MACHINE LEARNING AS AN ALTERNATIVE TO 3D PHOTOMODELING EMPLOYED IN ARCHITECTURAL SURVEY AND AUTOMATIC DESIGN MODELLING

C. Palestini[1], A. Basso[2], M. Perticarini[3]

[1] Dipartimento di Architettura, Università degli Studi "G. d'Annunzio" Pescara, Italy, [2] SAAD_Unicam, Università di Camerino,
[3] Università della Campania Luigi Vanvitelli
caterinapalestini@libero.it - alessandro.basso@unicam.it - maurizio.perticarini@unicampania.it

Commission II

KEY WORDS: Instant Nerf, image-based survey, neural networks, artificial intelligence, volume rendering, real time rendering

ABSTRACT:

This paper presents some experiments on the use of an alternative technique, Nerf, based on artificial intelligence, which can be used in the survey of architectural structures or parts of them. The Nvidia video cards supported by the new RTX architectures are now able to manage an enormous amount of data both in the calculation of lighting and the rendering of digital shaders, and in the management of the number of polygons that can be displayed in real time in the scene. The support of artificial intelligence has further improved the three-dimensional digitization process in order to support the hardware power of algorithms that use neural networks to optimize and reduce calculation times, improving various aspects of the graphics and also of the digital representation, also influencing the representation on "image based" 3D acquisition methods, currently the most used low-cost systems in the photomodeling of objects or three-dimensional scenes. At the current state of the art, many commercial companies and research organizations, first Nvidia with its Instant Nerf, are betting on algorithms based on the Neural Radiance Field. The paper highlights the criticalities of this new system and shows the discrete results and the expandable research scenarios that can involve the impossible survey with photomodeling methodology related to works of art or architecture, due to the intrinsic nature of the materials that they compose them or because it is impossible or very difficult to photograph them.

1. INTRODUCTION

In recent years there has been a notable development of GPU hardware which, by means of new architectures, are able to manage a large amount of data, both as regards the calculation of the lighting and the realistic rendering of the shaders, and in the management in real time of the number of polygons present in the scene. Nvidia, with its new RTXs, appears to be the most competitive company in the sector, combining the capabilities of the hardware with the resources that the use of artificial intelligence offers. In fact, neural networks can significantly improve normal graphics performance, facilitating processes and significantly lowering the calculation times of the various operations necessary to obtain a high level of realism.

Some examples of artificial intelligence applications are the use of denoisers to improve the rendering output by optimizing the times until obtaining clear images in real time, or the brand new Nanite algorithm (developed by Nvidia) which allows to manage a very high number of polygons within a 3D scene in real time, favouring remarkable photorealism and a high level of image detail. For a couple of years, several companies and research groups from universities and private entities have been experimenting with new methods for the digitization of reality and are focusing on new algorithms based on the Neural Radiance Field (Nerf) (Mildenhall et al. 2020): an acquisition method that exploits machine learning and neural networks to improve data management by lowering processing times and amplifying realism in the management of lighting effects, such as caustics, refractions and reflections on materials.

Nvidia has developed its code called Instant Nerf, which has in the same name the concept of immediacy of results through the direct use of an artificial intelligence, and the training of it through data entered by humans, capable of generating complex 3d scenes using reverse volumetric rendering through a rather small number of photos.

Although estimating the depth and appearance of an object based on a partial view is a natural skill for humans, it is a challenging task for AI. Creating a 3D scene. using traditional methods can take a long time, depending on the complexity, amount of data entered and the resolution of the display. Bringing artificial intelligence into reconstructive operations improves computation times.

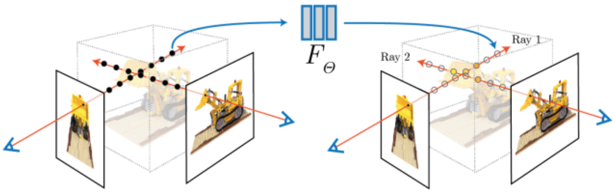


Figure 1. Diagram of how the Nerf rendering process works (image taken from" NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis")

