D3.1 - State-of-the-art of 3D heritage tools and methodologies

Deliverable Report n. 3.1: final version, issue date on 31/07/2025

Grant Agreement number:	101195149
Project acronym:	3D-4CH
Project title:	Online competence centre in 3D for Cultural Heritage
Funding programme:	Digital Europe
Project coordinator:	Marco Medici, INCEPTION
E-mail:	marco.medici@inceptionspinoff.com
Project website address:	www.3d4ch-competencecentre.eu

Title:	D3.1 - State-of-the-art of 3D heritage tools and methodologies
Issue Date:	31/07/2025
Produced by:	FBK
Main author:	Elisa Mariarosaria Farella (FBK), Fabio Remondino (FBK)
Co-authors:	Fotios Arnaoutoglou (AthenaRC), Anestis Koutsoudis (AthenaRC), Vangelis Nomikos (TALENT), Anthony Corns (DISC), Andrea Sterpin (INCEPTION), Iana Boitsova (Pixelated Reality), Fedir Boitsov (Pixelated Reality), Emanuel Demetrescu (CNR), Rafael J.Segura (UNIVERSITY OF JAEN), Antonio J. Rueda (UNIVERSITY OF JAEN)
Version:	v1.0
Reviewed by:	Marco Medici (INCEPTION), Matevz Straus (ARCTUR), Valentina Vassallo (CYI)
Approved by:	Marco Medici, INCEPTION
Dissemination:	Public

Colophon

3D-4CH is a Digital Europe project co-funded by the European Union under Grant Agreement n. 101195149.

Co-funded by the European Union

Document History

- 27.02.2025: First draft version (v0.1) with tentative table of contents
- 10.05.2025: Draft version (v0.3) with definition of the main section of the report
- 24.06.2025: Advanced draft verision (v0.6)
- 15.07.2025: Final version ready for review (v0.9)
- 31.07.2025: Final version ready for submission (v1.0)

Abbreviations and Acronyms

3D-4CH	Online Competence Centre in 3D for Cultural Heritage
AI	Artificial Intelligence
API	Application Programming Interface
AR	Augmented Reality
BIM	Building Information Modeling
CH	Cultural Heritage
CHIs	Cultural Heritage Institutions
DMP	Data Management Plan
EDM	Europeana Data Model
GCP	Ground Control Point
GSD	Ground Sample Distance
ICP	Iterative Closest Point
LiDAR	Light Detection and Ranging
MDE	Monocular Depth Estimation
MR	Mixed Reality
NeRF	Neural Radiance Fields
PBR	Physically Based Rendering
SfM	Structure from Motion
TLS	Terrestrial Laser Scanning
ToF	Time of Flight
UDIM	U-Dimension (advanced texturing system)
USD/OpenUSD	Universal Scene Description
UV	Texture coordinate system (U,V)

VR	Virtual Reality
XR	Extended Reality
OBJ	Wavefront OBJ 3D file format
FBX	Autodesk Filmbox 3D file format
DAE	COLLADA Digital Asset Exchange format
PLY	Polygon File Format / Stanford Triangle Format
STL	Stereolithography 3D file format
LAS/LAZ	LiDAR point cloud file formats (compressed/uncompressed)
PTX/PTS	Point cloud exchange formats
XYZ	Simple ASCII point cloud file format

Table of Content

Executive Summary

This document is a comprehensive review of established and emerging technologies and tools for 3D content creation and optimisation, and enhanced data fruition, with a focus on their applicability and significance in the Cultural Heritage (CH) sector. The deliverable framework is intended to serve as a guide to support CH professionals in their daily practice, particularly in the selection and comparison of existing solutions for the documentation, analysis, visualisation, and dissemination of CH content.

After providing a brief background and terminology overview (Section 1), the deliverable addresses:

- 3D content creation with more traditional range and image-based techniques, as well as emerging AI-based solutions (Section 2). These complementary or alternative methods are discussed in terms of principles, workflows, advantages, and limitations.
- 3D data post-processing, with some hints on co-registration techniques, and data optimisation, also AI-based (Section 3).
- Types of 3D data and formats (Section 4).
- Data fruition and visualisation through the latest Extended Reality (XR) platforms, increasingly used as educational tools and for creating immersive storytelling or virtual exhibitions (Section 5).
- Some clarifications on the re-assessment plan related to the automated translation of metadata (Section 6).
- Relevant tools and frameworks available for 3D data creation, editing, and XR-based experiences (Section 7).

1. Introduction

1.1 Background and aim

Digital transformation has stimulated significant progress and innovation in the CH domain over the past decades. The main advancements have involved heritage digitisation procedures, digital data processing, analysis, and data editing, as well as the introduction of innovative ways of presenting and valorising CH assets.

Digital technologies and tools are unavoidable in the work of many CH professionals, and their continuous and rapid evolution is increasingly tied to the sector's needs and demands. This link between professional needs and more and more evolved technologies is crucial for ensuring the real impact of these solutions on advancing knowledge in this sector.

At the same time, the CH specialists' needs are adapting to face new challenges in heritage preservation, from the automated processing and management of larger and more complex datasets to the growing demand for increased digital asset accessibility and interoperability.

Over the last decades, the integration of geomatic techniques in the CH sector, in particular, has deeply transformed heritage documentation practices, supplementing or replacing traditional recording methods. These techniques can be generally classified into active (ranges) and passive (images) categories, relying on different capturing sensors depending on the working scale.

The introduction of 3D reality-based surveying and modelling techniques has expanded data capturing and representing capabilities, enabling the derivation of more complete and complex information, and overcoming the limits of traditional approaches typically focusing on the acquisition of some basic physical and dimensional properties of heritage assets (e.g., richer texture and surface details).

The advent of 3D digital replicas has in-depth impacted the work of heritage conservators, architects, archaeologists, and curators, offering new powerful tools supporting more accurate condition assessment and conservation planning, finer interpretation and object analysis, and enabling high-fidelity visualisations and immersive experiences for increasing public awareness and engagement in the heritage sector.

Beyond documentation, and in a context where heritage is recognised as more and more at risk for conflicts, climate change, and time-related degradation, 3D digital models have created new opportunities for enabling remote and virtual access to distant or inaccessible heritage, and enlarging public participation and awareness through the advancement of Extended Reality (XR) solutions. The rise of these technologies - including Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR) - has opened unprecedented opportunities, especially in the education and tourism sectors, for enhancing heritage knowledge, understanding, and raising public awareness and participation in its preservation.

In parallel with the advancements of documentation and fruition solutions, the impressive and rapid growth of Artificial Intelligence (AI) algorithms has recently marked a further transformation within the heritage sector. The innovation driven by these techniques involves mainly automatic 3D content creation, data quality enhancement, and advanced interpretation and analysis of captured datasets. While these emerging solutions are increasingly promising for solving complex tasks in digital data processing, their generalisation and scalability still represent a significant challenge, as well as, in many cases, the interpretation, accuracy, and reliability of derived products.

Given the complexity of these evolving technological systems, the plethora of digital processing workflows, and the varying levels of maturity of available solutions, there is an emerging demand for clear guidance for the effective adoption and use of tools and methodologies currently available for the working practice of CH professionals.

This deliverable aims to offer a structured overview of key technologies and tools in the areas of 3D data creation, processing, and optimisation, as well as available solutions for enhancing and enlarging heritage access, understanding, and fruition. One of the desired outcomes is to help find a clear picture of how the data of digitised CH can be processed for exchange and integration through online platforms and APIs, i.e, for aggregation into the Data Space for Cultural Heritage[1].

[1] https://www.dataspace-culturalheritage.eu/en

1.2 Definitions - Terminology

This section provides some specifications on key terminology and technical parameters used throughout the document.

1.2.1 Reality-based 3D surveying and modelling

Active - Range-based techniques:

- *Targets*: artificial typically spherical, checkerboards, retro-reflective, or coded objects placed on or around the scanned object to support the processing and co-registration.
- *Triangulation*: systems suitable for small-scale and short-range applications. The 3D position of the points is derived by measuring the angles from two known positions, exploiting a laser emitter and a camera/detector.
- *Time-of-Flight (ToF)*: ideal for medium and long-range applications (terrestrial or airborne LiDAR systems), these systems rely on time measurements, i.e., the delay between transmitted and returned laser signals to determine distances. This delay can be measured:
    - directly, through short laser pulses (Pulsed Wave);
    - indirectly, by considering the phase shift of a continuous wave (Phase Shift).

Passive - Image-based techniques

- *Color targets:* standardised color reference charts positioned within the scene during the image acquisition and used for color calibration and correction.
- *Depth of field:* the range (distance between the nearest and farthest points) within the captured scene adequately sharp in the image and in focus.
- *Keypoints:* measured homologous points in the images. They identify the same object points and should be well distributed across the image to ensure a strong connection between images.
- *Image scale*: is the number defined by the ratio between the object distance (camera to object) and the principal distance (focal length).
- *Image overlap*: common/shared region captured by adjacent images.
- *Image spatial resolution*: is the smallest detail detectable by an imaging system.
- *Focal length*: the distance (millimeters) between the lens optical center and the image sensor, focusing at infinity.
- *Field of view*: the extent of the scene captured by a camera expressed as an angle. It is inversely proportional to the focal length.
- *GNSS data:* satellite-derived positioning information (geographic coordinates, elevation, timing).
- *Ground Control Points*: measured reference points with known 3D coordinates used to scale and georeference the model (absolute positional information in real-world coordinate systems).
- *GSD (Ground Sample Distance)*: is the ground distance (object space) corresponding to a single pixel in the image (image space). Higher is the GSD, lower is the spatial resolution of the image with fewer object details captured.
- *Ray intersection geometry*: the geometric condition describing the connection between two or more perspective rays (each connecting the camera projection center to an image point) converging at a unique 3D object point.
- *Optical distortions*: deviations from the ideal central perspective model (image geometry) due to imaging errors and limitations in the camera lens system.

Reality-based reconstruction products

- *Depth map:* a 2D image where each pixel represents the distance of each pixel from the camera. It is often visualized in false color and is normally produced through Dense Stereo Matching or depth sensors.
- *Dense point cloud:* a dense set of 3D points representing the scene and generated by dense stereo matching or multi-view stereo algorithms.
- *Dense Stereo Matching*: it computes pixel-wise depth by matching homologous points in stereo images.

- *Disparity*: difference in the position of corresponding pixel between images (inversely proportional to the depth).
- *Multi-view-stereo*: algorithms enabling the dense scene reconstruction by matching information from multiple overlapping images captured from different viewpoints. They estimate depth or disparity for pixels across images to generate dense point clouds.
- *Sparse 3D point cloud reconstruction:* a limited set of points generated by matched keypoints identified across multiple overlapping images.
- *Structured point cloud*: set of 3D points organized in a defined grid or pattern and with a clear and consistent relationship between points (typically generated by some LiDAR scanners, structured light sensors, and depth cameras).
- *Unstructured point cloud*: unordered and irregular set of points with different spatial density generated by most of the sensing methods (e.g., laser scanners and photogrammetry-based).

1.2.2 Quality measures and parameters

- *Accuracy:* the closeness of a measurement compared with a standard or reference value.
- *Artifacts (point clouds)*: systematic deviations or defects like discontinuities or duplicates due to processing and reconstruction errors.
- *Image sharpness:* it describes the clarity and detail of an image, and it is influenced by several factors, including the resolution, contrast, lens quality, focus accuracy, motion blur.
- *Measurement error:* it describes the deviation of the measure from a reference value. It includes random and systematic errors (that could be corrected if known).
- *Measurement uncertainty*: it defines the range within which the true value of a measure lies. It comprises all the unknown systematic and random errors.
- *Noise (point clouds)*: random deviations in the 3D data from the actual shape or surface of the object being reconstructed.
- *Precision:* it defines the spread of the measurements of a set of repeated measurements or an adjustment process (relative accuracy).
- *Reference value*: commonly used in working practice to compare measurements and estimate their quality. This value is measured with a system of higher order of accuracy and a smaller (5-10 times) uncertainty.
- *Resolution:* it describes the smallest change in the quantity to be measured and that the instrument can detect or display.
- *Sampling resolution*: the minimum distance between two consecutive measurements

1.2.3 Further definitions

- *Aliasing artifacts* (rendering): visual distorsions occurring in a scene when fine details are undersampled during rendering or NeRF training.
- *Baking (textures):* a technique to achieve photorealistic rendering by precomputing the lighting of a scene and storing the results as a texture map applied to 3D models.
- *Grid of UV tiles:* multiple square tiles generated by dividing the UV space and allowing the application of multiple textures (by assigning different UV islands to different tiles).
- *High-poly model:* 3D models with a large number of polygons and faces and highly-detailed surface representations.
- *Low-poly model*: 3D models with a relatively small number of faces, commonly used and efficient for real-time, VR, or games applications.
- *Octree representation:* a hierarchical data structure recursively dividing the 3D space into eight cubic regions (octans).
- *Physically Based Rendering (PBR)*: rendering technique used in computer graphics to simulate the flow of light, model the physical properties of materials, and produce photorealistic models under varying lighting conditions.
- *Radiance field*: a function describing and encoding the amount of light (radiance) passing through a point in the 3D space.
- *Ray tracing*: a rendering technique simulating the path of light rays passing through a 3D scene.
- *Specialised-augmented point clouds:* point clouds where each point is described not only by its coordinates but is augmented with additional attributes and features.
- *Volumetric representation:* a method to represent the scene as a continuous volume, storing

information such as color, density, and light properties at every point in the space.
- *Voxel representation:* a method to divide the 3D scene into a regular grid of cubic units called voxels, storing information about the scene's properties (such as color, radiance, or density).
- *Unwrapping*: the process of flattening a 3D model surface into a 2D UV map for the accurate texture mapping.
- *UV islands*: clusters of faces, representing separate and contiguous regions of a UV map and corresponding to connected areas of the 3D model's surface.
- *UV map*: 2D representation of the 3D model surface generated by the unwrapping. Each vertex of the 3D mesh is assigned coordinates (U,V) on a flat plane.
- *UV space:* the normalised coordinate system of the UV map.
- *Texture reprojection*: the process of transferring the textures from one view to another view, model, or UV map.

1.2.4 Source-based techniques

- *BIM-based modeling:* Building Information Modeling approach for comprehensive architectural reconstruction that integrates geometric representation with semantic information and metadata management.
- *Documentary sources:* archival materials including historical drawings, paintings, photographs, textual descriptions, archaeological reports, architectural plans, and other documentary evidence used as primary input for 3D reconstruction.
- *Interpretative modeling:* 3D modeling approach that combines multiple information sources to reconstruct hypothetical but scientifically grounded representations, requiring transparent documentation of interpretative choices.
- *NURBS modeling:* Non-Uniform Rational B-Splines technique for creating smooth, mathematically defined surfaces particularly suitable for architectural elements and complex curved geometries.
- *Paradata*: documentation that describes the process of interpretation and the reasoning behind decisions made during virtual reconstruction, ensuring methodological transparency and scientific rigor.
- *Parametric modeling:* modeling technique that creates flexible 3D models controlled by parameters, allowing systematic exploration of design variations and alternative reconstruction hypotheses.
- *Polygonal modelin*g: technique for creating 3D geometry through manual construction and manipulation of vertices, edges, and faces, commonly used in heritage reconstruction software.
- *Procedural modeling:* rule-based modeling approach using algorithms and shape grammars for automated or semi-automated generation of architectural elements based on defined parameters and constraints.
- *Scientific transparency*: requirement for clear documentation of sources, methods, interpretative choices, and uncertainty levels in virtual reconstruction projects to maintain scientific credibility.
- *Shape grammar:* formal rule-based system for generating complex architectural forms from simpler components, particularly useful for reconstructing repetitive architectural elements and stylistic patterns.
- *Source-based modeling:* 3D model creation through the interpretation and integration of documentary sources (drawings, photographs, texts, archaeological reports) rather than direct reality capture. The accuracy depends on source reliability and interpretative methodology rather than metric precision.
- *Uncertainty representation*: visual or textual indication of different confidence levels in various components of a virtual reconstruction, distinguishing between evidence-based and hypothetical elements.
- *Virtual reconstruction:* digital recreation of heritage contexts that no longer exist or are inaccessible, based on interpretation of available sources and scientific hypotheses rather than direct measurement.

2. 3D content creation (reality-based)

With the advancements of sensors and methodologies over the last decades, 3D digitisation with reality-based techniques has become a common practice in the CH sector to enhance documentation, preservation, and management capabilities of heritage assets. Regardless of the methodology and sensor, the key acquisition aspect is the sampling resolution, i.e., the minimum distance between two consecutive measurements. This is defined by the image Ground Sampling Distance (GSD) for image-based campaigns, and by the instrument specifications and performance characteristics for range-based acquisitions (Remondino et al., 2013).

The 3D digitisation pipeline consists of three main phases: design and planning, implementation, and delivery. The design phase is crucial to meet the digitisation project requirements and needs, and to ensure that the outcomes meet the expectations in terms of accuracy and data quality. The choice of the most suitable sensor (active-range, passive-image based) and technique is constrained by several factors, including digitisation scope and specifications, object size and complexity, accessibility and portability, time and budget. The following sections provide an overview of the main 3D reality-based surveying and modelling approaches, including emerging AI-based processing techniques, to guide CH people involved in digitisation tasks and empower them with forecasting abilities in selecting the most appropriate solutions based on their project requirements.

2.1 Range-based approaches

Range-based 3D surveying techniques enable the creation of accurate digital representations of *objects, monuments, and sites* by capturing their geometry in a non-contact, non-invasive manner. The quality and resolution of the resulting models depend on several factors, including the choice of technology, the intended use of the data, and the resources available for the project. High-resolution 3D recordings are crucial for monitoring, studying and disseminating cultural assets, as well as for ensuring that the data can be reprocessed or reused as technology advances.

2.1.1 Fundamentals of range-based techniques

Range-based 3D surveying techniques can be primarily classified into *terrestrial and airborne* methods, depending on the sensor and data acquisition platform (ground-based or aerial). 3D range-based or active techniques, in particular laser-based sensors (commonly known as 3D laser scanners), rely on different 3D acquisition principles depending on the object size and the sensor-to-object distance. These systems are also referred to as active techniques, as they rely on the emission and reception of signals (commonly laser beams) to determine distances. For small volumes, they are typically based on the *triangulation* principle, while for larger-scale contexts, *Time of Flight (ToF) or Phase Interference/ Phase Shift Scanners* are commonly exploited (Remondino and Stylianidis, 2016).
Among the most popular ToF-based systems are *LiDAR sensors (Light Detection and Ranging)*. These systems operate by emitting laser pulses toward the target and measuring the time taken for each pulse to return after reflecting off the surface. The distance is calculated based on the speed of light and the measured time interval, allowing the creation of a *dense point cloud* that represents the scanned geometry (3D coordinates and, and in some cases, further attributes such as intensity and color). To achieve a complete and accurate 3D model, multiple scans are typically performed from different positions around the object or site. This strategy helps to capture all surfaces and reduce areas of *occlusion*. Overlapping regions between scans are critical for successful alignment in post-processing. For extensive or complex environments, reference targets or ground control points are often incorporated to ensure geometric consistency.
A well-executed range-based scanning workflow combines technical precision with flexibility, allowing operators to efficiently capture accurate 3D models for CH documentation.

Key operational considerations include:

- Ensuring complete coverage by planning scan positions and overlaps.
- Using reference targets or ground control points for reliable alignment.
- Documenting acquisition parameters and environmental conditions for reproducibility.

- Adapting the workflow for fragile, complex, or immovable objects.

2.1.2 Sensor type and characterisation

Triangulation

Triangulation is a method used to determine a point's position by measuring angles from at least two known reference points. Key components of the range devices exploiting this measurement mechanism include a *laser source* (emitting a narrow and focused beam of light towards the target surface), a *scanning mechanism* (which directs the laser beam across the object/scene), and a *detector* (which captures the reflected laser light). The object point is measured through triangulation, by measuring the angle between the emitted and reflected beam, and knowing the geometry of the device (i.e., the distance between the source and the detector).
In triangulation-based systems, this principle can be extended by a single spot to a set of aligned points forming a segment. These segments can result in profiles that are straight lines if projected onto flat surfaces, or curves in the case of more complex object geometry. In these systems, different capturing positions will generate a set of arrays describing, strip by strip, the object geometry.
In pattern projection systems, multiple sheets of light are simultaneously projected instead of a single sheet.
Within this category, structured light scanning relies on projecting a sequence of known light patterns onto the object's surface and capturing their deformation with one or more cameras. The system uses triangulation principles to calculate the depth and surface contours with high precision, making this approach especially suitable for smaller objects or areas where fine surface detail is essential.

Time-of-Flight (ToF)

Time-of-Flight (ToF) scanners emit a *laser pulse*, measuring the time it takes for the laser to return after reflecting off an object. The distance is subsequently calculated based on the speed of light and the measured travel time. ToF techniques rely on both *short, pulsed laser emissions (Pulsed Wave - PW)*, or by considering *modulated continuous waves (Continuous Wave - CW, including AM-CV Phase Shift and FM-CW systems)*. ToF systems are commonly used in terrestrial laser scanners (TLS), mobile mapping systems, and airborne LiDAR. They are effective for large-scale CH sites and monuments recording and their accuracy is typically lower than short-range solutions.
Direct ToF (PW): In PW systems the distance is estimated based on short Pulsed Wave (PW) of light energy generated by a source and directed towards the target. The time interval between the emission and reception of the reflected pulse is measured by the system. The distance to the surface is then calculated directly, considering that the speed of the light is known and constant. These systems are suitable for long-range measurements (tens to hundreds of meters) and outdoor long-scale applications. They generally feature a lower resolution compared to triangulation or phase shift systems.
Indirect ToF (CW): In contrast to direct ToF scanners, indirect systems rely on the emission of a continuous Wave (CW) of laser light towards a target, whose intensity is modulated at a known frequency. These systems estimate distance based on the phase of frequency shift between the emitted and received signal.

- *Phase Shift (AM-CW - Amplitude Modulation)*:
  Phase shift systems emit a CW of laser light modulated at alternating frequencies and determine the distance to an object by measuring the phase difference between the emitted and reflected signals. The phase difference between the emitted/received signal is measured when the beam is reflected by the objects. The system returns the measurements since the phase shift is proportional to the distance. These scanners are generally used for medium-range applications (up to 100 meters) where high accuracy is required, and are suitable for architectural and interior scanning.
- *Frequency Modulated Continuous Wave (FM-CW)*:
  Unlike AM-CW, which computes distance based on phase delay, FM-CW systems determine distance from the frequency difference (beat frequency) between the transmitted and received signal. CW systems typically need a wavelength long enough to avoid ambiguity, and their performance is better when the wavelength is short. In FM-CW systems, the emitted signal is a continuous laser beam with linearly varying frequency over time (also known as a chirp). Distances are determined by evaluating the frequency difference between the emitted and reflected signal. FM-CW systems feature higher precision and better resolution over medium distances and less noise compared to AM-CW systems, and are suitable for highly accurate surveying within a medium range (generally up to 20-30 meters).

Method	Typical range	Resolution	Common applications
Triangulation	0.01 m - 2m	0.01 mm - 0.1 mm	Small objects - highly detailed scanning
Direct ToF (PW)	1 m - 5000 m	0.5 cm - 5 cm	Outdoor/large-scale sites, monuments
Indirect ToF (AM-CW/ Phase Shift)	up to ≈ 200 m	0.1 mm - 1 mm	Architectural, interiors, medium-scale sites
Indirect ToF (FM-CW)	up to 20-30 m	0.1 mm - 0.5 mm	Medium range, highly precise surveying

2.1.3 Range-based pipeline for 3D reconstruction

The range-based 3D reconstruction pipeline comprises three main steps: geometry acquisition planning, data collection, and processing.

Geometry acquisition planning

An optimal data capturing planning should ensure the lowest number of stations while covering the entire object surface, with sufficient overlap for data registration, and the achievement of the required geometric accuracy. The suitable scan station design should minimise occlusions and avoid issues during data registration due to a scarce overlap. The scanning incidence angle should also be taken into consideration, as accuracy decreases proportionally to the angle size.
Optimising data coverage and quality can not disregard sensor-specific considerations, such as the *sensor range, field of view and resolution*, and environmental object or asset constraints (like accessibility, lighting conditions, and obstacles). Ensuring consistent *point density* across the object is also critical, as variations in resolution can affect the following processing steps, as well as the overall quality of the reconstruction.
Two common strategies for optimising scan station placement are Multi-View Planning (MVP) (Munkelt et al., 2010) and Next Best View (NBV) (Trummer et al., 2010) approaches. MVP requires at least a coarse model of the scene and optimises all viewpoints simultaneously, while NBV selects views incrementally based on the current state of acquisition. Both methods remain relevant for image-based 3D reconstruction workflows. Recently, some advanced algorithms and AI and simulation-based solutions for automatic view planning started to be implemented (Chen et al., 2022b; Dharmi et al., 2023; Border and Gammel., 2024; Cabrera-Revuelta et al., 2024).

Data collection

Spatial data are acquired from the planned scan positions using active-range sensors. Depending on the project scale and needs, several sensors (Section 2.1.2) can be exploited. While TLS and structured light devices ensure higher accuracy, MMS (Mobile Mapping Systems) solutions (handheld or mounted on backpack or vehicle-based systems) are typically used for large-scale and complex environments. Consistent measurements during the acquisition sessions are ensured by sensor calibration.
To guarantee full object coverage and sufficient overlap between scans, data acquisition should follow the designed scan positions and trajectory planning. The positioning and capturing of artificial targets (typically spherical, checkerboards, retro-reflective, or coded targets) supports the following processing and co-registration phase. Real-time visualisation tools, implemented and available for several range-based commercial solutions, can assist in the identification of missing regions and the adjustments of acquisition parameters onsite. The collection of metadata and paradata is also crucial for supporting the processing step.

Data processing

Raw data is here converted into complete and coherent 3D models. A pre-processing step is sometimes required to remove noise and artifacts, or isolated points derived from non-collaborative surfaces (such as reflective or transparent). For scans registration, the main approaches include:

- methods based on artificial or natural targets, identified automatically or manually;
- feature-based matching algorithms using natural scene elements. These features can be geometric keypoints, edges, or surface descriptors, and they are matched across overlapping scans to compute their relative transformation. Common algorithms are *FPFH (Fast Point Feature Histogram)* (Rusu et al., 2009) and *SHOT (Signature of Histograms of Orientations)* (Salti et al., 2014).
- coarse alignment techniques followed by fine refinement. The most popular is the Iterative *Closest Point (ICP)* (Besl and McKay, 1992), which iteratively minimizes the distance between corresponding points or surfaces in overlapping scans. Many and increasingly robust ICP variants (Rusinkiewicz and Levoy, 2001; Zhang., 2021) have been implemented to handle outliers, point density variability, or scarce overlaps, and can be point-to-point or point-to-plane based, or rely on more complex metrics to increase convergence and accuracy (Li et al., 2020).

2.1.4 Advantages and limitations of range-based approaches

Range-based 3D scanning technologies provide a high level of measurement accuracy, which is one of their most significant advantages in CH documentation. Structured light scanners can achieve sub-millimeter precision, making them ideal for capturing fine surface details on small objects, while LiDAR systems are well-suited for large-scale environments, offering reliable accuracy over extended distances. The non-contact nature of both methods ensures that even the most delicate or valuable artifacts can be documented without risk of physical damage.
Despite these strengths, several limitations should be considered. *The costs* associated with range-based scanning can be considerable, encompassing not only the purchase or rental of advanced scanning devices but also the selection of appropriate scanning techniques. Expenses may further increase if specialized accessories, software, or trained operators are required. However, a well-planned scanning process can optimize resource use and significantly reduce both time and financial outlay. Efficient organization of scan positions, minimizing unnecessary captures, and ensuring proper coverage can streamline data acquisition and limit the need for extensive post-processing.
Challenging surfaces represent another area of concern. Highly reflective, transparent, or very dark materials can cause difficulties for both LiDAR and structured light systems, potentially resulting in incomplete or inaccurate data. In such cases, additional preparation, such as the application of removable matte sprays or the use of alternative scanning strategies, may be necessary.
Environmental conditions also play a crucial role. It is also important to maintain the cleanliness and stability of the scanning environment. For example, ensuring that the scanned object is free from moving people in the vicinity can help prevent the collection of extraneous data, reducing the time required for filtering and cleaning during post-processing. Attention to these practical aspects not only improves the quality of the final model but also contributes to a more efficient and cost-effective workflow. Some weather working conditions (e.g., extreme temperatures) can furtherly affect the performance of some capturing instruments.

2.2 Image-based approaches

Photogrammetry is the most important image-based technique enabling the derivation of accurate, metric, and semantic information from images (Remondino and Campana, 2014). Images for photogrammetric 3D reconstructions can be derived from terrestrial digital cameras or aerial and satellite imaging sensors (Section 2.2.2).
In order to have a clear understanding of the form and shape of an object or a scene, observation from multiple viewpoints is necessary. In this way each image, through its unique perspective, provides the essential information to determine the position of a point in space, via the process of triangulation. In order to achieve this, users, or nowadays computers thankfully, have to detect similar points across multiple images of the same subject taken from a slightly different point of view. With these points, through a process known as *bundle adjustment*, the position of each image and each point that is detected on more than one of them, can be projected in 3D space, resulting in a sparse representation of that scene. That sparse point cloud of detected significant image features, along with the images position and camera parameters that was used for their acquisition, are the fundamentals for the traditional 3D reconstruction process, where even more points are detected and triangulated in adjacent images, through the process of dense stereo matching.
This approach requires at least two 2D images from different viewpoints to derive 3D information by establishing geometrical relationships (i.e., a mathematical model approximating the physical world through a projection) between the captured real-world scenes and the imaging data (photos). Similarly to human vision,

with at least two images (stereoscopic view), capturing the object from different perspectives (parallax), 3D information can be derived in the overlapping area (Luhmann et al., 2023).
A distinction is generally made between terrestrial (for larger-scale structures), close-range (for small objects), and aerial applications.
Despite this technique's history being almost as long as that of photography itself, impressive advancements have been made in the digital era, based on the same fundamental concepts and models.

2.2.1 Photogrammetry - Fundamentals

Photogrammetric processing is based on the *collinearity principle*, which establishes the relationship between the image and object spaces through a straight line between the camera perspective centre, the image point P(x,y), and the object point P(X,Y,Z). A collinearity equation is written for each image point measured in the images (tie or homologous points), and all the equations are solved with an initial approximation of unknown parameters (exterior orientation parameters, 3D object coordinates, interior orientation parameters if unknown).
The *bundle adjustment* method enables the simultaneous determination of all parameters and the estimation of the precision of the unknowns (called *self-calibrating bundle adjustment* if the interior orientation parameters are unknown) and solved by various methods (Salvi et al., 2001).
Unlike range-based solutions, photogrammetric methods reconstruct object geometries without direct distance measurement. Consequently, in order to derive accurate metric models (scaled model) a spatial similarity transformation is needed, and it is usually achieved by introducing some *Ground Control Points* (GCPs - at least 3), at least one known distance, or GNSS data for aerial acquisitions.
The photogrammetric workflow comprises some key processing steps:

- camera calibration: to estimate interior camera orientation parameters;
- image orientation: to calculate exterior camera orientation parameters;
- scaling and georeferencing (optional);
- 3D dense point clouds generation;
- polygonal model generation;
- texture mapping.

