VISOR: Virtualization Service for Object Reconstruction

Iordanis Evangelou, Anastasios Gkaravelis, Nick Vitsas, Andreas A. Vasilakis

Phasmatic

Corresponding author: iordanis@phasmatic.com

Keywords: 3D reconstruction, neural networks, digital twin.

Introduction

Recent advances in deep learning, particularly in 3D geometric learning, combined with the growing availability of specialized hardware, have led to the development of powerful new methodologies, algorithms, and tools for 3D content creation. These innovations enable automatic generation, manipulation, and interpretation of complex spatial data at a computational cost that is manageable with today's commodity hardware. This work focused on a self-supervised training approach that leverages input data such as 2D images or other task-specific modalities captured from the physical world. The output is a digital twin in the form of a 3D triangular mesh, ready for real-time use in any conventional visualization framework (e.g., web, mobile, desktop). This type of 3D geometric learning algorithms has the potential to automate the creation of digital replicas of real-world objects. Consequently, these technologies can greatly accelerate the traditionally labor-intensive process of generating accurate virtual representations, streamlining workflows and enabling more efficient and high-fidelity novel content creation. However, architecting a reliable end-to-end workflow is a non-trivial task since one needs to strike balance between several key factors. These include output quality in terms of surface structure and associate properties (e.g., color), computation time and resources to prepare the input stream and generate the final mesh as well as transform the content of the output mesh such as quality is maximized while size is minimal to be transmitted and loaded to any conventional device.

This work introduces VISOR, a virtualization service for object reconstruction which builds upon the latest developments in the literature Neural Radiance Fields [1] (NeRFs) and stands as a complete pipeline for generating 3D content from real-world physical objects. The proposed pipeline integrates and extends state-of-the-art techniques at the intersection of computer vision, computer graphics, and neural optimization and it is structured into three distinct phases: pre-processing, optimization, and post-processing, each of which is discussed in detail below. Finally, we present empirical results on a diverse set of objects to demonstrate the effectiveness of the pipeline and discuss its current limitations along with potential future directions for improvement.

Methodology

VISOR is designed for the reconstruction of single entities from the physical world that can be captured by conventional recording devices such as a smartphone. In this section, we outline the key methodologies and design considerations required to construct such a workflow to transform a sequence of images, or a video stream, of a physical object into a 3D triangular mesh. Briefly, the VISOR pipeline can be described by three main stages: pre-processing, optimization, and post-processing, each designed to efficiently extract and refine spatial and appearance information from unstructured visual input.

Pre-processing. VISOR receives input from a finite stream of images or a video sequence capturing the target physical object. There are no restrictions on the type of images or the video encoding as well as the capturing resolution. To generate a plausible result, the input set of images can be as low as 50 unique and overlapping photos while the video sequence can be at least 30 seconds long. Since VISOR optimizes mesh structure from raw images, if a video stream is supplied instead, we automatically convert it to a finite image set by adaptively sampling the sharpest consecutive frames. To proceed with the NeRF optimization stage, camera calibration to extract both intrinsic and extrinsic camera parameters is required since these per-image parameters are typically unknown but implicitly defined by the user's capture device. VISOR recovers those parameters by applying the GLOMAP [2] calibration toolbox which balances between speed and accuracy. Finally, we isolate the object interest from the background for each image by applying a general-purpose dichotomous image segmentation neural network [3]. Each generated tuple of segmentation mask and viewport transformation is forwarded to the optimization stage.

