ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DEPARTMENT OF COMPUTER SCIENCE
AND ENGINEERING

ARTIFICIAL INTELLIGENCE

MASTER THESIS

in
Computer Vision

3D GAUSSIAN SPLATTING RECONSTRUCTION WITH DEPTH ENHANCED INITIALIZATION

CANDIDATE:
Jacopo Meglioraldi

SUPERVISOR:
Prof. Gustavo Marfia

CO-SUPERVISOR:
Pasquale Cascarano

Academic Year 2023/2024

Session 3rd

Abstract

This thesis proposes a novel pipeline incorporating segmentation masking and depth values to enhance the performance of Gaussian Splatting techniques. By leveraging 3D Gaussian Splatting's ability to achieve high-accuracy photorealistic reconstructions, the pipeline focuses on singular object reconstruction through segmentation masking, which removes unwanted backgrounds and accelerates the optimization process. Integrating recent advances in fast semantic segmentation using neural networks, the pipeline produces nearly ready-to-use models. Additionally, using a depth-sensing camera during acquisition allows for more accurate point cloud initialization with minimal overhead, leading to a significant boost in visual accuracy during the early optimization steps. The pipeline achieves a significant speedup, making it particularly advantageous for devices with limited computational capacity, though with a slight trade-off in the accuracy of the final results. The depth-enhanced initialization is carried out by sampling and projecting meaningful point information into the reconstructed space, offering a better approximation of under-sampled regions. This step also ensures the reconstruction is to scale, enabling precise measurements and further analysis.

Acknowledgements

I want to thank and express my sincere gratitude to my supervisor Prof. Gustavo Marfia and my co-supervisor Pasquale Cascarano. They made this project possible and helped me during the design, realization, and compilation of this work. Their expertise and skills in the research area have been a fertile ground where I grew in my professional skills and knowledge. In conjunction, I want to thank all the staff of the VARLab laboratory that created a passionate and dedicated environment in which people can express their capability and learn with mutual respect. In particular, I want to thank Giacomo Vallesciani, doctoral researcher and desk mate, together with the above-mentioned Pasquale Cascarano, for their laugh and chats during work breaks and for their support in the moments of discomfort. Some pastries and a joke is the recipe for a good friendship and a welcoming workplace.

The presented work is not only the results of the last months but of a long university course. A major thanks go to the Scouting Association of Albinea and its tireless members. Entering the professional world is not only a matter of knowledge acquisition and useful skills. It is also a matter of relations, organization, and compromises. I could not describe the impact and experience that volunteering in the scout association had on my formation as a student, but more importantly as a citizen and a human being. I want to specifically thank my division mates of the last years: Caterina, Fabio, Paolo, Elena, Francesco L., Francesco N., Lorenzo e Letizia. They supported me during my studies and were my lucky stars on the evenings before the exams. They taught me that the best way to change the world around you is to act and that taking care of the person next to you makes this place better.

I want to thank Francesca Bonacini, the psychotherapist who accompanied me last year helped me connect with my emotions, and supported me during my studies. Without her reaching the graduation would have been a climb too difficult for my strength alone.

Finally, but not least for importance, I want to thank my family: Barbara, Stefano, Riccardo M., Alice, and my long-lasting friends: Cristiano, Marialuisa, Luca, Danny, Gabriele Lon, Alessio, Sebastiano, Riccardo F., Mattia, Gabriele e Martina. They will never stop to accompany me in all my failures and all my achievements.

Contents

List of Abbreviations

2DGS 2D Gaussian Splatting.

3DGS 3D Gaussian Splatting.

GPU Graphics Processing Unit.

KDE Kernel Density Estimation.

LiDAR Light Detection and Ranging.

LPIPS Learned Perceptual Image Patch Similarity.

NeRF Neural Radiance Field.

NN Neural Network.

PDF Probabiliy Density Function.

PSNR Peak Signal-to-Noise Ratio.

RAM Random Access Memory.

RANSAC RANdom SAmple Consensus.

RGB Red Green Blue.

RGBD Red Green Blue Depth.

SDF Signed Distance Function.

SDS Score Distillation Sampling.

SfM Structure from Motion.

SH Spherical Harmonics.

SIFT Scale-invariant transform function.

SLAM Simultaneous localization and mapping.

SPAD Single Photon Avalanche Diode.

SSIM Structural Similarity Index Measure.

SuGaR Surface-Aligned Gaussian Splatting.

ToF Time of Flight.

List of Tables

List of Figures

Chapter 1

Introduction

In this work, it is presented a small but significant contribution to the 3D reconstruction research area. The demand for accurate 3D representation of real existing objects and scenes is increasing year by year, for all kinds of applications. Artificial intelligence (AI) breakthroughs in recent years revealed a digital world hungry of data. Democratizing automatic 3D reconstruction of real-life scenes will create a base for new techniques to build upon it. Nowadays reality is divided into two distinct layers that can interact with each other: a physical layer ruled by physical and mathematical laws and composed by matter in opposition with the artificial digital layer simulated in computers and represented with mathematical theory. Accessing the 3D representation of real-world objects allows analysis, synthesis and manipulation of their physical counterparts with high accuracy and efficiency. Sectors that benefit from it are the robotics industry for automatic manipulation and navigation, the entertainment industry for immersive and engaging content, the medical fields for learning and personalized health treatments.

In robotics arms manipulation tasks, developing a robust and adaptive sequence of movements for multiple objects requires the 3D reconstruction of the target object as well as the robot arm. Key information are the rotation and pose of the object in space with its geometric properties and shapes, to correctly simulate contact points and grasping conditions. With the advent of AI applications, a dataset of 3D objects to be manipulated is used for training and simulating the manipulation process [5]. In robotic navigation, the ability to dynamically reconstruct the space navigated and to detect the position of obstacles enable the robot to move across the environment with great accuracy and efficiency. In environments where space is shared with human operators, it also ensures a safer operation of the robots [37]. Scanning and digitalizing objects of relevant cultural values enable a better preservation of the real artworks. Additionally, the new representations led to advancements in the analysis techniques where sections or inside view visualization were impossible without damaging the exhibits [11]. Museums are facing a new challenging shift, integrating new technologies and 3D representation of objects in their exhibitions. XR applications, which often require accurate and immersive 3D models of the displayed objects, are integrated with the classical museum experience to increase engagement and innovate the storytelling of cultural heritage [14]. In the entertainment industry, 3D representation of objects is at the core of computer graphics history. Modern movies' special effects are indistinguishable from reality and handcrafted virtual 3D models are used side by side with real actors. Modern games aim for better and more realistic 3D representation pushing the research for more efficient and dynamic 3D displaying techniques. In medical fields, general 3D reconstruction of organs is widely used in training and for research purposes. Patient-specific reconstructions, such as x-ray reconstruction, are an important aid for accurate pathology identification, medical treatments and surgery planning [30]

Nowadays, obtaining a representation of real or imaginary objects is expensive and time-consuming. Manual 3D modeling and designing is still the best choice for a lot of applications, with the use of Computer-Aided Design (CAD) softwares or Digital Content Creation (DCC) softwares, but automatic reconstruction is rapidly gaining importance. New technologies and hardware innovation permit a faster and more accurate reconstruction of 3D objects starting from simple sensory information such as images, point clouds and laser reflection signals. Based on how the 3D object is represented, reconstructions can be divided into explicit or implicit representations:

- The main methods for displaying data include point clouds, voxels, and meshes, which are explicit representations. Explicit refers to a representation method that clearly defines geometric shapes and structures to directly describe the external or internal geometry of an object. Point clouds are an unordered collection of points in 3D space. Voxels are an unordered collection of unit cubes, called voxel, each storing spatial information such as position, color and density. Meshes are a structured representation with connected points, edges and faces that can be used to represent most topological characteristics. Explicit representation allows simple and intuitive manipulation, editing and extension to the different applications, but they are discrete approximations of the object characteristics, not suitable for complex and accurate simulations.

- Common implicit representations include implicit surfaces, Signed Distance Function (SDF), Occupancy Field and Radiance Field. Opposite to explicit methods, implicit expression of scenes does not rely on explicit storage of geometric data, but instead on functions and other mathematical primitives to store information. Implicit surfaces represent geometry by union of a set of equations describing surfaces such as Bézier curves. Simple but exact surfaces can be reconstructed in this way, suitable for accuracy-dependent applications. The Signed Distance Function is a continuous function assigning at each input 3D point in space the corresponding distance to the nearest surface, with negative values for points inside the object volume and positive otherwise. In Occupancy Field, points in space are mapped with probability value expressing their probability to be representing part of the object volume, usually accumulated in a 3D grid subdivision of space. Finally, Radiance Fields are representations that map points along a set of rays into a hyperspace describing light rays by intensity, wavelength and direction. The functions compute the color and visibility of all the points from the desired view. The major advantages of implicit representations are the lower requirement of storage space and the higher geometric expression capacity. In contrast, they require time-consuming computations and challenging manipulation operations.

Automatic reconstruction methods rely on simple sensory information to reconstruct an accurate 3D representation of the object. Opposite to hand-crafted 3D representation, sensory data are often noisy and a substantial pre-processing of input data is needed to extract clean and useful information. Smoothing filters over sensory signals can help improve the robustness of the reconstruction methods as well as missing values filling techniques. Segmentation and region of interest selection help remove unwanted background and avoid useless computation during reconstruction. The 2D to 3D reconstruction problem addressed in this thesis aims to use planar images to infer a 3D representation of a scene.

In 2D to 3D reconstruction task, one common problem with most acquisition techniques is the registration of the 2D images in a shared 3D space. Optimization algorithms like Global Bundle Adjustment and Pose Graph are used to globally optimize image pairs registration poses which are easier to compute but suffer a lot of limitations in the multi-way registration, such as drifts.

The presented work explores new cutting-edge techniques to reconstruct 3D information of a single object from an unordered set of 2D views. It expands the Gaussian splatting techniques [15], an implicit representation, by integrating its pipeline with additional depth information in order to reconstruct a better initial candidate. This work proposes a complete pipeline from unprocessed RGBD camera images of an object in the wild to the final mesh reconstruction. Objects and relative depth information are extracted from the images with segmentation. Images and point clouds registration algorithms are applied to sort in space the unstructured input images and extract a candidate point cloud. The 3D Gaussian splatting technique is used to reconstruct an accurate 3D representation of the object. An optional mesh extraction step is used for aligning the 3D object reconstructed with most used software solutions for 3D object applications. Results show that by leveraging depth information of the image acquired, the initial set of Gaussians can be refined to better approximate the final geometrical information in the initial representation with low computational overhead. A better initialization speed up the Gaussian splatting optimization technique while retaining the same quality in the early hundreds steps reconstruction, reducing the request for parallelized computations and hardware requirements. By pre-processing the data with segmentation masking extracted by fast neural networks, the foreground is separated in the initial stages, reducing the computational cost. It additionally avoids the manual or automatic step to extract the region by outputting an almost ready-to-use 3D model.

The work done is reported as follows: in **Chapter 2** the main relevant techniques addressed in this work are presented with their characteristics and limitations. In **Chapter 3** the proposed method pipeline and the preliminary studies supporting design choices are reported. In **Chapter 4** experimental setups and parameters are reported with a discussion of the obtained Results.

Chapter 2

3D Reconstruction Methods Overview

The 2D to 3D reconstruction is a well-known problem in literature for its importance and its intrinsic complexity. The objective of the task is to reconstruct a 3D representation from a set of 2D input information, usually RGB images. Acquiring images is the physical process of recording a discretized snapshot of a real scene. Images can be recorded from different sensors, the most common one is the three-channel visible light sensor cameras to record RGB pictures. Other cameras can record infrared, x-ray or electromagnetic signals. Acquiring images can be described as the mathematical process of projecting a space into a lower-dimensional one. The loss of information, in particular volumetric and spatial information, in this process is not negligible and defines the inverse problem of reconstructing the initial scene to be ill-posed. Research explored different solutions to this problem that can be grouped into 3 main categories.

- **Traditional Methods** try to reconstruct the geometric information of the scene by capturing multiple 2D representations of the scene and projecting pixel points from different views into the same space. Position of the 2D images in the shared space is fundamental to correctly match the projections and reconstruct the volumetric information thus camera position and parameters are required. In some methods known fixed camera poses are used in acquisition, while in others the camera positions are derived indirectly from the set of images. Usually, an optimization cycle to reduce the reprojection error is performed over the candidate points and camera positions are adjusted with respect the initial estimations. The main representatives of this group are classical Photogrammetry, Structure From Motion (SfM), Structured light and Light Detection And Ranging (LiDAR) . This methods will be described in detail in **Section 2.1**

- **Neural optimization methods** are newly introduced methods thanks to the advances in neural network computing and hardware parallelization. Neural networks are a computational model suitable for optimizing challenging functions that present high-order non-linear components with high dimensionality. These new methods show great accuracy and photorealism in the reconstruction of objects, gaining increasing attention by the research community. They are currently one of the hot topics of the computer vision research. In particular, NeRF methods are the main contributor in this family, followed by the more recent Gaussian Splatting methods. The full capabilities of these new models are yet to be discovered and in this work, we add a small but significant analysis of the method. They are be described in detail in **Section 2.2**

- **Prior knowledge Neural Networks methods** rely on the development of NN models trained on datasets containing 3D representations of objects and their images. These methods show great capabilities in the inference of shapes and information of the scene from few images, similarly to humans, based on previously seen data and prior knowledge. Although they can generate a 3D representation of the desired target even with a single image, the quality has yet to be challenging with respect to previous methods. Reconstructions are suitable for gaming industry applications where details of object in fictional worlds can be overlooked in favor of the overall gameplay experience. It is infeasible, though, for applications with high-quality reconstruction such as art and mechanical engineering. The two main limitations of these methods are the availability of good 3D representation of real-world objects leading to work mainly with synthetic data, and the amount of resources necessary to train such models. Nonetheless previous advancements in similar fields, in particular from the 2D image generation problem, showed the great potential of the architectures and pose a great interest for the research in this area. Example of models are DreamFusion, DreamGaussian, Magic3D and SDFusion [25, 36, 17, 6]

Reconstruction techniques could be further divided into *static* or *dynamic* reconstruction based on the type of scene inspected, creating a huge difference in the requested speed and accuracy of the method. In robotic navigation, we need constant updates of the geometry reconstructed in a continuous optimization cycle while the robot navigates new areas. The focus is on the spatial positioning of obstacles and targets while fine details and visual accuracy are of secondary importance. Usually, the representation has to be reconstructed in an online fashion thus requiring techniques that can reconstruct the scene with great speed. Simultaneous localization and mapping (SLAM) is the problem of reconstructing a space while keeping track of the position of the agent inside it. Usually, agents use multiple sensors such as camera, structured light, laser sensor or acoustic sensors and rely on variants of traditional reconstruction techniques. Static reconstruction aims to represent fixed scenes, like objects or places, and is divided into the acquisition phase and the subsequent offline computation phase to extract the reconstruction.

Starting from raw data, the reconstruction method from 2D images needs to solve sub-problems which can be grouped in the following three steps:

- **Pre-processing.** In this step raw data are processed to extract useful and meaningful information, removing noise and superfluous data. Segmentation removes unwanted additional data by selecting regions of interest from the raw image. Smoothing and filtering techniques remove unwanted noise coming from the acquisition process.