In *Structure from Motion (SfM)* workflows, camera calibration and image orientation are simultaneously estimated along with a sparse 3D object structure.

Key operational considerations include:

- *Planning the imaging configuration*: a careful planning of camera stations and network, by taking into account also the ray intersection geometry, is fundamental to ensure uniform coverage and minimize occlusions. The imaging configuration varies according to the object to be acquired, its scale, the planned *GSD*, and the image acquisition platform (terrestrial or aerial capturing). The geometry of the network may follow a convergent or parallel strips configuration, according to surveying needs. The convergent geometry strengthens depth estimation, while parallel strip acquisitions are typical in aerial applications. The network should ensure a sufficient image overlap (commonly 60-80% forward and 40-60% sideward) in both terrestrial and aerial cases. For convergent geometries, the imaging angle (at which an object is captured) should typically not be less than 20°. Too poor intersection angles between different images should also be avoided to not affect the reconstruction accuracy. Furthermore, the imaging configuration design is conditioned by the image scale, which in turn is determined by the *object distance, the focal length and related field of view*, and the desired final spatial resolution of the model. Particular attention should be paid to the *depth of field* when planning the acquisitions and the capturing distance, in order to ensure *image sharpness* throughout the object and avoid blurred areas affecting the quality of the reconstruction. *Color targets* could also be used to ensure proper harmonized colorization of the final 3D model and texture.
- *Ensure consistency in camera settings*: carefully set and lock exposure, focus, and white balance across all images. *A correct exposure is critical to guarantee coherence across the dataset* and prevent issues during the image matching step. In the operational activities, determine the proper exposure (by verifying the histogram) and lock it; define the aperture value (f-stop) which ensures a sufficient depth of field; *set a low ISO value (commonly 100 or 200) to minimize image noise;* try to *avoid motion blur effects* by adjusting the *shutter speed*. The focus should be set and kept fixed to *prevent variable image sharpness,* as well as *white balance* should be preset to ensure consistent colors, especially crucial in the feature detection and texture mapping operations.

- *Using coded targets, scale bars, GCPS, or GNSS data for model scaling and, eventually, georeferencing*. These spatial references need to be included within or around the scene. Coded targets and scale bars provide known distances for scaling the models, while GCPs (used both in terrestrial and aerial contexts) allow for spatial referencing of the model to a defined reference system. GNSS data are used in UAV (Unmanned Aerial Vehicle) or aerial photogrammetry often in combination or in place of GCPs for approximately determining camera positions and exterior orientation parameters, as well as for direct georeferencing applications.
- *Recording of paradata to ensure transparency and reproducibility*. This includes the description of camera models, lenses, image resolutions, acquisition time, but also information related to the acquisition settings (camera settings, environmental conditions, imaging geometry).

2.2.2 Sensor/lens types and characterization

The quality and the accuracy of photogrammetric 3D products are closely related to the sensors and lens characteristics, in addition to the image network geometry.
The main sensor types used in photogrammetry are:

- *Digital terrestrial cameras:* such as Digital Single-Lens Reflex (DSLR), mirrorless, panoramic, industrial, or action cameras. These are usually used in close-range photogrammetry for CH documentation, but can also be mounted on UAVs for low-altitude aerial imaging acquisition.
- *Aerial/aircraft imaging sensors:* typically classified as small, medium, and large format cameras. Metric cameras (built for photogrammetric applications) ensure very high geometric stability. They are used for capturing very high-resolution imagery over large areas and for supporting the documentation of widespread CH sites.
- *Satellite sensors:* satellite optical sensors (multispectral or panchromatic) operate at much higher altitudes, and provide data with varied spatial resolutions. Data can be used especially for large-scale site monitoring and documentation, and further geospatial analyses and assessments.

Lenses can be classified as:

- *Wide, super-wide, and fisheye lenses:* these short focal lengths are often used in CH documentation to cover a wide field of view - about 60-75° for wide, 80-120° for super-wide, and up to 180° for fisheye. The larger the field of view, the greater the optical distortions (radiant distortions) that need to be modelled and corrected.
- *Standard lenses:* typical focal lengths are in the range 35-50 mm, and they offer a good compromise in terms of covered area and distortions, and are common in close-range applications.
- *Zoom lenses:* they provide varying focal lengths in a single lens. In photogrammetric applications,. A change in the focal length requires a new interior orientation and the distortion estimation. Due to the low stability, they are rarely used in practice for highly-accurate documentation activities.
- *Tilt-shift lenses:* these enable the adjustment of the lens plane relative to the image plane and are employed for highly accurate CH documentation projects.
- *Telephoto lenses:* they feature narrow field of views (that vary based on the focal lengths and the sensor size), suitable for capturing distant objects with limited distortions (generally mounted on aircraft). Full-frame 90 mm-300 mm (practical) telephoto lenses are suitable for Macro Photogrammetry, due to the ability to shoot from a relatively vast distance (typically 1m +) and get a deeper depth of field while not closing an aperture too much to induce the diffraction effects.
- *Telecentric lenses:* they are designed to capture all object points at the same image scale regardless of their distance. In the CH context, they are used for highly detailed close-range applications and the documentation of small artefacts or other architectural elements to be reconstructed, requiring elevated geometric precision.
- *Panoramic lenses:* they are able to acquire ultra-wide or full 360° fields of view in a single shot or through a series of synchronized images. In the CH sector, they are used mainly for creating immersive and complete records of heritage assets.

The characterization of sensors and lenses allows the derivation and assessment of their optical and geometric properties which can condition the photogrammetric accuracy, This includes, as examples, the measurement of optical distortions (especially radial and tangential) and the evaluation of the image quality in terms of resolution and sharpness.

2.2.3 Image-based pipeline for 3D reconstruction

The key steps of the image-based photogrammetric pipeline for 3D reconstruction include:

- *Image pre-processing*. Before starting the reconstruction the quality of the dataset needs to be assessed, in terms of image *sharpness (focus), motion blur, noise level, and exposure correctness*. Poor image quality, if not corrected through image enhancement techniques, can affect the feature matching step and reduce the model accuracy. Image pre-processing also includes brightness, contrast, and white balance adjustment and colour optimisation.
- *Feature extraction and matching*. In this step, a set of distinctive features (keypoints) (corners, blobs, or edges) are primarily extracted in all images. *Several traditional (like SIFT, SURF, or ORB) or more recent learning-based detectors (such as SuperPoint, LoFTR, etc)* are available for their automatic identification and increasingly embedded in open-source and commercial photogrammetric software. Therefore, the detected keypoints are matched to established correspondences among overlapping images. This phase is critical to determine spatial geometric relationships (e.g., epipolar geometry) supporting the 3D reconstruction. Also for the *matching phase, new learning-based solutions (like SuperGlue, LoFTR, or LightGlue)* are emerging together with consolidated algorithms and strategies (such as *Brute-force matching*, *approximate nearest neighbour search*, and so on), and are especially promising when dealing with multi-temporal image datasets, and complex and low-texture CH environments.
- *Camera calibration & Image orientation / Structure from Motion (SfM)*. Taking as input the features matched in the previous step, SfM estimates camera pose (exterior orientation parameters) and, if unknown, camera interior parameters, often using self-calibration. A sparse 3D point cloud reconstruction is generated at the same time. SfM relies on collinearity equations and uses bundle adjustment to iteratively refine both the camera parameters and the 3D scene reconstruction. Different approaches have been implemented for this task, based on *incremental* (the reconstruction is performed by adding one image at a time), *global* (camera poses are estimated simultaneously), or *hierarchical* (combining both to manage complex and large datasets) approaches. Even though it remains an open-issue due to the lack of standards, several formats exist to exchange camera information (internal and external parameters) like *Agisoft XML, bundler (.out) or COLMAP* are the most used interchangeable formats.
- *Georeferencing and scaling*: Unlike range-based approaches, which measure distances directly, photogrammetric techniques require the insertion of additional information to produce a metric product. GCPs should be added in the georeferencing process to assign real-world coordinates and align the model with a geographical reference system. Alternatively, scaling can be performed using known distances between features visible in the acquired scene to assign a real-world scale.
- *Multi-View Stereo (MVS)*. MVS algorithms enable dense scene reconstruction, taking as input the camera poses and sparse point clouds derived from SfM algorithms (Stathopoulou and Remondino, 2023). With MVS, a combination of stereo images is created starting from oriented multi-image aggregation. Unlike traditional stereo matching, MVS leverages different images to generate detailed dense reconstructions. The estimated correspondences among images are used to extract depth information, and, through triangulation, depth data from different views are combined to compute 3D coordinates and create the dense scene. The redundancy of data from multi-view imagery supports the reconstruction and limits errors when dealing with occlusions, textureless areas, or limited viewing angles.
- *Polygonal model generation*: once the dense point cloud has been generated through MVS, it can be converted into a 3D mesh with a continuous surface by connecting point vertices. Common algorithms for this step are the *Poisson Surface Reconstruction, Delaunay Triangulation, or Voronoi-based mesh reconstruction*.
- *Mesh refinement and optimisation*: this step is often fundamental for improving the final model quality. It comprises different operations, from data decimation (reducing the number of polygons), smoothing, and hole filling for a complete model representation.
- *Texture mapping*: in this step, images of the object/scene are projected onto the 3D mesh model to create a realistic representation. The images are projected onto the model based on the camera poses. The first phase of texture mapping is *UV mapping*, i.e., the unwrapping of the 3D mesh into a 2D representation (2D UV coordinates). Therefore, images are projected onto the mesh using the created *UV map* and camera pose information.
- *Orthophoto generation*: as an additional photogrammetric product, *orthophotos* offer a 2D, scaled, and corrected representation of the scene. These products are essential in topographic mapping and when true-scale data are needed. In an orthophoto, distortions (such as those from camera tilt and perspective) have been geometrically corrected, preserving a consistent scale that allows for accurate measurements. The process of distortion correction is called *rectification*. After rectification,

an image mosaic is created by stitching multiple overlapping rectified images together, which are then projected onto a flat plane.

2.2.4 Advantages and limitations of image-based approaches

Image-based approaches offer a versatile and cost-effective solution for 3D reconstruction of CH assets. Together with range-based techniques, these approaches also feature several advantages and limitations. With a focus on photogrammetric reconstruction strategies:

Advantages

- *No-contact approach*: this ensures its large applicability to several CH assets and contexts for documentation and monitoring activities. Fragile heritage can benefit from this non-invasive approach to preserve the asset's integrity and minimize the risks during acquisition campaigns. In case of remote, challenging, or hardly accessible sites or monuments, it expands the possibilities of capturing data without physical presence.
- *Highly accurate 3D geometry*: when high-resolution images are properly acquired (taking into account the correct imaging configuration) and with the correct processing workflow, 3D photogrammetric products ensure high geometric accuracy, comparable to range-based results. Photogrammetric 3D models can provide precise measurements and spatial representations, making them suitable for documenting CH assets with a high level of detail and precision.
- *High-quality texture*: while range and image-based solutions can generate comparable accurate model representations on the geometry side, the high fidelity and quality of the derived texture is one of the main strengths of the photogrammetric technique. *Finer surface details, colors, material properties, and a higher realism* is typically derivable with the use of this technique.
- *Cost-effectiveness*: compared with range-based solutions, photogrammetry is typically a more affordable documentation method. This is essential in a documentation project with budget constraints and for increasing accessibility to a wider range of users. When referring to UAV-based acquisitions, extensive areas can be quickly acquired, reducing the costs of ground operations.
- *Flexibility and scalability*: this technique can be adapted to a broad range of CH needs and objects scale, from small artefacts to large-scale surveying (by adapting acquisition methodologies and equipment).
- *Automated processing workflow*: compared to an analogue/traditional photogrammetric processing workflow, current digital pipelines are largely automated. Feature extraction, image matching, camera calibration, and model generation can be performed transparently to the user.

Limitations

- *Image-quality dependence*: the quality of the final 3D models is strictly related to the quality of the input images used in the photogrammetric reconstruction process. The performance of feature extraction and image matching steps, as an example, can hardly be conditioned by blurred, low-resolution, or over- and under-exposed images. The geometric quality and final model appearance may also be affected by poor image quality.
- *Sensitiveness to low-texture surfaces*: In the image-matching step, the identification of distinctive object features is critical for the success of this operation. Surfaces with poor textures, such as monochromatic walls, or highly reflective materials, make this process particularly complex and challenging. The other steps of the pipeline are consequently affected by this critical operation, and the final 3D model can often show uncompleted areas in featureless regions.
- *Dependency on external information for generating a metric product*: GCPs or known distances are, unlike range-based techniques, always needed to obtain a scaled model. The quality of these reference measurements conditions the metric quality of the product.
- *Sensitivity to lighting conditions*: unlike range-based solutions, photogrammetric products are highly dependent on proper lighting conditions during the acquisition. Main reconstruction issues occur in the areas affected by shadows, with under- or over-exposed images, inconsistent lighting, which can lead to challenges in the image-matching phase, as well as can affect the color fidelity and quality during the texture mapping.
- *Contextual sensitivity and repeatability and reproducibility:* Image-based modeling techniques remain indirect measurement techniques. It means that the uncertainty of the measure itself relies on and is

derived from the capacity of the imaging system (capture and processing) in terms of accuracy and precision. Any change that will occur in the camera system, the imaging configuration or the context of the digitisation set-up will impact the measurement and consequently the resulting 3D model. This makes image-based models very sensitive to context variations affecting repeatability and reproducibility purposes. The same data-set processed with various software or different versions of the same software will lead to alternative results. Even the same object or scene digitized two times consecutively with the same camera and by the same operator and same context may lead to slightly different results. These repeatability issues could be extended to reproducibility concerns. *Nowadays those conditions are valid only if the acquisition is performed in a controlled environment and with an automated digitisation rig or set-up and fixed processing workflow.*

- *Dependency on the imaging configuration:* a proper planning of the camera network, in order to guarantee a sufficient image overlap, ensure an optimal coverage, and minimize the occlusions, is not always a straightforward operation. Environmental constraints can complicate the operational activities and prevent the acquisition of all needed images for a complete and accurate model reconstruction.
- *Dependency on accurate camera calibration:* The correct estimation of interior camera parameters is essential to prevent the generation of inaccurate 3D models. If these parameters are not available or not correctly estimated, achieving an accurate 3D model is challenging.
- *Requires specialised knowledge*: while automation has increased accessibility to this technique, it still requires a certain level of expertise to achieve highly-accurate and correct results. Inaccurate acquisition setup and configuration parameters, incorrect data capturing plan, insufficient knowledge of the processing workflow can lead to poor-quality reconstructions.

2.3 Alternative scene representations and AI-based 3D digitisation

In the context of 3D digital content storage and representation, special data structures for scene description and AI solutions have recently emerged, as complementary/alternative techniques to well-established approaches. Traditional methods (Sections 2.1 and 2.2) of 3D modelling and reconstruction have proven to be effective in many contexts, but, as highlighted in the previous sections, still have limitations. Alternative scene representations which rely on AI technology are increasingly emerging as a promising solution for overcoming these limitations and enabling the improvement of 3D data quality and the reconstruction even in challenging and complex scenarios. In the 3D reconstruction field, advances in Computer Vision and Machine Learning are driving innovation. The capabilities of these systems to process visual data and extract spatial information from different sources have significantly expanded the 3D reconstruction borders. As an example, AI-driven approaches can handle more complex scenarios, such as reconstructing thin, featureless, reflective, and transparent surfaces, or interpreting occlusions and varying lighting conditions. Trained on large datasets, these algorithms learn from patterns and can predict previously unseen data. Despite the evident advantages, these methods still face numerous challenges, such as the required computational resources for processing. Furthermore, the generation process often operates as a “black box”, making the decisions underlying the reconstruction uninterpretable. The lack of explainability makes these solutions hardly suitable when the control over the documentation outputs is critical.

2.3.1 Multi-Image Technologies

Traditional SfM techniques are typically effective in reconstructing scenes and objects that exhibit surfaces with rich texture and well-defined, randomly scattered features, such as the surface of a rock or a tree bark. However, when surface textures are poor and the environment presents complex lighting conditions, the detection of well distinguishable features across multiple adjacent images is frequently erroneous or impossible, resulting respectively in unsatisfactory 3D reconstructed surfaces that are noisy or contain holes (discontinuities).

In pursuit of overcoming the limitations inherent to traditional SfM techniques, particularly their tendency to produce incomplete or low-quality 3D reconstructions in challenging scenarios, alternative approaches have emerged. These advanced solutions leverage *volumetric or special point cloud based* representations to store significantly richer and more advanced information about a scene, compared to that obtained through conventional SfM and range-based methods. Basically, these techniques try to capture the way a point in space appears from different viewing angles, focusing on how it reflects and transmits light in a specific direction, quantified in terms of light radiance. By capturing this information, it becomes possible to generate

novel views with smooth transitions between viewpoints, resulting in a coherent and photorealistic visual output that exhibits intricate light interactions.
However, rendering this type of data requires highly specialised and sophisticated pipelines that are closely tied to the specific representation format, making their integration into mainstream applications challenging. In contrast, traditional 3D digitisation techniques, primarily based on the triangulation of simple spatial points, are capable of producing a far more universal and widely compatible representation of a subject, typically encoded as a straightforward set of X, Y, Z coordinates if referring to point clouds.
Most recent approaches to 3D documentation have increasingly relied on *radiance field representations*, with many leveraging machine learning and deep neural networks. These technologies play a key role in capturing and encoding the 3D information from multiple 2D images into a radiance field, as well as in rendering that information realistically. However, because deep neural networks can be computationally demanding, several alternative methods have emerged that aim to deliver similar levels of realism without relying on them, either throughout the entire process or at least during rendering. The goal of these approaches is to enable *interactive frame rates*, even on less powerful hardware, thereby democratising this modern type of visually appealing scene rendering.
In addition, the need for more compatible data that works with the available ecosystem of 3D graphics led to solutions for constructing volumetric representations that store the distance of a light ray from various solid surfaces as it is traveling through space, rather than how that ray is steered and deflected. These representations are called *Distance Function fields* and are able to resolve the distance of a given point in that field from a solid surface that is defined by that field. These Distance Function fields can be visualised directly using ray tracing, or alternatively they can be triangulated to produce a 3D mesh that can be exported in a common 3D file format that is usable by a plethora of software.
The table in Annex 1 offers a comprehensive overview of the most popular, recent, and promising technologies within this highly active research area.

2.3.2 MDE - Monocular Depth Estimation

Monocular Depth Estimation (MDE) is the computer vision challenge of extracting 3D information using just a single image captured from a stationary (monocular) camera, without any structured illumination, and relying solely on prior knowledge. This prior knowledge is based on the principle of visual cues, a concept that painters grasped and applied for centuries, helping them to create ultra-realistic paintings during the Italian Renaissance era. In a similar fashion, within computer vision, these visual cues are algorithmically identified with the advent of AI, and enable the automatic estimation of the distance of each pixel from the camera. This opens up a wide range of applications, including 3D scene reconstruction for robot vision tasks, such as ego motion and autonomous driving. It also has significant potential in the cultural heritage domain, enabling, as examples, the creation of 3D models of structures and artefacts lost or destroyed (using a single image from digitised legacy archives), or complete 3D assets for use in virtual exhibitions, games, and VR applications.
The recent advancements in deep learning and, more specifically, Convolutional Neural Networks and Vision Transformers, have significantly improved the accuracy and efficiency of MDE techniques. Recent works in AI-based depth estimation from a single image, exhibit an almost human-like ability to resolve depth information from visual cues. Models like MiDaS (Ranftl et al., 2020), Depth Anything (Yang et al., 2024), and DPT (Ranftl et al., 2021) leverage large-scale datasets and sophisticated neural network architectures to generate detailed depth maps from single image. Some of the more recent works even go a step further, imaging and reconstructing complete 3D models from just one image (Tochilkin et al., 2024; Lai et al., 2025; Wang et al., 2025).
All these AI models are trained on diverse datasets, enabling them to generalise well across various scenes and conditions. However, the inability to control certain aspects of the depth estimation process can lead to limitations. For instance, early MDE models struggle to accurately estimate depth in regions with low texture, occlusions, or reflective surfaces. A significant amount of research is steered to tackle such challenging conditions, including complex surfaces and transparent objects (Tosi et al., 2024). However, the reliance on large datasets to train these neural networks can introduce biases, which affect the model's performance across diverse environments and make true generalisation difficult to achieve.
A list of technologies that are able to estimate depth from a single image is provided by the Annex 2a.
Going beyond simply measuring the distance of each pixel in an image from the camera's center, AI can use prior knowledge to infer the full structure of an object and generate a complete 3D model from just a single image. A list of technologies that are able to synthesise a complete model from a single image is provided by the Annex 2b.

Going a step further, into the realm of the neural rendering techniques using radiance fields, there are some shy attempts to create such representations merely from a single image. However, these methods are not able to cope with generic input, as they rely on strong priors learned during training. As a result, the output is often less detailed and erroneous compared with techniques that use multiple images, due to scene structure ambiguities and the overfitting of pretrained data. A list of technologies capable of creating a neural scene representation from a single image is provided by the Annex 2c.

2.3.3 Advantages and limitations of AI-based solutions

The recent advent of AI technologies in computer vision has led to some significant breakthroughs in how a scene or a subject is 3D digitised, stored and rendered on screen. AI, through the introduction of Convolutional Neural Networks and more recently Vision Transformers, is able to provide computer vision human-level assessments of spatial structures and depth, even from a single image, by following visual cues and high-level semantic understanding. This ability enables photorealistic rendering and highly efficient data representations, drastically reducing the need for dense multi-view capture setups and extensive manual processing. AI methods have thus opened new possibilities in digital art, gaming, virtual and augmented reality, and media production, where rich visual realism with lower computational or acquisition costs is highly valued.

AI-based technologies are game-changing for photorealistic rendering and efficient data representation. However, they are not yet comprehensive replacements for full 3D digitisation pipelines in industries requiring precision, real-time interaction, or editable models. They can be used in applications like digital art, visual media, and virtual environments, but traditional methods still dominate fields where accuracy and interactivity are paramount. AI predictions can suffer from inaccuracies, hallucinations, or inconsistencies due to ambiguous or insufficient input data. AI models typically generate implicit or volumetric representations that are not always directly editable or compatible with existing workflows relying on polygonal meshes or point clouds. Furthermore, the real-time responsiveness needed for interactive applications still often exceeds current AI rendering speeds or requires specialised hardware, such as an expensive, modern, and powerful GPU. Another huge and really active debate on the use of AI, among many like that of the social impact, is the environmental impact that is needed mainly during training. Additionally, professional developers that create consumer application for content creation (3D editing for vfx animation etc.) or consumption (games) are also concerned, due to the additional characteristics and computational capabilities that are required in order to take advantage of such neural rendering techniques, but also their adoption to industry standardised pipelines that many organisations and companies have agreed on.

Since neural radiance rendering is a very active research area, it almost dictates that a wide adoption of a specific methodology will not yet be established in the same way, for example, as a triangular mesh or a point cloud is. A triangular mesh representation is established since more than 50 years and its rendering is hard-coded onto the hardware of many consumer electronics that have a screen and render some kind of graphics, including wrist watches nowadays. + All these advanced radiance rendering methodologies, either neural or not, are taking advantage of the programmable graphics rendering pipeline that started around the new millennium. Some early promising technologies like Gaussian Splatting and NeRFs are slowly being adopted into the existing ecosystem of some mainstream applications. However, Gaussian Splatting is the one that gains traction and a much bigger audience, since it can be generated faster than NeRFs, while at the same time it can be displayed at interactive speeds using modern low-cost consumer hardware. Still, there are many newer methodologies that surpass both of them in speed and visual quality as well. This states that the technology is not yet mature for wider adoption, and at the moment it will remain on the posh side for the majority and a niche for a few.

For serious applications where metric information is a top priority, as for example in the Cultural Heritage (CH) domain, where 3D archiving and accurate 3D documentation are crucial, the future likely lies in blending these techniques into a seamless pipeline that will augment traditional photogrammetry (Kim et al., 2025). Neural-based scene understanding can help traditional SfM methods achieve faster processing times and produce more complete final data.

Whether and when that kind of AI technology will be mainstream and easy to use by people who are willing to use this technology, like CH experts, is not yet clear. The learning curve in order to use a few of the AI technologies that are freely available is notably steep and demands a substantial degree of familiarity with the technology. Moreover, AI technologies demand powerful computing equipment that costs a lot at the moment. Having as an example widely accepted popular AI application, like those used for text, image and more recently music and video synthesis, it is obvious that the benefits of AI in CH domain will start coming initially as a service, provided by companies that already have an active role in the field of 3D digitisation and 3D data presentation through online repositories. This is already true, with some online 3D model repositories already supporting Gaussian Splatting rendering, while others advertise neural-assisted

digitisation using just your mobile phone, with your data passing through their centralised service for processing, stating that they are able to capture featureless, reflective and even transparent refractive objects.
Afterwards we report a short and generalised list of the advantages and limitations of AI-based solutions:

Advantages

- They can produce visually pleasant results, even from a minimum amount of input data, thus including just a single photo.
- Radiance field type representations, either volumetric or splat based, are able to create interactive photorealistic renderings showing complex light interaction, without requiring additional user effort for advanced materials and lighting setup, as it used to be with classic rendering pipelines based on polygonal meshes and image textures.
- They are able to infer missing information, based on prior knowledge that is acquired from a huge amount of training data.
- Neural Networks present human level image reasoning and are able to distinguish visual cues that were too difficult to extract from an image using classic computer vision algorithms and traditional programming. That helps them to accurately approximate structures and shapes from images that were never exposed before during training.
- They can produce polygonal meshes and textures as well as implicit and neural representations to facilitate applications that require such data.
- Novel mobile AI apps, which convert videos into 3D models or scenes, are usable by novices and require only low-cost equipment (i.e. a smartphone).

Limitations

- AI technologies are considered black box approaches that we have little control over their output. The output is mostly based on prior knowledge, which in some cases might introduce hallucination artifacts. That kind of artifacts are common when there is lack of information inside the pictures, like image features and visual cues that can cause problems for both multi and single image approaches. Multi image approaches can suffer even more, by a scarce number of input images, but also ambiguities and errors of how these images are posed in space.
- Most of these methodologies are computationally expensive and are taking advantage of the computer's Graphics Processing Unit (GPU) in order to perform the calculations in a reasonable amount of time. However, the major requirements concerned with this approach, is in the amount of GPU memory in combination with GPU architecture. Most of the approaches are having trouble processing input images at their full resolution, due to extreme GPU memory requirements. As of the current market state, the cost of a GPU capable of generating neural reconstructions from multiple images might be prohibitive for many users.
- The above statement is true for consuming such a content as well, since the majority of the AI techniques require advanced hardware for good quality interactive rendering of the generated neural representations. Nevertheless, there is a great scientific effort to democratise such content and provide solutions to speed up rendering even on commonly used low end hardware, such as mobile phones and common laptop and desktop computers.
- Editing and compositing such neural representations is difficult and their appearance depends on the lighting conditions during the capturing process. Relighting such representations is an active field of research.
- Extracting accurate metric information is not possible for almost every single image approach, but also for multi image approaches that hold their data in radiance field volumetric representations too.
- Most volumetric representations are prone to visual artifacts like aliasing that are caused by the partitioning of space and the sampling of the given information. Hierarchical based approaches are trying to tackle such problems but still space partitioning and high resolution sampling of information is a major issue.
- Dynamic scenes featuring deformable objects is not possible by many methods and for those that support it is not efficient at all, requiring vast amounts of GPU memory and computation time in order to generate.

- At the moment, there is a plethora of open data sets that were and still used to train many of these neural networks methodologies, however their context is somewhat limited and the majority of them feature low quality and low resolution data, leading to bad generalisation and low quality results. Retraining a neural network for a specific task, by feeding it with data that are targeting that case, requires vast computational resources, time and due to electrical power consumption money.

2.4 Source-based approaches

While the previous subsections have focused on reality-based 3D modeling techniques—where digital 3D models are created directly from sensor data captured from real-world objects or environments using active or passive sensors—there exists another significant approach within the 3D modeling domain that is particularly relevant for Cultural Heritage applications: *source-based modeling*.

Source-based modeling, a term that has entered the scientific literature relatively recently (Demetrescu, 2015), refers to the creation of 3D models through the interpretation and integration of various documentary sources rather than direct reality capture. This approach has become increasingly recognized in Cultural Heritage applications over the past decade and represents a crucial methodology, particularly when dealing with heritage contexts that no longer exist or are inaccessible for direct documentation.

2.4.1 Fundamentals of source-based techniques

Source-based 3D modeling encompasses the creation of digital reconstructions through the analysis, interpretation, and synthesis of diverse documentary materials including historical drawings, paintings, engravings, photographs, textual descriptions, archaeological reports, architectural plans, and other archival sources. Unlike reality-based approaches, these techniques rely on interpretative processes that combine multiple information sources to reconstruct hypothetical but scientifically grounded 3D representations.
The methodology employs various computer graphics techniques ranging from:

- **Polygonal modeling**: creating 3D geometry through manual construction and manipulation of vertices, edges, and faces.
- **NURBS modeling**: using *Non-Uniform Rational B-Splines* for creating smooth, mathematically defined surfaces.
- **Procedural modeling:** implementing rule-based systems and shape grammars for automated or semi-automated generation of architectural elements.
- **BIM-based modeling**: applying Building Information Modeling methodologies for comprehensive architectural reconstruction with semantic information integration.
- **Parametric modeling**: developing flexible models that can be modified through parameter adjustments.

2.4.2 Applications in Cultural Heritage

A paradigmatic example of source-based modeling are virtual historical reconstructions, where scholars combine archaeological evidence, historical sources, architectural treatises, and comparative analysis to recreate disappeared urban landscapes. These reconstructions serve multiple purposes including research, education, and public engagement, offering insights into historical contexts that would otherwise remain inaccessible.
Other applications include:

- Reconstruction of destroyed monuments and archaeological sites;
- Hypothetical restoration of incomplete architectural structures;
- Visualization of historical urban environments;
- Recreation of ancient landscapes and territorial configurations;
- Digital anastylosis of fragmented architectural remains.

2.4.3 Accuracy and validation approaches

The accuracy of source-based models is fundamentally different from reality-based reconstructions. Rather than being measured through geometric precision and metric accuracy, the quality of source-based models is evaluated based on:

- *Source reliability*: the credibility and historical accuracy of the documentary evidence used.
- *Methodological transparency*: clear documentation of interpretative choices and reconstruction hypotheses.
- *Scientific consistency*: coherence with archaeological, historical, and architectural knowledge.
- *Uncertainty representation*: explicit indication of different levels of confidence in various model components.

