A COMPARISON BETWEEN 3D RECONSTRUCTION USING NeRF NEURAL NETWORKS AND MVS ALGORITHMS ON CULTURAL HERITAGE IMAGES

F.Condorelli [1*], F. Rinaudo [1] , F. Salvadore [2] , S. Tagliaventi [2]

[1] DAD, Department of Architecture and Design, Politecnico di Torino, Italy - (francesca.condorelli, fulvio.rinaudo)@polito.it
[2] CINECA – HPC Department, Rome, Italy - (f.salvadore, s.tagliaventi)@cineca.it

Commission II, WG II/8

KEY WORDS: NeRF, Neural Networks, 3D Reconstruction, Photogrammetry, MVS algorithms, Cultural Heritage, Metric Quality Assessment

ABSTRACT:

In this research, an innovative comparison between 3D reconstructions obtained by means of Artificial Intelligence, in particular NeRF Neural Networks, and by Structure-from-Motion (SfM) and Multi-View-Stereo (MVS) open-source algorithms is proposed. The 3D reconstruction comparison is performed on two test cases, one of cultural interest, one useful only for technical discussion. It is known that the approaches are traditionally used with different objectives and in different contexts but they can however also be used with similar purpose, i.e., 3D reconstruction. In particular, we were interested in evaluating how NeRF reconstructions are accurate from a metric point of view and how the models obtained from the application of NeRF differ from the model obtained from the classical photogrammetry. By analyzing the results in the considered test cases, we show how NeRF networks, although computationally demanding, can be an interesting alternative or complementary methodology, especially in cases where classical photogrammetric techniques do not allow satisfactory results to be achieved. It is therefore suggested to expand efforts in this direction by exploiting, for example, the numerous improvement proposals of the original NeRF network.

1. INTRODUCTION: MOTIVATION AND STATE OF THE ART

Historical images are often the only remaining traces of monuments that have been lost or changed over time. Starting from images, however, it is possible to virtually reconstruct the cultural assets, thus providing a non-material way of preserving the goods of great use in terms of historical memory. 3D reconstruction methodologies have been proposed and used for several years and are traditionally carried out through algorithms and photogrammetry pipelines, i.e., Structure-from-Motion (SfM) and Multi-View-Stereo (MVS) approaches (Schönberger and Frahm, 2016). Actually, the 3D reconstruction starting from historical images poses several challenges such as the identification of the starting images from the vast archive material and the ability of photogrammetric algorithms to work with numerically reduced and low-quality images. We previously proposed an innovative approach based on deep learning image recognition, for the automatic selection of images of a specific monument from archive material, i.e., historical images or film footage. The strategy was designed to be used upstream of a standard photogrammetry pipeline and has proven effective in specific test cases (Condorelli et al., 2020).

In this work, we propose an exploratory application of an alternative 3D reconstruction technique that can be of help in the conditions of particular criticality of the starting material in which, as known, photogrammetric techniques can lead to reduced quality results. In particular we consider the recent NeRF neural networks (Mildenhall et al., 2020) as a complementary and alternative tool to traditional photogrammetry. Neural Radiance Fields (NeRF) approach models the radiance field and density of a scene from a set of input images within the weights of a neural network and can render high-resolution photorealistic novel views of real objects and scenes from RGB images captured in natural settings (Mildenhall et al., 2020). NeRF method aims to generate volumetric representations of the scene and is typically not preferred for 3D reconstruction due to its inherent limitations. However, starting from the trained network, it is possible to reconstruct the triangulated mesh of the analyzed object, using well-established computer graphics algorithms (e.g., marching cubes, Lorensen and Cline, 1987). The potentiality of NeRF for 3D reconstruction is that, differently from classical photogrammetry, it is able to 3D reconstruct objects that present features that could lead to a failure of the photogrammetric process, such as thin objects (trees, leaves), or reflecting (metal) objects, etc. The network can represent detailed scene geometry with complex occlusions, without any background isolation or masking.

