metrology

MDPI

Review

Learning-Based 3D Reconstruction Methods for Non-Collaborative Surfaces—A Metrological Evaluation

Ziyang Yan [1,2,*], Nazanin Padkan [1,3], Paweł Trybała [1], Elisa Mariarosaria Farella [1] and Fabio Remondino [1]

[1] 3D Optical Metrology Unit, Bruno Kessler Foundation (FBK), 38123 Trento, Italy; npadkan@fbk.eu (N.P.); ptrybala@fbk.eu (P.T.); elifarella@fbk.eu (E.M.F.); remondino@fbk.eu (F.R.)
[2] Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy
[3] Department of Mathematics, Computer Science and Physics, University of Udine, 33100 Udine, Italy
* Correspondence: ziyang.yan@unitn.it

Abstract: Non-collaborative (i.e., reflective, transparent, metallic, etc.) surfaces are common in industrial production processes, where 3D reconstruction methods are applied for quantitative quality control inspections. Although the use or combination of photogrammetry and photometric stereo performs well for well-textured or partially textured objects, it usually produces unsatisfactory 3D reconstruction results on non-collaborative surfaces. To improve 3D inspection performances, this paper investigates emerging learning-based surface reconstruction methods, such as Neural Radiance Fields (NeRF), Multi-View Stereo (MVS), Monocular Depth Estimation (MDE), Gaussian Splatting (GS) and image-to-3D generative AI as potential alternatives for industrial inspections. A comprehensive evaluation dataset with several common industrial objects was used to assess methods and gain deeper insights into the applicability of the examined approaches for inspections in industrial scenarios. In the experimental evaluation, geometric comparisons were carried out between the reference data and learning-based reconstructions. The results indicate that no method can outperform all the others across all evaluations.

Keywords: 3D reconstruction; deep learning; non-collaborative surfaces; quality control

Academic Editor: Ki-Nam Joo

Received: 14 February 2025
Revised: 14 March 2025
Accepted: 26 March 2025
Published: 3 April 2025

Citation: Yan, Z.; Padkan, N.; Trybała, P.; Farella, E.M.; Remondino, F. Learning-Based 3D Reconstruction Methods for Non-Collaborative Surfaces—A Metrological Evaluation. *Metrology* **2025**, *5*, 20. https://doi.org/10.3390/metrology5020020

1. Introduction

Traditional methods for image-based 3D reconstruction, such as photogrammetry and photometric stereo, have been employed for a long time to measure 3D shapes of objects in industrial scenarios. These methods have proven to be precise [1], cost-effective [2], light, and portable [3], as well as flexible [4]. They are used for quality control [5], reverse engineering [6], object inspection [7,8], or 3D micro-measurement [9]. However, achieving high-quality and consistent image-based 3D reconstruction of non-collaborative surfaces is still an open issue in the industrial field [10]. Objects with non-collaborative surfaces, such as glass, shiny metals, or transparent materials, are commonly found in production processes (Figure 1). Dealing with these types of objects is a pivotal issue for 3D reconstruction tasks. Due to the sensitivity of photogrammetric methods to texture properties, it is generally difficult to obtain accurate 3D reconstruction results from non-collaborative surfaces using traditional approaches [11] Indeed, all methods based on standard feature extraction and matching for image orientation hardly achieve satisfactory results in cases of a lack of a sufficient number and quality of image correspondences [12]. This issue becomes apparent when dealing with transparent or metallic objects with smooth and featureless surfaces [13], as 3D imaging is heavily influenced by refraction and specular reflections, resulting in noisy 3D reconstruction.

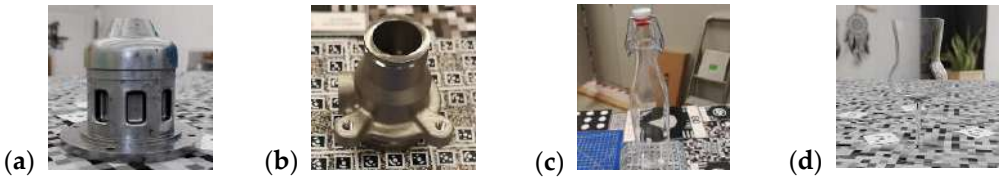


Figure 1. Examples of objects featuring non-collaborative surfaces: reflective and texture-less (**a**,**b**) or transparent and refractive (**c**,**d**).

Recently, different novel learning-based methods have been developed to overcome the aforementioned issues. They include Neural Radiance Fields (NeRF), learning-based Multi-View Stereo (MVS), Monocular Depth Estimation (MDE), Gaussian Splatting (GS), and image-to-3D generative AI. In previous studies, these methods have demonstrated their ability to estimate the geometry of indoor spaces or outdoor scenarios [14–16]. However, the literature still lacks an in-depth, critical, and quantitative analysis to understand the real potential of learning-based methods for the 3D reconstruction of texture-less, reflective, or transparent objects typical of the industrial sector. Although [14] provided an initial metric comparison between results obtained with popular NeRF frameworks and photogrammetry, they primarily looked into the results achieved by a single learning-based solution (instant-NGP). Due to the dynamic of the field, a growing number of methods are becoming available. Moreover, the industrial quality control sector has strict requirements. Therefore, there is a need to exhaustively assess the quality and reliability of a given method to identify advantages and limitations in handling non-collaborative surfaces. The findings of such analysis may impact the implementation of these methods in real applications and contribute to resolving issues identified during the process for future developments.

In this paper, we review and investigate the potential of diverse types of 3D reconstruction approaches for industrial object inspection and metric measurements. We applied NeRF, MVS, MDE, GS, and image-to-3D generative AI to a variety of industrial objects with problematic surface characteristics, such as texture-less, shiny, reflective, and transparent objects (Figure 1). Some of these objects are included in the NeRFBK dataset ([17]—https://github.com/3DOM-FBK/NeRFBK (accessed on 25 March 2025)). We used standard metrics that are commonly applied in photogrammetric processes to assess the quality of the 3D reconstruction results and analyzed the results generated by each technique in terms of reconstruction completeness, geometric accuracy, and precision. Therefore, the main objectives of this work were as follows:

(i) to report the available learning-based methods for the 3D reconstruction of industrial objects and, in general, non-collaborative surfaces.
(ii) to objectively evaluate the quality of 3D reconstructions generated by NeRF, MVS, MDE, GS, and generative AI methods.
(iii) to provide a clear summary of the advantages and limitations of such methods for 3D metrology tasks.

2. State of the Art

2.1. NeRF

Neural Radiance Fields (NeRF) is a family of view-synthetic methods. Each of them uses a set of images, together with their associated 3D camera positions and view directions (i.e., oriented images), as input and outputs for the volume density and view-dependent emitted radiance [13]. This principle is shown in Figure 2. All the NeRF-based approaches use a neural network, which learns the volumetric 3D representation of an object from multi-view 2D images. Then, by feeding the network a new camera position and view direction, it can perform so-called novel view synthesis, which predicts the emitted color

and volume density of the scene seen from the selected pose. To obtain the explicit 3D geometry, the depth maps of different views generated by applying maximal likelihood estimation of depth distribution in each camera ray can be used. They can be fused to directly derive the point clouds or can be fed into one of the surface generation algorithms, such as the Marching Cubes [18], to generate 3D meshes [14].

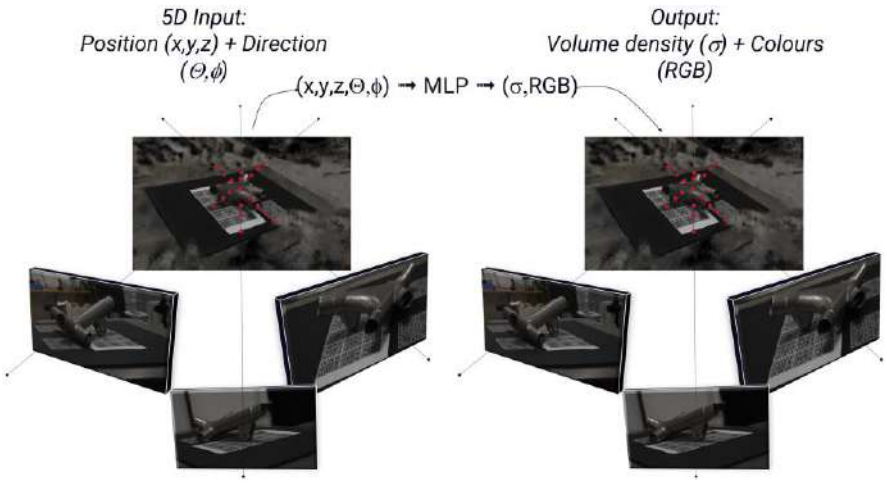


Figure 2. The basic concept of NeRF scene representation (built upon 19]).

NeRF is a type of neural implicit representation method which encodes a scene using an optimizable continuous function. Assuming that the 3D object is located at the center of the modeled space, for each 3D point coordinate $p = (x, y, z)$ and its view direction $d = (\theta, \phi)$, we can obtain the following relationship (Equation (1)):

$$(\sigma,\ c) = f\theta(p, d) \tag{1}$$

where $\sigma,\ c$ are the density and color of a point and θ is the parameter of the continuous function f.