This represents a shift from quantitative to qualitative assessment criteria, where the focus lies on the correctness and scientific validity of the sources and interpretative processes rather than metric precision.

2.4.4 Integration with reality-based approaches

Source-based and reality-based modeling approaches are increasingly being integrated in complex heritage projects. Reality-based data from existing remains can provide constraints and validation for source-based reconstructions, while hypothetical reconstructions can contextualize fragmented archaeological evidence captured through traditional surveying techniques. This hybrid approach combines the metric accuracy of sensor-based documentation with the interpretative richness of source-based modeling.

2.4.5 Software and technological considerations

The implementation of source-based modeling relies on a diverse ecosystem of software tools, each offering specific capabilities for different aspects of the reconstruction process. Open-source solutions like Blender (www.blender.org) or Unreal Engine (https://www.unrealengine.com/) have gained particular prominence in virtual heritage reconstruction projects due to their comprehensive capabilities, that uniquely combine semantic shape modeling with photorealistic material representation. This integration allows researchers to maintain both the scientific rigor required for scientific accuracy while achieving the visual communication effectiveness necessary for educational and dissemination purposes.

Other commonly employed software solutions include professional 3D modeling packages for NURBS and parametric modeling, BIM platforms for architectural reconstruction with semantic information management, and specialized procedural modeling tools for rule-based generation of complex architectural elements. The choice of software often depends on the specific requirements of the reconstruction project, the available expertise, and the intended final outputs.

Modern source-based modeling workflows increasingly integrate multiple software environments, leveraging the strengths of different tools while maintaining data interoperability through standardized 3D formats and protocols.

2.4.6 Characteristics and considerations

Source-based modeling presents unique characteristics that distinguish it from reality-based approaches. The technique enables reconstruction of lost or inaccessible heritage contexts and supports comparative studies across different historical periods, making it particularly valuable for educational and public engagement applications. However, the interpretative nature of the approach introduces inherent uncertainty that must be carefully managed through transparent methodological frameworks.

The dependency on availability and quality of documentary sources represents a key challenge, as does the risk of subjective bias in interpretation processes. Representing uncertainty and alternative hypotheses remains technically and conceptually challenging, while the potential for misrepresentation increases if methodological transparency is insufficient.

2.4.7 Methodological considerations

Source-based modeling requires rigorous methodological frameworks to ensure scientific validity. The London Charter (London Charter, 2006, Denard, 2013) and Seville Principles (Seville Principles, 2011) provide guidelines for computer-based visualization of heritage, emphasizing the importance of:

- Transparent documentation of sources and methods;
- Clear distinction between evidence-based and hypothetical elements;
- Provision of metadata and paradata describing reconstruction processes;
- Implementation of uncertainty visualization techniques;
- Peer review and interdisciplinary collaboration.

The integration of source-based approaches within the broader landscape of 3D heritage modeling acknowledges the complementary nature of different reconstruction methodologies and their specific contributions to heritage understanding, preservation, and dissemination.

3. 3D content co-registration, editing, and optimisation

Creating high-quality content is a complex process that requires many processing steps from data acquisition to model creation. Beyond the general processing pipelines highlighted in Section 2, in many cases, the editing and optimisation of the achieved results are unavoidable steps to meet the required levels of accuracy, detail, and realism of the reconstructed asset. *Raw 3D models often feature imperfections, artefacts, partially reconstructed areas, or redundant information*. These are mainly due to noisy data, misalignments, and other reconstruction errors that need to be corrected during or after the 3D content generation process. The data editing part is dedicated to the *correction of topological and textural errors* (filling holes, removal of unwanted artefacts and noise, correcting misalignments, and so on), while the *optimisation* (including, as an example, polygons decimation and simplification, texture size reduction) focuses on the improvement of the efficiency and management of the models, including their use in further platforms. Geometrical optimisation can also be applied to create smoother surfaces or more consistent geometry (i.e., through mesh refinement algorithms that smooth rough edges and correct some deformations generated during the reconstruction process). The following sections present more in-depth some of these techniques used for co-registering, editing, and optimising data.

3.1 Co-registration and data fusion

The spatial registration of the sensors and the collected data is intrinsically addressed, on the one hand, by the data processing flows from their respective techniques (range-based or image-based modeling) and on the other hand by the different 3D and 2D co-registration methods. According to the type of data and format, the registration of the spatial attributes may result in more or less complex processes. In a generic formalization, data registration aims to optimise spatial overlays by transforming two (or more) data sources. The registration process differs by means and objectives to the data fusion process which involve the creation of an added value, while registered data only solve spatial overlay.
Characteristics considered for registration and fusion can be *extrinsic*, as is the case with targets or calibrations from the acquisition device, or *intrinsic*, i.e., contained or extracted from the data source (points, corners, intensities, gradients, edges, regions, etc.). Although local models exist, global models applying to the entire source without splitting it into different parts are more common in CH applications. What is understood by data co-registration therefore consists in the application of calculation methods (geometric transformations) allowing the referencing of 3D or 2D data, in a common spatial coordinate system, using the projective relationship of common characteristics (2D/2D; 2D/3D-3D/2D; 3D/3D). There are different approaches to solve the registration problem of recalibration, that may rely on hardware (closer to the sensor), software (closer to the data), or hybrid implementation.

3.1.1 Hardware-guided registration

The registration of sensors and by extension of their data can be assisted by different techniques to integrate or guide the spatial-temporal repositioning of data. A distinction is made between methods based on sensors integrated or coupled with acquisition devices, the most common used in CH digitisation is *GNSS*. However, its limits in terms of fine positioning make it necessary to move towards *Ultra Wide Band*

(UWB) (Masiero et al., 2018a, Masiero et al., 2018b), especially for indoors application (Khoury and Kamat, 2009). This kind of approach is generally combined with other sensing methods like *Inertial Measurement Units (IMU) or geomagnetic* to define spatial orientation. This method is considered autonomous in contrast to methods based on target detection. This approach therefore involves integrating spatial cues and characteristics, more or less synchronously, into data collection. Although the accuracy of this registration is generally approximate, it remains very interesting in the pre-registration stage.

3.1.2 Intermediate registration based on hardware and software solutions

The use of extrinsic features at the sensing source, including the use of targets (optical or fiduciary) requires an action/interaction on the scene or object, but also the integration of a detection step based prior to the processing workflow. There are also techniques that apply in 4D/3D or 3D/2D contexts. We can first mention the *tracking* or *motion capture systems* (*Id Est Motion-Capture*) nowadays widely used in the field of Computer Vision (Lepetit et al., 2005) and more precisely in their applications for robotics (Smith and Singh, 2006) or virtual environments (Jankowski and Hachet, 2013) with only commercial and very competitive offers (such as*Vicon, OptiTrack, PhaseSpace*, etc.). These systems are based on the detection of spherical targets called *SMR* (*Id is Spherically Mounted Retroreflector*) themselves derived from the principle of angular reflectors (*Id is Corner Cube Reflector*) from the fields of metrology and photonics. Note that the use of targets is a preferred means for spatial alignment of many techniques used in the field of SP, including infrared thermography (Usamentiaga et al., 2014) or multi-spectral imaging (Chane et al., 2013a). The use of so-called optical targets is intrinsically linked to the practice of photogrammetry whose autonomous calibration methods (*Id is Self-Calibration*) are more recent (Remondino and Fraser, 2006; Stamatopoulos and Fraser, 2014). The «photogrammetric» targets for obtaining metric fulcrums are currently integrated into the methods of *SfM* (DeGol et al., 2018) and their importance from a metrological point of view is indisputable (Nocerino et al., 2014) beyond the optimisation of the network of connectivity of poses (Fraser, 1984). The use of coded targets (Ahn and Schultes, 1997; Calvet et al., 2016) has spread thanks to their robustness. They have become established over time as a solution for simultaneous mapping, localization and mapping methods (e.g. SLAM). These intermediate approaches can be combined with those of the previous type, in particular in the SLAM-visual case and its variations (Taketomi et al., 2017; Menna et al., 2022). Their drawback is the application to multi-temporal survey and impose an intervention on or around the survey area, which is sometimes impossible in real cases.

3.1.3 Data-based registration

The last type of approach is therefore that of data-based alignment and it is the most used in the CH digitisation process. If they do not require external media (integrated sensors, targets, etc.), they do require 2D and or 3D data sources with a certain overlapping threshold. These approaches are more commonly used in the CH field, because they offer more versatility and are implemented in real-based modeling software. However, there is a notable distinction — also valid for the above-mentioned approaches — between two families of methods. A simple or mono-modal case is for example that of the registration of *image to image* from a photogrammetric sequence in the visible domain or the *cloud to cloud* registration of two terrestrial laser scanner stations of the same scanning sequence. The registration of multimodal capture is more complex, because it takes into account the simultaneous variation of several parameters (sensors, resolutions, temporalities, scale jumps, radiocolorimetric drifts). This specific case concerns for example the TLS to photogrammetric digitisation process involving the co-registration of scan data and image set.

3.1.4 2D to 2D registration

2D/2D matching is an internal process of image-based modeling (photogrammetry, SfM, SLAM, etc.) achieved by detecting and matching features common to a pair or set of images. There are many methods from digital image processing (Zitova and Flusser, 2003), many of which are already applied in 3D scanning. Among the well-known and most used algorithms are *SIFT, A-SIFT, PCA-SIFT, SURF, FAST, ORB, BRIEF, BRISK, FREAK, A-KAZE* for the matching and estimation of photogrammetric poses (Apollonio et al., 2014), many competitive approaches (Caron et al., 2014) allow to substitute traditional methods based on the detection of areas of interest (*features detection*). Some methods developed recently cope the main limitation of these classic descriptors; MSD (*Id Est Maximal Self-Dissimilarities*) allows an extension in the spectral domain (Tombari and Di Stefano, 2014); the *SIRF* (Chen et al., 2015) method proposes a format suitable for data fusion; or *POP-SIFT* (Griwodz et al., 2018) proposes a GPU implementation of Lowe's robust algorithm for real-time applications. Some methods can also be used outside of 2D to 2D scope such

as the Mutual Information, a convincing method for 2D/3D (Palma et al., 2010). MI extracts a measure of similarity from statistical sciences to align images (Viola and Wells III, 1997). Without being completely outdated those methods are being challenged nowadays with AI-based ones.

Craft-based VS learning-based methods

This well-posed problem in Computer Vision is experiencing a resurgence of interest with the rise of deep neural networks particularly effective for this task. The above-mentioned algorithms are now categorized as traditional or craft methods as opposed to those based on machine learning (Stathopoulou and Remondino, 2023) also referred to as AI-based. There are many new entries in the literature comparing the most used algorithm, *SIFT*, to its competitors super-vitamins to artificial intelligence (*HP, DISK, LoFTR, SuperGlue, LightGlue*, etc.). Traditional methods are not obsolete, although they may be outclassed in some specific contexts (e.g., mapping of day and night photographs). Those up-to-date methods are becoming more and more accessible. PhotoMatch (Ruiz de Oña et al., 2023) and DIM (Deep-Image-Matching) (Morelli et al., 2024) are, for example, open source tools for testing and comparing the algorithms and offer the possibility to export the features for an image-based modeling purpose.

3.1.5 3D to 3D registration

The problem of 3D/3D data registration is well-known and benefits from decades of research in the reconstruction of a 3D model (Chen and Medioni, 1992). Nowadays the most used method for the case of point clouds is solved by an Iterative Closest Point algorithm (ICP) (Besl and McKay, 1992) calculating the affine transformation between two sets of points by minimizing the distance between a sampling of matching points. This algorithm gave rise to many variants (Rusinkiewicz and Levoy, 2001) evaluated for their resistance to noise, point density variability and robustness against quasi-planar surfaces (Low, 2004). Other work has more specifically addressed the use of PKI in the context of multi-stereoscopic correlation (Zhang, 1994) and their applicability on data acquired by 3D scanning (Pomerleau et al., 2013). However, the methods based on the ICP are not fully automatic, as they require an initial estimate, even a very rough one: sensor positioning, overlay hypothesis or manual presetting. This subproblem is addressed either by dimensional reduction, which consists either in using the multiplicity of 2D views that can be extracted from a 3D model (Huber and Hebert, 2003), or in exploiting the spherical correlation (Makadia et al., 2006). An alternative is based on the extraction of 4 coplanar and congruent points (Aiger et al., 2008), also implemented for automatic laser station consolidation (Theiler et al., 2014). Like its SIFT counterpart, the ICP algorithm is not obsolete despite these weaknesses, which deserve to be known and recognized. Its use will still cover the majority of current CH use cases.

3.1.6 2D to 3D registration

Because of the dimensional leap between data types, the problem of 2D/3D co-registration seems more complex, especially since it can be addressed bilaterally; 2D to 3D, by adding to the image a spatialization in a three-dimensional space or 3D to 2D, by reducing a three-dimensional model to an image plane. The 3D/2D approach is mainly used in the field of *shape analysis* where the selection of the best 2D views or representations (Dutagaci et al., 2010; Moratara et al., 2009) that can be extracted from a 3D model is exploited for pattern recognition and classification (Biasotti et al., 2015). Some of this work is also logically reinvested in the *Next Best-View planning* algorithms. The reverse approach, known as 2D/3D, generally consists of spatial image referencing within a 3D scene. At first, there were manual processes that require the coordinated manipulation of the image and a 3D scene (De Luca et al., 2010) in order to approximately satisfy the homographic relationship, then semi-automatic and now automatic procedures are the prerogative of advanced photogrammetric methods, and more recently of Computer Vision SfM implementation.

3.1.7 Data Fusion

Unlike the fields of remote sensing or medical imaging, which enjoy a more abundant literature, the applications and specificities of data fusion for heritage imaging is less studied. The difficulty in dealing with this problem is explained by the great diversity of heritage objects, their environments and the contexts of their survey, making it difficult to set up and reproduce automated fusion methods. Data fusion in the field of heritage is indeed a major challenge raising many obstacles induced by the accumulation of data variability (resolution gap, radiometric differential, spatio-temporal changes, multiplicity of scales of representation or

observation). Data fusion "appears in the literature in the 1960s as a mathematical model for data manipulation" (Esteban et al., 2005). Several classifications coexist, among which those of Whyte (Fig. 1), Dasarathy or the JDL Data Fusion Group (Steinberg et al., 1999). Data fusion is now understood as a multilevel issue that could be tackled from sensors, data, features, information, or even semantic-oriented layers like decisions. It is defined by various scientific communities, including ISPRS, which proposes « Data fusion is a formal framework in which the means and techniques for combining data from various sources are expressed ». The following definition derived from (Bostrom, 2003) and (Wald, 2002) include nevertheless an important precision for CH applications:

The fusion of data and/or information is the study of efficient methods for transforming variable sources into useful representations in order to increase their meanings.

It highlights that the result of a fusion process must be greater than the inputs cumulated, and differs in this sense from data registration. A data fusion procedure must include an improvement or an added value concerning different objectives:

- recognition (detection, identification of salient information);
- the estimation of a parameter obtained by combining values from different sources;
- the association of previous approaches.

If fusion remains a resolutely complex notion, it can nevertheless be simplified by the Aristotelian adage, "the whole is greater than the sum of its parts".

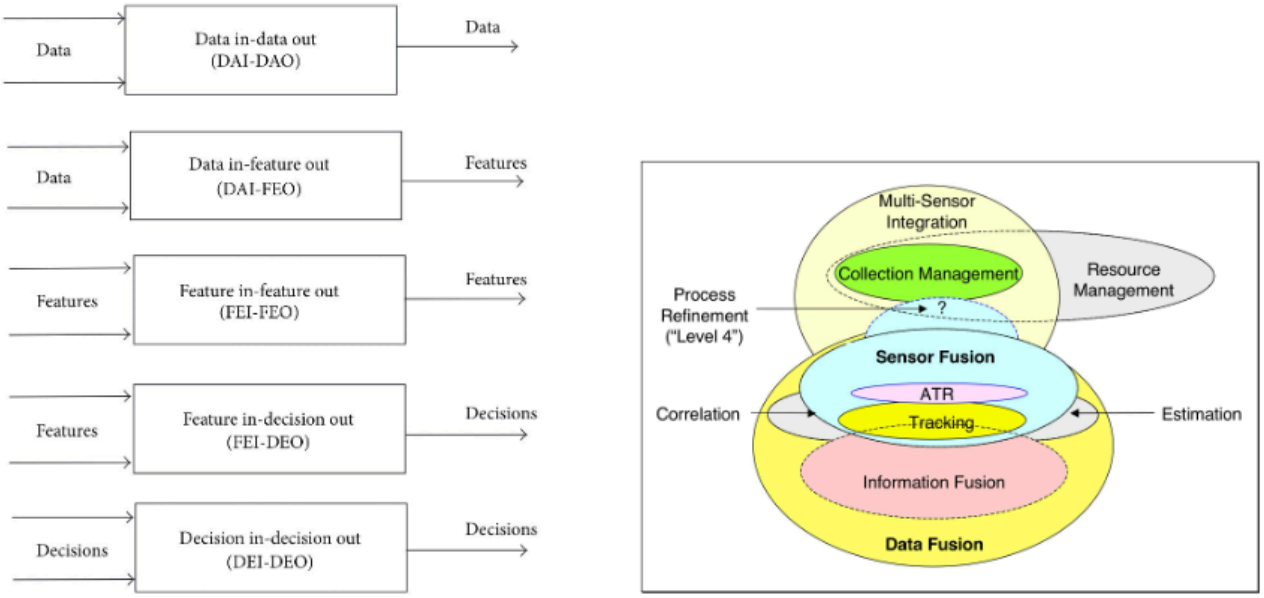

Figure 1– Classification of data fusion or related procedures (Castanedo, 2013; Steinberg et al., 2017).

Many aspects of the data influence fusion processes as analyzed in (Khaleghi et al., 2013), in which one could recognize some characteristics of real-based modeling data in the CH field, such as data imperfection, lack of ground truth, outliers, conflicting data, misalignments. In addition, a proper data fusion process must theoretically include two data fusion metrics enabling it to evaluate its performance. Input quality is estimated or calculated by a Degree of Confidence (DoC). It characterizes the reliability and credibility of sources through a rating system (score, ranking, etc.). For the output quality, the efficiency of the fusion must also be assessed, by the use of Measurement of Performance (MoP). Limits of data fusion are well-known and documented (Hall and Garga, 1999; Hall and Steinberg, 2001) throughout the scientific literature. Few works address data fusion in CH context from a systemic point of view (Ramos and Remondino, 2015; Pamart et al., 2023; Medici et al., 2024). The predominant classification is the one proposed by Ramos and derived from (Klein, 2012), in which the data fusion could intervene :

- at the intervention stage (low levels, intermediate levels, high levels);
- in data dimensions (3D/3D, 2D/3D, 2D/2D);
- based on the characteristics of the data used for fusion (points, characteristics, surfaces, etc.).

CH oriented fusion approaches were recently completed by review articles of (Adamopoulos and Rinaudo, 2019; Adamopoulos and Rinaudo, 2021), highlighting some sensor fusion practices in a wide range of imaging techniques. This study demonstrates and confirms the role of reality-based modelling in their ability to couple with other sources of imaging or measurement. More precisely, referring to (Medici et al., 2024)

only few methods implemented in range or image based software could be accepted as a data fusion process (i.e., when range data is cooperating with photogrammetric refinement of pose and geometry).

3.2 Geometric optimisation

Introduction to geometric optimisation

Geometric optimisation in 3D refers to the techniques and processes aimed at improving the structure and quality of 3D meshes or surfaces to meet specific criteria. One of these criteria is the reduction in 3D complexity, such as lowering polygon counts while preserving visual fidelity and essential geometric features. By optimizing geometric properties, we can create more efficient, accurate, and realistic representations of 3D objects. Several core processes can be carried out within the optimisation process which lead to improved models:

1. *Mesh simplification*, i.e. the process of collapsing edges and vertices whilst preserving object silhouette and curvature;
2. *Removing redundant geometry* such as unseen faced or duplicated vertices and edges;
3. *Regenerating meshes* to improve overall mesh topology for even triangle distribution;
4. *conversion of triangle meshes to quads* in case deformation modelling is required;
5. *Mesh smoothing* through vertex regulation and noise reduction.

The benefits of reducing the polygon count of detailed high polygon meshes, include:

- *Lower polygon count* (fewer vertices and faces) - rendering fewer polygons reduces the computational load on the GPU, resulting in faster frame rates and smoother performance, especially on lower-end devices;
- *Topology improvement* - ensures the even distribution of vertices and avoiding poor quality shapes like long skinny triangles (bad for shading or simulation);
- *Faster load times* - optimised geometry means smaller file sizes, leading to quicker loading and better performance particularly where 3D data is streamed;
- *Better Collision Detection and Physics* – where game engines are used, simpler meshes improve the performance of collision and physics systems by reducing calculations;
- *Improved Animation* – in models where character rigging is required, the use of simple models results in faster skeletal based animations.

General guidance for simplification

Before carrying out any mesh simplification it is generally better to check that the mesh is in the best possible condition which will ensure successful application of any optimisation routines. Several inspections should be carried out, including:

- check for and fix non-manifold geometry, duplicate vertices, and disconnected elements;
- eliminate zero-area faces;
- remove unreferenced or isolated vertices.

Afterwards, good practice includes:

- *Duplicate meshes* before simplifying it to avoid an overwrite of the original mesh;
- I*ncremental optimisation of 3D meshes* in order to:
    - *Preserves Mesh Integrity* - gradual changes help maintain the original topology and structure of the mesh. Whilst sudden or large-scale optimisations can introduce artifacts like self-intersections, non-manifold edges, or distorted geometry.
    - *Better Control Over Quality* - incremental steps allow for fine-tuning and quality checks at each stage. And allows you to monitor metrics like triangle quality, curvature preservation, or texture distortion and stop or adjust the process if needed.
    - *Computational Efficiency* - large meshes can be computationally expensive to optimise all at once.

3.2.1 Automatic decimation algorithms

The most common solutions are based on the *Adaptive Triangle Decimation* (Schroeder et al., 1992) *,* and can typically be found implemented in photogrammetry software or applications such as *Blender*. The algorithm simplifies triangle meshes by removing vertices and retriangulating the resulting holes through the process of vertex classification (simple, complex boundary) and vertex decimation. This method of vertex clustering and collapse is very fast but has less fidelity in detailed areas.

Another algorithm which is utilised within many software platforms is *Quadric Error Metrics (QEM)* (Garland and Heckbert, 1997), used to identify vertices or edges which can be collapsed with minimal distortion. QEM simplifies meshes by using an error metric based on distance from vertices to the original surface. Even though it was created in 1997, many modern approaches build upon or incorporate QEM including MeshLab, Blender, Autodesk 3ds Max ProOptimizer, Simplygon. QEM does have several limitations including its inability at preserving detailed features or textures and it doesn't account for perceptual relevance.
A new Enhanced QEM Algorithm (Lu et al., 2024) has been developed and is an advanced version of the widely used original Quadric Error Metrics (QEM) algorithm. Using *KD-tree principles* it searches for valid point pairs and evaluated them based upon several additional simplification factors including:

- *Vertex attribute preservation* (normal, colour, textures), which prevents texture distortion;
- *Shape preservation* by constraining additional vertex factors (vertex density, curvature and distance);
- The use of *weighted quadrics* to prioritise detailed areas and can be tuned to perceptual importance, preserving visually significant areas.

These new features produce optimisation improvements including superior feature preservation, efficient simplification and scalable performance particularly for high-poly models.
The most time-effective solution will always be *automatic decimation*, especially in conjunction with scripts for batch processing of the models, typically *Houdini*. Automatic mesh production approach requires a good-quality dataset to produce a 3D mesh without severe surface noise, misalignment artifacts, holes in the surface mesh, spikes, and non-manifold geometry. Contemporary applications such as *Reality Scan (Reality Capture)*, have a vast tool set for lossless smoothing and decimation of the resulting reconstructed 3D meshes before the export in a vast spectrum of available formats.

3.2.2 Semi-automatic retopology

State-of-the-art algorithms, which can be found as standalone open or closed-source applications such as *Quad remesher*, are available as a plugin for different “industry-standard” applications (originated from *Zbrush*), where there is some input from the user, but very minimal to guide the process of the retopology to a desirable direction. On other hand there are more sophisticated and controlled solutions such as *Instant meshes*, where *Position Field* generation based on curvature of the topology of the model is used to compute a smooth guidance field to direct the flow of the quads and then *Orientation field* generation step allows user to correct generated cross field with convenient GUI tools. Then the software solves a UV layout and aligns it to cross field, and converts the mesh into quads layout.
Mixed Integer Quadrangulation (MIQ) techniques are proposed by Bommes et al. (2009), while *"Instant Field-Aligned Meshes"* by Jakob et al. (2015).
The final step is fully automatic, and it converts the parametrized UV grid into quads and forms a quad mesh as the final result. There are several optimisation approaches which incorporate the use of machine learning, convolutional neural networks and directional fields some of which are outlined hereafter.

Instant Meshes
The *Instant Meshes algorithm* (Jakob et al, 2015) calculates a smooth direction field over the surface that aligns with geometric features (e.g., curvature, sharp edges) and determines where vertices would be optimally placed to follow the orientation of this field, with evenly spaced triangles or quads. This approach provides quick processing speed and scalability, being able to handle large datasets. It also preserves features well and will align and snap to features naturally. Instant Meshes has several shortcomings in that it may not achieve the same precision as globally optimised methods and complex geometries can lead to singularities in the orientation field, which may affect mesh quality. This software is available as open source (https://github.com/wjakob/instant-meshes) and has been implemented in Foundry’s Modo software providing their auto-retopology tool.

MeshCNN
MeshCNN (Hanocka et al., 2019) is a deep learning architecture designed specifically for 3D triangular meshes, and adapts traditional convolutional neural networks (CNNs) which are normally used to analyse images to work directly on mesh edges which are treated like pixels. Through mesh convolution learned edge collapse based upon either segmentation or classification tasks, resulting in a reduced mesh whilst preserving key features. Advantages of using this approach is the algorithm learns which parts of the mesh are important whilst maintaining mesh topology. The tool is limited to triangular meshes only and can be computationally intensive when applying it to large meshes. MeshCNN has not been integrated into any

desktop 3D software to date, but is available as an open-source PyTorch implementation on GitHub (https://github.com/ranahanocka/MeshCNN).

FlexiCubes
FlexiCubes (Shen, T. et al., 2023) is a 3D mesh representation and optimization technique developed by NVIDIA, integrated into their Kaolin library—a PyTorch-based framework for 3D deep learning. FlexiCubes is an isosurface extraction method designed for gradient-based mesh optimisation. It improves upon traditional methods like *Marching Cubes* (Lorensen & Cline, 1987) and *Dual Contouring* (Chen et al., 2022c) by introducing flexible learnable parameters that can be optimised during training or reconstruction, resulting in more accurate and detailed 3D meshes. This approach enhances mesh quality by introducing additional degrees of freedom (flexible vertex positioning, quad splitting, and grid deformation) whilst maintaining topological integrity. Advantages include the local adjustment of mesh vertices, producing better fine fit details and producing smoother more accurate meshes. The algorithm can also be used on both surface and volumetric meshes. The process is still experimental and requires training. FlexiCubes is already being used in several advanced tools and frameworks: including NVIDIA Kaolin (from v0.15.0) and GET3D, a generative AI modeller, and is available via GitHub (https://github.com/nv-tlabs/FlexiCubes).

ASimp
The ASimp (Automatic Simplification) (Lin et al., 2025) mesh optimisation algorithm is designed to simplify 3D models while preserving visual quality and user experience and bridges the gap between technical mesh reduction and human visual satisfaction. It uses the established QEM algorithm for actual mesh reduction but is guided by ASimpNet's that predicts optimal simplification ratios and has been trained on other high-poly meshes. ASimp incorporates Quality of Experience (QoE) metrics by analysing how users perceive visual differences in simplified models improves the outcome and achieves good results across a range of mesh criteria (chamfer distance, normal dissimilarity, watertightness, Laplacian Eigenvector Error). However, its effectiveness relies on the quality and diversity of the user perception data used to train it. ASimp is currently in prototype form and is therefore currently not available for public download as a standalone tool or open-source library.

3.2.3 Surface editing and Manual retopology of 3D models

When dealing with non-collaborative surfaces, it is common to get noisy imperfect results in terms of their geometric features. The problem with photogrammetric reconstructions consists mainly of the shininess of the surfaces of the objects captured at different viewing angles on the photographs. There are limited solutions to this issue such as cross polarization; however, it is not so effective in practice, and it is not possible to remove all shininess from the surfaces of the objects. Assuming good camera orientation (SfM) is already achieved (and we are meshing with the MVS algorithm), the resulting mesh typically will have some severe noise regions where the most shininess was present on photographs, severe bulb noise in shadow regions because of the featureless and noisy input data. Geometric features of the object may appear smeared, and thin parts may have protrusions or holes, all surrounded by a constellation of small floating geometry - "floaters".
In some cases, the automatic solution won't be effective and will lead to very smoothed out details, which also will lead to misprojections of the texture. In such cases, it is required to perform a *3D scan cleaning procedure*, which will fix all issues described above; however, it requires time-consuming manual sculpting to produce a scientific grade reconstruction of the high-poly model. In such cases, manual retopology of the automatically fixed and decimated mesh in photogrammetric software is the most time-efficient solution. It still requires a highly trained digital artist, but it takes less time because the manual retopology can be started right away, surpassing the time-consuming process of 3D mesh cleaning. The resulting Low-poly model will be easily unwrapped. UV islands are packed for the optimal texture reprojection and post-processing (if required), with fully controlled texel density for the chosen production environment.

3.2.4 Optimisation tools

Several software solutions are available to incorporate into the 3D processing pipeline which incorporate many of the optimisation functions described previously. The table below summaries some available tools and algorithms and outlines their key features. The range of tools can be used as part of a standalone software package, as an optional plugin or as a webservice/API.

