PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science
https://doi.org/10.1007/s41064-026-00412-y

RESEARCH

Recent Advances in Image-Based 3D Reconstruction: a Photogrammetric Perspective on Conventional and Learning-Based Techniques

Xin Wang[1] · Tengfei Wang[1] · Markus Hillemann[2] · Yifei Yu[1] · Zhe Shen[1] · Zongqian Zhan[1] · Rongjun Qin[3] · Markus Ulrich[2]

Received: 7 January 2026 / Accepted: 23 July 2026

Abstract

Image-based 3D reconstruction is vital in many applications, such as digital twins, smart cities, machine vision, and autonomous driving. In recent years, it has undergone a paradigm shift, propelled by advancements in both conventional photogrammetry and deep learning. This review provides a comprehensive photogrammetric perspective on both conventional and learning-based techniques, a viewpoint that prioritizes geometric fidelity, robustness, handling of uncertainty, and suitability for real-world applications. We first systematically revisit the fundamentals of traditional pipelines: Structure from Motion (SfM), Multi-View Stereo (MVS), and surface reconstruction. The review then details recent progress in conventional methods, highlighting innovations in scalable and efficient SfM, specialized camera models for MVS, and robust surface reconstruction algorithms. Subsequently, we explore the transformative evolution brought by learning-based techniques, including deep SfM, learning-based MVS, differentiable rendering-based scene representation methods (NeRF, 3DGS), groundbreaking feed-forward 3D reconstruction models (e.g., DUSt3R, VGGT), and surface reconstruction including explicit and implicit methods. Emphasis is placed on evaluating whether learning-based approaches genuinely meet photogrammetric requirements such as metric accuracy and reliability, rather than optimizing solely for visual realism. Furthermore, we conclude by identifying key challenges and research frontiers including generalization across domains, scalability to high-resolution imagery, real-time performance, and uncertainty quantification. By bridging the gap between classical photogrammetry and data-driven 3D vision, this work aims to guide future research toward robust, accurate, and certifiable 3D reconstruction systems suitable for engineering, industrial, and geospatial applications.

Keywords 3D Reconstruction · Photogrammetry · Machine Vision · Deep Learning · Structure from Motion · Multi-View Stereo · Surface Reconstruction

1 Introduction

Image-based three-dimensional (3D) reconstruction, the process of inferring the geometry and appearance of a scene from a collection of two-dimensional (2D) images, has long served as a foundational pillar across a diverse spectrum of scientific and industrial domains, including computer vision & graphics, photogrammetry, remote sensing, geoinformation science, machine vision, digital twins, smart city modeling, and autonomous driving.

Historically, conventional photogrammetry, grounded in projective geometry and optimization theory, provides well-established and interpretable pipelines, which are typically decomposed into Structure from Motion (SfM), Multi-View Stereo (MVS), and surface reconstruction. SfM serves as the initial sparse reconstruction phase, recovering camera poses and a 3D point cloud from unordered image collections through processes involving feature detection (e.g., SIFT (Lowe 2004)), matching, and robust optimization via bundle adjustment. Building upon the camera geometry pro-

✉ Markus Ulrich
markus.ulrich@kit.edu

[1] School of Geodesy and Geomatics, Wuhan University, Wuhan, China

[2] Institute of Photogrammetry and Remote Sensing, Karlsruhe Institute of Technology, Karlsruhe, Germany

[3] Geospatial Data Analytics Lab, The Ohio State University, Columbus, OH, USA

Published online: 10 August 2026

vided by SfM, MVS techniques aim to compute dense per-pixel correspondences, resulting in detailed depth maps or dense point clouds. Finally, surface reconstruction methods transform these discrete 3D samples—whether sparse from SfM or dense from MVS—into continuous, coherent surface representations, typically as triangle meshes, using either explicit algorithms like Delaunay triangulation (De Floriani 1987) or implicit functions like Poisson reconstruction (Kazhdan et al. 2006).

Over the past decade, conventional methods have continued to evolve, particularly to address large-scale datasets and efficiency bottlenecks. For incremental SfM, systems such as COLMAP (Schönberger and Frahm 2016) have become widely employed due to their reliable performance and end-to-end integration of feature processing, sparse reconstruction, and dense modeling. Meanwhile, global SfM has gained renewed attention as improved view-graph optimization and pose averaging strategies have narrowed the accuracy gap with incremental baselines while offering substantial speedups. Recent examples include advances (Pan et al. 2025) that revisit the global pipeline with better calibration and positioning in the view graph, achieving order-of-magnitude improvements in runtime for large collections. After SfM, MVS estimates dense depth and fuses it into a coherent 3D representation. Classical MVS relies on multi-view photometric and geometric consistency, along with robust handling of occlusions, view selection, and regularization to balance smoothness and edge preservation. Surface reconstruction then converts discrete samples (depth maps or point clouds) into a continuous surface representation using either explicit methods (e.g., mesh-centric approaches) or implicit methods (e.g., fields whose level sets define surfaces). Importantly, the conventional pipeline is not merely a sequence of engineering blocks: it embodies a philosophy of explicit geometry, interpretable assumptions, and statistical treatment of uncertainty (in many cases), which is central to high-precision mapping and surveying.

Concurrently, catalyzed by the rapid advancement of deep learning, a second, fast-moving axis of innovation has been witnessed. Data-driven approaches have increasingly challenged, and in certain contexts, surpassed traditional pipelines in terms of reconstruction completeness, robustness under weak texture, and inference speed (Yao et al. 2018; Kerbl et al. 2023; Wang et al. 2025a). Notably, learning-based SfM frameworks (e.g., ESfM (Moran et al. 2021), VGGSfM (Wang et al. 2024c), MVS architectures (e.g., MVSNet (Yao et al. 2018), R-MVSNet (Yao et al. 2019), and differentiable rendering-based scene representation methods (e.g., NeRF (Mildenhall et al. 2021)), 3D Gaussian Splatting (Kerbl et al. 2023) have redefined what is possible in end-to-end 3D vision. Most recently, feed-forward reconstruction models such as DUSt3R (Wang et al. 2024d), MASt3R (Leroy et al. 2024), and VGGT (Wang et al. 2025a) have emerged as a new cornerstone, capable of jointly estimating camera poses and scene geometry from unconstrained image sets within a single forward pass—bypassing iterative optimization altogether.

Yet, despite these impressive capabilities, learning-based methods raise critical concerns when evaluated by photogrammetric criteria and applications. First, metric accuracy and reliability remain nontrivial: methods optimized for visual realism may not preserve strict geometric correctness, especially under domain shift or biased training data. Furthermore, learning-based 3D reconstruction methods do not rely solely on the observed images of the object to be reconstructed, but they also leverage priors from previously seen training data, which can lead to hallucinated depth values. This effect is particularly pronounced in feed-forward generalizable methods (see Sect. 4.2.3), making it critical to consider for measurement tasks and safety-critical applications. Second, scalability is constrained by the data hunger and memory footprint of modern deep architectures, particularly transformer-based designs, when facing megapixel imagery and very large image collections. Third, "real-time" in photogrammetry often implies not only low latency but also consistent robustness in online measurement and dynamic feedback scenarios, a higher bar than many perception benchmarks. Finally, uncertainty quantification is a central bottleneck: while conventional pipelines can derive first-order statistics analytically, deep confidence scores are frequently over-confident, motivating research into reliable uncertainty estimation for learning-based 3D reconstruction.

Motivated by this landscape, this review provides a comprehensive photogrammetric perspective on both recent conventional and learning-based techniques. It can serve as a systematic survey of recent image-based 3D reconstruction techniques, specifically targeting photogrammetrists, geospatial engineers, industrial metrology specialists, and robotics researchers who seek to evaluate whether learning-based methods can replace or complement traditional pipelines in relevant applications. We revisit the fundamentals of SfM, MVS, and surface reconstruction (see Sect. 2), summarize recent progress in conventional methods (e.g., scalable SfM and specialized camera models, see Sect. 3), and then examine the transformative evolution of learning-based approaches, including deep SfM, learning-based MVS, differentiable rendering-based scene representation methods (NeRF, 3DGS), and emerging feed-forward 3D reconstruction models (see Sect. 4). Throughout, the emphasis is on whether these methods meet photogrammetric requirements—metric fidelity, robustness, and uncertainty handling—rather than optimizing solely for visual quality. Finally, we distill key research frontiers in generalization, scalability, real-time performance, and uncertainty

quantification, aiming to guide future work toward robust and certifiable 3D reconstruction systems for real-world photogrammetric applications (see Sect. 5).

2 Fundamentals of Photogrammetric Image-Based 3D Reconstruction

2.1 Structure from Motion

In the following, the term camera pose is used interchangeably with (exterior) camera orientation, and camera parameters are considered equivalent to interior orientation.

The task of SfM is to recover the 3D structure of a scene and the camera poses from a collection of input images. Furthermore, if the interior orientation of the cameras is unknown, it is estimated as well. The conventional SfM pipeline usually starts with the detection and description of point features in the input images, e.g., by applying the Scale-Invariant Feature Transform (SIFT) (Lowe 2004). They serve as the basis for a subsequent matching step between image pairs. The matching step is often split into an image retrieval task to efficiently find candidates of overlapping image pairs and a computationally more expensive feature correspondence search for these candidates. The feature correspondence search, which is typically embedded into a robust framework like RANSAC (Fischler and Bolles 1987), yields the geometry of each overlapping pair, which is either represented by a homography or a fundamental matrix depending on the type of scene and the type of camera motion. A homography is estimated for planar scenes that are acquired under a general camera motion as well as for general scenes that are acquired under a pure camera rotation. In contrast, a fundamental matrix is estimated for general scenes under general camera motion. In the calibrated case, i.e., if the interior orientation of the camera(s) is known, instead of the fundamental matrix, the essential matrix is directly estimated. The pairwise geometry is obtained by decomposing the respective matrix into a rotation and a translation component. Based on the pairwise geometries, typically a view graph is constructed, where nodes represent images and edges encode relative poses. This graph serves as the basis for the subsequent reconstruction strategy, which can follow either an incremental or a global approach.

2.1.1 Incremental Methods

In incremental SfM, the reconstruction is initialized from a carefully selected image pair or triplet with sufficient baseline and feature overlap. The model (comprising both camera poses and triangulated 3D points of the matching image features) is then expanded sequentially by registering additional images. For each newly added image, its camera pose is estimated relative to the existing 3D model using 2D-to-3D correspondences. New 3D points are triangulated from matched image features and the entire model (i.e., camera poses and 3D points) is refined using bundle adjustment. This optimization step is performed repeatedly throughout the reconstruction and jointly minimizes the reprojection error across all observations (i.e., image features). While this interleaved process ensures high accuracy and robustness, especially in unordered image sets with unknown overlap, it is computationally expensive and less scalable due to the frequent re-optimization via bundle adjustment of the growing model.

2.1.2 Global Methods

In contrast, global SfM aims to first estimate all camera poses simultaneously by leveraging the full set of pairwise geometries encoded in the view graph. In this step, typically no 3D points are triangulated. To make the problem tractable, the process is often decomposed into a rotation and a translation averaging step. In the rotation averaging step, absolute camera rotations are computed from pairwise relative rotations by using a robust optimization technique that minimizes the discrepancy between the estimated absolute rotations and all pairwise relative rotations. Once the global rotations are established, they are factored out from the camera poses and translation averaging is performed to robustly estimate the maximally consistent absolute camera positions from all pairwise relative translations. During this step, the scale ambiguity, which is inherently present in the SfM task, must be carefully taken into account. The thus obtained global camera pose estimates are then used to triangulate 3D points across the entire dataset. A final global bundle adjustment is applied to jointly refine all camera parameters and the 3D structure. Global SfM is significantly more scalable and computationally efficient than incremental methods, making it well-suited for large-scale or densely-connected image collections. However, it is generally more sensitive to outliers and requires high-quality feature matches and geometric consistency in the view graph.

2.2 Multi-View Stereo

The goal of MVS is to recover a dense 3D structure of the scene from the images, given the camera poses and the interior orientation, which are typically derived by a preceding SfM or calibration step (Szeliski 2022).

A typical MVS pipeline starts by selecting appropriate subsets of images for reconstruction. This first step often involves either a manual selection for pre-calibrated fixed camera setups (e.g., a machine vision setup for industrial

quality inspection) or view clustering or neighborhood selection (e.g., after applying SfM) to ensure sufficient overlap and baseline between images while maintaining computational efficiency.

For each selected subset, depth estimation is performed in a second step by exploiting multi-view constraints like geometric and photometric consistency to improve robustness and accuracy. The depth estimation in MVS is commonly formulated as an optimization problem that minimizes a photo-consistency measure across all views. One line of methods relies on pairwise matching strategies, where disparity maps are computed for selected camera pairs and converted into depth maps. The pairwise depth maps are then fused to a consistent 3D reconstruction (see below). A key advantage of these methods is their high efficiency, which results from the fact that the images can be rectified utilizing epipolar geometry, thereby significantly accelerating the pairwise matching. Since some recent advances relate to these methods, they are explained in more detail in Section 3.2.2.

Next to methods relying on pairwise matching, another line of methods takes more than two views into account when generating depth hypotheses. These methods can be grouped into three main categories: plane sweeping, patch-based methods, and volumetric approaches.

Plane sweeping, as introduced by (Collins 1996), evaluates candidate depth hypotheses by projecting image pixels onto hypothetical planes in 3D space and measuring photo-consistency across views. Patch-based methods, such as PatchMatch Stereo (Bleyer et al. 2011) and Patch-based Multi-View Stereo (Furukawa and Ponce 2010), propagate depth hypotheses locally and refine them iteratively, enabling efficient handling of large image sets and producing accurate depth maps even in complex scenes. Volumetric approaches, such as Space Carving (Kutulakos and Seitz 2000), discretize the scene into a voxel grid and iteratively remove inconsistent voxels based on consistency criteria to produce volumetric occupancy representations.

To handle occlusions and varying visibility, MVS algorithms incorporate strategies such as view selection, confidence weighting, and regularization. Regularization terms enforce smoothness in the reconstructed surface while preserving sharp edges, which is crucial for complex scenes. Some MVS methods also integrate geometric priors from the sparse SfM point cloud to guide depth estimation and reduce ambiguities.

After dense depth maps have been computed for individual views or clusters in the second step, they are fused into a consistent 3D scene representation in a third step. Fusion techniques range from simple point cloud merging to more sophisticated volumetric fusion, which aggregates depth information into a global model while resolving conflicts and removing outliers. The discrete output of MVS serves as input for subsequent surface reconstruction methods.

2.3 Surface Reconstruction

Surface reconstruction aims to transform the discrete 3D samples—typically sparse or dense point clouds or depth maps—generated by SfM and MVS into a continuous, coherent representation of scene geometry. Existing methods can be broadly classified into two paradigms: explicit approaches, which directly reconstruct surface geometry in the form of meshes or other discrete primitives, and implicit methods, which recover a continuous scalar field (e.g., signed distance or occupancy function) whose zero-level set defines the underlying surface.

2.3.1 Explicit Methods

Explicit surface reconstruction methods seek to produce a clear, well-defined geometric representation, most commonly in the form of a triangle mesh, directly from input point clouds or depth maps. Classical techniques often begin with a 3D Delaunay tetrahedralization (De Floriani 1987), from which a surface is extracted by selecting a subset of faces based on visibility, normal consistency, or geometric regularity criteria. Other prominent explicit approaches include region-growing strategies like the Ball Pivoting Algorithm (BPA) (Bernardini et al. 2002), which incrementally constructs a mesh by rolling a virtual ball over point samples to identify valid triangles. Additionally, multi-view visibility-based approaches, such as space carving and visual hull methods (Laurentini 2002), are also widely employed to reconstruct watertight surfaces. These explicit methods are particularly effective in preserving high-frequency geometric details and afford fine control over both mesh resolution and topology. However, they are generally sensitive to noise, outliers, and irregular point sampling densities, which can result in topological artifacts or incomplete surfaces. Consequently, extensive post-processing, such as mesh smoothing, hole filling, and simplification, is often required to generate robust and visually coherent results.

2.3.2 Implicit Methods

Implicit surface reconstruction methods represent a 3D scene by defining a continuous scalar field over space, where the reconstructed surface corresponds to a specific level set—most commonly the zero-level set of this field. Typical implicit representations include Signed Distance fields (SDF) and occupancy or density fields. These functions are usually constructed by interpolating or fitting the field to input data such as oriented point clouds or depth

maps, often under smoothness or regularization constraints. Once the implicit function is established, the next step is to convert it into an explicit surface for visualization or downstream use. This is typically accomplished through volumetric discretization: the 3D domain is partitioned into a regular grid of voxels, and the scalar field is evaluated at each voxel vertex. A triangle mesh is then extracted from this volumetric representation using algorithms such as Marching Cubes (Lorensen and Cline 1998). A prominent example of an implicit technique is Poisson surface reconstruction (Kazhdan et al. 2006), which treats oriented point samples as approximations of the gradient of an underlying indicator function. The surface is recovered by solving a global Poisson equation that best fits this vector field in a least-squares sense.

Implicit methods are widely appreciated for their ability to produce watertight, globally consistent, and smooth surfaces, even in the presence of noisy or incomplete data. They naturally handle complex topologies and topological changes without requiring explicit connectivity assumptions. However, these advantages often come at the cost of fine geometric detail, as strong regularization can lead to oversmoothing. Moreover, implicit approaches tend to be computationally demanding and memory-intensive—particularly when applied to large-scale scenes—due to the need for dense volumetric sampling or global optimization.

3 Recent Advances of Conventional 3D Reconstruction

This section presents recent technical developments in conventional image-based 3D reconstruction, focusing on SfM, MVS, and surface reconstruction.

3.1 Structure from Motion

SfM remains a cornerstone of image-based 3D reconstruction, enabling the recovery of camera poses and sparse scene geometry from unordered image collections. While the fundamental steps of feature detection, matching, pose estimation, triangulation, and bundle adjustment are well established (see Sect. 2.1), recent research has focused on improving robustness, scalability, and efficiency across the three main paradigms: incremental, global, and hybrid approaches. These developments target issues such as error accumulation, outlier sensitivity, and computational bottlenecks in large-scale reconstructions (Agarwal et al. 2009).

3.1.1 Incremental Methods

Incremental pipelines remain widely used due to their robustness and accuracy. They iteratively register images to a growing model while alternating camera pose estimation, triangulation, and bundle adjustment. COLMAP (Schönberger and Frahm 2016) exemplifies the state of the art in incremental SfM, offering a complete pipeline that includes feature extraction, exhaustive and sequential matching, sparse and dense reconstruction, and camera calibration. Its modular design allows users to switch between incremental and global SfM modes, and its GPU-accelerated dense reconstruction makes it suitable for high-resolution datasets. COLMAP's integration of photometric and geometric consistency checks further enhances the reconstruction quality. Built upon earlier tools like Bundler (Snavely 2008) and VisualSfM (Wu 2013), COLMAP has become a standard in photogrammetric workflows due to its consistent performance and open-source accessibility.

A recent advance in incremental SfM is hybrid feature integration: Leveraging image lines in addition to image points strengthens geometric constraints in weakly textured or structured scenes, with uncertainty propagation for 3D lines and improved localization (Liu et al. 2025b). Another innovation is to replace certain steps in the pipeline by deep-learning-based counterparts, e.g., by using deep learning to enhance correspondence search and accelerate optimization. For example, (Liu et al. 2024a) use learned feature matching coupled with faster bundle adjustment and depth-guided densification to improve completeness and efficiency.

3.1.2 Global Methods

Global methods solve camera geometry jointly via rotation and translation averaging followed by a single global bundle adjustment. (Wang et al. 2019a) proposes a time-efficient solution that integrates image retrieval and global pose estimation for large-scale ordered and unordered image sets. In particular, random k-d forests are employed to fast identify overlapping image pairs, global poses are solved by translation averaging and global consistent scale determination. GLOMAP (Pan et al. 2025) revisits the global pipeline with improved view-graph calibration and global positioning, achieving accuracy on par with the incremental baseline COLMAP but with speedups of one to two orders-of-magnitude. After rotation averaging, a joint global 3D point triangulation and camera position estimation is performed using normalized direction differences as an error measure during optimization. (Tao et al. 2025) address a critical gap in global SfM, handling multi-camera rigs common in autonomous vehicles. Their key innovation is a hybrid translation averaging objective that jointly optimizes dis-

tance-based and angle-based residuals. Unlike incremental methods that accumulate drift in long sequences, MGSfM's global formulation reduces pose inconsistency, which is particularly valuable for UAV swarms or vehicle-mounted multi-sensor platforms. However, the method assumes synchronized cameras and known intrinsics, its performance under asynchronous capture or rolling-shutter distortion remains unexplored. For industrial applications requiring submillimeter accuracy, further validation against calibrated ground truth is needed.

To enhance the robustness of Global SfM, some studies have focused on refining the input two-view geometries, specifically, relative rotations and translations. (Shah et al. 2018) formulate view-graph selection as a linear programming optimization problem, employing a minimum cost network-flow framework and investigating various cost models to guide the optimization. Meanwhile, (Wang et al. 2019b) and (Wang and Heipke 2020) address the degenerated relative orientations caused by repetitive structures and very short baselines. They introduce two complementary cost criteria and cast the outlier detection and removal task as a binary linear programming problem. In a related effort, (Xiao et al. 2021) exploit cycle consistency among relative orientations, leveraging loopy belief propagation to identify and eliminate inconsistent (i.e., erroneous) relative orientations. Based on the work of (Barath et al. 2021), which infers unknown two-view geometries from already estimated ones, (Gan et al. 2024b) present an efficient and robust view-graph generation method. A heuristic A* algorithm is applied to find a reliable and visible path between two unknown views, and multiple orthogonal minimum spanning trees are built to obtain redundant yet consistent geometric estimates, thereby improving overall robustness.

3.1.3 Hybrid Methods

Hybrid SfM combines global initialization with incremental refinement. HSfM (Cui et al. 2017) estimates rotations globally and positions incrementally to mitigate drift while retaining robustness. Hence, HSfM inherits robustness from its incremental part and efficiency from its global part. (Wang et al. 2021b) explore the mathematical characteristics of N-View essential matrix (Kasten et al. 2019), and propose hybrid global SfM method by separately solving collinear and non-collinear triplets. Collinear triplets are solved by rotation averaging and translation averaging, whereas the poses of non-collinear triplets can be directly estimated by averaging the corresponding N-View essential matrix. (Liu et al. 2023) apply partition-merge strategies to improve the scalability. (Chen et al. 2020) formulate large-scale reconstruction as a graph problem and apply a divide-and-conquer strategy. Images are clustered into subgraphs, reconstructed locally in parallel, and subsequently merged. Their approach achieves superior scalability and accuracy compared to prior hybrid pipelines. Hybrid SfM approaches that use partitioning, clusters, or subgraphs aim to split the image set into smaller groups for parallel local reconstruction and then merge these partial models. It should be noted that this clustering is primarily a computational strategy to improve scalability. However, it is not a hierarchical pipeline by itself, which can be represented, e.g., by a strict hierarchical tree structure.

3.1.4 Hierarchical Methods

Hierarchical pipelines on the other hand organize the reconstruction into a tree-like hierarchy. Local models are built at lower levels and progressively merged into higher levels, enforcing a structured merging order for global consistency. (Jiang et al. 2022) propose a parallel SfM pipeline tailored for UAV imagery. They reconstruct local clusters independently and merge them into a global model in a structured, multi-level fashion achieving remarkable speedups over traditional incremental SfM. Recent work by (Liang et al. 2023) introduces adaptive track selection and robust submap merging for UAV/oblique imagery, yielding speedups over COLMAP without accuracy loss.