The implementation structure of a NeRF mainly consists of an encoder and a decoder. The encoder usually leverages a convolutional neural network (CNN) which is responsible for extracting the spatial position and perspective features of each point in the scene from the input multiple-view images and camera parameters. Each convolutional layer in the encoder can map the input data from a low-dimensional space to a high-dimensional space and extract more complex feature representations.

The decoder is usually represented as an MLP network that generates a continuous 3D radiation field from the features extracted by the encoder. It accepts as its input the spatial location and perspective features of each point produced by the encoder, and outputs the color and density values of the point. Each MLP layer in the decoder can map the input data to another high-dimensional space and extract more complex representations. The original NeRF implementation [19], as well as subsequent derivative methods, utilize a non-deterministic stratified sampling approach, which is described by Equation (2). This method involves dividing the ray emitted from the direction of the camera into N equally spaced bins and uniformly drawing a sample from each bin. Finally, we obtain the ray color (i.e., the pixel color) by ray marching. Image rendering is performed by repeating the ray casting for each pixel.

$$\mathrm{C}(r) = \sum_{i}^{N} \mathrm{T}_i(1 - \exp(-\sigma_i\delta_i))c_i \tag{2}$$

where $T_i = exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right)$, and $\delta_i = t_{i+1} - t_i$ is the interval between adjacent sampled points.

The original method calculates the loss function that compares a rendered pixel value for camera ray r with the corresponding ground truth pixel value, C(gt), for all the camera rays of the target view with pose p. Thus, the loss function ℓ is given by Equation (3).

$$\ell = \sum_{r \in R(p)} \|\mathrm{C}(r) - C(gt)\|_2^2 \tag{3}$$

where $R(p)$ is the set of all camera rays of target pose p.

2.1.1. Multi-View Dependent NeRF

The NeRF approach uses an MLP network to represent a 3D scene as a learnable, continuous volumetric scene function and render the scene by optimizing the scene function. However, the original NeRF implementation can only deal with simple reflection scenarios and struggles with processing complex non-collaborative surfaces of industrial objects. Recently, more studies related to NeRF have started to pay attention to non-collaborative surfaces. Ref. [20] introduced NeRFReN which uses separate transmitted and reflected Neural Radiance Fields to process complex reflection scenes. Ref. [21] designed Dex-NeRF, which estimates the depth from transparent objects through a transparency-aware depth rendering method based on finding the first sample along the ray whose density is higher than the fixed threshold. Ref-NeRF [22] decomposes specular reflection and diffuse reflection from the target object and uses the viewing vector estimated by MLP to render the scenes. Nevertheless, it results in an enormous increment of parameters and computation. IBL-NeRF [23] ingeniously classifies indoor reflection from an indoor scene by prefiltered radiance fields. The limitation is that it is suitable for large-scale scene rendering rather than single object rendering, and this is especially relevant for the reconstruction of isolated transparent objects with perfect-mirror reflection. To improve the performance with less-observed and texture-less areas, MonoSDF [24] applies monocular geometry prediction and utilizes depth and normal cues predicted by monocular estimators. However, the reconstruction results are susceptible to the changes in the quality of the cues. Neuralangelo [25] is the advanced version of Instant-NGP [26], which combines multi-resolution 3D hash grids with neural surface rendering to reduce noise and then facilitate high-fidelity 3D surface reconstruction from large-scale scenes. Previous research [17] has shown that, in addition to outdoor and large-scale scenes, Neuralangelo also has great prospects in handling non-collaborative surfaces. However, multi-view dependent NeRF requires dozens or even hundreds of images to achieve high-quality scene rendering. The insufficient input of images or lack of images from specific perspectives will often cause geometric shape errors and missing surface areas [27].

2.1.2. Few/Single Shot NeRF

The limitation of multi-view dependent NeRF promotes the research and development of few-shot NeRF and even single shot NeRF [27–40]. Some of the existing methods achieve this goal by regularizing the geometry of the scene. DS-NeRF [27] utilizes sparse depth outputs from Structure-from-Motion (SfM [41]) as supervision, while DDP-NeRF [33] further obtains dense depth supervision from sparse inputs through a CNN network. RegNeRF [30] regularizes the geometry and appearance by proposing a depth smoothness loss and a pre-trained normalizing flow color model. Moreover, SimpleNeRF [35] trains two additional models, which, respectively, reduce positional encoding frequencies and remove view-dependent components. Pixelnerf [37] performs single-view reconstruction by first extracting the features from the input image through a CNN network and then projecting

the points sampled from the camera ray onto the image plane. The novel perspectives are then rendered by applying bilinear interpolation between pixel features to extract the corresponding image feature vector. Applying the diffusion model with NeRF is also a common way to achieve this goal. DiffusioNeRF [38] uses a trained diffusion model that can regularize the distribution of RGB-D patches from perturbed viewpoints. GANeRF [32] learns the patch distribution of the scene using an adversarial discriminator that provides feedback for radiation field reconstruction, thereby improving realism in a 3D consistent manner. ReconFusion [39] utilizes a diffusion prior-based NeRF by utilizing CLIP [42] to embed the feature vectors and PixelNeRF [37] to render a feature map. The latter includes the corresponding camera and geometric information, allowing the diffusion model to predict and generate novel perspectives.

However, for learning-based methods based both on geometric regularization and a diffusion prior, achieving high-quality rendering of non-collaborative surfaces with complex or fully transparent surfaces is still a tremendous challenge, especially when using only a single or a few views. A possible cause of this is that the pre-trained models of the aforementioned methods generally do not include a variety of non-collaborative surfaces in their training datasets [43].

2.2. Gaussian Splatting (GS)

Contrary to NeRF, Gaussian Splatting is an explicit representation method that directly and explicitly represents the geometric distribution of a surface/volume as a function with some parameters, such as a voxel grid or a set of 3D points. The concept of Gaussian Splatting was first introduced in EWA Splatting [44]. Ref. [45] proposed the application of 3D Gaussian Splatting to scene reconstruction and view synthesis, marking a significant milestone in advancing 3D reconstruction (Figure 3).

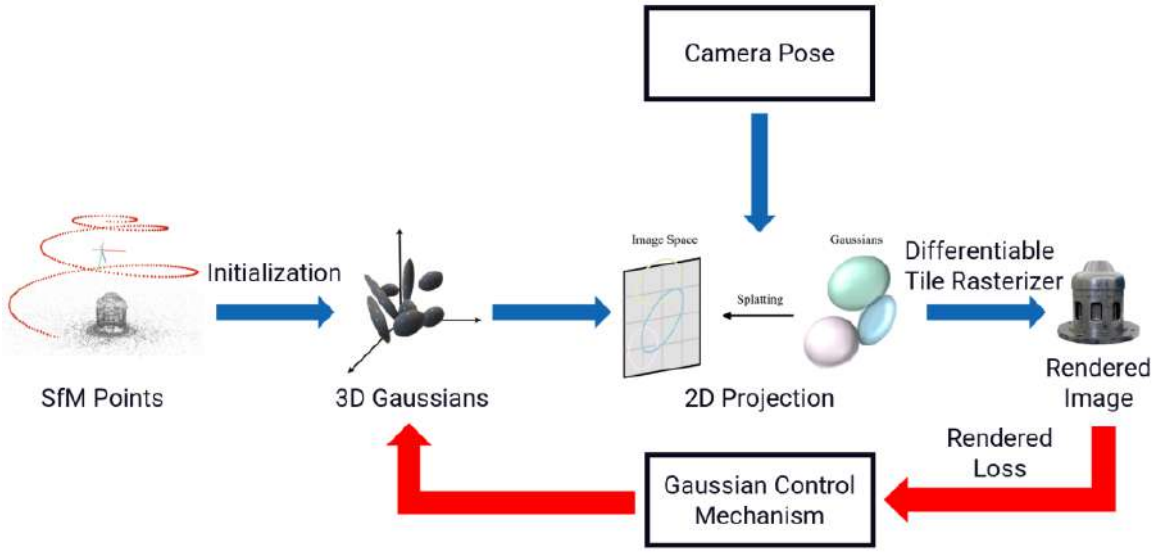


Figure 3. The basic concept of 3D Gaussian Splatting (based on [46]).