Our analysis focuses on the evaluation of two test cases – one properly of cultural interest, one useful only for technical discussion – and compares the reconstruction results using traditional photogrammetry against innovative NeRF strategy.

We will only consider the original NeRF network, devised for images of static subjects captured under controlled settings. However, a number of subsequent elaborations have already been proposed to overcome some found limitations of NeRF technology. For example, in order to adapt the network to other types of images acquired under different situations, NeRF-W was proposed and applied to the reconstruction of famous monuments starting from the Photo Tourism dataset (Snavely et al., 2006). NeRF-W managed to synthesize novel views of complex outdoor scenes using only unstructured collections of in-the-wild photographs in which variable illumination or

[*] Corresponding author

https://doi.org/10.5194/isprs-archives-XLIII-B2-2021-565-2021 |

transient occlusions are present (Martin-Brualla et al., 2020). Another remarkable attempt to address the different lighting condition issues was proposed in Srinivasan et al., 2020. The method takes as input a set of images of a scene illuminated by unconstrained known lighting, and produces as output a 3D representation that can be rendered from novel viewpoints under arbitrary lighting conditions. These studies could be particularly interesting in the case of historical images that present low resolution, black and white colour and bad illumination. Another interesting modification of the NeRF network is related to single shot reconstruction. PixelNeRF (Yu et al., 2020) is a learning framework that predicts a continuous neural scene representation conditioned on one or few input images allowing novel view synthesis and single image 3D reconstruction of a specific scene by implicitly encoding volumetric density and colour through a neural network. This network potentially addresses one of the most serious issues of image-based processes since photogrammetry requires two or more images in order to achieve reconstruction. Finally, it is worth mentioning the NeRFs related to the video process. D-NeRF could be of help for the 3D reconstruction of images taken from historical film footage. D-NeRF is a method that extends neural radiance fields to a dynamic domain, allowing to reconstruct and render novel images of objects under rigid and non-rigid motions from a single camera moving around the scene (Pumarola et al., 2020). Another version of D-NeRF allows the reconstruction of non-rigidly deforming scenes using photos/videos captured casually from mobile phones from arbitrary viewpoints (Park et al., 2020).

2. MATERIAL AND METHODS

2.1 Photogrammetric pipeline

The photogrammetric pipeline chosen as reference in this research is the COLMAP (Schönberger and Frahm, 2016) open-source Structure-from-Motion and Multi-View Stereo (MVS) algorithm implementation, developed by ETH of Zurich (https://github.com/colmap/colmap, 2020). The SfM sequential processing pipeline is: 1) Feature detection and extraction, 2) Feature matching and geometric verification, 3) Structure and motion reconstruction (Schönberger and Frahm, 2016).

After that, the MVS implementation in COLMAP is used for the generation of the mesh. Multi-View Stereo (MVS) in COLMAP uses the output of SfM to compute depth and/or normality information for each pixel in an image. Merging the depth and normal maps from multiple 3D images then produces a dense point cloud of the scene. Using the depth and normal information from the merged point cloud, algorithms such as Poisson surface reconstruction can recover the 3D surface geometry of the scene (Schönberger and Frahm, 2016).

2.2 Neural Radiance Fields (NeRF) for View Synthesis

According to the Neural Radiance Field (NeRF) approach (Mildenhall et al., 2020), a continuous scene can be represented as a 5D function whose inputs are a 3D location $\boldsymbol{x} - (x, y, z)$ and a 2D viewing direction. The viewing direction can be expressed in terms of angles (ϑ, ϕ) or using a normalized vector $\boldsymbol{d}$. The outputs are an emitted RGB color $\boldsymbol{c} - (\boldsymbol{r}, \boldsymbol{g}, \boldsymbol{b})$ and volume density $\boldsymbol{\sigma}$.

The 5D scene representation is approximated with a deep fully-connected neural network, also known as multilayer perceptron (MLP):

$$F_{\Theta} : (\boldsymbol{x}, \boldsymbol{d}) \rightarrow (\boldsymbol{c}, \sigma) \qquad (1)$$

where Θ are the weights to be optimized.

