

COMPARISON OF STATE-OF-THE-ART MULTI-VIEW STEREO SOLUTIONS FOR CLOSE RANGE HERITAGE DOCUMENTATION

A. Murtiyoso¹, J. Markiewicz², A. K. Karwel², P. Grussenmeyer³, P. Kot⁴

¹ Forest Resources Management, Institute of Terrestrial Ecosystems, Department of Environmental Systems Science, ETH Zurich, Switzerland – amadi.murtiyoso@usys.ethz.ch

² Faculty of Geodesy and Cartography, Warsaw University of Technology, Warsaw, Poland – (jakub.markiewicz, artur.karwel)@pw.edu.pl

³ Université de Strasbourg, INSA Strasbourg, CNRS, ICube Laboratory UMR 7357, Photogrammetry and Geomatics Group, 67000, France – pierre.grussenmeyer@insa-strasbourg.fr

⁴ Built Environment and Sustainable Technologies (BEST) Research Institute, Liverpool John Moores University, Liverpool, United Kingdom - p.kot@ljmu.ac.uk

Commission II

KEY WORDS: Close-range, Documentation, Heritage, Learning-based, NeRF, MVS, Photogrammetry.

ABSTRACT

In recent years novel 3D reconstruction methods have been developed to improve the conventional image-based point cloud generation techniques. These novel methods generally attempt to address various challenges encountered in conventional methods, namely, the reconstruction of reflective surfaces and the amount of processing time required, both of which are major bottlenecks in heritage documentation and especially those related to large and complex objects. In this paper, we identified three types of 3D image-based reconstruction techniques and tested their usage on heritage datasets, namely (1) conventional multi-view stereo (MVS), (2) learning-based MVS, and (3) neural radiance fields (NeRF). The aim of this study is to determine the capabilities of these methods in reconstruction of three different heritage-related datasets with different challenges. Our results show that conventional MVS is nowadays a reliable solution for 3D reconstruction, in many instances recording good results relative to the reference terrestrial laser scans (TLS) when properly deployed. When applied to a challenging highly reflective scene, conventional MVS fared well using the PatchMatch algorithm (reaching an object completeness rate of 99.05%), while NeRF's best performance was 99.98%. However, NeRF suffered from noisy data, some of which may stem from its radiance field-to-point cloud conversion method. The results show that there is great potential in using specific methods for specific cases, and research in combining them may yield interesting results in the future.

1. INTRODUCTION

The application of photogrammetry for heritage documentation has a long-established history. Over the past few decades, photogrammetry has experienced a renewed interest in this field, primarily owing to significant advancements in image processing technology. Image-based reconstruction is popularly used in heritage documentation due to its lower cost and high fidelity, although it requires more knowledge both in terms of data acquisition and data processing. The use of photogrammetry was classically linked to 3D plots of objects but has since seen a significant shift towards the creation of point clouds and 3D meshes, which require less user interaction. Furthermore, the emergence of Multi-View Stereo (MVS) methods enabled the creation of dense point clouds from photogrammetry although several challenges can still be identified in specific cases, such as texture-less or reflective surfaces owing to MVS depends on handcrafted features at the pixel level, thus generating ambiguity when encountering such cases.

Several authors tried to solve this problem, for example, by employing semantic constraints (Murtiyoso et al., 2022; Stathopoulou et al., 2021). Another recent attempt to improve MVS was to depart from classical methods and use learning-based methods instead (Stathopoulou and Remondino, 2023). Further approach to the problem involves the use of neural radiance fields (NeRF) (Mildenhall et al., 2020). NeRF was originally developed to render novel viewpoints in a 3D space, but using computer graphics methods, it is possible to convert it into point clouds (Martin-Brualla et al., 2021). Since NeRF computes a density function of discrete 3D spaces instead of projecting image pixels directly, such as in MVS, it behaves

differently when faced with cases that MVS usually struggles with.

The use of these techniques in heritage documentation is becoming attractive. MVS has a long establishment in heritage documentation (Bedford, 2017; Grilli and Remondino, 2019; Murtiyoso and Grussenmeyer, 2017), with several researchers experimenting with learning-based methods (Wei et al., 2020). NeRF is a relatively new concept, although several authors have also attempted to use it for heritage recording purposes (Balloni et al., 2023; Croce et al., 2023; Murtiyoso and Grussenmeyer, 2023; Vandenabeele et al., 2023).