Name	Optimisation Algorithm(s)	Key Features	Type
Blender	Decimate Modifier	● Collapse Mode - best for general mesh simplification ● Un-Subdivide Mode - best for meshes that were previously subdivided ● Planar Mode - merges faces that lie on the same plane (good for architectural models) ● Optional RapidPipeline Add-On	Software
MeshLab	Quadric Error Metrics (QEM) Clustering Decimation	● Quadric Edge Collapse Decimation & Clustering Decimation ● Repair tools also available	Software
Geomagic (Design X & Wrap)	N/A Propriety	● Decimate – reduces the number of polygons while preserving the overall shape and detail (includes controls for balancing quality and performance) ● Enhance Shape - sharpens corners and smooths flat or rounded areas ● Global Remesh - reconstructs the mesh with a uniform triangle size ● Optimize Mesh - refines both the structure and geometry of the mesh	Software
Rhino	Advancing-front algorithm	● TRmesh Plugin - remeshing and topology optimisation for Breps, meshes, and point clouds ● Render Mesh Analysis Tools	Software
Instant Meshes	Instant Field-Aligned Meshes	● Produces clean, low-polygon mesh that follows the shape and features of the original ● Integrated in Foundry Modo Software	Software Library
RapidPipeline	Quadric Error Metrics (QEM) Edge Length Method	● Automated solution ● Mesh Decimation with advanced controls for balancing quality and performance. ● Remeshing – uses voxelization or shrink wrap to create new meshes	Web Platform, API, Blender Add-On
ZBrush	N/A Propriety Voxel-based algorithm Clustering Decimation	● ZRemesher - automatically retopologizes high-resolution sculpts into clean, low-poly meshes ● Local ZRemesher - enables remeshing on specific regions. ● DynaMesh – Maintains uniform polygon distribution during sculpting (good for organic shapes) ● Decimation Master - Reduces polygon count while preserving surface detail which supports batch processing and maintains UVs and textures	Software
Autodesk Maya	Quadric Error Metrics (QEM) & Vertex Collapse	● Reduce Mesh Tool - adjustable reduction (%) and options to preserve UVs, borders, and hard edges ● Paint Reduce Weights function – enables selective mesh reduction by painting areas to preserve or simplify ● Additional plugin available	Software
Autodesk 3ds Max	Quadric Error Metrics (QEM) & Vertex Collapse	● ProOptimizer Modifier – has relative (%) or absolute (count) reduction control and can optionally preserve normals, UVs and materials ● Additional plugin available	Software

Reality Capture / Reality Scan	N/A Propriety	• Mesh Simplification - has relative (%) or absolute (count) reduction control • Options to control minimal triangle edge length to minimise the creation of distorted triangles • Border Decimation – option to retain border detail • Density Equalization – enables consistent distribution of vertices in mesh	Software
Agisoft Metashape	Collapse Edges Quadric Edge Collapse	• Decimate/ Mesh Simplification - has relative (%) or absolute (count) reduction control • Option to preserve mesh boundaries and UV available	Software
Simplygon	Volumetric remeshing (Voxel)	• Focus on 3d gaming assets where target polygon counts are based on screen size of final object • ReductionProcessor - Heuristic based vertex and triangle reduction • Triangle Reducer & Quad Reducer - reduces the number of triangles and vertices. Option to take into account UV coordinates, tangents, normal and vertex colors	Software Plugin
Houdini		• PolyReduce Surface Operator (SOP) - reduces geometry complexity (useful for creating LODs) • Remesh SOP - rebuilds geometry with uniform triangles (improves mesh quality before retopology) • Fuse SOP - merges nearby points to clean up geometry • Clean SOP – removes unused ponts and mesh artefacts • Normal & Facet SOPs – Recalculate and smooth normal to fix shading issues • Custom VEX scripts allow procedural geometry manipulation	Software

3.3 Texture mapping and optimisation

Texturing in 3D is a crucial technique used to enhance the visual realism and detail of digital models. While a 3D mesh defines the shape and structure of an object, textures bring it to life by adding surface characteristics such as colour, roughness, and reflectivity characteristics which we normally see with our eyes when observing the real world. The process of applying textures enables us to simulate materials like wood, metal, skin, or fabric without increasing the model's geometric complexity. By applying textures, 3D assets become more immersive and believable, bridging the gap between raw geometry and rich visualisation making digital objects and environments feel tangible and engaging.

In reality-based 3D modelling texturing methods play a vital role in accurately representing the real-world appearance of objects and environments. These methods focus on capturing and applying surface details like colour, wear, and material properties and can inform the viewer of the condition of the object as much as the geometry does.

Texturing covers a range of methods which can be integrated linearly into a pipeline, with each method playing an integral part to the creation of effective, realistic and efficient 3D objects. The normal order or texture processing following the creation of clean mesh models is:

1. UV Unwrapping;
2. Texture Creation;
3. Baking Maps;
4. Material Setup;
5. Application & Adjustment;
6. Texture Optimization;
7. Texture Testing & Review.

3.3.1 Texturing methods in photogrammetry software

View Selection Methods

Single-View Projection or Best-View Selection: they are based on picking the *best photograph for each triangle* of the 3D mesh where the best is defined by different quality metrics. This method can produce very sharp textures since it uses original, unblended pixel data, however often results in visible, harsh seams between adjacent faces that were textured from different photos with different lighting or white balance.

View-Angle Selection (Normal-based): it selects the photograph where the camera's viewing direction is most parallel to the surface normal of the mesh face minimizing distortion.

Resolution-based Selection: This algorithm calculates which photograph will provide the most detail for a given mesh face. Useful when it is required to separate close-up shots from wide shots.

Sharpness-based Selection (Focus/Blurriness Metric): it analyzes the sharpness of the source images by measuring high-frequency detail or Laplacian variance to discard blurry images.

Occlusion-aware Selection: before projecting a texture, the algorithm ensures there is a clear line-of-sight from the camera to the mesh face, with no other parts of the mesh self occluding.

Data Blending methods

These are more advanced methods that combine pixel data from multiple suitable photographs to create a smooth, seamless texture.

Weighted Averaging: a simple blending method. For each pixel on the final texture, it takes a weighted average of the colors from all valid source photos. The weights are typically based on the quality metrics from Category 1 (view-angle, resolution, etc.). Photos that are better contribute more to the final color.

Multi-Band Blending: this method decomposes each image into different spatial frequency bands. It then blends the low-frequency bands (colors, tones) over a wide area and the high-frequency bands (details, edges) over a narrow area. This creates a transition that is imperceptible to the human eye.

Graph-Cut Optimization (Seam Finding): this approach treats the problem as finding an optimal *seam* that cuts across the texture. The algorithm finds the path preferring to place seams in low-contrast or low-detail areas where they will be less noticeable.

3.3.2 UV Mapping and methodologies

UV mapping is a foundational process in 3D texturing that involves projecting a 2D image texture onto a 3D model. Since textures are inherently two-dimensional, they need a coordinate system to wrap correctly around a three-dimensional surface. This is where UV coordinates come in—"U" and "V" represent the horizontal(X) and vertical axes (Y) of the 2D texture space. During UV mapping, the 3D model is "unwrapped" into a flat 2D layout, much like unfolding a cardboard box or wrapping a present. This layout allows the 3D modeller to paint or apply textures with precision, ensuring that every part of the model receives the correct visual detail without distortion.

The fundamentals of UV mapping include defining seams or strategically placed cuts around the mesh surface, which follow vertices and edges, unwrapping the mesh, and arranging these individual pieces or islands efficiently within the UV space. Once unwrapped, the resulting UV islands are scaled, rotated, and packed to make optimal use of the texture area.

A well-constructed UV map limits the potential texture distortion through the minimization of stretching and overlapping, which is crucial for achieving high-quality, realistic textures. This process is especially important in photogrammetric, and reality based workflows that rely on photographic textures or detailed surfaces to accurately model the surface characteristics of heritage objects.

Unwrapping methods in UV mapping are critically important because they directly affect how well a 2D texture conforms to the surface of a 3D model. Different unwrapping techniques—such as planar, cylindrical, spherical, or manual seam-based unwrapping—are chosen based on the shape and complexity of the model. A well-chosen method ensures that the UV layout minimizes distortion, stretching, and overlapping, which are common issues that can degrade texture quality. For example, using cylindrical unwrapping for a pipe-like object or planar unwrapping for flat surfaces helps maintain texture fidelity and alignment. One of the challenges with heritage objects is they often don’t conform to simple geometric primitives and are highly organic shapes with careful strategy to preserve texture fidelity and minimize distortion. These shapes often have complex curves and irregular topology, making standard projection methods (like planar or cylindrical) insufficient. Usually, manual seam placements are utilised to guide the unwrapping process placed in less visible areas (to allow the mesh to unfold naturally into manageable UV islands. This approach helps maintain proportional texture distribution and reduces stretching.

Moreover, the unwrapping method influences how efficiently the texture space is used. A clean, organized UV layout allows for better resolution distribution across the model, which is especially important in the detailed modelling of heritage objects. Poor unwrapping can lead to wasted texture space, visible seams, or mismatched details, making the model look unrealistic or unprofessional. Therefore, selecting the right unwrapping method is a foundational step in achieving high-quality, visually accurate 3D texturing.

Automatic UV Unwrapping

Automatic unwrapping algorithms provide a procedural solution for generating UV layouts. These methods are most effective for models with simple geometry and a limited number of complex topological features. While they are highly time-efficient and can produce satisfactory results for many applications, they often lack the precision required for complex assets or specific texturing requirements. These methods use algorithms to automatically place seams, unwrap the mesh, and pack the resulting UV islands. While they may not offer the same level of control as manual unwrapping, they are incredibly useful for speeding up workflows or handling large datasets.

Below is an overview of a selection of relevant automated UV-unwrapping algorithms, the major 3D tools that implement them, and their respective benefits and shortcomings and what type of heritage objects utilise these methods.

Method	Algorithm	Software	Benefits	Shortcomings	Relevant Heritage Objects
Projection-Based Unwrapping	Planar, cylindrical, spherical, cubic/box projections	Blender Cinema 4D Modo Ultimate Unwrap 3D	Extremely fast, real-time results No manual seam placement required Ideal for simple or predominantly flat geometry	High distortion on curved or complex surfaces Overlapping UVs are common without manual adjustments Limited packing optimization—often requires manual island arrangement	Cylindrical or Conical (vases, amphora) Architectural features (columns) Geometrically simple shapes (barrel vaulted ceilings)
Energy-Based Flattening (EBF)	Angle-Based Flattening Least Squares Conformal Map;	Blender Modo Angle Based Unwrap Cinema 4D RizomUV	Minimizes angular or area distortion across the UV map Produces globally smoother unwraps for organic shapes	Relies on user-placed seams for optimal results Increased compassion times with mesh complexity Can create localized stretching without further manual tuning	Sculptural Reliefs and Figurines Rock Art and Petroglyphs Architectural Fragments Fragmented or Deformed Objects
Iterative Stretch Minimization (SLIM)	Spectral Least Isometric Maps (SLIM) iteratively optimizing the surface parameterization to reduce	Blender UVPackmaster plugin RizomUV	Reduces distortion User control of quality vs computation time Offers a standalone “Minimize Stretch” operator to improve existing	High-quality (many iterations) runs can be slower than simpler methods Newer integration therefore the tool is still	Highly Detailed Sculptures and Figurine Engraved or Embossed Artifacts Organic or

	local stretching		unwraps	maturing	Freeform Objects
Heuristic, Topology-Aware One-Click Unwrapping	Rule-based methods that make educated guesses to solve problems efficiently. Identify optimal seam placement / Minimize texture distortion / Preserve features Topology-Aware - Option of either Hard surface or organic topologies	Ministry of Flat Unwrella Plugin RizomUV	Eliminates manual seam-setting and island-tweaking Consistent, production-ready UVs for diverse asset types Scalability for batch processing and pipeline integration	Limited fine-tuning—acts as a “black box” Can sometimes create too many seams and islands than necessary	Moderately Complex Artifacts Museum-Scale Collections - One-click unwrapping enables batch processing

Manual and Semi-Automatic Unwrapping
Manual or semi-automatic unwrapping is essential when precise control over texture application is required. This approach is necessary for models exhibiting complex topology, such as concavities, perforations, or non-spherical forms (e.g., a handle on a vessel). Manual control is indispensable for:

- *Controlling Texture Projection:* ensuring that textures are applied without distortion or seams in visually critical areas;
- *Optimizing Texture Space:* efficiently stitching and arranging UV islands to maximize the use of the texture area;
- *Facilitating Post-Processing:* creating a logical and human-readable UV layout that simplifies subsequent editing of the texture maps in image manipulation software;
- *Fixing problematic auto-unwraps*: when automatic methods produce overlapping, inverted or stretched UV maps and ensuring UV islands are within gutter area of the map;
- *Critical Areas*: where accurate mapping of surface features (e.g., cracks, pigment traces) is critical manual editing of the UV map maybe required.

Manual UV mapping provides granular control over the final appearance of the textured asset. One potential combination of approaches is to initially use a range of automatically produce a UV map which can then be manually edited and fine tuned to improve its effectiveness and efficiency. Some UV tools such as RizomUV and 3ds Max UVW Unwrap Modifier have the ability to display area and angle deformation which can be manually reduced through a combination of UV vertex editing, creation of additional seams or islands. Depending on the size of the mesh this can be computationally difficult to dynamically display in 3ds Max however RizomUV handles this better.

3.3.3 Resolution and Texel Density

*Resolution o*f texture maps details the dimensions of a texture map, typically expressed as width × height (e.g., 2048×2048). It determines the level of detail that can be displayed on the surface of the model. Higher resolutions allow for finer details, such as object details but they also increase memory usage and rendering time. Balancing resolution with performance requirements is a decision made based on where and how a 3D asset will be used. Where users have access to high end rendering capabilities or where the number of models is limited in a scene, higher resolutions should be used. Where performance and memory efficiency are an issue such as mobile or VR platforms or where large scale environments are used where many assets are rendered simultaneously lower resolution texture maps are better.

Texture map sizes typically follow standardised resolutions that are powers of two, ensuring compatibility with most rendering engines, mipmap-compatible and optimises memory usage and performance. Here are the most common sizes:

Texture Resolution	Environment Use Case	Single Object Use Cases
256×256	Low-detail game assets, background elements	Not normally used
512×512	Moderately detailed props, mobile and web assets	Not normally used
1024×1024 (1K)	Common for medium-detail assets	Mobile models
2048×2048 (2K)	High-quality game assets	Common for low-quality mobile models
4096×4096 (4K)	Cinematics renders, close-up game assets	Common for Normal-quality model
8192×8192 (8K)	large scene environments, tiled surfaces	High-quality model
16384×16384 (16K)	Very large-scale environments	Very high-quality model

When deciding on a suitable texture resolution consideration should also be made on how this will impact what level of detail can be transferred through the baking process from the high-poly model. The resolution size proportionally limits the maximum number of polygons in the high-poly model which can be represented in the normal or displacement maps, e.g. the maximum number of polygons an 8K texture can represent in the texture map is equal to the number of pixels (approximately 67 million polygons).
Texel density is a critical metric that defines the resolution of a texture as applied to a 3D surface, typically measured in pixels per meter (px/m) or pixels per centimeter (px/cm). It is analogous to dots per inch (DPI) in print media but is applied to a model's real-world scale.
For the digitisation of small artifacts, the objective is often to maximize texel density to capture the finest surface details. High-resolution photography used in the photogrammetric texturing process can yield a very high texel density. Depending on project requirements 100-400 px/cm² is considered *High-Fidelity archival quality* for digital visualization.
For large-scale models, such as buildings or landscapes, maintaining a uniform and adequate texel density across the entire surface with a single large *megatexture* is computationally prohibitive. A more effective strategy is to segment the model and assign multiple materials which correspond to a separate UV layout and texture set, allowing for high texel density to be achieved across vast surfaces by distributing the texture data over dozens of UV tiles.

3.3.4 Advanced Texturing Systems: UDIM (U-Dimension)

The UDIM system is an advanced texturing technique that extends the standard 0-to-1 UV space. It allows a single 3D model to utilize a grid of *UV tiles,* with each tile holding a separate texture map. This methodology is particularly advantageous for assets requiring extremely high resolution, such as vast landscape 3D models or architectural visualizations.

Key benefits of the UDIM workflow include:

- *Massive Texture Resolution:* enables the application of numerous high-resolution textures to different parts of a mesh, achieving very high texel density on large-scale objects, enabling efficient memory usage .
- *Modular Material Control:* allows for distinct shader properties and material attributes to be assigned to each UDIM tile, providing granular control over the final look of an asset.

It is important to follow good practice when creating UDIMs, including:

- *Planning the content of each texture map* to correspond to different components of your heritage object, e.g., A UDIM for a historical building could have each room texture on separate UDIM tiles;
- *Use naming conventions* which correspond to the positions in UV space;
- *Keep texel density consistent* across UDIM tiles.

3.3.5 Baking Texture Maps

Texture baking is a process in texture creation for n 3D models which enables the process of transferring complex surface details, lighting, and shading information from a high-poly model or shader setup onto an optimised low-poly model using 2D image textures. This results in the optimisation textures for real-time rendering producing visually rich results with minimal geometry. Several core baking techniques are commonly used:
Normal Map: captures surface detail from the high-poly model such as fine bumps, grooves, and sculpted features by encoding them into a tangent-space (RGB colour values) or object-space normal map.
Ambient Occlusion (AO): greyscale map which simulates how ambient light does illuminate recessed areas, and enhances depth and realism by adding soft shadows, e.g. shaded areas in the folds of a sculpture.
Curvature: greyscale map highlights the convex and concave nature of a 3D model's surface with dark areas representing concave regions, bright areas being convex and flat surfaces as grey.
Position: captures the 3D object space coordinates of each point (X, Y, and Z position) on a model's surface into RGB colour values.
Additional maps can be generated depending upon your software platform including *ID* (used to isolate/mask specific areas of a model texture), *Cavity* (similar to AO maps but used to shade micro-surface features), *Height/Displacement* (a more resources intensive alternative to normal maps and represents elevation or depth of surface of features and the silhouette) and *Thickness* maps (grayscale texture representing the distance between the front and back surfaces of a model, used in model where light transmission through the object is key, e.g. thin bone artefact).
There are several software solutions to texture baking including dedicated products (Adobe Substance Painter, Marmoset Toolbag and X Normal) and functions within regular 3D modelling software such as 3dsMax. ZBrush and Blender. Current advances in texture baking include:

- *Real-Time Baking Engines* - enable preview and realtime control of baking variables including cage adjustment (Marmoset Tooolbag);
- AI-Assisted Baking - can automatically reduce baking artefacts (light bleeds, shadow seams) and adjust baking parameters (cage settings);
- Advanced Lightmap and AO Baking - material aware lightmap baking (RapidPipeline);
- Multi-UV Baking - support for UDIMS with the baking (Marmoset Tooolbag, Blender, Substance Painter) which normally requires a workaround.

3.3.6 Material and Texture Map Generation for Physically Based Rendering

Physically Based Rendering (PBR) is the prevailing methodology for creating realistic materials in 3D visualization. It aims to simulate the interaction of light with surfaces in a physically plausible manner, ensuring that assets react realistically under any lighting condition. This is achieved through a combination of a specialized shader and a set of texture maps that describe the physical properties of a surface.

The two primary PBR workflows are:

- *Metallic-Roughness:* widely adopted in real-time applications such as video games and web viewers. It uses base color (*Albedo*), *Metallic, Roughness, Normal and Ambient Occlusion (AO) maps.*
- *Specular-Glossiness:* historically common in offline rendering for VFX, as it can offer more nuanced control over non-metallic reflectivity. It uses Diffuse color, Specular, Glossiness*, Normal and Ambient Occlusion (AO) maps.*

Modern real-time engines, such as *Unreal Engine 5*, have advanced their shader capabilities to incorporate elements of both workflows, for instance, by enabling direct control over *Fresnel reflectance* (F0, F90) utilizing new shader framework *Substrate*.

Common PBR texture maps

1. *Albedo/Base Color:* the diffuse colour of the surface, theoretically without lighting or shadow information.

2. *Normal:* simulates how light is rendered on a surface creating fine detail without increasing geometric complexity.
3. *Height/Displacement:* modifies the actual geometry to create significant surface details.
4. *Metallic:* a grayscale map defining which parts of a surface are metallic (white) or dielectric/non-metallic (black).
5. *Roughness (Metallic-Roughness workflow):* a grayscale map controlling the microsurface scattering of light, defining how rough or smooth a surface appears.Plays crucial role in a Transmission of the light in a material (see Transmission map below).
   In Specular-Glossiness workflow this map has an inverted grayscale value.
6. *Ambient Occlusion (AO):* simulates soft shadowing in vast crevices and occluded areas, adding depth.
7. *Transmission (PBR - Based Transparency):* this is the modern, physically accurate approach for materials that light can pass through, like glass, water, jewels, or clear plastic. It simulates how light refracts (bends) and gets absorbed as it travels through the medium.
8. *Emissive*: a map which defines which part of the 3D model appear to glow as if emitting light. Values can range from 0 (black) to white or RGB value if the glow has an associated colour.

- Typically controlled by a slider from 0-1 where 0 - opaque and 1- fully transparent.
- Can be assigned to a shader as a black and white map where 0 - black and 1 - white.
- It has input for an IOR (index of refraction) to control the type of behaviour of material.
- Attenuation (Absorption color) which simulates light being absorbed while it travels through media (material volume).

There are some differences for the *Metallic and Smoothness workflows*, however, they are similar. In both cases the metallic channel map has to be set to non-metallic and Color Albedo or Diffuse map do not have a direct influence on the color of the fully transparent glass. All effects as frosted glass or smooth reflection controlled by a Roughness-Smoothness (Glossiness) map.
Translucency/Subsurface Scattering (SSS): Simulates light penetrating the surface of a material and scattering within it, essential for organic materials like skin or wax.

There are other maps for both PBR workflows, however, they are not commonly used in the CH visualization applications or have more technical implementations.

3.3.7 PBR Workflow in Photogrammetry

A specialized PBR workflow can be employed in photogrammetry to capture material properties directly from imagery:

1. An Albedo texture is generated from images captured using cross-polarized lighting, which minimizes specular reflections and surface shadows resulting in close to a true albedo color of the captured surface.
2. A separate set of images is captured using linear-polarized light.
3. By computationally extracting the Albedo data (from the cross-polarized set) from the linear-polarized images, the remaining data represents the surface's specularity. This result is converted to a grayscale map that functions as a Glossiness (Smoothness map) and if required can be inverted to Roughness (depending on workflow), providing a “ground truth” basis for the material's surface reflectivity.

3.3.8 Textures Format and Optimisation

Texture Formats and Bit Depth

- *Common Formats:* lossless formats like PNG and TIFF are preferred for source and data-critical maps. Lossy formats like JPG are used for delivery where file size is a primary concern.
- *Bit Depth:* for maps containing subtle gradients or precise data, higher bit depth is crucial.
  - *Normal Maps:* typically 16-bit to prevent banding artifacts.
  - *Displacement Maps:* typically 32-bit floating-point (e.g., in TIFF or EXR format) to represent true displacement values accurately.

Channel Packing
To optimize memory usage and shader draw calls (texture fetches), single-channel grayscale maps (like Metallic, Roughness, AO) are often packed into the Red, Green, and Blue channels of a single RGB image. In Unreal Engine, this is commonly referred to as an *ORM texture (Occlusion in Red, Roughness in Green, Metallic in Blue)*. In Unity, this is often termed a *texture mask.*

Platform-Specific Considerations (Asset Dressing)
Assets must be correctly prepared for their target rendering engine. A critical example is the normal map format, which differs between rendering APIs:

- DirectX (used by Unreal Engine) interprets the Y-axis (Green channel) as pointing down (-Y).
- OpenGL (used by Unity, Blender) interprets the Y-axis as pointing up (+Y).
  Normal maps must be generated or converted to match the target engine's standard to render correctly.

Advanced Data Integration and Processing
Data Fusion for Texturing: This technique involves aligning datasets from different acquisition methods to leverage the strengths of each. For example, a geometrically precise but untextured model from a structured light or laser scanner can be aligned with a less accurate but photorealistic model from photogrammetry. The high-resolution color data from the photogrammetry model is then projected, or "baked," onto the geometrically superior mesh, resulting in a final asset with high geometric and textural fidelity.
AI-Enhanced Post-Processing: AI models are increasingly used for texture post-processing tasks, such as intelligent upscaling (super-resolution) to increase texture dimensions while preserving or generating plausible detail, and for artifact removal or denoising.

Texture Data Optimisation via Compression
High-resolution textures present a significant data management challenge. Traditional image storage formats offer various compression methods, but these are primarily used for reducing file size on the storage disk(HDD/SSD). Lossless formats such as TIFF using LZW compression or PNG provide great quality but must be fully decompressed into memory for GPU access. This also is true for formats like TGA, (with optional RLE compression) and format like JPEG, while effective for storage, requires decompression to raw bitmap data before being loaded into Video RAM (VRAM), thereby consuming an enormous amount of VRAM. Hardware texture compression mitigates this bottleneck by utilizing specialized, block-based algorithms. Natively decoded by the GPU, these formats allow texture data to remain compressed in VRAM, substantially reducing its memory footprint which leads to benefits such as:

- Less VRAM Usage: compressed textures take up far less video memory, allowing more assets to be loaded at once for richer, more detailed scenes.
- Higher Memory Bandwidth: smaller textures free up the data pipeline between the GPU and its memory resulting in higher and more stable frame rates (FPS).
- Quicker Loading: smaller files load faster from storage into memory, reducing wait times.
- Effective GPU Caching: more of the compact texture data can fit into the GPU’s high-speed cache, which minimizes delays from fetching data from slower VRAM.
- Power Savings: on mobile devices, less data movement means less energy consumption, leading to longer battery life.

4. Types of 3D data and formats

The choice of appropriate 3D data formats is important for ensuring long-term preservation, accessibility, and reusability of digitized cultural heritage assets. The selection of formats significantly impacts data interoperability, workflow efficiency, and the ability to share and reuse 3D content across different platforms and applications over time. Within the framework of the Data Space for Cultural Heritage, it becomes even more important to have a clear and consistent understanding of how these formats are handled, especially when they are exchanged, visualised, or integrated through online platforms and APIs. Therefore, defining and adopting clear technical standards for 3D file formats, already addressing their choice based also on MIME (Multipurpose Internet Mail Extensions) types, a standardized way for managing file in web technologies, is a key step towards ensuring that digital cultural resources remain accessible, usable, and interoperable for current and future generations.

4.1 Classification of 3D Data Types

3D cultural heritage data can be categorized into several fundamental types, each serving specific documentation and visualization purposes:

- **Point Clouds**: collections of 3D coordinates representing the surface geometry of objects or environments. These are typically unstructured data (e.g., as output of the photogrammetric workflow), and serve as the foundation for further 3D model generation.
- **Mesh Models**: Structured 3D representations consisting of vertices, edges, and faces that define the surface geometry. Meshes can be further enhanced with texture mapping, material properties, and other visual attributes to create photorealistic digital replicas.
- **Parametric Models**: Mathematical representations that describe 3D geometry through parameters and constraints, commonly used in Computer-Aided Design (CAD) applications for technical documentation and engineering purposes.
- **Volumetric Data**: Three-dimensional arrays of data points representing properties throughout a volume, often used for scientific analysis and non-destructive investigation of heritage objects.

4.2 3D File Format Analysis

The landscape of 3D file formats presents a complex array of options, each optimized for specific use cases and workflows. The following analysis examines key formats relevant to cultural heritage applications:

- **Open and Standardized Formats**
  LAS (LASer File Format): An open binary format standardized by *ASPRS* (*American Society for Photogrammetry and Remote Sensing*), widely used in geomatics. Data is stored in an unstructured way: all points are recorded in a single point cloud, without distinction between individual scans. The header contains rich metadata, and the format is extensible via additional fields (extra bytes). It is the most widely used format for storing, analyzing, and exchanging *unstructured point clouds.*
  LAZ (Compressed LAS): A compressed and lossless version of the LAS format, created using the LASzip tool. Although not formally standardized, it is considered a de facto standard for the transmission and preservation of compressed point cloud data. It retains the same structure, header, and metadata as LAS, and is fully compatible with libraries that support LASzip. Ideal for online publishing or efficient transfer of large datasets.
  E57: A binary file format with an XML header, standardized by *ASTM Internationa*l (formerly *American Society for Testing and Materials*) to ensure interoperability between scanners from different manufacturers. It stores data in a *structured* form, preserving the organization of individual scans (poses, timestamps, associated images), and supports the integration of multiple point clouds. Metadata is extensible via custom XML tags. Particularly suitable for the complete and structured archiving of complex 3D laser scanning data.
  PLY (Polygon File Format): An open format developed at *Stanford*, available in both binary and ASCII versions. Originally designed for 3D meshes, it is also widely used for *unstructured point clouds* thanks to its flexible and extensible structure: the header explicitly defines the fields included (coordinates, RGB color, intensity, etc.). Its simplicity and widespread support make it ideal for archival purposes and data exchange between different software platforms.
  OBJ (Wavefront OBJ): A widely adopted open format supporting mesh geometry, materials, and textures through companion MTL (*Material Template Library*) files. Despite its limitations in supporting advanced features like animations or scene graphs, OBJ remains a reliable choice for static 3D model archiving due to its universal compatibility.
  STL (STereoLithography): Primarily designed for 3D printing applications, STL stores mesh geometry without color, texture, or material information. While limited in scope, it serves as an effective format for rapid prototyping and physical reproduction of heritage objects.
  X3D (Extensible 3D): An *ISO standard (ISO/IEC 19775)* built on XML, supporting comprehensive 3D scene descriptions including geometry, materials, lighting, animations, and interactivity. X3D's standardized nature and extensive feature set make it suitable for long-term archival and web-based heritage applications. Despite being an active standard, X3D has seen reduced adoption in modern applications, largely superseded by more efficient formats like glTF for web-based heritage visualization.
  glTF (GL Transmission Format): Developed by the *Khronos Group* as an open standard for efficient 3D content transmission and rendering. glTF utilizes a JSON-based scene graph structure that makes it inherently extensible and customizable through a well-defined extension mechanism.

This graph-based architecture, similar to JSON's hierarchical data organization, enables flexible representation of complex 3D scenes while maintaining readability and programmatic accessibility. glTF 2.0 provides comprehensive support for:

- Physically Based Rendering (PBR) materials
- Detailed surface characterization through normal maps (macro-surface properties)
- Microsurface properties via roughness and metallic parameters
- Embedded or external texture resources
- Animation and scene hierarchy
- Custom extensions for domain-specific requirements

The format extensibility allows cultural heritage institutions to develop specialized extensions for **metadata**, provenance information, or conservation-specific data while maintaining compatibility with standard viewers and tools.

GLB (Binary glTF format): Developed by the *Khronos Group* as a binary variant of the glTF (GL Transmission Format).

COLLADA (3D Asset Exchange Schema): Developed by the *Khronos Group*. COLLADA™ defines an XML-based schema to make it easy to transport 3D assets between applications, enabling diverse 3D authoring and content processing tools to be combined into a production pipeline. The intermediate language provides comprehensive encoding of visual scenes including: geometry, shaders and effects, physics, animation, kinematics, and even multiple version representations of the same asset. COLLADA FX enables leading 3D authoring tools to work effectively together to create shader and effects applications and assets to be authored and packaged using OpenGL® Shading Language, Cg, CgFX, and DirectX® FX.

OpenUSD/ USD (Universal Scene Description): Originally developed by *Pixar* and now open-source, USD provides a robust framework for complex 3D scene composition, asset management, and collaborative workflows. USD excels in:

- Hierarchical scene organization
- Non-destructive workflow composition
- Advanced material and shading networks
- Time-varying data and animation
- Large-scale asset management
- Extensible schema for custom data types

For cultural heritage applications, USD's ability to manage complex collections and maintain data integrity across collaborative workflows makes it valuable for institutional archives and comprehensive heritage documentation projects.