In order to obtain a multiview consistent representation, the first part of the network predicts the volume density as a function of only the location $\boldsymbol{x}$. This part of the network is composed by 8 fully-connected layers, ReLU activations and 256 channels per layer and outputs, in addition to σ, a 256-dimensional feature vector. In the second part of the network, the input is the concatenation between the previously identified feature vector and the camera viewing direction and is passed to one additional fully-connected layer: the final output is the view-dependent RGB color.

Following principles from classical volume rendering (Kajiya and Herzen, 1984), the volume density can be interpreted as the differential probability of a ray terminating at an infinitesimal particle at location $\boldsymbol{x}$. To perform volume rendering, the expected color $\boldsymbol{C}(\boldsymbol{r})$ of camera ray $\boldsymbol{r}$ may be computed integrating, along the ray, the volume density weighted by the accumulated transmittance from near to far bounds. To compute the integral, tailored quadrature rules have to be implemented.

The adopted rendering function is differentiable so that it is possible to optimize the scene representation by minimizing the residual between synthesized and ground truth observed images. Furthermore, two additional improvements were designed to enable the representation of high-resolution complex scenes.

The first improvement is known as *positional encoding* and is used to enhance the reproduction of high-frequency variations in color and geometry.

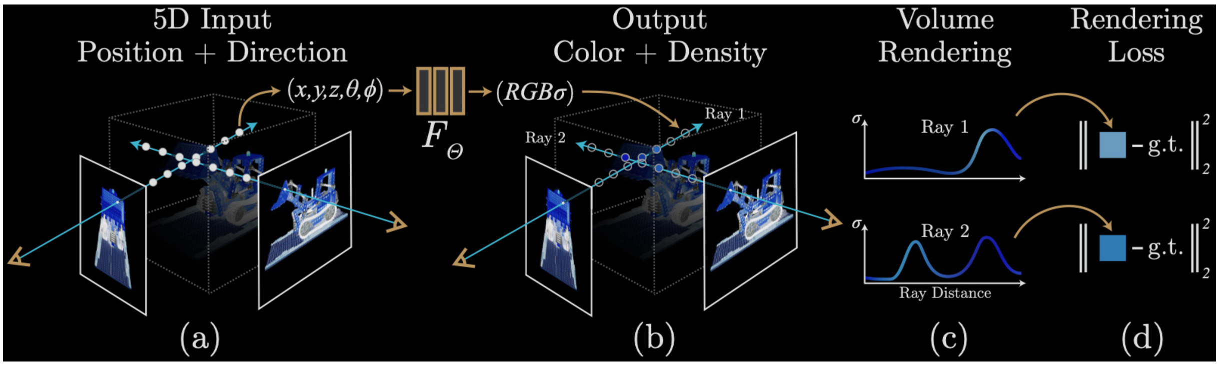


Figure 1. Overview of neural radiance field scene representation. Images are synthetized by sampling 5D coordinates (location and viewing direction) along camera rays, feeding these locations into an MLP to produce a color and volume density (left and center). Volume rendering techniques are used to composite these values into an image.

To this aim, F_Θ is reformulated as the composition of two functions $F'_\Theta \circ \gamma$, where γ is a predefined mapping from $\mathbb{R}$ into a higher dimensional space $\mathbb{R}^{2L}$ and F'_Θ is still simply a regular MLP. The γ encoding function is chosen as:

$$\gamma(p) = (\sin(2^0 \cdot \pi \cdot p), \cos(2^0 \cdot \pi \cdot p), \sin(2^1 \cdot \pi \cdot p), \cos(2^1 \cdot \pi \cdot p), \dots, \sin(2^L \cdot \pi \cdot p), \cos(2^L \cdot \pi \cdot p)) \tag{2}$$

and is applied to each of the three coordinate values in $\boldsymbol{x}$ and to the three components of $\boldsymbol{d}$. The second improvement is called *hierarchical volume sampling*. In practice, two networks are simultaneously optimized: one "coarse" and one "fine". The procedure manages to allocate more samples to regions we expect to contain visible content.