- **Extraction of 3D information.** Depending on the technique and sensor used, 3D information, usually point clouds, is extracted from single images or pairs. Multiple 3D representations are extracted from the data in an unstructured or semi-structured format. In other words, each image or pair of images contributes to the 3D reconstruction by spatial information extracted from the view. Different techniques have pros that make them suitable for specific applications. In general, the objective is to estimate the depth values of each point in the picture from the sensing camera. Naive aggregation of this information, such as subsequently linked chain of transformations, often incorrectly reconstructs the 3D scene, thus the third subproblem exists.

- **Global registration.** To accurately intersect the different spatial information from different views, an optimization step is often performed to coherently adjust the representation and reduce the error from the multiple views.

In this chapter all the theoretical foundations and literature works necessary to an organic understanding of the proposed work are reported. First, the different techniques and sensory acquisition methods to extract spatial information from 2D data are analyzed according to the above-mentioned division in traditional, neural optimization and learned prior methods. Then, the undistortion and segmentation pre-processing steps are analyzed. Finally, the widespread global registration methods of Global Bundle Adjustment and Pose graph algorithm are analyzed.

2.1 Traditional reconstruction methods

In this section, traditional 3D reconstruction techniques used to derive spatial information from images and sensor data are described. Starting with classical Photogrammetry, the section focuses on other approaches like Structure from Motion, structured light reconstruction, and LiDAR, highlighting their principles, advantages, and limitations across various applications.

2.1.1 Classical Photogrammetry

Photogrammetry is a technique that uses the relative position of known calibrated cameras to compute the positions of the points of the scene by multi-image triangulation. The most common model is the stereo camera model in which two close cameras capture simultaneous pictures of the scene from two different viewpoints. An extension of this technique is used for multiple views of the same object by multiple cameras [38]. This technique heavily relies on the good estimation of camera positions and intrinsic camera parameters during calibration. A colored image is a 2D projection of color points from the 3D space. The simplest mathematical model of a camera is the well-known *pinhole camera* where any point $X = (x, y, z)$ in 3D space coordinates is transformed into the pixel space point $U = (u, v)$ according to Equation 2.1 and represented in Figure 2.1. The model takes into account the focal length f of the camera and the discretization parameters due to real pixel dimensions ρ_u and ρ_v.

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \begin{bmatrix} \frac{f}{\rho_u} & 0 & u_0 & 0 \\ 0 & \frac{f}{\rho_v} & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \multicolumn{3}{c}{R} & T \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} \tag{2.1}$$

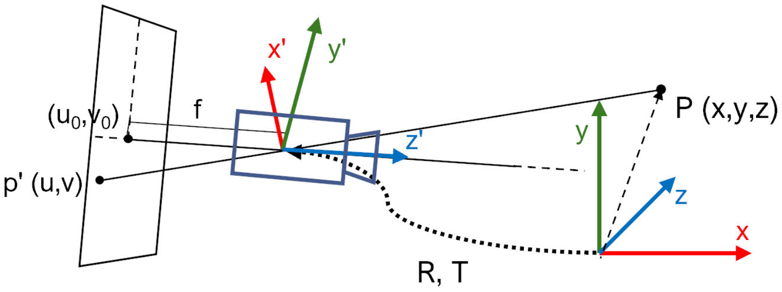

Figure 2.1: Simple Pinhole camera model with intrinsics and extrinsics parameters. From simple geometry construction the relation between points in Camera Refernce Frame (CRF) and pixel space is derived. The transform between World Reference Frame (WRF) and CRF is derived using standard algebra transformation

More refined models which take into account camera lens distortion and manufacturing structure, have been developed to better approximate real cameras. In order to obtain the source of the projection recorded from cameras, for each image a bundle of projection lines is cast in the shared 3D space starting from the principal point of each camera and passing through the pixels. By intersecting the different projection lines, real 3D positions of points are found.

The stereo camera pair case is depicted in Figure 2.2. In real applications, projection rays do not intersect perfectly due to irreducible noise in the estimated camera parameters and manufacturing imperfections. To reduce the error and estimate good intersection points, the photogrammetry scanning procedure takes place in a controlled environment with well-calibrated cameras in fixed positions around the object. Additionally, numerical values and intersection points are refined with optimization steps. The whole process is expensive due to the hardware setup which usually includes multiple high-resolution cameras, structural supports and controlled light conditions as well as high standard calibration and manufacturing precision. A cheaper solution is often used through a rotating support and a single fixed camera.

A lot of applications are satisfied with the simple stereo pair setup and extracting only

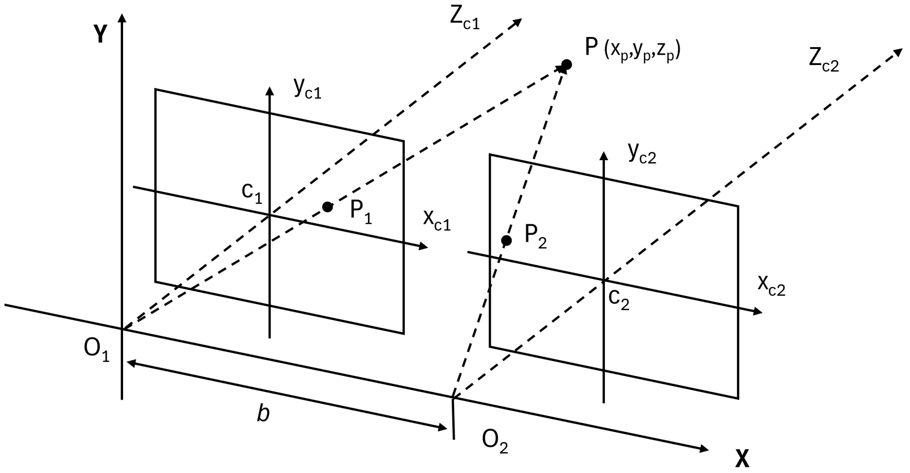

Figure 2.2: Scheme of a Stereo camera setup. The known distance *b* and the position of point P_{c1} and P_{c2} define univocally the triangulation of point *P* in world space, given the two reference frames aligned on the x-axis

the depth information. For example, in map reconstruction from aerial or satellite images [33], high-resolution cameras are used from the distance and altitude of the ground-level landscapes are recorder, as well as building positioning. Classical photogrammetry multi-view setup generates a high accuracy reconstruction which is well suited for various industrial applications where controlled setups can be integrated into production lines [2].

2.1.2 Structure from Motion

Structure from Motion [35] is a reconstruction techniques that compute the 3D structure of an object from an unstructured set of images taken from several unknown cameras. To achieve these results, a series of offline computational steps are required over the set of images to estimate the camera pose intrinsic and extrinsic parameters and to triangulate image points into the 3D space. The steps are summarized in the Figure 2.3

First, a set of keypoints are extracted from each image. They are pixel points selected for their great information and easiness of being recognized in different images. Usually, keypoints are extended with additional information about their neighboring pixels' structure, called descriptors. SIFT features [18] are commonly used to detect keypoints and extract meaningful descriptors which are scale, rotation and light condition invariant. Using feature matching, all possible keypoints between two images are detected by computing distances in the feature space. Key-points are often called feature points for

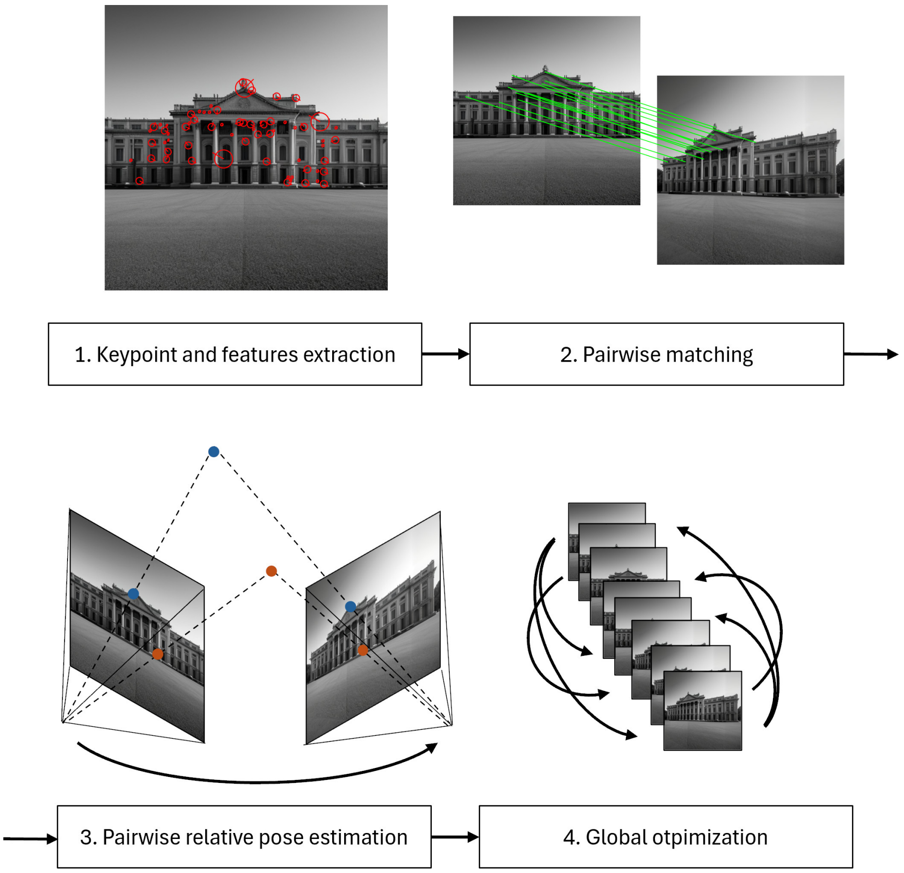
1. Keypoint and features extraction
2. Pairwise matching
3. Pairwise relative pose estimation
4. Global otpimization

Figure 2.3: Structure from Motion algorithm scheme

simplicity. Pairs of images sharing enough keypoints are detected and organized into a new data structure. The pairs are then used to reconstruct the camera intrinsics and relative camera extrinsic parameters using RANSAC-based algorithms [23]. After the last step, the problem can be solved as a classical triangulation points problem as in photogrammetry, reconstructing the 3D position only of keypoints for computational efficiency. Points are triangulated and a consecutive bundle adjustment optimization step is performed to reduce the error of reconstruction.

This method shows great capabilities in reconstructing large areas, such as city monuments [1], even from unstructured collections of images captured at different times by multiple agents. Industrial applications use SfM technique with the additions of known calibrated cameras for better precision of the reconstructed surfaces. This method is one of the best in terms of cost, relying on simple RGB pictures, without the need for calibrated expensive cameras and with ideally no acquisition time based on the availability of different pictures of the scene in previous times. Although the theoretical basis of this method was presented at the end of the last century, only recent breakthroughs in hardware optimization and parallel computing made the time-cost of the offline computing steps competitive with other techniques. Another important limitation of this method is the low precision of reconstruction since it relies on sparse feature points triangulation, opposite to denser points techniques. Additional steps are necessary to reconstruct a dense representation.

2.1.3 Structured light reconstruction

Structured light reconstruction is a technique that uses active light projection of known patterns (usually a grid, parallel stripes or a dot matrix) over the object to reconstruct its surface based on the deformation of the light spots. Knowing the position of the light source and the position of the sensing camera, a point can be triangulated in the 3D space with standard camera analytical computation. The setup is visually reported in Figure 2.4. Structured light scanners are a standard in 3D reconstruction for industrial settings: surface reconstruction scanners are integrated with production lines to inspect manufacturing errors [41], hand-held portable scanners are used for custom scans of objects with different sizes [7], multi-camera setups enhance the classical photogrammetry with the additional accuracy of structured light triangulation [40].

This technique is a fast and accurate method thanks to point triangulation within single images, but it presents some limitations inherent to the material properties of the object scanned. Reflective surfaces, semi-transparent surfaces and bad light conditions distort the projected light introducing not negligible error in the reconstruction. The acquisition phase in a controlled environment as well as applying opaque coatings to the material before the scan [19] drastically reduce the limitations of this technique, making it flexible and reliable. Structured light cameras are still expensive requiring a calibrated camera and a projector with high manufacturing accuracy, not affordable for mass users. The procedure of scanning large objects is time-consuming for a single hand-held camera sensor which can take hours to correctly scan all the parts of the object while maintaining good overlap between frames for correct positional tracking of the camera.

2.1.4 Light Detection and Ranging

LiDAR is a laser-based sensing technology that is able to reconstruct distances of object points from the emitting station. It is composed of an active sensor that emits laser waves

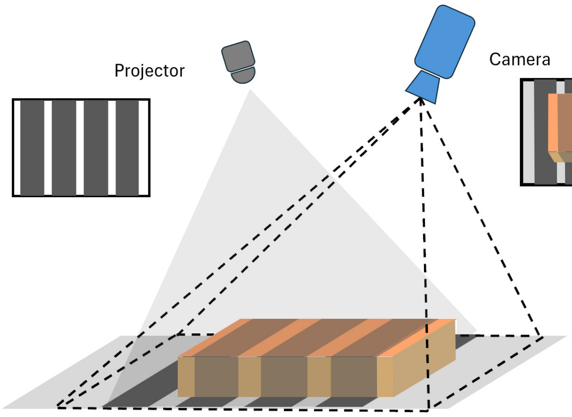

Figure 2.4: Scheme of the structured light reconstruction method. The setup is composed by a light projector and a sensing camera.

and a passive sensor that reconstructs distances based on laser beam travel time. Since mechanical laser scanners, introduced in the 1990s, LiDAR technology has been widely studied to increment accuracy, reliability and portability of the technique. Depending on the application, LiDAR technology differs greatly in complexity, cost and capabilities [3]. Compact time-of-flight (ToF) cameras are used in robot navigation due to their capability of measuring simultaneously and multiple times each second the whole scene [26]. Aerial LiDAR is used for mapping large areas with high distance laser accuracy, allowing the creation of digital elevation models for geographical information systems[39]. Nowadays, low energy and fast LiDAR sensors using Single Photon Avalanche Diode (SPAD) can be found in consumers' smartphones, for common user applications such as spatial measurements and improved media recording features.

To measure distance, the sensor records the energy of the returning signal from the reflected laser by specialized electric components. Most techniques use the Time of Flight (ToF) to retrieve the distance of surfaces from the emitter. Multiple data are then combined to obtain depth maps and point clouds. Depending on the LiDAR technology used, different properties of the material scanned can be extracted. In particular, lasers are highly susceptible to reflectiveness properties of the material. Opposite to structured light technology, High reflective surfaces mean a clearer reflected signal allowing for precise measuring even from long distances. Depth information of a scene is usually paired with standard RGB images by mapping the two sensory information in a unified RGBD image. The cost of acquisition can vary depending on desired accuracy and application, from high-precision high-distance LiDAR sensors mounted on satellites, to cheap low-resolution setups mounted in pairs with high-resolution cameras on smartphones. LiDAR sensors are widely spread in dynamic scene reconstruction applications where a high sampling rate is required for fast-moving objects and with a wide variety of objects with complex material properties appearing in the scene. LiDAR sensors are also used to augment standard photogrammetry setup providing additional depth information and easier triangulation with minimal cost.

2.2 Neural Networks optimization methods

This section focuses on Neural Radiance Fields (NeRF) and 3D Gaussian Splatting (3DGS). Different from traditional 3D reconstruction techniques, these methods leverage neural networks and optimized geometric primitives to achieve photorealistic novel view synthesis and detailed scene reconstruction. In the following, their principles, advantages, and recent advancements are discussed, highlighting their role in modern computer vision and 3D modeling applications.