3.1.5 Real-Time Methods

Real-time SfM seeks to combine the robustness of traditional SfM (capable of handling high-resolution images with limited overlap and producing accurate point clouds) with the real-time capability of visual simultaneous localization and mapping (SLAM) systems. Unlike offline SfM pipelines, real-time approaches aim to deliver incremental reconstructions during image acquisition, which is critical for UAV and mobile mapping scenarios. (Zhao et al. 2022b) introduce RTSfM, a SLAM-inspired online SfM pipeline for sequential aerial images with limited overlap. Their method integrates hierarchical feature matching, multi-homography estimation, and graph-based optimization with GPS constraints to achieve real-time performance while maintaining geometric consistency. On-the-Fly SfM (Zhan et al. 2024, 2025) enables real-time incremental reconstruction as images are captured. (Gan et al. 2024a) propose LVG-SfM, further improving the robustness of input view-graph by integrating learning-based correspondence generation (Morelli et al. 2024) and two-view geometry disambiguation—Doppelgangers (Cai et al. 2023). These online SfM approaches involve a fast image retrieval with tailored learning-based features (Hou et al. 2023), a correspondence refinement method using least squares matching, and an efficient local bundle adjustment using hierarchical weights to guarantee a fast yet robust online SfM. In contrast to RTSfM, On-the-Fly SfM does not require spatiotemporal

continuity of the input images nor GPS/IMU data, making it well-suited for unordered image collections across diverse applications.

3.1.6 Pose Accuracy in Metric Reconstruction

Accurate camera poses are not merely an intermediate output, they fundamentally dictate the geometric fidelity of subsequent dense reconstruction and surface modeling. Even sub-pixel reprojection errors in pose estimation can propagate into systematic depth biases during MVS, particularly in narrow-baseline or low-overlap configurations. Recent advances in featuremetric refinement (e.g., Lindenberger et al. (2021)) demonstrate that optimizing alignment directly on image gradients, rather than relying solely on discrete feature keypoints, can significantly reduce pose drift and improve scale consistency. This paradigm has been increasingly adopted by learning-based SfM frameworks, yet a critical gap remains: while deep models excel at initial pose estimation under extreme viewpoint changes, they often lack the iterative, gradient-based refinement steps that classical bundle adjustment provides. For photogrammetric workflows targeting millimeter-level accuracy, hybrid approaches that combine learning-based initialization with featuremetric or photometric BA remain essential. See more detailed surveys on learning-based SfM in Section 4.1.

3.2 Multi-View Stereo

This section reviews recent advances in conventional MVS-based 3D reconstruction with a particular emphasis on machine vision applications. To provide context, we first discuss the role of MVS in industrial machine vision and outline the typical MVS processing pipeline.

3.2.1 The Role of MVS in Machine Vision

In machine vision, MVS pipelines are typically optimized for computational efficiency, as machine vision represents one key technology in manufacturing. It is widely employed for tasks such as automated quality inspection to detect and remove defective products from production lines or for machine control, for example by guiding an industrial robot that assembles products (Steger et al. 2018). Since machine vision is one integral step within the production process, it is required that the run-time of the machine vision algorithms can keep up with the production speed. Depending on the specific application, run-time requirements usually range from a few milliseconds to a few seconds.

In recent years, 3D data has become increasingly important for a wide range of industrial applications. For example, objects can be detected and their 6D pose estimated based on 3D point clouds enabling bin picking or pick-and-place applications where robots grasp and manipulate parts during automated assembly (Drost et al. 2010). 3D geometry is also crucial for the inspection of complex product shapes to verify dimensional accuracy or surface geometry (Michel and Ulrich 2025). Furthermore, in electronics manufacturing, printed circuit boards can be examined in 3D to ensure completeness or correct component placement.

Besides MVS, several other principles exist for generating 3D data for machine vision applications (Steger et al. 2018). Sheet-of-light systems project a laser line onto the object and analyze its deformation in the camera image to infer depth. Consequently, for a complete reconstruction the object must be moved relative to the sheet-of-light system and the individual reconstructions need to be registered. Structured-light methods project coded patterns (e.g., Gray codes) onto the scene. The codes are used in the camera image to establish correspondences between the projected pattern and image points from which the depth is calculated by triangulation. Achieving higher accuracy typically requires projecting multiple patterns sequentially and capturing each with the camera. Finally, time-of-flight cameras emit radiation and determine depth by measuring the time delay between emission and reception of the reflected signal.

Compared to these principles, MVS offers the following advantages: it does not require relative movement (unlike sheet-of-light systems), requires only a single image per camera and hence allows a fast 3D reconstruction even of dynamic objects (unlike sheet-of-light and structured light systems), and generally achieves higher reconstruction accuracy (compared to time-of-flight sensors). However, since MVS is based only on passive sensor data (i.e., camera images), it requires a sufficient amount of texture on the object surfaces. This limitation can be mitigated by projecting a random texture onto the scene. These advantages make MVS still the preferred solution for many machine vision applications that require 3D data.

3.2.2 Typical MVS Pipeline in Machine Vision

The typical 3D reconstruction pipeline in machine vision is described in (Steger et al. 2018, Chapter 3.10). In most machine vision applications, a MVS system comprises at least two cameras that are rigidly mounted to observe the workspace.

In a first step, the cameras are calibrated often by using a planar calibration target with calibration marks (e.g., circular marks or corners of a checkerboard pattern) with accurately known positions. For this purpose, the target is placed in the workspace at about 10 to 20 different poses and for each target pose, images are acquired from all cameras. The calibration is performed by minimizing the repro-

jection error of all visible calibration marks in all images of all cameras. The results of the calibration are the interior orientations including lens distortion parameters of all cameras as well as the relative orientations of the cameras with respect to a selected reference camera. Lens distortions are typically modeled by taking radial as well as decentering components into account (Brown 1971). Unlike MVS approaches that rely on a SFM result, this procedure ensures a highly accurate and consistent calibration even in challenging scenes that lack robust image features and avoids the scale-ambiguity, thereby enabling metric 3D reconstruction. Furthermore, since the cameras are rigidly mounted, the calibration step needs to be performed only once in the offline phase. Thus, the computationally expensive SfM step during the online-phase can be omitted.

In a second step, camera pairs are selected for which binocular disparity images are to be computed. A well-chosen set of pairs is essential for reconstruction quality. Pairs with small baselines suit binocular disparity algorithms, while larger baselines yield more accurate depth. Selected pairs should cover different viewpoints to ensure a complete surface coverage. Since runtime grows linearly with the number of pairs, the goal is to choose as many as necessary but as few as possible to balance completeness and efficiency. For each selected camera pair, a rectification map is computed for both images. The rectification map stores the geometric image transformation that rectifies the original images of both cameras into the epipolar standard configuration and simultaneously eliminates lens distortions. Consequently, corresponding points have the same row coordinate in both rectified images. The rectification maps may also include weights for bilinear image interpolation to ensure a fast and accurate rectification. The rectification maps also need to be computed only once in the offline phase.

In the online phase, a 3D reconstruction of the object(s) in the workspace is performed. Each camera acquires an image. For each selected camera pair, the two corresponding images are rectified to the epipolar standard configuration by applying the pre-computed rectification maps. Rectified images are an essential prerequisite for efficient disparity computation. The goal of disparity computation is to determine, for each pixel in one image, the column offset to the corresponding pixel in the second image. The pixel-wise disparities can directly be transformed into a depth value and hence can be transformed into a dense 3D reconstruction. A variety of approaches exist for disparity estimation. The fastest are still often based on template-matching approaches that use a gray-value-based similarity measure, e.g., the sum of absolute gray value differences (SAD), the sum of squared gray value difference (SSD), or the normalized cross correlation (NCC) (Steger et al. 2018, Chapter 3.10.1.6). A computationally more demanding but widely used approach is semi-global matching (Hirschmüller 2005, 2008) that minimizes an energy function along multiple one-dimensional paths across the image.

Typically, one or more of the following improvements are applied to this standard procedure: Since the disparity is inversely related to the depth of a point, a significant speed-up is achieved by restricting the disparity search space on the epipolar line in accordance with the depth range of the objects within the workspace. To accelerate computation, similarity measures can be computed recursively, yielding run-times that are independent of the template-matching window size and thereby enabling real-time performance. To improve the robustness of the disparity estimation, a threshold is applied to the maximum (NCC) or the minimum (SAD, SSD) similarity measure, for example, to recognize occlusions. The accuracy of the 3D reconstruction is improved by subpixel disparity estimation, for example, by fitting a parabola through the three pixel-wise disparities at the optimum of the similarity measure and extracting its extremum. To improve the robustness to weak texture or occlusions, additional techniques are sometimes applied: To increase reliability, image areas of weak texture can be identified (e.g., low standard deviation of the gray values) and excluded from the disparity estimation. Occlusions can be identified by applying a consistency check that switches the order of the stereo image pair and performs the disparity estimation a second time. Only those disparities are included in the result that are consistent in both computations.

In the final step, the 3D reconstructions for each camera pair are transformed into a common coordinate system, for example, the coordinate system of the reference camera, by using the calibrated relative orientations. A consistent smooth surface without outliers can be efficiently obtained, for example, by minimizing an energy functional with a total variation regularization (Zach et al. 2007).

3.2.3 Recent Advances in Conventional MVS

Conventional MVS methods have continued to develop around the PatchMatch framework, which remains one of the most effective paradigms for dense depth estimation from calibrated multi-view images. In PatchMatch-based MVS, per-pixel depth and normal hypotheses are iteratively initialized, propagated, evaluated, and refined using photometric and geometric consistency. Recent advances have mainly improved this pipeline by enhancing hypothesis propagation, adaptive view selection, and depth recovery in weakly textured or incomplete regions. Xu and Tao (2019) proposed ACMH and ACMM to address two limitations of earlier PatchMatch MVS methods: inefficient hypothesis propagation and unreliable view aggregation. ACMH introduces adaptive checkerboard sampling and

multi-hypothesis joint view selection, enabling more effective propagation and pixel-wise aggregation-view selection. ACMM further incorporates multi-scale geometric consistency, allowing reliable depth estimates from coarse scales to guide fine-scale reconstruction in low-texture regions. Following this direction, subsequent methods have focused more explicitly on textureless planes and unreconstructed areas. Sun et al. (2022) proposed an MRF-based depth inference strategy for large-scale scene reconstruction, in which unreconstructed pixels are assigned candidate depth hypotheses from neighboring reliable estimates and then optimized using a Markov random field. This strategy improves completeness in large textureless regions that are often discarded by conventional photometric consistency filtering. MP-MVS (Tan et al. 2023) further extends this line of work by combining multi-scale window PatchMatch with planar-prior-assisted depth estimation. It improves checkerboard sampling by reducing unreliable sampling in distant regions and uses cross-view geometric consistency to select reliable triangulated vertices for planar correction, thereby improving depth estimation in untextured and piecewise planar areas.

Another important line of research focuses on improving the local matching support in PatchMatch, because fixed planar patches are often inadequate for curved surfaces, illumination variations, and large textureless regions with ambiguous matching costs. Song et al. (2020) proposed a local quadric window for PatchMatch-based MVS, replacing the fixed planar support assumption with a more flexible quadric surface model. By representing local geometry with a 3D quadric function and separating photometric and geometric properties, this method improves the robustness of matching costs under complex surface shapes and illumination changes. Rather than modifying the surface model, APD-MVS adapts the patch support itself. Wang et al. (2023b) observed that unreliable pixels in textureless regions require a larger receptive field, whereas uniformly enlarging all patches increases matching ambiguity and memory consumption. APD-MVS therefore detects ambiguous matches and adaptively deforms the patches of unreliable pixels, using reliable neighboring pixels as anchors to expand the effective matching region only where necessary. Its extension, APDe-MVS (Zeng et al. 2025), further separates pixels into reliable and unreliable groups and applies different adaptive patch deformation strategies to improve robustness against matching ambiguity within the non-learning PatchMatch framework.

Beyond RGB-based photometric consistency, PolarPMS (Zhao et al. 2024) incorporates polarization information into PatchMatch-based MVS. By jointly evaluating photometric and polarimetric consistency for depth and normal hypotheses, it introduces additional physical constraints for textureless or reflective surfaces where color-based matching is unreliable. Overall, recent advances in conventional MVS reveal a coherent progression from improved hypothesis propagation and adaptive view selection to geometry-aware priors, adaptive matching support, and physically meaningful image cues. These developments enhance reconstruction completeness and robustness while preserving the interpretability and full-resolution scalability required for photogrammetric applications.

Examples for recent developments in the field of machine vision focus on leveraging specialized lenses and sensors as well as on the use of industrial robots:

Hypercentric lenses have their entrance pupil in front of the lens so that the object lies between entrance pupil and lens, producing a converging view in which object parts closer to the camera appear smaller in the image. This enables the simultaneous imaging of an object's top and lateral surfaces (see Fig. 1). This is particularly advantageous for inspection tasks where this surface coverage would otherwise require multiple cameras or object rotations. (Ulrich and Steger 2019) provide a multi-view camera model for hypercentric lenses, a calibration procedure, as well as an image rectification procedure that minimizes distortions and enables an efficient multi-view 3D reconstruction by applying the approach described in Section 3.2.2.

Line-scan cameras are widely used in machine vision due to their lower cost per resolution compared to area-scan cameras. This makes line-scan cameras ideal for inspecting small objects over a large field of view or for applications requiring very high accuracy. As mentioned before, for stereo reconstruction, rectifying the image pair into the epipolar configuration is crucial for efficient correspondence search. However, for line-scan cameras with conventional (entocentric) lenses, such rectification is not possible in general (Kim 2000). (Steger and Ulrich 2022) propose a multi-view line-scan camera model for cameras with telecentric lenses. The authors show that an accurate and efficient epipolar rectification of line-scan images is possible if telecentric lenses are used, again enabling a fast disparity estimation. They also propose an efficient algorithm to compute 3D coordinates from the estimated disparities.

In addition to the two established approaches for determining the camera poses in MVS (either using SfM or a prior calibration step), an alterative is available when an industrial robot arm is used in the application: the camera can be mounted on the robot's end effector and precisely moved to predefined poses with it. In this case, MVS can be applied to the acquired images even with a single camera and without having to apply a prior SfM step (Hillemann et al. 2024). For this, a hand-eye calibration must have been performed before, e.g., (Horaud and Dornaika 1995; Dornaika and Horaud 1998; Ulrich and Steger 2016; Steger et al. 2018). Almost all approaches do not explicitly model

the uncertainty of the robot. While industrial robots are known for their high repeatability, their absolute accuracy is typically much lower. As shown by (Ulrich and Hillemann 2024), neglecting this uncertainty can significantly degrade the result of the hand-eye calibration. They therefore propose a method that incorporates this uncertainty into the calibration process, which substantially improves the accuracy of the resulting hand-eye poses, and hence, the accuracy of robot-based multi-view stereo.

3.3 Surface Reconstruction

Similar to Sect. 2.3, recent advances in conventional surface reconstruction can also be broadly categorized into explicit and implicit methods.

3.3.1 Explicit Methods

Explicit methods directly assemble polygonal surfaces from 3D points, often supplemented with multi-view visibility cues. These approaches offer precise local control over sampling density and element quality.

Among these methods, Delaunay triangulation (DT) and its dual Voronoi diagram (VD) (De Floriani 1987) are the most classical approaches, where they select the subset of simplices that best approximate the underlying geometry while avoiding the creation of elongated triangles. However, DT and VD are sensitive to noise, non-uniform sampling, and sparsely observed regions, meaning that the resulting surface quality heavily depends on the quality of the input data. Several classical variants have been proposed, including α-shapes, Crust (Ni and Ma 2010), Power-Crust (Amenta et al. 1998), and Cocone (Amenta et al. 2000). α-shapes generalize the convex hull by retaining Delaunay simplices whose circumsphere radius is below a user-specified scale parameter, whereas Crust and Power-Crust use Voronoi poles and power diagrams, respectively, to select simplices that approximate the medial axis and improve robustness to non-uniform sampling. Cocone and Tight-Cocone (Dey and Goswami 2003) further refine this selection by keeping triangles whose dual Voronoi edges lie within a cone around the estimated normal direction, enabling fast and topologically reliable reconstructions under appropriate sampling conditions.

Region-growing surface reconstruction initializes from a single seed triangle and incrementally extends the mesh by selecting a third vertex under prescribed topological criteria, continuing this process until all points have been examined and a complete surface is formed. Owing to its conceptual simplicity, low computational cost, and good scalability to large point sets, this strategy has been widely adopted. Representative methods include the ball pivoting algorithm (BPA) (Bernardini et al. 2002), intrinsic property driven (IPD) approaches (Lin et al. 2004), and enveloping-sphere schemes (Di Angelo et al. 2011). BPA rolls a virtual ball along the edges of the mesh, inserting a triangle whenever the ball touches three points without enclosing others, thus interpolating the surface. IPD methods select the next vertex based on intrinsic measures of sampling uniformity, yielding a minimum-weight triangulation that approximates the input surface while constrained to a two-dimensional manifold. Enveloping-sphere schemes choose new points that satisfy a Gabriel 2-simplex criterion, producing Delaunay-like triangulations with high tessellation rates and good reconstruction quality on well-sampled point clouds.

More closely related to photogrammetric pipelines are multi-view visibility-based methods. Space-carving (Kutulakos and Seitz 2000) and visual-hull approaches (Laurentini 2002; Matusik et al. 2000) reconstruct the mesh by successively removing voxels or polyhedral cells that contradict silhouette or photometric constraints, yielding watertight surfaces that closely follow the observed outlines. However, concave regions that are invisible from all cameras cannot be recovered unless additional priors or depth information are introduced. Following the same visibility principle, visibility-consistent surface reconstruction methods (Labatut et al. 2007, 2009) employ Delaunay triangulation for adaptive space decomposition and graph-cut optimization to classify tetrahedra as either free space or occupied, efficiently extracting the separating interface as a watertight mesh. These methods simultaneously enforce geometric regularity and photometric consistency, but their effectiveness is highly dependent on accurate calibration and robust visibility estimation.

Furthermore, building reconstruction often benefits from methods that explicitly incorporate structural priors. Edge- and curve-guided approaches (Schnabel et al. 2007; Nan and Wonka 2017) leverage the regularity of man-made architecture by first detecting and fitting basic line and plane primitives, and then stitching them into polygonal surfaces using coplanarity and angular constraints. By utilizing rooflines, ridges, and façade edges as high-level cues, these techniques produce compact meshes with polygons closely aligned to structural features. Overall, such explicit formulations prioritize accurate geometry, controllable topology, and sharp-feature preservation, especially when point sampling and multi-view observations are reliable.

3.3.2 Implicit Methods

Implicit methods, unlike explicit methods, do not directly connect sample points into a mesh; instead, they estimate a continuous scalar field, with its zero-level set defining the surface. This approach inherently promotes watertight reconstructions while smoothly averaging out noise. Implicit

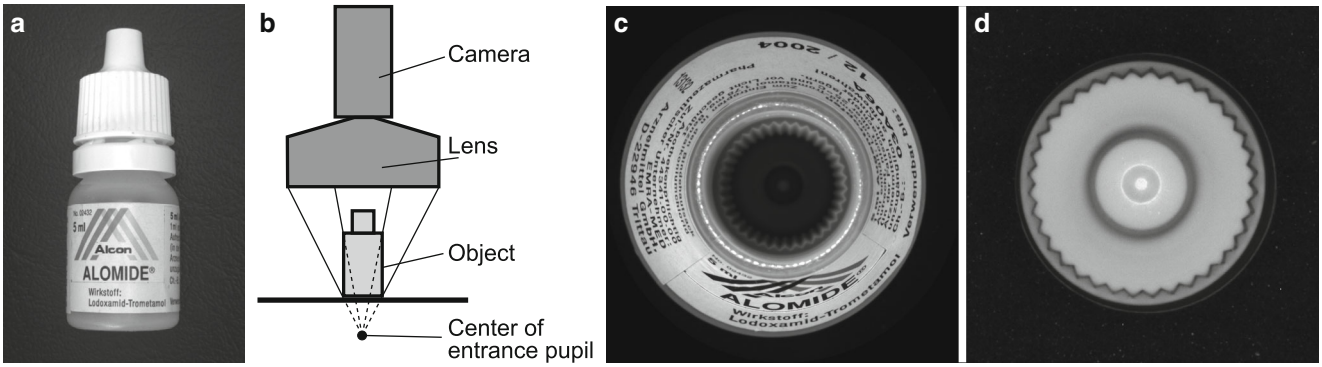

Fig. 1 Imaging the top and side of a bottle using a hypercentric lens. The label of the bottle (**a**) can be inspected by using a camera with a hypercentric lens that is positioned above the bottle (**b**). The bottle is placed between the front of the lens and the entrance pupil, creating a converging view. (**c**) Resulting image captured with a hypercentric lens. (**d**) For comparison, image captured with a conventional perspective lens with the same camera pose. Source: (Ulrich and Steger 2019)

methods can be further categorized into those based on Signed Distance Fields (SDF) and those based on Poisson equations.

SDF-Based Surface Reconstruction SDF-based surface reconstruction methods model the surface as the zero-level set of a smooth signed distance field, with the primary distinction among methods lying in the approach used to estimate this field from data. Early approaches, such as from (Hoppe et al. 1992), fit local tangent planes using Principal Component Analysis and interpret the points as samples of the gradient of an implicit function, from which an SDF and its zero set are derived. Variational and level-set methods (Zhao et al. 2001) evolve the SDF by minimizing an energy functional via Partial Differential Equations, balancing data fidelity with smoothness, and naturally handling topology changes, although at high computational cost. A large class of methods approximates the SDF with scattered-data interpolants: implicit Moving Least Squares surfaces (Kolluri 2008) and Radial Basis Function (RBF)-based (Carr et al. 2001) implicit surfaces fit a global or piecewise-global distance function, whose zero level reproduces the point set. Multi-level partition-of-unity (MPU) (Ohtake et al. 2005) schemes decompose the domain into overlapping patches, fit local RBFs, and blend them hierarchically to obtain a scalable SDF for large point clouds. Closely related projection-based approaches (Levin 2003; Pauly 2003) define an implicit function by minimizing a weighted projection error of sample points onto the unknown surface, producing a signed distance-like scalar field, whose zero level is the reconstructed surface. After estimating the SDF field, the zero-level surface is typically extracted using Marching Cubes (Lorensen and Cline 1998) or its dual, dual contouring (Ju et al. 2002): Marching Cubes places mesh vertices on grid edges where the SDF changes sign, while dual contouring places vertices inside grid cells and connects them across cells, better preserving sharp features.