For each scene during the training stage a separate neural continuous volume representation network is produced. To perform the training, a dataset of RGB images of the scenes is required, alongside the camera poses, intrinsic parameters, and scene bounds. These quantities can be estimated by means of COLMAP structure-from-motion package, as it happens in the photogrammetric pipeline. At each training iteration, random samples of camera rays are extracted from the set of all pixels of the dataset, and hierarchical sampling is performed. Then, volume rendering procedure is used to calculate the color of each ray from both sets of samples. The final loss includes both coarse and fine sampling:

$$\mathcal{L} = \sum_{r \in \mathcal{R}} \left[|C_c(r) - C(r)|^2 + |C_f(r) - C(r)|^2 \right] \tag{3}$$

where $\mathcal{R}$ is the set of rays in each batch, $C(\boldsymbol{r})$, $C_f(\boldsymbol{r})$, $C_c(\boldsymbol{r})$ are the ground truth, fine volume predicted and coarse volume predicted RGB colors for ray $\boldsymbol{r}$, respectively. As concerns the training configuration, each batch includes 4096 rays, each sampled at $N_{coarse} = 64$ and $N_{fine} = 128$. Adam optimizer is adopted (Diederik et al., 2015).

We adopted the publicly available implementation on GitHub (https://github.com/bmild/nerf) which is deeply described in Mildenhall et al., 2020. Starting from the trained network it is easily possible to generate novel views. To also obtain a 3D reconstruction, the implemented software takes advantage of the marching cube algorithm (Lorensen and Cline, 1987), a well-established computer vision method able to extract a triangulated surface starting from a field known in an equispaced Cartesian mesh.

Besides the evident potentiality of these algorithms for 3D reconstruction in general, in the architectural and Cultural Heritage field it is interesting to investigate how much these reconstructions are accurate from a metric point of view and how the models obtained from the application of NeRF differ from the models obtained from the classical photogrammetry.

In order to assess the quality of the 3D reconstruction obtained from NeRF and to compare it with the model obtained from photogrammetry, the NeRF pipeline and the SfM/MVS pipeline were tested in parallel on the same case studies.

3. CASE STUDIES AND DATASETS

The comparison between NeRF and MVS reconstruction methods was performed using a case study approach.

The first test was carried out on a dataset specifically created for this investigation. The dataset name is "flower" and consists of 10 images acquired by the authors (Figure 2).

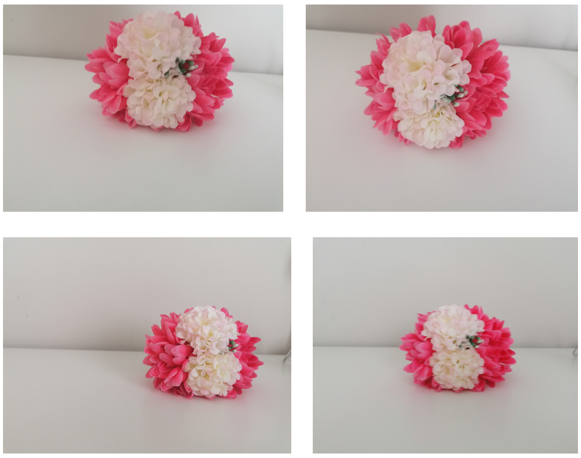


Figure 2. A selection of the images contained in the dataset "flower".

Given the geometric characteristics of the object, i.e., the high-frequency of similar patterns, photogrammetry is expected to perform poorly on such a dataset and the failure of the whole reconstruction procedure is likely. In view of NeRF training, as explained in Mildenhall et al., 2019, the optimal way to acquire data is to take a set of images of a static scene, where the maximum disparity between views is no more than ~1/8 of the horizontal field of view between images. The images were acquired following a rough grid pattern, starting from the top to the bottom.