2.2.1 Neural Radiance Fields

Neural Radiance Field (NeRF) models are an implicit 3D reconstruction technique, first proposed in 2020 for novel view synthesis [20]. They leveraged the computational capability of neural networks to approximate complex functions in order to predict the radiance field of a scene: a 5D function that for each position in space (x, y, z) and for each view orientation in polar coordinated (θ, ϕ) returns a color c and a density value σ. The model is trained in a self-supervised manner over a set of pictures to estimate the implicit representation in a two-stage prediction model. In the baseline NeRF model, the first stage predicts density σ and a feature vector of size 256, starting from the input position and view orientation. The second stage takes as input the feature vector and the viewing orientation to estimate the color information c.

Given the radiance field, reconstructed views are computed with volume rendering of camera rays for each pixel in the target image. For any camera ray $r(t) = o + td$, with camera position o and viewing direction d, the resulting color $C(r)$ is

$$C(r) = \int_{t_1}^{t_2} T(t) \cdot \sigma(r(t)) \cdot c(r(t), d) \cdot dt, \tag{2.2}$$

where $c(r(t),d)$ and $\sigma(r(t))$ are the color and density at specific ray point $r(t)$. The integration is computed along the ray direction of movement dt. $T(t)$ is the accumulated transmittance from t_1 to t which represents the probability of the ray traveling to the point t without being intercepted and is computed as

$$T(t) = exp(-\int_{t1}^{t} \sigma(r(u)) \cdot du) \tag{2.3}$$

A loss is used to guide the training and is computed as the square error loss between predicted pixel colors of the reconstructed views and the ground truth ones. For a single image, the loss is

$$L = \sum_{r \in R} ||\hat{C}(r) - C_{gt}(r)||_2^2 \tag{2.4}$$

where R is the batch of rays associated with the synthesized image and $C_{gt}(r)$ is the ground truth color associated with the single ray r. The whole technique is summarized in Figure 2.5.

NeRF implicit representation achieves state-of-the-art reconstruction of 3D scenes, with photorealistic novel view synthesis. Color variation due to illumination and material properties are implicitly reconstructed in the representation by taking into account camera view orientation during training, even though they are not separable from the rest of the representation. Dynamic relighting techniques have been proposed with good results [29]. NeRF reconstruction baseline pipeline is slow both in rendering new views and training, which takes hours or days to achieve state-of-the-art results on the scene. In its first step, registration of camera positions and an initial candidate of the point cloud is extracted using SfM technique, followed by the actual training of the radiance field. This method maintains the advantages and disadvantages of the SfM techniques requiring a large set of images with high overlap (advised to be $> 70\%$) but with no camera parameter needed. A subsequent work tries to speed up this technique by a faster rendering process through optimization of the ray sampling [9] achieving real-time rendering and faster training. Another limitation of implicit representation is the difficulty of integrating them in industrial 3D model pipelines, for example in applications like animation or editing. Subsequent works proposed different solutions, one of which is realistic animatable avatar reconstruction using NeRF [42] achieving good results in deformation and mathematical expressions to handle radiance fields manipulation. Still, converting the reconstruction in explicit representation is the most common approach for real application with an irreducible loss of accuracy in the process coming from insertion of artifact and discretization.

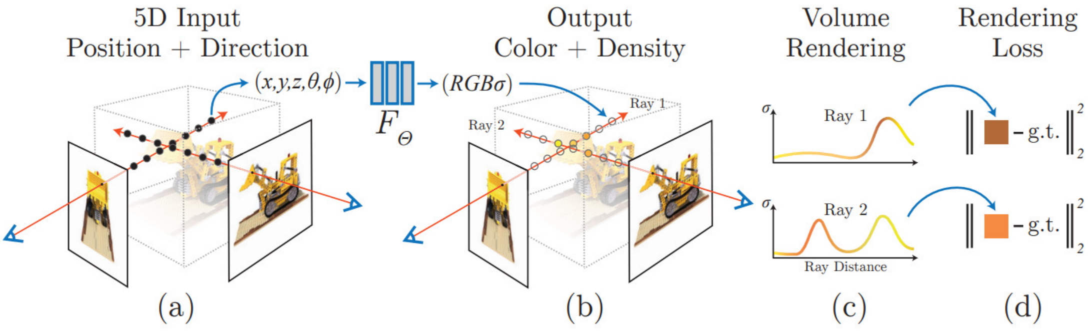

Figure 2.5: NeRF's pipeline scheme from original paper [20]. (a) Images are rendered by projecting rays in the 5D space, (b) color and volume density are estimated with an MLP for points along the rays, and (c) volume rendering allow to reconstruct the final image for each pixel, all of which is end-to-end differentiable (d).

2.2.2 Gaussian Splatting

3D Gaussian Splatting (3DGS) is a new implicit representation that achieved state-of-the-art results proposed in 2023 by Kerbl, et al. [15]. It optimizes a set of 3D Gaussian positioned in the 3D space to achieve photorealistic reconstruction and fast rendering of new views. Gaussian primitives are chosen for their fast α-blending rendering without need of computing normal: given a Gaussian $G(x)$ it can be fully expressed with a covariance matrix Σ and centered at point μ

$$G(x) = e^{-\frac{1}{2}(x)^T\Sigma^{-1}(x)} \tag{2.5}$$

For each primitive, projections in camera space can be found with

$$\Sigma' = JW\Sigma W^T J^T \tag{2.6}$$

where Σ' is the covariance matrix in camera coordinates, W is the viewing transformation and J is the Jacobian of the affine approximation of the projective transformation. To correctly optimize the covariance matrix Σ which has physical meaning only when it is positive definite during training, a new decomposition of the matrix is used to enforce the constraint

$$\Sigma = RSS^T R^T \tag{2.7}$$

where S and R are respectively a scaling and rotating matrix. These two values are optimized independently. Color is parametrized using SH coefficients to correctly capture the different viewing point color functions and an α parameter controls the Gaussian transparency in a multiplicative fashion. Additionally, during the training process, a step called *Adaptive Control* of Gaussians is performed to densify areas in the learned representation. They found that areas missing geometric features and areas represented only by a few large Gaussians both have large view-space positional gradients and they are the focus of the Adaptive control. This step adds new primitives by splitting large ones or by populating the region of interest randomly. Elements with α value lower than a threshold α_τ are also removed during this step to maintain only useful Gaussian contributions.

Optimization is performed by iteratively projecting Gaussians to desired views, compute loss and backpropagate the gradient to correctly estimate position, rotation, color, transparency and number of Gaussians in the representation. The whole process is summed up in the Figure 2.6.

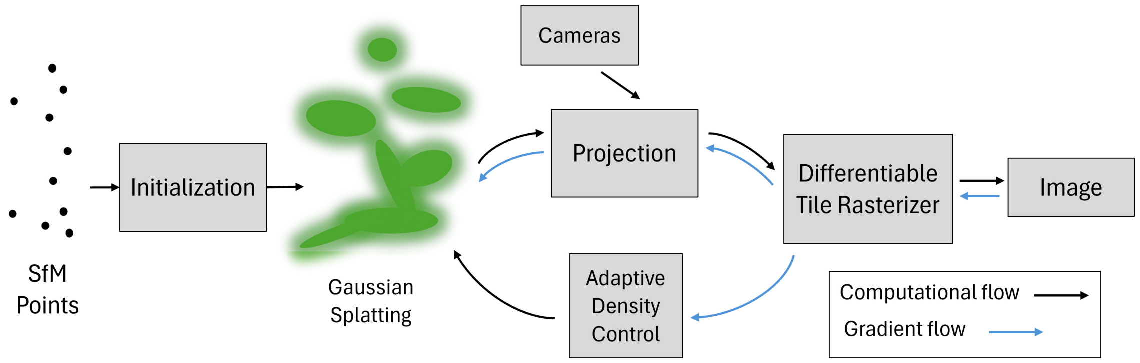

Figure 2.6: Gaussian pipeline scheme [15]. Starting from the point cloud extracted with structure from Motion, the Gaussian splatting model is initialized. The optimization is performed by iteratively computing the projection of Gaussians according to camera poses and compared with ground truth views for backpropagation of gradient.

As for NeRF models, the initial candidate of the Gaussian Splatting technique is derived from an SfM stage over the input set of images. The camera parameters for each view are unknown and will be estimated during this step. The points in the extracted point cloud are used as centers μ of initial candidate Guassians with randomly assigned covariance matrices Σ. This initial step still requires a large number of overlapping images to correctly extract camera parameters and correctly register input images. Instead of NeRF models, the Gaussian Splatting baseline already achieves fast rendering of novel views thanks to the simple computation of the projection matrix of each Gaussian. Visualization at each stage can be run in real-time. From the results reported in the original paper, the optimization of the Gaussians is reported to be faster than the one required for training NeRF model and it reconstructs the model with higher accuracy, scoring 0.815 SSIM, 27.21 PSNR and 0.214 LPIPS on the Mip-NeRF360 Dataset, with a training time of 41 minutes. The best performing M-NeRF360 technique scores instead 0.792 SSIM, 27.69 PSNR and 0.237 LPIPS on the same dataset with a training time of 48 hours. Moreover, from visual inspection, the reconstruction of Gaussian Splatting better represents thin structures with high occlusion of the scene, like foliage and wheel's ray of bicycles.

Commonly to other implicit representation methods, it is tricky to perform editing of the representation for standard applications, such as animation. Opposite to NeRF representation, Gaussian still uses a set of geometric primitives that can be more easily understood since they have clear spatial positioning, rotation and scale similar to other common explicit representations. Some work have tried to animate Gaussian Splatting representations using video data and learning the movement of the Gaussians over time as well as relighting the representation [16]. Some other works have tried to extract explicit representation from the Gaussian one for easy reconstruction and integration in standard industrial applications. SuGaR [12] proposed additional loss and a refinement step to better align the Gaussians to the ground truth surface, followed by sampling of the points for Poisson surface reconstruction. 2D Gaussian Splatting (2DGS) [13] proposed to use two-dimensional Gaussians instead of 3D ones and align them with the surface of the object to better approximate the geometric properties. They then used a marching cube approach to extract a mesh representation of the scene.

In my work, we analyzed the problem of reconstructing close-range objects using 3DGS starting from depth stereo camera input for ready-to-use representation. This is achieved by leveraging the depth information from the sensing camera to estimate better point cloud initialization of the Gaussian Splatting pipeline. It also adds segmentation masking preprocessing removing the background and correlated unwanted computations. Results showed high accuracy reconstructed 3D representation of the object robust to noisy input images and complete isolation of the target from the scene. It additionally showed a boost in performance of the early hundreds iterations thanks to the depth-enhanced initialization, useful for low computational capacity machines.

2.3 Neural Networks prior models

Presented methods so far can achieve reconstruction of objects only partially if a small number of images is taken into consideration. Human beings, otherwise, can infer and guess the whole 3D structure of an object from one or few images by using experience and stored knowledge of real objects. Since the explosion of research on 2D generative models and thanks to recent breakthroughs in neural network architectures such as transformers, 3D generative models become the next step in generative AI research. Single-image-to-3D or text-to-3D generative models have been proposed in the past years [25, 21, 27]. The key idea is to train a neural network in a supervised manner having as input an image, few images or only a text prompt and infer the explicit surface representation of the ground truth 3D model. This is possible due to the proposed mathematical computation of the gradient flowing back from the image comparison to the actual representation. One example is Score Distillation Sampling (SDS) proposed as part of Dreamfusion [25] to use the Gradient Descent algorithm to steer a NeRF model and an Image Diffusion model to generate a 3D representation from a text prompt.

Although promising results, the quality of the extracted reconstruction is still not comparable with the photorealistic performance of other proposed methods. It is difficult for the learned model to exactly reconstruct the real scene or object, especially in the details where no 2D image information is given. Their application is then limited to the designing process instead of the actual reconstruction of real objects. No methods of this type will be analyzed in this work, although they are important to mention as one of the main areas of research for 3D reconstruction in the next years.

2.4 Image pre-processing

Image pre-processing is a fundamental step in working with 2D image data. Pixels are a discrete representation of the real scene information, discretized both in the spatial and intensity values by the capturing sensors. Some pre-processing is performed at the hardware level of the cameras while some is performed at the software level over the captured image. We will focus on the second part. In particular, we will focus on the Undistortion and Segmentation processing step used for close-range object reconstruction as the scope of this project.

2.4.1 Image Undistortion

As seen in **Section 2.1**, RGB camera pixel positions are used to triangulate the 3D spatial position of pixel points, according to Pinhole Camera. Real cameras use lenses to define the field of view and the focal length with which the sensors capture the scene. This increased flexibility comes at the cost of a distortion of the standard projections into the camera space, introducing errors in the subsequent inverse problem. The problem has been extensively studied and distortion models have been proposed to correct the positional shift of pixel positions in the captured images. Undistortion is the process of reconstructing the real scene image removing the lens effect by applying the inverse distortion transformation. When calibrating a camera, the distortion parameter could be estimated. During SfM, distortion parameters are automatically computed based on multiple image correspondences. Brown's distortion model [4] is one of the most used ones for classical RGB cameras, taking into account both radial and tangential non-linear distortions.

2.4.2 Segmentation

In some 2D to 3D reconstructions, such as aerial photogrammetry, all the pixels of the image are used for reconstruction. In close-range object reconstruction, instead, the task focuses on the main subject, usually appearing in the foreground. Segmentation is useful in this process to mask some part of the image, usually the background, and use only useful information about the target without additional computation cost. Automatic segmentation algorithms are commonly used instead of hand-made ones, reducing the time and cost of this step. Neural Network approaches have conquered the semantic segmentation task consisting of masking part of the image depending on the semantics of the object captured. Fast and accurate models can be used to segment images based on text prompts, foreground and background distinction or with simple manual annotations. Moving from RGB images to other sensory information, segmentation can be processed using depth value associated with captured pixels. A simple or complex threshold filter can be implemented to separate pixels from the foreground to the background, usually enough for close-range reconstruction applications.

2.5 Global Optimization Methods

This section explores optimization techniques for refining camera poses and 3D points in reconstruction tasks. Global Bundle Adjustment (BA) and Pose Graph Optimization (PGO) are highlighted as critical methods for ensuring accuracy and global consistency.

2.5.1 Global Bundle Adjustment

Global Bundle Adjustment is an optimization method that uses the reprojection error to optimize a set of points and camera positions at the same time. Given the error from the points reprojected in the camera space, according to classical camera equations, and the actual points in camera space, the Jacobians of the error are computed. Then using the Levenberg-Marquardt method for non-linear least squares to solve the Normal equation, the correction of camera poses and point positions is computed. Finally, the adjustment is applied and actual positions are outputted. This process is repeated until convergence is reached. Global bundle adjustment needs a good initial pose estimation to ensure convergence and can be computationally costly for a large collection of poses to be optimized. This method computes accurate pose reconstruction and ensures coherence across the whole scene. It is often used as the final step of the Structure from Motion (SfM) algorithm.

2.5.2 Pose Graph Optimization

Pose graph is an optimization technique to improve the estimated poses of cameras. It is often used in SLAM robot navigation applications and it can efficiently run even on thousands of poses. Each estimated pose is represented as a node in the graph and edges represent constraints between poses. The relative position of estimated poses, gyroscope information or other constraint can be used to define edges. One important constraint is the loop closure, where two estimated poses are considered to be very close at the beginning and at the end of a loop around the scene. The optimization is performed by iteratively computing the error of each estimated pose according to the selected cost function, using the Levenberg-Marquardt method for non-linear least squares to solve the Normal equation. Required Jacobians are computed on the edges around the node. Solving the equations, correction adjustment of the poses are derived and new poses are computed.