2. HOW THE INSTANT NERF WORKS

The "Neural Radience Field" is a neural network aimed at generating new views of complex three-dimensional scenes based on a partial set of 2D images. The Nerf uses the input images already present in the dataset - and therefore acquired by humans - interpolating them to create a complete 3D scene. It basically tries to solve what is called Novel View Synthesis: it synthesizes an image characterized by an arbitrary point of view (camera target), from a series of source images each characterized by its own point of view. Nerf has therefore particularly increased the state of the art by probably replacing the GANs used up to now to achieve similar results. As evidenced by recent experiments, the Nerf system is 2.5 times faster than GANs. To optimize rendering times, Nerf implements various optimization processes: the process called Multi-layer Perceptron (MLP) in order not to calculate the density of a given point several times; positional coding and generic sampling (Xie et al. 2021)(Mildenhall et al. 2020)(Tancik et al. 2020). Early NeRF models rendered sharp scenes without artifacts in minutes, but still took hours to train. Instant NeRF, however, greatly reduces render time, as it relies on a technique called multi-resolution hash grid encoding, which is optimized to work efficiently on independent NVIDIA GPU system, without requiring complex machine learning on external servers and using a new and faster input coding method consisting of a single continuous 5D coordinate (spatial position (x, y, z) and observation direction (θ, φ)) and whose output is the volume density and the emitted radiance depending on the spatial position (RGBσ).

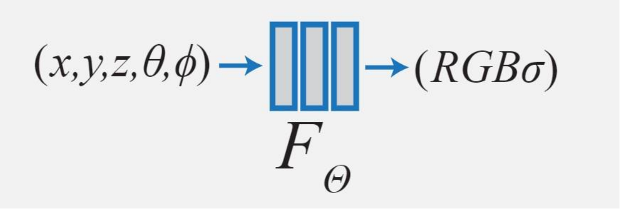


Figure 1. Algorithmic formula for increasing the quality of volumetric rendering defined by Nvidia

This formula is necessary to understand how important it is to know the direction of the point of view of an image: in the case of a completely Lambertian surface (which assumes in the rendering an ideal "opaque" surface, such as terracotta) it would not be so necessary, as an object with a similar characteristic has more or less the same colour regardless of the point of view from which it is framed; the question changes regarding a mirroring surface (perfect mirror), in this case each ray of light is reflected in a single direction and this means that a different reflection is shown depending on the point of view. Volume rendering allows the creation of a 2D projection of a 3D dataset. Each position of the camera corresponds to RGB data and information relating to the density "σ" for each voxel in space, through which the beam coming from the camera passes. The opposite process creates a 3D object from a series of 2D images that show the object from multiple angles and therefore from different perspectives. From these data the system predicts the depth and density of the framed objects. The Coldmap model, based on its own algorithms of 3d photogrammetry, enhanced by artificial intelligence, finally takes care of systematizing the information relating to each photographic shot (reverse raytracing) with the other shots, generating the volumetric cloud that returns the framed object in its entirety. By synthesizing the views and interrogating the 5D coordinates along the camera beams, the classic volume rendering techniques can be used to project the output colours and densities into an explorable model. In volumetric rendering, the processing does not stop at the specific data of the objects, but the acquisition rays penetrate inside them. it is to be considered the evolution of traditional ray tracing which only affects the illuminated surfaces of a 3d space by exploiting only a single beam combined with the camera to obtain information relating to the light sources, their intensity and any occlusions to the interception of objects. Volumetric rendering, working no longer with surfaces but with volumes and exploiting multiple Raytracing shots put into the system at the same time, does not need the calculation of global illumination, in which the rebounds of the secondary rays reproduce the correct ambient lighting. In summary, the new RTX graphics cards enhance the final phases of volumetric rendering and support the initial phase of machine learning for the "image based" acquisition of complex visual information, put into the system simultaneously to generate the rendering of a three-dimensional space.