The second test was carried out on a case study concerning the cultural heritage field, this being the main application field of our investigation. The selected monument is the Tour Saint Jacques in Paris. This bell tower is in flamboyant gothic style and it has been inscribed in the UNESCO Heritage List since 1998 for its historical importance.

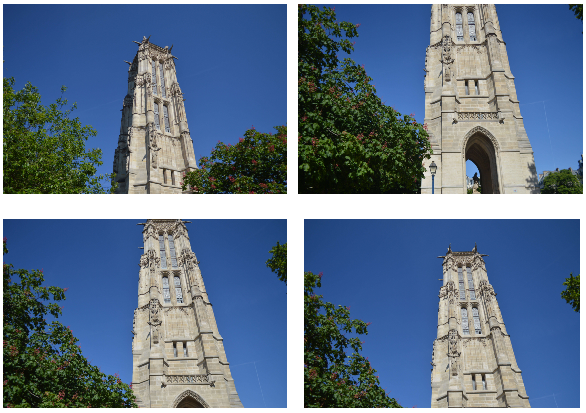


Figure 3. A selection of the images contained in the dataset "tower".

The building, although existing today, has been moved from its initial configuration and is therefore a significant representative of a monument that requires historical material for documentation. In fact, the tower is present in several historical archive images, both as photographs and as film footage. On the other hand, since it is still accessible, it is also possible to carry out surveys or in any case obtain high quality images. In short, given its hybrid nature, it represents an ideal case study for testing reconstruction methods under optimal conditions as well as under realistic conditions from a historical point of view.

The dataset "tower" includes images which were acquired by the authors during a survey on the place. They were not acquired in view of NeRF methodology, and for this reason they do not follow the standard acquisition rules for the NeRF. The main challenge related to the usage of this dataset with respect to NeRF reconstruction is that, given the considerable height of the tower, the dataset images were taken very close to the object. As a consequence, the conditions significantly different from the laboratory images considered for NeRF training and discussed in Mildenhall et al., 2020. The dataset, however, represents a more realistic scenario when dealing with cultural heritage and for this reason it was considered suitable for this type of investigation. Indeed, the aim of this experiment is to test the NeRF reconstruction under challenging conditions somehow similar to the real-world historical case that typically involves the usage of random images of an object.

4. RESULTS AND DISCUSSION

As explained in Section 2, for both the datasets the poses were generated by means of the SfM pipeline implemented in COLMAP. After this step, the two reconstruction methodologies were applied.

As concerns NeRF, the network training required a considerable computing power and was performed using High Performance Computing resources, i.e., Marconi100 cluster equipped with NVIDIA V100 GPUs available at Cineca HPC Department. Around 300,000 iterations were needed to complete the training and the complete run lasted about 24 hours. This is a clear limitation of NeRF approach if compared to photogrammetry which is in general much more lightweight from the computational point of view.

As anticipated in section 3, the MVS reconstruction step failed when applied to the "flower" dataset. All the other attempts completed successfully at least from the software point of view. The summary of the attempts/outcomes is provided in Table 1.

Dataset	NeRF	MVS	Reference
Flower	✓	✗	Direct survey
Tower	✓	✓	UAVs

Table 1. Summary of the reconstruction methods applied to the two case studies: success/failure of each attempt is provided.

In the following figures, the results of the NeRF trained using the "flower" dataset are reported: in Figure 4 some example images automatically generated by the network application as novel views of the same training scene are shown.

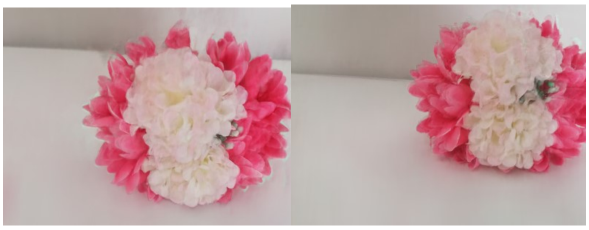


Figure 4. Images of two novel views generated by NeRF trained using the "flower" dataset.