The aim of this article is to evaluate three distinct approaches to 3D reconstruction: (1) classical MVS, (2) learning-based MVS, and (3) NeRF. To this end, three different datasets highlighting different challenges encountered within the context of heritage documentation will be presented and compared. Notably, of the three datasets two will involve synthetic datasets made from terrestrial laser scanner (TLS) point clouds and one from terrestrial photogrammetry. The use of synthetic data is chosen in order to better design the image network and thus better control the parameters of the experiment. Meanwhile, the real-world data will present additional environmental challenges, e.g. irregular image acquisition.

Of the three datasets, one will focus on a reflective surface, which is a particularly challenging scenario for image-based methods. Another focuses on frescoes or wall paintings, while the last one is a 3D model of an exterior building façade.

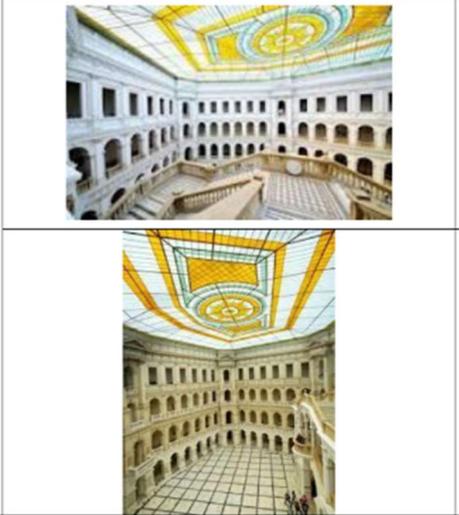
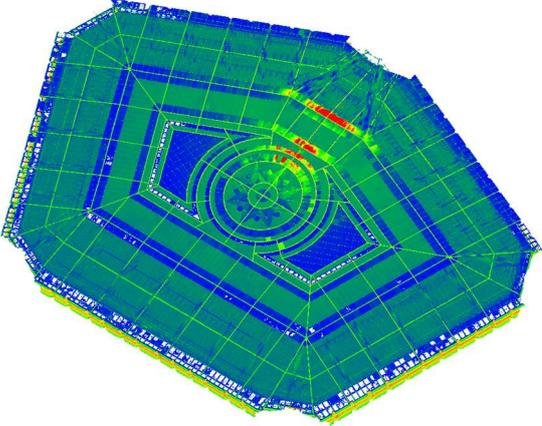
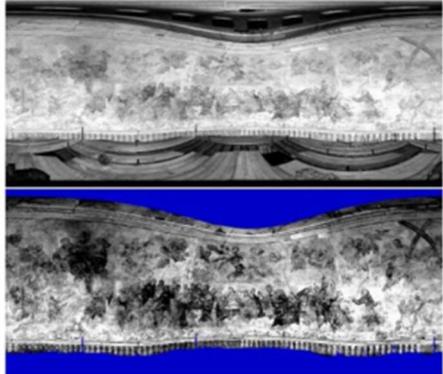
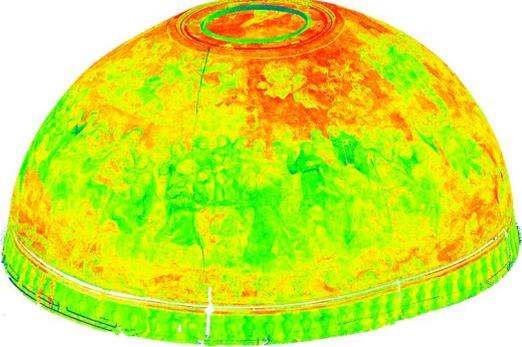
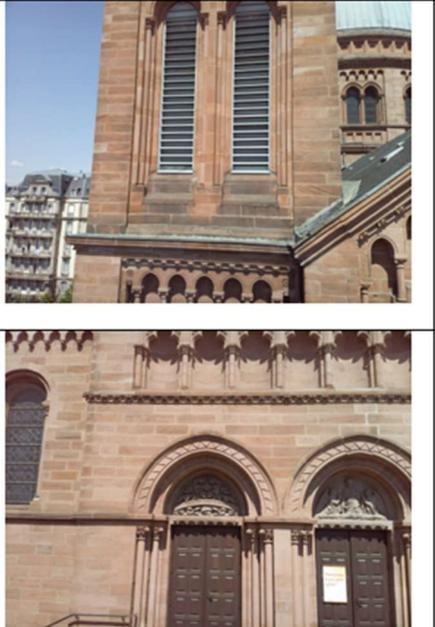
Dataset	Images	Reference TLS point cloud
"Hall"		
"Chapel"		
"Facade"		

Table 1. The three datasets used in this study. The "Hall" and "Chapel" datasets consist of synthetic images generated from TLS point cloud, whereas "Facade" was taken using a terrestrial close-range camera.