When the correction values are below a certain threshold, the optimization is considered to have reached convergence. Pose graph optimization can correctly reconstruct camera positions and ensure global consistency across the scene.

Chapter 3

RGBD Initialized Gaussian Splatting

Gaussian Splatting stands out as one of the most effective techniques for balancing photorealistic 3D reconstruction with efficient rendering performance. As a recently introduced method, its full potential and capabilities are still being explored. In this chapter, a novel pipeline is presented. The developed method leverages Gaussian Splatting to achieve highly accurate and photorealistic reconstructions of small to medium-sized objects. The approach integrates advanced mask filtering for precise segmentation and depth-enhanced initialization to improve geometric accuracy and optimization efficiency for accessible and high-quality 3D modeling.

3.1 Experimental Pipeline

The pipeline to obtain a full photorealistic reconstruction of an object is composed of two distinct sections: the acquisition part and the offline processing. From acquisition which can be performed with different strategies, a video of the desired object with RGB and depth channels is retrieved. This first step has been designed to be as simple as possible using a standard recording device easily available in the market and it will be described in detail in **Section 4.1**. The video will record the desired object from a wide number of positions ensuring that all parts of the object are seen at least once. After the acquisition, the data are processed in an offline pipeline composed of 4 stages:

- **Frame extraction and segmentation**. A fixed set of frames with RGBD channels is extracted and processed for the next steps, discarding the leftover information. During these steps, the segmentation mask for extracted frames is computed.

- **Structure from Motion (SfM)**. Using only the RGB channels of the image set, the different camera intrinsic and extrinsic parameters are estimated in pairs with a point cloud based on projected keypoints in the 3D space.
- **Depth based point cloud**. The initial estimated point cloud is replaced with a more accurate and dense point cloud based on depth information from single images, projecting points according to their distance value into the 3D space. An estimation step to correctly scale the SfM space into real measurement is also performed in this stage.
- **Gaussian Splatting optimization**. Gaussians are initialized based on the extracted point cloud and optimized according to the original paper technique, resulting in an accurate reconstruction of the object already separated from the background scene.

The pipeline is visually reported in Figure 3.1. The presented pipeline offers great flexibility in each step allowing for the replacement of single stages in a modular fashion without compromising the whole process. This is particularly useful depending on the application: offloading some stages of the computation to external machines can help to implement the pipeline even on low computing power devices. Each stage is reported in detail in the next sections.

3.1.1 Frame extraction and segmentation

The input video is composed of N^* image frames each with dimensions $H \times W \times 4$. A fixed set of frames N is extracted from the video by sampling the signal with a fixed step $\Delta ts = \frac{N^*}{N}$, starting from the $k-th$ frame with $k = \frac{\Delta ts}{2}$. Since Structure from Motion techniques rely on feature matching for correctly registering the different camera poses, a good overlap between extracted frames is needed for the next stage. As a rule of thumb, N should be ≥ 100 for medium-sized objects recorded with a looping movement around them. The design choice to sample with fixed distance comes from the acquisition policy used in experiments, where the camera position follows multiple loops around the object. Moving around the target takes more or less the same time for each loop, so the fixed timestep sampling ensures a well-distributed collection of images across the possible views. The initial frames which often contain noisy frames due to initial camera grabbing and setup are skipped by starting at the k-th frames.

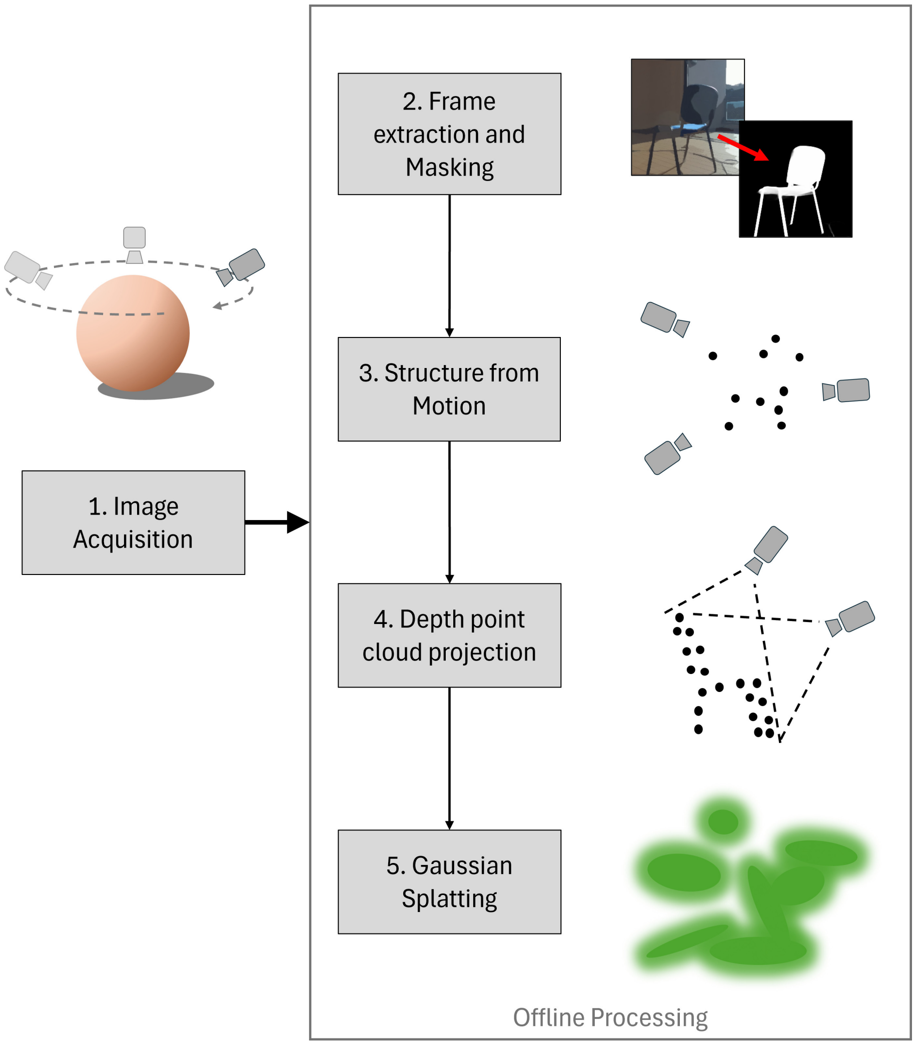

Figure 3.1: The proposed pipeline for full reconstruction of 3D object with Gaussian splatting representation.

After extracting the frames, the background and pixels not belonging to the object should be segmented out, in order to let the Gaussian Splatting optimization focus only on the desired reconstruction. To do so, a state-of-the-art pre-trained neural network for semantic segmentation is used. Given an image frame I, the networks compute a segmentation mask M with a pixel value $m_{ij} \in [0, 1]$ indicating if the pixel in position (i, j) belongs to the foreground object or not. During Gaussian optimization, only the pixel belonging to the object will contribute to the gradient during the propagation of error. No temporal coherence is used while evaluating the masks to increase the flexibility of the pipeline to work with images acquired from different cameras at different times, even though it could reduce noise in the mask estimation. Examples of segmentation masks are depicted in Figure 4.5. In the experiments, it has been used the python library named *rembg* [10] where different pre-trained network models are collected in a unified and ready-to-use environment.

3.1.2 Structure from Motion

The standard Structure from Motion algorithm is described in detail in **Section 2.1**. In the proposed pipeline, according to the original 3D Gaussian splatting technique, COLMAP [31, 32] library has been used to implement the registration algorithm with some modifications. Equal to the standard pipeline, only RGB channels of image frames are used in the algorithm to extract features for matching. In COLMAP implementation, SIFT features [18] and keypoints are extracted for each image, being them invariant to scale, orientation and color. With the addition of segmentation masks, only keypoints belonging to the region of interest are used in the matching step, while others are not considered. Depending on the object scanned, it could be necessary to avoid using the segmentation mask in this step. In fact, objects with a low number of keypoints due to uniform flat areas, as well as the opposite case with a high number of misleading ones from reflective surfaces, can make the matching problem hard. In these cases it has been found that it is better to impose the segmentation mask only at the end of this stage, allowing keypoints and features to be extracted from the background too, with a higher chance of correct registration.

The global optimization step for adjusting the different camera poses is the Global Bundle Adjustment (see **Section 2.5**), performed over the set of matches to minimize reprojection error. Computations are handled using blocks of 50 images and then aggregated together, according to standard COLMAP practice.

As an output from this stage, each image I_i is associated with its corresponding camera pose in the form of a rotational quaternion vector Q_i and a translation vector T_i which univocally describes the pose according to the global reference frame. Camera intrinsics parameters are also computed and associated with each image. As a complementary output, the point cloud P_{COLMAP} used for reprojection error computation is delivered. Color information and links to the images in which they appear are associated with each point, allowing direct relation with input images.

3.1.3 Depth point cloud projection and scale estimation

Structure from Motion output is accurate and robust against noise in input images, but the reconstruction is not in scale with the real object. The point cloud generated for the reprojection error is also sparse, including only correctly matched keypoints that are usually present only on the edges of objects and in high-contrast textured areas. As one of the main contributions of this work, in this stage the depth information retrieved from the sensing camera during acquisition is used to correctly scale the reconstruction and to populate the point cloud with additional projection points, allowing for a full surface reconstruction.

Given an RGBD image I_i and its associated intrinsic matrix A and extrinsic matrix E, any pixel in the camera space can be projected to the 3D global reference frame by inverting Eq. 2.1. First 3D coordinates $\tilde{X} = (\tilde{x}, \tilde{y}, \tilde{z})$ in the camera reference frame are computed with $\tilde{z}$ being known from the depth camera and the other coordinates computed as:

$$\begin{pmatrix} \tilde{x} \\ \tilde{y} \\ \tilde{z} \end{pmatrix} = \tilde{z}A^{-1} \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \tilde{z} \begin{bmatrix} \frac{\rho_u}{f} & 0 & \frac{-\rho_u u_0}{f} \\ 0 & \frac{\rho_v}{f} & \frac{-\rho_v v_0}{f} \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} \tag{3.1}$$

Then the coordinates in the global reference frame are computed as

$$\begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} = E^{-1} \begin{pmatrix} \tilde{x} \\ \tilde{y} \\ \tilde{z} \\ 1 \end{pmatrix} = \begin{bmatrix} & R^T & & -R^T T \\ & & & \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} \tilde{x} \\ \tilde{y} \\ \tilde{z} \\ 1 \end{pmatrix} \tag{3.2}$$

From each image I_i, any pixel with a depth value can be projected in the 3D space creating a point cloud P_{I_i}. To solve the problem of scale alignment, the depth information of points appearing in both representations (COLMAP and depth derived) is used. Given a point X_{COLMAP} of the COLMAP point cloud, the corresponding pixel point U_i appearing in image I_i is also given. The depth value of that point from the camera view i can be derived by computing $\tilde{z}$ in the camera reference frame. The scale value $s_{U,i}$ for the point U_i is computed as the ratio between the reconstructed and real depth values. Each point can appear in multiple images leading to a set of different scale estimations for each point in the point cloud and for each image in which it appears. To solve this problem and retrieve the actual scale s^* a probability density function (pdf) is computed using Kernel

Density Estimation (KDE) with a Gaussian kernel. Camera pose and COLMAP point cloud are scaled according to the scale $s*$ matching the maxima value of the estimated pdf.

The final point cloud used in the initialization is the union of all the point clouds projected from each RGBD image, creating a dense point cloud distribution over all the surfaces of the object. Depth maps from easily affordable cameras usually rely on stereo vision to compute depth values and present a lot of holes in the output depth map. Even ignoring pixels with no associated depth values from the camera sensor, the point cloud generated from a single high-resolution image is in the range of thousands of elements, generating a heavy global point cloud in the order of hundreds of thousands of elements when merged with other images. Since the input set is composed of images with high overlap, there is a high redundancy of points over the surface. To reduce the number of elements in the final point cloud, a voxel downsampling step is used before the union of point clouds from the different images. Segmentation mask is also applied during projection of point clouds effectively reducing the points belonging to the background. A comparison between the point cloud reconstructed with COLMAP techniques only against the proposed method is portrayed in Figure 4.6.

3.1.4 3D Gaussian Splatting optimization

The Gaussian splatting optimization method is described in detail in **Section 2.2**. The proposed pipeline uses the original 3DGS paper to optimize the implicit representation but it is affected by the previous steps of estimating the new initial point cloud and segmenting the images. Initialization of the scene is performed by instantiating a Gaussian primitive in each point described in the point clouds using the positions as centers. The optimization steps try to move and deform Gaussians to approximate target geometry. Starting from a point cloud that better approximates the shape of the object, a boost in the first steps in the accuracy of reconstruction is recorded. Moreover, the images that are used to guide the optimization process are segmented, considering only the foreground object, and removing background-related computation in the optimization cycle. As output from the pipeline, depending on the quality of the segmentation mask, the 3D reconstruction is completely separated from the background with minimal need for manual post-processing. Still, some artifacts represented by outlier Gaussians are present in the final reconstruction due to irreducible noise. Gaussian Splatting is an implicit representation and could be difficult to use in standard workflows for animation and editing. Some methods, like SuGaR and 2DGS, can be used to extract a mesh from the optimized representation. Thanks to the flexibility of the pipeline, the mesh can be extracted just by replacing this stage with the desired Gaussian optimization method and the mesh can be obtained.

3.2 Preliminary Studies

The proposed pipeline and the design choice to use Gaussian Splatting representation for 3D reconstruction against other methods come from a set of preliminary studies executed over an object with cultural importance. To define a baseline and a comparison between reconstruction methods, an elaborate theatrical outfit of the famous singer Pavarotti was used as a test subject in 4 different systems all different in economic cost and acquisition time cost. The outfit is part of a collection held by the *Luciano Pavarotti Foundation* [8]. The evaluated methods are:

- Structured light hand-held scanner *Artec Eva*. This device is state-of-the-art for industrial reconstruction, capturing 16 frames per second with up to 18 million data points with high precision (up to 1.0mm) and 1.3MP textures. It uses a pulsed light approach with a known pattern to triangulate geometry features (as described in **Section 2.1**) and it uses an online registration method to directly align frames while recording. The reconstruction is handled by the Artec Studio software with functions for merging different scans and editing the mesh. The output has been used as the geometrical ground truth for the comparison due to its high accuracy in reconstruction. The high cost of the technology and the tedious acquisition strategy may be prohibitive for some users and applications. Also, poor lighting conditions and highly reflective surfaces can impact negatively the quality of the scan.

- Iphone’s LiDAR camera with *Scaniverse* app [22]. The iPhone 12 PRO is equipped with a LiDAR sensing camera which emits laser pulses measuring distances from the object (see **Section 2.1**). The Scaniverse app leverages the RGBD data collected by the smartphone during the acquisition to create a dense point cloud and enhance it by applying the high-resolution texture from 12MP cameras. Scaniverse scan is quick and easy to use, visualizing the model created in real-time and allowing for editing or exporting of the 3D reconstruction to the most known formats. During the scan, the app also leverages other smartphone sensors such as accelerometer and gyroscope data to obtain a better online registration. This method has some limitations, mostly due to the poor resolution of the 24x24 LiDAR scanning grid which cannot capture small geometrical details. Moreover, only the iPhone 12 Pro series and newer models are equipped with the LiDAR sensor, while most smartphones are not, making it unaffordable for some users.