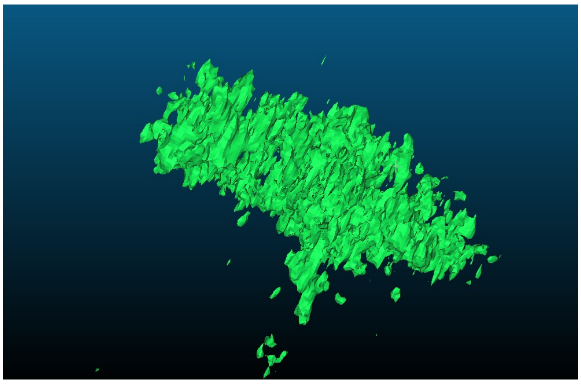


Figure 5. A lateral view of the mesh generated by the NeRF on the "flower" dataset.

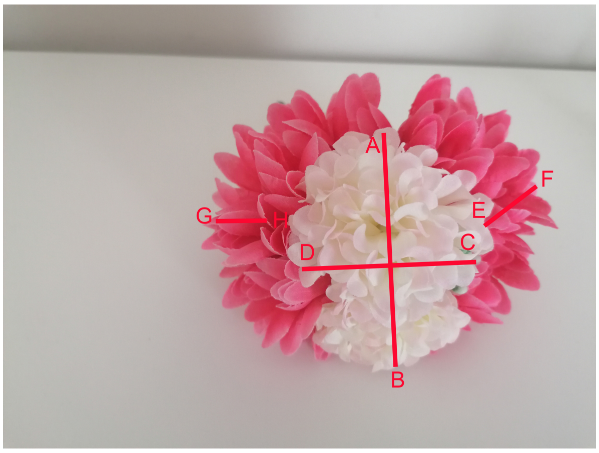


Figure 6. Distances directly measured on the surface of the "flower" object.

Distance	Mesh [cm]	Reference[cm]	Residuals[cm]
AB	0.09	0.10	0.01
DC	0.04	0.05	0.01
EF	0.04	0.06	0.02
GH	0.05	0.07	0.02

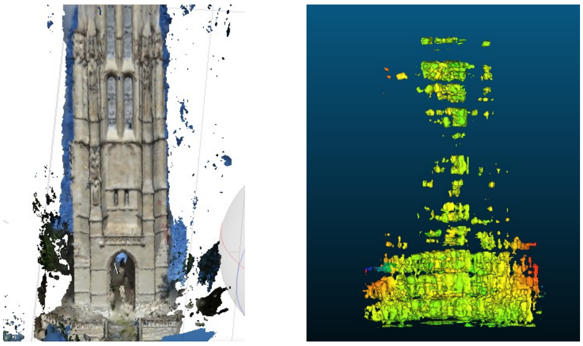
Table 2. Distances extracted from the mesh and from the object (reference) and the computation of the residuals.

To discuss the metric results of the NeRF reconstruction applied to this dataset, a direct survey on the object was carried out. Some distances were directly measured on the object surface (Figure 6) and reported in Table 2, chosen as reference for the comparison. Then, the same distances were measured in the obtained mesh and used, together with the reference measures, for the computation of the residual values between the two reconstructions (Table 2).

As shown in Table 2, the difference between the two reconstructions is between 1 cm and 2 cm. Given the limited values of the residuals, it is proved that the reconstruction obtained from a NeRF is suitable for the extraction of matrix information.

As concerns the "tower" dataset, both the NeRF and the MVS reconstructions were successfully completed.

The results are shown in Figure 7. It is worth noting that, since the images of the dataset were not sufficiently distributed, MVS was unable to reconstruct the lateral views of the tower. On the other hand, the NeRF reconstruction was able to achieve better results with respect to this feature. However, the mesh resolution is higher in the MVS model if compared to the NeRF one.



(a) (b)

Figure 7. Models obatained processing the same image dataset "tower"of the Tour Saint Jacques: (a) mesh from MVS in which only one façade was reconstructed; (b) mesh from NeRF, in which two views were reconstructed.