Poisson Surface Reconstruction Poisson Surface Reconstruction (PSR) (Kazhdan et al. 2006) treats oriented samples as gradients of a binary occupancy function; solving the associated Poisson equation recovers a globally consistent implicit surface that is robust to noise and irregular sampling, with a smoothness-fidelity trade-off primarily controlled by the octree depth. Screened Poisson Surface Reconstruction (Kazhdan and Hoppe 2013) enhances this approach by adding pointwise positional constraints, "screening" the Poisson equation to ensure that the implicit function both fits the noisy data and remains globally smooth. This screened variant improves adherence to the input samples and better preserves fine structures, while maintaining the efficiency of the multigrid octree solver. Building on this foundation, Stochastic Poisson Surface Reconstruction (Sellán and Jacobson 2022) reinterprets PSR in a probabilistic context, replacing the implicit function with a Gaussian-distributed scalar field. Each point in space is assigned a mean and variance, enabling statistical queries such as collision likelihood, next-best-view planning, and uncertainty-aware surface repair. These Poisson-based methods naturally generate watertight surfaces and effectively handle gaps or redundant measurements, but they depend on accurate normal orientations and still require additional feature-aware processing to fully preserve sharp edges.

4 Recent Advances of Learning-Based 3D Reconstruction

This section presents recent technical developments in learning-based 3D reconstruction, focusing on SfM, MVS, and surface reconstruction.

4.1 Structure from Motion

In learning-based SfM, recent advances can be grouped by network inputs. The first family comprises correspondence- or tie-point-based architectures that mirror conventional SfM by consuming local feature matches. The second group, image-based architectures, takes the original images as input that is similar to feed-forward 3R methods (see more details in Sect. 4.2.3).

4.1.1 Tie-Point-Based Architectures

(Moran et al. 2021) propose ESfM, a deep permutation-equivariant framework that remains invariant to the ordering of multi-view correspondences: tie-point tracks are packed into a tensor, a permutation-equivariant encoder ensures order invariance, and two heads regress camera projection matrices and 3D point coordinates. Both supervised and unsupervised training regimes are explored for calibrated and uncalibrated settings. Building on ESfM, (Gong 2023) analyzes performance by modifying the MLP layers within the permutation-equivariant encoder, while (Khatib et al. 2025) introduce RESfM, which follows the ESfM design but adds a third head for tie-point outlier detection. To handle large-scale UAV datasets, (Chen et al. 2024c) present DeepAAT, which also ingests GPS priors: images are clustered into sub-blocks, their tie points and GPS cues are processed to predict camera poses and 3D points, and the sub-solutions are merged into a single block. (Brynte et al. 2024) model the observation set as a heterogeneous graph with nodes for views, 3D points, and their projections, connected by visibility edges; stacked Graph Attention Network layers aggregate over this graph to infer camera poses (R_i, t_i) and point coordinates X_j in a permutation-equivariant manner. In practice, these correspondence-driven models typically append a canonical global bundle adjustment (BA) initialized from the network outputs for final refinement.

In contrast, VGGSfM (Wang et al. 2024c) proposes a fully differentiable, learning-based SfM pipeline that dispenses with correspondence generation and canonical bundle adjustment (BA). Building upon the work of Lindenberger et al. (2021), it employs a coarse-to-fine tracker architecture to localize accurate tie points. Subsequently, two dedicated deep transformer networks are introduced: one for joint estimation of camera poses, conditioned on both global image feature maps and local descriptors of the tie points, and another for 3D point reconstruction. A differentiable Levenberg-Marquardt layer is further integrated to refine the geometric estimates in an end-to-end manner. Despite its architectural innovation and differentiability, VGGSfM faces practical scalability limitations: it struggles to process datasets comprising more than a few thousand high-resolution images—a scale routinely encountered in real-world photogrammetric applications.

4.1.2 Image-Based Architectures

SfM-Net (Vijayanarasimhan et al. 2017) learns scene structure and camera motion directly from video frames. It employs two convolutional/deconvolutional sub-networks: one predicts a pixel-wise depth map of the first frame, which can be lifted into a dense point cloud, while the other estimates the 3D rotation and translation of neighboring frames relative to the reference. Similar to ESfM (Moran et al. 2021), SfM-Net supports both unsupervised and supervised training paradigms. Another notable approach is BA-Net (Tang and Tan 2019), which formulates SfM as a learning problem based on feature metric error minimization. Given a sequence of video frames, BA-Net first generates a set of basis depth maps for the reference frame and linearly combines them to produce the final depth estimate. A differentiable bundle adjustment (BA) layer—implemented via the Levenberg-Marquardt (LM) algorithm—is then introduced to jointly refine the scene structure and camera poses, using the estimated depth and feature maps extracted by a DRN-54 backbone (Yu et al. 2017). In contrast, DeepSfM (Wei et al. 2020) adopts a cost-volume-based architecture to simultaneously estimate depth and camera pose. It utilizes two separate 2D CNNs to extract feature representations from input frames, which are then used to construct distinct cost volumes. These volumes are processed by two dedicated 3D CNNs: one incorporates a context network and regression head to refine the depth map, while the other predicts relative camera poses. More recently, DRO (Gu et al. 2021) proposes a more unified framework that integrates depth and pose estimation into a single cost volume. Instead of decoupling the two tasks, DRO employs a Gated Recurrent Unit (GRU) to iteratively refine both depth and pose in a joint optimization loop, leading to improved consistency and accuracy.

However, the aforementioned image-based deep learning architectures are generally unsuitable for processing conventional real-world photogrammetric datasets. This limitation stems from three main factors: First, these methods are primarily designed for video sequences with strong spatiotemporal continuity, i.e., they assume extremely high inter-frame overlap, which is often not satisfied in typical photogrammetric acquisitions (e.g., aerial or close-range surveys with sparse or irregular sampling). Second, these approaches adopt a fixed-reference strategy, where the first frame serves as the anchor and all other views are aligned to it. Such a scheme can fail in complex scenarios, particularly in close-range photogrammetry where no single view provides a globally consistent reference due to occlusions, viewpoint diversity, or limited coverage. Finally,

Table 1 Photogrammetric comparison of recent reconstruction methods.

Technique family	Metric fidelity	Robustness	Uncertainty	Scalability	Main challenge
Conventional MVS	Strong, traceable	Texture-sensitive	Interpretable	Full-resolution strong	Weak-texture sensitivity
Depth Map-Based Methods	Strong in-domain	Completeness strong	Weakly calibrated	Cost-volume limited	Domain transfer
NeRF-Based Methods	Regularization dependent	Appearance strong	Weakly quantified	Optimization limited	Shape-radiance ambiguity
3DGS-Based Methods	Regularization dependent	Appearance efficient	Weakly quantified	Memory limited	Illumination sensitivity
Feed-Forward Generalizable Methods	Accuracy uncertain	Prior dependent	Reliability unverified	Resolution limited	Pose-scale consistency

most of them typically rely on dense cost volumes and per-frame depth estimation, which incur substantial memory and computational demands. As a result, they struggle to scale to large photogrammetric datasets comprising hundreds or thousands of high-resolution images, especially under practical GPU memory constraints.

4.2 Multi-View Stereo

MVS has emerged as a cornerstone of image-based 3D reconstruction, leveraging multiple images captured from diverse viewpoints to synthesize a coherent and detailed 3D representation. Owing to its balance of efficiency, accuracy, and scalability, MVS has become the de facto standard for dense reconstruction in complex real-world environments. In recent years, the advent of deep learning has spurred the development of numerous learning-based MVS approaches that significantly outperform classical methods in both accuracy and robustness. Inspired by (Wang et al. 2024a), these methods can be mainly categorized into three paradigms: *Depth Map-based Methods, differentiable rendering-based scene representation methods* including Neural Radiance Fields (NeRF) (Mildenhall et al. 2021) and 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023), and *Feed-Forward Generalizable Methods*.

To situate these recent reconstruction categories within a photogrammetric evaluation framework, Table 1 compares them with conventional MVS using a concise set of criteria. Rather than ranking individual methods, the table uses short evaluative labels to summarize their typical strengths and limitations. Metric fidelity denotes the ability to support measurement-grade geometry; robustness describes stability under weak texture, sparse observations, reflective surfaces, illumination changes, or domain shift; uncertainty indicates whether the reliability of depth, pose, or reconstructed geometry can be interpreted, calibrated, or quantified; and scalability refers to the ability to process full-resolution images, large image collections, and large-scale scenes. The final column identifies the main challenge that limits each reconstruction category in photogrammetric applications.

As summarized in Table 1, the key distinction is between geometry-driven reconstruction and learning-driven completeness or visual plausibility. Conventional MVS remains advantageous in metric fidelity, interpretability, and scalability to full-resolution imagery, but its performance can degrade in weakly textured, reflective, or radiometrically unstable regions. Depth map-based methods improve reconstruction completeness through learned matching and depth regularization; however, their reliability is still affected by domain transfer, confidence calibration, and cost-volume scalability. NeRF- and 3DGS-based methods further strengthen appearance reconstruction and novel-view synthesis, yet their metric geometry usually depends on additional depth, normal, SDF, or multi-view constraints. Feed-forward generalizable methods offer a more direct route to sparse-view or weakly constrained reconstruction, but their pose-scale consistency and measurement reliability still require systematic validation. The following subsections review these reconstruction categories in detail: depth map-based reconstruction, differentiable rendering-based scene representation methods represented by NeRF and 3DGS, and feed-forward generalizable reconstruction methods.

4.2.1 Depth Map-Based Methods

Depth map-based methods explicitly estimate pixel-wise dense depth maps for each view from calibrated multi-view images, which are subsequently fused into a coherent 3D representation. This section first reviews supervised and unsupervised learning paradigms for depth map estimation, followed by a discussion of recent extensions specifically designed for photogrammetric applications.

Supervised Methods Supervised depth map-based methods depend on the availability of high-quality ground-truth depth maps, typically obtained from active sensors such as LiDAR or structured-light scanners, to provide

direct supervision during training. These reference depth maps enable pixel-wise or patch-wise learning objectives that guide the network toward accurate and geometrically consistent depth estimation across the entire scene. Such methods are commonly categorized into online and offline pipelines, depending on their input requirements, computational demands, and adherence to real-time constraints.

Online methods, exemplified by MVDepthNet (Wang and Shen 2018) and DeepVideoMVS (Duzceker et al. 2021), are designed for real-time or near-real-time deployment and typically process streaming video sequences or short image clips. To meet stringent latency requirements, they construct compact 3D cost volumes and employ lightweight 2D or shallow 3D convolutional architectures. These approaches often adopt simple plane-sweeping strategies to build cost volumes and place strong emphasis on robust multi-view feature extraction, temporal coherence modeling, and efficient depth refinement mechanisms—striking a careful balance between reconstruction accuracy and computational efficiency.

In contrast, offline methods, such as foundational MVS-Net (Yao et al. 2018) and its extension R-MVSNet (Yao et al. 2019), are designed to process high-resolution image sets captured from arbitrary viewpoints, prioritizing reconstruction accuracy over real-time performance. These methods typically construct dense 3D cost volumes via plane sweeping and regularize them using powerful 3D CNNs or RNN modules. Depth estimation is then obtained through either soft-argmax regression or discrete probability classification over depth hypotheses. Subsequent advancements have sought to refine depth geometry and improve robustness in challenging regions. For instance, DMVNet (Ye et al. 2023) introduces a dual-depth representation, employing saddle-shaped cost volume cells that dynamically oscillate to capture both local precision and global context, thereby yielding more accurate depth maps. Building on this, RRT-MVS (Jiang et al. 2025a) enhances regularization through a Recurrent Regularization Transformer, which decouples processing along depth and spatial dimensions. It leverages Recurrent Self-Attention (R-SA) to effectively aggregate global matching costs and suppress spurious feature correlations, significantly improving handling of depth discontinuities and occlusion boundaries. DI-MVS (Jiang and Ma 2025), on the other hand, incorporates a depth-aware iterative refinement strategy combined with a hybrid loss function, optimizing depth maps in a coarse-to-fine manner. This design enhances robustness in low-texture regions and improves generalization across various benchmarks.

Unsupervised Methods The scarcity and practical difficulty of acquiring ground truth depth data have significantly limited the applicability of supervised depth map-based methods, particularly in large-scale or real-world photogrammetric scenarios. This challenge has spurred growing interest in unsupervised depth estimation, which offers promising solutions for 3D reconstruction without requiring labeled data. Unsupervised methods can be broadly classified into two categories: multi-stage methods and end-to-end methods, depending on whether pseudo-labels are generated in advance.

Multi-stage methods typically begin by generating pseudo-depth maps or intermediate geometric cues serving as pseudo-labels, which are subsequently used for self-supervised training. Prominent approaches include CVP-MVSNet (Yang et al. 2020), which constructs a cost-volume pyramid to iteratively refine depth predictions, and KD-MVS (Ding et al. 2022), which leverages a knowledge distillation framework to transfer geometric knowledge from a pre-trained teacher model to a more efficient student model, thereby improving depth map accuracy without ground-truth supervision. Additionally, (Wang et al. 2024h) proposes a multi-stage self-training strategy that produces pseudo-labels from a reference view to enhance training stability and generalization. DualNet (Wang et al. 2025e) adopts a hybrid paradigm, combining self-supervised teacher learning with pseudo-label-guided student training to better enforce feature metric consistency across views—addressing a key challenge in unsupervised multi-view geometry.

In contrast, end-to-end methods eliminate the need for pre-generated pseudo-labels. Instead, they primarily use photometric consistency between reconstructed and observed views as the core supervisory signal. Over time, this paradigm has been extended with complementary constraints to improve robustness and accuracy. For instance, RC-MVSNet (Chang et al. 2022) incorporates neural rendering to improve depth map consistency by synthesizing reference views, addressing occlusions, and non-Lambertian surfaces. CL-MVSNet (Xiong et al. 2023) employs contrastive learning to enforce cross-view feature consistency, yielding more reliable depth estimates in low-texture and reflective regions. Similarly, DIV-MVS (Rich et al. 2024) couples view-synthesis with a depth-smoothness prior to reduce artifacts and enhance depth map accuracy, particularly in occluded or complex regions.

Applications in Photogrammetry Depth map-based methods have been increasingly adopted across diverse photogrammetric applications, demonstrating strong adaptability to domain-specific challenges. In large-scale terrain modeling, TS-SatMVSNet (Zhang et al. 2025c) introduces a slope-aware depth estimation module that explicitly incorporates terrain gradient priors, significantly improving elevation accuracy in satellite-based topographic reconstruction. In ecological and forested environments where

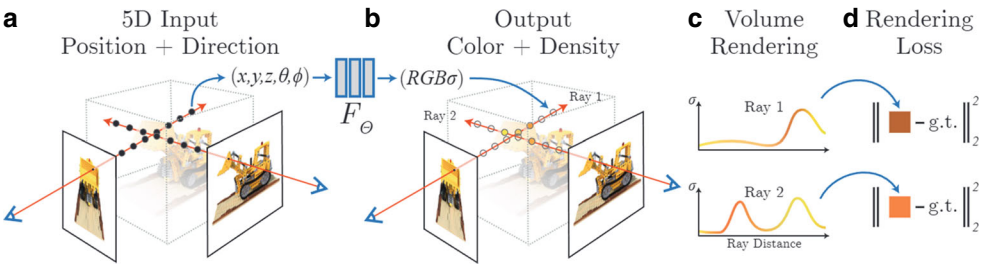

Fig. 2 Pipeline of Neural Radiance Fields (NeRF). Given camera rays, NeRF samples 5D inputs consisting of 3D position and viewing direction, and uses a neural network to predict color and volume density at each sampled point. The predicted radiance and density values are integrated along each ray through differentiable volume rendering, and the network is optimized by minimizing the rendering loss between synthesized and observed images. Source: (Mildenhall et al. 2021)

complex canopies and sparse textures pose significant challenges, CDP-MVS (Liu et al. 2024d) enhances the completeness and geometric fidelity of reconstructed point clouds through an advanced confidence propagation mechanism, underscoring the viability of learning-based MVS in vegetation-rich scenes. Building on this, FS-MVSNet (Chen et al. 2025d) integrates multi-scale feature pyramids with attention-based fusion to refine depth predictions in forested areas, not only boosting reconstruction accuracy but also enabling efficient extraction of dendrometric parameters such as diameter at breast height (DBH). For urban modeling and large-scale aerial photogrammetry, EG-MVSNet (Zhang et al. 2024) proposes an edge-aware depth inference architecture that preserves sharp geometric boundaries of buildings, thereby delivering more robust and structurally coherent 3D city models. Complementing these learning-based approaches, MVP-Stereo (Yan et al. 2024) achieves competitive reconstruction quality through a classical yet highly optimized pipeline that combines a multi-view dilated ZNCC (Zero-mean Normalized Cross-Correlation) similarity measure with a parallelized PatchMatch strategy, offering a scalable solution for high-resolution aerial datasets.

4.2.2 Differentiable Rendering-Based Scene Representation Methods

Compared with traditional photogrammetric reconstruction pipelines, differentiable rendering-based scene representation methods provide an alternative paradigm by directly optimizing scene representations from multi-view observations. These methods establish the relationship between image observations and 3D scene models through differentiable rendering. Specifically, Neural Radiance Fields (NeRF) represent scenes using implicit neural fields optimized via differentiable volume rendering, whereas 3D Gaussian Splatting (3DGS) adopts an explicit representation based on optimized Gaussian primitives and differentiable rasterization. In this section, we review NeRF-based and 3DGS-based approaches from three perspectives closely related to classical photogrammetry: **geometric fidelity, efficiency**, and **large-scale reconstruction**.

NeRF Neural Radiance Fields (NeRF) (the vanilla workflow is shown in Fig. 2) and their variants integrate multi-view geometry and appearance into a differentiable volume-rendering framework. This framework couples volumetric density with view-dependent radiance and minimizes image-domain errors through backpropagation. While NeRF naturally addresses view dependence and complex occlusions, it also introduces challenges, such as shape-radiance coupling, slow convergence, and limited geometric interpretability. These limitations have motivated a large body of follow-up NeRF-based methods for reconstruction, whose chronological development is summarized in Fig. 3.

High Fidelity Although NeRF excels at novel-view synthesis, its density-based formulation is misaligned with the surface-centric requirements of photogrammetry, which demand accurate and watertight geometry. NeuS (Wang et al. 2021a) and VolSDF (Yariv et al. 2021) address this by reformulating NeRF as a signed distance function (SDF) field with an adapted volume-rendering integral, so that the SDF zero level set is explicitly optimized as the scene surface under only RGB supervision. Geo-NeuS (Fu et al. 2022) further incorporates multi-view geometric constraints from structure-from-motion depth and photometric consistency, yielding globally more consistent surfaces and better reconstruction of thin structures and smooth regions. Neuralangelo (Li et al. 2023) combines multi-resolution hash-grid features, neural volume rendering, and normal-gradient regularization in a coarse-to-fine scheme to recover high-frequency details and support large-scale outdoor reconstructions from handheld video.

Another line of work strengthens neural implicit surfaces by injecting richer multi-view cues and priors. MVCPS-NeuS (Santo et al. 2024) embeds a constrained multi-view photometric stereo setup into NeuS, jointly resolving lighting ambiguity and surface geometry; by coupling light sources with the camera, it makes photometric stereo more

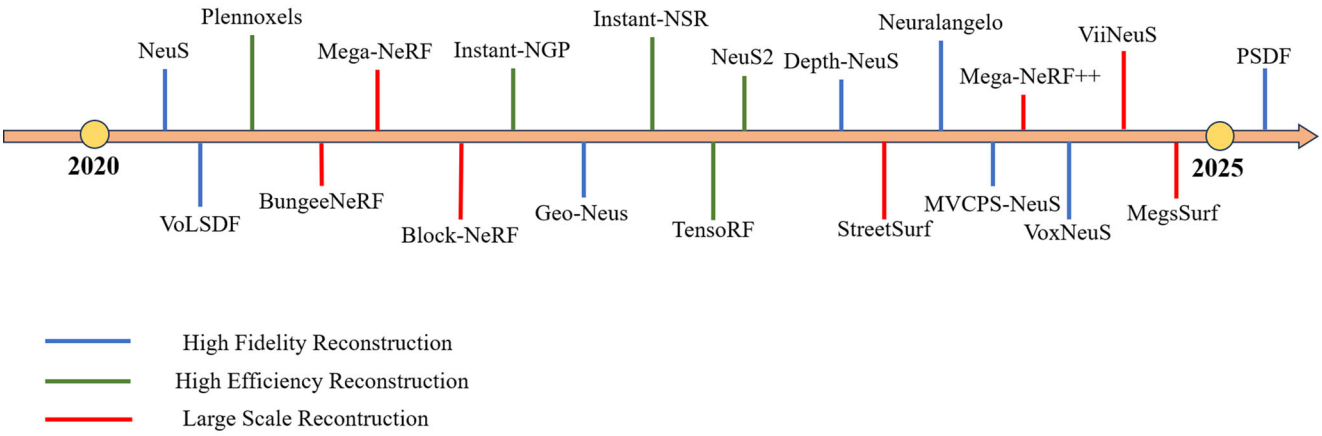

Fig. 3 Timeline of NeRF-based methods for high fidelity, high efficiency, and large-scale reconstruction (until November 2025)

practical and markedly improves fine-scale shape recovery. PSDF (Su et al. 2024) instead fuses geometric priors from an external MVS network into an SDF-based model, using visibility-aware feature consistency and depth-guided sampling to steer the SDF towards accurate surface intersections, particularly in textureless or specular regions. For weak-texture indoor scenes, Depth-NeuS (Jiang et al. 2023) introduces depth supervision and geometric consistency losses to stabilize reconstruction when color cues are unreliable. Together, these advances show that, when augmented with SDF formulations, multi-view geometry, and carefully designed priors, NeRF-style models can evolve from pure view-synthesis engines into accurate neural surface reconstruction frameworks that better satisfy photogrammetric requirements.

High Efficiency In photogrammetric applications, computational efficiency is as critical as geometric accuracy, particularly for large-scale city modeling, scenarios requiring frequent updates, or machine vision applications. Early acceleration methods for NeRF therefore replaced coordinate MLPs with explicit voxel or factorized representations. Plenoxels (Fridovich-Keil et al. 2022) uses a sparse 3D grid of spherical-harmonic coefficients optimized from calibrated views, reaching NeRF-level quality while reducing optimization time by two orders of magnitude. Instant-NGP (Müller et al. 2022) employs multi-resolution hash encoding that maps coordinates to compact feature tables, enabling high-quality fields to be trained in seconds on a single GPU. TensoRF (Chen et al. 2022a) and related tensor- or lattice-factorized fields further cut memory and computation while retaining high-frequency detail.

Meanwhile, these acceleration techniques have been extended from pure view synthesis to neural surface reconstruction. Instant-NSR (Zhao et al. 2022a) adopts the Instant-NGP hash grid for SDF learning and replaces automatic differentiation of the volume-rendering integral with finite-difference gradients, avoiding costly second-order derivatives while preserving surface quality. NeuS2 (Wang et al. 2023a) re-implements NeuS with multi-resolution hash encodings and a progressive training schedule, achieving roughly two orders of magnitude speed-up over NeuS with comparable results in static and dynamic scenes. VoxNeuS (Liu et al. 2024b) revisits explicit voxel SDF, identifying discontinuous trilinear gradients as a key instability source and proposing a gradient-interpolation and geometry-radiance disentangling scheme that improves convergence and enables minute-level reconstructions. Streaming RGB/IMU data from a HoloLens 2 to a GPU workstation for on-the-fly Instant-NeRF training, (Haitz et al. 2023) deliver an end-to-end real-time mobile-mapping pipeline whose specialized inference outperforms grid-based NeRF sampling by orders of magnitude, underscoring substantial efficiency gains for practical NeRF-based 3D reconstruction. Together, these methods transform NeRF-based surface reconstruction from an hours-long offline optimization into a practically deployable component of photogrammetric pipelines.