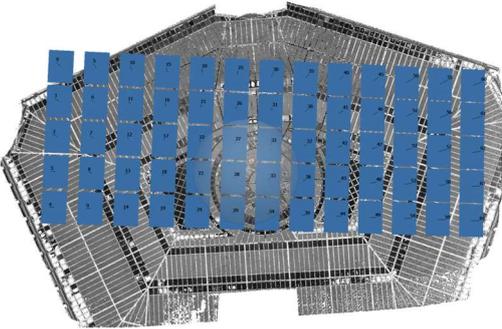
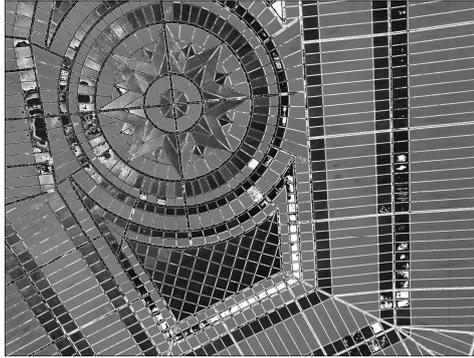
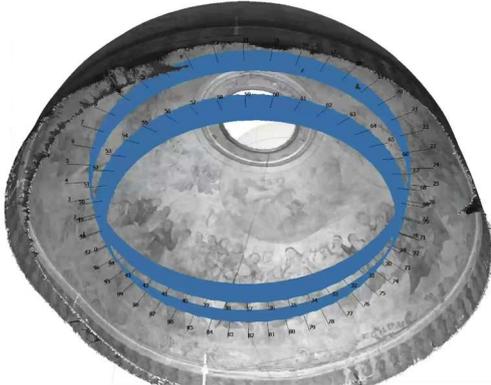
Dataset	Images distribution	Sample image
"Hall"		
"Chapel"		

Table 2. The synthetic images are distributed on a sphere ("Hall") and parallel to the object ("Chapel") used in this study with sample images.

From these methods and data, 3D point clouds were computed and a benchmarking against data from TLS was performed to assess the geometrical quality and aptitude of each image-based reconstruction method. The final part of the article presents several important take-home messages and inputs for further study in this area.

2. EXPERIMENT SETUP

In this paper, three case studies of heritage sites recorded using close-range photogrammetry will be presented. Dense point clouds from the case studies will be generated using three methods: conventional MVS implemented in commercial software, in this case Agisoft Metashape (<https://www.agisoft.com/>, accessed 15 January 2024) and from the open-source library OpenMVS 2.0 (<https://github.com/cdcseacave/openMVS>, accessed 15 January 2024), learning-based MVS (Vis-MVSNet) (Zhang et al., 2020), and NeRF. Terrestrial laser scanning (TLS) datasets were also used as reference data. Each case study presents different scenarios with different challenges related to 3D reconstruction, especially with respect to conventional MVS. These case studies comprise thus the following sites (Table 1): (1) the interior main hall of Warsaw University of Technology ("Hall") with a remarkable glass ceiling, walls and another architectural elements made of marble, sandstone and bricks (Markiewicz & Zawieska, 2015); (2) the dome of the Ladislas chapel in the church of St. Anne in Warsaw, Poland ("Chapel") consisting of detailed Rococo wet-painted frescoes (Górecka et al, 2022); and (3) the main facade of the St-Pierre-le-Jeune church in Strasbourg, France ("Facade"), which consists of close-range images of Neo-Byzantine church architecture.

In order to generate synthetic data on the basis of point clouds from terrestrial laser scanning, the image synthetic simulator was used (Markiewicz, et al., 2023). To generate the dataset, individual images were generated with user-specified parameters encompassing image size and resolution, focal length, camera positions, and interior orientation parameters. In the initial step, the projection centre and camera angle were set. Then, the point clouds are reprojected onto the reference plane, taking into account the model transformation matrix, projection, and observation range. The texture was derived from TLS intensity including the appropriate image depth buffer.

The distribution of the images was based on the shape of the analysed test site (Tab. 2). For the "Hall" dataset, the images were distributed parallel to the ceiling. The distance between the images in a row and between rows of approximately 2.5 m was assumed, and in total 65 images were generated, with resolution of 1910 x 1450 pixels, focal length 1035.407 px, and camera lens angle of 70 degrees. It was decided to generate images without the geometric distortion.