The Gaussian Splatting pipeline takes a sparse point cloud estimated from SfM [41] as its input to initialize a Gaussian set. Each Gaussian point x is then represented by a full 3D covariance matrix Σ in world space and its center position μ (Equation (4)):

$$G(x) = e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)} \quad (4)$$

To ensure the validity of Σ, it is decomposed into the scaling matrix S and the rotation matrix R to characterize the geometry of a 3D Gaussian ellipsoid (Equation (5)):

$$\Sigma = RSS^T R^T \tag{5}$$

Then, the 3D Gaussians are projected to 2D for rendering by computing the camera space covariance matrix Σ′ (Equation (6)):

$$\Sigma' = JW\Sigma W^T J^T \tag{6}$$

where J is the Jacobian matrix of the affine approximation of the projection transformation and W is the viewing transformation. The color of each pixel can then be calculated by applying alpha blending with sorted depths of these Gaussians (Equation (7)):

$$C = \sum_{i}^{N} c_i a_i \prod_{j}^{i-1} (1 - a_j) \tag{7}$$

where c_i is the rendered color of a 3D Gaussian, and a_i is the product of an evaluated 2D Gaussian projection and its corresponding opacity.

Gaussian Splatting achieves real-time rendering by explicitly representing scenes as a collection of Gaussians [47]. It not only retains the easy-to-optimize characteristics of continuous scene representation functions established by NeRF but also applies a fast GPU sorting algorithm together with tile-based rasterization that supports anisotropic splatting [48] Then, the Gaussian parameters are optimized via a loss calculated by the stochastic gradient descent (SGD) [49]. The optimized Gaussian is finally rendered through differentiable Gaussian rasterization and outputs the color and opacity of each pixel. After the publication of the first modern approach presented by [47], a large number of GS-based research works [50–58] have sprung up in just a few months after the launch of Gaussian Splatting. To further optimize Gaussian Splatting, Ref. [59] proposed a multi-scale Gaussian Splatting method that reduces the aliasing produced in signal sampling by adjusting the size of the Gaussians based on the image resolution.

Similar to NeRF, the development of Gaussian Splatting is also moving towards few/single views as the input. Ref. [60] proposed a dense depth map generated using a pre-trained MDE model to mitigate the overfitting rendering problem that occurs in the novel viewpoints. SparseGS [61] incorporates depth and diffusion constraints along with an artifact's removal technique to further improve the quality of generated novel perspectives. Using only a few views as the input for SfM, the generated point clouds will be very sparse. FSGS [62] addresses this issue by introducing a Proximity-guided Gaussian Unpooling algorithm to densify the initial sparse point cloud. This method applies KNN to grow external 3D points based on the Euclidean distance from the closest original 3D points.

The efficiency of Gaussian Splatting has led to its swift adoption across various domains, such as 3D geometry generation [63–66], dynamic scene rendering [67–69], the creation of animatable 3D human models [70–74], real-time surgical reconstruction [75], and SLAM [55,76,77]. However, there is little research that specifically investigates the performance of 3D Gaussian models in reconstructing non-collaborative surfaces. Scaffold-GS [78] introduces a hierarchical 3D Gaussian scene model with anchor points initialized from SfM to enhance the ability to capture scene local details, especially for reflective, transparent, or texture-less regions. GaussianShader [79] employs a streamlined shading function on 3D Gaussians to improve the accuracy of normal estimation, thus elevating rendering quality in scenes featuring reflective surfaces. In our experiments (S. 4), we applied original 3D Gaussian Splatting [45], implemented in the Nerfstudio [80] platform,

FSGS [62], GaussianShader [79], and Scaffold-GS [78], to the metallic and transparent objects included in the NeRFBK dataset to assess their reconstruction capabilities.

2.3. Learning-Based MVS

MVS is a wide term encompassing all methods utilizing multiple images, taken from known poses, to perform a dense 3D reconstruction of a scene [15]. Given that SfM, commonly applied to orient the image set, creates only a sparse point cloud which describes the object geometry with an insufficient level of detail for most tasks, MVS is often used as the next step of the photogrammetric 3D reconstruction process. The general principle of MVS is based on a search for corresponding points in the 3D space of every pixel in the input images. In the standard workflow, the correctness of found matches and their 3D positions is facilitated using consistency metrics such as the Sum of Squared Differences (SSD), the Sum of Absolute Differences (SAD), and Normalized Cross-Correlation (NCC) [81].

Even if this traditional approach can achieve 3D high precision under favorable conditions, it still struggles with scenes with specular reflection or weakly textured surfaces due to the over-reliance on the geometric consistency and the camera-point visibility intersections [82]. With a strong matching ability, the CNN-based 3D reconstruction can better introduce global semantic information [83]. Ref. [84] introduced SurfaceNet, the first learning-based MVS, which uses voxel-wise view selection to precompute the cost volume and to use the CNN network to represent surface voxels. Since this, learning-based methods have shown remarkable improvements compared to traditional methods. The most typical characteristic of learning-based MVS is that it usually estimates the dense depth map using deep CNNs. MVSNet [85] leverages an end-to-end deep learning architecture to infer depth maps, and then build and regularize the 3D cost volume by feature warping using 3D CNNs, establishing a foundation for future advancements in the field. Further research works have built upon its structure [86–96].

Taking the original MVSNet as an example, the common workflow of a learning-based MVS can be described as follows (Figure 4): firstly, a feature map including the deep features of each input image is extracted by a convolutional network. By using a variable differential homography transformation, a 3D feature volume is created from a 2D feature map. The variance calculation method is applied to merge N feature volumes into a cost volume. Next, the 3D convolution process is employed to calculate the probability of each depth value, followed by the use of the weighted average of the depth to derive the predicted depth information. Photometric and geometric consistencies with original images are then combined to optimize the reconstruction results.

Figure 4. The common steps in a learning-based MVS workflow.

According to the types of 3D surface representation, we can classify existing MVS methods into four categories: volumetric-based [84,97], direct point cloud [87,98], mesh-based [99], and depth map-based [100–103]. Compared to the others, the depth map-based approach is more flexible and robust due to its decomposition of the reconstruction task into two stages of per-view estimation and multi-view fusion [85].

To optimize memory consumption, refs. [86,94] leveraged recurrent networks to regularize cost volumes. Gbi-Net [104] includes a discrete binary search in MVS to further reduce memory consumption in 3D cost volume calculations and achieve a better trade-off between efficiency and accuracy. In MVSTER [105], an epipolar Transformer architecture is utilized to aggregate multi-view features and speed up the training by significantly

reducing the demand for depth hypotheses. GeoMVSNet [106] adopts geometric priors and embeddings to eliminate external dependencies of cost marching. It also enhances the full-scene depth perception using Gaussian-Mixture Model (GMM) distribution instead of traditional, uniform depth distribution.

2.4. Monocular Depth Estimation (MDE)

Monocular Depth Estimation (MDE) refers to the ill-posed problem of estimating depth from a single RGB image. It has a wide potential range of applications, such as aiding in scene comprehension, 3D modeling, robotics, and autonomous driving. Given the undeniable power of deep learning across various fields of computer vision, recent progress has also had a significant impact on MDE. The first MDE algorithm based on deep neural networks was introduced by [107], which was based on a coarse-to-fine framework. This approach involves using two deep network stacks, which make a coarse global prediction on the entire image and then another refines the output locally.

Since then, numerous researchers have turned their attention to the development of Monocular Depth Estimation algorithms rooted in deep learning, leading to the creation of several novel approaches. First, the algorithms employed CNN-based architectures [107,108], but after the introduction of visual transformers, many authors replaced CNNs with transformers [109–112].

The pipeline of learning-based Monocular Depth Estimation typically involves an encoder–decoder network, where the only input is an RGB image [113]. The computed depth map is often an inverse relative depth map, i.e., an array with maximum values for closest objects and zeros for the farthest pixels, which helps avoid computational issues with infinite distances. The architecture of the ZoeDepth algorithm, as an example of an MDE algorithm, is presented in Figure 5.

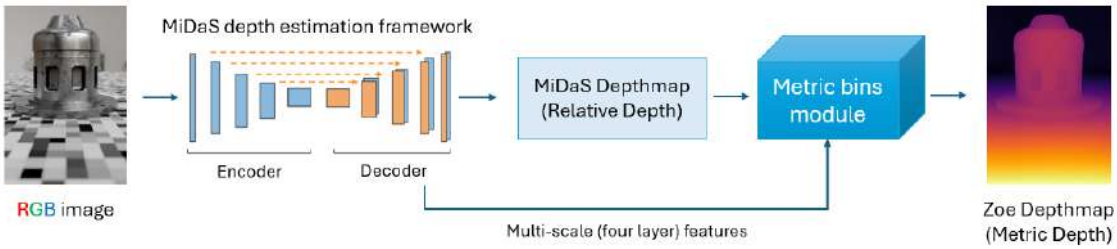


Figure 5. The general architecture of the ZoeDepth algorithm.

The RGB image is given to the MiDaS depth estimation algorithm to compute the relative depth. Then, the bottleneck and four hierarchy levels of the MiDaS decoder are connected to the metric bins module. The metric bins module computes per-pixel depth bin centers from the MiDaS decoder outputs, which are then combined to produce the final metric depth output.