2.1 Preparation of the architecture for Volume Rendering with Neural Radiance Fields

In order to experiment with the algorithm, it is necessary to prepare the environment in which to run the code in the hardware structure, which is based on a particular framework called Tiny CUDA, already partially used in some unbiased real time rendering engines, such as Otoy Octane Render. Using the Anaconda platform and the Python language, the algorithm begins a primary training phase, reading and interpreting all the images entered in the work directory, preparing the basis for the second phase, Volumetric Rendering, in which the " software”- directly from Anaconda through a specific ngp interface - which allows the processing of the 3d scene visually in parallel or perspective projection. Unlike traditional photogrammetry, the software does not generate a point cloud, but a NUBE voxel (volumetric pixels) and the scene appears as a solid volume composed of voxels that gradually thin out defining the various three-dimensional elements. Through some commands in the software interface, it is possible to optimize the final product through the bounding box and finally proceed with the export of a triangulated polygonal model. Based on the resolution and the number of polygons that can be selected directly from the "Instant NGP" interface, corresponding to the polygonal weight of the output mesh, the exported model can be quite fragmented. It is the product of a volumetric rendering that works by involving volumes and is often structured in irregular polygons (Voxel) not offering good results in defining objects. The dynamic vertex colour component gives the elements realism, bypassing the texturing or mapping phase and dynamically changing according to the framing factor during real-time exploration of the scene. The quality of the vertex colour will depend on the quantity of polygons that make up the mesh. Comparing this method with the photogrammetry technique, which allows the generation of realistic models that can be exported to any rendering or modelling platform, the big limitation of Nerf currently consists of the export phase of the model, which needs to be worked on and optimized later. The model is also devoid of textures generated directly from the photos used in the process, which is very often essential to configure a realistic model that can be used in other rendering platforms. A trick is to treat the model obtained on external software such as Z-brush sculpting, capable of managing hundreds of thousands of polygons and which has some plug-ins useful for mending gaps and correcting shapes, such as Z-Remesher, Dynamesh o Voxel Tessellation. Another experimental method is to export the Voxel model generated by the Anaconda platform to Cinema 4d and use the LAZPOINT 2 plugin capable of generating point clouds from surfaces with vertex colour (ignoring the internal density) from any polygon mesh model. Once the point cloud has been generated, it is possible to export it in the most common formats in photomodeling software and start the mesh generation process, usually based on the POISSON algorithm. In this case the result obtained is clean and the model can be used directly for rendering. Although these initial management problems, presumably resolvable in some time, in relation to the rendering and direct representation in the field of spatial architectural representation, the Instant-Nerf NGP method is much more precise in relation to objects made of reflective or metallic materials. In the experiments carried out, summarized in the next chapter, this system has offered discrete results even using photographs with different light exposure taken through generic shots (and therefore not structured in a precise photogrammetric survey plane) with obstacles that hide parts of the framed object or that they portray moving subjects, images affected by motion blur (photographs therefore unusable in traditional photomodeling processes). Two recent studies conducted by Google Research and the University of Washington demonstrate how it is possible to carry out a Nerf survey using simple selfies with your phone (which in photogrammetry would be impossible since the framed subject would move and the correspondence of the points would disappear between frames) or using existing photographs taken at different times and conditions (Lombardi et al. 2019)(Park et al. 2022) (Chen et al. 2021). This opens up very articulated research scenarios, which can involve the today impossible survey of works of art or architecture placed in inaccessible war zones, buildings that are now destroyed but photographed in the past and preserved in archives, graphic studies on temporal structural evolution of architectures or in a broader sense of entire cities only using historical and current photographic or aerial photography data (Martin-Brualla et al. 2021).

3. CASE STUDIES AND EXPERIMENTS

The investigation to highlight and explore the potential of the new low-cost systems for surveying focused on experimenting with some case studies on a variable scale, relating to some architectures and a more contained object of an artistic nature. The tools offered by instant nerf NGP offer interesting alternative methods to three-dimensional photomodeling with much lower costs than canonical "range based" methods. Surely, they can be included in the "image based" survey, today the most used low-cost system, in relation to which comparisons and reasoning on the differences and affinities between the two methods analysed have been made. After having installed the “neural graphics primitives” based on the neural radiance fields (NeRF), some photographic sets used in the past for architectural photomodeling work were taken.

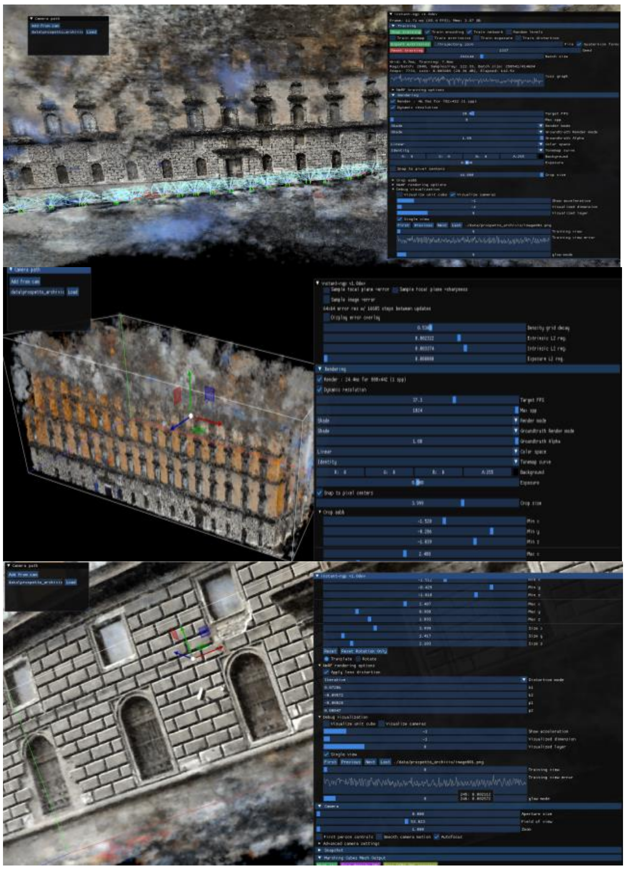


Figure 3. The Nerf survey of the façade of Santi Severino in Naples.