Large-Scale Reconstruction In photogrammetry, city-scale image collections routinely contain tens of thousands of high-resolution views, whereas standard NeRF is constrained to small, bounded scenes due to GPU memory limits and expensive per-scene optimization. To address this scalability gap, several large-scene NeRF variants have been proposed. Block-NeRF (Tancik et al. 2022) decomposes an urban environment into a grid of spatial blocks, each modeled by a compact NeRF; at inference time, only relevant sub-NeRFs are queried, while appearance embeddings and learned blending mitigate block boundaries and support incremental updates. BungeeNeRF (Xiangli et al. 2022) targets extreme multi-scale settings by progressively

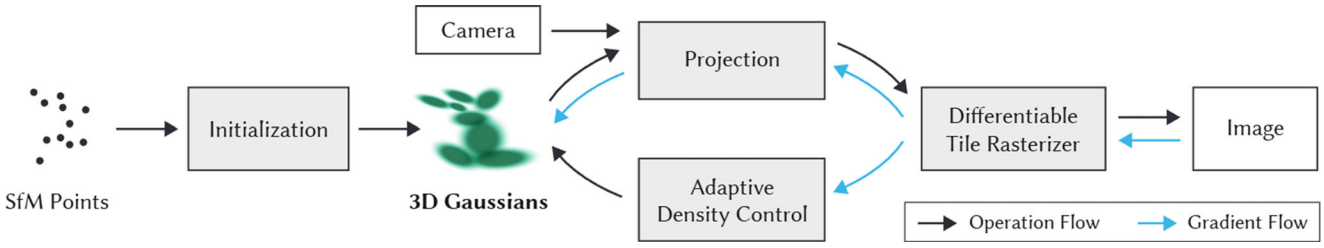

Fig. 4 Pipeline of 3D Gaussian Splatting (3DGS). The scene is initialized from sparse SfM points and represented by a set of anisotropic 3D Gaussians. During optimization, the Gaussians are projected to the image plane and rendered by a differentiable tile-based rasterizer. Image reconstruction loss is back-propagated to update Gaussian attributes, while adaptive density control progressively refines the Gaussian distribution through densification and pruning. Source: (Kerbl et al. 2023)

growing the representation: training starts from distant, low-frequency views and incrementally adds NeRF blocks and higher-frequency channels as closer views appear, enabling level-of-detail rendering over several orders of magnitude. Mega-NeRF (Turki et al. 2022) further improves scalability for drone-captured cities via a partition into visibility-aware spatial cells and a sparse network of specialized submodules, trained in a data-parallel manner. For high-resolution photogrammetric imagery, Mega-NeRF++ (Xu et al. 2024) jointly fine-tunes overlapping sub-NeRFs with an overlap-aware loss that enforces cross-block color and geometry consistency, thereby reducing seams while maintaining block-level detail.

These ideas have been extended from view synthesis to large-scale surface reconstruction. StreetSurf (Guo et al. 2023) adapts implicit surfaces to long, unbounded street trajectories by partitioning space into road, façade, and sky regions, using aligned cuboid hash grids and monocular geometric priors to recover detailed street-level geometry without LiDAR. MegaSurf (Wang et al. 2024g) focuses on aerial city reconstruction, introducing a learnable geometric guider that derives sampling fields from MVS-style cues and a divide-and-conquer training scheme that alleviates shape-radiance ambiguity at scale. ViiNeuS (Djeghim et al. 2025) addresses driving scenes with limited overlap via a hybrid initialization that jointly models volumetric density and SDF fields, using a self-supervised depth-probability formulation to gradually transition from volume to surface representation, yielding accurate urban meshes with substantially reduced training time.

3DGS As mentioned above, many NeRF-based reconstruction methods have been introduced. Although 3DGS (Kerbl et al. 2023) (the original workflow is shown in Fig. 4) is a explicit representation, it shares similar reconstruction goals with NeRF and has already inspired numerous studies in photogrammetry-related applications, as summarized in Fig. 5.

High Fidelity 3D Gaussian Splatting (3DGS) is an explicit point-based representation whose Gaussian layout directly encodes scene geometry, motivating methods that extract surface meshes directly from trained fields. One line of work refines geometry through self-regularization and redesigned splatting. SuGaR (Guédon and Lepetit 2024) aligns Gaussians to the underlying surface and applies Poisson reconstruction, enabling fast extraction of editable meshes with sharper details than NeRF-based meshing. Building on surface-attached primitives, 2DGS (Huang et al. 2024) collapses volumetric Gaussians into oriented disks and optimizes them with depth- and normal-consistency losses, while PGSR (Chen et al. 2024a) adopts planar Gaussians with unbiased depth rendering and multi-view geometric regularization; all three reconstruct meshes via TSDF fusion of rendered depth maps, which can locally over-smooth geometry. Gaussian Opacity Fields (GOF) (Yu et al. 2024b) instead derives an opacity volume from 3D Gaussians and applies adaptive Marching Tetrahedra, producing compact, topology-aware meshes even in unbounded scenes. FeatureGS (Jäger et al. 2025) optimizes additional geometric losses derived from eigenvalues of the covariance matrix of Gaussians, like the planarity loss, to refine geometry and reduce floater artifacts.

A complementary direction is to leverage pretrained models for semantic segmentation, depth estimation, and normal prediction as additional priors to guide the optimization of the Gaussians. HighSurf (Dai et al. 2024) exploits normal maps predicted by a pretrained monocular deep neural network as priors and penalizes the gradient of the rendered normals to regularize the curvature of the reconstructed surface. GS2Mesh (Wolf et al. 2024) keeps the standard 3DGS model but renders stereo-aligned novel views for a pre-trained stereo network and fuses the resulting depths into globally consistent meshes. Masked Gaussian Fields (MGFs) (Wang et al. 2025d) specialize 3DGS for buildings via façade segmentation, boundary-aware losses, and a tailored tetrahedral meshing strategy, while HRGS (Li et al. 2025a) tackles high-resolution scenes by learning a global coarse Gaussian prior and hierarchi-

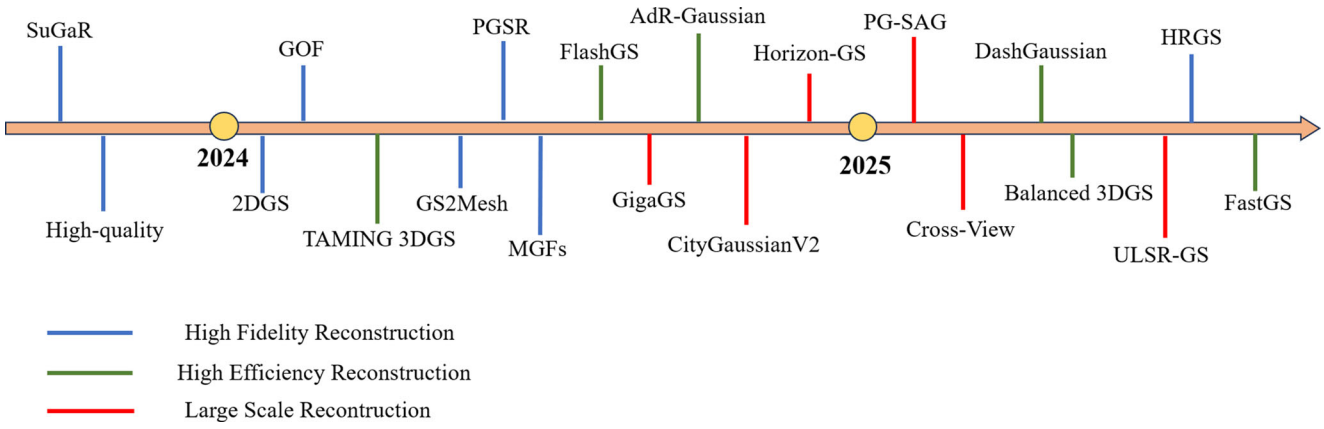

Fig. 5 Timeline of 3DGS-based methods for high fidelity, high efficiency, and large-scale reconstruction (until November 2025)

cally refining block-wise partitions, scaling 3DGS to large datasets with improved cross-block consistency.

High Efficiency Compared with NeRF, the speed of 3D Gaussian Splatting (3DGS) is governed not only by the rasterization pipeline but also by densification, where millions of Gaussians are spawned and refined. Recent work, therefore, accelerates 3DGS by either controlling densification or optimizing rendering.

On the densification side, Taming-3DGS (Mallick et al. 2024) introduces guided constructive densification with contribution-based scoring, an explicit Gaussian budget, and lightweight gradient updates, yielding 4–5× reductions in model size and training time. AdR-Gaussian (Wang et al. 2024f) shrinks the effective radius of Gaussian-tile pairs and moves culling into a parallel-friendly preprocessing stage, substantially boosting FPS without degrading PSNR. DashGaussian (Chen et al. 2025c) progressively increases resolution while adjusting the active primitive count via a closed-form complexity model, enabling large scenes to be optimized in a few hundred seconds. FastGS (Ren et al. 2025) unifies these ideas into a multi-view-consistent densification-pruning framework, achieving over 3× training speed-up on standard benchmarks and order-of-magnitude gains on challenging scenes.

On the rendering side, Taming-3DGS already introduces alternative backpropagation and parallelization schemes that reduce gradient-evaluation overhead. FlashGS (Feng et al. 2025) systematically analyzes the original rasterizer and applies kernel-level optimizations, redundancy removal, pipelining, and memory-access reorganization, reporting roughly 4× acceleration on commodity GPUs. Balanced-3DGS (Gui et al. 2024) tackles load imbalance across streaming multiprocessors through Gaussian-wise parallel rendering, fine-grained tiling, dynamic workload redistribution, and self-adaptive kernel selection, improving kernel efficiency by up to 7.5×. Together, these methods push 3DGS beyond its initial "heavy but real-time" regime toward genuinely budget-aware reconstruction and rendering on practical hardware.

Large-Scale Reconstruction As an explicit Gaussian-based scene representation, 3DGS models a scene with numerous Gaussian primitives, making it naturally amenable to divide-and-conquer strategies for large-scale environments. Existing work can be broadly grouped into rendering-oriented and surface-reconstruction methods.

For large-scale rendering, VastGaussian (Lin et al. 2024) partitions scenes into visibility-aware cells that are optimized separately and then merged, reducing training time while preserving urban detail. CityGaussian (Gao et al. 2025) further adopts a level-of-detail pipeline guided by global priors and adaptive data selection, enabling real-time city-scale rendering with compressed multi-scale Gaussians. Horizon-GS (Jiang et al. 2025b) and Block-Gaussian (Wu et al. 2025b) unify aerial and ground views through hierarchical or block-wise partitioning with visibility-aware scheduling, while CrossView-GS (Zhang et al. 2025a) targets cross-view data with strong height and trajectory variations via branch-wise Gaussian reconstruction and gradient-aware fusion.

Beyond view synthesis, 3DGS has been extended to large-scale surface reconstruction. GigaGS (Chen et al. 2024b) enforces multi-view photometric and geometric consistency within a block-wise Level-of-Detail framework, yielding detailed yet scalable meshes. PG-SAG (Wang et al. 2025c) and CityGaussianV2 (Liu et al. 2024c) adapt 3DGS and 2DGS for urban modeling through semantic grouping, edge-aware geometric losses, and redesigned densification, achieving strong compression with high geometric accuracy. ULSR-GS (Li et al. 2025c) further tailors 2DGS for oblique-view mapping using point-to-patch partitioning and multi-view consistency, obtaining survey-grade surfaces at competitive cost. Collectively, these methods

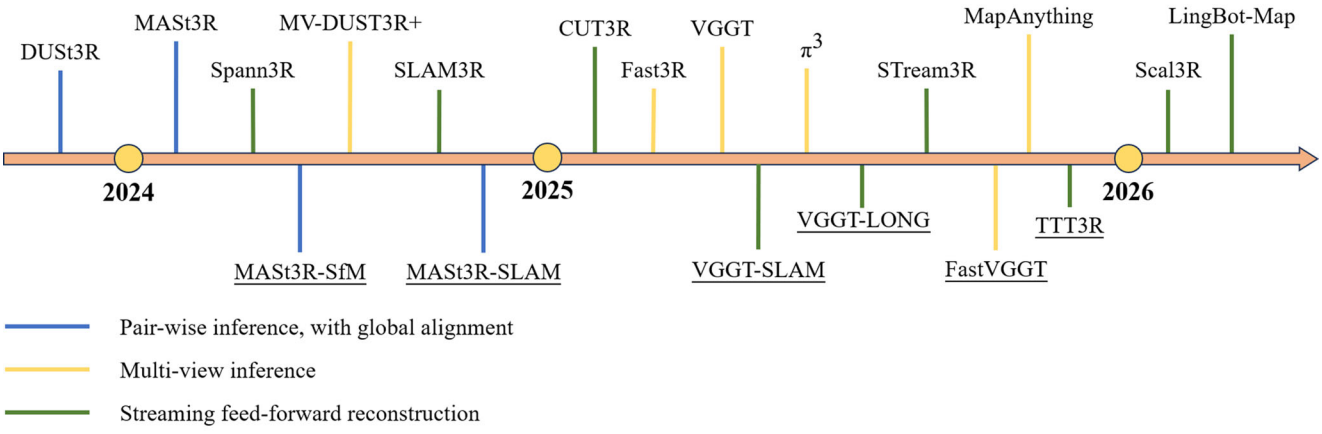

Fig. 6 Timeline of some representative feed-forward 3R methods (until April 2026). The method marked with an underline indicates that this method does not require training

show that 3DGS can move from view synthesis to accurate, scalable, large-scene reconstruction, though a direct comparison with traditional photogrammetric pipelines is still missing.

4.2.3 Feed-Forward Generalizable Methods

With the rapid development of vision foundation models, recently, the field of 3D reconstruction has experienced a paradigm shift—from iterative optimization to an end-to-end inference manner. Conventional pipelines typically decompose the 3D reconstruction process into a sequence of sub-tasks, as explained in the previous sections. Although these sub-tasks have strong interpretability, they are practically computation-intensive and susceptible to error accumulation across sub-tasks. Recently, feed-forward 3D reconstruction frameworks, exemplified by DUSt3R (Wang et al. 2024d), have demonstrated the power to infer both camera poses and scene geometry from a set of unconstrained images within a single forward pass of a pre-trained neural network; notably, visually reasonable reconstruction can be generated even in the case of image pairs with extreme baselines. By internalizing geometric reasoning into a Transformer-based backbone, these approaches learn implicit geometric priors in a fully data-driven manner. Consequently, they eliminate the need for complex iterative optimization, achieving totally end-to-end reconstruction with higher robustness—representing a significant breakthrough in the field of 3D reconstruction. Inspired by DUSt3R, many follow-up works have been proposed in the last two years, and the timeline is shown in Fig. 6.

DUSt3R is the first fully end-to-end 3D reconstruction framework, which innovatively reformulates the two-view reconstruction problem as a point-map regression task. For multiple views, DUSt3R employs a global alignment optimization to register all predicted point maps into a consistent global coordinate frame. This pioneering formulation establishes the foundation for feed-forward 3D reconstruction and opens a new paradigm in this field. However, its global optimization step remains computationally expensive and limits scalability. MASt3R (Leroy et al. 2024), built upon the DUSt3R network, introduces an additional feature matching head, an extra matching loss, and leverages larger-scale training data to improve matching precision and geometric consistency. Derived from this framework, MASt3R-SfM (Duisterhof et al. 2025) employs the MASt3R encoder for image retrieval and sparse scene graph construction to accelerate the reconstruction, but still relies on pairwise inference followed by global alignment. Similarly, MASt3R-SLAM (Murai et al. 2025) integrates MASt3R's reconstruction as a prior and incorporates local point-map fusion and loop-closure detection to maintain global consistency. While these methods mark an important step toward end-to-end reconstruction, they still depend on pairwise inference and computation-expensive global alignment, which severely limits the scalability of feed-forward approaches.

To move beyond the paradigm of pairwise inference and global alignment toward larger-scale scene reconstruction, several subsequent works explore the way to process multi-view images in arbitrary order within a single forward inference. MV-DUSt3R (Tang et al. 2025) introduces a multi-view decoder that directly processes multiple images within a single feed-forward pass. It predicts all point maps in a shared reference coordinate system, thereby removing the need for global alignment. On the other hand, Fast3R (Yang et al. 2025) encodes each image to a set of patch features and passes the concatenated patch features to a fusion transformer that performs all-to-all self-attention to provide the full context of all images. Thereby, it reasons simultaneously and jointly over all frames without assumption of the image order. In the meanwhile, VGGT (Wang

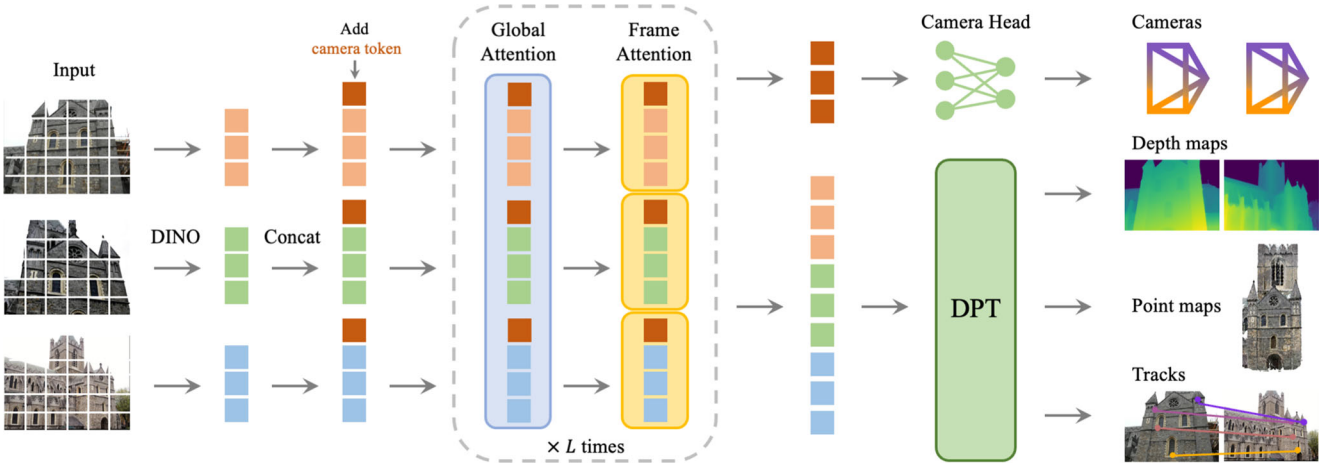

Fig. 7 Architecture of Visual Geometry Grounded Transformer (VGGT). VGGT encodes multiple input images into visual tokens, adds camera tokens, and alternates global and frame-wise attention for multi-view geometric reasoning. The model outputs camera parameters, depth maps, point maps, and point tracks through task-specific heads. Source: (Wang et al. 2025a)

et al. 2025a) represents a major breakthrough in feed-forward 3D reconstruction. It tokenizes all the input images and incorporates camera tokens to jointly infer both camera poses and point maps via alternating frame-wise and global self-attention layers (the architecture of VGGT is shown in Fig. 7). Benefiting from its well-designed network architecture, VGGT is capable of handling up to hundreds of views and simultaneously producing depth maps, point clouds, camera parameters, and 3D trajectories in a single forward pass. Despite its remarkable performance and generalization capability, its high memory consumption still limits its scalability to very large-scale scenes, which are very common in the field of photogrammetry. To reduce VGGT's computational burden caused by dense global attention and accumulated errors, FastVGGT (Shen et al. 2025a) introduces a token-merging mechanism for 3D tasks, which efficiently reduces redundant computation and GPU memory consumption without retraining, while maintaining reconstruction quality. Following the success of the VGGT, several recent works explore more generalizable feed-forward 3R architectures. Unlike previous approaches that depend on a fixed reference view to determine the coordinate system, π^3 (Wang et al. 2025f) employs a fully permutation-equivariant architecture that predicts affine-invariant camera poses and scale-invariant local point maps without any designated reference frame, thus achieving inherent robustness to image input order. Furthermore, targeting heterogeneous inputs, MapAnything (Keetha et al. 2025) supports arbitrary numbers of images and optional geometric priors (such as poses, intrinsics, or depth). Via a factorized scene representation, it enables scalable, decoupled multi-view reasoning and directly outputs metric-scale 3D reconstructions.

Another direction for improving feed-forward 3R methods focuses on streaming reconstruction from sequential image inputs. Spann3R (Wang and Agapito 2025) addresses this issue by introducing a spatial memory module to store previously inferred 3D information, which is then used to predict the global point map for upcoming frames. This memory-based design eliminates the need for explicit global optimization and significantly accelerates inference. However, its frame-by-frame prediction strategy leads to cumulative drift over long sequences. On the other hand, CUT3R (Wang et al. 2025b) adopts a stateful recurrent model that continuously updates its latent state with new input image, integrating information to generate globally aligned point maps in a single forward pass. Nevertheless, it suffers from memory decay and performance degradation as the image sequence length increases. Different from these methods, SLAM3R (Liu et al. 2025c) employs a dual-level hierarchical architecture comprising Image-to-Point (I2P) and Local-to-World (L2W) modules. It predicts local point clouds within a sliding reference window and incrementally fuses them into a consistent global coordinate system. By jointly processing multiple frames per window, SLAM3R effectively reduces drift and achieves efficient feed-forward reconstruction without explicit global optimization. Some subsequent works also further extend VGGT to streaming reconstruction. VGGT-SLAM (Maggio et al. 2025) introduces SL(4) manifold optimization to resolve projection ambiguities arising from uncalibrated cameras, achieving consistent reconstruction across submaps that cannot be aligned by simple Sim(3) transformations. While, VGGT-LONG (Deng et al. 2025) extends VGGT to long sequences via block-wise alignment, processing overlapping chunks, performing robust sub-block alignment, and applying loop-closure correction. This design scales feed-forward reconstruction to kilometer-level outdoor scenarios, significantly improving scalability at the cost of slower inference speed.

Table 2 Photogrammetric comparison of feed-forward generalizable methods.

Family	Input	Scalability (in 24G GPU)	Memory Complexity	Global Consistency	Runtime (in 4090GPU)	Online Capability	Main challenge
Pair-wise inference, with global alignment	Unordered images	Tens	Very High	Rely global alignment	0.5–1 FPS	Offline	Expensive global alignment
Multi-view inference	Unordered images	Hundreds	High	Strong	1–10 FPS	Offline	Limited scalability
Streaming feed-forward reconstruction	Ordered images	Thousands	Medium	Drift-prone	2–30 FPS	Online	Long sequence drift

In contrast to block-wise alignment strategies and simplistic memory mechanisms with limited scalability, STream3R (Lan et al. 2025) employs a causal attention mechanism together with a KV cache to enable long-sequence memory, effectively improving the efficiency and robustness of streaming reconstruction. Meanwhile, TTT3R (Chen et al. 2025a) proposes a plug-and-play memory update strategy that derives a closed-form learning rate from the alignment confidence between the memory state and new images, thereby dynamically balancing historical information retention and adapting to new image. This enables effective long-sequence generalization without additional training. More recently, Scal3R (Xie et al. 2026) and LingBoT-Map (Chen et al. 2026) leverage global context representation and geometric context attention, respectively, to effectively preserve long-term and critical context information during streaming reconstruction, extending streaming reconstruction to thousands of images and kilometer-scale 3D scenes.