In the case of the "Chapel" dataset, it was decided to arrange the images on the sphere in two rows with an average distance of 40 cm between them. In total 90 images, with resolution of 4392 x 3043 pixels, focal length 2172.927 px, and camera lens angle of 70 degrees, were generated. Similar to the previous dataset, the images were free of geometrical distortion.

The "Facade" dataset, was meant to represent a real-world case and was taken using a Canon EOS 6D DSLR camera equipped with a 28 mm fixed lens, yielding 78 images with 20 megapixels resolution. Point clouds were generated for each case study using the three aforementioned methods. Conventional MVS point

clouds were processed using the Semi-Global Matching (**OpenMVS-SGM**) and PatchMatch (**OpenMVS-PM**) method implemented in the OpenMVS library. Agisoft **Metashape** was used to represent commercial conventional MVS reconstruction. The learning-based MVS method from **Vis-MVSNet** (Zhang et al., 2020) was also used to process the datasets. As a training dataset the BlendedMVS (Yao et al., 2019) were used. BlendedMVS is a large-scale MVS dataset for generalised multi-view stereo networks. The dataset contains 17k MVS training samples covering a variety of 113 scenes, including architectures, sculptures and small objects. The NeRF approach used by default the Nerfacto method (Tancik et al., 2022), implemented within the **Nerfstudio** API. For Nerfstudio, the initial image orientation was carried out using Metashape, before conversion into Nerfstudio format.

Geometric comparisons were performed using the CloudCompare software. The availability of TLS reference point clouds enabled the geometric analysis to be based on the Cloud-to-Mesh (C2M) functionality, which yields signed values for the errors. For this purpose, a 3D model mesh was created from the reference TLS point cloud to which each tested method's results will be compared.

For the "Hall" dataset, further analysis based on point cloud completeness (relative to the reference) particularly in challenging elements, e.g. glasses, were considered. The Horizontal and vertical profiles of the dome in the "Chapel" dataset were also extracted and an analysis performed on each method's capability to reconstruct the dome geometry. In order to determine the percentage of completeness in the "Hall" dataset, orthoimages of the point clouds from the five tested methods and one from the reference TLS dataset were generated. The orthophotos were then inputted into a Matlab[®] script which counted the number of empty pixels and computed a percentage of non-empty pixels relative to the total number of pixels in the orthoimage. Finally, for the "Facade" dataset an additional density analysis was also performed.

3. RESULTS AND DISCUSSIONS

3.1. "Hall" dataset

For the "Hall" dataset, a point cloud completeness analysis was performed. The completeness value was obtained by counting the number of non-empty pixels in the orthophotos generated by each method, and visually represented in Figure 1. A visual inspection of Figure 1 shows that the reflective nature of the dataset posed indeed a challenge for 3D reconstruction. Even the reference TLS dataset was not able to reach a 100% completeness rate in this regard, with minor holes present in some of the brightest glass panels on the dataset.

Among the MVS methods, Vis-MVSNet suffered from the worst output. Indeed, Vis-MVSNet had difficulties not only on minor locations, but also more generally as it was not able to properly construct the left part of the ceiling (see Figure 1). Overall, Vis-MVSNet yielded a completeness of 90.84%. OpenMVS-SGM suffered from a similar problem but to a lesser degree; it managed to score 93.69% on its completeness rate. Both Metashape and especially OpenMVS-PM generated arguably satisfactory results with completeness scores of 97.76% and 99.05%, respectively. In the case of OpenMVS-PM, these results mimic the reference TLS the most, barring several minor problems.

Nerfstudio, on the other hand, scored an impressive 99.98% on the completeness score, even outperforming TLS. These results must however be considered with caution due to several aspects. First of all, the results of Nerfstudio used synthetic images generated by TLS; the higher score is therefore suspect to geometric errors. Secondly, despite the high completeness score, Nerfstudio also generated the most noise among the other contenders which lowers its usability from a heritage documentation perspective, despite the impressive potential especially in dealing with this challenging scenario. In the case of "Hall", more than 50% of the points generated by Nerfstudio may be considered outliers, compared to an average of 15% in conventional MVS and 35% in Vis-MVSNet (Table 3).