As for the metric comparison between the two reconstructed meshes, the model obtained from a recent UAV survey (by Iconem) was used and chosen as reference, because of the missing parts in the MVS reconstruction.

The comparison was performed using the Mesh-to-Mesh distance comparison algorithm in Cloud Compare software (https://www.danielgm.net/cc/, April 2021). This open-source software allows the comparison of point clouds by estimating their distances using the Multiscale Cloud Model Comparison (M3C2) plug-in, which uses the normal directions of one of the two surfaces to calculate local distances and provides estimations of the confidence intervals for each measurement (Lague et al., 2013).

The results are shown in Figure 8. It is noted that the distance values between the two meshes are generally very small (around 0.003 m). Concluding, the results are acceptable in the architectural context.

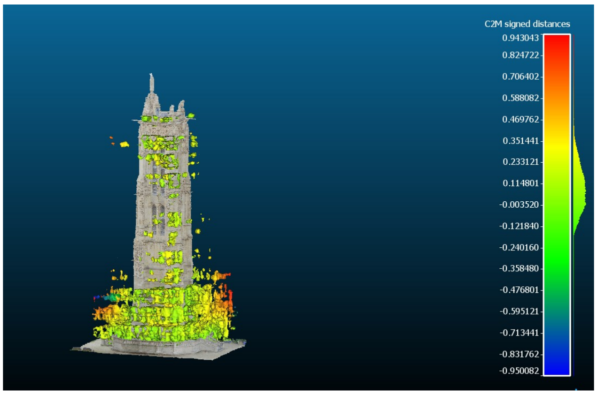


Figure 8. Mesh-to-Mesh distance comparison between the mesh from the NeRF and from photogrammetry (chosen as reference).

The potentiality of the use of NeRF is that it could help to reconstruct other views of the building that normally photogrammetry struggles to obtain for different reasons: dataset not acquired for the scope of the reconstruction, images from the web, historical images, etc. The combination of the two techniques is the strength of the proposed method.

5. CONCLUSIONS

In this work we proposed an exploratory analysis of comparison between two methodologically very different techniques for obtaining 3D reconstructions starting from images. In particular, we considered traditional photogrammetry pipelines based on Structure-from-Motion (SfM) and Multi-View-Stereo (MVS) open-source algorithms. Then, we considered a recent approach based on the use of Neural Radiance Fields (NeRF) neural networks. Both approaches rely on input image datasets but, while photogrammetry is geared towards 3D reconstruction, NeRF networks are designed to generate volumetric representations of the scene and render high-resolution photorealistic novel views of real objects and scenes. However, through classical computer vision techniques, it is possible to obtain 3D reconstructions also starting from the results of suitably trained NeRF networks. Analyzing the results of the comparison in a first test case, we have highlighted how NeRF networks are able to provide a 3D reconstruction, albeit of limited quality, even in conditions in which the photogrammetry algorithms are unable to offer any results. Considering a monument test case, we have instead highlighted how both methods allow to achieve reasonable results for 3D reconstruction, each with specific strengths and weaknesses. We also attempted to make a metric comparison between the two results and to interpret the obtained results. Overall, the analyzed test cases offer interesting insights relating to our specific field of interest which is the 3D reconstruction of lost cultural assets. In fact, it is known in this context that the number and quality of available images can be very limited, and this means that NeRF networks can be an alternative and complementary tool for the most difficult cases.

The proposed analysis is a preliminary analysis based on the original NeRF network structure. The work can and deserves to be extended in different directions. For example, in addition to the variations of the NeRF network presented in the introduction, it is interesting to consider other recently proposed evaluation studies such as those proposed by Zhang et al. 2020, or, in the same line of thought, in Wu et al., 2021 or Giegler and Koltun, 2020.

ACKNOWLEDGEMENT

This paper is dedicated to the memory of our friend and colleague Stefano Tagliaventi who choose the topic and devised the research, carrying out the first experiments.

