—Describe, Don't

Detect for Local Feature Matching. Proc. *International Conference on 3D Vision (3DV)*, pp. 148-157.

Elias, M., Kehl, C. and Schneider, D., 2019. Photogrammetric water level determination using smartphone technology. *The Photogrammetric Record*, *34*(166), pp.198-223.

Elias, M., Weitkamp, A. and Eltner, A., 2023. Multi-modal image matching to colorize a SLAM based point cloud with arbitrary data from a thermal camera. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, *9*, p.100041.

Fiorillo, F., Jimenez Fernandez-Palacios, B., Remondino, F., Barba, S., 2012. 3D Surveying and modeling of the archaeological area of Paestum, Italy. Proc. 3rd Arquelogica 2.0.

Glira, P., Pfeifer, N. and Mandlburger, G., 2019. Hybrid orientation of airborne lidar point clouds and aerial images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *4*, pp.567-574.

Guarnieri, A., Remondino, F. and Vettore, A., 2006. Digital photogrammetry and TLS data fusion applied to Cultural Heritage 3D modeling. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci*, *36*, pp.1-6.

Jiang, X., Ma, J., Xiao, G., Shao, Z. and Guo, X., 2021. A review of multimodal image matching: Methods and applications. *Information Fusion*, 73, pp.22-71.

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

Jonassen, V.O., Kjørsvik, N.S. and Gjevestad, J.G.O., 2023. Scalable hybrid adjustment of images and LiDAR point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, *202*, pp.652-662.

Julin, A., Kurkela, M., Rantanen, T., Virtanen, J.-P., Maksimainen, M., Kukko, A., Kaartinen, H., Vaaja, M.T., Hyyppä, J., Hyyppä, H., 2020. Evaluating the Quality of TLS Point Cloud Colorization. *Remote Sens.*, 12(17), 2748.

Kehl, C., Buckley, S.J., Viseur, S., Gawthorpe, R.L. and Howell, J.A., 2017. Automatic illumination-invariant image-to-geometry registration in outdoor environments. *The Photogrammetric Record*, *32*(158), pp.93-118.

Li, J., Hu, Q. and Ai, M., 2019. RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform. *IEEE Transactions on Image Processing*, 29, pp.3296-3310.

Li, M., Qin, Z., Gao, Z., Yi, R., Zhu, C., Guo, Y. and Xu, K., 2023. 2d3d-matr: 2d-3d matching transformer for detection-free registration between images and point clouds. *Proc. ICCV*, pp. 14128-14138.

Luhmann, T., 2019. Combination of Terrestrial Laserscanning, UAV and Close-Range Photogrammetry for 3D Reconstruction of Complex Churches in Georgia. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-2/W11, 753-761.

Lindenberger, P., Sarlin, P.E. and Pollefeys, M., 2023. Lightglue: Local feature matching at light speed. *Proc. CVPR*, pp. 17627-17638.

Marelli, D., Morelli, L., Farella, E.M., Bianco, S., Ciocca, G. and Remondino, F., 2023. ENRICH: Multi-purposE dataset for beNchmaRking In Computer vision and pHotogrammetry. *ISPRS Journal of Photogrammetry and Remote Sensing*, *198*, pp.84-98.

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

Medici, M., Perda, G., Sterpin, A., Farella, E.M., Settimo, S. and Remondino, F., 2024. Separate and Integrated Data Processing for the 3D Reconstruction of a Complex Architecture. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, *48*, pp.249-256.

Morelli, L., Mazzacca, G., Trybała, P., Gaspari, F., Ioli, F., Ma, Z., Remondino, F., Challis, K., Poad, A., Turner, A. and Mills, J.P., 2024b. The Legacy of Sycamore Gap: The Potential of Photogrammetric AI for Reverse Engineering Lost Heritage with Crowdsourced Data. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, *48*, pp.281-288.

Morelli, L., Ioli, F., Maiwald, F., Mazzacca, G., Menna, F. and Remondino, F., 2024a. Deep-image-matching: a toolbox for multiview image matching of complex scenarios. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, *48*, pp.309-316.

Moussa, W., Abdel-Wahab, M., and Fritsch, D, 2012. An automatic procedure for combining digital images and laser scanner data. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XXXIX-B5, 229–234.

Ramos, M., Remondino, F., 2015. Data fusion in Cultural Heritage – A Review. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Vol. XL-5/W7, pp. 359-363.

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

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

Suwardhi, D., Menna, F., Remondino, F., Hanke, K., Akmalia, R., 2015. Digital 3D Borobudur – Integration of 3D surveying and modeling techniques. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, Vol. XL-5/W7, pp. 417-423.

Teza, G., Pesci, A. and Ninfo, A., 2016. Morphological analysis for architectural applications: comparison between laser scanning and Structure-from-Motion photogrammetry. J. *Surv. Eng.*, 142(3), 04016004.

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

Zhao, X., Wu, X., Chen, W., Chen, P.C., Xu, Q. and Li, Z., 2023. Aliked: A lighter keypoint and descriptor extraction network via deformable transformation. *IEEE Transactions on Instrumentation and Measurement*, *72*, pp.1-16.