The problem of estimating depth using a single RGB image can be viewed as follows. Let $I \in \mathbb{R}^{w \times h}$ be a RGB image with size $w \times h$. $D \in \mathbb{R}^{w \times h}$ is the corresponding depth map with the same size as I. For the training set τ, a non-linear mapping $\psi : I \rightarrow D$ can then be learned by the network (Equation (8)):

$$\tau = \{(I_n, D_n)\},\ I_n \in \mathbb{R}^{w \times h} \text{ and } D_n \in \mathbb{R}^{w \times h} \tag{8}$$

The majority of the MDE methods include three categories: supervised [114–116], semi-supervised [117–119], and self-supervised [120]. The formulation described above is usable for supervised learning-based MDE algorithms, where the pixel-level ground truth is available [121]. This has been the prevailing approach in recent years. In contrast, previous

works also explored self-supervising methods, which utilize only synchronized stereo pairs [122,123] or monocular videos [124] to learn how to estimate the depth from novel monocular images, and unsupervised methods, trained on images acquired by multiple cameras to generate novel viewpoints [125]. Most state-of-the-art algorithms are supervised or semi-supervised and are trained on large available datasets such as NYUv2 and KITTI.

ZoeDepth [126] integrates both absolute and relative depth estimation methods in a two-stage process. Initially, it trains an encoder–decoder model to predict relative depths from diverse datasets, benefiting from MiDaS's training strategy for scale and shift invariant loss. In the second stage, ZoeDepth enhances depth estimates by incorporating absolute depth information through metric fine-tuning on indoor and outdoor datasets like NYU Depth v2 [127,128].

MiDaS [129] leverages several depth estimation models and originates from a critical study on relative depth, highlighting the value of dataset integration for better zero-shot performance. Depth prediction occurs in disparity space, considering scale and shift, and uses invariant losses for uncertain depth labels. The MiDaS models merge different datasets, evolving with increasing data integration over subsequent iterations. The method follows a standard encoder–decoder structure based on ResNet and it is suitable for real-time applications [130].

Depth Anything [131] introduces a practical method for accurately estimating depth from single images without introducing new technical components. Through extensive dataset expansion and the implementation of effective strategies like challenging optimization objectives and supplementary supervision, it displays remarkable adaptability across different datasets and real-world scenarios. Furthermore, fine-tuning with precise depth information from NYUv2 and KITTI datasets results in an improvement of the results of both depth estimation and its applications, such as ControlNet [132].

2.5. Generative AI

Recently, generative AI (also called AIGC: AI-generated content) has achieved remarkable progress and received widespread attention. Novel generative AI methods focused on vision have gained popularity in the fields of text-to-image [132–136] and text-to-video [137–142]. Recent advantages in text-to-image using diffusion models have attracted researchers to explore the potential of applying diffusion priors in 3D vision, including text-to-3D [143–146] and image-to-3D [10,63,147–151]. Previously, the generalizability of most 3D native generation methods was limited to specific datasets [152], like constructing text-3D pairs based on ShapeNet [153], which contains only fixed object categories [154–157].

2.5.1. Diffusion Model

Inspired by the Denoising Autoencoder [158] and Score Matching [159], the diffusion model, which is a parameterized Markov chain [160] trained using variational inference to produce samples matching the data after a finite time, was first introduced by [161]. Normally, a diffusion model runs two processes. In the forward diffusion stage (Equation (9)), the distribution of Gaussian noise $q(x_t|x_0)$ is calculated and the noise is gradually applied to the image until the image is completely masked (Figure 6).

$$q(x_t|x_0) = N\left(x_t;\ \sqrt{\overline{a_t}}x_0,\ (1-\overline{a_t})\mathrm{I}\right) \tag{9}$$

where N and I are the normal distribution and identity matrix, respectively, $x_0, \ldots x_t$ are the latent representations up to a timestep t, and $\overline{a_t} = \prod_1^t 1-\beta_t$ and β_t are learnable hyperparameters.

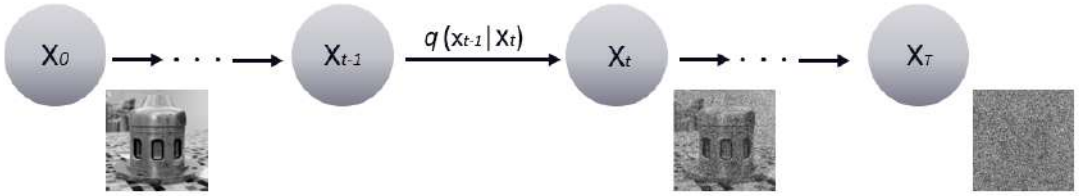


Figure 6. Schematic diagram of forward diffusion stage.

Then, in the reverse diffusion stage (Figure 7), the model learns how to restore the original image from the Gaussian noise ∈ by minimizing a variational bound for the Langevin-like reverse process [162]. In this procedure (Equation (10)), x_t serves as the input to approximate the mean and variance of a Gaussian distribution. We then randomly sample from this distribution based on the predictions to obtain x_{t-1}. By iteratively predicting and sampling, we eventually generate a genuine image. Here, $\mu_\theta(x_t, t)$ is the reverse process mean function approximator (Equation (11)), $\in_\theta$ acts as another approximator aimed at predicting ∈ from x_t, and $\sum_\theta(x_t, t)$ is a variance estimator.

$$p_\theta(x_{t-1}|x_t) = N\left(x_{t-1}; \mu_\theta(x_t, t), \sum_\theta(x_t, t)\right) \tag{10}$$

$$\mu_\theta(x_t, t) = \frac{1}{\overline{a_t}}(x_t - \frac{\beta_t}{\sqrt{(1-\overline{a_t})}}\in_\theta(x_t, t)) \tag{11}$$

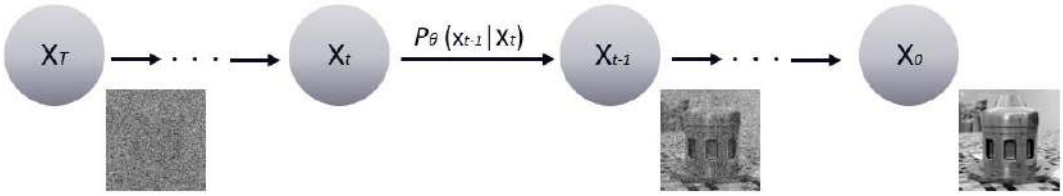


Figure 7. A schematic diagram of the backward diffusion stage. (P_θ is the distribution of the Gaussian noise).

2.5.2. Image-to-3D by Diffusion Prior

Although most of the state-of-the-art text-to-3D methods can generate diverse shapes under the guidance of prompts, the output geometry usually suffers from the low level of details and shape ambiguity of the object described by a short text. In addition, non-collaborative surfaces are often untextured and have complicated geometry, which is difficult to describe in detail. Therefore, applying text-to-3D in an industrial setting to objects with non-collaborative surfaces poses a great challenge.

However, the development of domain transfer learning, such as zero-shot learning and multimodal large language models (MLLMs) makes it possible to fine-tune an MLLM using some specific datasets and to enable it to learn multi-view synthesis and view control. Ref. [39] applied the diffusion model in view synthesis utilizing the ShapeNet dataset. Ref. [163] fine-tuned Stable Diffusion [164], a diffusion model trained by billion-level text-image pairs, and thus proposing the Zero123 model. Their approach can extract the object's geometric features from a single input image and infer novel perspectives by parsing its semantic information. Instead of fine-tuning the 2D diffusion model, MVDream [165] converts the original 2D self-attention layer into 3D by connecting different views in the self-attention layer and adding the camera embedding of each input view into temporal embedding as a residual to optimize the accuracy of generated novel perspectives. To improve the geometric inferential capability of the fine-tuning diffusion models, SyncDreamer [166], Consistent 1-to-3 [167], and Zero123plus [168] have sought to enhance multi-view consistency through joint diffusion processes. ImageDream [169] does this by additionally adding

a textual prompt as a constraint. PC2 [170] facilitates single image object generation from a randomly generated spherical Gaussian point set to a specific 3D geometry by applying a ViT [171] to extract the 2D representation and to use it as a clue to control Gaussian point generation. Even though they were only trained on the CO3D dataset [172], the development of efficient MLLM fine-tuning methods [173–176] makes it promising for applications in other fields. Based on the reliability of inferential views, the final dense 3D shape representation can be easily obtained by connecting its output to a NeRF [145,177–184], Gaussian Splatting [63,66,185–187], or other 3D reconstruction approaches.

3. Analysis and Evaluation Methodology

This section presents a critical evaluation of the abovementioned learning-based 3D reconstruction methods by objectively measuring their capability in dealing with non-collaborative surfaces. To accomplish this, some industrial objects of different sizes and surface characteristics are considered, including texture-less, metallic, translucent and transparent (Table 1). The proposed evaluation strategy and metrics aim to support researchers in understanding the strengths and limitations of each approach.

Table 1. A summary of the objects used in our analyses and available in the NeRFBK dataset [17].