Dataset	C2M Parameter	Metashape	OpenMVS-SGM	OpenMVS-PM	Vis-MVSNet	Nerfstudio
"Hall"	Avg. Error (cm)	0.00	0.80	0.70	1.00	3.70
	Std. Deviation (cm)	0.50	0.90	0.70	1.00	3.30
	Outlier (%)	0.33	15.87	1.99	35.60	52.72
"Chapel"	Avg. Error (cm)	0.00	0.10	0.10	0.30	0.00
	Std. Deviation (cm)	0.30	0.30	0.20	0.40	0.70
	Outlier (%)	0.32	0.24	0.32	0.48	10.35
"Facade"	Avg. Error (cm)	0.40	0.90	0.50	0.00	1.90
	Std. Deviation (cm)	2.00	1.70	1.60	0.02	4.80
	Outlier (%)	10.11	0.65	5.58	1.00	7.36

Table 3. Quality parameters of each dataset compared to their respective TLS references using the C2M method.

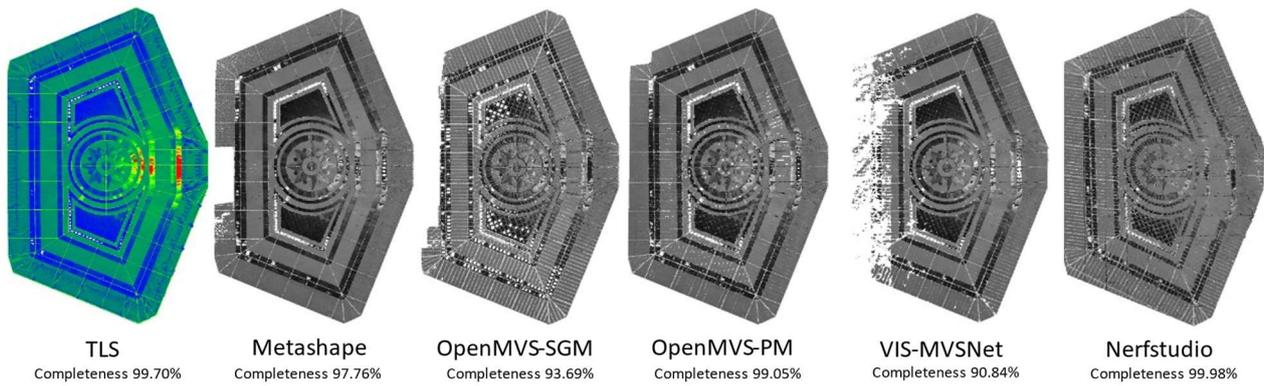


Figure 1. Analysis of point cloud completeness from the five tested methods, plus TLS.