- Structure from Motion photogrammetry with Android smartphone. Using a simple Android smartphone, with the help of the *Polycam* app [24] an accurate reconstruction from high-resolution RGB images can be computed using the Structure from Motion algorithm. The acquisition consists of capturing multiple images of the objects from different views, with a minimum of 20-30 images per object. In the reported analysis it has been used a Fairphone 4 android smartphone with 48MP. Polycam's software analyzes the images, registering them by features point matching and creating an initial point cloud. Then it performs additional optimization steps and computes a mesh representation of the object with high-resolution texture from the captured images. The surface smoothing step is a critical step where any rough or uneven surface is smoothed out, as well as the hole filling step where the mesh is corrected to be watertight. For complex reconstruction with intricate details, a lot of images are needed during the offline reconstruction. Any Android device with a camera can be used for reconstruction, resulting in an affordable technique for a wide variety of users. The Polycam app offers a free trial version for reconstructions (as used in this analysis) and a paid subscription version with additional features such as exporting, editing and cloud storing.

- SuGaR with Android smartphone recording. This pipeline is the implementation of the original paper Surface Aligned Gaussian Splatting [12] using frames extracted from a video recorded by a standard Android smartphone as input images. For the acquisition, a video of the object while moving in loops around it with different altitudes and inclinations has been recorded. Around 180 frames were extracted from the video and used as input of the Gaussian Splatting technique which starts by registering the images using Structure from Motion and then optimizing an implicit representation of the object by 3D Gaussian primitives. The input video was recorded using a Fairphone 4 android device with 48MP as in the previous scenario, while offline computations were performed on a 24GB GPU machine. SuGaR technique additionally to the standard 3DGS technique (**Section 2.2**) optimizes the Gaussians to be aligned with the reconstructed surface, allowing for a better geometry approximation of the target and subsequently a better mesh reconstruction. The technique uses Poisson surface reconstruction from datapoints sampled from surface probability estimation of the Gaussians themselves to reconstruct the mesh geometry. This technique requires a high-capacity GPU machine in addition to the acquisition device, but the computations can be offloaded to servers similar to Polycam app, making it easily affordable for users. Also, the acquisition technique starting from a video is faster than taking a large number of pictures with a high overlap of the same object.

As a design choice, the study didn't use any NeRF technique for comparison for the following reasons. First, Gaussian Splatting is an implicit representation technique as well as NeRF and in extensive analysis done by the authors of the original papers emerges that 3DGS can achieve higher quality reconstruction than state-of-the-art NeRF models. In particular, it can reconstruct accurately the geometric structure of thin occluded objects such as foliage as well as intricate details, a desired quality in high-quality reconstruction. Additionally, the Gaussian Splatting technique shows high performance in rendering new views, allowing for real-time visualization of the representation. Aiming to use the implicit representation to its full extent without necessarily exporting to other explicit representations, Gaussian Splatting was a better candidate. Finally, Gaussian Splatting representations are easier to edit than NeRF's ones for easier integration in industrial applications.

3.2.1 Comparison Results

The presented techniques were compared by approximate costs, both in terms of spent time for the reconstruction and in economic terms and reported in Table 3.2. The quality of the reconstruction has been assessed quantitatively by confronting the rendered views of the reconstructed mesh objects with images manually selected from the pool of acquisitions. Examples of the view are reported in Figure 3.2. PSNR and SSIM scores were

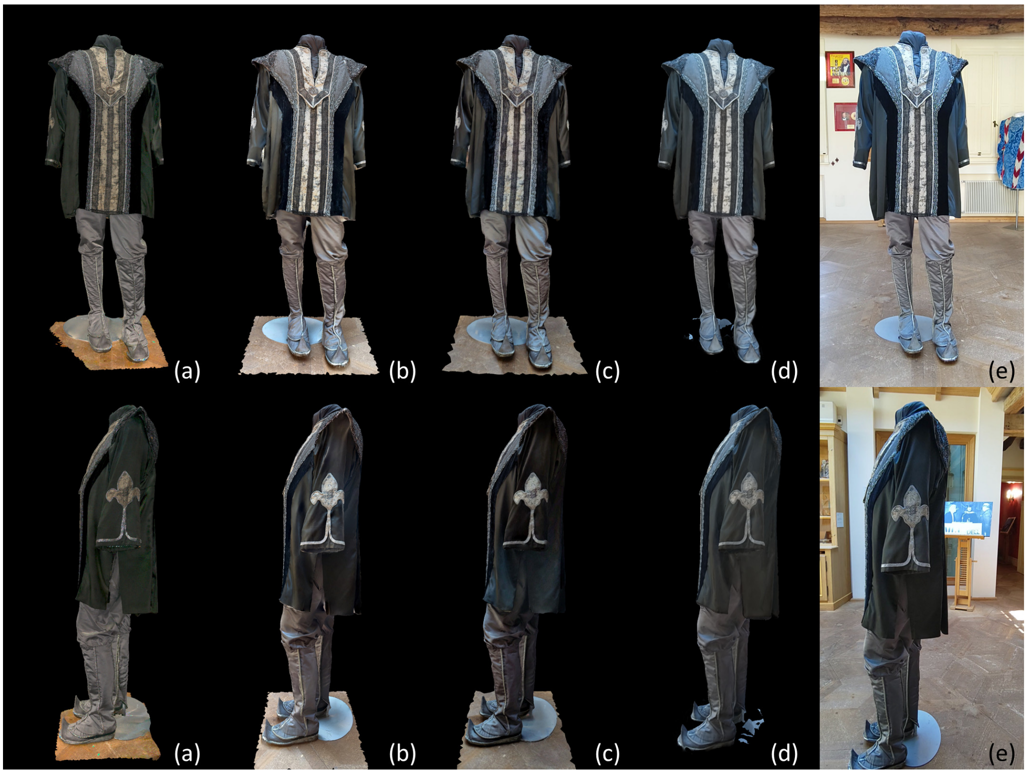

Figure 3.2: Reconstructed model from two different views: front (top row) and side (bottom row). The images are rendered from the resulting mesh reconstruction of a Pavarotti theatrical outfit depicted in (e). The different techniques reported for visual comparison are: (a) Artec Eva (b) Scaniverse (c) Polycam (d) Gaussian Splatting.

used to quantify the error between the two views and collected in Table 3.1. An alignment step between images was used to ensure that the small mismatch in object position between the real image and the rendered ones had no impact on the numerical computation. Additionally, only the foreground pixels were used in the evaluation by segmentation masking.

Geometric reconstruction performance was assessed qualitatively and reported in Figure 3.3. In techniques that reconstructed the background in combination with the desired object, a simple 3D box Region of Interest selection around the object was used to separate the background. From the quantitative results, reported in the Table, we can see that Gaussian Splatting techniques perform better than other techniques in terms of both PSNR and SSIM. The colors of the reconstruction and the high resolution of details match the input images more accurately than the other techniques. The Artec reconstruction performs the worst in terms of new view comparison, especially due to the low-resolution camera used in the handheld device and in the use of pulsed light resulting in a more ac-

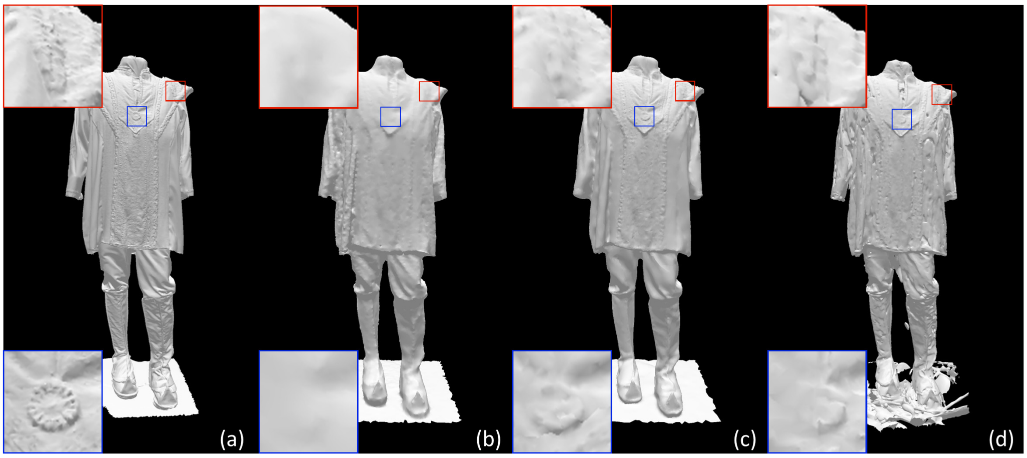

Figure 3.3: Visual inspection of geometrical reconstruction from the four analyzed methods. The Artec reconstruction (a) is considered as ground truth for its high geometrical resolution. Following in sequence Scaniverse (b) misses the small geometrical features, Polycam (c) and SuGaR (d) presents an attempt to reconstruct the fine details.

Method	**PSNR**	**SSIM**
Artec Eva	20.80	0.8118
Scaniverse	22.06	0.8372
Polycam	21.64	0.8305
SuGaR	**24.42**	**0.8738**

Table 3.1: PSNR and SSIM comparisons among the analyzed 3D reconstruction methods.

curate but less natural coloration of the object. Still from qualitative analysis of the mesh, the Artec technique shows great details and accuracy compared to the other ones, accurately reconstructing small details and intricate patterns of the theatrical outfit. LiDAR iPhone's scan presents a high-quality visual reconstruction of the object, mainly due to the texturing and high resolution of the camera, while the geometry of the object is of low resolution, missing important details. Structure from Motion only from the Polycam app presents an acceptable reconstruction in terms of geometry but scores lower than LiDAR's technique in terms of visual reconstruction. Gaussian Splatting emerges as the best trade-off between the quality of the geometric reconstruction and visual reconstruction. Considering also the costs of the techniques, Gaussian Splatting is an easy and affordable technique for 3D reconstruction of objects, democratizing the production of accurate 3D representation of everyday objects. The full extent of the technique has yet to be studied,

Method	Acquisition time	Processing Time	Estimated cost
Artec Eva	$1h$	$40min$	14000€
Scaniverse	$5-10min$	$40min$	500€
Polycam	$5-10min$	$20min$	500€
SuGaR	$5-10min$	$1h\ 20m$	500€

Table 3.2: Approximative temporal and economical costs of the analyzed methods.

given it only recent proposal, hoping for an increase in performance and accuracy of the reconstruction in future studies.

3.3 Expectations

The proposed pipeline aims to achieve photorealistic accurate 3D automatic reconstruction from a simple acquisition strategy generating an almost ready-to-use representation of the desired object. Segmentation masking is a key factor in the process allowing for computational cost reduction by focusing on the foreground target. As an additional contribution, segmentation masking almost completely removes the need for an additional manual polishing step of the representation, usually needed to extract the foreground object from the surrounding scene. The integration of the depth information from input acquisition should bring two main advantages. First, the 3D reconstruction will be on scale, allowing measures and analysis of the geometry of the object to be meaningful as they were executed on the real scene. Second, by adding points that accurately approximate the geometry surface it should help the optimization steps to converge faster to an accurate reconstruction. In the long run, using 30000 steps the accuracy of reconstruction using the depth enhanced initialization compared to the normal one should not produce any noticeable difference in the reconstruction error since the method has plenty enough time to correct the underrepresented regions. It may certainly arrange Gaussians in a more uniform way across flat surfaces thanks to initial positioning instead of interactions from gradients only. In fact, although the COLMAP starting point cloud presents holes in these regions, the adaptive control steps effectively populate them with a sufficient number of Gaussians to ensure a good reconstruction with negligible computational cost. In the short run, however, from 100 to 3000, the better point cloud initialization should definitely improve the reconstruction accuracy. This is useful in applications with low GPU capacity devices, where the optimization steps of the Gaussian Splatting technique are significantly slower and performing a high number of steps is not feasible. The higher performance in the early steps of reconstruction is paid with a fixed time cost for integrating the depth information in the pipeline. The overhead cost should not be burdening as it does not require GPU capability, making it very suitable for this kind of application. It is also useful for online applications where the computational time for each partial representation is limited and the increased accuracy of the reconstruction in the first steps plays a major role in the overall performance.

Chapter 4

Evaluation and Experimental Results

The proposed method is evaluated through both quantitative and qualitative analyses under various experimental parameters. To demonstrate its effectiveness, experiments were conducted on two representative objects, achieving detailed 3D reconstructions that are nearly ready for practical use. This chapter starts with a description of the experimental setup and parameters to ensure reproducibility, followed by an analysis of performance metrics and their implications.

4.1 Acquisition strategy and Experimental setup

The experiments were done over two common objects, chosen for their geometric properties. The first object is a chair showing flat lowly textured surfaces plus complex thin structures and holes. The second item reconstructed is a large gym ball with a foam pedestal that shows textured round surfaces with little details. Both acquisitions were done under outside natural lighting conditions. The videos were captured with ZED 2 device [34], a mid-tier cost RGBD camera, and a custom script using the ZED API. Although the ZED 2 camera has multiple sensors for registering the position and movement of the camera during acquisition, only the left camera RGB images and the depth values from the stereo setup were used, increasing the flexibility of the technique. The RGBD information was then transferred to an AMD Ryzen 9 5900X 12-core processor machine with 64GB of RAM and 24GB of GPU for offline processing. The acquisition is performed by moving the viewpoint around the object while trying to keep the center of the camera cone over the target. No visual feedback of the recording was given during the acquisition and noise is inevitably introduced in the forms of motion blur distortion or partial crop. These frames, if extracted, were kept nevertheless and results show the robustness of the approach to noisy input images, suitable for application *in-the-wild*. For the first object, two circular loops were performed around the object, one at a central height compared to the target center and the second one from a higher perspective view angle. For the gym ball, three loops were used thanks to the pedestal that allowed an extra loop from a lower perspective angle. The loops can be visualized from the registration reconstruction of the camera poses depicted in Figure 4.6. The recordings were respectively of $31s$ and $72s$ duration, both with 15 fps frame-rate and resolution of 1280×720.

Pre-processing was performed by extracting 100 frames from the captured videos and computing the masks using the open-source *isnet-general-use* [28] pretrained model from the publicly available *rembg* library [10]. For both of the experiments, segmentation masks were used during the Structure from Motion (SfM) algorithm implemented in the open COLMAP software, considering only keypoints inside the segmentation masking. The masking and registration process combined took respectively $235.43s$ and $209.72s$ for the chair and the gym ball experiments. The next stage of the pipeline projected the point cloud from depth images to the global reference frame. For both experiments the voxel downsample step to reduce the number of elements is performed with a 0.05 voxel radius. For the chair experiments, this stage took $30.19s$ and increased the number of points from 9670 to 56479. For the gym ball experiments, it took $28.78s$ and increased the point cloud from 7807 to 32764 points.

In the last step, the 3D Gaussian Splatting original implementation is used by executing the open source code and all the parameters were kept the same except for the white background option. The option uses the white color to represent the background segmentation mask, avoiding computation over that region. To study the effect of the different stages of the pipeline, for each experiment, three different reconstructions are compared. First, as a baseline, Gaussian Splatting with unmasked input image and original point cloud reconstruction is computed. Then the segmentation masking of input data is introduced both in the COLMAP estimation and in the optimization stages. Last, the depth information is integrated in addition to the segmentation masking for the final pipeline. The training behaviors are compared for the two pipelines using masked input images, while the original baseline is compared only with visual inspection and for time baseline.