These feed-forward methods mark a new stage in photogrammetric 3D reconstruction, shifting the focus from optimization-based pipelines to unified network inference. By enabling efficient, scalable, and geometry-aware reconstruction directly from image datasets, they create a possibility to establish a solid foundation for the development of next-generation automated feed-forward photogrammetric systems.

In photogrammetry field, some recent studies have systematically evaluated and extended feed-forward 3R methods in real-world aerial photogrammetry scenarios. (Wu et al. 2025a) conducted a comprehensive assessment of DUSt3R, MASt3R, and VGGT on conventional photogrammetric aerial image blocks (large-scale nadir-view imagery), discussing their advantages and limitations from the perspectives of camera pose estimation and dense 3D reconstruction that photogrammetrists care about. From the data perspective, AerialMegaDepth (Vuong et al. 2025) proposed a data-generation framework that fuses real ground-level images with pseudo-synthetic aerial views, and constructed a hybrid dataset to fine-tune existing feed-forward 3R models, achieving notable improvements in reconstruction quality on real aerial-to-ground image pairs. Focusing on the memory limitations encountered by feed-forward 3R methods when applied to large-scale aerial datasets, Shen et al. (Shen et al. 2025b) introduced a divide-and-conquer framework that extends the applicability of these models to larger scenes. Extensive experiments on various photogrammetric datasets are conducted to investigate their processing time, GPU memory usage, and accuracy, further explore the feasibility and potential of applying feed-forward 3R methods within the domain of photogrammetry.

According to the above description, feed-forward generalizable methods can be categorized into three types: pair-wise inference with global alignment, multi-view inference and streaming feed-forward reconstruction. Table 2 provides a detailed comparison of these three categories across several key metrics. Input indicates whether the method requires a specific ordering of images input. Scalability refers to the number of images that can be processed on a 24 GB GPU. Memory Complexity denotes the memory consumption during inference. Global Consistency evaluates the geometric consistency of the predicted point clouds. Runtime measures computational efficiency on an RTX 4090 GPU. Online Capability reflects the ability to perform real-time online processing. Finally, the main challenges of each category are also summarized.

As shown in Table 2, pair-wise inference with global alignment methods, although not sensitive to input ordering, is often constrained by expensive global alignment procedures. As a result, they typically handle only dozens of images, exhibit relatively poor runtime efficiency, and are difficult to deploy in real-time scenarios. These methods are therefore mainly suitable for object-centric small-scale scenes. Multi-view inference methods, on the other hand, already support inference over hundreds of unordered images in a single forward pass on an RTX 4090 GPU, producing results with strong global geometric consistency and relatively competitive runtime performance. Nevertheless, these methods are mostly offline and memory-intensive, which limits their scalability to larger image datasets. In contrast, streaming feed-forward reconstruction methods demonstrate significant advantages in processing speed and scalability, and are well-suited for real-time online applications. However, they are highly dependent on a fixed input order and are prone to drift over long sequences.

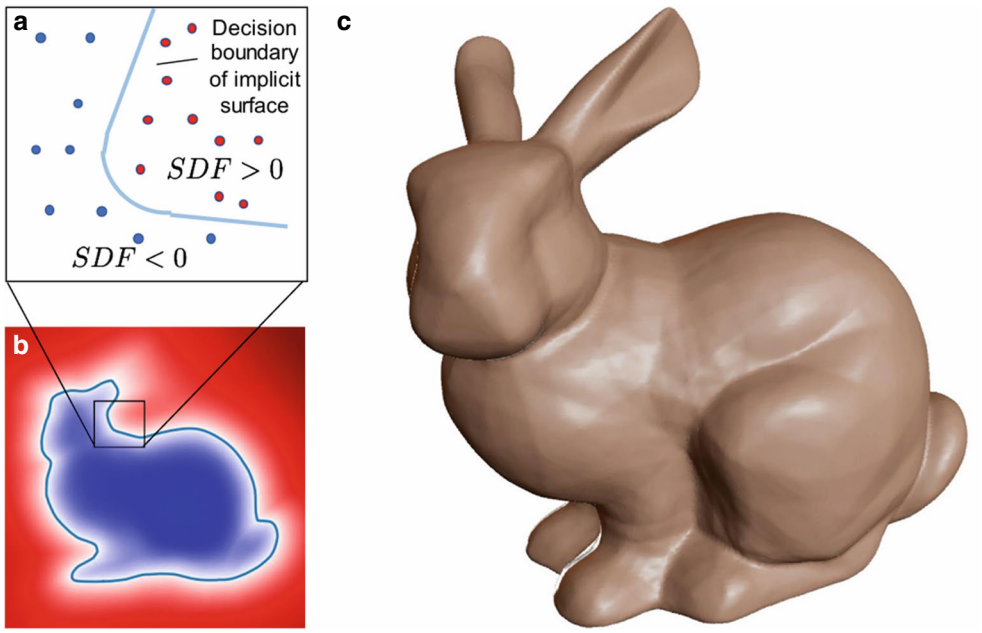

Fig. 8 The description of SDF. Source: (Park et al. 2019)

Discussion While feed-forward models such as VGGT and MapAnything demonstrate remarkable robustness to viewpoint changes and texture scarcity, their training objectives are primarily supervised through point-map or depth losses, rather than explicitly enforcing the globally consistent geometric constraints (e.g., epipolar geometry, bundle adjustment) required for photogrammetric reconstruction. As a result, despite producing visually convincing reconstructions, they do not inherently guarantee globally consistent or metrically verifiable 3D geometry across multi-view observations. This divergence raises critical concerns for photogrammetric applications requiring certifiable accuracy. *First*, unlike classical SfM/MVS, which derive geometry solely from observed pixel correspondences and epipolar constraints, foundation models implicitly leverage priors learned from massive training datasets. In low-texture, occluded, or repetitive-structure regions, this can manifest as hallucinated depth—geometrically coherent but metrically incorrect surfaces that deviate from the true scene geometry. *Second*, classical pipelines propagate analytical covariance matrices via bundle adjustment, yielding statistically rigorous confidence intervals. In contrast, learning-based confidence scores are frequently overconfident due to dataset biases and the deterministic nature of single-pass inference. Recent work attempts to attach uncertainty heads to depth predictors, but calibrating these estimates under domain shift remains an open challenge (see more detailed discussion below). *Third*, for digital twin visualization or autonomous navigation preview, visually convincing reconstructions may suffice. However, for industrial inspection, cadastral mapping, or structural deformation monitoring, hallucinated geometry and uncalibrated uncertainties are unacceptable. Future photogrammetric systems must either enforce hard geometric constraints (epipolar, scale, known intrinsics) within feed-forward architectures, or adopt hybrid verification workflows where neural outputs are rigorously validated against sparse classical control points or LiDAR references.

For safety-critical applications such as autonomous driving, medical imaging, and industrial machine vision, calibrated uncertainty in the estimated depth is indispensable: it enables risk-aware rejection of unreliable estimates, adaptive sensor fusion, and robust behavior under domain shift, without compromising the throughput required for real-time deployment. Since learning-based 3D reconstruction or depth estimation is inherently a regression problem, uncertainty corresponds to the standard deviation of the predicted depth values, a concept that represents a fundamental task in photogrammetry and is crucial for assessing measurement reliability. In this context, (Landgraf 2025) develops efficient predictive-uncertainty estimation for deep learning-based vision, including lightweight calibration and decision-layer integration for aleatoric and epistemic uncertainty components. Building on foundation models, (Landgraf et al. 2025c), synthesize uncertainty quantification with DepthAnythingV2 (Yang et al. 2024), outlining integration strategies and evaluation protocols that make large-scale monocular depth predictors more reliable in safety-critical use. Complementing this, a comparative study demonstrates multi-task uncertainty calibration across semantic segmentation and monocular depth, highlighting which estimators deliver trustworthy confidence at operational frame rates (Landgraf et al. 2025a). Finally, an efficient multi-task framework introduces lightweight uncertainty heads

for joint segmentation-depth models, achieving well-calibrated uncertainties with minimal computational overhead suitable for edge and embedded platforms (Landgraf et al. 2025b).

4.3 Surface Reconstruction

As discussed in Sect. 3.3, conventional surface reconstruction methods are commonly divided into explicit and implicit approaches. Learning-based surface reconstruction follows the same distinction, but replaces hand-crafted geometric rules or variational formulations with data-driven connectivity prediction, implicit field learning, or learned surface priors. Accordingly, this section first reviews recent learning-based explicit surface reconstruction methods, then discusses learning-based implicit methods, and finally provides a comparative discussion of conventional and learning-based surface reconstruction categories under common surface-quality criteria.

4.3.1 Explicit Methods

Beyond purely geometric formulations, recent advances in mesh generation have shifted towards learning explicit surface connectivity directly from point cloud data, PointTriNet (Sharp and Ovsjanikov 2020) and IER Meshing both employ neural networks to directly predict triangle meshes. PointTriNet utilizes a two-stage neural network: the first stage classifies the likelihood of triangle existence, while the second proposes additional candidate triangles, enabling efficient large-scale mesh generation. In contrast, IER Meshing (Liu et al. 2020) leverages intrinsic and extrinsic ratios to predict local point triplets, which form valid mesh faces, thus enhancing efficiency, particularly with sparse and noisy point clouds. However, the performance of both methods significantly degrades when processing sparse point clouds. To address this issue, MergeNet (Wang et al. 2024e) improves robustness and efficiency by predicting edges to reconstruct the mesh, particularly excelling in sparse point cloud scenarios. Finally, Point2Quad (Li et al. 2025b) extends triangle meshes to quadrilateral meshes by predicting candidate faces using local geometric features, optimizing the mesh for planarity and squareness, and offering an efficient method for quadrilateral mesh generation.

Another learning-based approach to explicit reconstruction utilizes Delaunay triangulations as a structural prior, shifting the focus from directly modeling visibility to labeling tetrahedral cells. DeepDT (Luo et al. 2021) applies this concept by learning to classify the inside/outside labels of Delaunay tetrahedra using the point cloud and its corresponding triangulation. However, it faces several challenges, including limited local geometry perception due to insufficient use of tangent plane features, ambiguous tetrahedral classification arising from the simple combination of point features, and scalability issues when dealing with large-scale data. To address these limitations, DMNet (Zhang et al. 2023) adopts a fully local approach that enables efficient reconstruction while capturing fine details. By encoding each tetrahedron and triangular facet individually, DMNet enhances local geometry understanding through a graph feature encoder and a Local Graph Iteration network. This network refines connectivity by iteratively processing one-hop node features without relying on global adjacency matrices, resulting in improved robustness and accuracy, particularly when working with sparse point clouds. Building on DMNet, DMNet+ (Zhang and Tao 2025) introduces a new graph feature encoder that integrates both global geometry and multi-scale local features into the point feature encoder. This enhancement reduces reliance on complex manual features, streamlining network input and significantly improving both the efficiency and accuracy of mesh reconstruction.

4.3.2 Implicit Methods

Recent advancements in implicit neural representations for surface reconstruction have led to significant progress, with these methods generally categorized into three main approaches: Signed Distance Fields (SDF), Occupancy Networks, and Unsigned Distance Fields (UDF).

SDF DeepSDF (Park et al. 2019) was one of the earliest methods to introduce a continuous approach for 3D surface representation. By training an autoencoder to learn the Signed Distance Field (SDF), this method bypasses traditional voxel grids, enabling high-fidelity shape modeling. It predicts SDF values from point samples, from which zero-level set surfaces are derived. The description of SDF is shown in Fig. 8. Building on this foundation, DiGS (Ben-Shabat et al. 2022) improved surface reconstruction accuracy by integrating SIREN (Sitzmann et al. 2020) and introducing a divergence penalty. This approach enables the learning of smoother, more accurate surfaces, even without the use of normal vectors, making it particularly effective for modeling complex and diverse shapes. The introduction of MVSDF (Zhang et al. 2021) further advanced the field by incorporating multi-view stereo (MVS) constraints into the SDF framework. This enhancement improved the model's ability to handle complex scene reconstructions by leveraging the consistency between multiple views, thus enhancing geometric accuracy, particularly for concave and intricate shapes. Additionally, the integration of SDF with implicit neural representations like NeRF (Mildenhall et al. 2021) led to the development of methods such as NeuS (Wang et al. 2021a) and VolSDF (Yariv et al. 2021). These

Table 3 Photogrammetric comparison of conventional and learning-based surface reconstruction categories.

Category	Fidelity	Topology	Robustness	Scalability	Primary limitation
Conventional Explicit	Detail strong	Topology controllable	Noise-sensitive	Large-scale feasible	Dense input required
Conventional Implicit	Smooth, detail-limited	Watertight strong	Gap-tolerant	Computation-heavy	Oversmoothing
Learning-Based Explicit	Local accuracy variable	Connectivity variable	Domain-sensitive	Inference efficient	Weak transferability
Learning-Based Implicit	Continuity strong	Topology flexible	Prior-sensitive	Sampling-heavy	Low interpretability

methods introduced neural rendering techniques that use volume rendering for multi-view reconstruction, improving robustness against self-occlusions and thin structures, while simultaneously enhancing surface reconstruction efficiency and quality. Finally, the combination of explicit Gaussian-based representations, such as 3DGS (Kerbl et al. 2023), with SDF in methods like GSDF (Yu et al. 2024a) and GS-SDF (Liu et al. 2025a) leveraged the strengths of Gaussian splatting for rendering alongside SDF for geometric tasks. These methods employ dual-branch networks that optimize both rendering and geometry simultaneously, addressing the quality-efficiency trade-offs observed in earlier approaches. These advancements reflect the continuous evolution of SDF for 3D reconstruction, making them more accurate, efficient, and applicable to a wide range of complex tasks.

Occupancy Network Occupancy Networks (Mescheder et al. 2019) present a novel approach to 3D reconstruction by representing surfaces as the continuous decision boundary of a neural network, avoiding the discretization of 3D space inherent in traditional voxel or mesh methods. This efficient and high-resolution representation learns the occupancy probability for each point in space, with meshes generated through a multi-resolution isosurface extraction algorithm using octree structures and the Marching Cubes algorithm. Convolutional Occupancy Networks (Peng et al. 2020) build upon this concept by incorporating convolutional encoders and implicit decoders, facilitating the integration of local geometric features and enabling the reconstruction of more complex 3D scenes. These networks also scale efficiently for large-scale reconstructions by introducing inductive biases, such as translational equivariance. To enhance reconstruction efficiency, Lightweight 3D Convolutional Occupancy Networks (Tonti et al. 2024) optimize the process for real-time applications on edge devices, balancing speed and accuracy. Extending this approach to the field of photogrammetry, (Chen et al. 2025b) applies it to building reconstruction from UAV point clouds, achieving high-efficiency results with fewer facets, which makes it ideal for urban planning and disaster assessment. (Zhang et al. 2025b) combines signed distance fields (SDF) with occupancy networks for underwater 3D reconstruction, improving accuracy in shadowed areas and addressing limitations of traditional SDF methods. For Digital Surface Model (DSM) reconstruction, Occupancy Networks play a key role. Deep-FG-DSM (Hou et al. 2024) uses deep implicit occupancy networks on UAV imagery for fine-grained DSM generation, combining high-resolution UAV images with point cloud data to produce detailed reconstructions, even with sparse data.

UDF To model general non-watertight surfaces, (Chibane et al. 2020) introduced Neural Unsigned Distance Fields (NDF), marking a significant advancement in 3D surface representation. NDF employs neural networks to predict unsigned distance fields from sparse point clouds, enabling the representation of complex shapes with arbitrary topologies. Unlike traditional methods that are limited to closed surfaces, NDF extends to open surfaces, such as walls and objects with internal structures, which were previously challenging to model. This approach overcomes resolution limitations and improves the ability to represent diverse topologies, offering a more flexible solution for 3D reconstruction. Building upon NDF, DUDE (Venkatesh et al. 2020) further refined UDF for high-fidelity surface reconstruction. It employs a disentangled representation model, enabling the use of unsigned distance embeddings to accurately capture arbitrary shapes, including both open and closed surfaces. To further enhance geometric accuracy, (Ren et al. 2023) introduced GeoUDF, a geometry-guided learning approach that improves UDF precision by explicitly modeling the geometric structure of point clouds. The integration of edge-based marching cube techniques facilitated detailed surface generation, overcoming the limitations of previous methods in handling open surfaces. Further refinement came with 2S-UDF (Deng et al. 2024), a two-stage UDF learning method that ensures surface reconstruction quality by addressing occlusions and biases during the learning process. Most recently, (Xu et al. 2025) introduced DEUDF, which improves upon existing UDF learning techniques by addressing gradient issues and enhancing surface detail extraction through integrated alignment strategies. These advancements demonstrate the ongoing evolution of UDF methods, improving both the accuracy and applicability of 3D surface reconstruction across various domains.

4.3.3 Discussion

The preceding subsections show that learning-based surface reconstruction inherits the explicit-implicit distinction of conventional surface reconstruction, while introducing data-driven prediction mechanisms. To clarify the relative strengths and limitations of different surface reconstruction categories, Table 3 compares conventional explicit, conventional implicit, learning-based explicit, and learning-based implicit methods under a common set of surface-quality criteria. The comparison uses short evaluative labels. Fidelity describes the ability to preserve local details, smoothness, or surface continuity; topology refers to mesh controllability, watertightness, connectivity stability, or topological flexibility; robustness reflects sensitivity to noise, gaps, sparsity, or domain shift; scalability indicates computational feasibility for large point clouds or large scenes; and the final column summarizes the primary limitation of each reconstruction category.

As summarized in Table 3, the central trade-off is between explicit geometric controllability and implicit surface continuity. Conventional explicit methods are advantageous when sharp local details and controllable mesh structures are required, but they depend on dense and reliable samples. Conventional implicit methods are more suitable for producing smooth and watertight surfaces from incomplete data, although they may weaken sharp structures and increase computational cost. Learning-based explicit methods can improve local surface assembly through data-driven prediction, but their accuracy and connectivity may vary under domain shift. Learning-based implicit methods provide strong continuity and flexible topology through learned fields, but their dependence on learned priors makes metric interpretation and reliability assessment more difficult. Therefore, learning-based surface reconstruction offers important opportunities for sparse, noisy, or incomplete observations, but its photogrammetric reliability still requires careful validation in terms of metric accuracy, generalization, and interpretability.

5 Discussion

In this section, we first provide a comparative discussion that contrasts conventional and learning-based paradigms in terms of feature matching, dense reconstruction quality, scalability, and uncertainty estimation. Then, we seek to delineate the current key research frontiers and identify persistent challenges.

5.1 Comparison of Conventional and Learning-Based Techniques

Both the sparse and dense reconstruction processes have experienced a paradigm shift as they progress into this deep learning era. Although existing image-based 3D reconstruction systems, e.g., either open-source or commercial software packages, still rely on conventional methods, deep learning-based approaches have shown significant advancement in many studies in certain conditions. Therefore, this subsection will introduce commonly used conventional and learning-based methods for both tasks together with their respective advantages and limitations.

5.1.1 Sparse Feature Points Extraction and Matching

Conventional Methods Sparse matching is a critical step in image-based 3D reconstruction. It provides reliable correspondence across multiple convergent images to build reliable observations for incremental relative orientation (or SfM), and subsequently the bundle adjustment. A classic two-step process for feature correspondences first detects interest points and then computes a descriptor per point for use in matching these extracted points across different images. SIFT (Lowe 2004) and its similar variants, including SURF (Speeded Up Robust Features) (Bay et al. 2008), FAST (Features from Accelerated Segment Test) (Rosten et al. 2010), BRIEF (Binary Robust Independent Elementary Features) (Calonder et al. 2010), ORB (Oriented FAST and Rotated BRIEF) (Rublee et al. 2011), and FREAK (Fast Retina Keypoint) (Alahi et al. 2012), have been the most widely used approach due to its robustness, particularly in aerial images from crewed air surveys and unmanned aerial vehicles (UAV). Despite it no longer ranks among the top in the leadboard in computer vision benchmarks, its relatively low requirement on hardware (by today's standard), effectiveness, generalization capability (classic method, no training biases) on aerial and close-range photogrammetric dataset, put the SIFT one of the most implemented algorithm in many 3D reconstruction/modeling software. However, there are still many scenarios where the conventional approaches (using handcrafted features) fail to perform, such as in processing data captured at varying lighting environment, large view perspective differences, texture-less images etc. For example, in processing satellite stereo with consistent lighting (images with the same-day capture), and small intersection angles, SIFT can still be very competitive. However, when processing multi-date satellite images or images in machine vision applications where large angle differences, lighting/shadowing variations, and transient or specular objects exist, it often fails to provide sufficient matches for further processing.

Learning-Based Approaches Compared to conventional feature detector/matchers, the deep learning-based approaches learn a higher-level feature representation using examples, which have shown promising results in some of the challenging cases. This learning process can present at any of the keypoint detection and descriptor level in a two-step process, or it can be in an end-to-end fashion (from image to matches) as so called detector-free approaches. Hence, the learning-based methods can be further divided into the learning-based detector-descriptor approaches, and detector-free methods, described below.

Learning-Based Detectors/Descriptors Both the detector and descriptors can be learnt in a neural network, in both cases, the networks intake an input image and output a pixel-grid of likelihood, indicating the potential locations of the keypoint and thus subsequent/parallel networks in outputting the potential feature descriptors associated with each pixel (Yi et al. 2016; Ono et al. 2018; DeTone et al. 2018; Dusmanu et al. 2019; Revaud et al. 2019; Tyszkiewicz et al. 2020). Among the many algorithms developed in the past few years, the line of work in SuperPoint and SuperGlue, including its variant LightGlue, stood out among the most highly cited (DeTone et al. 2018; Sarlin et al. 2020; Lindenberger et al. 2023). The former uses parallel networks for both keypoint detection and description, followed by descriptor matching; the latter uses subsequent networks, especially graph-neural networks as a subsequent network for matching keypoints, fully utilizing the 2D topological relationship among points and its near and far neighbors. In fact, the SuperGlue and LightGlue line of work, by learning attentions among points (instead of image contexts), gain very generalization capabilities, and have been gradually incorporated into contemporary structure-from-motion pipelines and exhibit rather robust performance (Sarlin et al. 2020, 2019; Lindenberger et al. 2023; Schönberger and Frahm 2016). In addition to these seminal works in this category, there are many other variants of these approaches published in recent years, non-exhaustively listed here: LIFT (Learned Invariant Feature Transform) (Yi et al. 2016), LF-Net (Learning Local Features from Images) (Ono et al. 2018), D2-Net (Trainable CNN for Joint Description and Detection) (Dusmanu et al. 2019), R2D2 (Repeatable and Reliable Detector and Descriptor) (Revaud et al. 2019), and DISK (Learning Local Features with Policy Gradient) (Tyszkiewicz et al. 2020).