The first experiment relating to the survey of the façade located in Via del Grande Archivio of the former Monastero dei Santi Severino e Sossio in Naples, the current seat of the national historical archive, with a total length of about 100 meters did not give exceptional results. The model created with the volumetric rendering shows correct proportions, but the voxel cloud, in some cases very dense, tends to hide some parts of the building and to generate a confused and blurred model, especially in the upper parts, as the survey was carried out by the street level and not with the help of a drone. Furthermore, the quality of the images obtained does not reach satisfactory results when approaching details, such as windows or portals. Even the export in mesh model does not provide a quality comparable to what can be obtained with the canonical photomodeling methodology. Despite this, the proportional calculation of the building appears correct in its entirety. Each part has been aligned and intersected and the shape, albeit poorly defined in the upper parts, is closed, returning a complete model up to the edges. Exporting images in photoplan view is an excellent alternative for scaled vector redrawing. The "ngp" interface also allows a somewhat awkward but effective cropping to extrapolate only the facade. There is no metadata recognition capability for the georeferencing of the model, which appears to be translated and rotated: in this development phase of the Nerf system, it is not possible to implement georeferenced data that allow a correct initial positioning of the architectural model.

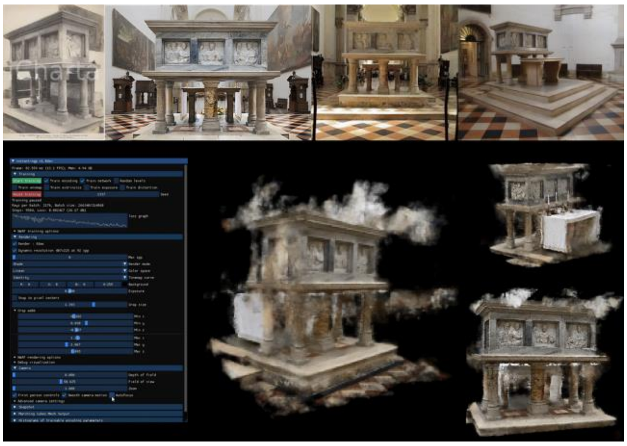


Figure 4. Volumetric rendering of Altare di San Mathias at the church of S._Giustina in Padua.

The case study of Altare di San Mathias at the church of Santa_Giustina in Padua represents a survey of a medium-sized element compared with closed, complex and articulated spaces typical of a large basilica. The images used for the calculation have been appropriately selected to halve the number compared to those used in the canonical “image based” calculation. The materials used in the construction of the altar, for the most part opaque, have some parts of glossy marble coating. Thanks to the diffused natural light (weak due to the size of the church) that filters through large windows, the luminous conditions of the object are sufficient to intervene by means of calibration to make the images usable without losing their quality. In this case the result is quite good, the volumetric cloud is concentrated at a sufficient distance from our object allowing an excellent extrapolation through cropping. The glossy elements thanks to the volumetric rendering do not lose their reflective characteristic according to the movement of the camera. The sculptural and bas-relief elements retain a sufficiently good quality for a mesh export that can also be used in visualization on other software. Despite the use of a smaller number of frames, each element generated by the volumetric rendering appears well proportioned, correctly structured without any overlapping errors of elements or gaps (Yu et al. 2021).

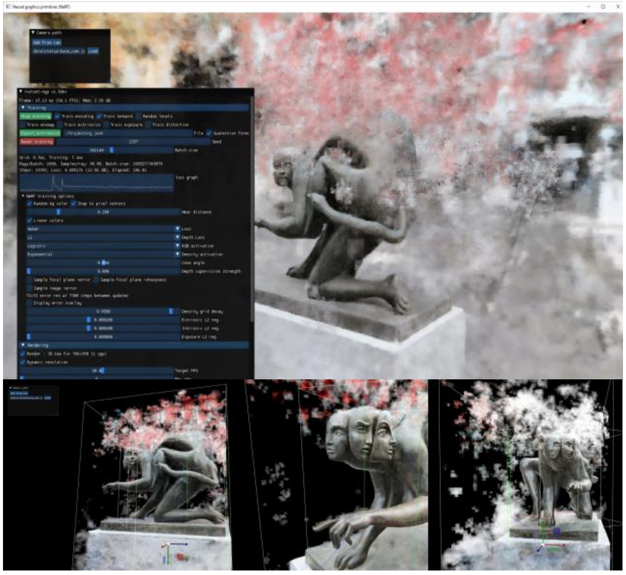


Figure 5. first experiments whit Ai calculi for the survey of the Gonzalez statue.