The quantitative comparison is not possible for the baseline as it requires manual extraction of the target from the scene introducing unmeasurable effects. The Gaussian Splatting optimization was performed for 30000 steps, with Gaussian adaptive control each 3000 steps till the 15000 steps mark as standard Gaussian Splatting practice. No manual polishing or post-processing of the representation has been used before qualitative analysis. If an explicit 3D reconstruction is needed for the desired application, the last stage of optimization can be replaced with SuGaR or 2DGS variants of Gaussian Splatting, extracting the desired mesh representation at the end of the optimization process.

4.2 Results

The proposed pipeline produces a high-accuracy photorealistic reconstruction of the experimental objects. The metrics used to assess the quality of the reconstruction are L1 error and PSNR between the image frames used in the generation and the actual rendering of the representation with the same camera parameters. All the results are schematized in the Table 4.1. Both experiments score higher than 27 in the PSNR measure and lower than 0.007 in the L1 error measure after the 30000-th optimization step, achieving results in pair with the 3DGS technique. In particular the chair experiments score 29.065 in the PSNR score and $6.014 \cdot 10^{-3}$ of L1 error, while the gym ball respectively scores 27.058 and $6.660 \cdot 10^{-3}$. Is important to notice that the results presented in this report are for masked objects, granting the metrics to be accurate over the region of interest instead of a mean measure of the whole scene, in opposition to the 3DGS original paper. Both trainings reached the last steps under the 8 minutes mark, with a total pipeline time cost below 12 minutes. Because it has been used a machine with the same GPU capacity as the one used in the paper, the time cost of optimization is comparable. The proposed pipeline computes the reconstructed representation with a speed up in the order of $2\times$ with respect to the baseline as it is also highlighted in Table 4.2 . Background regions are a big portion of the information present in the input image frames and preventing computation for these pixels grants the speedup.

The reconstruction under visual inspection presents some white artifacts around the borders of the object. There are two kinds of artifacts, one appears as a white color in Gaussian that belongs to the object but is visible only under some view angles, and the other is composed of completely separate white outlier Gaussians. The presence of these artifacts is due to the segmentation masking data. Both experiments with segmentation applied show the artifacts while the whole scene reconstruction does not. The explanation is that, during the pipeline, images are segmented by applying a flat white background around the object according to the extracted mask. Segmentation masks usually are not 100% accurate, especially around the border of the object. Even humans sometimes cannot agree if some pixels at the edge belong to the foreground or the background, so it is common practice to add a third special label in the ground truth of segmentation datasets representing unknown pixels. Neural Networks trained over these datasets inherently can misclassify some border pixels where ground truth is not accurate. Additionally, no temporal coherent masking algorithm is used, introducing noise in the segmentation mask of the object from different view angles. During Gaussian Splatting optimization some 3D points will be reprojected back to images and will be compared to areas presenting the white background. These points are correctly represented in some views while in others are excluded. The optimization pushes the reconstruction to address these special pixels to appear white (as the masked background) even though they are outside of the mask. In not-reported experiments where segmentation backgrounds were represented as black, the artifacts showed a black coloration instead of a white one. A controlled environment that achieves a clear distinction between the foreground and the background region in terms of color can help the segmentation step to reduce the noise introduced in the representation. If a full depth map without holes can be extracted from the sensing device, using the depth information could create a better segmentation mask by imposing thresholds in the depth space. An example of a noisy mask is depicted in Figure 4.5.

By visual inspection of the $30k$ step representation between the baseline and the proposed method, the accuracy of reconstruction in the latter case is sensibly better. The baseline chair object shows bad reconstruction in the thin structures with blurred edges. The gym ball does not present horizontal texture lines and the pedestal details are not accurate. The baseline visual results are reported in Figure 4.4.

The addition of a dense point cloud in the initialization step does increase the representation quality performances in the early stages of the optimization. From visual inspection, the point cloud generated with the addition of depth information describes better the object and it is denser over the surface of interest. Additionally, it reduces the number of outlier points that the COLMAP initial point cloud produces. In fact, even if segmentation masking is applied during the keypoints extraction step, the matching and reprojection of the points during global optimization lead to a large number of points identified in the background. The proposed depth enhanced point cloud is not free from outlier points but they are bounded by a depth threshold filter which filters out all the points further from the camera according to a threshold value. They also are derived only from noisy segmentation masks point clouds, being a small portion of the union of point clouds. A visual comparison of the point cloud generated is depicted in Figures 4.6. From visual inspection, it also emerges that PSNR is probably not the best metric for comparing the whole reconstruction but an additional metric should be used. In fact, by looking at the PSNR scores for gym ball experiments at 500 steps, the two values are almost similar, scoring 24.8933 for the proposed pipeline and 24.8926 for the one only with segmentation input. Looking at the same iteration in Figure 4.2 the difference is clearly visible instead.

About quantitative evaluation, the proposed method achieves fast and accurate reconstruction of the object. From the results reported in Table 4.1 and the relative plots depicted in Figure 4.3, it is appreciable the difference in the PSNR and L1 scores in contrast to the original initialization point cloud, scoring a 2 point increase in the PSNR at the 100th optimization step. In the subsequent steps, the discrepancy in the quality of the results thins out and the experiments converge to similar quality scores with some small negligible fluctuations. From visual inspection, the difference in the quality of the early results brought by the proposed technique is very evident and depicted in Figure 4.1 and Figure 4.2. The 30k step representation scores above 27 of PSNR score for both experiments quantitative demonstrating the accuracy of the reconstruction. In only 300 steps it reaches at least 23 of PSNR score with the proposed method. Although the scores are very similar at the 500 step reconstruction between the depth initialized method and the masked only one, visual inspection of the results shows a noticeable difference in the quality of reconstruction. This is probably because a metric more related to the geometric features of the reconstruction and less about the visual rendering should be used for better evaluation. No such metric has been used in this report in alignment with the original Gaussian Splatting paper.

By visual inspection, the distribution of the center point μ of the Gaussians is analyzed at the 12000-th steps and reported in Figure 4.7. There is a small but noticeable difference in the presence of Gaussians inside regions corresponding to large flat surfaces. The proposed method shows a more dense and better-distributed presence of Gaussians in these regions. A higher density of elements over surfaces can increase the ability of the model to behave correctly under distortions, for example during animation. In the context of a fast and more affordable reconstruction, the good performances of our model with only a few optimization steps can be useful in application scenarios where the computational power of the machine is low. In particular, the Gaussian Splatting technique heavily relies on the GPU parallelization capability during training for fast computation of the Jacobians and Neural Network gradients. In the case of a low GPU capacity, the 30k optimization steps can take hours to complete. Computing only 500 steps and still achieving a PSNR score of 24 can democratize the technique for such devices. The additional point cloud estimation using the depth map does not significantly slow down computation since it does not use GPU or parallel optimization techniques, adding an overhead time cost below $30s$ from reported results. Additionally, it provides further control over the resolution of the point cloud and the subsequent number of Gaussians of the representation with the voxel downsampling radius parameter.

The whole reconstruction time is reported in Table 4.2 showing the acquisition, preprocessing and Gaussian optimization times. A full 3D photorealistic reconstruction of a medium-sized object with the proposed pipeline takes less than $10min$ from start to end, creating a new setup for fast 3D reconstruction.

An explicit reconstruction mesh, if needed, can be extracted by using 2DGS technique. The chair object for example has been processed to extract its mesh with a 7k steps optimization and the visual result is reported in Figure 4.8. The reconstruction accurately represents the object even in hard regions such as thin holes and thin structures.

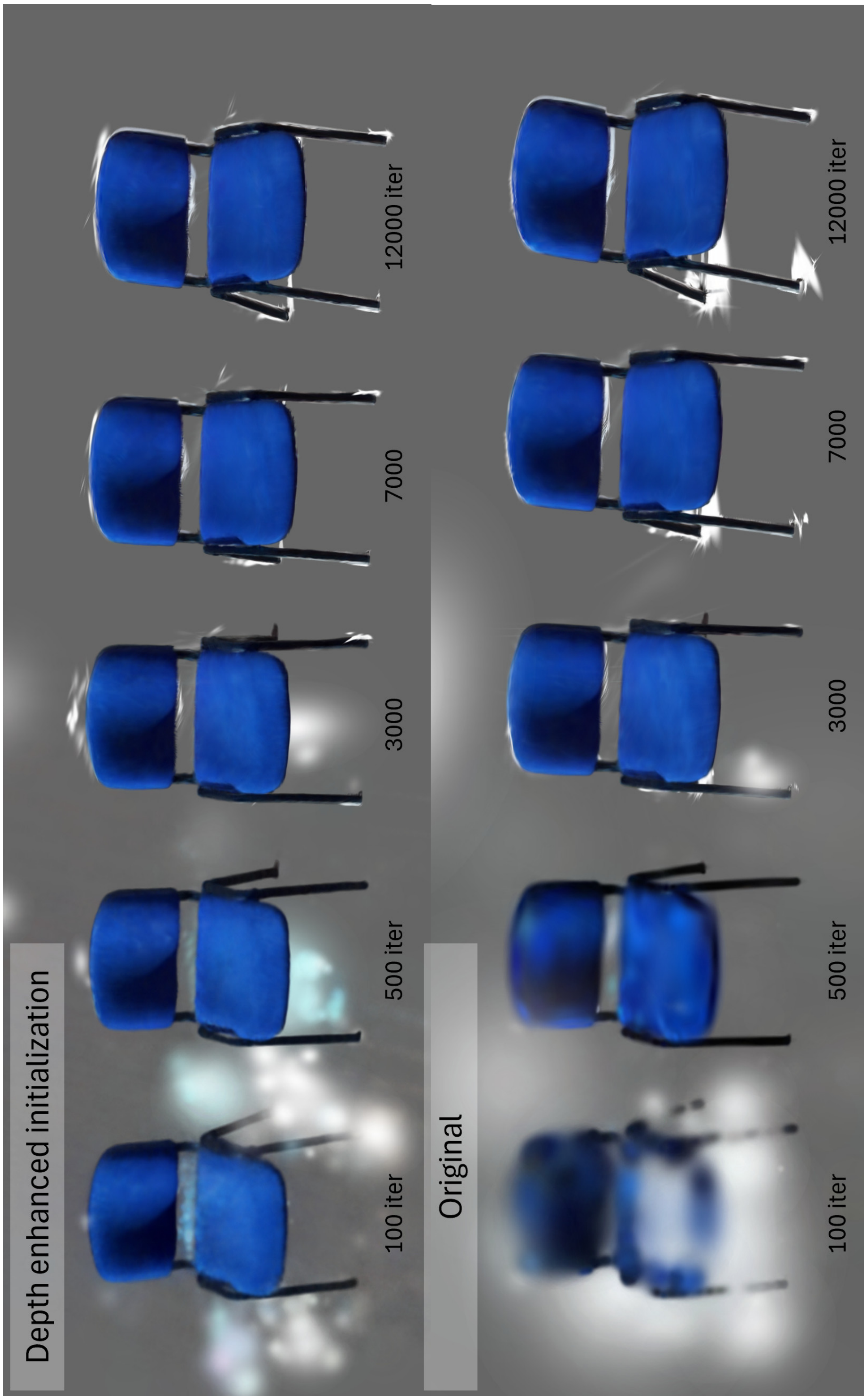

Figure 4.1: Visual comparison of Gaussian splatting with or without depth enhanced initialization, at different time steps for the chair experiment. At 100 and 500 steps the effect of the different initialization is visible with a higher accuracy in the reconstruction. Manual removing of some outlier Gaussians has been applied to better visualize the results in the first steps

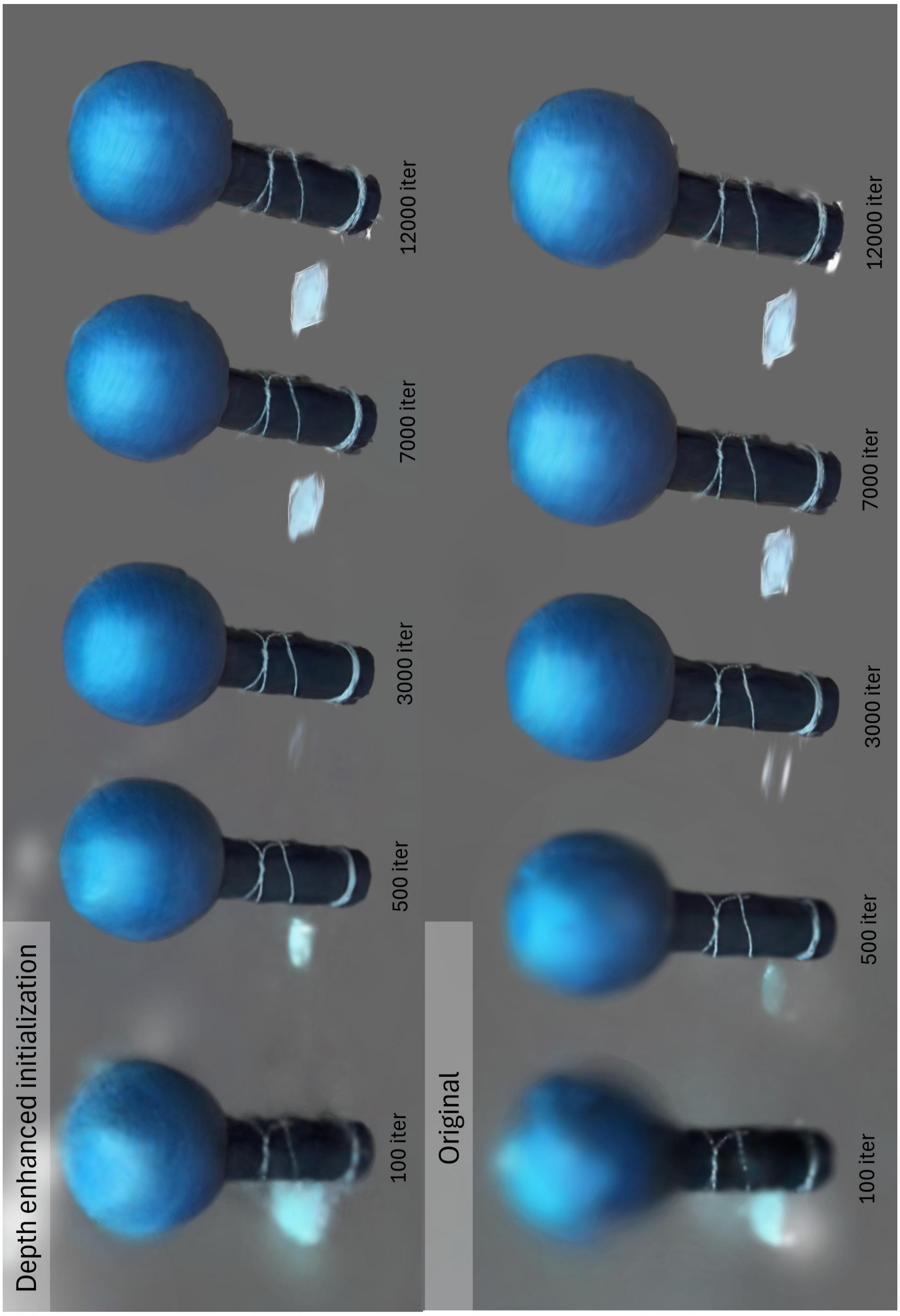

Figure 4.2: Visual comparison of Gaussian splatting with or without depth enhanced initialization, at different time steps for the gym ball experiment. At 100 and 500 steps the effect of the different initialization is visible with a higher accuracy in the reconstruction. Manual removing of some outlier Gaussians has been applied to better visualize the results in the first steps