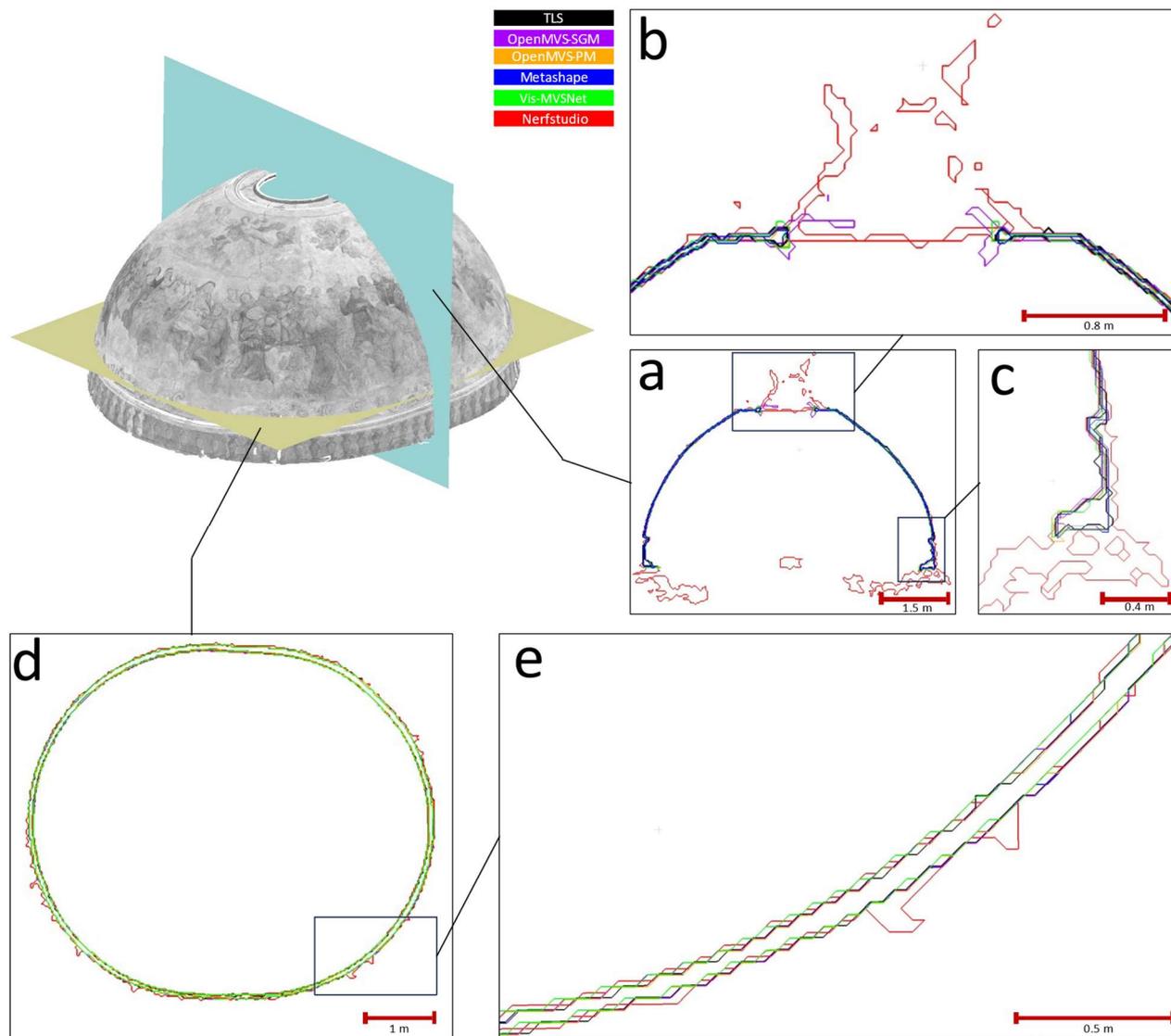


Figure 2. Vertical (a) and horizontal (d) profiles of the dome in the “Chapel” dataset. In (b), a subset of the vertical profile shows the part where the dome’s oculus is represented, while (c) shows the base of the dome where it meets the drum. In (e), a subset of the horizontal profile showcases in more detail the geometry of the dome.

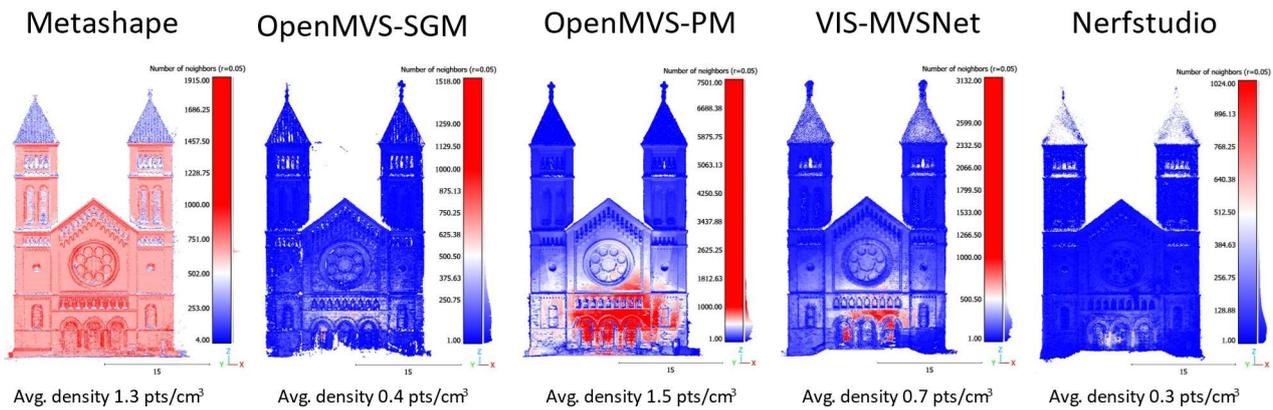


Figure 3. Density analysis on the “Facade” dataset. For all results shown in this figure, the saturation of the colour legend was fixed to show a gradation between red, white and blue from 1,000 points to 1 point (per the set neighbouring radius of 5 cm).

3.2. “Chapel” dataset

For the “Chapel” dataset, a dome was reconstructed using the tested methods. From the C2M analysis (Table 3) conducted from the results, most of the tested algorithms managed to create satisfactory results. The average error ranges from 0 to 3 mm, at which point the differences may be safely attributed to random errors.

The standard deviation values yielded similarly good results across the methods, with Nerfstudio registering a slightly higher value (7 mm) compared to the other four which averaged at 3 mm. Finally, using the set tolerance of 2.7 cm Metashape, OpenMVS and Vis-MVSNet only generated negligible amounts of outliers at less than 1%. However, Nerfstudio registered 10.3% of its points as outliers.