The case study relating to the smallest, but also more complicated object, focuses on the survey of the "Sfinge e Colomba" statue by Alba Gonzales using the Nerf calculation. The Roman sculptress is known for a style that tends to the stylization and fusion of forms of different nature and the use of different materials, including bronze. The work, located in Piazza Statuto in Pietrasanta, is in fact a bronze statue characterized by the harmony of the anthropomorphic forms and the shiny and reflective material.

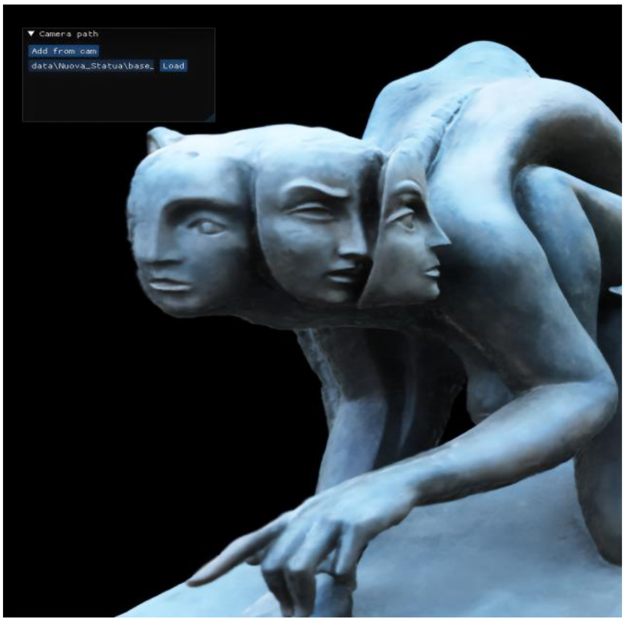


Figure 6. Figure placement and numbering

The light conditions encountered during the survey and the material characteristics, constitute the ideal opportunity for experimenting with this new technology. The results are satisfactory, thanks to the use of an RTX 3090ti graphics card. The main difficulty of the experiment is encountered during the output phase: being a model generated by a volumetric rendering, it is complex for the software to convert it into a mesh with the correct topology. The definition of the mesh can be chosen through a slider that establishes the number of polygons that compose it. This experimental method is in an embryonic phase both from the point of view of extrapolation of the data obtained and in relation to the interoperability of formats that can be used on other software. In conclusion, it can be said that this innovative method can possibly solve various problems related to the lighting conditions of the environment and the reflectance of the materials, reducing the conception time of the actual three-dimensional space compared to the canonical photogrammetric technique. The system, however, in the last year has had an exponential improvement, it is therefore desirable that soon it will replace canonical photomodeling, becoming soon stable and within everyone's reach.

In relation to the Nerf NGP interface supporting Anaconda, which at this stage of development of the tool is already comfortable and intuitive, soon it will certainly have a graphical improvement and many more visual processing and mesh export tools. Some of the most interesting functions are the cubic selection operation of the volumetric cloud (Crop Cloud), AI sharp features, based on the new neural systems DLSS 2, Deep Learning Super Sampling, which can increase the quality of the resolution, depth management tools field and saturation adjustment based on various colour spaces (Linear-Rgb, S-RGB).

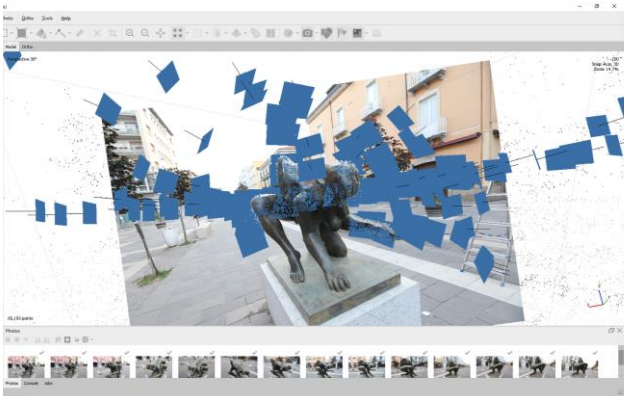


Figure 7. Photomodeling on Agisoft Metashape platform of the same case study.