	Industrial_A	**Synthetic Metallic**	**Synthetic Glass**
			
Numb. images, resolution	290 images 1280 × 720 px	300 images 1080 × 1920 px	300 images 1080 × 1920 px
Ground truth (GT)	Triangulation-based laser scanner	Synthetic data	Synthetic data
Characteristics	Texture-less/small and complex	Texture-less/complex/reflective	Transparent/highly refractive

3.1. Proposed Assessment Methodology

The assessment procedure is shown in Figure 8. The 3D reconstructions from three different datasets are compared with the available ground truth (GT) data in order to derive quantitative metrics.

All collected images or videos required camera poses in order to generate a 3D reconstruction, either with MVS, GS, or NeRF-based methods. Starting from the available unoriented images, camera poses were retrieved, for all datasets using COLMAP [188]. Then, the selected learning-based methods were applied to generate dense 3D geometries.

For evaluating the MDE method, a subset of images (4–16) for each object taken from different viewpoints was selected and used as the input into the candidate networks. Then, the predicted depth maps and known camera parameters were used to create 3D point clouds. Finally, all produced point clouds were co-registered and rescaled with respect to the available ground truth (GT) data in Cloud Compare using an Iterative Closest Point algorithm [189], and a quality evaluation was performed. To provide an unbiased evaluation of geometric accuracy, different well-establish photogrammetric criteria were applied [2,190], including best plane fitting, cloud-to-cloud comparison, profiling, accuracy, and completeness.

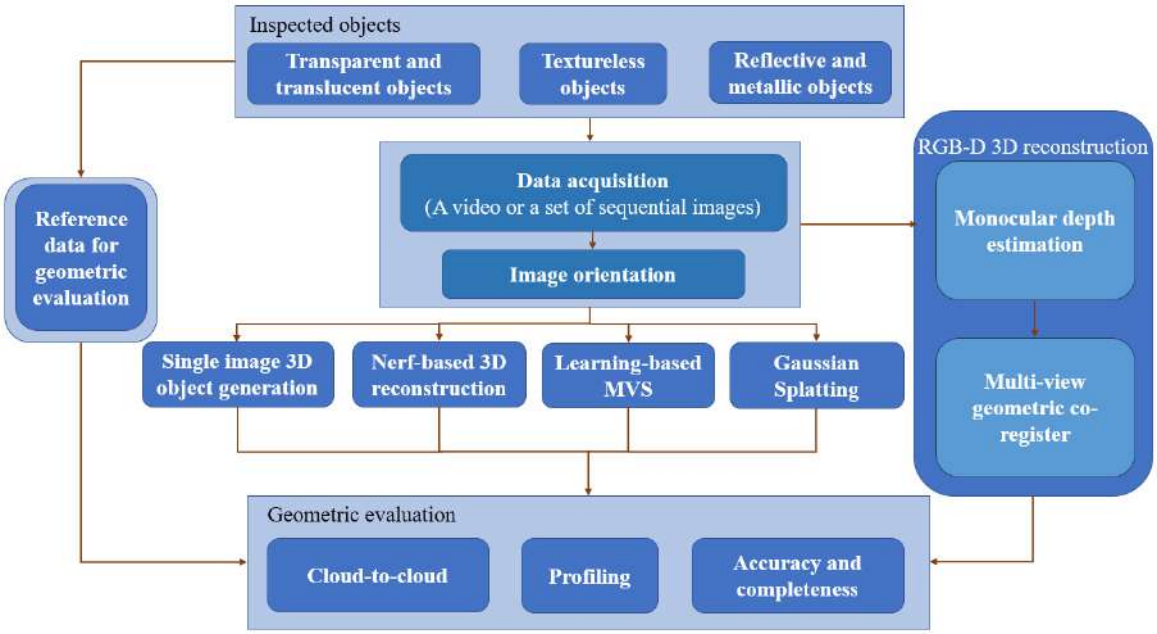


Figure 8. An overview of the proposed procedure to assess the performance of different 3D reconstruction methods.

For some selected objects, profiling was additionally conducted by extracting a cross-section from the 3D data to highlight complex geometric details of the reconstructed surface. An inspection of profiles allowed us to evaluate the performance of a method in preserving geometric details, such as edges and corners, and avoiding smoothing effects. Cloud-to-cloud (C2C) comparisons refer to the measurement of the Euclidean distance between corresponding closest points in the evaluated and GT point cloud.

3.2. Metrics

To quantitatively compare the differences between methods, similarly to other works in the industrial community, we applied statistical metrics for cloud-to-cloud and plane fitting processes, including the mean error (Mean_E, or $\overline{X}$) (Equation (11)), the standard deviation (STD) (Equation (12)), the root-mean-square deviation (RMSD) (Equation (13)), and the mean absolute error (MAE) (Equation (14)). The mean error indicates the average difference between reconstructed surfaces, while RMSD represents the general level of compliance with the GT model. MAE can be used to reflect the discrepancies between the actual and predicted point positions, and STD measures the precision of the reconstructed surface.

$$Mean\_E = \overline{X} = \frac{(X_1 + X_2 + \cdots X_j)}{N} \tag{12}$$

$$STD = \sqrt{\frac{\sum_{j=1}^{N}(X_j - \overline{X})^2}{N}} \tag{13}$$

$$RMSD = \sqrt{\frac{\sum_{j=1}^{N}(X_j)^2}{N}} \tag{14}$$

$$MAE = \frac{\sum_{j=1}^{N}|X_j|}{N} \tag{15}$$

where N denotes the number of observed point clouds and X_j denotes the closest distance of each point to the corresponding reference point or surface.

Other important metrics include accuracy (Equation (15)) and completeness (Equation (16)), sometimes also called precision and recall [14,190,191]. The accuracy measures what ratio of the reconstructed points lies within a certain distance from the GT, while completeness reflects the percentage of points that have been reconstructed within a given tolerance. The metrics were calculated using different threshold distances Th to obtain the percentage of points that fall within it, which allowed for the plotting of accuracy and completeness curves and a detailed description of the quality of the reconstruction.

$$Accuracy = \frac{\sum_{i=1}^{S}(DisT_i < Th)}{S} \quad (16)$$

$$Completeness = \frac{\sum_{i=1}^{T}(DisS_i < Th)}{T} \quad (17)$$

where $DisT$ denotes the distances between the points of the evaluated 3D point clouds to the closest points in ground truth, and $DisS$ represents the distance for an opposite relationship. S and T are the total number of points in the investigated point cloud and ground truth, respectively.

4. Comparison and Analysis

4.1. Testing Objects and Methods

To achieve the study objectives, two datasets of industrial objects and one of a transparent glass available in the NeRFBK dataset [17] were used (Table 1). They feature objects of distinctive characteristics and surface types, with different lighting conditions, materials, sensor types, scales, and resolutions.

We considered four categories (NeRF, Gaussian Splatting, learning-based MVS, MDE, and generative AI) and a total of 30 methods (Table 2) chosen among open-source codes available in Github repositories. The NeRF methods were integrated in NeRFStudio [80] and SDFStudio [192], whereas all other tools are available from Github repositories. It is crucial to mention that the generative AI methods achieve a 3D reconstruction with just a single image as input. Therefore, after background removal, the most representative view of each object was selected for processing.

Table 2. The evaluated methods for the 3D reconstructions of texture-less, metallic, translucent, and transparent objects.

NeRF (Section 4.2)				
Instant-NGP [26]	Mono-Neus [192]	MonoSDF [24]	Mono-Unisurf [192]	Nerfacto [80]
Neuralangelo [25]	NeuS [193]	Nerfacto (w/depth) [80]	Nerfacto (w/o depth) [80]	Unisurf [194]	VolSDF [195]
Gaussian Splatting (Section 4.2)				
FSGS [62]	GaussianShader [79]	Gaussian Splatting [45]	Scaffold-GS [78]	
Learning-based MVS (Section 4.2)				
DI-MVS [196]	ET-MVSNet [197]	GBi-Net [104]	GeoMVSNet [106]	
KD-MVS [198]	MVSFormer [199]	MVStudio [200]	TransMVSNet [201]	
MDE (Section 4.3)				
ZoeDepth [126]	MiDaS [130]	Depth Anything [131]		
Generative AI (Section 4.4)				
One-2-3-45 [147]	DreamGaussian [63]	Magic123 [149]	Zero-1-to-3 [163]	

All experiments were performed with a single NVIDIA GeForce A40, A6000, or RTX3080TI GPU. To enable efficient comparison, only the top-ranked methods in each category are afterwords presented.

4.2. The 3D Results from Multi-View Image Sequences (NeRF, GS, Learning-Based MVS)

4.2.1. Industrial_A Object

Out of the 22 considered methods based on multi-view image sequences (Table 2), all methods, with some exceptions with five NeRF-based approaches (Mono-Unisurf, NeuS-facto, Unisurf, VolSDF, and Instant-NGP that failed to reconstruct at least 60% of the object), successfully reconstructed the object's geometry. The comparison results are reported in Table 3 for the best approach of each evaluated 3D reconstruction method. Neuralangelo achieved the best results among the NeRF methods, and MVSFromer ranked second among all methods. Their RMSDs were 0.57 mm and 0.85 mm, respectively. From a visual inspection (Table 3) it is evident that the surface reconstructed by Neuralangelo is highly consistent with the GT, as evidenced by the predominance of dark green points in the comparison. In contrast, MVSFormer shows extensive red regions, indicating significant errors, along with some scattered noise near the surface. Gaussian Splatting, the best Splatting method, exceeded 1 mm in the RMSD, achieving worse results than half of the tested NeRF and MVS methods due to its noticeably uneven surface, which significantly deviated from the GT. More visual and numerical details of the comparison are reported in Figure A1 in Appendix A.