Experiment	**Iterations**	**PSNR**	**L1**	**Elasped time**
Chair depth enhanced	100	17.6237	0.08943	$2s$
	300	23.3934	0.01679	$5s$
	500	23.9571	0.01187	$8s$
	1k	24.7086	0.00987	$14s$
	3k	25.1355	0.00850	$37s$
	7k	26.5785	0.00705	$1m30s$
	12k	27.6660	0.00601	$2m45s$
	30k	29.0652	0.00522	$7m06s$
Chair mask only	100	15.3352	0.09109	$2s$
	300	21.7330	0.02719	$6s$
	500	22.9820	0.01727	$10s$
	1k	24.6081	0.01057	$16s$
	3k	25.2012	0.00848	$39s$
	7k	26.7603	0.00690	$1m31s$
	12k	28.3657	0.00521	$2m49s$
	30k	29.9596	0.00417	$7m24s$
Gym ball depth enhanced	100	20.7613	0.05435	$1s$
	300	24.2537	0.01555	$4s$
	500	24.8933	0.0112	$7s$
	1k	25.2035	0.00863	$14s$
	3k	25.7515	0.00789	$44s$
	7k	26.0844	0.00759	$1m34s$
	12k	26.3997	0.00731	$2m30s$
	30k	27.0577	0.00666	$5m26s$
Gym ball mask only	100	18.9597	0.08921	$2s$
	300	24.7588	0.01534	$5s$
	500	24.8926	0.01116	$7s$
	1k	24.9556	0.00911	$12s$
	3k	25.6631	0.00798	$32s$
	7k	26.0918	0.00761	$1m19s$
	12k	26.3948	0.00729	$2m23s$
	30k	27.0730	0.00663	$5m18s$

Table 4.1: PSNR and L1 metrics during training of the experiments. The proposed pipeline presents increased metrics in the early steps, while it converges to similar results in the long run with respect to the pipeline using only segmentation masking

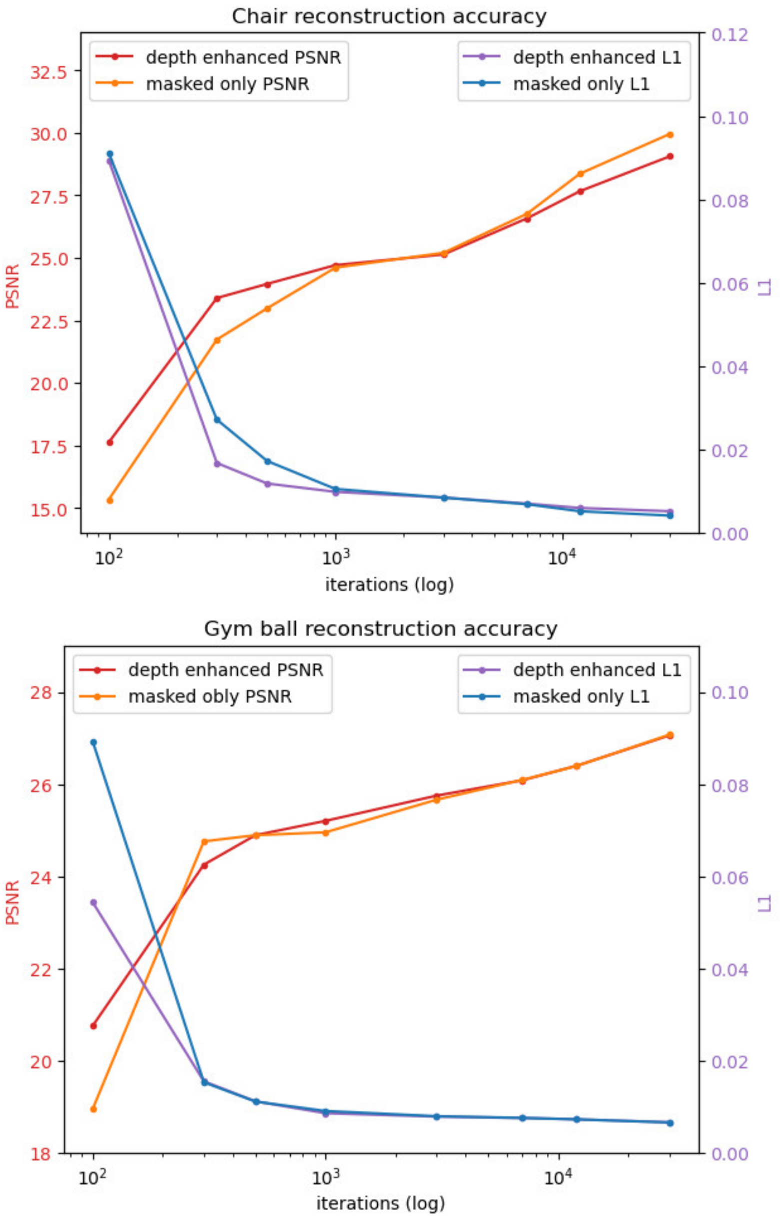

Figure 4.3: Plots of the reconstruction metrics reported for chair (top plot) and gym ball (bottom plot) experiments. In each plot, the measures are paired for a direct comparison. The boost in accuracy in the early steps is appreciable, while they present similar scores in the latest iterations.

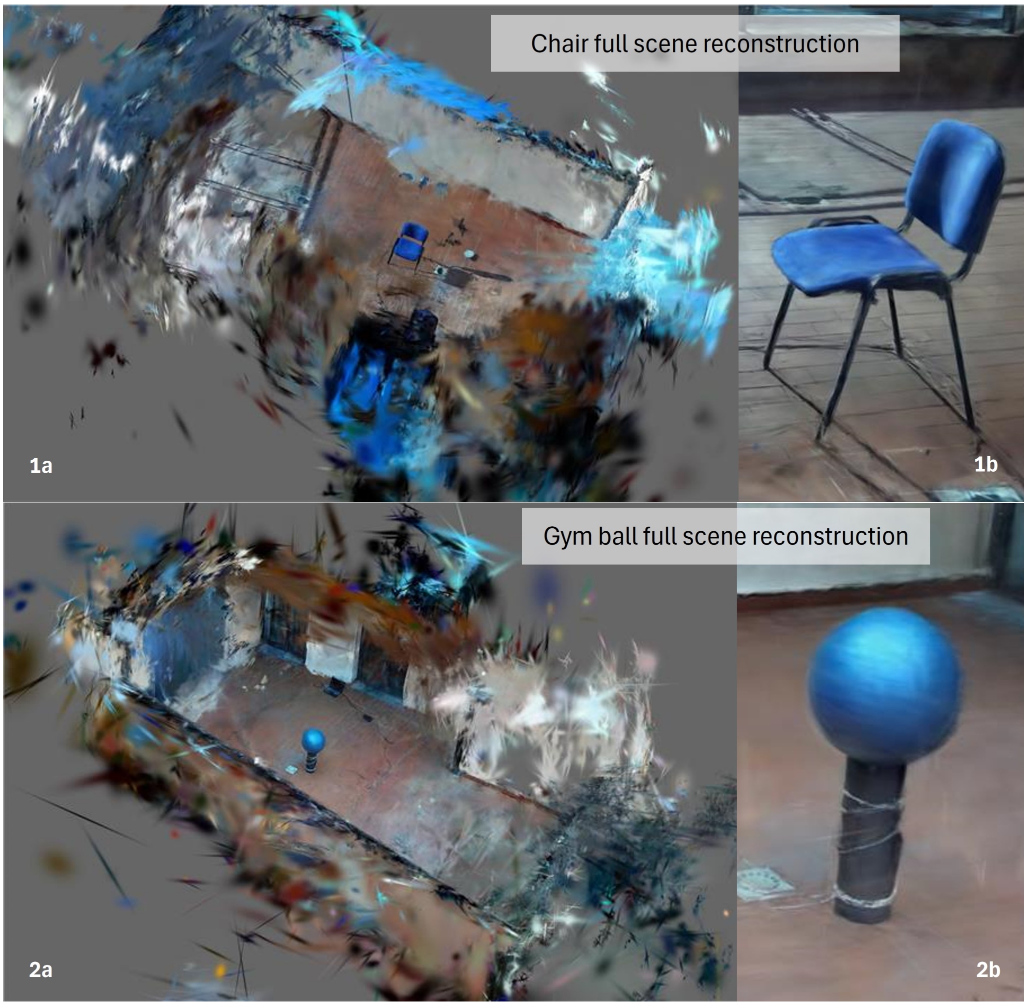
Chair full scene reconstruction
1a
1b
Gym ball full scene reconstruction
2a
2b

Figure 4.4: Visual results of baseline method over the two experiments at 30k iteration step. The whole scene is reconstructed with high accuracy (1a and 2a) at the cost of some accuracy of the target object (1b and 2b).

Experiment	**Acquisition time**	**Pre-processing and SfM time**	**Depth projection time**	**3DGS time**	**Total time**
Chair depth enhanced	35*s*	3*min* 30*s*	29*s*	7*min* 6*s*	11*min* 40*s*
Chair masked only	35*s*	3*min* 30*s*	–	7*min* 24*s*	11*min* 29*s*
Chair baseline	35*s*	1*min* 26*s*	–	14*min* 14*s*	16*min* 15*s*
Gym ball depth enhanced	1*min* 12*s*	3*min* 55*s*	20*s*	5*min* 26*s*	11*min* 3*s*
Gym ball masked only	1*min* 12*s*	3*min* 55*s*	–	5*min* 18*s*	10*min* 25*s*
Gym ball baseline	1*min* 19*s*	1*min* 19*s*	–	12*min* 6*s*	14*min* 36*s*

Table 4.2: Recorder time cost of the different pipelines. The segmentation pipeline has a great effect in increasing the performances with respect to the original Gaussian splatting baseline, with a speedup of 2×. No appreciable time difference is introduced in the optimization step as effect of the depth projection step. The 3DGS time costs reported refer to the 30k step reconstruction for all presented pipelines

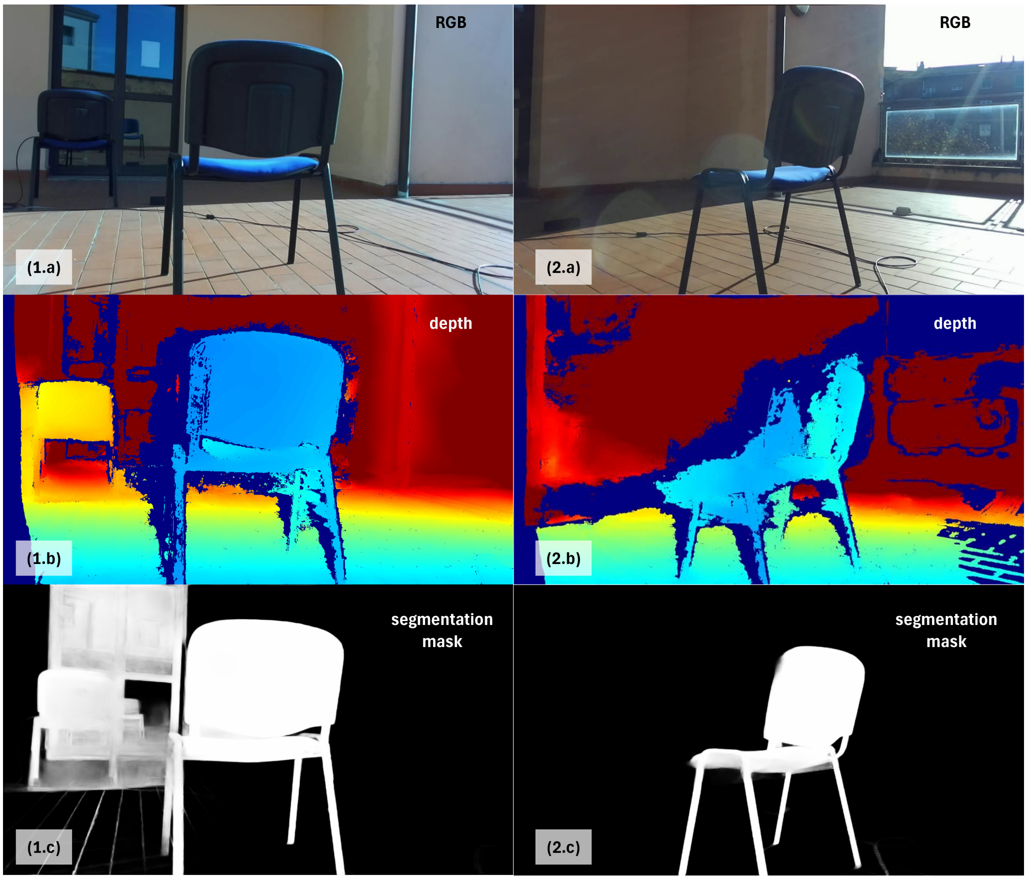

Figure 4.5: Example input views and masks of the chair experiments. The RGB images (1.a and 2.a) are recorded under natural outside light conditions. The depth values (1.b and 2.b) are recorded from a stereo camera setup and show holes and noise. Segmentation masks are predicted using NN models. In 1.c a noisy mask is reported showing additional background regions considered, while in 2.c a more accurate mask is reported

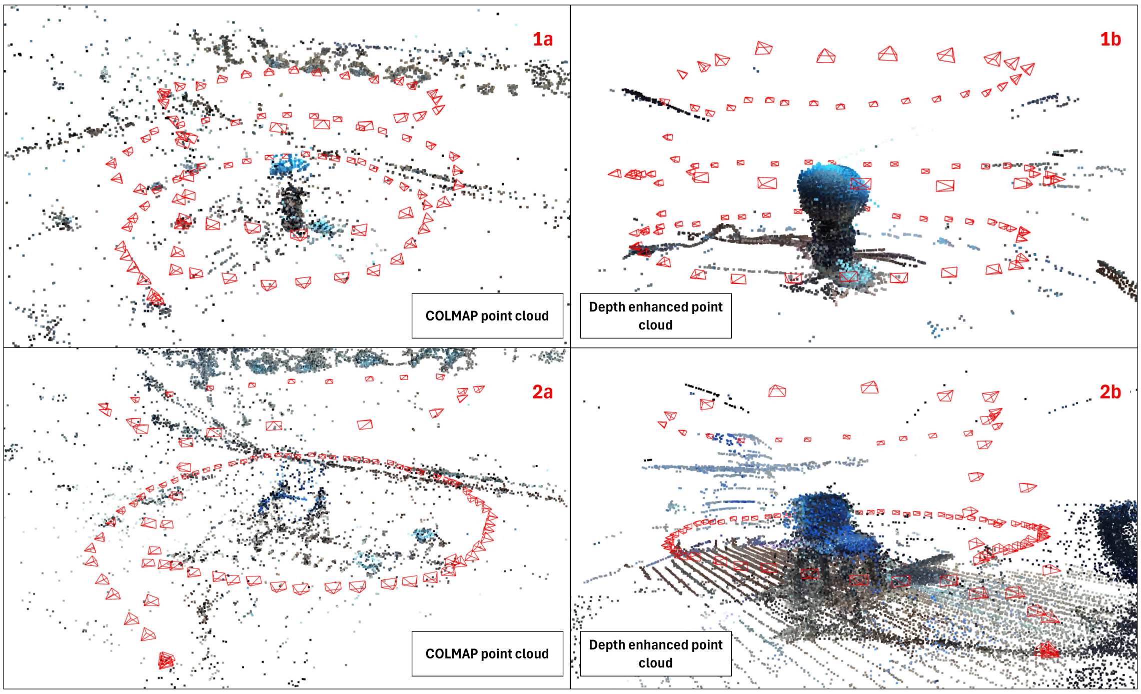

Figure 4.6: Visual comparison of point clouds with or without the depth enhancing step. From visual comparison, in the chair example (2a and 2b) the point cloud extracted from COLMAP shows holes in the structure of the chair in the center. The same happens with the ball reconstruction (1a and 1b), where only the top part of the point cloud is present in the COLMAP reconstruction. For both examples, the depth enhanced version better approximates the target object.