Through some commands in the AI software interface in real time, it is also possible to optimize the final image through the bounding box and finally proceed to export the mesh with a specific function that generates a mesh voxel with colour vertex at various resolutions. By synthesizing the views and interrogating the 5D coordinates (x, y, z + angular (θ, φ)) along the camera beams, it is thus possible to use the classic volumetric rendering techniques to project colours and output densities. In volumetric rendering, based on "multi-resolution hash grid coding", which does not require complex machine learning operations on external servers and uses Nvidia's new and faster input encoding method, the processing does not stop at the data specific to the surfaces of the objects, but the capture rays that penetrate inside are also considered important. It differs from traditional ray tracing (rendering that concerns only surfaces) because, as already mentioned, it uses a single ray combined with a single frame to obtain information on light sources and their intensities; this method therefore does not need the rebounds of the secondary rays to reproduce the overall illumination of an environment (global illumination).

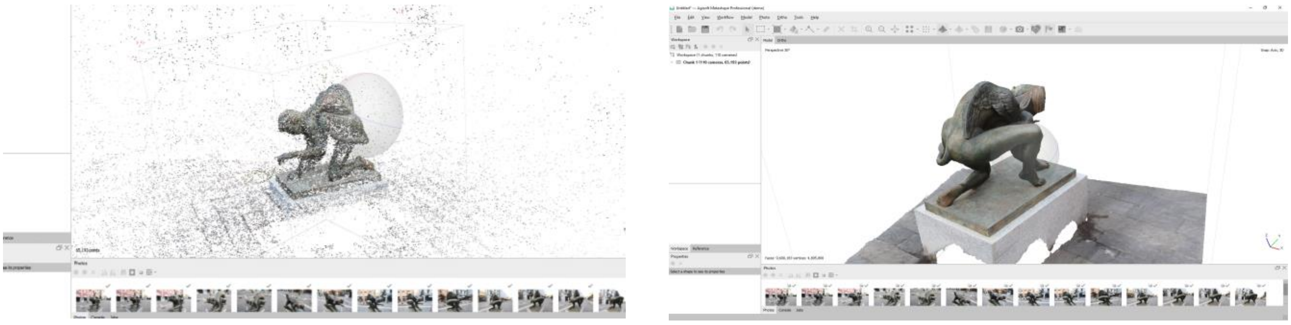
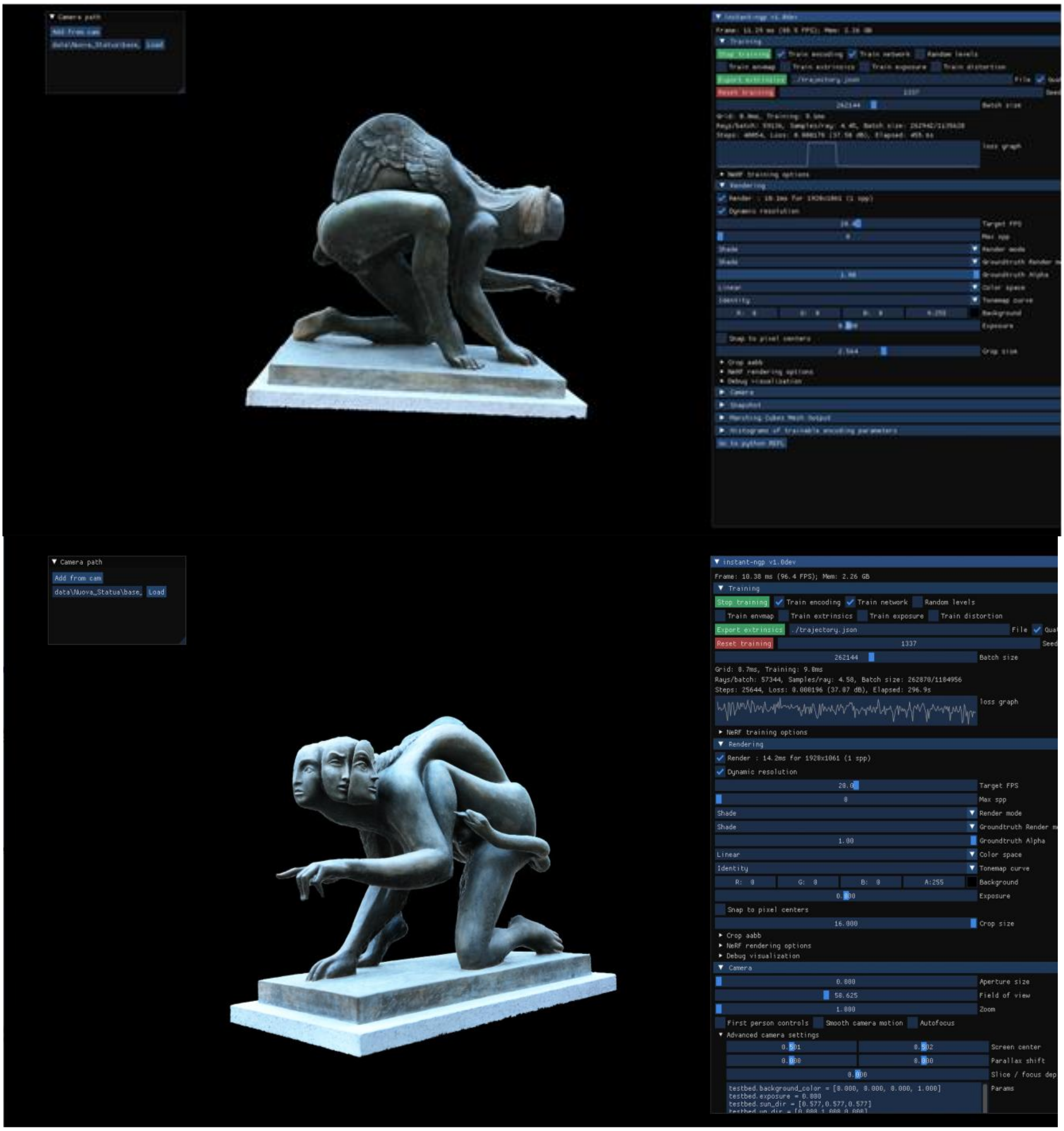