Table 3. Metrics for the cloud-to-mesh comparisons of the best-performing methods from each category applied to the Industrial_A object.

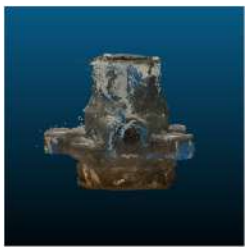
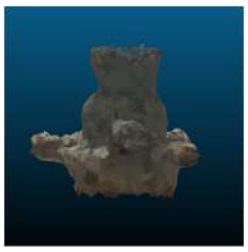
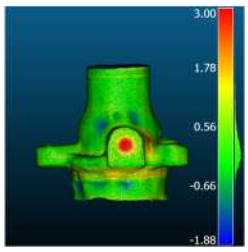
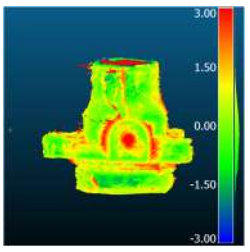
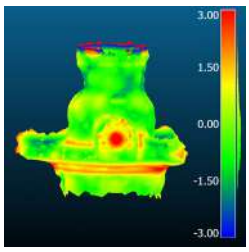
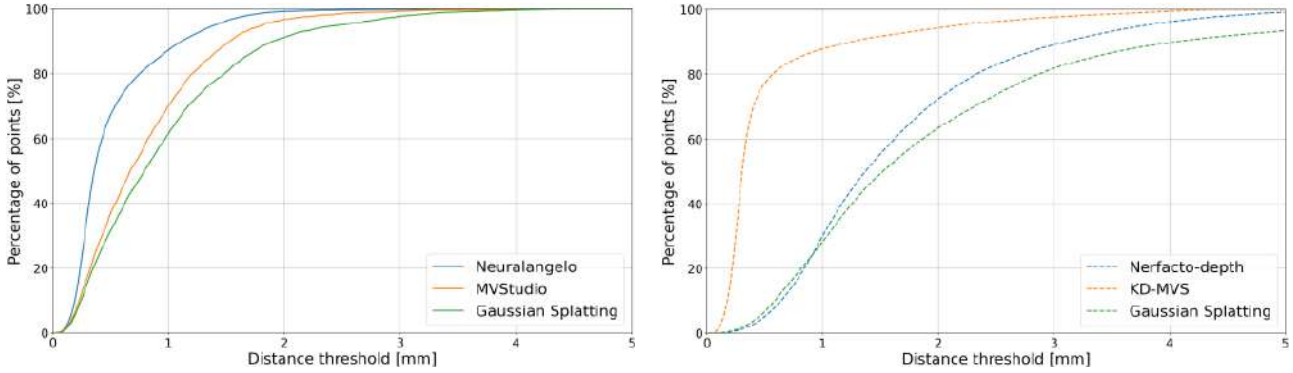
		NeRF	**Learning-Based MVS**	**Gaussian Splatting**
3D geometry				
Comparison result [mm]				
Method		Neuralangelo	MVSFormer	Gaussian Splatting
Metric [mm]	RMSD	0.57	0.85	1.11
	MAE	0.43	0.69	0.89
	STD	0.37	0.49	0.66
	Mean_E	0.13	−0.19	0.14

Figure 9 shows the accuracy and completeness of all tested methods on the Industrial_A dataset. Similarly, for the convenience of comparison, we consolidated the top-performing methods in terms of accuracy and completeness in each category for comprehensive analysis. For the other methods, please refer to Figures A2–A5 in Appendix A. In terms of accuracy, Neuralangelo outperforms the other methods, followed by MVStudio.

KD-MVS is the winner in terms of completeness, achieving higher than 85% and 90% recall within 1 mm and 2 mm, respectively, far ahead of other methods. However, KD-MVS is inferior to other methods in completeness, since its output point cloud is high density, but also contains a substantial noise.



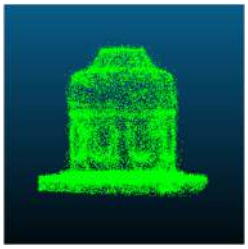
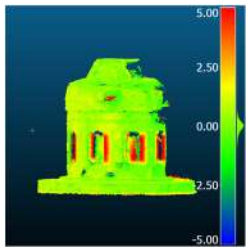
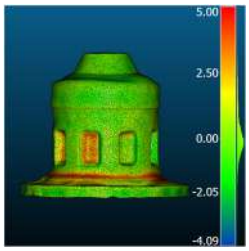
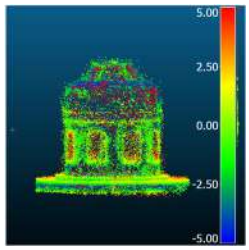
(**a**) (**b**)

Figure 9. The best accuracy (**a**) and completeness (**b**) for all tested methods using the Industrial_A object (Table 3).

4.2.2. Metallic Object

The Synthetic Metallic dataset contains 300 images (1080 × 1920 pixels) of a synthetic, reflective, texture-less, and metallic object created in Blender. We applied the same processing steps as reported in Section 4.2.1 to Industrial_A object. All approaches successfully reconstructed the geometry of the object, and the comparison results are shown in Table 4 and Figure A6 (Appendix A). Among the MVS methods, GBi-Net achieved the best results, with KD-MVS and MVSFormer ranking second and third. The RMSD of GBi-Net was 0.7 mm, almost twice as low as the first-ranked NeRF-based method (Mono-Neus). FSGS achieved a relatively inferior quality in the experiment, with the RMSD equal to 1.92 mm. Overall, the worst performing method was Instant-NGP, which generated a very noisy result, with an RMSD of 5.8 mm. The top-performing results of the accuracy and completeness of each category applied to the Synthetic Metallic dataset are presented in Figure 10, while full results for each examined method are shown in Figures A7–A10 (Appendix A). GBi-Net achieved the highest accuracy among all methods, followed by Mono-Neus and FSGS. This can be attributed to GBi-Net; as an MVS-based method, it benefits from depth supervision, allowing it to capture more precise geometric details when reconstructing objects with regular sizes. While minor estimation errors remain in the concave regions at the object's center due to depth estimation inaccuracies, its overall reconstruction accuracy was significantly higher than Mono-Neus (which exhibited geometric distortion at the object's base and chassis connection) and FSGS (which produced excessive noise points above and on top of the object). About completeness, TransMVSNet demonstrated outstanding results within a 1 mm distance threshold but ultimately was surpassed by Nerfacto and Scaffold-GS as the distance threshold exceeded 3 mm. Similarly to the findings in Industrial_A, the Gaussian Splatting methods exhibited lower accuracy and completeness in a low threshold range (<=2 mm) compared to NeRF and MVS. This discrepancy could be attributed to a limited number of initialized Gaussian points sampled from highly reflective surfaces.

Table 4. Metrics for the cloud-to-mesh comparisons of the best-performing methods from each category applied to the synthetic metallic object.

	Learning-Based MVS	NeRF	Gaussian Splatting
3D geometry			
Comparison result [mm]			

Method		GBi-Net	Mono-Neus	FSGS
Metric [mm]	RMSD	0.70	1.38	1.92
	MAE	0.61	1.10	1.49
	STD	0.35	0.83	1.22
	Mean_E	0.00	0.32	−0.41

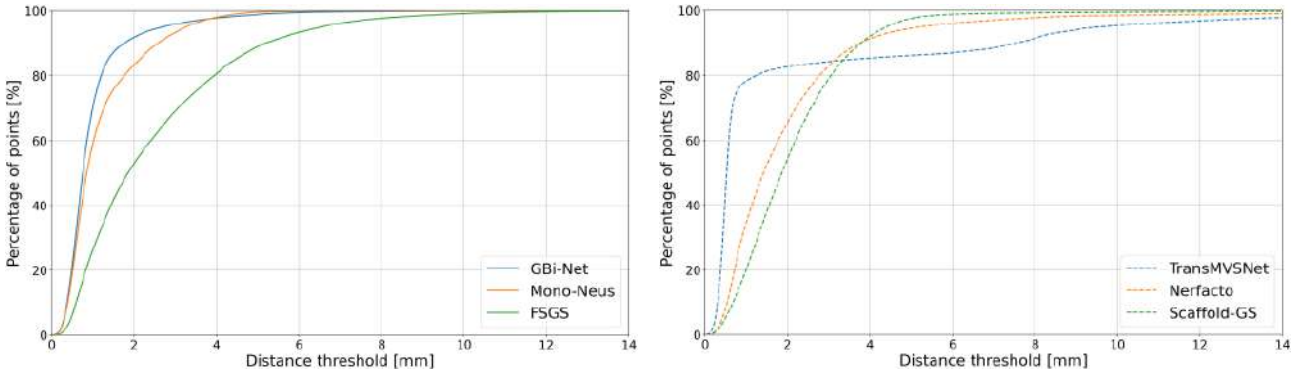


Figure 10. The best accuracy (**a**) and completeness (**b**) methods among all tested approaches for the Industrial_A object.