OpenUSD/USD is more an open source project than a (open) standard.

- **Proprietary and Specialized Formats**

  PTS: An ASCII format for *unstructured point clouds*, developed by *Leica*. Each line represents a point with XYZ coordinates, intensity, and optionally RGB values. Lacking a header and metadata, it is not formally extensible. Its simple structure makes it suitable for quick exchanges and intermediate conversions, although file sizes tend to be large.

  PTX: An ASCII format for *structured point clouds*, also developed by *Leica*. Each block represents a single scan and is preceded by a header specifying the grid dimensions (in rows and columns), along with the scanner's position and orientation. This is followed by a list of points (XYZ, intensity, RGB), ordered as acquired. The format is not extensible and includes only minimal metadata, but it preserves the original scan structure, making it well-suited for registration and comparison operations between scans.

  XYZ: A simple ASCII format for *unstructured point clouds*, consisting of one point per line, typically with XYZ coordinates and optionally additional attributes such as intensity or RGB. It lacks a header and any standardized metadata, and the field order is not formally defined, requiring prior knowledge or manual interpretation. Due to its minimal structure, it is not extensible and poorly suited for long-term archiving, but it remains widely used for quick inspection, basic interoperability, and intermediate data exchange thanks to its simplicity and human readability.

  FBX (FilmBoX): *Autodesk's* proprietary format supporting comprehensive 3D scene data including animations, cameras, lighting, and physics properties. While feature-rich, its proprietary nature raises concerns for long-term preservation and data accessibility.

  STEP (Standard for the Exchange of Product Data): *An ISO standard (ISO 10303)* for product data exchange in engineering and manufacturing contexts. STEP files are valuable for technical documentation and dimensional analysis of heritage structures and objects.

4.3 Format Selection Criteria and Open Science Principles

The selection of appropriate 3D formats for cultural heritage applications must balance technical requirements with long-term preservation goals and adherence to open science principles. This strategic approach ensures that digitisation investments provide maximum value for research, education, and preservation efforts.

- **Technical and Functional Criteria**
  The choice between formats depends on 3D data type (point cloud or mesh), specific project needs, material representation, animation support, metadata inclusion, etc. Web-based applications and real-time visualization require formats optimized for transmission and rendering efficiency. For example, for comprehensive digital replicas of objects requiring detailed material characterisation, glTF and USD could be suitable choices due to their robust material models, extensibility, and industry support. On one side, glTF particularly excels with native PBR (Physically Based Rendering) support and efficient compression algorithms like Draco, making it optimal for web-based heritage applications and online exhibitions. On the other hand, USD serves as the preferred solution for desktop application interoperability and complex production pipelines.
- **FAIR Principles Implementation**
  The FAIRification of 3D heritage data requires careful format selection to ensure:
  - **Findability**: Open formats with embedded metadata capabilities enable better discovery through search engines and data repositories. For example, formats like glTF and USD support rich metadata integration, facilitating automated indexing and cataloguing.
  - **Accessibility**: Standardized, open formats ensure data remains accessible across different software platforms and over extended time periods, crucial for enabling broad participation in heritage research and education.
  - **Interoperability**: Open standards promote seamless data exchange between research groups, institutions, and software ecosystems. For example, the JSON-based structure of glTF facilitates integration with web technologies and modern data processing pipelines.
  - **Reusability**: Well-documented, open formats with clear licensing frameworks enable researchers to build upon existing work, accelerating scientific progress and maximising digitisation investments.
- **Long-term Preservation and Sustainability**
  Ensuring long-term sustainability of 2D and 3D heritage data is one of the principles of the London Charter[2]. Open, standardized formats provide critical advantages:
  - **Future-proofing**: Well-documented specifications ensure accessibility as technologies evolve.
  - **Cross-platform compatibility**: Standards enable data sharing across software ecosystems and institutional boundaries.
  - **Scientific reproducibility**: Standardized data formats enable independent verification of research results.
  - **Collaborative research**: Format standardization facilitates multi-institutional collaborations and large-scale heritage studies.

It should be noted that this assessment reflects the state-of-the-art as of 2025, and format preferences may evolve as new technologies emerge and existing standards mature. Historical precedents, such as the transition from Collada to current preferred formats, demonstrate the dynamic nature of 3D data standards. This integrated approach to format selection aligns with the 3D-4CH project's objectives of promoting standardised, sustainable practices for cultural heritage digitisation while fostering interoperability across the European Data Space for Cultural Heritage and supporting open science goals.

4.4 MIME Types for 3D Formats

A MIME type (Multipurpose Internet Mail Extensions type) is a standardized way to indicate the nature and format of a file so that computers, web browsers, and applications know how to handle it correctly. Originally developed for email attachments, MIME types are now fundamental in web technologies: they tell browsers

[2] https://londoncharter.org/

and APIs whether a file is an image, video, text document, 3D model, or any other data type. For example, a .jpg image is usually served as image/jpeg, and a .gltf 3D model as model/gltf+json. **Using the correct MIME type ensures that files are delivered, interpreted, and displayed properly by different systems and devices.**
A precise definition of the corresponding MIME types for 3D file formats — including both officially registered standards and widely used de facto practices — is fundamental to guarantee that 3D files are correctly recognised by systems, transmitted without errors, and interpreted reliably by different software solutions and web-based viewers. Given the growing role of web services, APIs, and distributed platforms within the Data Space for Cultural Heritage, clear and standardised MIME type information becomes a technical prerequisite for smooth data exchange and integration.
To address this need, a dedicated table has been developed and is provided in **Annex 3** of this document. **The table offers a detailed overview of the MIME types linked to the most common 3D model and point cloud file formats used in the cultural heritage sector.** By consolidating this information in one place, the appendix serves as a practical reference for institutions, developers, aggregators, and all stakeholders involved in 3D digitisation workflows. This resource will help them make informed choices when preparing, publishing, or reusing 3D datasets, supporting long-term data sustainability, technological compatibility, and the principles of open and reusable digital heritage.

5. XR solutions

In today's digital era, advanced technologies play a crucial role in the preservation, interpretation, and dissemination of cultural heritage. The integration of immersive digital tools allows users to engage with heritage in unprecedented ways, offering not just passive observation but active, engaged participation. Through the virtual reconstruction of traditional crafts, rituals, and historical sites, audiences can explore and experience cultural narratives across time and space, often accessing places or artifacts that are physically remote, fragile, or no longer exist.

One of the most transformative advancements in this field is the rise of *Extended Reality (XR)*, a collective term that includes *Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR)*. These technologies have seen widespread adoption in museums, archaeological sites, and educational institutions, where they are used to create immersive environments that enhance learning, storytelling, and emotional connection with the past.

For instance, VR can transport users into fully reconstructed historical environments, AR overlays digital information onto physical objects in real-time, and MR blends real and virtual elements in interactive spaces. Such tools not only improve accessibility for broader audiences - including people with limited mobility or geographical constraints - but also support cultural sustainability by documenting intangible heritage elements, such as oral histories, craftsmanship, and performative traditions.

VR, AR, and MR are technologies widely used today in the dissemination of cultural heritage, although their maturity and level of adoption varies. Collectively known as XR, these technologies are intended to combine the physical and a digital world, giving users an immersive experience whose primary application is the virtual reconstruction of the past.

There is intense scientific debate about the appropriate way to approach these virtual reconstructions (Ferdani et al., 2020), which has led to the proposal of guidelines and best practices in the field of scientific visualisation of the past, such as the London Charter (London Charter, 2006) and the Principles of Seville (Seville Principles, 2011). In accordance with these principles and documents, it is clear that XR technologies applied to heritage must pursue a scientific purpose, drawing on mature disciplines like archaeology, history or architecture as a foundation and support for their hypotheses.

5.1 VR platforms

Virtual reality provides users with a fully immersive digital environment that simulates a three-dimensional space, often using headsets or goggles. By completely replacing the user's real-world surroundings, VR allows individuals to explore and interact with virtual environments, making it particularly popular in gaming and entertainment.

The VR market has evolved rapidly, offering immersive digital environments used across industries - from gaming and education to healthcare and heritage preservation. VR's roots can be traced back to the 1960s, with early systems like the "Sensorama" and "Sword of Damocles". Over decades, the technology advanced through academic and military applications before entering mainstream use with devices like the Oculus Rift in the 2010s. Today, the market is defined by a diverse range of platforms, categorized into standalone headsets, PC-based systems, and mobile VR solutions.

- Standalone VR headsets are all-in-one devices that require no external hardware. Some examples are the Meta Quest 3 provided by Meta and Pico 4 from Bytedance.
- PC-based VR systems offer high-fidelity experiences, essential for detailed visualization. Examples include Valve Index and HTC Vive Pro, and Varjo XR Series.
- Mobile VR solutions offered entry-level VR and were used in early educational and museum projects. However, they are now discontinued in favor of standalone VR headsets.

XR today is being shaped by several hardware and software trends, which will continue to impact it over the next few years:

- **Hardware Evolution:** The move to wireless high-fidelity VR has been a game-changer. Standalone headsets, such as the **Meta Quest 3** mentioned above, offer an appealing combination of freedom of movement and graphical capability, making VR more accessible than ever. At the high end, professional-grade headsets like those from **Varjo** are pushing the boundaries of visual fidelity with "retina-resolution" displays that are crucial for detailed professional applications. A key breakthrough is the implementation of **foveated rendering** (Krajancich et al., 2023), a technique that uses eye-tracking to render the user's gaze area in high resolution while reducing detail in the peripheral vision. This significantly reduces the computational load, allowing for more complex and realistic simulations on less powerful hardware.
- **Haptic Feedback:** The sense of touch is a critical component of immersion. Beyond the simple vibrations of handheld controllers, the development of advanced haptic feedback suits and gloves is creating more profound tactile experiences. There are vests with multiple feedback points (bHaptics.com) that can simulate a range of sensations, from the impact of a raindrop to the force of a virtual punch. These peripherals are transitioning from niche gaming accessories to valuable tools for training and simulation, enabling more realistic and impactful virtual interactions.
- **Multisensory Immersion:** As VR continues to evolve, new methods to engage more of the user's senses are being explored to create richer, more immersive experiences. Beyond visual, tactile, and auditory inputs, sensations such as smell are now being introduced into virtual environments. Devices like Escents from Scentient aim to enhance realism and deepen user engagement by releasing scents that complement the virtual experience.
- **Software and Development:** The integration of **Artificial Intelligence (AI)** is revolutionising content creation for VR. AI-powered tools are now capable of generating 3D assets from text prompts or 2D images, drastically reducing the time and cost of developing virtual environments. This democratisation of content creation is empowering smaller developers and institutions to build bespoke VR experiences. Furthermore, the rise of **WebXR**, an API that allows VR experiences to be delivered directly through a web browser, is removing the barrier of app downloads and installations, making VR content as accessible as a webpage.

5.1.1. Virtual Reality in Cultural Heritage

VR offers the unparalleled ability to transport users to places and times that would otherwise be inaccessible:

- **Virtual Reconstruction and Preservation:** VR allows for the digital reconstruction of historical sites that have been lost to time or are too fragile for public access. For example, a user could virtually walk the streets of Roman London or explore the interior of a Neolithic tomb as it would have appeared thousands of years ago. This serves not only as a powerful educational tool but also as a form of digital preservation, safeguarding our heritage for future generations. The concept of **"digital twins"** in cultural heritage allows for the creation of highly detailed, data-rich virtual replicas of heritage sites for monitoring, research, and conservation planning.
- **Enhanced Accessibility:** VR can break down the physical barriers that prevent many people from experiencing cultural heritage. People with mobility issues, those in remote locations, or those who cannot afford to travel can experience world-class museums and historical sites from the comfort of their own homes. There are numerous examples of VR being used to access remote sites.
- **Immersive Storytelling:** VR provides a powerful medium for storytelling, allowing curators to create narrative experiences that bring history to life. Instead of passively reading about a historical event, a user can witness it unfold around them, fostering a deeper, more emotional connection to the past.

However, like any technology, its application also comes with a series of risks.

- **The "Wow" Factor vs. Substantial Engagement:** There is a risk that VR experiences in cultural heritage will focus too heavily on spectacle at the expense of genuine historical research. As noted earlier, the image must serve research, simulation, and interpretation. A visually stunning virtual environment does not automatically equate to a meaningful learning experience.
- **Cost and Technical Expertise:** Creating high-quality, historically accurate VR experiences can be expensive and time-consuming, requiring a specialised skill set that many cultural institutions may not possess in-house. In some cases, the necessary digitisation processes are very costly. This can lead to a digital divide, where only the largest and best-funded institutions can afford to develop meaningful VR content.
- **Authenticity and Interpretation:** The digital reconstruction of historical sites inevitably involves a degree of interpretation. Decisions about materials and the placement of objects can influence a user's understanding of the past. There is an ethical responsibility to be transparent about these interpretations and to avoid presenting a sanitised or misleading version of history.

5.2 AR platforms

Augmented Reality superimposes computer-generated information onto a user's view of the real world. This is most commonly experienced through smartphones and tablets, but the long promise of the arrival of lightweight and affordable devices that we can wear unobtrusively would introduce this technology in our daily lives.

Three elements define AR (Azuma, 1997): a combination of real and virtual, real-time interaction, and 3D registration. These are the basic problems to be solved in this technology, and significant progress has been made on them in recent years.

The evolution of AR is focused on seamless integration with the user's environment and the delivery of contextually relevant information:

- **Miniaturisation and Hardware Style:** A significant barrier to the widespread adoption of AR glasses has been their bulky and conspicuous design. However, recent advances have led to the development of more discreet and aesthetically pleasing models. While a true all-day consumer AR device is still on the horizon, the progress in miniaturising projectors, sensors, and batteries is undeniable.
- **Simultaneous Localisation and Mapping (SLAM):** The ability of an AR device to understand and map its physical environment in real time is fundamental to creating convincing AR experiences. Modern SLAM algorithms, which can take advantage of low-cost **LiDAR sensors** integrated into recent mobile devices, are more robust and accurate than ever, allowing for persistent AR content that can be anchored to specific real-world places and objects. This is crucial for cultural heritage applications, where digital reconstructions must be accurately superimposed on historical ruins. These techniques eliminate the need for traditional targets for registration, which always alter the real environment—a particularly sensitive aspect in heritage contexts. However, the changing environmental conditions that can occur in outdoor settings still pose a challenge for these algorithms, so it is still necessary to improve the adaptive capabilities of SLAM.
- **Visual Positioning Systems (VPS):** Advancements in computer vision algorithms to use and analyze 3d maps from images of the user's surroundings have significantly improved the accuracy of AR experiences. VPS is a vision-based localization system that can determine the user's position with high precision using pre-mapped visual data, enabling the AR content to be spatially anchored with greater reliability, enhancing digitally enabled navigation and site-specific storytelling.
- **AI and Contextual Awareness:** As with VR, AI algorithms are used to recognise objects and scenes in the real world, triggering the display of relevant digital information. This "contextual awareness" is the key to AR's potential as a personal assistant, providing everything from real-time language translation of street signs to interactive assembly instructions superimposed on a piece of machinery. AI can also play a role in improving the adaptive capabilities of the SLAM algorithms we mentioned earlier.
- **WebAR (web-based augmented reality)** enables AR experiences to run directly within web browsers, eliminating the need for specialized apps. While still limited regarding performance and advanced functionalities compared to native AR apps, WebAR is evolving rapidly, promising to make AR more accessible, scalable, and easier to deploy across a variety of platforms.

5.2.1. Augmented Reality in Cultural Heritage

AR enhances our experience of physical heritage sites by superimposing digital information onto our view of the real world, highlighting the following possibilities.

- **In-Situ Visualisation:** AR can be used to overlay historical reconstructions onto existing ruins, allowing visitors to see what a site would have looked like in its heyday. The **"London AR Trail for Heritage Quarter"** is an excellent example of this, using AR to bring famous sites to life for visitors (LondonAR, 2025). This provides a powerful sense of context and helps visitors better understand the scale and significance of what they are seeing.
- **Interactive and Gamified Experiences:** AR can be used to create interactive trails and games that encourage visitors to explore a heritage site in a more engaging way. This can be particularly effective for younger audiences, transforming a museum visit into a scavenger hunt or a mystery to solve.
- **Access to Hidden Information:** AR can provide access to a wealth of information that cannot be displayed on traditional signage. By pointing their smartphone at an artefact, a visitor could access 3D models, videos of the object in use, or detailed information about its history and provenance.

Among the risks and negative aspects of using this technology, we can highlight the following:

- **Device Dependence:** At present, most AR experiences in cultural heritage are delivered via smartphones or tablets. This means visitors must have a compatible device with a charged battery and may be more focused on their screen than on the physical environment around them.
- **Environmental Limitations:** As mentioned earlier, AR applications can be affected by environmental factors such as poor lighting, inclement weather, or unrecorded changes in the environment's configuration. The accuracy of the AR overlay can also be compromised if the user moves too quickly or if the device's camera is obstructed.
- **Intrusion and Distraction:** A poorly designed AR experience can be intrusive and distracting, drawing the visitor's attention away from the authentic heritage site. The digital overlay must enhance, not overshadow, the real-world experience.

5.3 MX techniques

Mixed Reality, as its name suggests, is a hybrid of VR and AR. It allows users to interact with virtual objects that are aware of and can interact with the real world in real time. This is often achieved through headsets equipped with high-resolution cameras that transmit a view of the real world to the user, over which digital elements are rendered.

MR is arguably the most ambitious of the three technologies, and its recent advancements are bringing its futuristic promise closer to reality:

- **High-Fidelity Passthrough:** The quality of the "passthrough" video feed is critical for a convincing MR experience. Modern MR headsets, such as the **Apple Vision Pro** and the **Varjo XR-4**, feature high-resolution, low-latency colour passthrough that makes the fusion of real and virtual elements feel almost seamless. This is a significant leap from the grainy, black-and-white passthrough of earlier devices.
- **Hand and Eye Tracking:** The ability to interact with virtual objects using bare hands is a cornerstone of the MR vision. Advanced hand-tracking technology, often combined with eye-tracking for gaze-based interactions, is making this a reality. This intuitive form of interaction is far more natural than using handheld controllers and is essential for tasks that require fine motor skills. In this regard, tools like **Google Mediapipe** represent a significant advance in capturing users' gestures, actions, and movements within work environments.
- **The Role of 5G:** The rollout of **5G networks** is a crucial enabler for the future of all XR technologies, but particularly for MR. 5G's high bandwidth and ultra-low latency will allow much of the computational load to be offloaded from the headset to the cloud. This will make it possible to create lighter, more powerful headsets and to stream highly complex, photorealistic virtual objects into the user's environment without delay. In any case, there is also the possibility of hybrid rendering distributed between the glasses and a server, with balancing capabilities depending on the power of the glasses used.

5.3.1. Mixed Reality in Cultural Heritage

MR offers the most integrated and interactive approach to blending the digital and physical, with significant potential for cultural heritage. Its possibilities, which are still largely unexplored, are as follows:

- **Tangible Interaction with Virtual Artefacts:** MR allows users to interact with virtual objects as if they were real. A visitor could "pick up" a virtual Roman vase, turn it over in their hands, and even feel its texture through haptic feedback. This provides a level of engagement and understanding that is impossible with traditional museum exhibits.
- **Collaborative Experiences:** MR is an inherently multi-user technology, allowing groups of visitors to share the same mixed-reality experience. A tour guide could lead a group through a virtual reconstruction of a historical building, with all participants able to see and interact with the same digital elements. Unfortunately, most collaborative solutions use a gamified graphical environment (such as Meta's Horizon), which in most cases prevents the incorporation of realistic elements into the environment.
- **Dynamic and Responsive Exhibitions:** MR exhibitions can be dynamic and respond to user actions. For example, a virtual character could appear and tell the story of a particular artefact when a visitor approaches it, creating a more personalised and engaging experience. In this and other fields, advances in AI and intelligent virtual assistants and their integration with VR, AR, and MR tools are essential for achieving interactive agents.

Unfortunately, there are a number of important challenges that hinder the mass adoption of MR in heritage dissemination.

- **High Hardware Cost:** MR headsets are currently very expensive, making them inaccessible to most people and a significant investment for cultural institutions.
- **Ethical Considerations of Representation:** As with VR, creating MR experiences in cultural heritage raises ethical questions about representation and authenticity. A recent paper (DeHass et al., 2025) highlights the importance of collaboration and respecting the cultural property rights of indigenous communities when creating digital replicas of their heritage.
- **Technical Complexity:** Developing and implementing MR experiences are even more complex than VR or AR, requiring a high level of technical expertise and a robust infrastructure.

6. Automated translation of metadata

The project consortium has **reassessed the plan to analyze metadata translation tools** in light of Europeana's existing multilingual infrastructure. The Europeana platform already provides a mature, automated metadata translation service, so developing a separate solution would be redundant. Notably, Europeana's translation tool builds on the comprehensive pipeline developed during the *Europeana Translate* project, which successfully translated **over 25 million** metadata records into English. These English metadata enrichments are now available on Europeana. In addition, Europeana offers an **on-demand translation** feature that lets users translate item metadata or search results on the fly. This service currently leverages the **Google Translate** system for real-time machine translation, supporting all 24 official EU languages. Together, these capabilities represent a ready-made solution that **eliminates the need for the project to identify or evaluate additional translation tools** as originally proposed.

From an operational perspective, **metadata translation is handled in the post-aggregation phase** – that is, after metadata records have been ingested into the Europeana platform. This places the task firmly under the purview of the Europeana Foundation (a key project partner). Relying on Europeana's established service avoids duplicating expertise and infrastructure within the project. Conducting a parallel analysis of other translation solutions would add little value, since the consortium can utilize Europeana's proven pipeline instead. This approach also ensures seamless integration with the broader *Common European Data Space for Cultural Heritage* initiative, by leveraging a centralized service for multilingual metadata across the platform. In essence, delegating metadata translation to **Europeana's existing infrastructure** enables consistent, high-quality translations while aligning with Europeana's role as the data space operator.

By deferring to Europeana's translation services, the **3D Competence Centre** can concentrate on its core mandate: advancing 3D digitisation technologies and methodologies. Instead of expending resources on redundant translation tool development, the Competence Center will focus on **domain-specific**

contributions. This includes monitoring and validating any new **3D-specific metadata fields** introduced for cultural heritage objects, and ensuring these new fields integrate properly into Europeana's translation workflow. In practice, the 3D Competence Center will coordinate with Europeana to make sure that any metadata fields unique to 3D content are recognized and automatically translated by the platform's services. This guarantees that multilingual access is preserved even as the project innovates in 3D metadata standards, and it takes advantage of Europeana's existing multilingual **infrastructure for continuity and efficiency**.

It should be noted that Europeana's multilingual strategy is **evolving**. For example, Europeana staff currently translate some editorial content (such as blog posts or exhibitions) with the help of *eTranslation* (https://commission.europa.eu/resources/etranslation_en) – the European Commission's neural machine translation service – combined with manual curation. Plans are underway to integrate such translation services more tightly into Europeana's Content Management System. Moreover, Europeana has recently developed a dedicated **Translation API** to act as a broker for various machine translation engines. This API will allow Europeana to **switch between translation providers** (e.g. Google Translate, eTranslation, or others) as needed, improving resilience and optimizing quality and cost. All automatic translations would be mediated through this single API, which also introduces caching to avoid redundant translations. These enhancements, however, **do not change the project's decision** to rely on Europeana's translation infrastructure. On the contrary, they reinforce that Europeana's platform is well-equipped to handle multilingual metadata translation. By using Europeana's proven solution, the project ensures that its metadata will be accessible in English (and other languages on demand) without reinventing the wheel, allowing the consortium to focus on innovation in 3D content and leave translation tasks to the established, state-of-the-art services provided by Europeana.

7. Tools and frameworks

7.1 Image-based

The following section provides an overview of the main image-based software used for 3D digital reconstruction of cultural heritage assets. The selection includes both proprietary and open-source solutions, reflecting the diversity of tools available for academic research and professional practice.
Each software description covers key aspects such as target users, application domains, and both core and supplementary functionalities.

Proprietary solutions

Metashape by Agisoft **Target users:** Academic and professional. **Adoption level:** High.	**Application domains:** Environment, agriculture, industrial, AEC, cultural heritage, product design. **Core functionalities:** SfM-based photogrammetric processing, including camera alignment, dense point cloud generation, mesh reconstruction, texturing, and orthoimage production. Generates detailed reports for camera orientation and georeferencing. **Supplementary functionalities:** Includes point cloud classification tools; supports the co-registration of laser-derived point clouds and photogrammetric data.
3DF Zephyr by 3Dflow **Target users:** Academic and professional. **Adoption level:** High.	**Application domains:** Environment, agriculture, industrial, AEC, cultural heritage, product design. **Core functionalities:** SfM-based photogrammetric processing, including camera alignment, dense point cloud generation, mesh reconstruction, texturing, and orthoimage production. Generates detailed reports for camera orientation and georeferencing. **Supplementary functionalities**: Includes point cloud classification tools; supports the co-registration of laser-derived point clouds and photogrammetric data.

RealityScan (ex RealityCapture) by Epic Games **Target users:** Academic and professional. **Adoption level:** High.	**Application domains:** Environment, agriculture, industrial, AEC, cultural heritage, product design. **Core functionalities:** SfM-based photogrammetric processing, including camera alignment, dense point cloud generation, mesh reconstruction, texturing, and orthoimage production. Generates detailed reports for camera orientation and georeferencing. **Supplementary functionalities**: Includes classification tools; supports the co-registration of laser-derived point clouds; provides seamless integration with Unreal Engine for real-time visualization and rendering workflows.
Pix4Dmatic by Pix4D **Target users:** Academic and professional. **Adoption level:** Medium.	**Application domains:** Environment, agriculture, industrial, AEC, cultural heritage. **Core functionalities:** SfM-based photogrammetric processing, including camera orientation, dense point cloud generation, mesh reconstruction, texturing, and orthoimage production. Generates detailed reports for camera orientation and georeferencing. **Supplementary functionalities:** Includes point cloud classification tools; supports the co-registration of laser-derived point clouds and photogrammetric data.
DJI Terra PRO by DJI **Target users:** Academic and professional. **Adoption level:** Primarily tied to DJI hardware.	**Application domains:** Environment, agriculture, industrial, AEC, cultural heritage. **Core functionalities:** SfM-based photogrammetric processing, including camera orientation, dense point cloud generation, mesh reconstruction, texturing, and orthoimage production. Supports the import and processing of aerial LiDAR data with kinematic trajectory. Enables combined photogrammetric and LiDAR-based workflows. Generates detailed reports for orientation and georeferencing. **Supplementary functionalities:** Includes point cloud classification tools; natively integrates with DJI drone platforms and flight planning software.

Open source solutions

COLMAP **Target users:** Academic. **Adoption level:** Specialized.	**Application domains:** Cultural heritage, product design. **Core functionalities:** SfM-based photogrammetric processing, including camera orientation and dense point cloud generation. Generates detailed reports for camera orientation.
MicMac by IGN **Target users:** Academic and professional (geospatial only). **Adoption level:** Specialized.	**Application domains:** Environment, agriculture, cultural heritage. **Core functionalities:** SfM-based photogrammetric processing, including camera orientation, dense point cloud generation, mesh reconstruction, texturing, and orthoimage production. Generates detailed reports for camera orientation and georeferencing. **Supplementary functionalities:** Includes a modular and highly configurable processing pipeline.
Meshroom by AliceVision **Target users:** Academic. **Adoption level:** Specialized.	**Application domains:** Cultural heritage, product design. **Core functionalities:** SfM-based photogrammetric processing, including camera orientation, dense point cloud generation, mesh reconstruction, and texturing. Generates detailed reports for camera orientation. **Supplementary functionalities:** Includes a modular and highly configurable processing pipeline.
Regard3D **User target:** Academic. **Adoption level:** Low.	**Application domains:** Cultural heritage, product design. **Core functionalities**: SfM-based photogrammetric processing, including camera orientation, dense point cloud generation, mesh reconstruction, and texturing.

7.2 Range-based

The following section offers an overview of the main range-based software solutions used for 3D digital reconstruction in cultural heritage, with a focus on systems based on Time-of-Flight (ToF) technologies. Only proprietary solutions are included, due to the limited availability of effective open-source alternatives. This is primarily because raw data processing typically requires manufacturer-specific proprietary software tools, which also tend to provide all the reconstruction functionalities commonly required.
Each software description covers key aspects such as the developer, license model, target users, application domains, operational scale, and both core and supplementary functionalities.
Software related to triangulation-based systems is excluded from the list, given the wide variety of available hardware and the tight integration with proprietary software environments, where data processing and alignment usually occur in real time. This specific configuration reduces the role of software as an autonomous component. For this reason, it would be more appropriate to evaluate the technical and performance characteristics of the acquisition device rather than those of the reconstruction software itself. However, this falls outside the scope of the present analysis.