Figure 8. A).sparse point cloud and B) mesh model from photogrammetry



```
▼ instant-ngp v1.0dev
Frame: 10.38 ms (96.4 FPS); Mem: 2.26 GB
▼ Training
Stop training  ☑ Train encoding  ☑ Train network  □ Random levels
□ Train envmap  □ Train extrinsics  □ Train exposure  □ Train distortion
Export extrinsics  ./trajectory.json  File  ☑ Qua
Reset training  1337  Seed
262144  Batch size
Grid: 8.7ms, Training: 9.8ms
Rays/batch: 57344, Samples/ray: 4.58, Batch size: 262870/1184956
Steps: 25644, Loss: 0.000196 (37.07 dB), Elapsed: 296.9s
loss graph
► NeRF training options
▼ Rendering
☑ Render : 14.2ms for 1920x1061 (1 spp)
☑ Dynamic resolution
28.0  Target FPS
8  Max spp
Shade  ▼ Render mode
Shade  ▼ Groundtruth Render m
1.00  Groundtruth Alpha
Linear  ▼ Color space
Identity  ▼ Tonemap curve
R: 0  G: 0  B: 0  A:255  Background
0.000  Exposure
□ Snap to pixel centers
16.000  Crop size
► Crop aabb
► NeRF rendering options
► Debug visualization
▼ Camera
0.000  Aperture size
50.625  Field of view
1.000  Zoom
□ First person controls  □ Smooth camera motion  □ Autofocus
▼ Advanced camera settings
0.501  0.502  Screen center
0.000  0.000  Parallax shift
0.000  Slice / focus dep
testbed.background_color = [0.000, 0.000, 0.000, 1.000]  Params
testbed.exposure = 0.000
testbed.sun_dir = [0.577,0.577,0.577]
testbed.up_dir = [0.000,1.000,0.000]
```

Figure 9. ngp interface and volumetric rendering model of the bronze statue in real time view (prospective visualization).

Unfortunately, if the input images are not well structured and clean (in this case our goal was the representation of the statue), an excess of volumetric fog can be generated which tends to disturb the display and even more the export of the model. A bit like what happens in photomodeling and point cloud laser scanners. Escamotage to improve the cleaning of the Nerf cloud consists in treating the reference photos with a masking (with manual Photoshop techniques or automatic AI) of the element we are interested in rendering, isolating it from the context. In this case the images with PNG transparency allow a much more detailed and precise rendering and a considerably more precise model. Unlike traditional 3d photogrammetry, the software does not generate a point cloud, but a cloud of voxels (volumetric pixels) and the scene is presented as a solid volume composed of voxels. In fact, if the techniques based on photogrammetry are compared, Nerf still does not produce excellent results and it is necessary to use other programs such as Zbrush. On the other hand, traditional photogrammetry fails to capture mirrors, glossy black surfaces, reflective metal surfaces and there are other limitations, (even the computation time of 3d photogrammetry is longer than volumetric rendering). Furthermore, not using global illumination but directly the reference inputs entered in the AI calculation, photographic references that exactly reproduce the behaviour of diffused, reflected, shadow-generating light, etc., the Nerf model will be extremely more precise in reproducing such luminous phenomena, such as refraction, reflection, the exact behaviour of the glossy surfaces to the movement of the camera (precisely because it has acquired this information directly from reality and the artificial intelligence interprets it based on the amount of information that we can enter into the calculation to obtain a more precise model than to that generated by photogrammetric procedures). In traditional photomodeling the colour, then transformed into texture, does not store the metadata relating to the type of material, in this case bronze with its reflection coefficient, which is possible through the AI volumetric rendering. The model of the statue, despite being well made and with few gaps, therefore has a colour that does not react in a realistic way, resulting in opaque / Lambertian. The methodology that uses the Structure for Motion algorithm, to achieve good results, needs more than double the photographs given compared to the input images used in the Nerf, following the rule of frame overlap of about 80%. The three experimental case studies make us reflect on the potential of this new "image Based" method, faster and more precise but still under development, in dealing with elements that can range from architectural to smaller scales. Excellent for managing indoor and outdoor environments with complex or insufficient lighting or consisting of geometrically articulated and unmanageable elements.