In addition to cloud to mesh comparisons and accuracy and completeness analyses, some cross-sections were extracted from the best-reconstructed geometries in each category (Mono-Neus for NeRF, FSGS for Gaussian Splatting, and GBi-Net for MVS) to check whether small geometric details could be reconstructed. The section location and the profiles are shown in Figure 11. The MVS profile (green point) resulted in a better match with ground truth than the others (red, purple, and blue lines) due to the depth-based nature of MVS. However, the result was full of noise on the surface. The result of NeRF (blue point) was slightly inferior to MVS. The geometric features of the cavities of the object were not accurately reconstructed. Compared to the other methods, FSGS (red points) had trouble in the geometric reconstruction of both the convex and concave parts.

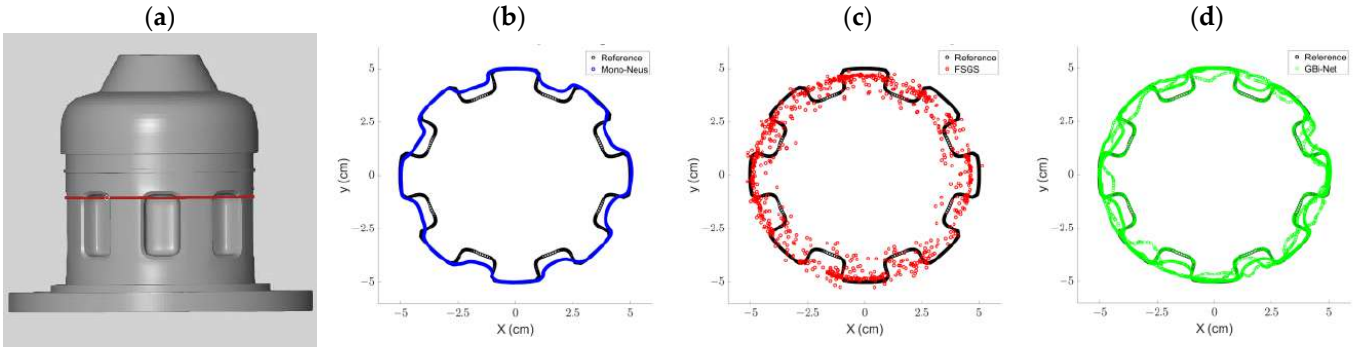
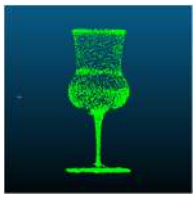
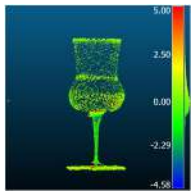
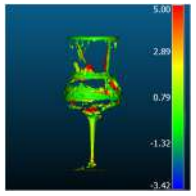
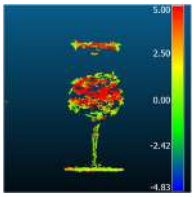


Figure 11. Cross-section profiles on the Synthetic_Metallic object reconstructed with the best method from each category: (**a**) The location of the profile on the Synthetic_Metallic object, (**b**) Mono-Neus for NeRF, (**c**) FSGS for Gaussian Splatting, and (**d**) GBi-Net for MVS.

4.2.3. Transparent Object

To evaluate the ability to deal with transparent and refractive surfaces, the methods reported in Section 4.1 were tested with the Synthetic_Glass dataset. Table 5 reports the results for the top-performing methods of each category: Gaussian Splatting achieved the best results, with 1.54 mm in RMSD, 1.22 mm in MAE, and 0.93 mm in STD. Neuralangelo and MVStudio ranked second and third with RMSD of 2.29 mm and 3.14 mm, respectively. These results also demonstrated that Gaussian explicit representation is more effective for transparent objects, whereas MVS-based methods struggle due to their reliance on depth estimation. This phenomenon was also corroborated by our MDE experiments (Section 4.3), where all MDE methods failed to reconstruct the geometry of the Synthetic_Glass object.

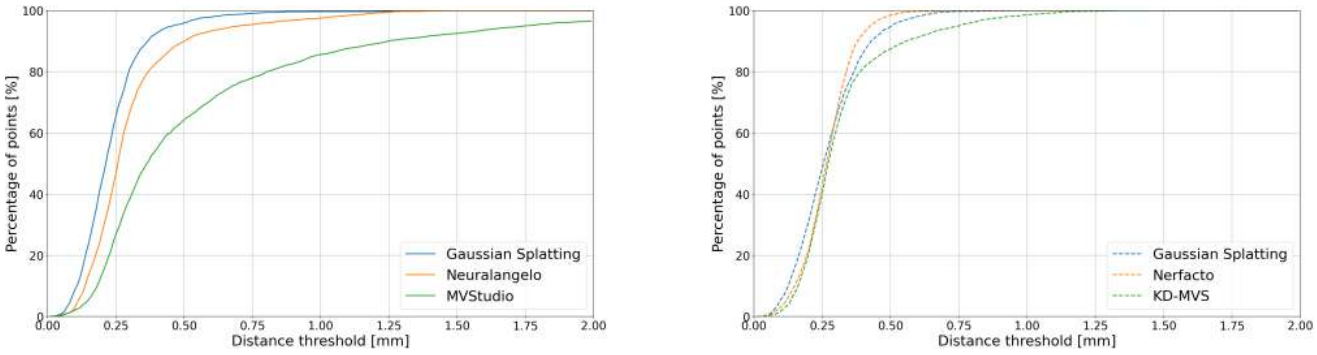
Table 5. Metrics for the cloud-to-mesh comparisons of the best-performing methods from each category applied to the transparent glass object.

		Gaussian Splatting	**NeRF**	**Learning-Based MVS**
3D geometry				
Comparison result [mm]				
Method		Gaussian Splatting	Neuralangelo	MVStudio
Metric [mm]	RMSD	1.54	2.29	3.14
	MAE	1.22	1.72	1.69
	STD	0.93	1.51	1.43
	Mean_E	0.44	1.19	0.93

For MVS methods, GeoMVSNet and GBi-Net failed to reconstruct the geometry, whereas MVStudio consistently outperformed the other methods across all metrics, despite the fact that its completeness was generally low.

Accuracy and completeness results are presented in Figures 12 and A11, Figures A12–A14 in Appendix A. In contrast to the performance observed on industrial objects, Gaussian

Splatting surpassed all competitors in accuracy, while in terms of completeness, the best method resulted from Nerfacto. It is worth mentioning that Gaussian Splatting not only outperformed all other methods in terms of accuracy but also ranked second in cross-category comparison of completeness, highlighting its notable ability to handle transparent objects. In contrast, even if Nerfacto performed worse in accuracy, it achieved better results than other methods in completeness, indicating the capability to reconstruct transparent surfaces in a denser but noisier way.



(**a**) (**b**)

Figure 12. The best accuracy (**a**) and completeness (**b**) methods among all tested approaches for the transparent glass object.

4.3. The 3D Results from Monocular Depth Estimation (MDE)

Different from other 3D reconstruction methods, an object's 3D shape can also be obtained from a single RGB image by applying the MDE method. Assuming to have different viewpoints of the object, MDE outputs are inferred depth maps per viewpoint; hence, point clouds can then be generated, e.g., using the Open3D library [202], and finally all clouds can be co-registered to create a unique 3D reconstruction of the object. As Zoedepth provides metric depth estimates, while other methods infer relative depths, we rescaled the estimated depths of the MiDaS and Depth Anything methods by employing linear regression to establish a linear relationship between pixel values in the depth image and their respective distances in meters.

The reported results refer to Industrial_A and Synthetic_Metallic objects, as no MDE methods could derive successful results on the transparent glass. The results for Industrial_A are presented in Figure 13 and Table 6. Clearly, ZoeDepth achieved better outcomes compared to MiDaS and Depth Anything. Although MiDaS attained the lowest error in View_01, its notably high standard deviation indicates relative algorithmic instability. However, it is worth noting that the accuracy of these results may be questionable due to significant geometric distortion observed in some of the generated point clouds. This distortion could potentially lead to inaccurate geometric matching when applying ICP for point cloud co-registration.

On the other hand, the results for Sythetic_Metallic are shown in Figure 14 and Table 7. Depth Anything slightly exceeded ZoeDepth in RMSD and STD, whereas the opposite trend is observed for MAE. Additionally, the average and standard deviation also indicate that, besides the lag in error metrics, the instability of MiDaS is once again noticeable in the Synthetic_Metallic object.