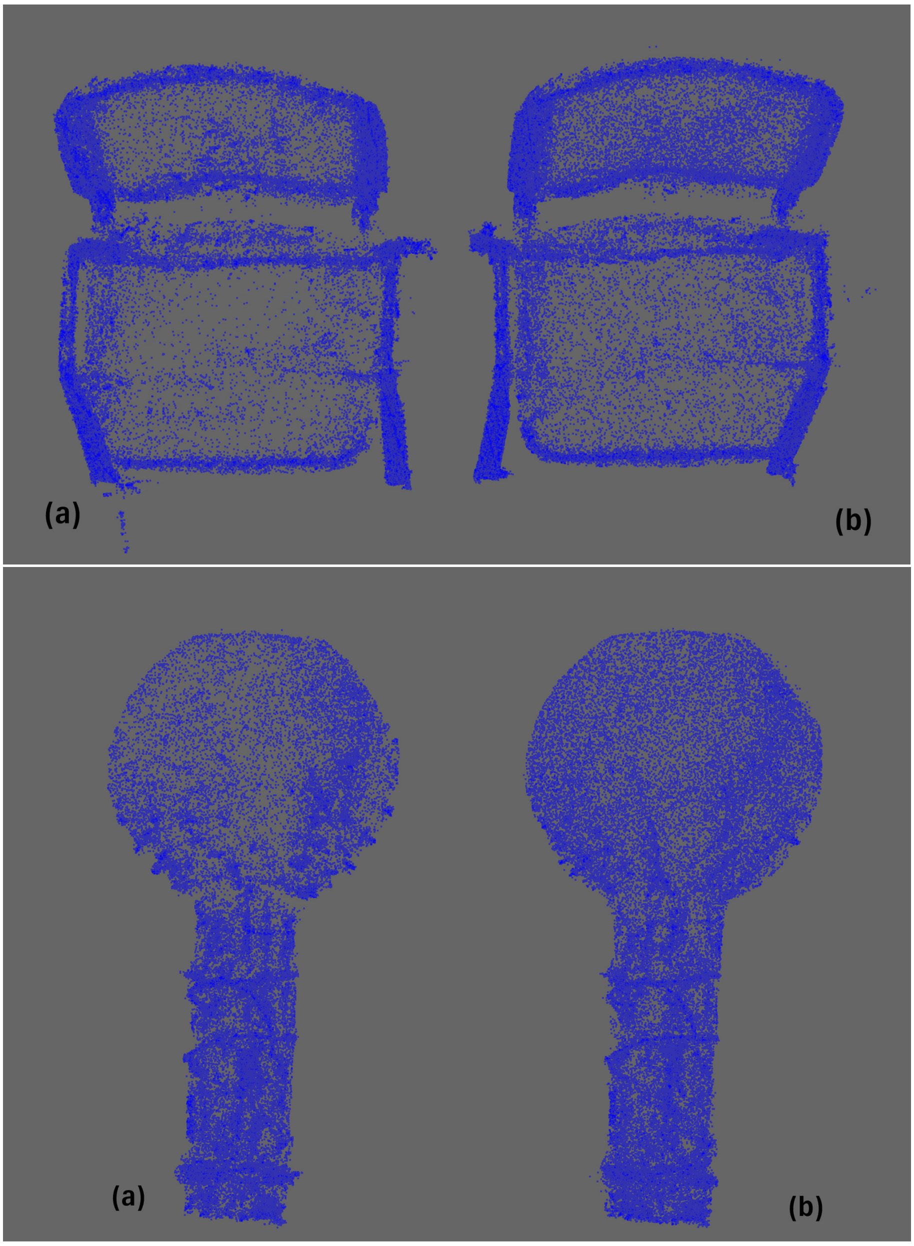

Figure 4.7: Visual comparison across the distribution of centers of Gaussian in the 30k steps reconstruction. In the top image, the chair experiment is reported, while in the bottom image the gym ball one. The left side (a) for both experiments reports the masked only pipeline results showing sparser Gaussians in flat uniform regions with respect to the depth enhanced initialization ones (b)

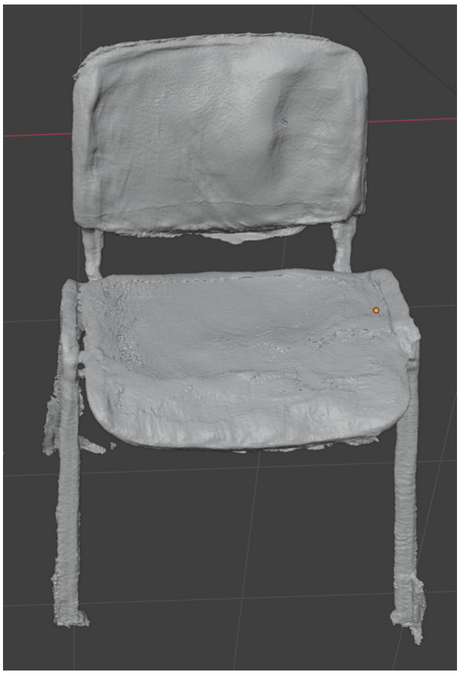

Figure 4.8: Reconstructed mesh using 2DGS technique applying segmentation masking and depth enhanced optimization. The mesh is extracted from the 7000 step snapshot without any further manual processing. The geometry of the object is reconstructed with accuracy even in holes and thin structures. The texture of the surface still presents some artifacts and imperfections.

Chapter 5

Conclusion

The presented work represents a significant study of the capabilities of Gaussian Splatting techniques. Thanks to the proposed pipeline one can achieve photorealistic and ready-to-use models from a simple acquisition pipeline with a depth and color-sensing camera. The proposed pipeline grants flexibility and modularity, allowing future improvement to single stages of the reconstruction. The technique and the proposed methods are not yet affordable for all users, still requiring a depth camera and and high computational power machine for post-processing, but it makes a step toward democratizing the technology. Acceptable reconstructions are derived with only a few hundred steps of optimization, reducing the GPU capacity needed while maintaining acceptable times. The segmentation masking of the object increases the final polish-ness of the reconstruction removing unwanted background and creating a representation almost ready to use. Some small artifacts appear in the final representation due to noisy inputs, but they can be manually removed. Expanding Gaussian Splatting usage and applications is key in the probable transition from explicit mesh representation to implicit ones in some areas of industry. With respect NeRF technique Gaussian Splatting emerges as a more friendly representation, working with more easily understandable Gaussian primitives and scoring high accuracy in the reconstruction even in hard regions such as thin structures.

5.1 Future Works

During the work, some areas of improvement have been individuated, as well as alternative approaches to some of the stages. First, the depth values acquired from the sensing camera are used only in the initialization of the Gaussian Splatting optimization. An additional step could be the integration of the depth information inside the loss and gradient computation during actual optimization steps, allowing in theory for a faster and more accurate convergence of the representation. To achieve a more polished reconstruction and remove the artifacts, a final post-processing stage could be added at the end of the pipeline. Techniques such as outliers removal or good parametrized thresholds could be implemented, although no post-processing directly over Gaussians splatting has yet been implemented in literature. As an alternative to reduce the artifact appearance, better masks with less noise could be extracted from the frames removing the regions during optimization where gradients conflict across different frames. To make the reconstruction process more affordable, simple RGB cameras from smartphones could be used and the depth map could be generated in the post-processing computation. A key requirement would be the temporal coherence of the depth values generated and the retaining of the real measures of the scene for a correct reconstruction. Generated depth maps do not present holes or noisy areas from real sensors making it suitable for better processing of the depth information. Still, the realism and coherence of values across frames are required. Since the method proposed achieves good reconstruction within a few hundred steps, it could be implemented in low computational capacity devices. The actual implementation will bring challenges about the segmentation mask extraction or the increase in time costs of the actual optimization steps but in theory, the full reconstruction could be carried out in an acceptable time, thanks to the depth-enhanced initialization which does not require parallelization and gives a boost in the reconstruction in the early steps.

Bibliography

[1] Sameer Agarwal et al. “Building rome in a day”. In: *Communications of the ACM* 54.10 (2011), pp. 105–112.

[2] Juan-José Aguilar, F Torres, and MA Lope. “Stereo vision for 3D measurement: accuracy analysis, calibration and industrial applications”. In: *Measurement* 18.4 (1996), pp. 193–200.

[3] Cihan Altuntas. “Review of Scanning and Pixel Array-Based LiDAR Point-Cloud Measurement Techniques to Capture 3D Shape or Motion”. In: *Applied Sciences* 13.11 (2023), p. 6488.

[4] Duane Brown. “Decentering distortion of lenses”. In: *Photogrammetric engineering* 32.3 (1996), pp. 444–462.

[5] Berk Calli et al. “The ycb object and model set: Towards common benchmarks for manipulation research”. In: *2015 international conference on advanced robotics (ICAR)*. IEEE. 2015, pp. 510–517.

[6] Yen-Chi Cheng et al. “Sdfusion: Multimodal 3d shape completion, reconstruction, and generation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 4456–4465.

[7] Artec Europe. *Artec 3D Portable Scanners*. `https://www.artec3d.com`. Access date: 11-20-2024. 2024.

[8] Luciano Pavarotti Foundation. *Luciano Pavarotti Foundation Website*. `https://www.lucianopavarottifoundation.com/en/`. Access date: 11-20-2024. 2024.

[9] Stephan J Garbin et al. “Fastnerf: High-fidelity neural rendering at 200fps”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 14346–14355.

[10] Daniel Gatis. *rembg*. `https://github.com/danielgatis/rembg`. 2024.

[11] Leonardo Gomes, Olga Regina Pereira Bellon, and Luciano Silva. “3D reconstruction methods for digital preservation of cultural heritage: A survey”. In: *Pattern Recognition Letters* 50 (2014), pp. 3–14.

[12] Antoine Guédon and Vincent Lepetit. “Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 5354–5363.

[13] Binbin Huang et al. “2d gaussian splatting for geometrically accurate radiance fields”. In: *ACM SIGGRAPH 2024 Conference Papers*. 2024, pp. 1–11.

[14] Antonios Kargas, Nikoletta Karitsioti, and Georgios Loumos. “Reinventing museums in 21st century: Implementing augmented reality and virtual reality technologies alongside social Media’s logics”. In: *Virtual and augmented reality in education, art, and museums*. IGI Global, 2020, pp. 117–138.

[15] Bernhard Kerbl et al. “3D Gaussian Splatting for Real-Time Radiance Field Rendering.” In: *ACM Trans. Graph.* 42.4 (2023), pp. 139–1.

[16] Zhe Li et al. “Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 19711–19722.

[17] Chen-Hsuan Lin et al. “Magic3d: High-resolution text-to-3d content creation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 300–309.

[18] David G Lowe. “Object recognition from local scale-invariant features”. In: *Proceedings of the seventh IEEE international conference on computer vision*. Vol. 2. Ieee. 1999, pp. 1150–1157.

[19] Radomír Mendřickỳ. “Impact of applied anti-reflective material on accuracy of optical 3D digitisation”. In: *Materials science forum*. Vol. 919. Trans Tech Publ. 2018, pp. 335–344.

[20] Ben Mildenhall et al. “Nerf: Representing scenes as neural radiance fields for view synthesis”. In: *Communications of the ACM* 65.1 (2021), pp. 99–106.

[21] Paritosh Mittal et al. "Autosdf: Shape priors for 3d completion, reconstruction and generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 306–315.

[22] Niantic. *Scaniverse - 3D Scanner*. `https://scaniverse.com/`. Access date: 11-20-2024. 2024.

[23] David Nistér. "An efficient solution to the five-point relative pose problem". In: *IEEE transactions on pattern analysis and machine intelligence* 26.6 (2004), pp. 756–770.

[24] Polycam. *Polycam: 3D Scanner & Editor*. `https://poly.cam/`. Access date: 11-20-2024. 2024.

[25] Ben Poole et al. "Dreamfusion: Text-to-3d using 2d diffusion". In: *arXiv preprint arXiv:2209.14988* (2022).

[26] Xiaojuan Qi et al. "Structural dynamic deflection measurement with range cameras". In: *The Photogrammetric Record* 29.145 (2014), pp. 89–107.

[27] Guocheng Qian et al. "Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors". In: *arXiv preprint arXiv:2306.17843* (2023).

[28] Xuebin Qin et al. "Highly accurate dichotomous image segmentation". In: *European Conference on Computer Vision*. Springer. 2022, pp. 38–56.

[29] Viktor Rudnev et al. "Nerf for outdoor scene relighting". In: *European Conference on Computer Vision*. Springer. 2022, pp. 615–631.

[30] Mriganka Sarmah, Arambam Neelima, and Heisnam Rohen Singh. "Survey of methods and principles in three-dimensional reconstruction from two-dimensional medical images". In: *Visual computing for industry, biomedicine, and art* 6.1 (2023), p. 15.

[31] Johannes Lutz Schönberger and Jan-Michael Frahm. "Structure-from-Motion Revisited". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[32] Johannes Lutz Schönberger et al. "Pixelwise View Selection for Unstructured Multi-View Stereo". In: *European Conference on Computer Vision (ECCV)*. 2016.

[33] Kevin V Stefanik et al. “UAV-based stereo vision for rapid aerial terrain mapping”. In: *GIScience & Remote Sensing* 48.1 (2011), pp. 24–49.

[34] StereoLabs. *StereoLabs — AI perception for automation*. `https://www.stereolabs.com/en-it`. Access date: 11-20-2024. 2024.

[35] Peter Sturm and Bill Triggs. “A factorization based algorithm for multi-image projective structure and motion”. In: *Computer Vision—ECCV’96: 4th European Conference on Computer Vision Cambridge, UK, April 15–18, 1996 Proceedings Volume II 4*. Springer. 1996, pp. 709–720.

[36] Jiaxiang Tang et al. “Dreamgaussian: Generative gaussian splatting for efficient 3d content creation”. In: *arXiv preprint arXiv:2309.16653* (2023).

[37] Sebastian Thrun et al. “Robotic mapping: A survey”. In: (2002).

[38] V Uffenkamp. “State of the art of high precision industrial photogrammetry”. In: *Third international workshop on accelerator alignment*. Springer Berlin/Heidelberg, Germany. 1993.

[39] Xue Wang and Peijun Li. “Extraction of urban building damage using spectral, height and corner information from VHR satellite images and airborne LiDAR data”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 159 (2020), pp. 322–336.

[40] Michael Weinmann et al. “A multi-camera, multi-projector super-resolution framework for structured light”. In: *2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission*. IEEE. 2011, pp. 397–404.

[41] Jing Xu et al. “Rapid 3D surface profile measurement of industrial parts using two-level structured light patterns”. In: *Optics and Lasers in Engineering* 49.7 (2011), pp. 907–914.

[42] Yihao Zhi et al. “Dual-space nerf: Learning animatable avatars and scene lighting in separate spaces”. In: *2022 International Conference on 3D Vision (3DV)*. IEEE. 2022, pp. 1–10.