In order to do a more detailed inspection on these results, two profiles of the dome were generated (Figure 2). From both the horizontal and vertical profile, it may be seen that as far as the dome structure is concerned, all five methods were able to reconstruct it properly. Already from the horizontal profile (Figure 2(e)), one may observe that Nerfstudio generated more noise than the other methods.

However, some errors were registered at two points of change: one at the bottom of the dome towards the base and the drum, and another at the summit at which an opening (an oculus) is present. From Figure 2(b), it may be seen that OpenMVS-SGM and Nerfstudio encountered problems in identifying the opening, indeed both generated points where there should not be any. Nerfstudio goes even further and generates points across the profile and beyond the oculus, points which were definitely registered as outlier in the preceding analysis.

This important presence of noise may be attributed to two potential error sources. First, the inadequate post-processing of Nerfstudio point cloud, which most likely depended on a simple SOR-based noise cleaning algorithm as opposed to more sophisticated methods used in conventional and learning-based MVS. Secondly, the nature of the NeRF reconstruction depends on discrete neural radiance elements which are a function of object density and by extension transparency; the parameters of which density threshold constitute geometric point may therefore play a role. This may further be complicated by the fact that the “Chapel” dataset consisted of synthetic grayscale images. In both cases, these observations point to problems not precisely linked to the NeRF method per se, but rather to the conversion method between the neural radiance and the point cloud, which remains a bottleneck as also observed by Croce et al. (2023).

3.3. “Facade” dataset

For the real world “Facade” dataset, C2M analysis showed that all methods manage to yield average errors of less than 2 cm (Table 3). These results are satisfactory when considering that the resolution of the reference TLS dataset is 1 cm. In terms of standard deviation, OpenMVS-PM performed best. Nerfstudio was able to both visually and numerically achieve much better results compared to the synthetic datasets, registering a standard deviation value of 4.8 cm.

In terms of noisy points considered as outliers, OpenMVS-SGM performed best with less than 1% of points considered as noise. Nerfstudio performed very well compared to Metashape, registering only 7.36% to Metashape’s 10.11%. This may indicate that the NeRF method, like conventional photogrammetry, also benefits from less challenging environments. Indeed, the nature of this dataset (sandstone) provides an ample case for dense matching due to the presence of textures.

For the “Facade” dataset, a further density analysis was performed (Figure 3). In line with findings in Murtiyoso et al. (2023), Metashape presented a very homogeneous density. This indicated a post-processing which may involve point cloud subsampling by the software. This behaviour is not found in the four other open-source algorithms. Overall, OpenMVS-PM and Metashape generated denser results (between 1.3 and 1.5 points/cm³), while OpenMVS-SGM and Nerfstudio tend to produce less dense point clouds (0.4 and 0.3 points/cm³, respectively), consistent with visual inspection on this dataset as well as the other two.

4. CONCLUSIONS

In this paper, five state-of-the-art 3D reconstruction algorithms were tested on three heritage datasets. **Metashape** still proved to be a versatile software when dealing with image-based reconstruction, performing well on the reflective surface challenge of “Hall” both in terms of geometric accuracy and completeness. However, due to its commercial nature, it is naturally not possible to understand the inner workings of this software. That being said, Metashape showed signs of integrating state-of-the-art MVS methods since it showed similar results to OpenMVS-PM.

OpenMVS as an open-source method provided to variants in this experiment. **OpenMVS-SGM** performed well with real world applications where texture is available, but worked less effectively against reflective surfaces. SGM also presented more

noise in the case of the “Chapel” dataset. **OpenMVS-PM** performed very well in all datasets and may be considered the state-of-the-art of modern MVS (Stathopoulou & Remondino, 2023).

The learning-based method used in this paper, **Vis-MVSNet**, struggled to reconstruct the “Hall” dataset. Although it managed to score good geometric values, the amount of noise and incomplete point cloud means that it did not perform sufficiently for this challenging scenario. It did, however, work very well with more conventional scenarios as well as the real-world “Facade” dataset. Indeed, for the “Facade” scenario, **Vis-MVSNet** outperformed all the other tested methods in all the computed quality parameters.

Nerfstudio worked differently from the other MVS-based methods by benefitting from neural radiance fields. In theory, this would mean that it would work better for challenging surfaces such as the one encountered in the “Hall” dataset. While it did score the best completeness rate, it suffered from high levels of noise to the point of it being unusable for metric applications. However, **Nerfstudio** performed better in the “Facade” dataset, even scoring comparable results to **Metashape**.