4. CONCLUSIONS

In conclusion, it can be said that this innovative method can possibly solve various problems of the photogrammetric imaged based survey related to the light conditions of the environment and the reflectance of the materials, reducing the conception time of the actual three-dimensional space compared to the canonical photogrammetric technique. This experimental method is still in an initial state from the point of view of extrapolating the data obtained on other rendering applications and in relation to the interoperability of formats that can be used on other software. Despite these initial management problems, presumably resolvable in some time, in relation to rendering and direct representation, this method is much more accurate in traditional survey based on images in relation to reflective or metallic objects. Furthermore, in the last year the system has had an exponential improvement, also thanks to the power of the video cards on the market, it is therefore desirable that in the future it will replace photomodeling as we know it today, soon becoming stable and within everyone's reach. The software, even though it already has an intuitive interface and allowing easy navigation within the 3D scene, still presents a difficult installation (if it can be defined as such) since there is no real setup file and a relative executable, but the software must be started through code (using Anaconda and the Python language). This obviously creates a big limitation, because it is necessary to have an advanced knowledge of the machine on which to start the software or even ask for the support of an IT.

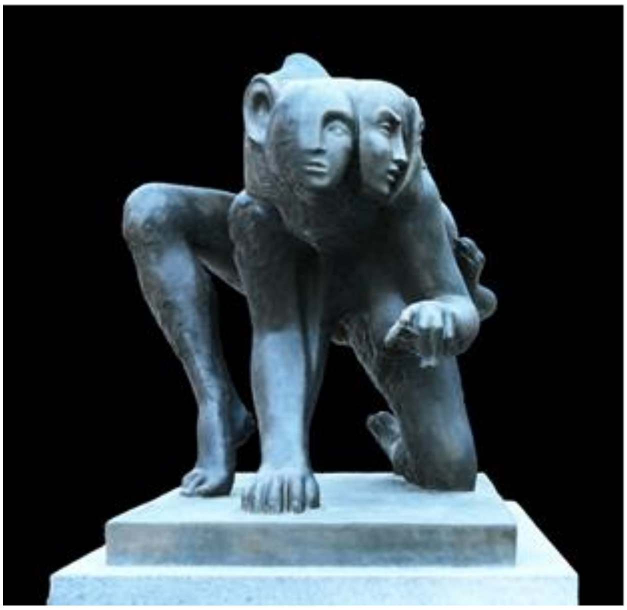


Figure 10. Neural volumetric raytracing of model of the statue of Alba Gonzalez, front elevation.

REFERENCES

Chen, Anpei, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. 2021. "MVSNeRF." *Iccv*, 14124–33. http://arxiv.org/abs/2103.15595.

Lombardi, Stephen, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. 2019. "Neural Volumes: Learning Dynamic Renderable Volumes from Images." *ACM Transactions on Graphics* 38 (4). https://doi.org/10.1145/3306346.3323020.

Martin-Brualla, Ricardo, Noha Radwan, Mehdi S.M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. 2021. "NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections." *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* 7206–15. https://doi.org/10.1109/CVPR46437.2021.00713.

Mildenhall, Ben, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis." *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12346 LNCS: 405–21. https://doi.org/10.1007/978-3-030-58452-8_24.

Park, Keunhong, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. 2022. "Nerfies: Deformable Neural Radiance Fields," 5845–54. https://doi.org/10.1109/iccv48922.2021.00581.

Tancik, Matthew, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. 2020. "Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains." *Advances in Neural Information Processing Systems* 2020-December: 1–24.

Xie, Yiheng, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. 2021. "Neural Fields in Visual Computing and Beyond" 41 (2). https://doi.org/10.1111/cgf.14505.

Yu, Alex, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. 2021. "PixelNeRF: Neural Radiance Fields from One or Few Images." *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 4576–85. https://doi.org/10.1109/CVPR46437.2021.00455.