Optimization. VISOR's reconstruction methodology follows the signed distance field (SDF) formulation for which the reference 3D surface lies at the zero-level set. SDF essentially defines a real function that maps 3D coordinates to real values indicating their closest distance from the reference surface. In the special case of the zero-level set, these distances must be equal to zero. This paradigm is a common modelling approach in object reconstruction due to its continuous, flexible and mathematically grounded form. The task of the VISOR service is to learn and recover this function from the input stream of images. VISOR optimizes the SDF by following the current state-of-the-art paradigm based on NeRF theory [1]. Briefly, it leverages volume rendering designed for novel view synthesis tasks reparameterized in a way that allows for SDF optimization at the same time. As is typical in this optimization regime, ray batches are iteratively generated from the calibrated viewports of the input images. These rays are then sampled across their span in a predefined granularity and the volume rendering equation is evaluated in a per-pixel basis. VISOR follows practices from the most recent literature [4], which demonstrates how opaque objects can be formulated as volumetric entities by adopting stochastic geometry theory. Then, by reparameterizing the volume rendering equation the novel synthesis task can be reformulated as an SDF optimization problem, effectively retrieving the zero level-set of the reference object. VISOR follows this paradigm as its core optimization methodology. To make this approach tractable in terms of resources and computation time, VISOR is implemented based on state-of-the-art hardware accelerated neural encoding modules [5] as well as adopting specific design choices for the underlying neural-based modules for this task [6] to further accelerate the training phase. As an additional optimization

and to further accelerate convergence, VISOR incorporates loss-driven adaptive ray generation. This scheme effectively decouples the input image resolution from the actual physical object which is only depicted on a small portion of it.

Post-processing. In the final phase, VISOR processes the signed distance values sampled from the optimized field to generate an initial triangular 3D mesh. An isosurface extraction algorithm (in our case, Marching Cubes [7]) is used to construct the polygonal surface. Notably, the chosen resolution to discretize space directly influences both quality and file size. Since we aim to extract a triangular surface that closely resembles the optimized SDF we prioritize quality over size in the first step. Then, the surface is appropriately simplified to reduce primitive redundancies [8] to drastically reduce the initial triangle count. To also embed surface properties such as color information we apply UV-unwrapping [9], allowing for texture baking. Since the simplified geometry includes new points, we need to resample the radiance in the hemisphere around the local normal vector using the already trained model to estimate the per-point appearance. Surface colors are encoded into the texture map we have already created and enable realistic shading of the mesh. The completed 3D model is then exported in a standard file format (e.g., FBX or GLB), making it ready for real-time rendering in 3D game engines such as Unity or Unreal Engine.

Results

In this section we provide some experimental results demonstrating the efficiency of the VISOR service. The full workflow of VISOR can be executed in under 10 minutes on commodity GPU hardware, such as NVIDIA RTX 4090, for a single object instance. In Figure 1, we depict some indicative outputs from our service. For each example case, we include the size of the reference object in physical units, the total time required to get the output 3D mesh along with the final file size and triangle count of the actual content.

Conclusion & Future Work

The VISOR pipeline offers an end-to-end solution for 3D reconstruction from a set of input images or a video stream. While each module is designed with research-driven methodologies, certain limitations remain, primarily due to the quality of input data and implementation-specific assumptions embedded within intermediate stages. For instance, early-stage errors, such as motion blur or insufficient coverage of the target object, can propagate through the pipeline and significantly degrade the fidelity of the final reconstruction. To support accessibility and ease of use, VISOR is deployed as a web-based service, allowing users to upload videos or image collections and receive a reconstructed 3D asset in return. The planned future updates aim to align the pipeline more closely with the latest research. Improvements are anticipated in the calibration stage, enhancing robustness to challenging capture conditions, and in surface appearance modeling to better approximate physically accurate rendering. However, the most critical area of ongoing development is the optimization phase, with a focus on improving the converged SDF quality and reducing the time required to achieve high-quality reconstructions.

Acknowledgements. This work was funded by the European Union under the CORTEX2 project (Grant Agreement no. 01070192).