Detector-Free Feature Matching Detector-free methods bypass the need for performing a separate feature detection stage, while directly output feature correspondences given an input of an image pair. This type of approaches typically utilizes the Transformer architecture (Vaswani et al. 2017) to learn attentions (both self-attention and cross-attention) among potential candidate points. In the case of detector-free methods, these potential candidates are full pixel grids. Due to the large demand for GPU memory, this type of approach also limits the resolution (by down-sampling high-resolution images) of the input images for training and testing. The detector-free methods, due to that it takes all pixels as candidates with long-range dependence through the Transformer architecture, it naturally performs very well for low-texture regions, with a great success in data such as 3D reconstruction of indoor areas, or on satellite images of the field where repetitive textures or low textures areas are dominant. Representatives detector-free methods LoFTR (Local Feature TRansformer), ASpanFormer (Adaptive Span Transformer), and DKM (Dense Kernelized Feature Matching) (Sun et al. 2021; Chen et al. 2022b; Edstedt et al. 2023). In the respective litereature, it was reported that their accuracy, as compared to the conventional methods or learning-based detector-descriptor methods (Lowe 2004; Sarlin et al. 2020), such an approach excel in some benchmarks at a notable margin. For example, LoFTR outperforms SuperGlue on MegaDepth with an AUC at 10 degrees of 69.19 versus 61.16, that is a 13% point gain (Sun et al. 2021).; Aspanformer further raises MegaDepth performance to 71.5 AUC at 10 degrees, about +2.3 points over LoFTR and +10.3 over SuperGlue. Dense Kernelized Feature Matching (DKM) sets a new state of the art on MegaDepth 1500, improving AUC at 5 degrees by +4.9 over the best sparse method ASpanFormer and by +8.9 over the best prior dense method, and also shows +3.6 AUC at 3 pixels on HPatches homography (Edstedt et al. 2023). However, the detector-free methods face two major limitations. First, the high GPU memory demand constrains use at full resolution images, especially for large frame images like aerial or satellite imagery; second, detector-free methods require a pair of image as an input and optimize correspondences for that specificalpairs rather than enforcing multi-view consistency (Chen et al. 2022b). In image-based 3D reconstruction, the quality of the geometry strongly relies on multi-ray points (more than 2). In such a context, although detector-free methods can generate relatively dense correspondences, these may not yield stable multi-ray points in structure from motion, reducing feature track length (He et al. 2024). Hence, these two aspects in detector-free methods still have considerable rooms for improvement.

5.1.2 Dense Reconstruction

The process of dense reconstruction computes the per-pixel correspondence across a pair or multiple images for triangulating 3D points for surface reconstruction. Depending on the number of input images for feature correspondence, dense reconstruction has two types of basic processing, (1) stereo matching, in which two images are densely

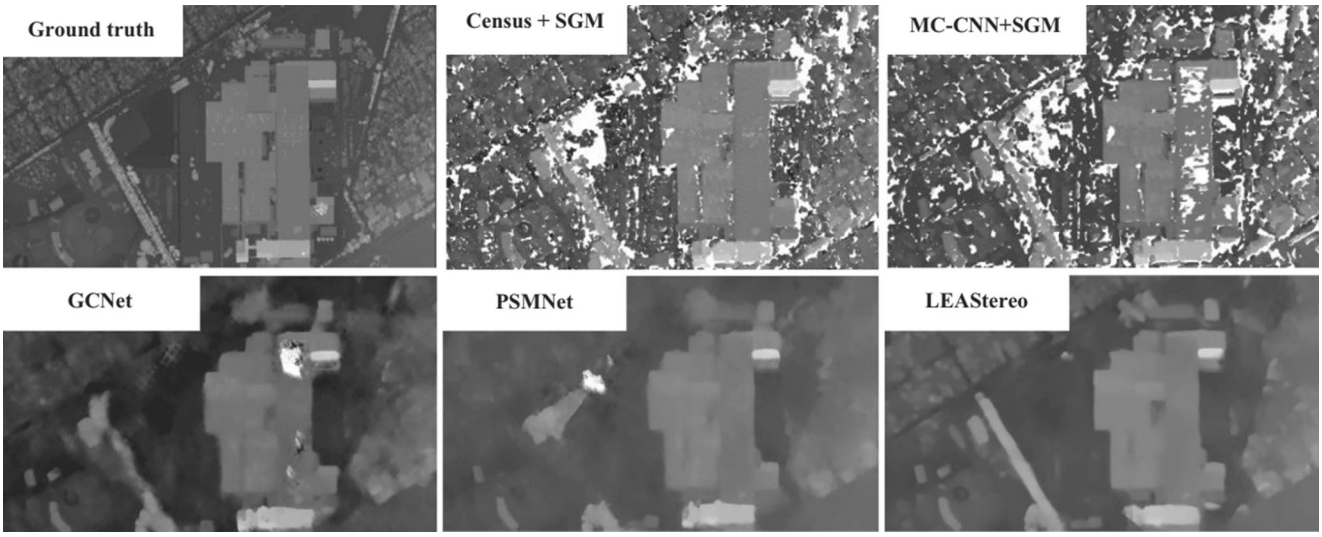

Fig. 9 Sample DSMs from different stereo dense image matching algorithms. White pixels indicate missing values. Source: (Albanwan and Qin 2022a)

matched at one time, and (2) multi-view stereo matching, in which more than two images are densely matched simultaneously.

Binocular Stereo Matching Classical stereo methods firstly take a pair of images, and epipolarly rectify them to convert the 2D pixel search problem into 1D (Hartley and Zisserman 2003). They then build a cost volume by testing many disparities per pixel and take the disparity with the best score after an optimization, often imposing a regularization for smoothness. Among classic approaches, some rely on local matching such as normalized cross-correlation, while others add global optimization, such as graph-cut optimization (Boykov et al. 2001). Semi-global matching (Hirschmüller 2008) stands in between by optimizing along on linear sequence of pixels and has converged as a standard baseline due to its balance of speed and accuracy (Scharstein and Szeliski 2002). Learning-based methods either learn to compare pixel patches to improve the patch-level similarity score, such as the MC-CNN (Matching Cost with Convolutional Neural Networks) (Zbontar and LeCun 2015), or learn contextual similarity end-to-end. Examples of such networks include PSMNet (Pyramid Stereo Matching Network) (Chang and Chen 2018), GCNet (End-to-End Learning of Geometry and Context for Deep Stereo Regression) (Kendall et al. 2017), LEAstereo (Learning Efficient Architecture Stereo) (Cheng et al. 2020b), along with other variants such as GA-Net (Guided Aggregation Network) (Zhang et al. 2019) and DeepPruner (a real-time stereo matching model with differentiable PatchMatch) (Duggal et al. 2019). Figure 9 shows an example of a comparison of both the conventional and deep learning method applied on a satellite stereo pair, extracted from a comparative work (Albanwan and Qin 2022a). It clearly shows that the SGM and the MC-CNN underperform the deep learning methods by producing a large number of gapped pixels, while the results of three end-to-end methods in the second row, clearly show their advantage in generating a coherent and gap-free disparity map, making it a reliable reference for DSM generation (Albanwan and Qin 2022a, b). However, like many other DL based methods, this type of end-to-end learning approaches suffers strongly from the generalization problems, being that models learned from data of one geographical region can be hardly applied to another one with large land-scape difference. For example, Fig. 10. shows DSMs that indicate the end-to-end algorithms, specifically LEAStereo, GCNet and PSMNet, do not generalize well across different scenes (Albanwan and Qin 2022a).

Multi-View Stereo Matching Multi-view stereo (MVS) refers to the process of simultaneously matching correspondences across multiple images, while some of the literature also classifies pair-wise stereo matching followed by fusion as MVS. In this section, we refer MVS as the former. MVS utilize the idea of sweeping planes (Collins 1996; Werner and Zisserman 2002) for constructing cost-volumes to support finding the solution that achieves photoconsistency across multiple images. This process, with stronger geometric support, typically produces cleaner point clouds with higher accuracy, while due to the increased chances of occlusions when considering multiple images, the resulting point clouds often contain more gaps. The optimization of an MVS process, like stereo matching, often involves smooth regularization to reduce noisy output by assuming surface smoothness. Representative and conventional methods are the class of patch-match based methods (Furukawa

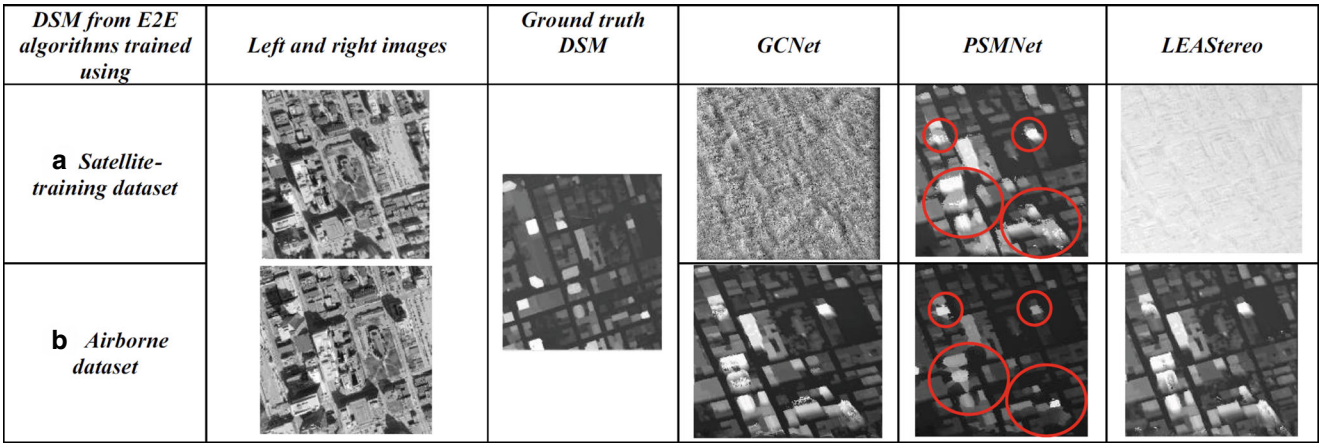

Fig. 10 Sample DSMs generated by end to end (E2E) algorithms trained and tested on different datasets. GCNet and LEAStereo produce inconsistent DSMs, while PSMNet is more consistent but still shows noise and outliers (red circles). Source: (Albanwan and Qin 2022a)

and Ponce 2010; Bleyer et al. 2011; Schönberger et al. 2016) which utilize variants of belief propagations to regularize surface smoothness on top of photoconsistency. This class of method, due to memory efficiency and simplicity, remains one of the primary lines of work in practice for handling full-resolution images. The deep learning variant of MVS primarily focuses on extracting more robust and representative image-level features for constructing more informed cost-volumes, where the image-level features are either pretrained through existing datasets with ground truth or through self-supervision (Wang et al. 2024a; Laga et al. 2022). Examples of deep learning-based include are MVSNet, CasMVSNet, UCSNet etc (Yao et al. 2018; Gu et al. 2020; Cheng et al. 2020a), and these variants mainly differ based on training schemes, level of supervision and network architecture etc.

Feedforward Reconstruction Models The feedforward reconstruction methods, as introduced in Section 4.2.3, aim to fit both the sparse and dense matching into a deep learning model, mostly using the Transformer architecture. Examples of such models include DUSt3R, MASt3R, and VGGT, which learn from a large amount of data, typically from millions of images, with specific training schemes using a mix of data with and without groundtruth. It has shown a remarkable advantage in reconstructing complete 3D point clouds from very difficult cases such as data with drastically different viewpoints and even works well with single-view images. A well-trained model can process dozens of images or more in a single forward pass. Due to the nature of the image-based Transformer architecture being memory-demanding, current feed-forward models typically operate on images resized to about 512 pixels on the longest side. Recent models scale the number of input images from a few to hundreds or even thousands, yet they train and infer at about 512 pixels on the longest side (Yang et al. 2025). When it comes to processing real-scale photogrammetric datasets at fullest resolution, traditional approaches are still the top choice. A comparison is shown in Fig. 11 using the UseGeo UAV dataset (Nex et al. 2024), in which the feedforward reconstruction methods show higher completeness but lower accuracy, with significantly incorrect camera poses. Newer models, like VGGT, can be provided with hundreds of views, while older models like DUSt3R and MASt3R can only process a dozen of them (Wu et al. 2025a). Hence, like many other learning-based methods, the scalability of these models to robustly, accurately reconstruct 3D using the images at the fullest resolution is still a major challenge for this type of approach. Additionally, the level of uncertainty of either the pose estimation or 3D point clouds, defined in traditional means (circular or linear errors), can be extremely challenging to characterize, lifting the cost of producing trustworthy point clouds for applications in geospatial engineering and machine vision.

5.2 Future Research Directions

The past five years have witnessed the exciting progress of using data-driven approaches in image-based 3D reconstruction, which have arguably demonstrated their ability to quickly, completely reconstruct scenes from relatively sparse and low-quality images. These advancements have certainly enabled novel applications for real-time 3D reconstruction in rather complex scenes at which conventional approaches fail, while it largely holds limitations for classic photogrammetric 3D tasks and applications such as high-precision mapping with certifiable accuracy. Hence, we consider that the following research paths can be of interest to further reconcile existing limitations in this field.

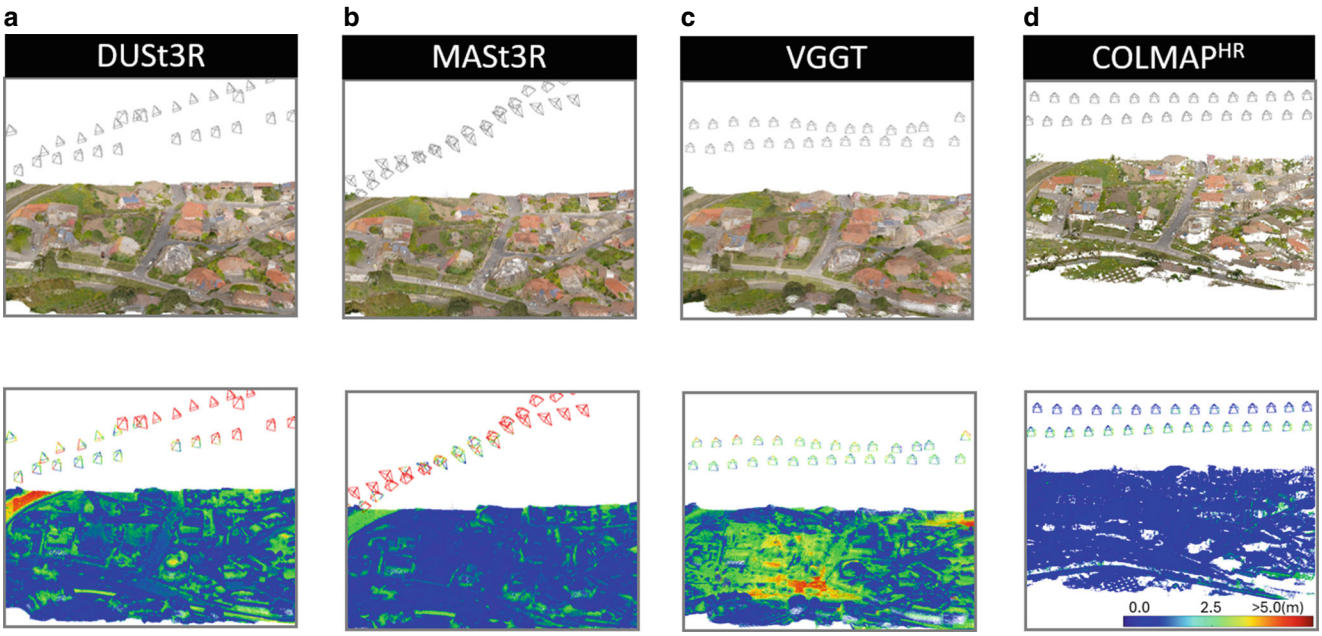

Fig. 11 Results from DUSt3R (**a**), MASt3R (**b**), VGGT (**c**), and COLMAPHR (**d**), where COLMAPHR denotes COLMAP results obtained from high-resolution inputs. The top row presents the dense point cloud and the estimated camera poses (represented in gray), while the bottom row displays the error map, comparing the results to ground truth LiDAR data. Camera poses are color-coded based on their distance from the ground truth. Source: (Wu et al. 2025a)

5.2.1 Deeper Integration of Conventional and Deep Learning Approaches

Despite significant progress in incorporating geometric principles into learning-based reconstruction, the degree of integration between data-driven models and classical photogrammetric constraints varies substantially among different approaches. For example, learning-based MVS methods such as MVSNet (Yao et al. 2018) explicitly incorporate camera geometry through differentiable plane-sweep cost volumes, while MVSTER (Yao et al. 2022) further exploits epipolar geometry through transformer-based feature aggregation. These methods demonstrate that learning-based reconstruction can benefit from the combination of learned representations and physical geometric constraints. However, some recent highly over-parameterized end-to-end architectures and foundation models tend to rely more heavily on learned priors and implicit geometric reasoning, where explicit projective constraints and camera models are relatively weaker. Therefore, further integrating physics-based concepts, including perspective projection, epipolar constraints, and accurate camera models that capture real-world imaging characteristics, into future data-driven reconstruction frameworks remains an important direction for improving metric fidelity and reducing errors caused by model bias.

5.2.2 Scalability of 3D Reconstruction Algorithms

Existing DL methods are data-hungry and also require huge graphics memory, especially transformer-based architecture. The demand for high-end hardware limits the existing methods when scaling at mega-pixel, high-resolution images. Thus, developing lightweight models that handle high-resolution images are a promising direction for DL algorithms.

5.2.3 Improving Generalization Capabilities

DL approaches with large/complex models can learn contextual correspondences beyond photo-consistency, while, as with general DL models, they suffer from generalization issues when the datasets used for training are distinct from the datasets used for deployment. Therefore, it is important to investigate methods and strategies that reduce the risk of biased predictions by either 1) developing methods with stronger generalization capability or 2) developing approaches that predicts model and data biases to enable application with confidence. This challenge becomes particularly evident in machine vision scenarios where input images often differ significantly from those typically used to pretrain DL models.

5.2.4 Reliable Real-Time Performance

Thanks to fast-forward inference enabled by powerful GPUs, DL models achieved remarkable success in a wide range of real-time tasks, such as object detection, semantic segmentation, and geometry reasoning. However, many practical photogrammetric applications, including real-time 3D reconstruction, online measuring, and dynamic feedback system (Lou et al. 2025), demand even higher levels of reliability and robustness in real-time performance. To address this need, it is promising to either develop task-specific DL architectures tailored to these demanding scenarios or to fine-tune existing lightweight models to enhance their accuracy, stability, and real-time responsiveness.

5.2.5 Uncertainty Estimation

Uncertainty estimation for conventional approaches is well covered using first-order statistics derived from analytical formulations of image-based 3D reconstruction. Metrics such as covariance matrices for poses and 3D point clouds are pivotal in high-precision mapping. However, although DL-based methods naturally provide confidence scores, these scores are frequently considered as being overly confident because of over-parameterized networks used to fit image-based 3D reconstruction problems. Therefore, performing reliable uncertainty estimation for DL-based 3D reconstruction is of particular interest. While first very promising approaches exist for specific applications, the problem remains unresolved in practice in general, especially under domain shift and real-time constraints. Strong baselines such as Monte-Carlo dropout and deep ensembles typically require multiple forward passes or maintaining several model instances, incurring substantial memory and computation overhead. Hence, methods that deliver well-calibrated depth uncertainties in a single pass with negligible latency and footprint are still needed for embedded/edge deployment and high-throughput reconstruction.

6 Conclusion

This review surveys recent advances in image-based 3D reconstruction from a photogrammetric standpoint, bridging conventional pipelines (SfM-MVS-Surface Reconstruction) with emerging learning-based paradigms. We first revisit the classical workflow, and summarize recent developments that enhance robustness, scalability, and computational efficiency, particularly for large-scale image collections. The discussion then turns to learning-based progress in SfM, MVS, and surface reconstruction, with a focus on major research frontiers including NeRF, 3D Gaussian Splatting (3DGS), and feed-forward 3D reconstruction methods up to November 2025.

We further identify key open challenges that are critical for practical deployment, including reliable uncertainty estimation for deep models, scaling to hundreds or thousands of high-resolution images under realistic GPU constraints, and achieving dependable real-time performance. Finally, we outline directions toward next-generation automated photogrammetric systems that synergistically integrate data-driven learning with explicit geometric priors and rigorous accuracy evaluation.

Author contributions We describe contributions to the paper using the CRediT taxonomy. X.W. and T.W. contributed equally to this work. Conceptualization: X.W. and M.U.; Writing — original draft: X.W. (1, 4.1, 6), T.W. (2.3, 3.3, 4.3), M.H. (2.2), Y.Y. (4.2), Z.S. (4.2), R.Q. (5), and M.U. (2.1, 2.2, 3.1, 3.2) ; Writing — review & editing: X.W., T.W., M.H., Y.Y., Z.S., Z.Z., R.Q., and M.U.

Funding This research was supported by National Natural Science Foundation of China under Grant 42301507.