REGISTER 360 PLUS by Leica Geosystems **Target users:** Academic and professional. **Adoption level:** Primarily tied to Leica hardware.	**Application domains:** Industrial, AEC, cultural heritage. **Core functionalities:** Processes TLS and SLAM scans from Leica; supports the import of third-party scans and point clouds in open formats. Performs automatic or semi-automatic registration using cloud-to-cloud, target-based, or visual alignment methods. Includes advanced cleaning tools and generates detailed registration reports. **Supplementary functionalities:** Supports advanced point cloud classification tools. Integrates with Leica CYCLONE 3DR for mesh generation, advanced editing, and texturing.
FARO Scene by FARO Technologies **Target users:** Academic and professional. **Adoption level:** Primarily tied to FARO hardware.	**Application domains:** Industrial, AEC, cultural heritage. **Core functionalities:** Processes FARO TLS scans and, optionally, FARO SLAM scans (after FARO Connect pre-processing); supports the import of third-party scans and point clouds in open formats. Performs automatic or semi-automatic registration using cloud-to-cloud, target-based, or visual alignment methods. Includes advanced cleaning tools and generates detailed registration reports. **Supplementary functionalities:** Includes basic tools for mesh generation and texturing.
RealWorks by Trimble **Target users:** Academic and professional. **Adoption level:** Primarily tied to Trimble hardware.	**Application domains:** Industrial, AEC, cultural heritage. **Core functionalities:** Processes TRIMBLE TLS scans; supports the import of third-party scans and point clouds in open formats. Performs automatic or semi-automatic registration using cloud-to-cloud, target-based, or visual alignment methods. Includes advanced cleaning tools and generates detailed registration reports. **Supplementary functionalities:** Includes tools for mesh generation and texturing. Supports point cloud classification tools.
Reconstructor by Gexcel **Target users:** Academic and professional. **Adoption level:** Primarily tied to Gexcel hardware.	**Application domains:** Industrial, AEC, cultural heritage. **Core functionalities:** Processes Gexcel SLAM data (after HERON pre-processing); imports laser scans and point clouds in open formats. Performs automatic or semi-automatic registration using cloud-to-cloud, target-based, or visual alignment methods. Includes advanced cleaning tools and generates detailed registration reports. **Supplementary functionalities:** Includes tools for mesh generation and texturing (via the Color add-on integrated with 3DF Zephyr). Supports manual point cloud classification tools.
RiSCAN PRO by RIEGL **Target users:** Academic and professional. **Adoption level:** Primarily tied to RIEGL hardware.	**Application domains:** Industrial, AEC, cultural heritage. **Core functionalities:** Processes RIEGL TLS scans; supports the import of third-party scans and point clouds in open formats. Performs automatic or semi-automatic registration using cloud-to-cloud, target-based, or visual alignment methods. Includes advanced cleaning tools and generates detailed registration reports. **Supplementary functionalities:** Includes basic tools for mesh

	generation and texturing. Supports point cloud classification with limited capabilities.
RiPROCESS by RIEGL **Target users:** Academic and professional. **Adoption level:** Primarily tied to RIEGL hardware.	**Application domains:** Environment, agriculture, industrial, AEC, cultural heritage. **Core functionalities:** Processes RIEGL kinematic scans; performs automatic or semi-automatic registration using cloud-to-cloud and target-based methods. Includes advanced cleaning tools and generates detailed registration reports. **Supplementary functionalities:** Supports point cloud classification tools.
ReCap PRO by Autodesk **Target users:** Academic and professional. **Adoption level:** High.	**Application domains:** Environment, industrial, AEC, cultural heritage. **Core functionalities:** Imports laser scans and point clouds in open formats. Performs automatic or semi-automatic registration using cloud-to-cloud and target-based methods. Includes manual cleaning tools and generates detailed registration reports. Integrates well into SCAN2CAD and SCAN2BIM workflows within the Autodesk ecosystem. **Supplementary functionalities:** Includes basic tools for mesh generation and texturing. Supports point cloud classification with limited capabilities.
PointCab Origins 3D + Registration by PointCab GmbH **Target users:** Professional. **Adoption level:** Medium.	**Application domains:** Environment, industrial, AEC, cultural heritage. **Core functionalities:** Supports the import and registration of point clouds from various sources (open formats) using cloud-to-cloud, target-based, or visual alignment methods. Includes basic cleaning tools and generates detailed registration reports. Well suited for SCAN2CAD and SCAN2BIM workflows. **Supplementary functionalities:** Includes basic tools for mesh generation.

7.3 3D editing

The following section provides an overview of the main 3D modeling, processing, and optimisation software solutions employed in the refinement and enhancement of digital models within the context of cultural heritage. The selection includes both proprietary and open-source tools widely adopted across academic and professional domains. These software solutions support a broad range of tasks, including mesh editing, UV mapping, texture generation, and advanced material definition.
Each description includes key information regarding the developer, license model, target users, application domains, operational scale, and both core and supplementary functionalities.

Cyclone 3DR by Leica Geosystems **Target users:** Academic and professional. **Adoption level:** Primarily tied to Leica hardware.	**Application domains:** Industrial, AEC, cultural heritage. **Core functionalities:** Offers tools for post-processing of point clouds pre-registered in Leica REGISTER 360. Provides mesh generation from point clouds and mesh editing (cleaning, decimation, remeshing). Supports texture projection from single images or textured models. **Supplementary functionalities:** Includes advanced analysis and classification tools for point clouds.
Geomagic Wrap / Design X by 3D Systems **Target users:** Academic and professional. **Adoption level:** High.	**Application domains:** Cultural heritage, product design. **Core functionalities:** Provides mesh generation from point clouds and mesh editing (cleaning, decimation, remeshing). Supports texture projection from images or textured models and UV mapping. **Supplementary functionalities:** Design X adds NURBS surface fitting and CAD reconstruction. Includes advanced mesh analysis tools.
ZBrush by Maxon	**Application domains:** Cultural heritage, product design, CGI & digital media. **Core functionalities:** Provides 3D modeling, advanced mesh editing

Target users: Professional. **Adoption level:** High.	(cleaning, decimation, remeshing, sculpting) and advanced retopology tools. Supports basic UV mapping and texture editing.
Maya by Autodesk **Target users:** Professional. **Adoption level:** High.	**Application domains:** Cultural heritage, product design, CGI & digital media. **Core functionalities:** Provides tools for 3D modeling, advanced mesh editing (cleaning, decimation, remeshing, sculpting) and retopology tools. Supports texture projection from images or textured models, advanced UV mapping, texture editing and PBR material generation. Includes tools for animation and rigging. **Supplementary functionalities:** Offers basic sculpting tools.
Houdini FX / Indie / Apprentice by SideFX **Target users:** Academic and professional. **Adoption level:** High.	**Application domains:** Cultural heritage, product design, CGI & digital media. **Core functionalities:** Provides procedural modeling, mesh generation from point clouds, advanced mesh editing (cleaning, decimation, remeshing, sculpting) and advanced retopology tools. Supports texture projection from images or textured models, advanced UV mapping, texture editing and PBR material generation. Includes tools for animation and rigging.
Rhinoceros 3D by McNeel & Associates **Target users:** Academic and professional. **Adoption level:** High.	**Application domains:** Environment, industrial, AEC, cultural heritage, product design. **Core functionalities:** Primarily focused on NURBS modeling, also provides mesh modeling tools for reverse engineering workflows, including mesh generation from point clouds and mesh editing (cleaning, decimation, remeshing). Supports advanced UV mapping and material assignment. **Supplementary functionalities:** Provides CAD-style vector operations and parametric plugin support.
CloudCompare by CloudCompare Project **Target users:** Academic and professional. **Adoption level:** High.	Application domains: Environment, agriculture, industrial, AEC, cultural heritage, product design. Core functionalities: Offers tools for point cloud and mesh registration (cloud-to-cloud or target-based). Supports mesh generation from point clouds and mesh editing (cleaning, decimation). Provides advanced tools for point cloud and mesh analysis, segmentation, and scalar field computation. Supplementary functionalities: Offers a wide ecosystem of community plugins for extended functionality.
MeshLab by ISTI (CNR) **Target users:** Academic and professional. **Adoption level:** High.	**Application domains:** Cultural heritage, product design. **Core functionalities:** Provides mesh generation from point clouds and mesh editing (cleaning, decimation, remeshing). Supports texture projection from images, oriented cameras from Structure-from-Motion software or textured models, and advanced UV unwrapping. **Supplementary functionalities:** Offers mesh analysis tools and export of normal and curvature maps.
Blender by Blender Foundation **Target users:** Academic and professional. **Adoption level:** High.	**Application domains:** Environment, cultural heritage, product design, CGI & digital media. **Core functionalities:** Provides tools for 3D modeling, advanced mesh editing (cleaning, decimation, remeshing, sculpting) and advanced retopology tools. Supports texture projection from images or textured models, advanced UV mapping, texture editing, and PBR material generation. Includes tools for animation and rigging. **Supplementary functionalities:** Offers a wide ecosystem of community plugins for extended functionality.
ShapeLab by LeoPoly **Target users:** Academic and professional. **Adoption level:** Low.	**Application domains:** Environment, cultural heritage, product design, CGI & digital media. **Core functionalities:** Provides full-fledged tools for 3D sculpting in VR mode, advanced mesh editing (cleaning, decimation, remeshing, sculpting) and advanced retopology tools. Supports texture export from polypaint. Includes tools for animation and rigging.

	Supplementary functionalities: Offers a wide selection of formats for export as FBX.
InstaMAT by ABSTRACT **Target users:** Academic and professional. **Adoption level:** Low.	**Application domains:** Environment, cultural heritage, product design, CGI & digital media. **Core functionalities:** Provides a full suite of tools for 3D texturing and parametric modeling, model optimization. Advanced mesh editing (decimation, remeshing), advanced texture baking, node-based PBR workflow, export to any imaginable format including 32-bit EXR. **Supplementary functionalities:** Offers a wide selection of other Abstract products such as InstaLOD.
RizomUV by Rizom-Lab **Target users:** Academic and professional. **Adoption level:** High.	**Application domains:** Environment, cultural heritage, product design, CGI & digital media. **Core functionalities:** Provides a full suite of best-in-class automatic, semi-automatic, and manual tools for UV layout, texel density, and island orientation. Supports several UV mapping channels and UDIMs. **Supplementary functionalities:** Offers a wide selection of plugins and bridges for all major industry-standard software. Includes 'Rizom UV Real Spaces' for 1:1 layout in real-world scale for digital fabrication.
GigaMesh by Hubert Mara **Target users:** Academic and professional. **Adoption level:** Low.	**Application domains:** Environment, cultural heritage, archaeology. **Core functionalities:** Suitable for archaeological artifacts unwrapping. Offers cone, sphere, or cylinder-based 3D model rollout unwrapping, decimation, cross-section generation, distance measuring, rendering, distance visualization, volume calculation.

8. Summary and conclusions

The digital transformation of CH has significantly advanced, driven by innovations established in 3D digitisation, immersive technologies and inevitably AI-assisted tools. The intersection of sciences and humanities led to an evolving ecosystem of technologies focused on 3D documentation, analysis and dissemination. This deliverable outlined the core and emerging technologies while highlighting the integration challenges in the 3D digitisation pipeline. Particular emphasis was given on the interoperability, visual fidelity and usability across various reality-based modelling.

Currently, two principal digitisation approaches dominate the domain. Range-based techniques (*active*) such as the well-established structured light and terrestrial laser scanning are hardware-dependent, vendor-locked and costly, but they deliver dense, metrically reliable data sets making them valuable for detailed documentation. On the other hand, image-based approaches (*passive*), especially SfM and MVS when coupled with high-resolution aerial and terrestrial imagery, offer more accessible and scalable pipelines. However, despite advances in photogrammetric calibration processes and robust feature detection, these still remain sensitive to textureless surfaces and lighting conditions. A panacea, while still not available, is being approximated by the scientific community through sophisticated workflow co-registration and data fusion. Meanwhile, traditional techniques such as the Iterative Closest Point still remain essential, while recent advancements such as deep-learning based registration and SLAM indicate paths with improved autonomy and robustness. Multimodal integration of popular methods such as TLS and SfM continue to present technical challenges but they point towards that critical area where high fidelity data and scenarios relevant to long-term monitoring of CH co-exist.

Beside more conventional methods, over the past years, the research domain has witnessed a paradigm shift with the experimentation and slow adoption of AI and learning-based approaches. Monocular Depth Estimation, Neural Radiance Fields and Visual Geometry Grounded Transformers demonstrate the ability to approximate 3D reconstruction of real world scenes based on limited or even single-view inputs. In some cases they deliver where traditional triangulation fails due to reflective, transparent or featureless surfaces. The black-box nature of these approaches require extensive computational power/resources and lack explainability. The latter is a characteristic that opposes the requirements and prerequisites of archaeology or cultural heritage. These are auditability and scientific process transparency.

Furthermore, the 3D digitisation pipeline includes tasks such as data post-processing, usage-based optimisation and quality assurance. These are integral components of the pipeline and may address mesh cleaning, decimation, remeshing as well as texture optimisation. All are considered crucial not only for enhancing the 3D digital assets in terms of realism but also to ensure their compatibility and utilisation in the target platforms for which they were originally captured for. Careful consideration is required whether the produced 3D digital assets are intended for long-term archiving and monitoring or downstream use in Web visualisation and VR/AR deployment.

Building up the foundations of accurate 3D digitisation, immersive technologies (VR,AR,MR) emerged as impressive technologies for the dissemination and the experiential engagement with content derived from the CH domain. Such modalities enable interactive storytelling, spatial exploration and contextual understanding. Hence, they contribute towards bridging the gap between scientific material, scholarly documentation and the general public. A successful 3D digitisation will provide the metrically accurate and visually detailed assets which through immersive deployment will acquire narrative depth, emotional resonance as well as information temporal layering. Semantic structuring is also utilised by these technologies to deliver content according to user profiles. The continuous advancements in real-time rendering, procedural and AI-based scene composition introduce responsive virtual experiences across a wide range of platforms ranging from head mount displays up to Web-based XR. These advancements, alongside evolving technologies such as 3D printing significantly lowered the technical barriers for CH institutions, stakeholders, museums, etc. to repurpose their digitised assets.

In this regard, attention must also be paid to metadata enrichment, semantic labeling (shared ontologies based on CIDOC CRM) and cross-platform interoperability using standards (e.g. Dublin Core, Europeana EDM) and efficient file formats. It is a fact that sustainability and reusability of 3D digital assets do not solely depend on their geometric fidelity or immersive potential, but also on metadata frameworks and infrastructure ecosystems in which they live. Digital assets are involved in CH tasks and practices, such as managing, discovering, contextualisation and sharing, though rich, structured and interoperable metadata and paradata. These need to address multiple types of information ranging from technical such as acquisition methods and device settings, to semantic such object identity and context, to provenance and more. Spatially allocated user annotations also fulfill aspects of the above need.

The recent shift towards cloud-native architectures, containerised microservices, and modular digital twin environments underpins a more sustainable and scalable approach to 3D CH data management. Such infrastructures are envisioned to support long-term archiving and almost real-time data delivery if aimed to be addressed by interactive and immersive applications. To support both cases, hybrid systems have emerged in which high-resolution assets are maintained in secure digital repositories, while lightweight derivatives are dynamically created or pre-baked to supply WebXR applications.

Viewed as a whole, the developments on 3D digitisation, AI-driven reconstruction, immersive technologies as well as metadata infrastructure point towards a pivotal transformation of the way our CH thesaurus is documented, preserved, monitored and disseminated. As with any transformation, in order to be embraced by the scientific community and professionals of the CH domain it must reflect scientific rigor, sustainability and accessibility. Interdisciplinary collaboration is essential to move forward and address challenges ranging from technical to methodological and semantic. FAIR data usage principles and open standards are key pointers to that direction. Similarly, end-users' needs should be met and aligned with the various digitisation pipelines.

It is this holistic approach that has to be followed to transform digital CH from a fragmented and vast technological domain into a coherent ecosystem which depicts coherence and impact. To this end, 3D-4CH attempts to respond to this need by establishing best practices and tools that empower the above.

9. References

- Azuma, R. A., 1997. A survey of Augmented Reality. *Presence: Teleoperators and Virtual Environments, 6*, pp. 355-385.
- Adamopoulos, E. and Rinaudo, F., 2021. Close-range sensing and data fusion for built heritage inspection and monitoring—a review. *Remote Sensing, 13*(19), p.3936.
- Adamopoulos, E. and Rinaudo, F., 2019. 3D interpretation and fusion of multidisciplinary data for heritage science: A review. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 42*, pp.17-24.
- Ahn, S.J. and Schultes, M., 1997. A new circular coded target for the automation of photogrammetric 3D-surface measurements. *Optical 3-D measurement techniques IV*, pp.225-234..
- Aiger, D., Mitra, N.J. and Cohen-Or, D., 2008. 4-points congruent sets for robust pairwise surface registration. In *ACM SIGGRAPH 2008 papers* (pp. 1-10).
- Apollonio, F.I., Ballabeni, A., Gaiani, M. and Remondino, F., 2014. Evaluation of feature-based methods for automated network orientation. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 40*, pp.47-54.
- Bala, M., Cui, Y., Ding, Y., Ge, Y., Hao, Z., Hasselgren, J., Huffman, J., Jin, J., Lewis, J.P., Li, Z. and Lin, C.H., 2024. Edify 3d: Scalable high-quality 3d asset generation. *arXiv preprint arXiv:2411.07135*.
- Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P. and Hedman, P., 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5470-5479).
- Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P. and Hedman, P., 2023. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 19697-19705).
- Besl, P. J. and McKay, N. D.,1992. Method for registration of 3-d shapes. In *Sensor Fusion IV : Control Paradigms and Data Structures*. International Society for Optics and Photonics. Volume 1611, pp. 586–607.
- Bhat, S.F., Birkl, R., Wofk, D., Wonka, P. and Müller, M., 2023. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*.
- Biasotti, S., Cerri, A., Falcidieno, B. and Spagnuolo, M., 2015. 3D artifacts similarity based on the concurrent evaluation of heterogeneous properties. *Journal on Computing and Cultural Heritage (JOCCH), 8*(4), pp.1-19.
- Border, R. and Gammell, J.D., 2024. The surface edge explorer (see): A measurement-direct approach to next best view planning. *The International Journal of Robotics Research*, *43*(10), pp.1506-1532.
- Boström, H., Andler, S.F., Brohede, M., Johansson, R., Karlsson, A., van Laere, J., Niklasson, L., Nilsson, M., Persson, A., Ziemke, T., 2003. On the Definition of Information Fusion as a Field of Research 8. *Report HS-IKI-TR-07-006*.
- Bruneau, R., Brument, B., Quéau, Y., Mélou, J., Lauze, F.B., Durou, J.D. and Calvet, L., 2025. Multi-view Surface Reconstruction Using Normal and Reflectance Cues. *arXiv preprint arXiv:2506.04115*.
- Cabrera-Revuelta, E., Tavolare, R., Buldo, M. and Verdoscia, C., 2024. Planning for terrestrial laser scanning: Methods for optimal sets of locations in architectural sites. *Journal of Building Engineering*, *85*, p.108599.
- Calvet, L., Gurdjos, P., Griwodz, C. and Gasparini, S., 2016. Detection and accurate localization of circular fiducials under highly challenging conditions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 562-570).
- Cao, J., Wang, H., Chemerys, P., Shakhrai, V., Hu, J., Fu, Y., Makoviichuk, D., Tulyakov, S. and Ren, J., 2023. Real-time neural light field on mobile devices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8328-8337).
- Caron, G., Dame, A., Marchand, E., 2014. Direct model based visual tracking and pose estimation using mutual information. *Image and Vision Computing 32*, pp.54–63.
- Castanedo, F., 2013. A Review of Data Fusion Techniques. *The Scientific World Journal 2013*, pp.1–19.
- Chen, A., Xu, Z., Geiger, A., Yu, J. and Su, H., 2022a. Tensorf: Tensorial radiance fields. In *European conference on computer vision* (pp. 333-350). Cham: Springer Nature Switzerland.
- Chen, C., Li, Y., Liu, W., Huang, J., 2015. SIRF: Simultaneous Image Registration and Fusion in A Unified Framework. *IEEE Transactions on Image Processing 24*, 4213–4224.
- Chen, Y., Medioni, G., 1992. Object modelling by registration of multiple range images. *Image and Vision Computing, Range Image Understanding 10*, pp. 145–155.

- Chen, Z., Tagliasacchi, A., Funkhouser, T. and Zhang, H., 2022c. Neural dual contouring. ACM Transactions on Graphics (TOG), 41(4), pp.1-13.
- Chen, Z., Zhang, W., Huang, R., Dong, Z., Chen, C., Jiang, L. and Wang, H., 2022b. 3D model-based terrestrial laser scanning (TLS) observation network planning for large-scale building facades. *Automation in Construction*, *144*, p.104594.
- Dai, P., Xu, J., Xie, W., Liu, X., Wang, H. and Xu, W., 2024, July. High-quality surface reconstruction using gaussian surfels. In *ACM SIGGRAPH 2024 conference papers* (pp. 1-11).
- Darmon, F., Bascle, B., Devaux, J.C., Monasse, P. and Aubry, M., 2022. Improving neural implicit surfaces geometry with patch warping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6260-6269).
- DeHass, M.C., Collins, L., Taitt, A., Raymond-Yakoubian, J., Doering, T., Ellanna, L.N., Hollinger, E., Gonzalez, J., John Jr, E., Martinez, D. and Tapqaq, M.S., 2025. Ethical Considerations in the Use of 3D Technologies to Preserve and Perpetuate Indigenous Heritage. *American Antiquity, 90*(2), pp.282-306.
- De Luca, L., Busarayat, C., Stefani, C., Renaudin, N., Florenzano, M. and Véron, P., 2010, June. An iconography-based modeling approach for the spatio-temporal analysis of architectural heritage. In *2010 Shape Modeling International Conference* (pp. 78-89). IEEE.
- DeGol, J., Bretl, T. and Hoiem, D., 2018. Improved structure from motion using fiducial marker matching. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 273-288).
- Demetrescu, Emanuel. 2015. «Archaeological Stratigraphy as a formal language for virtual reconstruction. Theory and practice». Journal of Archaeological Science 57: 42–55.
- Denard, H. (2013). Implementing best practice in cultural heritage visualisation: the London Charter. In C. *Corsi, B. Slapšak & F. Vermeulen (Eds.), Good practice in archaeological diagnostics: non-invasive survey of complex archaeological sites*, pp. 255–268.
- Duckworth, D., Hedman, P., Reiser, C., Zhizhin, P., Thibert, J.F., Lučić, M., Szeliski, R. and Barron, J.T., 2024. SMERF: Streamable memory efficient radiance fields for real-time large-scene exploration. *ACM Transactions on Graphics (TOG), 43*(4), pp.1-13.
- Dutagaci, H., Cheung, C.P. and Godil, A., 2010, October. A benchmark for best view selection of 3D objects. In *Proceedings of the ACM workshop on 3D object retrieval* (pp. 45-50).
- Esteban, J., Starr, A., Willetts, R., Hannah, P. and Bryanston-Cross, P., 2005. A review of data fusion models and architectures: towards engineering guidelines. *Neural Computing & Applications, 14*(4), pp.273-281.
- Ferdani D., Demetrescu E., Cavalieri, M., Pace, G., Lenzi, S., 2020. 3D modelling and visualization in field archaeology: from survey to interpretation of the past using digital technologies. *Groma Documenting Archaeology, 4*.
- Fraser, C.S., 1984. Network design considerations for non-topographic photogrammetry. *Photogrammetric Engineering and Remote Sensing, 50*(8), pp.1115-1126.
- Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B. and Kanazawa, A., 2022. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5501-5510).
- Garland, M. and Heckbert, P.S. (1997) 'Surface simplification using quadric error metrics', in Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques. USA: ACM Press/Addison-Wesley Publishing Co. (SIGGRAPH '97), pp. 209–216. Available at: https://doi.org/10.1145/258734.258849.
- Griwodz, C., Calvet, L. and Halvorsen, P., 2018, June. Popsift: A faithful SIFT implementation for real-time applications. In *Proceedings of the 9th ACM Multimedia Systems Conference* (pp. 415-420).
- Gui, M., Schusterbauer, J., Prestel, U., Ma, P., Kotovenko, D., Grebenkova, O., Baumann, S.A., Hu, V.T. and Ommer, B., 2025, April. DepthFM: Fast Generative Monocular Depth Estimation with Flow Matching. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 39, No. 3, pp. 3203-3211).
- Hanocka, R. et al. (2019) 'MeshCNN: a network with an edge', ACM Transactions on Graphics, 38(4), pp. 1–12. Available at: https://doi.org/10.1145/3306346.3322959.
- Hall, D.L., Garga, A.K., 1999. Pitfalls in data fusion (and how to avoid them), in: *Proceedings of the Second International Conference on Information Fusion* (Fusion'99). pp. 429–436.
- Hall, D. L. and Steinberg, A. (2001). Dirty secrets in multisensor data fusion. In *Hall, D., . L.-J., editor, Multisensor Data Fusion (1st ed.)*. CRC Press, Boca Raton.
- Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C. and Shen, S., 2024. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*
- Huber, D.F., Hebert, M., 2003. Fully automatic registration of multiple 3D data sets. *Image and Vision Computing 21*, pp. 637–650.

- Jun, H. and Nichol, A., 2023. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463.*
- Kondo, N., Ikeda, Y., Tagliasacchi, A., Matsuo, Y., Ochiai, Y. and Gu, S.S., 2021. Vaxnerf: Revisiting the classic for voxel-accelerated neural radiance field. *arXiv preprint arXiv:2111.13112.*
- Jain, A., Tancik, M. and Abbeel, P., 2021. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 5885-5894).
- Jakob, W., Tarini, M., Panozzo, D. and Sorkine-Hornung, O., 2015. Instant Field-Aligned Meshes. In ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia) 34(6). 189:1–189:15.
- Jankowski, J. and Hachet, M., 2013, May. A survey of interaction techniques for interactive 3D environments. In *Eurographics 2013-STAR*.
- Ke, B., Qu, K., Wang, T., Metzger, N., Huang, S., Li, B., Obukhov, A. and Schindler, K., 2025. Marigold: Affordable Adaptation of Diffusion-Based Image Generators for Image Analysis. *arXiv preprint arXiv:2505.09358.*
- Kerbl, B., Kopanas, G., Leimkühler, T. and Drettakis, G., 2023. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph., 42*(4), pp.139-1.
- Khaleghi, B., Khamis, A., Karray, F.O., Razavi, S.N., 2013. Multisensor data fusion: A review of the state-of-the-art. *Information Fusion 14*, pp. 28–44.
- Khoury, H.M., Kamat, V.R., 2009. Evaluation of position tracking technologies for user localization in indoor construction environments. *Automation in Construction 18*, pp. 444–457.
- Kim, J., Park, G. and Lee, S., 2025. Multiview Geometric Regularization of Gaussian Splatting for Accurate Radiance Fields. arXiv preprint arXiv:2506.13508.
- Klein, L.A., 2012. Sensor and Data Fusion: A Tool for Information Assessment and Decision Making, Second Edition. *SPIE*.
- Krajancich, B., Kellnhofer, P., Wetzstein, G., 2023. Towards Attention–aware Foveated Rendering. *ACM Trans. Graph. 42*, 4, 77.
- Lai, Z., Zhao, Y., Liu, H., Zhao, Z., Lin, Q., Shi, H., Yang, X., Yang, M., Yang, S., Feng, Y. and Zhang, S., 2025. Hunyuan3D 2.5: Towards High-Fidelity 3D Assets Generation with Ultimate Details. *arXiv preprint arXiv:2506.16504.*
- Lepetit, V. and Fua, P., 2005. Monocular model-based 3d tracking of rigid objects: A survey. *Foundations and Trends in Computer Graphics and Vision, 1*(1), pp.1-89.
- Levoy M. and Hanrahan P., 1996. Light field rendering. Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, *SIGGRAPH*, pp. 31-42.
- Li, P., Wang, R., Wang, Y. and Tao, W., 2020. Evaluation of the ICP algorithm in 3D point cloud registration. *IEEE access*, *8*, pp.68030-68048.
- Li, Z., Müller, T., Evans, A., Taylor, R.H., Unberath, M., Liu, M.Y. and Lin, C.H., 2023. Neuralangelo: High-fidelity neural surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8456-8465).
- Lin, L., Kang, H., Shi, Y., Duan, H., El Saddik, A. and Cai, W., 2025. ASimp: Automatic High-Poly 3D Mesh Simplification for Preprocessing Based on QoE, In *IEEE International Conference on Multimedia & Expo (ICME'25)*, Nantes, France, June 30 – July 4, 2025
- Liu, M., Shi, R., Chen, L., Zhang, Z., Xu, C., Wei, X., Chen, H., Zeng, C., Gu, J. and Su, H., 2024. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10072-10083).
- Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S. and Vondrick, C., 2023b. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9298-9309).
- Liu, Y.T., Wang, L., Yang, J., Chen, W., Meng, X., Yang, B. and Gao, L., 2023a. Neudf: Leaning neural unsigned distance fields with volume rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 237-247).
- LondonAR. London AR Trail for Heritage Quarter. 2025. URL: https://zubr.co/case-study/london-ar-trail/.
- London Charter. The London Charter for the Computer-based Visualisation of Cultural Heritage. URL: https://londoncharter.org (2006).
- Long, X., Lin, C., Liu, L., Liu, Y., Wang, P., Theobalt, C., Komura, T. and Wang, W., 2023. Neuraludf: Learning unsigned distance fields for multi-view reconstruction of surfaces with arbitrary topologies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 20834-20843).
- Lorensen, W.E. and Cline, H.E. (1987) 'Marching cubes: A high resolution 3D surface construction algorithm', SIGGRAPH Comput. Graph., 21(4), pp. 163–169. Available at: https://doi.org/10.1145/37402.37422.