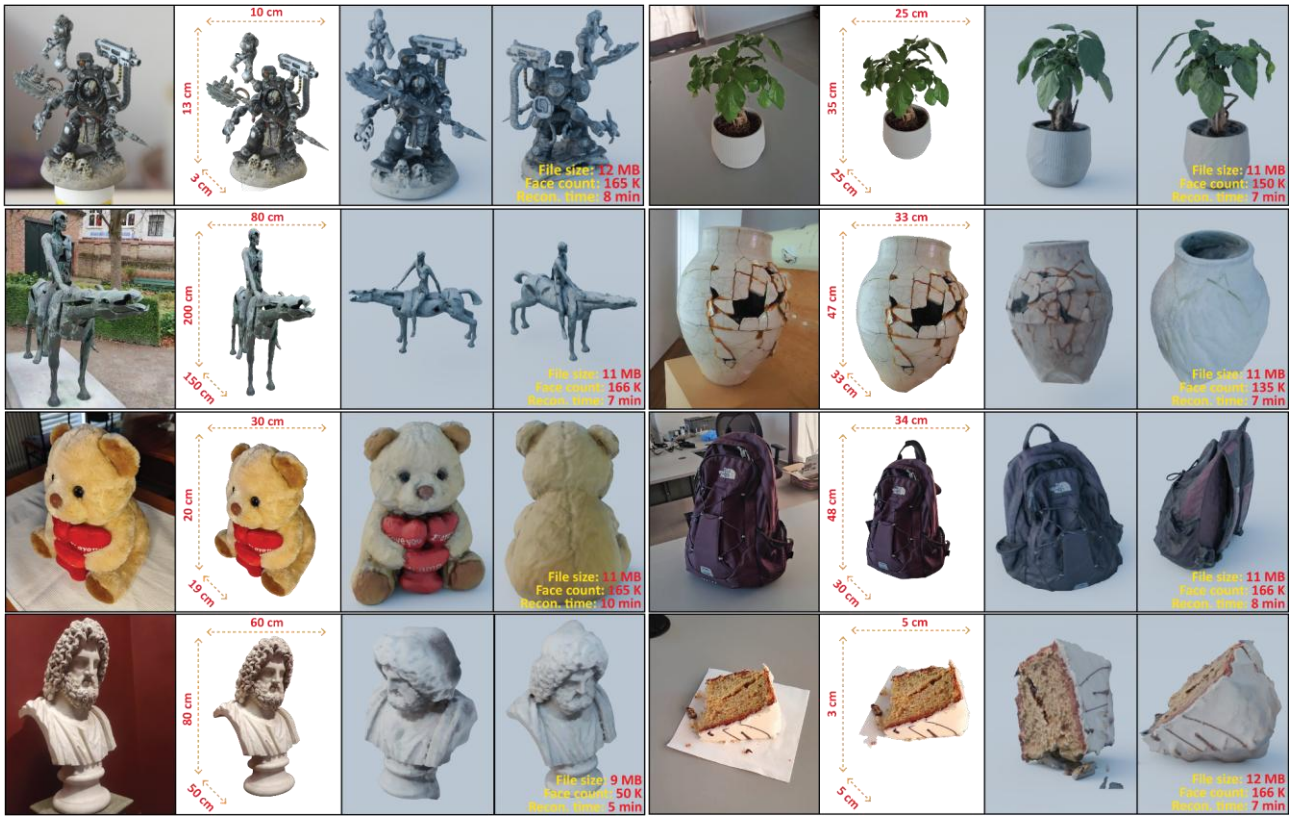

Figure 1. Indicative outputs from the VISOR service.

References

[1] Mildenhall, Ben, et al. "Nerf: Representing scenes as neural radiance fields for view synthesis." Communications of the ACM 65.1 (2021): 99-106.

[2] Pan, Linfei, et al. "Global structure-from-motion revisited." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.

[3] Yu, Qian, et al. "Multi-view aggregation network for dichotomous image segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

[4] Miller, Bailey, et al. "Objects as volumes: A stochastic geometry view of opaque solids." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

[5] Müller, Thomas, et al. "Instant neural graphics primitives with a multiresolution hash encoding." ACM transactions on graphics (TOG) 41.4 (2022): 1-15.

[6] Wang, Yiming, et al. "Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

[7] Lorensen, William E., and Harvey E. Cline. "Marching cubes: A high resolution 3D surface construction algorithm." Seminal graphics: pioneering efforts that shaped the field. 1998. 347-353.

[8] Kapoulkine, Arseny. "Mesh Optimizer", Available: https://github.com/zeux/meshoptimizer

[9] Young, Jonathan. "Xatlas", Available: https://github.com/jpcy/xatlas