REFERENCES

Condorelli, F., Rinaudo, F., Salvadore, F., Tagliaventi, S., 2020. A Neural Networks Approach to Detecting Lost Heritage in Historical Video, *ISPRS Int. J. Geo-Inf. 2020*, 9(5), 297; https://doi.org/10.3390/ijgi9050297.

Diederik, P., Kingma, Ba, J., 2015. Adam: A Method for Stochastic Optimization. In: *3rd International Conference for Learning Representations*, San Diego, 2015.

Gkioxari, G., Malik, J., Johnson, J., 2019. Mesh R-CNN, arXiv:1906.02739.

Kajiya, J., Herzen, B., 1984. Ray tracing volume densities. In: *Proceedings of the 11th annual conference on Computer graphics and interactive techniques (SIGGRAPH '84)*. Association for Computing Machinery, New York, NY, USA, 165–174. DOI:https://doi.org/10.1145/800031.808594

Lague, D., Brodu, N., Leroux, J., 2013. Accurate 3D comparison of complex topography with terrestrial laser scanner: Application to the Rangitikei canyon (N-Z). In: *ISPRS Journal of Photogrammetry and Remote Sensing*, 82, pp. 10-26, https://doi.org/10.1016/j.isprsjprs.2013.04.009.

Martin-Brualla, R., Radwan, N., Sajjadi, M., Barron, J., Dosovitskiy, A., Duckworth, D., 2020. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In: *CVPR*, 2021.

Mildenhall, B., Srinivasan, P., Tancik, M., Barron, J., Ramamoorthi, R., Ng, R., 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In: *ECCV 2020*.

Mildenhall, B., Srinivasan, P., Ortiz-Cayon, R., Khademi Kalantari, N., Ramamoorthi, R., Ng, R., Kar, A., 2019. Local Light Field Fusion: Practical View Synthesis with Prescriptive Sampling Guidelines. In: *ACM Transactions on Graphics (TOG)*, 2019.

Park, K., Sinha, U., Barron, J., Bouaziz, S., Goldman, D., Seitz, S., MArtin-Brualla, R., 2020. Deformable Neural Radiance Fields, arXiv:2011.12948.

Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W., Johnson, J., Gkioxari, G., 2020. Accelerating 3D Deep Learning with PyTorch3D, arXiv:2007.08501.

Riegler, G., Vladlen, K., 2020. Free View Synthesis. In: European Conference on Computer Vision, 2020.

Snavely, N., Seitz, S., Szeliski, R., 2006. Photo tourism: Exploring photo collections in 3D. In: *ACM Transactions on Graphics (SIGGRAPH Proceedings)*, 25(3), 2006, 835-846.

Schönberger, J. L., Frahm, J. M., 2016. Structure-from-motion revisited. In: *IEEE Conference on Computer Vision and Pattern Recognition*, CVPR 2016, Vol. 2016, 4104-4113, IEEE Computer Society.

Srinivasan, P., Deng, Zhang, X., Tancik, M., Mildenhall, Barron, J., 2020. NeRV: Neural Reflectance and Visibility Fields for Relighting and View Synthesis. In: *CVPR*, 2021.

Pumarola, A., Corona, E., Pons-Moll, G., & Moreno-Noguer, F., 2021. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE.

Wu, C., Liu, J., Huang, X., Li, Z., Yu, Ye, J., Zhang, J., Zhang, Q., Dou, X., Goyal, V., Xu, F., Pan J., 2021. Non–line-of-sight imaging over 1.43 km. In: *Proceedings of the National Academy of Sciences*, 118 (10) e2024468118; doi: 10.1073/pnas.2024468118.

Yu, A., Ye, V., Tancik, M., Kanazawa, A.: PixelNeRF: Neural Radiance Fields from One or Few Images, arXiv:2012.02190.

Zhang, Riegler, G., Snavely, N., Koltun, V., 2020. NeRF++: Analyzing and Improving Neural Radiance Fields, 2020. arXiv:2010.07492.