- Low, K.L., 2004. Linear least-squares optimization for point-to-plane icp surface registration. *Chapel Hill, University of North Carolina, 4*(10), pp.1-3.
- Lu, H., Wu, C., Tan, H., & Yan, H. (2024, December). Mesh Simplification Method Based on Fast Retrieval and Region-Preserving Quadric Error Metrics. In 2024 4th International Symposium on Artificial Intelligence and Intelligent Manufacturing (AIIM) (pp. 923-928). IEEE.
- Luhmann, T., Robson, S., Kyle, S. and Boehm, J., 2023. *Close-range photogrammetry and 3D imaging*. Walter de Gruyter GmbH & Co KG.McGlone, J.C. (2013)
- Makadia, A., Patterson, A. and Daniilidis, K., 2006, June. Fully automatic registration of 3D point clouds. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* (Vol. 1, pp. 1297-1304). IEEE.
- Masiero, Andrea, Fissore, F., Guarnieri, A., Pirotti, F., Visintini, D., Vettore, A., Masiero, A., Fissore, F., Guarnieri, A., Pirotti, F., Visintini, D., Vettore, A., 2018a. Performance Evaluation of Two Indoor Mapping Systems: Low-Cost UWB-Aided Photogrammetry and Backpack Laser Scanning. *Applied Sciences 8*, 416.
- Masiero, A., Fissore, F., Guarnieri, A. and Vettore, A., 2018b. Indoor photogrammetry aided with Uwb navigation. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 42*, pp.683-690.
- Medici, M., Perda, G., Sterpin, A., Farella, E.M., Settimo, S. and Remondino, F., 2024. Separate and Integrated Data Processing for the 3D Reconstruction of a Complex Architecture. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 48*, pp.249-256.
- Meng, X., Chen, W. and Yang, B., 2023. Neat: Learning neural implicit surfaces with arbitrary topologies from multi-view images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 248-258).
- Menna, F., Torresani, A., Battisti, R., Nocerino, E., Remondino, F., 2022. A Modular and Low-cost Portable VSLAM System for Real-time 3D Mapping: From Indoor and Outdoor Spaces to Underwater Environments. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. XLVIII-2/W1-2022*, pp.153–162.
- Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R. and Ng, R., 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM, 65*(1), pp.99-106.
- Morelli, L., Ioli, F., Maiwald, F., Mazzacca, G., Menna, F. and Remondino, F., 2024. Deep-image-matching: a toolbox for multiview image matching of complex scenarios. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 48, pp.309-316.
- Mortara, M., Spagnuolo, M., 2009. Semantics-driven best view of 3D shapes. *Computers & Graphics 33*, pp. 280–290.
- Müller, T., Evans, A., Schied, C. and Keller, A., 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG), 41*(4), pp.1-15.
- Munkelt, C., Breitbarth, A., Notni, G., Denzler, J., 2010. Multi-view planning for simultaneous coverage and accuracy optimisation. *Proc. British Machine Vision Conference.*
- Nichol, A., Jun, H., Dhariwal, P., Mishkin, P. and Chen, M., 2022. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751.*
- Niemeyer, M., Mescheder, L., Oechsle, M. and Geiger, A., 2020. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3504-3515).
- Nocerino, E., Menna, F. and Remondino, F., 2014. Accuracy of typical photogrammetric networks in cultural heritage 3D modeling projects. *The international archives of the photogrammetry, remote sensing and spatial information sciences, 40*, pp.465-472.
- Oechsle, M., Peng, S. and Geiger, A., 2021. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 5589-5599).
- Palma, G., Corsini, M., Dellepiane, M. and Scopigno, R., 2010, November. Improving 2D-3D Registration by Mutual Information using Gradient Maps. In *Eurographics Italian Chapter Conference* (pp. 89-94).
- Pamart, A., Abergel, V., de Luca, L., Veron, P., 2023. Toward a Data Fusion Index for the Assessment and Enhancement of 3D Multimodal Reconstruction of Built Cultural Heritage. *Remote Sensing 15*, 2408.
- Peng, A., 2024. Efficient neural light fields (enelf) for mobile devices. *arXiv preprint arXiv:2406.00598.*
- Pomerleau, F., Colas, F., Siegwart, R., Magnenat, S., 2013. Comparing ICP variants on real-world data sets: Open-source library and experimental protocol. *Autonomous Robots 34*, pp.133–148.
- Qian, G., Mai, J., Hamdi, A., Ren, J., Siarohin, A., Li, B., Lee, H.Y., Skorokhodov, I., Wonka, P., Tulyakov, S. and Ghanem, B., 2023. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843.*

- Ramos, M.M., Remondino, F., 2015. Data fusion in Cultural Heritage - A Review. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XL-5/W7*, pp. 359–363.
- Ranftl, R., Lasinger, K., Hafner, D., Schindler, K. and Koltun, V., 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence, 44*(3), pp.1623-1637.
- Ranftl, R., Bochkovskiy, A. and Koltun, V., 2021. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 12179-12188).
- Remondino, F., Fraser, C., 2006. Digital camera calibration methods: considerations and comparisons. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences 36*, pp.266–272.
- Remondino, F., Menna, F., Koutsoudis, A., Chamzas, C. and El-Hakim, S., 2013, October. Design and implement a reality-based 3D digitisation and modelling project. In *2013 Digital Heritage International Congress (DigitalHeritage)* (Vol. 1, pp. 137-144). IEEE.
- Remondino, F. and Campana, S., 2014. *3D recording and modelling in archaeology and cultural heritage.* Oxford: British Archaeological Reports.
- Remondino, F. and Stylianidis, E., 2016. *3D recording, documentation and management of cultural heritage (Vol. 2)*. Dunbeath, UK: Whittles Publishing.
- Rosu, R.A. and Behnke, S., 2023. Permutosdf: Fast multi-view reconstruction with implicit surfaces using permutohedral lattices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8466-8475).
- Ruiz de Oña, E., Barbero-García, I., González-Aguilera, D., Remondino, F., Rodríguez-Gonzálvez, P., Hernández-López, D., 2023. PhotoMatch: An Open-Source Tool for Multi-View and Multi-Modal Feature-Based Image Matching. *Applied Sciences 13*, 5467.
- Rusinkiewicz, S. and Levoy, M. (2001). Efficient variants of the ICP algorithm. In *3-D Digital Imaging and Modeling, 2001*. Proceedings. Third International Conference on, pages 145–152. IEEE.
- Rusu, R. B., Blodow N., and Beetz, M., 2009. Fast Point Feature Histograms (FPFH) for 3D registration. *IEEE International Conference on Robotics and Automation*, Kobe, Japan, 2009, pp. 3212-3217.
- Salti, S., Tombari, F. and Di Stefano, L., 2014. SHOT: Unique signatures of histograms for surface and texture description. *Computer vision and image understanding*, *125*, pp.251-264.
- Salvi, J., Armangué, X., Batlle, J., 2002. A comparative review of camera calibrating methods with accuracy evaluation. *Pattern Recognition 35*, pp. 1617–1635.
- Seville Principles. Principles of Seville. International Principles of Virtual Archaeology. URL: https://icomos.es/wp-content/uploads/2020/06/Seville-Principles-IN-ES-FR.pdf (2011).
- Shen, T. et al. (2023) 'Flexible Isosurface Extraction for Gradient-Based Mesh Optimization', ACM Trans. Graph., 42(4). Available at: https://doi.org/10.1145/3592430.
- Schroeder, W. J., Zarge, J. A. & Lorensen, W. E. (1992). Decimation of triangle meshes. ACM SIGGRAPH '92, Computer Graphics, Vol. 26 No. 2, pp. 65–70.
- Chane, S., C., Mansouri, A., Marzani, F.S., Boochs, F., 2013. Integration of 3D and multispectral data for cultural heritage applications: Survey and perspectives. *Image and Vision Computing 31*, 91–102.
- Spencer, J., Russell, C., Hadfield, S. and Bowden, R., 2024. Kick back & relax++: Scaling beyond ground-truth depth with slowtv & cribstv. *arXiv preprint arXiv:2403.01569.*
- Stathopoulou, E.K. and Remondino, F., 2023. A survey on conventional and learning-based methods for multi-view stereo. *The Photogrammetric Record, 38*(183), pp.374-407.
- Steinberg, A.N. and Bowman, C.L., 2017. Revisions to the JDL data fusion model. In *Handbook of multisensor data fusion* (pp. 65-88). CRC press.
- Sun, C., Choe, J., Loop, C., Ma, W.C. and Wang, Y.C.F., 2025. Sparse Voxels Rasterization: Real-time High-fidelity Radiance Field Rendering. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 16187-16196).
- Sun, C., Sun, M. and Chen, H.T., 2022a. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5459-5469).
- Sun, J., Chen, X., Wang, Q., Li, Z., Averbuch-Elor, H., Zhou, X. and Snavely, N., 2022b. Neural 3d reconstruction in the wild. In *ACM SIGGRAPH 2022 conference proceedings* (pp. 1-9).
- Sun, J., Peng, C., Shao, R., Guo, Y.C., Zhao, X., Li, Y., Cao, Y., Zhang, B. and Liu, Y., 2024. Dreamcraft3d++: Efficient hierarchical 3d generation with multi-plane reconstruction model. *arXiv preprint arXiv:2410.12928.*
- Taketomi, T., Uchiyama, H., Ikeda, S., 2017. Visual SLAM algorithms: a survey from 2010 to 2016. IPSJ *Transactions on Computer Vision and Applications 9*, 16.

- Tang, J., Ren, J., Zhou, H., Liu, Z. and Zeng, G., 2023. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653.*
- Theiler, P.W., Wegner, J.D., Schindler, K., 2014. Keypoint-based 4-Points Congruent Sets – Automated marker-less registration of laser scans. *ISPRS Journal of Photogrammetry and Remote Sensing 96*, pp.149–163.
- Tochilkin, D., Pankratz, D., Liu, Z., Huang, Z., Letts, A., Li, Y., Liang, D., Laforte, C., Jampani, V. and Cao, Y.P., 2024. Triposr: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151.*
- Tombari, F., Di Stefano, L., 2014. Interest Points via Maximal Self-Dissimilarities, in: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (Eds.), *Computer Vision -- ACCV 2014.* Springer International Publishing, Cham, pp. 586–600.
- Tosi, F., Ramirez, P.Z. and Poggi, M., 2024, September. Diffusion models for monocular depth estimation: Overcoming challenging conditions. In *European Conference on Computer Vision* (pp. 236-257). Cham: Springer Nature Switzerland.
- Trummer, M., Munkelt, C., Denzler, J., 2010. Online next-best-view planning for accuracy optimization using an extended e-criterion. *Proc. IEEE International Conference on Pattern Recognition (ICPR'10)*, pp. 1642–1645.
- Usamentiaga, R., Venegas, P., Guerediaga, J., Vega, L., Molleda, J., Bulnes, F., 2014. Infrared Thermography for Temperature Measurement and Non-Destructive Testing. *Sensors 14*, 12305–12348.
- Viola, P., Wells III, W.M., 1997. Alignment by Maximization of Mutual Information. *International Journal of Computer Vision 24*, pp.137–154.
- Wald, L., 2002. *Data fusion: definitions and architectures ; fusion of images of different spatial resolutions.* Les Presses de l'École des Mines, Paris.
- Wan, Z., Paschalidou, D., Huang, I., Liu, H., Shen, B., Xiang, X., Liao, J. and Guibas, L., 2024. Cad: Photorealistic 3d generation via adversarial distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10194-10207).
- Wang, H., Ren, J., Huang, Z., Olszewski, K., Chai, M., Fu, Y. and Tulyakov, S., 2022, October. R2l: Distilling neural radiance field to neural light field for efficient novel view synthesis. In *European Conference on Computer Vision* (pp. 612-629). Cham: Springer Nature Switzerland.
- Wang, P. and Shi, Y., 2023. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201.*
- Wang, Y., Han, Q., Habermann, M., Daniilidis, K., Theobalt, C. and Liu, L., 2023. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 3295-3306).
- Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C. and Novotny, D., 2025. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 5294-5306).
- Wang, Y., Huang, D., Ye, W., Zhang, G., Ouyang, W. and He, T., 2024. Neurodin: A two-stage framework for high-fidelity neural surface reconstruction. *Advances in Neural Information Processing Systems, 37*, pp.103168-103197.
- William, S., 1992. Decimation of triangle meshes. In Proc. of SIGGRAPH 92 (pp. 65-70).
- Wu, Q., Wang, K., Li, K., Zheng, J. and Cai, J., 2023. Objectsdf++: Improved object-compositional neural implicit surfaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 21764-21774).
- Xu, D., Jiang, Y., Wang, P., Fan, Z., Shi, H. and Wang, Z., 2022, October. Sinnerf: Training neural radiance fields on complex scenes from a single image. In *European Conference on Computer Vision* (pp. 736-753). Cham: Springer Nature Switzerland.
- Xu, Q., Wang, W., Ceylan, D., Mech, R. and Neumann, U., 2019. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *Advances in neural information processing systems, 32.*
- Xu, J., Cheng, W., Gao, Y., Wang, X., Gao, S. and Shan, Y., 2024. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191.*
- Yan, J., Zhao, H., Bu, P. and Jin, Y., 2021, December. Channel-wise attention-based network for self-supervised monocular depth estimation. In *2021 International Conference on 3D vision (3DV)* (pp. 464-473). IEEE.
- Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J. and Zhao, H., 2024. Depth anything v2. *Advances in Neural Information Processing Systems, 37*, pp.21875-21911.
- Yariv, L., Gu, J., Kasten, Y. and Lipman, Y., 2021. Volume rendering of neural implicit surfaces. *Advances in neural information processing systems, 34,* pp.4805-4815.

- Yariv, L., Hedman, P., Reiser, C., Verbin, D., Srinivasan, P.P., Szeliski, R., Barron, J.T. and Mildenhall, B., 2023, July. Bakedsdf: Meshing neural sdfs for real-time view synthesis. In *ACM SIGGRAPH 2023 conference proceedings* (pp. 1-9).
- Yu, A., Li, R., Tancik, M., Li, H., Ng, R. and Kanazawa, A., 2021a. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 5752-5761).
- Yu, A., Ye, V., Tancik, M. and Kanazawa, A., 2021b. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4578-4587).
- Yu, Z., Peng, S., Niemeyer, M., Sattler, T. and Geiger, A., 2022. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems, 35*, pp.25018-25032.
- Yuan, W., Gu, X., Dai, Z., Zhu, S. and Tan, P., 2022. New crfs: Neural window fully-connected crfs for monocular depth estimation. *arXiv preprint arXiv:2203.01502.*
- Zhang, Z., 1994. Iterative point matching for registration of free-form curves and surfaces. *International Journal of Computer Vision 13*, pp. 119–152.
- Zhang, Z., 2021. Iterative closest point (ICP). In *Computer vision: a reference guide* (pp. 718-720). Cham: Springer International Publishing.
- Zitová, B., Flusser, J., 2003. Image registration methods: a survey. *Image and Vision Computing 21*, 977–1000.
- Zou, Z.X., Yu, Z., Guo, Y.C., Li, Y., Liang, D., Cao, Y.P. and Zhang, S.H., 2024. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10324-10335).

Annex 1 - Multi-image technologies

Technology	Description
DVGO Direct Voxel Grid Optimization (Sun et al., 2022a) [https://sunset1995.github.io/dvgo/]	It is a method that is able to accelerate NeRFs by two orders of magnitude, using a *hierarchical voxel-based representation (3D grid of cubes)*, where it stores view-dependent information that is encoded with the help of a neural network. That hierarchical data structure, along with further optimisations and the use of a much simpler neural network, makes *DVGO* 300 times faster than NeRFs, without sacrificing photorealism and visualisation quality.
Gaussian Splatting (Kerbl et al., 2023) [https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting]	*Gaussian Splatting (GS)* is a new approach based on a rather old idea for representing spatial information using *specialised-augmented point clouds*. The core idea behind GS involves representing each 3D point as a small, coloured, oriented Gaussian volume. During rendering, these Gaussians are projected (splatted) and blended onto the camera's 2D focal plane, creating smooth and realistic images. In this way, GS can efficiently render complex 3D scenes with varying densities and materials, by approximating these complexities in a computationally efficient way, making GS suitable for both volume and surface rendering. Similar to NeRFs, GS are generated from multiple images taken with known camera positions and orientations, combined with a sparse point cloud representation of the scene. However, unlike NeRFs, the radiance information from the scene is encoded and rendered without the need for deep neural networks, making GS an order of magnitude faster than NeRFs.
Instant-NGP Instant Neural Graphics Primitives (Muller et al., 2022) [https://nvlabs.github.io/instant-ngp/]	Presented in 2022 by *Nvidia* researchers, *Instant-NGP (iNGP)* is a technology that provides near-instant training of neural graphics primitives on a single GPU. As graphics primitives, it is capable of tackling 2D images, *Signed Distance Function (SDF) fields*, NeRFs and other radiance and density volumetric field representations. By utilising a smaller and more efficient neural network architecture, along with a multiresolution structure that holds features from the training process, iNGP is able to speed up both the encoding and rendering of the encoded information. In the case of NeRFs, iNGP is capable of capturing the same level of detail in a matter of minutes or even seconds instead of hours that are required by the early implementations of NeRFs representation.
Mip-NeRF 360 Neural Radiance Fields (Barron et al., 2022) [https://github.com/google-research/multinerf]	It is a method based on neural radiance field rendering similar to NeRF. However, this method tries to improve visualisation quality by eliminating *aliasing artifacts* during rendering. This is achieved by representing the scene at a continuous range of scales rather than just points along rays. Also it can provide visualisation for open and large unbounded scenes, in great detail, without significant visualisation artifacts like ghosting and blurry backgrounds. Moreover, the proposed neural network used to encode the scene is 22 times faster than that used initially by NeRF.

NeRF Neural Radiance Fields (Mildenhall et al., 2021) [https://github.com/bmild/nerf]	It is a method for capturing and representing the 3D world using a volumetric approach powered by AI. More specifically *NeRF* are able to compress and encode the information from numerous images, depicting the same subject into a volumetric 3D scene representation known as radiance field, using *Coordinate Based Neural Networks.* NeRF works by inputting a collection of images, along with their corresponding camera positions and orientations, into a neural network. Using that data, the neural network is trained in order to precisely model how light travels through a 3D space, and is represented as a continuous volumetric function. As a result, NeRF can generate high-quality, photorealistic images of the scene from completely new viewpoints, offering a smooth and continuous rendering experience.
NeLFs Neural Light Fields (Wang et al., 2022) [https://snap-research.github.io/R2L/]	*NeLFs* is a new neural rendering approach that combines deep learning with a concept first introduced in 1996 by (Levoy and Hanrahan, 1996). NeLFs use the classic light slab (two-plane) representation, introduced back then, that parameterises each light ray as an ordered pair of intersection points with two fixed planes. This representation enables efficient grid-based encoding of the light field, allowing NeLFs to learn a direct mapping from rays to pixel colors with reduced computational complexity. By revisiting this classic model, NeLFs can produce continuous and accurate scene representation, while achieving faster rendering than NeRFs, making them well-suited for real-time applications, especially on mobile devices (Cao et al., 2023; Peng, 2024). NeLFs implements a two-step approach: i) the training of a NeRF model for extracting synthetic rendered images to train NeLF; ii) the NeLF finetuning using the real data (captured images), extending even further the visual quality of the synthetic images.
Zip-NeRF Anti-Aliased Grid-Based Neural Radiance Fields (Barron et al., 2023) [https://jonbarron.info/zipnerf/]	*Zip-NeRF* is a neural radiance field rendering method similar to NeRF but two orders of magnitude faster, with an increased rendering quality. This is achieved by combining characteristics of *Mip-NeRF 360* and those of *Instant-NGP* (Muller et al., 2022). As a result, Zip-NeRF is able to provide continuous photorealistic visualisation for large open scenes without the computation overhead that traditional NeRF approach demands. This enables NeRF generation on modern high-end hardware in a couple of hours instead of days.
VaxNeRF Voxel-AcceleratedNeural Radiance Field (Kondo et al., 2021) [https://github.com/naruya/VaxNeRF]	*VaxNeRF* is an approach to speed up the generation of NeRFs by utilising a simple volumetric representation of the scene. This representation denotes which areas of the scene should be included in the radiance field for neural network training, thereby significantly reducing the amount of computation required. This volumetric representation is created using a classic image-based 3D digitisation technique called *shape-from-silhouette*, which is able to approximate the 3D model of a scene with a *3D visual hull*. The visual hull is not an exact 3D model of the scene, as it is a solid formed by projecting the subject's 2D silhouettes from the multiple viewpoints into 3D space and then taking the intersection of all these projections, keeping only the common volume shared by

	them, using voxels (volumetric pixels / grid of cubes).
SMERF Streamable Memory Efficient Radiance Fields (Duckworth et al., 2024) [https://smerf-3d.github.io]	*SMERF* is a cutting-edge novel view synthesis approach proposed by *Google*, which shows the potential use of Radiance Field visualisation for real time navigation of large scenes. Using neural network distillation and hierarchical volume data structures, SMERF can render extremely large scenes with very fine details inside the browser on common devices, like laptops and smartphones. This technology illustrates how future versions of Google Street View could render their global-scale data.
TensoRF Tensorial Radiance Fields (Chen et al., 2022a) [https://apchenstu.github.io/TensoRF]	This is an alternative to the NeRF approach that encodes the radiance field representation of a scene into a simple mathematical object called *Tensor*, without using computationally expensive deep neural networks. *TensorRF* is taking a much simpler machine learning and multilinear algebra approach to efficiently encode, store and render a volumetric radiance representation of the 3D scene. Despite the computation simplicity of this methodology and the compact representation that is able to produce, TensorRF is able to match or even surpass the visual quality of NeRF in a matter of minutes instead of days.
SVASTER Sparse Voxels Rasterization: Real-time High-fidelity Radiance Field Rendering (Sun et al., 2025) [https://svraster.github.io]	Svaster is an efficient radiance field rendering algorithm that uses adaptive sparse voxels representation for scene storage, combined with a customised rasterisation process for rendering, without relying on neural networks. Its two key innovations are adaptively allocating sparse voxels at multiple levels of detail to capture fine scene features at very high resolution while maintaining high frame rates. Svaster achieves state-of-the-art novel view synthesis from multiple posed images, with significantly better quality and over tenfold rendering speed improvement than previous neural-free voxel methods. Apart from visualisation, SVASTER is able to reconstruct a 3D polygonal mesh of the captured scene. However, it might suffer from high-frequency artifacts when there are abrupt colour changes appearing on the subject's surface texture.
StopThePop Sorted Gaussian Splatting for View-Consistent Real-time Rendering (Radl et al., 2024) [https://github.com/r4dl/StopThePop]	This is a method that tries to improve the rendering of GS (Kerbl et al., 2023) by eliminating visual popping and blending artifacts that appear during novel view synthesis. This method introduces a hierarchical rasterisation technique that efficiently resorts and culls splats per pixel, eliminating those artifacts without heavy computation. *StopThePop* improves view consistency and prevents cheating view-dependent effects, achieving similar image quality while being only slightly slower than the original approach. However, it reduces the number of Gaussians by half without quality loss, nearly doubling rendering speed and halving memory usage.
LightGaussian Unbounded 3D Gaussian Compression (Fan et al., 2024) [https://lightgaussian.github.io]	This is a novel method that compresses 3D Gaussian representations for more efficient and compact scene reconstruction by identifying and removing Gaussians that contribute little to the scene, reducing redundancy while

	preserving visual quality through a pruning and recovery process. Also the remaining information is compressed, using quantisation and further distillation with view dependent functions like spherical harmonics. This gives great efficiency and speed improvements during rendering, while not sacrificing visual quality to the average human perception level.
CompGS Smaller and Faster Gaussian Splatting with Vector Quantization (Navaneet et al., 2023) [https://github.com/UCDvision/compact3d]	This is a method for Gaussian Splatting optimisation that manages to compress their data using neural network-based clustering, in order to group similar gaussians together into quantised indexes that are further compressed using classical compression methods. This method allows for a reduction of the required data storage volume by 40 to 50 times, while doubling at least the rendering speed with only a subtle reduction in visual quality.
RadSplat Radiance Field-Informed Gaussian Splatting for Robust Real-Time Rendering (Niemeyer et al., 2025) [https://m-niemeyer.github.io/radsplat/]	*RadSplat* is a method that combines neural fields with point-based 3D representations to enable fast, high-quality rendering of complex scenes. It uses radiance fields as a guide to improve the accuracy and stability of optimising point-based models. *RadSplat* introduces a pruning technique that reduces scene size while enhancing visual quality by avoiding ghosting artifacts and making the scene more compact at the same time. Additionally, it uses a test-time filtering process that speeds up rendering even more without sacrificing quality. This approach achieves state-of-the-art results on common benchmarks and renders scenes up to 3 orders of magnitude faster than previous methods when rendered using a high-end gaming-grade GPU.
2DGS 2D Gaussian Splatting for Geometrically Accurate Radiance Fields (Huang et al., 2024) [https://surfsplatting.github.io]	This method is a novel approach to improve the accuracy of neural surface representation provided by the classical 3D Gaussian Splatting methods by collapsing the 3D volume into oriented 2D Gaussian disks, which inherently model surfaces and maintain view consistency. This approach enables noise-free and detailed geometry reconstruction and textured mesh extraction, while at the same time allowing radiance field-based real-time rendering with competitive visual quality, without presenting ghosting artifacts. However, this technique relies on multiview constructed depth maps for surface extraction that favors opaque surfaces with plenty of texture features that do not present strong light interactions.
SuGaR Surface-Aligned Gaussian Splatting for Efficient 3D Mesh Reconstruction and High-Quality Mesh Rendering (Guedon et al., 2024) [https://imagine.enpc.fr/~guedona/sugar/]	This is a novel method that is able to extract a 3D polygonal mesh model from a 3D Gaussian Splatting representation, in a matter of minutes on a single GPU. This is possible by the close alignment of the gaussians onto the surface of the scene through a regularisation term. Furthermore, a new rendering technique is introduced that applies colors and materials to the underlying extracted geometry through gaussian splatting, combining the advantages of both worlds, enabling photorealistic GS visualisation but at the same time the flexibility of a polygonal mesh model that can be edited, lit with various ways, be part of scene compositions, as well as rigged and animated.
Plenoxels	*Plenoxels* is a *view-dependent sparse voxel representation* for

(Fridovich-Keil et al., 2022) [https://alexyu.net/plenoxels/]	novel view synthesis and photorealistic scene capture and rendering that can rival NeRFs, since it is 2 orders of magnitude (100x) faster. Similar to GS above, Plenoxels representation does not depend on neural networks for scene encoding and rendering. It uses instead a discrete 3D grid, where the scene is encoded in stacked cubes that hold colour and density information that varies in dependence to the viewing direction. This combination of volume rendering with view-dependent characteristics enables photorealistic rendering with complex light interactions.
Plenoctrees (Yu et al., 2021a) [https://alexyu.net/plenoctrees/]	Like Plenoxels (Fridovich-Keil et al., 2022), it is a method that tries to speed up NeRFs rendering with a hierarchical data structure. It encodes the 3D visual information from multiple photos into an *octree representation* that holds view-dependent values of the captured scene. Despite the use of neural networks, *Plenoctrees* manage an increased speed up for the generation of its radiance fields structure, taking advantage of the hierarchical characteristics of the octree structure. A wise decision was taken to differentiate the rendering procedure from the encoding, in order to avoid the use of the computationally heavy neural network rendering approach. Instead, the neural network is used to produce a data structure in a certain way that can be efficiently rendered with conventional rasterisation methods, even on a web browser using common hardware.
Radiant Foam Real-Time Differentiable Ray Tracing (Govindarajan et al.,2025) [https://radfoam.github.io]	This is an ingenious scene representation technique for real-time photorealistic rendering of a scene from a collection of 2D images. The difference from techniques like NeRFs and GS lies in the way the volumetric data structure is formed, used for storing the appearance of the 3D scene. Radiant Foam subdivides space using a volumetric partitioning that resembles the foam by partitioning space using Voronoi tessellation, so that each foam bubble accomodates a significant point that is used for the scene reconstruction. Radiant Foam is capable of producing high-quality visuals, comparable to NeRFs and GS, with much greater performance, without requiring neural network inference during rendering. Instead, this foam-like partitioning of space is ray-traced efficiently using modern consumer gaming hardware.
Neuralangelo High-Fidelity Neural Surface Reconstruction (Li et al., 2023) [https://research.nvidia.com/labs/dir/neuralangelo/]	It is a neural surface reconstruction method proposed by *Nvidia* and is based on their previous work presented as *Instant-NGP*. Similarly, *Neuralangelo* takes advantage of a *hierarchical grid representation* of SDF, that are encoded into hashes using neural networks, in order to extract detailed surface geometry from a set of images with known position and orientation in space. Due to the hierarchical structure, it is able to encode large scenes in great detail, without extreme memory requirements. Nevertheless, as a neural network-based technology, encoding and rendering of the data structure requires high-end hardware that is costly at the moment and power hungry.
NeuS2	*NeuS2* is a fast neural implicit surface reconstruction method

Fast Learning of Neural Implicit Surfaces for Multi-view Reconstruction (Wang et al., 2023) [https://vcai.mpi-inf.mpg.de/projects/NeuS2/]	designed for multi-view 3D scene reconstruction, for static and dynamic scenes. It significantly accelerates the training process compared to previous methods like NeRFs, achieving a two orders of magnitude speedup without sacrificing reconstruction quality, working with a hierarchical network architecture, similar to *Instant-NGP*. Unlike NeRFs and Gaussian Splatting, NeuS2 is not targeting photorealistic visualisation through radiance fields. Instead, it is producing a volumetric representation of SDF. This type of representation can then be visualised directly using *ray tracing* or converted into a triangular representation, which can be exported in a common 3D mesh format that is compatible for use in conventional 3D applications and more practical uses like 3D printing.
RNb-NeuS2 Multi-View Surface Reconstruction Using Normal and Reflectance Cues (Bruneau et al., 2025) [https://robinbruneau.github.io/publications/rnb_neus2.html]	This is a recent state-of-the-art method for multiview neural geometry reconstruction that is implemented within the NeuS2 framework above. However, it augments NeuS2 radiance surface reconstruction by the introduction of light reflectance visual cues. These cues are provided using a technique similar to photometric stereo, that requires taking pictures of the same subject from the same camera position while moving the light source illuminating the scene. This helps in the identification and reconstruction of miniscule surface geometric features such as bumps, cracks and grooves providing results that are by far more detailed even when compared to state of the art Multi View Stereo (MVS) techniques that are not taking into consideration the angle of the light source. Furthermore, RNb-NeuS2 speeds up neural surface reconstruction close to two orders of magnitude compared to NeuS2, by working on small image patches, similar to MVS. This is a technique capable of producing high-fidelity 3D surface reconstruction for complex surfaces that features extreme details. However, in order to do so requires controlled lighting conditions and somewhat specialised lighting rig equipment.
NeUDF Leaning Neural Unsigned Distance Fields with Volume Rendering (Liu et al., 2023a) [http://geometrylearning.com/neudf/] **NeuralUDF** Learning Unsigned Distance Fields for Multi-view Reconstruction of Surfaces with Arbitrary Topologies (Long et al., 2023) [https://www.xxlong.site/NeuralUDF/]	Similar to *NeuS2* (Wang et al., 2023), *NeUDF* and *NeuralUDF* are both neural implicit surface reconstruction methods that are especially targeted to overcome the limitation of watertight surface generation that characterises NeuS2 and is imposed by the use of SDF. To overcome this limitation, both of these methods are using an *Unsigned Distance Function (UDF) field* representation, achieving high fidelity surface reconstruction of complex shapes with open boundaries.
NeAT Leaning Neural Implicit Surfaces with Arbitrary Topologies from Multi-view Images (Meng et al., 2023) [https://xmeng525.github.io/xiaoxumeng.github.io/projects/cvpr23_neat]	It is a neural implicit surface reconstruction method that relies on a neural representation of a *UDF* field, similar to *NeuS2*. However, despite using UDF, it is capable of coping also with arbitrary open surfaces, using a neural network that validates surface existence at a specific point in space.

Unisurf Unifying Neural Implicit Surfaces and Radiance Fields for Multi-View Reconstruction (Oechsle et al., 2021) [https://moechsle.github.io/unisurf/]	This is a multi-image based approach that unifies both novel view and implicit 3D surface reconstruction using the same volumetric representation. This is performed by *narrowing volumetric sampling intervals*, shifting from a broad volume sampling to focused surface sampling, enabling in this way efficient learning of both geometry and appearance in a single model.
Binary Opacity Grids Capturing Fine Geometric Detail for Mesh-Based View Synthesis (Reiser et al., 2024) [https://creiser.github.io/binary_opacity_grid/]	This method improves similar volumetric scene representation methods by changing how they represent surfaces. Thus, instead of using smooth density, they use a grid of opacity values that sharply switch from transparent to opaque right at the surface. They also cast multiple rays per pixel to better capture edges and tiny structures without blur. By encouraging the opacity to become clearly solid or empty but not something in between (like NeRF and other volumetric approaches), it facilitates the extraction of clean, precise surfaces. Furthermore, this technique is able to generate simplified, compact meshes that can be rendered quickly, even on mobile devices, while producing much clearer and more accurate 3D views than older mesh-based techniques. Another great advantage of mesh reconstruction is that it can be exported to a format compatible with common 3D applications, ranging from content creation to 3D printing.
VolSDF Volume Rendering of Neural Implicit Surfaces (Yariv et al., 2021) [https://lioryariv.github.io/volsdf/]	Similar to *Unisurf* above, *VolSDF* is a neural volumetric representation that enables both novel view generation as well as implicit 3D surface reconstruction of a photographed scene. VolSDF demonstrates superior surface reconstruction quality on challenging multi-view datasets compared to prior neural volume rendering methods, while maintaining excellent novel view synthesis, from a sparse set of input images.
NeuralReconW Neural 3D Reconstruction in the Wild (Sun et al., 2022b) [https://zju3dv.github.io/neuralrecon-w/]	It is a method for neural surface reconstruction designed to work with image sequences that exhibit significant motion between frames (which is a challenging case not only for MVS-SfM methods but for other neural network based approaches as well). As for neural methods-based cases, it is using a volumetric representation of SDF that are produced from the posed image depth maps. *NeuralReconW* is able to produce dense image depth maps for pictures with considerable diversity, in terms of view point and angle, lighting conditions, and camera characteristics. By leveraging a *Transformer Neural Network* architecture for the dense image feature matching, it enables the creation of high detailed 3D surface reconstruction from diverse image collections depicting the same subject, such as those gathered through *crowdsourcing*.
NeuRodin A Two-stage Framework for High-Fidelity Neural Surface Reconstruction (Wang et al., 2024) [https://open3dvlab.github.io/NeuRodin/]	*NeuRodin* is a neural surface reconstruction method that combines Signed Distance Function fields with density fields in order to produce highly detailed 3D models of arbitrary topology. The produced 3D models are on par or better compared to SfM methods, capturing subtle geometric details but also reconstructing smooth featureless surfaces. Obviously, the reconstructed surface can be extracted as a polygonal mesh using a universal format, in order to be used in mainstream 3D applications.