In conclusion, at the moment of writing **OpenMVS-PM** seems to provide the best overall result on the three tested datasets. **Metashape** also delivered solid results. **Nerfstudio** and **Vis-MVSNet** worked better in conventional environments but still encountered the same (or even worse) problems than conventional MVS when dealing with reflective surfaces, especially when geometric quality is concerned. More investigation on these two novel methods is required; specifically for **NeRF** an interesting further work would be the study on the conversion method between the neural radiance fields and geometric point clouds.

ACKNOWLEDGEMENTS

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. The authors also wish to thank Nicolas Hoffmann (Master student at INSA Strasbourg) for his help in acquiring and processing the images for the “Facade” dataset.

REFERENCES

Balloni, E., Gorgoglione, L., Paolanti, M., Mancini, A., Pierdicca, R., 2023. Few shot photogrammetry: a comparison between NeRF and MVS-SfM for the documentation of cultural heritage, in: ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. pp. 25–30.

Bedford, J., 2017. Photogrammetric Applications for Cultural Heritage. Historic England, Swindon.

Croce, V., Caroti, G., Luca, L. De, Piemonte, A., 2023. Neural radiance fields (NeRF): review and potential applications to digital cultural heritage, in: ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. pp. 25–30.

Górecka, K., Łapiński, S., & Markiewicz, J. (2022). Example of using long-term monitoring and non-invasive measurements for Structural Health Monitoring. Cause study of St. Anna’s Church in Warsaw. *Journal of Modern Technologies for Cultural Heritage Preservation*, 1(1).

Grilli, E., & Remondino, F. (2019). Classification of 3D digital heritage. *Remote Sensing*, 11(7), 1–23.

Markiewicz, J.S. and Zawieska, D., 2015, June. Quality assessment of the TLS data in conservation of monuments. In *Optics for Arts, Architecture, and Archaeology V* (Vol. 9527, pp. 219–228). SPIE.

Markiewicz, J., Kowalczyk, M., Karwel, K., Kot, P., & Markiewicz, Ł. (2023). The evaluation of structure-from-motion workflow with the TLS synthetic images simulator – the cultural heritage approach. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII(June), 25–30.

Martin-Brualla, R., Radwan, N., Sajjadi, M.S.M., Barron, J.T., Dosovitskiy, A., Duckworth, D., 2021. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections, in: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. pp. 7206–7215.

Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R., 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, in: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. pp. 405–421.

Murtiyoso, A., & Grussenmeyer, P. (2017). Documentation of heritage buildings using close-range UAV images: dense matching issues, comparison and case studies. *The Photogrammetric Record*, 32(159), 206–229.

Murtiyoso, A., Grussenmeyer, P., 2023. Initial assessment on the use of state-of-the-art NeRF neural network 3d reconstruction for heritage documentation, in: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. pp. 25–30.

Murtiyoso, A., Pellis, E., Grussenmeyer, P., Landes, T., & Masiero, A. (2022). Towards Semantic Photogrammetry: Generating Semantically Rich Point Clouds from Architectural Close-Range Photogrammetry. *Sensors*, 22(3).

Stathopoulou, E. K., Battisti, R., Cernea, D., Remondino, F., & Georgopoulos, A. (2021). Semantically derived geometric constraints for MVS reconstruction of textureless areas. *Remote Sensing*, 13(6), 1–19.

Stathopoulou, E. K., & Remondino, F. (2023). A survey on conventional and learning-based methods for multi-view stereo. *Photogrammetric Record*.

Tancik, M., Weber, E., Ng, E., Li, R., Yi, B., Wang, T., Kristoffersen, A., Austin, J., Salahi, K., Ahuja, A., McAllister, D., Kanazawa, A., 2022. Nerfstudio: A Framework for Neural Radiance Field Development [WWW Document]. <https://github.com/nerfstudio-project/nerfstudio>.

Vandenabeele, L., Häcki, M., Pfister, M., 2023. Crowd-sourced surveying for building archaeology: the potential of structure from motion (SFM) and neural radiance fields (NeRF), in: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. pp. 25–30.

Wei, Z., Wang, Y., Yi, H., Chen, Y., Wang, G., 2020. *Applied Sciences* (Switzerland) 10, 1275.

Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., & Quan, L. (2019). *BlendedMVS: A Large-scale Dataset for Generalized Multi-view Stereo Networks*. <http://arxiv.org/abs/1911.10127>

Zhang, J., Yao, Y., Li, S., Luo, Z., & Fang, T. (2020). *Visibility-aware Multi-view Stereo Network*. <http://arxiv.org/abs/2008.07928>