Open Access funding enabled and organized by Projekt DEAL.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

Agarwal S, Snavely N, Simon I, Seitz SM, Szeliski R (2009) Building Rome in a day. Proc IEEE Int Conf Comput Vis. https://doi.org/10.1109/ICCV.2009.5459148

Alahi A, Ortiz R, Vandergheynst P (2012) FREAK: Fast retina keypoint. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. https://doi.org/10.1109/CVPR.2012.6247715

Albanwan H, Qin R (2022a) A comparative study on deep-learning methods for dense image matching of multi-angle and multi-date remote sensing stereo-images. Photogramm Rec 37(180):385–409. https://doi.org/10.1111/phor.12430

Albanwan H, Qin R (2022b) Fine-tuning deep learning models for stereo matching using results from semi-global matching. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences V-2-2022:39–46. https://doi.org/10.5194/isprs-annals-V-2-2022-39-2022

Amenta N, Bern M, Kamvysselis M (1998) A new Voronoi-based surface reconstruction algorithm. In: Proceedings of the 25th annual conference on Computer graphics and interactive techniques, pp 415–421

Amenta N, Choi S, Dey TK, Leekha N (2000) A simple algorithm for homeomorphic surface reconstruction. In: Proceedings of the sixteenth annual symposium on Computational geometry, pp 213–222

Barath D, Mishkin D, Eichhardt I, Shipachev I, Matas J (2021) Efficient initial pose-graph generation for global SfM. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 14541–14550

Bay H, Ess A, Tuytelaars T, Van Gool L (2008) Speeded-up robust features (SURF). Comput Vis Image Underst 110(3):346–359. https://doi.org/10.1016/j.cviu.2007.09.014

Ben-Shabat Y, Koneputugodage CH, Gould S (2022) DiGS: Divergence guided shape implicit neural representation for unoriented point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 19323–19332

Bernardini F, Mittleman J, Rushmeier H, Silva C, Taubin G (2002) The ball-pivoting algorithm for surface reconstruction. IEEE transactions on visualization and computer graphics 5(4):349–359

Bleyer M, Rhemann C, Rother C (2011) PatchMatch stereo—stereo matching with slanted support windows. BMVC 11:1–11

Boykov Y, Veksler O, Zabih R (2001) Fast approximate energy minimization via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(11):1222–1239. https://doi.org/10.1109/34.969114

Brown DC (1971) Close-range camera calibration. Photogrammetric Engineering 37(8):855–866

Brynte L, Iglesias JP, Olsson C, Kahl F (2024) Learning structure-from-motion with graph attention networks. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp 4808–4817

Cai R, Tung J, Wang Q, Averbuch-Elor H, Hariharan B, Snavely N (2023) Doppelgangers: Learning to disambiguate images of similar structures. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 34–44

Calonder M, Lepetit V, Strecha C, Fua P (2010) BRIEF: Binary robust independent elementary features. In: Daniilidis K, Maragos P, Paragios N (eds) Computer Vision—ECCV 2010. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 778–792

Carr JC, Beatson RK, Cherrie JB, Mitchell TJ, Fright WR, McCallum BC, Evans TR (2001) Reconstruction and representation of 3d objects with radial basis functions. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques, pp 67–76

Chang JR, Chen YS (2018) Pyramid stereo matching network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Chang D, Božič A, Zhang T, Yan Q, Chen Y, Süsstrunk S, Nießner M (2022) RC-MVSNet: Unsupervised multi-view stereo with neural rendering. In: European conference on computer vision, Springer, pp 665–680

Chen A, Xu Z, Geiger A, Yu J, Su H (2022a) TensoRF: Tensorial radiance fields. In: European conference on computer vision. Springer, pp 333–350

Chen D, Li H, Ye W, Wang Y, Xie W, Zhai S, Wang N, Liu H, Bao H, Zhang G (2024a) PGSR: Planar-based Gaussian splatting for efficient and high-fidelity surface reconstruction. IEEE Transactions on Visualization and Computer Graphics

Chen H, Luo Z, Zhou L, Tian Y, Zhen M, Fang T, McKinnon D, Tsin Y, Quan L (2022b) ASpanFormer: Detector-free image matching with adaptive span transformer. In: Avidan S, Brostow G, Cissé M, Farinella GM, Hassner T (eds) Computer Vision—ECCV 2022. Springer Nature Switzerland, Cham, pp 20–36

Chen J, Ye W, Wang Y, Chen D, Huang D, Ouyang W, Zhang G, Qiao Y, He T (2024b) GigaGS: Scaling up planar-based 3d Gaussians for large scene surface reconstruction. arXiv preprint https://arxiv.org/abs/2409.06685

Chen LZ, Gao J, Chen Y, Cheng KL, Sun Y, Hu L, Xue N, Zhu X, Shen Y, Yao Y et al (2026) Geometric context transformer for streaming 3d reconstruction. arXiv preprint https://arxiv.org/abs/260414141

Chen X, Cheng Y, Han X, Zhao B, Tao W, Ozdemir E, Pan D (2025b) A compact surface reconstruction method for buildings based on convolutional neural network fitting implicit representations. Journal of Computing in Civil Engineering 39(3):04025024

Chen X, Chen Y, Xiu Y, Geiger A, Chen A (2025a) Ttt3r: 3d reconstruction as test-time training. arXiv preprint https://arxiv.org/abs/2509.26645

Chen Y, Shen S, Chen Y, Wang G (2020) Graph-based parallel large scale structure from motion. Pattern Recognit 107:107537. https://doi.org/10.1016/j.patcog.2020.107537

Chen Y, Jiang J, Jiang K, Tang X, Li Z, Liu X, Nie Y (2025c) DashGaussian: Optimizing 3d Gaussian splatting in 200 seconds. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp 11146–11155

Chen Z, Li J, Li Q, Dong Z, Yang B (2024c) DeepAAT: Deep automated aerial triangulation for fast UAV-based mapping. International Journal of Applied Earth Observation and Geoinformation 134:104190

Chen Z, Dai L, Wang D, Guo Q, Zhao R (2025d) Fs-mvsnet: A multi-view image-based framework for 3d forest reconstruction and parameter extraction of single trees. Forests 16(6):927

Cheng S, Xu Z, Zhu S, Li Z, Li LE, Ramamoorthi R, Su H (2020a) Deep stereo using adaptive thin volume representation with uncertainty awareness. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Cheng X, Zhong Y, Harandi M, Dai Y, Chang X, Li H, Drummond T, Ge Z (2020b) Hierarchical neural architecture search for deep stereo matching. Advances in Neural Information Processing Systems 33:22158–22169

Chibane J, Pons-Moll G et al (2020) Neural unsigned distance fields for implicit function learning. Advances in Neural Information Processing Systems 33:21638–21652

Collins R (1996) A space-sweep approach to true multi-image matching. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. https://doi.org/10.1109/CVPR.1996.517097

Cui H, Gao X, Shen S, Hu Z (2017) HSfM: Hybrid structure-from-motion. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. https://doi.org/10.1109/CVPR.2017.257

Dai P, Xu J, Xie W, Liu X, Wang H, Xu W (2024) High-quality surface reconstruction using Gaussian surfels. In: ACM SIGGRAPH 2024 conference papers, pp 1–11

De Floriani L (1987) Surface representations based on triangular grids. The Visual Computer 3(1):27–50

Deng J, Hou F, Chen X, Wang W, He Y (2024) 2S-UDF: a novel two-stage UDF learning method for robust non-watertight model reconstruction from multi-view images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5084–5093

Deng K, Ti Z, Xu J, Yang J, Xie J (2025) VGGT-Long: Chunk it, loop it, align it—pushing vggt's limits on kilometer-scale long RGB sequences. arXiv preprint https://arxiv.org/abs/2507.16443

DeTone D, Malisiewicz T, Rabinovich A (2018) Superpoint: Self-supervised interest point detection and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 224–236

Dey TK, Goswami S (2003) Tight cocone: a water-tight surface reconstructor. In: Proceedings of the eighth ACM symposium on Solid modeling and applications, pp 127–134

Di Angelo L, Di Stefano P, Giaccari L (2011) A new mesh-growing algorithm for fast surface reconstruction. Computer-Aided Design 43(6):639–650

Ding Y, Zhu Q, Liu X, Yuan W, Zhang H, Zhang C (2022) KD-MVS: Knowledge distillation based self-supervised learning for multi-view stereo. In: European conference on computer vision, Springer, pp 630–646

Djeghim H, Piasco N, Bennehar M, Roldão L, Tsishkou D, Sidibé D (2025) ViiNeuS: Volumetric initialization for implicit neural surface reconstruction of urban scenes with limited image overlap. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp 11854–11863

Dornaika F, Horaud R (1998) Simultaneous robot-world and hand-eye calibration. IEEE Transactions on Robotics and Automation 14(4):617–622. https://doi.org/10.1109/70.704233

Drost B, Ulrich M, Navab N, Ilic S (2010) Model globally, match locally: Efficient and robust 3d object recognition. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. https://doi.org/10.1109/CVPR.2010.5540108

Duggal S, Wang S, Ma WC, Hu R, Urtasun R (2019) DeepPruner: Learning efficient stereo matching via differentiable patchmatch. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 4383-4392

Duisterhof BP, Zust L, Weinzaepfel P, Leroy V, Cabon Y, Revaud J (2025) MASt3R-SfM: a fully-integrated solution for unconstrained structure-from-motion. In: 2025 International Conference on 3D Vision (3DV), IEEE, pp 1–10

Dusmanu M, Rocco I, Pajdla T, Pollefeys M, Sivic J, Torii A, Sattler T (2019) D2-Net: A trainable cnn for joint description and detection of local features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Duzceker A, Galliani S, Vogel C, Speciale P, Dusmanu M, Pollefeys M (2021) DeepVideoMVS: Multi-view stereo on video with recurrent spatio-temporal fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 15324–15333

Edstedt J, Athanasiadis I, Wadenbäck M, Felsberg M (2023) DKM: Dense kernelized feature matching for geometry estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 17765–17775

Feng G, Chen S, Fu R, Liao Z, Wang Y, Liu T, Hu B, Xu L, Pei Z, Li H, et al. (2025) FlashGS: Efficient 3d Gaussian splatting for large-scale and high-resolution rendering. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp 26652–26662

Fischler MA, Bolles RC (1987) Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. In: Fischler MA, Firschein O (eds) Readings in Computer Vision. Morgan Kaufmann, San Francisco (CA), pp 726–740 https://doi.org/10.1016/B978-0-08-051581-6.50070-2

Fridovich-Keil S, Yu A, Tancik M, Chen Q, Recht B, 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

Fu Q, Xu Q, Ong YS, Tao W (2022) Geo-Neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. Advances in Neural Information Processing Systems 35:3403–3416

Furukawa Y, Ponce J (2010) Accurate, dense, and robust multiview stereopsis. IEEE Trans Pattern Anal Mach Intell 32(8):1362–1376. https://doi.org/10.1109/TPAMI.2009.161

Gan W, Yu Y, Perda G, Morelli L, Xia R, Zhan Z, Wang X, Remondino F (2024a) LVG-SfM: Learning-based view-graph generation for robust on-the-fly SfM. In: European conference on computer vision Workshops. Springer, pp 158–174

Gan W, Zhan Z, Wang X (2024b) Efficient feature matching and pose-graph initialization for sfm. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLVIII-1-2024:165–171

Gao Y, Li H, Chen J, Zou Z, Zhong Z, Zhang D, Sun X, Han J (2025) CityGS-X: A scalable architecture for efficient and geometrically accurate large-scale scene reconstruction. arXiv preprint https://arxiv.org/abs/2503.23044

Gong J (2023) Structure from motion with a neural network. Master's Theses in Mathematical Sciences, Mathematics (Faculty of Engineering), Lund University, ISSN 1404-6342, https://lup.lub.lu.se/student-papers/search/publication/9126530

Gu X, Fan Z, Zhu S, Dai Z, Tan F, Tan P (2020) Cascade cost volume for high-resolution multi-view stereo and stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Gu X, Yuan W, Dai Z, Tang C, Zhu S, Tan P (2021) DRO: Deep recurrent optimizer for structure-from-motion. arXiv preprint https://arxiv.org/abs/2103.13201.

Guédon A, Lepetit V (2024) Sugar: Surface-aligned Gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5354–5363

Gui H, Hu L, Chen R, Huang M, Yin Y, Yang J, Wu Y, Liu C, Sun Z, Zhang X et al (2024) Balanced 3DGS: Gaussian-wise parallelism rendering with fine-grained tiling. arXiv preprint https://arxiv.org/abs/2412.17378

Guo J, Deng N, Li X, Bai Y, Shi B, Wang C, Ding C, Wang D, Li Y (2023) StreetSurf: Extending multi-view implicit surface reconstruction to street views. arXiv preprint https://arxiv.org/abs/2306.04988

Haitz D, Jutzi B, Ulrich M, Jäger M, Hübner P (2023) Combinbin HoloLens with Instant-NeRFs: Advanced real-time 3D mobile mapping. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLVIII-1/W1-2023:167–174. https://doi.org/10.5194/isprs-archives-XLVIII-1-W1-2023-167-2023

Hartley R, Zisserman A (2003) Multiple View Geometry in Computer Vision. Cambridge University Press, Cambridge

He X, Sun J, Wang Y, Peng S, Huang Q, Bao H, Zhou X (2024) Detector-free structure from motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 21594–21603

Hillemann M, Langendörfer R, Heiken M, Mehltretter M, Schenk A, Weinmann M, Hinz S, Heipke C, Ulrich M (2024) Novel view synthesis with neural radiance fields for industrial robot applications. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLVIII-2-2024:137–144. https://doi.org/10.5194/isprs-archives-XLVIII-2-2024-137-2024

Hirschmüller H (2005) Accurate and efficient stereo processing by semi-global matching and mutual information. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2:807–814. https://doi.org/10.1109/CVPR.2005.56

Hirschmüller H (2008) Stereo processing by semiglobal matching and mutual information. IEEE Trans Pattern Anal Mach Intell 30(2):328–341. https://doi.org/10.1109/TPAMI.2007.1166

Hoppe H, DeRose T, Duchamp T, McDonald J, Stuetzle W (1992) Surface reconstruction from unorganized points. In: Proceedings of the 19th annual conference on computer graphics and interactive techniques, pp 71–78

Horaud R, Dornaika F (1995) Hand-eye calibration. Int J Rob Res 14(3):195–210. https://doi.org/10.1177/027836499501400301

Hou Q, Xia R, Zhang J, Feng Y, Zhan Z, Wang X (2023) Learning visual overlapping image pairs for SfM via CNN fine-tuning with photogrammetric geometry information. International Journal of Applied Earth Observation and Geoinformation 116:103162

Hou Y, Zhang H, Wang X, Zhan Z (2024) Deep-FG-DSM: Fine-grained DSM for high-resolution UAV imagery based on deep implicit occupancy networks. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 48:209–216

Huang B, Yu Z, Chen A, Geiger A, Gao S (2024) 2d Gaussian splatting for geometrically accurate radiance fields. In: ACM SIGGRAPH 2024 conference papers, pp 1–11

Jäger M, Hillemann M, Jutzi B (2025) FeatureGS: Eigenvalue-feature optimization in 3d gaussian splatting for geometrically accurate and artifact-reduced reconstruction. ISPRS Open Journal of Photogrammetry and Remote Sensing. https://doi.org/10.1016/j.ophoto.2025.100100

Jiang J, Ma H (2025) Efficient multi-view stereo with depth-aware iterations and hybrid loss strategy. Pattern Recognition p 112500

Jiang H, Zeng C, Chen R, Liang S, Han Y, Gao Y, Wang C (2023) Depth-NeuS: Neural implicit surfaces learning for multi-view reconstruction based on depth information optimization. arXiv preprint https://arxiv.org/abs/2303.17088

Jiang J, Wang L, Yu H, Hu T, Chen J, Ma H (2025a) RRT-MVS: Recurrent regularization transformer for multi-view stereo. Proceedings of the AAAI Conference on Artificial Intelligence 39:3994–4002

Jiang L, Ren K, Yu M, Xu L, Dong J, Lu T, Zhao F, Lin D, Dai B (2025b) Horizon-GS: Unified 3d Gaussian splatting for large-scale aerial-to-ground scenes. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp 26789–26799

Jiang S, Li Q, Jiang W, Chen W (2022) Parallel structure from motion for UAV images via weighted connected dominating set. IEEE Geosci Remote Sens Lett 60:1–13. https://doi.org/10.1109/TGRS.2022.3222776

Ju T, Losasso F, Schaefer S, Warren J (2002) Dual contouring of hermite data. In: Proceedings of the 29th annual conference on Computer graphics and interactive techniques, pp 339–346

Kasten Y, Geifman A, Galun M, Basri R (2019) Algebraic characterization of essential matrices and their averaging in multiview settings. In: ICCV 2019, pp 5895–5903

Kazhdan M, Hoppe H (2013) Screened poisson surface reconstruction. ACM Transactions on Graphics (ToG) 32(3):1–13

Kazhdan M, Bolitho M, Hoppe H (2006) Poisson surface reconstruction. In: Proceedings of the Fourth Eurographics Symposium on Geometry Processing, pp 61–70

Keetha N, Müller N, Schönberger J, Porzi L, Zhang Y, Fischer T, Knapitsch A, Zauss D, Weber E, Antunes N et al (2025) MapAnything: Universal feed-forward metric 3d reconstruction. arXiv preprint https://arxiv.org/abs/2509.13414

Kendall A, Martirosyan H, Dasgupta S, Henry P, Kennedy R, Bachrach A, Bry A (2017) End-to-end learning of geometry and context for deep stereo regression. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV)

Kerbl B, Kopanas G, Leimkühler T, Drettakis G (2023) 3d Gaussian splatting for real-time radiance field rendering. ACM Trans Graph 42(4):139–141

Khatib F, Kasten Y, Moran D, Galun M, Basri R (2025) RESfM: Robust deep equivariant structure from motion. In: Proceedings of The International Conference on Learning Representations, https://openreview.net/forum?id=wldwEhQ7cl

Kim T (2000) A study on the epipolarity of linear pushbroom images. Photogrammetric Engineering & Remote Sensing 66(8):961–966

Kolluri R (2008) Provably good moving least squares. ACM Transactions on Algorithms (TALG) 4(2):1–25

Kutulakos KN, Seitz SM (2000) A theory of shape by space carving. International journal of computer vision 38(3):199–218

Labatut P, Pons JP, Keriven R (2007) Efficient multi-view reconstruction of large-scale scenes using interest points, delaunay triangulation and graph cuts. In: 2007 IEEE 11th international conference on computer vision, IEEE, pp 1–8

Labatut P, Pons JP, Keriven R (2009) Robust and efficient surface reconstruction from range data. Computer graphics forum 28(8):2275–2290

Laga H, Jospin LV, Boussaid F, Bennamoun M (2022) A survey on deep learning techniques for stereo-based depth estimation. IEEE Trans Pattern Anal Mach Intell 44(4):1738–1764. https://doi.org/10.1109/TPAMI.2020.3032602

Lan Y, Luo Y, Hong F, Zhou S, Chen H, Lyu Z, Yang S, Dai B, Loy CC, Pan X (2025) STream3R: Scalable sequential 3d reconstruction with causal transformer. arXiv preprint https://arxiv.org/abs/250810893

Landgraf S (2025) Efficient estimation and exploitation of predictive uncertainties in deep learning-based machine vision. PhD thesis. Karlsruher Institut für Technologie (KIT) https://doi.org/10.5445/IR/1000183395

Landgraf S, Hillemann M, Kapler T, Ulrich M (2025a) A comparative study on multi-task uncertainty quantification in semantic segmentation and monocular depth estimation. Tech Mess 92(7-8):298–310. https://doi.org/10.1515/teme-2025-0004

Landgraf S, Hillemann M, Kapler T, Ulrich M (2025b) Efficient multi-task uncertainties for joint semantic segmentation and monocular depth estimation. In: Cremers D, Lähner Z, Moeller M, Nießner M, Ommer B, Triebel R (eds) Pattern Recognition. Springer Nature Switzerland, Cham, pp 348–364

Landgraf S, Qin R, Ulrich M (2025c) A critical synthesis of uncertainty quantification and foundation models in monocular depth estimation. arXiv preprint https://arxiv.org/abs/2501.08188

Laurentini A (2002) The visual hull concept for silhouette-based image understanding. IEEE Transactions on pattern analysis and machine intelligence 16(2):150–162

Leroy V, Cabon Y, Revaud J (2024) Grounding image matching in 3d with MASt3R. In: European Conference on Computer Vision, Springer, pp 71–91

Levin D (2003) Mesh-independent surface interpolation. In: Geometric modeling for scientific visualization. Springer, pp 37–49

Li C, Zhu H, Chen H, Zhang J, Chen T, Yang S, Shao S, Dong W, Zhang B (2025a) HRGS: Hierarchical Gaussian splatting for memory-efficient high-resolution 3d reconstruction. arXiv preprint https://arxiv.org/abs/2506.14229

Li Z, Müller T, Evans A, Taylor RH, Unberath M, Liu MY, Lin CH (2023) Neuralangelo: High-fidelity neural surface reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 8456–8465

Li Z, Yao S, Wu T, Yue Y, Zhao W, Qin R, Garcia-Fernandez AF, Levers A, Ralph J, Zhu X (2025c) ULSR-GS: Urban large-scale surface reconstruction Gaussian splatting with multi-view geometric consistency. ISPRS J Photogramm Remote Sens 230:861–880. https://doi.org/10.1016/j.isprsjprs.2025.10.008

Li Z, Qi Z, Wang W, Wang Z, Duan J, Lei N (2025b) Point2Quad: Generating quad meshes from point clouds via face prediction. IEEE Transactions on Circuits and Systems for Video Technology

Liang Y, Yang Y, Fan X, Cui T (2023) Efficient and accurate hierarchical SfM based on adaptive track selection for large-scale oblique images. Remote Sensing. https://doi.org/10.3390/rs15051374

Lin HW, Tai CL, Wang GJ (2004) A mesh reconstruction algorithm driven by an intrinsic property of a point cloud. Computer-Aided Design 36(1):1–9

Lin J, Li Z, Tang X, Liu J, Liu S, Liu J, Lu Y, Wu X, Xu S, Yan Y, et al. (2024) VastGaussian: Vast 3d Gaussians for large scene recon-

struction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5166–5175
Lindenberger P, Sarlin PE, Larsson V, Pollefeys M (2021) Pixel-perfect structure-from-motion with featuremetric refinement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 5987–5997
Lindenberger P, Sarlin PE, Pollefeys M (2023) LightGlue: Local feature matching at light speed. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 17627–17638
Liu J, Wan Y, Wang B, Zheng C, Lin J, Zhang F (2025a) GS-SDF: LiDAR-augmented Gaussian splatting and neural SDF for geometrically consistent rendering and reconstruction. arXiv preprint https://arxiv.org/abs/2503.10170
Liu L, Wang C, Feng C, Gong W, Zhang L, Liao L, Feng C (2024a) Incremental SFM 3D reconstruction based on deep learning. Electronics. https://doi.org/10.3390/electronics13142850
Liu M, Zhang X, Su H (2020) Meshing point clouds with predicted intrinsic-extrinsic ratio guidance. In: European conference on computer vision, Springer, pp 68–84
Liu S, Qiao P, Ye Z, Li W, Dou Y (2024b) VoxNeuS: Enhancing voxel-based neural surface reconstruction via gradient interpolation. arXiv preprint https://arxiv.org/abs/2406.07170
Liu S, Gao Y, Zhang T, Pautrat R, Schönberger JL, Larsson V, Pollefeys M (2025b) Robust incremental structure-from-motion with hybrid features. In: Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G (eds) Computer Vision—ECCV 2024. Springer Nature Switzerland, Cham, pp 249–269 https://doi.org/10.1007/978-3-031-72764-1_15
Liu Y, Luo C, Mao Z, Peng J, Zhang Z (2024c) CityGaussianV2: Efficient and geometrically accurate reconstruction for large-scale scenes. arXiv preprint https://arxiv.org/abs/2411.00771
Liu Y, Dong S, Wang S, Yin Y, Yang Y, Fan Q, Chen B (2025c) SLAM3R: Real-time dense scene reconstruction from monocular RGB videos. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp 16651–16662
Liu Z, Qv W, Cai H, Guan H, Zhang S (2023) An efficient and robust hybrid SfM method for large-scale scenes. Remote Sensing. https://doi.org/10.3390/rs15030769
Liu Z, Chen Z, Zhang X, Cheng S (2024d) CDP-MVS: Forest multi-view reconstruction with enhanced confidence-guided dynamic domain propagation. Remote Sensing 16(20):3845
Lorensen WE, Cline HE (1998) Marching cubes: A high resolution 3d surface construction algorithm. In: Seminal graphics: pioneering efforts that shaped the field, pp 347–353
Lou L, Li W, Gan W, Yu Y, Wang T, Wang X, Zhan Z (2025) On-the-fly feedback SfM: Online explore-and-exploit UAV photogrammetry with incremental mesh quality-aware indicator and predictive path planning. arXiv preprint https://arxiv.org/abs/2512.02375
Lowe DG (2004) Distinctive image features from scale-invariant keypoints. International journal of computer vision 60(2):91–110
Luo Y, Mi Z, Tao W (2021) DeepDT: Learning geometry from delaunay triangulation for surface reconstruction. Proc AAAI Conf Artif Intell 35(3):2277–2285. https://doi.org/10.1609/aaai.v35i3.16327
Maggio D, Lim H, Carlone L (2025) VGGT-SLAM: Dense RGB SLAM optimized on the sl (4) manifold. arXiv preprint https://arxiv.org/abs/250512549
Mallick SS, Goel R, Kerbl B, Steinberger M, Carrasco FV, De La Torre F (2024) Taming 3DGS: High-quality radiance fields with limited resources. In: SIGGRAPH Asia 2024 Conference Papers, pp 1–11
Matusik W, Buehler C, Raskar R, Gortler SJ, McMillan L (2000) Image-based visual hulls. In: Proceedings of the 27th annual conference on Computer graphics and interactive techniques, pp 369–374
Mescheder L, Oechsle M, Niemeyer M, Nowozin S, Geiger A (2019) Occupancy networks: Learning 3d reconstruction in function space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4460–4470
Michel C, Ulrich M (2025) Automatic inspection of punched metal plate fasteners on timber-to-timber joints with image-based 3d reconstruction. 6th Joint International Symposium on Deformation Monitoring (JISDM 2025), Karlsruhe, 07.–09.04.2025, Karlsruher Institut für Technologie (KIT). https://doi.org/10.5445/IR/1000180376
Mildenhall B, Srinivasan PP, Tancik M, Barron JT, Ramamoorthi R, Ng R (2021) NeRF: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM 65(1):99–106
Moran D, Koslowsky H, Yoni K, Maron H, Galun M, Basri R (2021) Deep permutation equivariant structure from motion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 5976–5986
Morelli L, Ioli F, Maiwald F, Mazzacca G, Menna F, 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 XLVIII-2/W4-2024:309–316. https://doi.org/10.5194/isprs-archives-XLVIII-2-W4-2024-309-2024
Müller T, Evans A, Schied C, Keller A (2022) Instant neural graphics primitives with a multiresolution hash encoding. ACM transactions on graphics (TOG) 41(4):1–15
Murai R, Dexheimer E, Davison AJ (2025) MASt3R-SLAM: Real-time dense slam with 3d reconstruction priors. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp 16695–16705
Nan L, Wonka P (2017) Polyfit: Polygonal surface reconstruction from point clouds. In: Proceedings of the IEEE international conference on computer vision, pp 2353–2361
Nex F, Stathopoulou E, Remondino F, Yang M, Madhuanand L, Yogender Y, Alsadik B, Weinmann M, Jutzi B, Qin R (2024) UseGeo-A UAV-based multi-sensor dataset for geospatial research. ISPRS Open Journal of Photogrammetry and Remote Sensing 13:100070
Ni TG, Ma ZH (2010) A fast surface reconstruction algorithm for 3D unorganized points. 2010 2nd International Conference on Computer Engineering and Technology, Chengdu, China: V7-15-V7-18, https://doi.org/10.1109/ICCET.2010.5485908
Ohtake Y, Belyaev A, Alexa M, Turk G, Seidel HP (2005) Multi-level partition of unity implicits. In: ACM Siggraph 2005 Courses, pp 173–es
Ono Y, Trulls E, Fua P, Yi KM (2018) LF-Net: Learning local features from images. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 31, https://proceedings.neurips.cc/paper_files/paper/2018/file/f5496252609c43eb8a3d147ab9b9c006-Paper.pdf
Pan L, Baráth D, Pollefeys M, Schönberger JL (2025) Global structure-from-motion revisited. In: Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G (eds) Computer Vision—ECCV 2024. Springer Nature Switzerland, Cham, pp 58–77
Park JJ, Florence P, Straub J, Newcombe R, Lovegrove S (2019) DeepSDF: Learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 165–174
Pauly M (2003) Point primitives for interactive modeling and processing of 3D geometry. Hartung-Gorre, Süderdithmarschen, Germany
Peng S, Niemeyer M, Mescheder L, Pollefeys M, Geiger A (2020) Convolutional occupancy networks. In: European Conference on Computer Vision, Springer, pp 523–540
Ren S, Hou J, Chen X, He Y, Wang W (2023) GeoUDF: Surface reconstruction from 3d point clouds via geometry-guided distance rep-

resentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 14214–14224
Ren S, Wen T, Fang Y, Lu B (2026) FastGS: Training 3d Gaussian splatting in 100 seconds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 2604-26103
Revaud J, De Souza C, Humenberger M, Weinzaepfel P (2019) R2D2: Reliable and repeatable detector and descriptor. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R (eds) Advances in Neural Information Processing Systems, vol 32. Curran Associates, Inc.,
Rich A, Stier N, Sen P, Höllerer T (2024) Smoothness, synthesis, and sampling: Re-thinking unsupervised multi-view stereo with div loss. In: European Conference on Computer Vision, Springer, pp 380–397
Rosten E, Porter R, Drummond T (2010) Faster and better: A machine learning approach to corner detection. IEEE Trans Pattern Anal Mach Intell 32(1):105–119. https://doi.org/10.1109/TPAMI.2008.275
Rublee E, Rabaud V, Konolige K, Bradski G (2011) ORB: An efficient alternative to SIFT or SURF. Proc IEEE Int Conf Comput Vis. https://doi.org/10.1109/ICCV.2011.6126544
Santo H, Okura F, Matsushita Y (2024) MVCPS-NeuS: multi-view constrained photometric stereo for neural surface reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 20475–20484
Sarlin PE, Cadena C, Siegwart R, Dymczyk M (2019) From coarse to fine: Robust hierarchical localization at large scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
Sarlin PE, DeTone D, Malisiewicz T, Rabinovich A (2020) Superglue: Learning feature matching with graph neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4938–4947
Scharstein D, Szeliski R (2002) A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International journal of computer vision 47(1):7–42
Schnabel R, Wahl R, Klein R (2007) Efficient RANSAC for point-cloud shape detection. Computer graphics forum 26(2):214–226
Schönberger JL, Frahm JM (2016) Structure-from-motion revisited. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. https://doi.org/10.1109/CVPR.2016.445
Schönberger JL, Zheng E, Frahm JM, Pollefeys M (2016) Pixelwise view selection for unstructured multi-view stereo. In: Leibe B, Matas J, Sebe N, Welling M (eds) Computer Vision—ECCV 2016. Springer International Publishing, Cham, pp 501–518
Sellán S, Jacobson A (2022) Stochastic poisson surface reconstruction. ACM Transactions on Graphics (TOG) 41(6):1–12
Shah R, Chari V, Narayanan PJ (2018) View-graph selection framework for SfM. In: European Conference on Computer Vision, Springer
Sharp N, Ovsjanikov M (2020) PointTriNet: Learned triangulation of 3d point sets. In: European conference on computer vision, Springer, pp 762–778
Shen Y, Zhang Z, Qu Y, Cao L (2025a) FastVVT: Training-free acceleration of visual geometry transformer. arXiv preprint https://arxiv.org/abs/2509.02560
Shen Z, Shu M, Wang G, Yu Y, Zhan Z, Wang X (2025b) Surveys on feed-forward 3r methods for high-resolution photogrammetric images via image divide-and-conquer strategy. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLVIII-1/W5-2025:109–116
Sitzmann V, Martel J, Bergman A, Lindell D, Wetzstein G (2020) Implicit neural representations with periodic activation functions. Advances in neural information processing systems 33:7462–7473
Snavely KN (2008) Scene reconstruction and visualization from internet photo collections, Dissertation edn. University of Washington
Song H, Park J, Heo S, Kang J, Lee S (2020) Patchmatch based multiview stereo with local quadric window. In: Proceedings of the 28th ACM International Conference on Multimedia, pp 2664–2672
Steger C, Ulrich M (2022) A multi-view camera model for line-scan cameras with telecentric lenses. J Math Imaging Vis 64(2):105–130. https://doi.org/10.1007/s10851-021-01055-x
Steger C, Ulrich M, Wiedemann C (2018) Machine Vision Algorithms and Applications, 2nd edn edn. Wiley-VCH, Weinheim
Su W, Zhang C, Xu Q, Tao W (2024) PSDF: Prior-driven neural implicit surface learning for multi-view reconstruction. IEEE Transactions on Visualization and Computer Graphics 31(9), pp 5229-5244
Sun J, Shen Z, Wang Y, Bao H, Zhou X (2021) LoFTR: Detector-free local feature matching with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 8922–8931
Sun S, Xu D, Wu H, Ying H, Mou Y (2022) Multi-view stereo for large-scale scene reconstruction with mrf-based depth inference. Computers & Graphics 106:248–258
Szeliski R (2022) Computer vision: algorithms and applications. Springer Nature
Tan R, Wang Q, Wang X, Yan C, Sun Y, Feng Y (2023) Mp-mvs: Multi-scale windows patchmatch and planar prior multi-view stereo. arXiv preprint https://arxiv.org/abs/2309.13294
Tancik M, Casser V, Yan X, Pradhan S, Mildenhall B, Srinivasan PP, Barron JT, Kretzschmar H (2022) Block-NeRF: Scalable large scene neural view synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8248–8258
Tang C, Tan P (2019) BA-Net: Dense bundle adjustment networks. In: Proceedings of The International Conference on Learning Representations, https://openreview.net/pdf?id=B1gabhRcYX
Tang Z, Fan Y, Wang D, Xu H, Ranjan R, Schwing A, Yan Z (2025) MV-DUSt3R+: Single-stage scene reconstruction from sparse views in 2 seconds. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp 5283–5293
Tao P, Cui H, Tu D, Shen S (2025) MGSfM: Multi-camera driven global structure-from-motion. In: ICCV 2025, pp 5232–5241
Tonti CM, Papa L, Amerini I (2024) Lightweight 3-d convolutional occupancy networks for virtual object reconstruction. IEEE Computer Graphics and Applications 44(2):23–36
Turki H, Ramanan D, Satyanarayanan M (2022) Mega-NeRF: Scalable construction of large-scale NeRFs for virtual fly-throughs. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12922–12931
Tyszkiewicz M, Fua P, Trulls E (2020) DISK: Learning local features with policy gradient. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in Neural Information Processing Systems, vol 33. Curran Associates, Inc, pp 14254–14265
Ulrich M, Hillemann M (2024) Uncertainty-aware hand-eye calibration. IEEE Trans Robot 40:573–591. https://doi.org/10.1109/TRO.2023.3330609
Ulrich M, Steger C (2016) Hand-eye calibration of SCARA robots using dual quaternions. Pattern Recognit. Image Anal. 26(1):231–239. https://doi.org/10.1134/S1054661816010272
Ulrich M, Steger C (2019) A camera model for cameras with hypercentric lenses and some example applications. Mach Vis Appl 30:1013–1028. https://doi.org/10.1007/s00138-019-01032-w
Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in Neural Information Processing Systems, vol 30. Curran Associates, Inc,

Venkatesh R, Sharma S, Ghosh A, Jeni L, Singh M (2020) DUDE: Deep unsigned distance embeddings for hi-fidelity representation of complex 3d surfaces. arXiv preprint https://arxiv.org/abs/2011.02570

Vijayanarasimhan S, Ricco S, Schmid C, Sukthankar R, Fragkiadaki K (2017) SfM-Net: Learning of structure and motion from video. arXiv preprint https://arxiv.org/abs/1704.07804

Vuong K, Ghosh A, Ramanan D, Narasimhan S, Tulsiani S (2025) AerialMegaDepth: Learning aerial-ground reconstruction and view synthesis. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp 21674–21684

Wang H, Agapito L (2025) 3d reconstruction with spatial memory. 2025 International Conference on 3D Vision (3DV). https://doi.org/10.1109/3DV66043.2025.00013

Wang K, Shen S (2018) MVDepthNet: Real-time multiview depth estimation neural network. In: 2018 International conference on 3d vision (3DV), IEEE, pp 248–257

Wang X, Heipke C (2020) An improved method of refining relative orientation in global structure from motion with a focus on repetitive structure and very short baselines. Photogrammetric Engineering & Remote Sensing 85:299–315

Wang F, Zhu Q, Chang D, Gao Q, Han J, Zhang T, Hartley R, Pollefeys M (2024a) Learning-based multi-view stereo: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 48(6), pp 6241-6260

Wang J, Karaev N, Rupprecht C, Novotny D (2024c) VGGSfM: Visual geometry grounded deep structure from motion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 21686–21697

Wang J, Chen M, Karaev N, Vedaldi A, Rupprecht C, Novotny D (2025a) VGGT: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp 5294–5306

Wang P, Liu L, Liu Y, Theobalt C, Komura T, Wang W (2021a) NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. arXiv preprint https://arxiv.org/abs/2106.10689

Wang Q, Zhang Y, Holynski A, Efros AA, Kanazawa A (2025b) Continuous 3d perception model with persistent state. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp 10510–10522

Wang S, Leroy V, Cabon Y, Chidlovskii B, Revaud J (2024d) DUSt3R: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 20697–20709

Wang T, Zhan Z, Xia R, Ji L, Wang X (2025d) Masked Gaussian fields for automated building surface meshing from multi-view images. Automation in Construction 180:106502

Wang T, Wang X, Hou Y, Xu Y, Zhang W, Zhan Z (2025c) PGSAG: Parallel Gaussian splatting for fine-grained large-scale urban buildings reconstruction via semantic-aware grouping. PFG-Journal of Photogrammetry, Remote Sensing and Geoinformation Science pp 1–16

Wang W, Deng Y, Li Z, Liu Y, Lei N (2024e) MergeNet: Explicit mesh reconstruction from sparse point clouds via edge prediction. In: 2024 IEEE International Conference on Multimedia and Expo (ICME), IEEE, pp 1–6

Wang X, Rottensteiner F, Christian H (2019a) Structure from motion for ordered and unordered image sets based on random k-d forests and global pose estimation. ISPRS J Photogramm Remote Sens 147:19–41. https://doi.org/10.1016/j.isprsjprs.2018.11.009

Wang X, Xiao T, Gruber M, Heipke C (2019b) Robustifying relative orientations with respect to repetitive structures and very short baselines for global SfM. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops

Wang X, Teng X, Yoni K (2021b) A hybrid global structure from motion method for synchronously estimating global rotations and global translations. ISPRS J Photogramm Remote Sens 174:35–55. https://doi.org/10.1016/j.isprsjprs.2021.02.002

Wang X, Yi R, Ma L (2024f) AdR-Gaussian: Accelerating Gaussian splatting with adaptive radius. In: SIGGRAPH Asia 2024 Conference Papers, pp 1–10

Wang Y, Han Q, Habermann M, Daniilidis K, Theobalt C, Liu L (2023a) 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 Y, Zeng Z, Guan T, Yang W, Chen Z, Liu W, Xu L, Luo Y (2023b) Adaptive patch deformation for textureless-resilient multi-view stereo. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1621–1630

Wang Y, Zhou K, Zhang W, Xiao C (2024g) MegaSurf: Scalable large scene neural surface reconstruction. In: Proceedings of the 32nd ACM International Conference on Multimedia, pp 6414–6423

Wang Y, Zheng J, Zhang C, Zhang Z, Li K, Zhang Y, Hu J (2025e) Dualnet: Robust self-supervised stereo matching with pseudo-label supervision. Proceedings of the AAAI Conference on Artificial Intelligence 39:8178–8186

Wang Y, Zhou J, Zhu H, Chang W, Zhou Y, Li Z, Chen J, Pang J, Shen C, He T (2025f) π3: Permutation-equivariant visual geometry learning. arXiv preprint https://arxiv.org/abs/2507.13347

Wang Z, Luo H, Wang X, Zheng J, Ning X, Bai X (2024h) A contrastive learning based unsupervised multi-view stereo with multi-stage self-training strategy. Displays 83:102672

Wei X, Zhang Y, Li Z, Fu Y, Xue X (2020) DeepSFM: Structure from motion via deep bundle adjustment. In: European Conference on Computer Vision, Springer, pp 230–247

Werner T, Zisserman A (2002) New techniques for automated architectural reconstruction from photographs. In: Heyden A, Sparr G, Nielsen M, Johansen P (eds) Computer Vision—ECCV 2002. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 541–555

Wolf Y, Bracha A, Kimmel R (2024) GS2Mesh: Surface reconstruction from Gaussian splatting via novel stereo views. In: European Conference on Computer Vision, Springer, pp 207–224

Wu C (2013) Towards linear-time incremental structure from motion. 2013 International Conference on 3D Vision—3DV 2013. https://doi.org/10.1109/3DV.2013.25

Wu X, Landgraf S, Ulrich M, Qin R (2025a) An evaluation of DUSt3R/MASt3R/VGGT 3D reconstruction on photogrammetric aerial blocks. Geo Spat Inf Sci. https://doi.org/10.1080/10095020.2025.2597491

Wu Y, Qi Z, Shi Z, Zou Z (2025b) BlockGaussian: Efficient large-scale scene novel view synthesis via adaptive block-based Gaussian splatting.

Xiangli Y, Xu L, Pan X, Zhao N, Rao A, Theobalt C, Dai B, Lin D (2022) BungeeNeRF: Progressive neural radiance field for extreme multi-scale scene rendering. In: European conference on computer vision, Springer, pp 106–122

Xiao T, Deng F, Wang X, Heipke C (2021) Sequential cycle consistency inference for eliminating incorrect relative orientations in structure from motion. PFG—Journal of Photogrammetry, Remote Sensing and Geoinformation Science 89:233–249

Xie T, Yang P, Jin Y, Cai Y, Yin W, Ren W, Zhang Q, Hua W, Peng S, Guo X et al (2026) Scal3R: Scalable test-time training for large-scale 3d reconstruction. arXiv preprint https://arxiv.org/abs/2604.08542

Xiong K, Peng R, Zhang Z, Feng T, Jiao J, Gao F, Wang R (2023) CL-MVSNet: Unsupervised multi-view stereo with dual-level contrastive learning. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 3769–3780

Xu Q, Tao W (2019) Multi-scale geometric consistency guided multi-view stereo. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5483–5492

Xu C, Hou F, Wang W, Qin H, Zhang Z, He Y (2025) Details enhancement in unsigned distance field learning for high-fidelity 3d surface reconstruction. Proceedings of the AAAI Conference on Artificial Intelligence 39(8):8806–8814

Xu Y, Wang T, Zhan Z, Wang X (2024) Mega-NeRF++: An improved scalable nerfs for high-resolution photogrammetric images. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 48:769–776

Yan Q, Kang J, Xiao T, Liu H, Deng F (2024) MVP-stereo: A parallel multi-view patchmatch stereo method with dilation matching for photogrammetric application. Remote Sensing 16(6):964

Yang J, Mao W, Alvarez JM, Liu M (2020) Cost volume pyramid based depth inference for multi-view stereo. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4877–4886

Yang J, Sax A, Liang KJ, Henaff M, Tang H, Cao A, Chai J, Meier F, Feiszli M (2025) Fast3R: Towards 3d reconstruction of 1000+ images in one forward pass. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 21924–21935

Yang L, Kang B, Huang Z, Zhao Z, Xu X, Feng J, Zhao H (2024) Depth Anything V2. Advances in Neural Information Processing Systems 37:21875–21911. https://doi.org/10.52202/079017-0688

Yao Y, Luo Z, Li S, Fang T, Quan L (2018) MVSNet: Depth inference for unstructured multi-view stereo. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 785-801

Yao Y, Luo Z, Li S, Shen T, Fang T, Quan L (2019) Recurrent MVSNet for high-resolution multi-view stereo depth inference. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5525–5534

Yao Y, Luo Z, Li S, Fang T, Quan L (2022) MVSTER: Epipolar transformer for efficient multi-view stereo. In: European Conference on Computer Vision, Springer, pp 573–591

Yariv L, Gu J, Kasten Y, Lipman Y (2021) Volume rendering of neural implicit surfaces. Advances in neural information processing systems 34:4805–4815

Ye X, Zhao W, Liu T, Huang Z, Cao Z, Li X (2023) Constraining depth map geometry for multi-view stereo: Adual-depth approach with saddle-shaped depth cells. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 17661–17670

Yi KM, Trulls E, Lepetit V, Fua P (2016) LIFT: Learned invariant feature transform. In: Leibe B, Matas J, Sebe N, Welling M (eds) Computer Vision—ECCV 2016. Springer International Publishing, Cham, pp 467–483

Yu F, Koltun V, Funkhouser T (2017) Dilated residual networks. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp 472–480

Yu M, Lu T, Xu L, Jiang L, Xiangli Y, Dai B (2024a) GSDF: 3DGS meets SDF for improved neural rendering and reconstruction. Advances in Neural Information Processing Systems 37:129507–129530

Yu Z, Sattler T, Geiger A (2024b) Gaussian opacity fields: Efficient adaptive surface reconstruction in unbounded scenes. ACM Transactions on Graphics (ToG) 43(6):1–13

Zach C, Pock T, Bischof H (2007) A globally optimal algorithm for robust tv-l1 range image integration. Proc IEEE Int Conf Comput Vis. https://doi.org/10.1109/ICCV.2007.4408983

Zbontar J, LeCun Y (2015) Computing the stereo matching cost with a convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Zeng Z, Wang Y, Guan T (2025) Matching ambiguity-resilient multi-view stereo via adaptive patch deformation. Pattern Recognition p 112593

Zhan Z, Xia R, Yu Y, Xu Y, Wang X (2024) On-the-fly SfM: What you capture is what you get. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences. https://doi.org/10.5194/isprs-annals-X-1-2024-297-2024

Zhan Z, Yu Y, Xia R, Gan W, Xie H, Perda G, Morelli L, Remondino F, Wang X (2025) SfM on-the-fly: A robust near real-time SfM for spatiotemporally disordered high-resolution imagery from multiple agents. ISPRS J Photogramm Remote Sens 224:202–221. https://doi.org/10.1016/j.isprsjprs.2025.04.002

Zhang C, Tao W (2025) Learning meshing from delaunay triangulation for 3d shape representation. International Journal of Computer Vision 133(6):3413–3436

Zhang C, Yuan G, Tao W (2023) DMNet: Delaunay meshing network for 3d shape representation. In: 2023 IEEE/CVF international conference on computer vision (ICCV), IEEE Computer Society, pp 14372–14382

Zhang C, Cao Y, Zhang L (2025a) CrossView-GS: Cross-view Gaussian splatting for large-scale scene reconstruction. arXiv preprint https://arxiv.org/abs/2501.01695

Zhang F, Prisacariu V, Yang R, Torr PH (2019) GA-Net: Guided aggregation net for end-to-end stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Zhang J, Yao Y, Quan L (2021) Learning signed distance field for multi-view surface reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 6525–6534

Zhang M, Hu Z, Zhang T, Han X, Wei Z, Zhang S, Dong J (2025b) Learning neural implicit surfaces from sonar image based on signed distance functions combined with occupancy representation. Expert Systems with Applications 270:126505

Zhang S, Wei Z, Xu W, Zhang L, Wang Y, Zhang J, Liu J (2024) Edge aware depth inference for large-scale aerial building multi-view stereo. ISPRS Journal of Photogrammetry and Remote Sensing 207:27–42

Zhang S, Wei Z, Xu W, Zhang L, Wang Y, Zhang J, Liu J (2025c) TS-SatMVSNet: Slope aware height estimation for large-scale earth terrain multi-view stereo. arXiv preprint https://arxiv.org/abs/2501.01049

Zhao F, Jiang Y, Yao K, Zhang J, Wang L, Dai H, Zhong Y, Zhang Y, Wu M, Xu L et al (2022a) Human performance modeling and rendering via neural animated mesh. ACM Transactions on Graphics (TOG) 41(6):1–17

Zhao HK, Osher S, Fedkiw R (2001) Fast surface reconstruction using the level set method. In: Proceedings IEEE workshop on variational and level set methods in computer vision, IEEE, pp 194–201

Zhao J, Oishi J, Monno Y, Okutomi M (2024) Polarimetric patchmatch multi-view stereo. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 3476–3484

Zhao Y, Chen L, Zhang X, Xu S, Bu S, Jiang H, Han P, Li K, Wan G (2022b) RTSfM: Real-time structure from motion for mosaicing and dsm mapping of sequential aerial images with low overlap. IEEE Geosci Remote Sens Lett 60:1–15. https://doi.org/10.1109/TGRS.2021.3090203