ObjectSDF++ Improved Object Compositional Neural Implicit Surfaces (Wu et al., 2023) [https://wuqianyi.top/objectsdf++]	This is an improved neural implicit surface reconstruction technique from multi-view images that uses *object masks guidance* to confine errors and improve the overall reconstruction quality of scenes and individual objects. These masks are produced automatically using a neural network to segment the individual objects of the scene.
PermutoSDF Fast Multi-View Reconstruction with Implicit Surfaces using Permutohedral Lattices (Rosu and Behnke, 2023) [https://github.com/RaduAlexandru/permuto_sdf]	This is a novel neural surface reconstruction methodology that is able to create novel views and geometry from multi-view images of a subject by utilising a different volumetric representation of the scene. Instead of using a grid of cubes similar to a plethora of voxel-based approaches, it creates a lattice of special geometric structure called a *permutohedron*. This special geometric partitioning of space is used to hold SDF and colour information of a scene, encoded into hashes with the help of neural networks. This approach cannot capture complex light interactions like neural radiance fields do. However, it is able to encode the 3D information of a scene in a matter of minutes and render it at interactive rates using consumer gaming-grade hardware.
Gaussian Surfels (Dai et al., 2024) [https://turandai.github.io/projects/gaussian_surfels/]	This approach aims to leverage the flexible optimisation capabilities of 3D *Gaussian points* for improved surface reconstruction quality, using a novel point-based representation called *Gaussian surfels*. This technique is able to demonstrate a superior surface reconstruction and neural volume rendering for a given set of images.
NeuralWarp Improving neural implicit surfaces geometry with patch warping (Darmon et al., 2022) [https://imagine.enpc.fr/~darmonf/NeuralWarp/]	This is a neural surface reconstruction method that tries to improve reconstruction results using a method called patch warping. This method warps local patches onto the surface during training. This warping aligns patches better with the underlying geometry, allowing the neural network to learn more precise surface details.
BakedSDF Meshing Neural SDFs for Real-Time View Synthesis (Yariv et al., 2023) [https://bakedsdf.github.io]	As the name denotes, *BakedSDF* tries to exploit the technique of *baking* in order to speed novel view synthesis and photorealistic rendering of SDFs. Baking is a method well known from games pre-calculating the lighting of a scene and storing the results on the surface of the 3D models, using a bitmap image. BakedSDF uses a similar approach, by baking the view-dependent appearance of the scene onto the surface of a high-quality 3D model with the help of spherical Gaussians. That 3D model is generated using SDF, calculated by a neural network, and is stored in a volumetric representation. This allows BakeSDF to provide both detailed polygonal models and novel view synthesis, with photorealistic rendering and complex light interactions, from a set of pre-aligned images with known position and orientation. Rendering of BakedSDF is very efficient and can provide photorealistic novel view synthesis to low-end hardware like laptops and mobile phones.
MonoSDF Exploring Monocular Geometric Cues for Neural Implicit Surface Reconstruction (Yu et al., 2022) https://niujinshuchong.github.io/monosdf	This is a method that can generate 3D geometry from sparse image data by incorporating additional geometric clues from monocular (single-image) depth and surface normal predictions. These predictions help guide the reconstruction process by providing more structure-aware information. Using

	these monocular cues significantly boosts the quality and speed of neural implicit reconstructions across various scenarios that range from small single objects to large multi-object scenes, regardless of the specific neural surface representations used. This method is able to reconstruct accurate geometry using very few images showing the same scene, even when these are not in close proximity. This is a really difficult case that most Structure from Motion methods cannot cope with.

Annex 2

a) Monocular Depth Estimation

Technology	Description
MiDaS Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer (Ranftl et al., 2020) [https://github.com/isl-org/MiDaS]	Midas is a neural network approach for predicting depth information from a single image. This method can generalise quite well due to the diversity of the datasets being used for the training of the neural network, enabling the production of relative depth maps that capture the 3D layout of a scene effectively, handling challenges like occlusions or complex environments. The model performs well in zero-shot scenarios, meaning that it can generalise to new, unseen images without needing retraining.
Depth-Anythin-V2 (Yang et al., 2024) [https://github.com/DepthAnything/Depth-Anything-V2]	The second version of Depth-Anything is based on the previous work of Depth-Anything, which was inspired by MiDaS (Ranftl et al., 2020). It outperforms the previous version in fine-grained details and robustness. These improvements were achieved not only by improving the network architecture but also by enhancing its training process with the use of vast amounts of synthetic data along with existing real-life datasets. This helped to improve both generalisation and finer detail resolution, while providing the ability to extract metric information out of a single image. Compared to other state-of-the-art methods that are based on more complex stable diffusion neural network models, this method is an order of magnitude faster, while presenting more accurate results.
Marigold Affordable Adaptation of Diffusion-Based Image Generators for Image Analysis (Ke et al., 2025) [https://github.com/prs-eth/Marigold]	This is a novel technique that makes use of Diffusion-Based neural network Image Generators for Monocular Depth Estimation. It is capable of generating both distance (depth) and orientation (normal) information for every pixel of the image, even for unseen content during the training process. However, extracting accurate metric information from a depth image is neither straightforward nor reliable.
DPT Dense Prediction Transformers	DPT is a neural network architecture that can predict depth from a single image. It differentiates from other methods since

(Ranftl et al., 2021) [https://github.com/isl-org/DPT]	it replaces traditional convolutional network backbones with Vision Transformer networks (ViT), enabling it to generalise better with cases that were left out from training. It is able to achieve better density and operate on images with much higher resolution.
ZoeDepth Zero-shot Transfer by Combining Relative and Metric Depth (Bhat et al., 2023) [https://github.com/isl-org/ZoeDepth]	This is another AI-based depth estimation method that can work with many general cases, while at the same time being able to provide metric information. By utilising a small, smart component called the metric bins module, ZoeDepth can estimate depth accurately for different types of images without relying on similar datasets during training. It can resolve fine details and also provide accurate and stable metric information from depth maps.
DepthFM Fast Monocular Depth Estimation with Flow Matching (Gui et al., 2025) [https://depthfm.github.io]	This is a fast, versatile generative model for monocular depth estimation, delivering state-of-the-art results faster than a lot of other methods, while not sacrificing depth accuracy and details.
NeW CRFs (Yuan et al., 2022) [https://github.com/aliyun/NeWCRFs]	NeW CRFs is a neural network approach for depth map generation from a single image. This method tries to remove the complexity of the neural network that is used for guessing the depth of the given image, by utilising a technique called Conditional Random Fields (CRFs), which helps make better sense of how different parts of the image relate to each other. However, applying CRFs to the whole image at once is not efficient. This work optimises the process by dividing the image into smaller sections where CRFs are applied in parallel, taking advantage of modern hardware and making the process faster and more practical. Also, they propose a smart attention system that is based on a Vision Transformer network, in order to detect object relationships in the image, helping to improve the resulting depth, outperforming many other techniques, and showing impressive results even when applied on panoramic pictures.
CADepth-Net (Yan et al., 2021) [https://github.com/kamiLight/CADepth-master]	This work proposes a smart attention system in order to estimate a detailed depth map from the whole scene captured in an image. This depth map is then refined with the inference of finer details appearing in the image, like smaller objects, corners etc. This two-steps method enables long range depth estimation, while maintaining fine grained details on the foreground.
Kick Back & Relax++ (Spencer et al., 2024) [https://github.com/jspenmar/slowtv_monodepth]	This is an AI method for estimating depth from a single camera, based on a modern transformer-based neural network architecture and working without needing labeled data for training. The authors introduce two new datasets, SlowTV and CribsTV, made up of videos from YouTube and showing a variety of environments. They use these datasets to train a model that can estimate depth in new, unseen environments without any extra training. This neural network model performs better than current methods that don't require supervision, as well as some of the best supervised models. To improve the model's ability to generalise to different situations, the authors use a few innovative techniques like

	learning camera settings, applying stronger data augmentation, mixing up training frames and flexible motion tracking.
Metric3Dv2 (Hu et al., 2024) [https://github.com/yvanyin/metric3d]	This method proposes a versatile geometric foundation model in order to estimate both depth and normal of each pixel appearing in a single image, which is crucial for creating 3D mesh models of objects for real world applications. The advantage of this method is that they trained their neural network on different camera models on a large dataset, in order to distil diverse data knowledge from metric depth. As a result, this method enables accurate recovery of metric 3D structures on randomly collected internet images, which is important for plausible single-image metrology, meaning that we can use this method to measure the distance between objects in a 3D space from a single image.
VGGT Visual Geometry Grounded Transformer (Wang et al., 2025) [https://vgg-t.github.io/]	This is an AI approach that detects camera parameters and extracts depth and 3D point cloud data from a single image, as well as multiple images showing the same scene. It is based on a modern transformer-based feed-forward neural network that can very fast detect accurate camera parameters and infer the depth information from images.
Diffusion Models for Monocular Depth Estimation (Tosi et al., 2024) [https://diffusion4robustdepth.github.io/]	This is a method to estimate depth from a single image, especially in complex and challenging scenarios, like transparent and reflective objects. It is able to enhance the accuracy of depth estimation through iterative training using a diffusion neural network model, ensuring robust performance across diverse conditions. The results demonstrate significant improvements over existing techniques, making this one a promising solution for applications requiring precise single-image depth estimation in complex scenes.

b) Monocular single image 3D model reconstruction

Technology	Description
Zero123 Zero-shot One Image to 3D Object. (Liu et al., 2023b) [https://github.com/cvlab-columbia/zero123]	This is a method that can generate a textured 3D model using just a single image. It works by exploiting a huge diffusion-based neural network model that can synthesise new views from a single input image. Then these images are used to construct a volumetric radiance field, which can be triangulated in order to extract the 3D polygonal mesh. Despite the use of a synthetic 3D dataset for training, this method can generalise well and cope with real-life cases as well. Nevertheless, this technique struggles with images featuring complex and natural backgrounds, which need to be removed in order to get a decent result.
DreamGaussian Generative Gaussian Splatting for Efficient 3D Content Creation. (Tang et al., 2023) [https://dreamgaussian.github.io]	DreamGaussian is a groundbreaking 3D content generation framework that strikes the perfect balance between speed and quality, since it is able to produce high-quality results in just 2 minutes, from a single image input. Unlike conventional methods, this approach offers a more efficient alternative, by using Gaussian Splatting (GS). This method proposes the

	densification of the 3D Gaussians using a diffusion-based neural network that can generate the missing information, based on prior knowledge. Afterwards, the resulting 3D GS model is used for the final 3D mesh extraction, which includes baked image textures, using advanced techniques.
Edify 3D Scalable High-Quality 3D Asset Generation (Bala et al., 2024) [https://arxiv.org/abs/2411.07135]	This is an advanced solution that is designed for high-quality 3D asset generation using just a single image or a text prompt. By leveraging a Visual Transformers neural network architecture, this method is able to reconstruct both detailed geometry with clean topology and photorealistic materials with high-resolution textures, even for unseen surfaces, based on prior knowledge. Nevertheless, the use of transformer neural network architecture helps the method to generalise well with unseen images.
One-2-3-45++ Fast Single Image to 3D Objects with Consistent Multi-View Generation and 3D Diffusion (Liu et al., 2024) [https://github.com/SUDO-AI-3D/One2345plus]	One-2-3-45++ is a method to turn any single photo into an accurate and detailed 3D model within just about one minute. Similar to other methods, this approach starts by fine-tuning a system that can generate consistent views of the same object from different angles using only a single image, Then it elevates this information into a full 3D model, using a Signed Distance Function Field. This method not only creates high-quality and varied 3D models with texture, but also ensures they closely resemble the original photo used as input. This makes One-2-3-45++ incredibly useful for anyone needing to quickly create realistic digital versions of real-world objects.
ImageDream Image-Prompt Multi-view Diffusion for 3D Generation (Wang and Shi, 2023) [https://github.com/bytedance/ImageDream]	ImageDream is a novel 3D object generation method that uses image-prompt multi-view diffusion. It excels in generating high-quality 3D models compared to other state-of-the-art image-conditioned efforts. It uses a canonical camera coordination and multi-level image-prompt controllers to enhance control and address geometric inaccuracies. However, future improvements could focus on further reducing texture blurriness in the generated models.
Magic123 One Image to High-Quality 3D Object Generation Using Both 2D and 3D Diffusion Priors (Qian et al., 2023) [https://github.com/guochengqian/Magic123]	This method can create detailed 3D models with textures from just a single photo. It is a two-step process and starts by making a rough 3D shape of the object, which then refines in order to add finer details and realistic textures. This method uses a smart balance of different techniques to make sure the 3D model looks both creative and accurate. It also includes ways to keep the model consistent when viewed from different angles and to avoid errors in the shape. Tests on both computer-generated and real photos show that Magic123 produces much better 3D models than previous methods, making it a strong tool for turning ordinary pictures into lifelike 3D objects.
CAD Photorealistic 3D Generation via Adversarial Distillation (Wan et al., 2024) [http://raywzy.com/CAD/]	This method supports a variety of 3D tasks, including reconstructing objects from a single view that can be optionally strengthened using a text prompt. It is capable of producing a wide range of diverse 3D models. Tests show this approach outperforms older methods, offering higher quality and richer details in generic 3D content generation. However, the method is tested at the moment on single object extraction

	that is pictured in a clean image without a background, meaning that the user should somehow isolate the subject in order to get a good result.
Dreamcraft3D++ Efficient Hierarchical 3D Generation with Multi-Plane Reconstruction Model (Sun et al., 2024) [https://dreamcraft3dplus.github.io/]	DreamCraft3D++ is an advanced method for 3D asset generation from a single input image. With that image as a guide, an image diffusion neural network creates multiple views of the depicted object along with the same views displaying the normal. Then, with that picture, the AI is able to generate 3D mesh models with textures in a matter of minutes rather than hours. As is true with other similar methods, DreamCraft3D++ requires a clean background.
DVR Differentiable Volumetric Rendering (Niemeyer et al., 2020) [https://is.mpg.de/avg/publications/niemeyer2020cvpr]	This method is able to generate a 3D textured watertight object from one or multiple images depicting that object in a clean background. It does so without the need for supervision. The proposed neural network architecture can learn 3D shapes from plain images, without the need to train on existing 3D datasets. This provides great generalisation while keeping computation requirements relatively low in comparison to more advanced training pipelines.
Shape-E (Jun and Nichol, 2023) [https://github.com/openai/shap-e]	This AI method is able to generate a neural radiance field and, furthermore, generate a textured 3D mesh by using that field. Shape-E can either work by a text prompt, that is used to generate an image using an image diffusion neural network, or by feeding a clean image of an object that lacks background information. The proposed method presents a significant advancement in the efficiency and flexibility of AI-driven 3D asset generation, bridging the gap between quality, speed, and usability previously unaddressed by other 3D generative models.
Point-E (Nichol et al., 2022) [https://openai.com/index/point-e/]	This is a method for text and image conditional 3D object generation that is dramatically faster than previous approaches. The process works either by creating a synthetic image from the text prompt using a diffusion model or by providing a picture of an object with removed background. Then, Point-E proceeds by generating a 3D point cloud conditioned on the given or the generated image, using another diffusion model. While sample quality is not yet on par with the very best methods, this approach is one to two orders of magnitude faster, making it a highly practical alternative for applications where speed is critical.
DISN Deep Implicit Surface Network (Xu et al., 2019) [https://github.com/laughtervv/DISN]	DISN is a neural network designed to reconstruct high-quality, detailed 3D objects from a single 2D image by predicting the underlying signed distance fields (SDFs). Unlike previous methods, DISN combines both global features from the entire image and local features extracted from the area where each 3D point projects onto the image, enabling it to capture fine structural details such as holes and thin parts. This dual-feature strategy allows DISN to deliver state-of-the-art single-view 3D reconstructions that retain intricate details and work effectively on both synthetic and real images
Hunyuan3D 2.5 Towards High-Fidelity 3D Assets	This is a powerful suite of AI-driven 3D diffusion neural network models that are designed to generate highly detailed

Generation with Ultimate Details (Lai et al., 2025) [https://github.com/Tencent-Hunyuan/Hunyuan3D-2]	and realistic textured 3D assets from a single or multiple images. It is based on a two-stage pipeline that enables the generation of sharp, intricate 3D geometry that closely follows its input while maintaining clean and smooth mesh surfaces. This method is able to generate textures for physically-based rendering (PBR) using a novel multi-view approach, resulting in materials that look far more photorealistic. The system achieves fast model generation speeds, improved mesh topology and enhanced stability for complex models, narrowing the gap between AI-generated and human handcrafted 3D assets. Hunyuan3D 2.5 outperforms prior approaches in both geometric precision and texture fidelity, producing industry-ready assets suitable for applications like VR, games and animation workflows.
TripoSR Fast 3D Object Reconstruction from a Single Image (Tochilkin et al., 2024) [https://github.com/VAST-AI-Research/TripoSR]	This is a neural network-based approach that can reconstruct high-quality textured 3D assets from a single image. It is based on a Vision Transformer network architecture to encode the 3D information in a triplanar NeRF representation, presenting great performance characteristics that enable mesh generation in less than a second when using a workstation-grade GPU accelerator.
InstantMesh (Xu et al., 2024) [https://github.com/TencentARC/InstantMesh]	InstantMesh refers to “Efficient 3D Mesh Generation from a Single Image with Sparse-view Large Reconstruction Models” It is a state-of-the-art method for generating textured 3D assets from a single image in a matter of seconds, without requiring expensive workstation-grade GPU acceleration. It is able to generalise and provide diverse but also high-quality game and VR-ready 3D assets.

c) Monocular single image NeRF generation

Technology	Description
pixelNeRF Neural Radiance Fields from One or Few Images. (Yu et al., 2021b) [https://github.com/sxyu/pixel-nerf]	pixelNeRF is a method that learns to reconstruct the whole 3D structure of a scene from one or just a few images. Instead of starting from scratch for every new scene, like older methods, this one uses knowledge from many previously seen scenes to guess a volumetric NeRF representation of the depicted content. The advantage of this technique is that, during training, it learns to recognise patterns (like cars have wheels, or chairs have legs), by figuring out depth and structure from 2D images acquired from different angles. As a result, this method can generalise quite well and can reconstruct single objects as well a whole scene that contains multiple objects of familiar classes (for example, tables, chairs, cars). However, the output is based on NeRF representation that requires significant computational resources for their rendering, but also cannot be converted to 3D polygonal meshes very reliably.

SinNeRF Single View NeRF (Xu et al., 2022) [https://github.com/VITA-Group/SinNeRF]	This is a novel method based on Vision Transformers neural networks that is designed to train neural radiance fields (NeRFs) for complex scenes using only a single reference image as input, without relying on dense multi-view inputs typically required by traditional NeRF methods. This method works by extracting an initial depth map of the input image and then tries to wrap that depth map and other extracted pseudo labels to novel synthesised views, enforcing multi-view geometric consistency despite the lack of real multiple images. As a result, the applied methodology is able to perform photo-realistic novel-view synthesis, even without pre-training on multi-view datasets.
DietNeRF Putting NeRF on a Diet: Semantically Consistent Few-Shot View Synthesis. (Jain et al., 2021) [https://github.com/codestella/putting-nerf-on-a-diet]	DietNeRF is a 3D neural scene representation that estimate and produce novel views with as few as one observed image. When pre-trained on a multi-view dataset, it is able to produce plausible completions of completely unobserved regions. DietNeRF introduced a new type of guidance during training called a semantic consistency loss. It is trained primarily to accurately recreate the scene from the input viewpoint, and then keep the overall meaning and important features consistent when viewed from different random angles. This helps the model generate realistic images even from new angles it hasn't seen before. However, its single-view performance is inferior to newer methods that are based on transformer neural network architecture like SinNeRF (Xu et al., 2022).
TGS Triplane Meets Gaussian Splatting (Zou et al., 2024) [https://zouzx.github.io/TriplaneGaussian/]	TGS is a really efficient approach that uses two transformer-based neural networks that work together in order to reconstruct an object from a single image using a hybrid "Triplane Gaussian" representation. The advantage of this method is that it can generate a Gaussian Splatting representation of the object for high-quality rendering. Furthermore, if a 3D mesh model is needed, the Gaussian representation can be exploited, along with the densified sparse point cloud that it uses, in order to reconstruct a 3D polygonal mesh. One big disadvantage of the method is that the unseen side of the object is blurry, since it is challenging for 3D Gaussians to recover missing information.

Annex 3 - MIME Types

Format	Official IANA MIME Type	Common/De facto MIME Type(s)	Notes (Web/API usage)
LAS (LiDAR point cloud)	**application/ vnd.las** (IANA registered).	Often served as generic binary (**application/octet-stream**) if not configured.	LAS is a binary LiDAR point cloud format (ASPRS standard). Browsers don't natively render LAS; web apps (e.g. Potree) fetch it as binary data. No text variant exists (ASCII exports use other formats like XYZ).
LAZ (LAZzip compressed LAS)	**application/ vnd.laszip** (IANA registered).	Typically treated as binary (**application/octet-stream**) in practice (if not using the official type).	LAZ is the lossless compressed form of LAS. It's binary; web viewers use a decoder (e.g. laszip.js) to unpack LAZ in-browser. Servers often default to octet-stream for .laz files due to lack of built-in recognition. No text version (must decompress to LAS for use).
XYZ (Point cloud text file)	No official IANA type.	Usually **text/plain** or **text/csv** (since it's an ASCII list of coordinates).	"XYZ" files list point coordinates in plain text. There is no registered MIME; they are generally handled as simple text files. (Note: .xyz is also used in chemistry with unofficial type **chemical/x-xyz** , unrelated to point clouds.) In web contexts, these are downloaded or parsed as text – binary vs text is not an issue since XYZ is inherently text.
PTS (Leica/ Faro ASCII points)	No official IANA type.	Treated as text (e.g. **text/plain**); sometimes **.pts** is auto-detected as ASCII data	PTS is an ASCII point cloud exchange format (each line is X Y Z [and optional data]). Common in LiDAR software (e.g. Cyclone) for exporting merged scans. Browsers/APIs typically handle it as plain text. (Very large PTS files may be zipped for transfer, then served as e.g. **application/zip**)
PTX (Leica Cyclone format)	**No official IANA type.**	Treated as text; e.g. often handled as **text/plain** for .ptx files.	PTX is an ASCII structured format from Leica Cyclone . It includes header info and point lists in text form. No formal MIME exists; it's usually imported via software, not directly rendered on web. Being text, it should use a text- based content type if served via HTTP.
E57 (ASTM E57)	**model/e57** (IANA registered)	Occasionally seen as **model/e57** (now official); older practice was generic binary (**application/ octet-stream**) before registration.	E57 is a **binary+XML** container format for 3D imaging (point clouds + images), standardized by ASTM. It has an official model/* MIME. Web browsers don't natively support E57; files are typically handled in point-cloud software or libraries. No separate text version (format contains an

			XML metadata section but file is overall binary).
DXF (AutoCAD Drawing eXchange Format)	**image/vnd.dxf** (registered vendor type)	Often served as **application/dxf** (unofficial) or sometimes as legacy “image/x-dwg” for older plugins.	A CAD vector format not natively rendered by browsers. Historically proposed under the “image” category for web plugins. If hosting DXF files, configure the server with the official type for consistency; otherwise, browsers will download it as a generic file.
OBJ (Wavefront .obj)	**model/obj** (standards-tree type registered in 2020)	Before 2020, usually treated as plain text (default **text/plain**) or given a custom type like **application/ object**. Some systems also used personal-tree types *(e.g. text/ prs.wavefront-obj).*	A widely-used plain-text 3D geometry format. Three.js and similar libraries load OBJ files via AJAX/text, so the MIME type was often ignored. With the official model/obj now available, servers can explicitly identify OBJ files. Material files (.mtl) use **model/mtl.**
FBX (Autodesk Filmbox)	*No official IANA type.*	Commonly defaults to **application/octet**-stream (binary data). Some applications use **application/fbx** or **model/x-fbx** as a custom type, but these are not standardized.	A proprietary 3D exchange format. Browsers do not recognize .fbx types (e.g. the file input type will be empty). If delivering .fbx from a server, using **application/octet-stream** is typical (since no specific type exists). Web engines like three.js provide loaders for FBX, but they rely on parsing the file content (or file extension) rather than MIME type.
DAE (Collada)	**model/ vnd.collada+xml** (XML-based COLLADA, registered 2011)	Previously often served as generic XML: **application/xml** or **text/xml**. (Some systems used model/collada+xml before registration.)	An XML format for 3D assets (ISO PAS 17506). WebGL libraries (e.g. three.js ColladaLoader) parse the XML document; thus servers should send the official type (or at least an XML type so that clients know it’s XML).
PLY (Stanford Polygon File)	*No official MIME type.*	ASCII PLY files are usually detected as **text/plain** by systems. Binary .ply files may fall back to **application/octet-stream**. Some use an unofficial **application/ply** or **model/x-ply** internally.	A 3D point cloud/mesh format (ASCII or binary). Not supported natively in browsers; used via libraries. When serving PLY, ensure the correct mode in the client (binary vs text) is specified. The MIME type itself is typically not checked by loaders – many workflows simply rely on file extension or user selection.
STL (Stereolithography)	**model/stl** (registered March 2018)	Historically used **application/sla** (unofficial “stereolithography”). Also sometimes misidentified as **application/vnd.ms-pki .stl** (which actually refers to a certificate trust list, not 3D model).	A simple mesh surface format widely used in 3D printing. As of 2018, *model/stl* is the proper type. Before that, “application/sla” was a de facto standard and may still appear in older server configs. Modern OS tools (e.g. Windows 3D Viewer, macOS Preview) can open STL, but on the web it’s

			typically downloaded or handled by script.
IGES (Initial Graphics Exchange Spec, *.igs)	**model/iges** (registered, replaces old application type)	Earlier standard was **application/iges**. Many systems still recognize *application/iges* for *.igs files.	A classic CAD exchange format (ANSI/ISO Neutral). MIME registration moved from application to model in the late 1990s. Web use is rare (usually downloaded or converted server-side). If served, using the model/iges type is recommended for consistency.
STEP (ISO 10303 STEP, *.stp)	**model/step** (and related: **model/step+xml**, **model/step+zip**, etc., registered 2021)	Older practice was **application/step** for .stp files. Some implementations still use unofficial types like **model/x-step**.	A complex CAD 3D exchange format. The IANA-registered suite covers plain text Part 21 (.stp) and XML encodings (STEP-XML) and zipped variants. Browsers won't render STEP data; it's typically downloaded or processed with CAD plugins. Ensure the MIME matches the content (e.g. *model/step+xml* for XML-based.stpx) if delivering via API.
VRML (Virtual Reality Modeling Language, *.wrl)	**model/vrml** (official, per RFC 2077)	Legacy browsers/ plugins **used x-world/ x-vrml** (an experimental type). Also seen: **application/x-world** (variant of the same).	An older 3D web format (VRML97). Modern browsers have dropped support for VRML, but it can be viewed via standalone plugins. For historical content, servers should send *model/vrml*, though many VRML files still carry the old x- world/x-vrml for compatibility.
X3D (Extensible 3D, XML-based VRML successor)	**model/x3d+xml** (classic XML encoding), **model/ x3d+fastinfoset** (binary encoding), **model/x3d-vrml** (VRML-style encoding).	Prior to registration (circa 2013), X3D files might be served as generic XML (**application/xml**) or not recognized at all. Some systems incorrectly use **model/x3d** without a suffix (non-standard).	An XML-based 3D scene format (X3D is the ISO successor to VRML). It defines three encodings with distinct MIME types. X3D viewers (or embedded X3DOM in HTML5) expect the correct type – e.g. an *.x3d* file as *model/x3d+xml*. Servers and Apache configs were updated to include these types. Content negotiation by "+xml" allows XML tools to handle .x3d where needed.

glTF (GL Transmission Format, *.gltf JSON* and *.glb binary*)	**model/gltf+json** (for .gltf, JSON scene); model/ gltf-binary (for .glb, binary).	In early adoption some servers treated .gltf as **application/json** (due to JSON content) or defaulted to octet-stream. However official types were and tools have widely adopted them	A modern efficient 3D asset format by Khronos (web- friendly). Both MIME types are officially registered and should be used accordingly. WebGL frameworks (three.js, Babylon.js, etc.) and Web APIs (like WebGL and WebXR pipelines) natively support glTF. Browsers don't render glTF files by themselves, but with the correct MIME, engines can (E.g. a <model> HTML element or <a-asset> in A-Frame would rely on these types.) Ensure .gltf is served with *model/gltf+json* so that it's recognized as JSON data, and .glb with *model/gltf-binary* for proper binary handling.
USD (Universal Scene Description)	for specific subtypes: **model/ vnd.usda** for ASCII .usda files; **model/vnd.usdz+zip** for USDZ package files. *(No single generic type for binary .usd/.usdc; these often use the .usd extension.)*	Before registration, USD assets were often served as **application/octet-stream** (or even **application/zip** for USDZ). Now **model/vnd.usdz+zip** is used for USDZ on the web (e.g. Safari AR Quick Look) and **model/vnd.usda** for USDA text layers.	USD can be text or binary. **.usda** are plain text (UTF-8) scene files, while **.usdc** ("crate") files are binary; **.usd** extension may be either (detected by software). In practice, web/AR platforms primarily use **USDZ**, which is a zipped archive of USD data and resources. Apple's AR Quick Look requires .usdz files to be served with the correct MIME **model/vnd.usdz+zip** for in-browser AR viewing. Standard WebGL/three.js does not natively support USD(Z) yet – plugins or converters are used, so the MIME mainly matters for download/launch behavior.