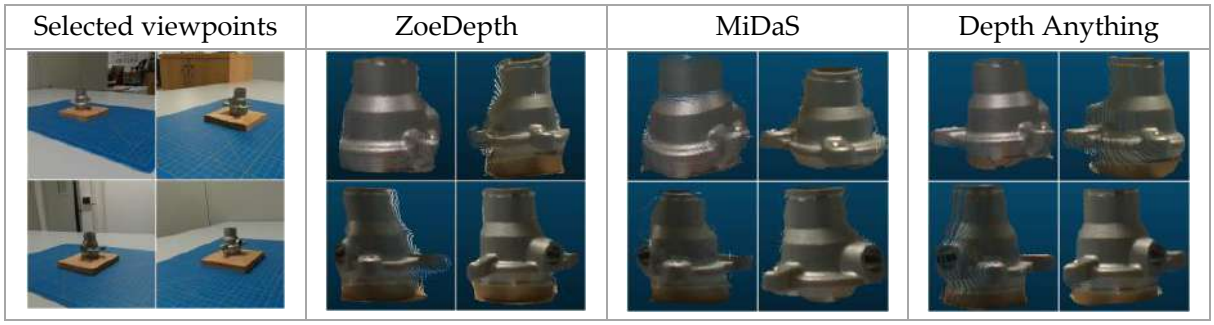


Figure 13. Visualization of MDE results for Industrial_A object.

Table 6. Metrics [mm] for the cloud-to-mesh comparisons of the tested MDE methods applied to the Industrial_A object.

Method	ZoeDepth			MiDaS			Depth Anything		
Metric [mm]	**RMSD**	**MAE**	**STD**	**RMSD**	**MAE**	**STD**	**RMSD**	**MAE**	**STD**
View_01	1.67	1.22	1.14	0.89	0.68	0.58	1.95	1.30	1.44
View_02	1.41	1.09	0.90	1.46	1.12	0.94	1.67	1.22	1.14
View_03	1.19	1.01	0.86	2.08	1.56	1.38	1.28	0.99	0.82
View_04	1.35	1.11	0.76	1.77	1.16	1.32	1.17	0.88	0.77
Average	1.41	1.11	0.92	1.55	1.13	1.06	1.52	1.10	1.04
Standard deviation	0.20	0.09	0.16	0.51	0.36	0.37	0.36	0.20	0.31

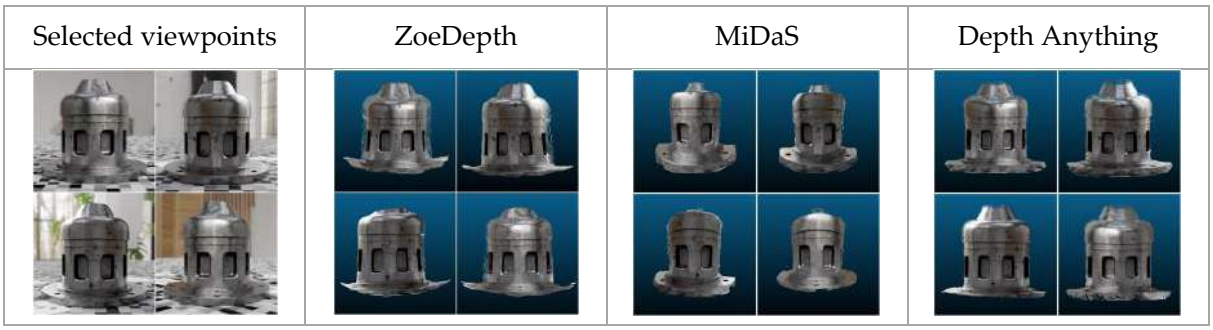


Figure 14. Visualization of MDE results of Synthetic_Metallic object.

Table 7. Metrics [mm] for the cloud-to-mesh comparisons of examined MDE methods applied to the Synthetic_Metallic object.

Method	ZoeDepth			MiDaS				Depth Anything	
Metric [mm]	**RMSD**	**MAE**	**STD**	**Metric [mm]**	**RMSD**	**MAE**	**STD**	**Metric [mm]**	**RMSD**
View_01	3.95	2.80	2.79	View_01	3.95	2.80	2.79	View_01	3.95
View_02	3.71	2.72	2.52	View_02	3.71	2.72	2.52	View_02	3.71
View_03	3.08	2.28	2.07	View_03	3.08	2.28	2.07	View_03	3.08
View_04	4.22	2.62	3.31	View_04	4.22	2.62	3.31	View_04	4.22
Average	3.74	2.61	2.67	Average	3.74	2.61	2.67	Average	3.74
Standard deviation	0.49	0.23	0.52	Standard deviation	0.49	0.23	0.52	Standard deviation	0.49

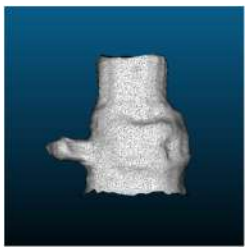
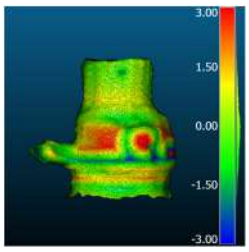
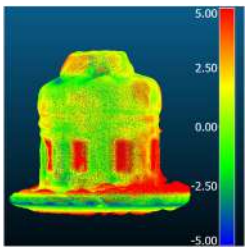
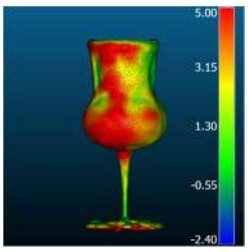
4.4. The 3D Results from Novel View Synthesis (Generative AI)

Generative AI methods use only one initial input image, firstly executing novel view inference and synthetizing new views based on the input text prompt. Then, they reconstruct the object in the same way as the multi-view methods reported in Section 4.2.

Table 8 reports the results for the top-performing generative AI methods of each tested object: Magic123 achieved the best results for the Industrial_A object, with 1.12 mm in RMSD, 0.88 mm in MAE, and 1.26 mm in STD. While Zero-1-to-3 outperformed for

both Synthetic_Metallic and Synthetic_Glass objects, the RMSD of them was 3.08 mm and 3.32 mm, respectively. The complete comparison results of each method are shown in Tables A1–A3. Since generative AI methods reconstruct full geometry from a single input image through multi-view inference, their performance is heavily influenced by the pre-trained foundation models. If these models are trained on a large amount of similar data, they are more likely to achieve better results for that specific type of object.

Table 8. Metrics for the cloud-to-mesh comparisons of the generative AI methods applied to the tested objects.

Object		**Industrial_A**	**Sythetic_Metallic**	**Sythetic_Glass**
Best Method		**Magic123**	**Zero-1-to-3**	**Zero-1-to-3**
3D geometry				
Comparison result [mm]				
Metric [mm]	RMSD	1.12	3.08	3.32
	MAE	0.88	2.46	2.76
	STD	0.68	1.84	1.85
	Mean_E	−0.04	1.26	2.39

5. Discussion

Table 9 summarizes the experimental results for each considered method and object. Synthetic_Glass presented the most significant challenge, failing for eight NeRF-based methods, two learning-based MVS, and all MDE methods. Industrial_A also posed some challenges, and six NeRF-based methods failed to produce correct 3D data. Conversely, Synthetic Metallic proved to be the easiest object, with all methods being able to reconstruct its geometry.

The results reported in Section 4 indicate that none of the AI-based methods always outperformed the others in all tested scenarios, although in each category, certain approaches emerge as clear winners. For NeRF-based methods, Mono-Neus stands out as the undisputed champion, achieving first place in both Industrial_A and Synthetic_Metallic objects despite its shortcomings in dealing with transparent objects. In generative AI, Zero-1-to-3 took the crown by securing the top spot in Synthetic_Glass and Synthetic_Metallic, and second place in Industrial_A, showing its outstanding generalization across different objects. However, determining a definitive winner in learning-based MVS proves challenging as no single approach demonstrates outstanding performance across multiple object types.

In terms of accuracy, no approach can be unequivocally deemed as the winner due to the challenge of achieving consistently stable and excelling performance across all test scenarios. Nerfacto emerged as a frontrunner in completeness within the tested scenarios,

even though it had shortcomings in accuracy performance. Although learning-based MVS achieved better results in reconstructing geometric details on the surface, it tended to introduce more noise into the results compared to NeRF-based methods and Gaussian Splatting. Generative AI methods, limited by the number of input images, may struggle to accurately capture and represent complex geometry. However, these methods typically produce results with less noticeable noise on the surface of the generated object.

Table 9. A summary of the evaluated methods for the three different non-collaborative industrial objects.

	Method	Synthetic Metallic	Industrial_A	Synthetic_Glass
NeRF	Instant-NGP	✔	✘	✘
	Mono-Neus	✔	✔	✘
	MonoSDF	✔	✘	✘
	Mono-Unisurf	✔	✘	✘
	Nerfacto(w/depth)	-	✔	✔
	Nerfacto(w/o depth)	✔	✔	✔
	Neuralangelo	✔	✔	✔
	NeuS	✔	✔	✘
	Neus-Facto	✔	✘	✘
	Unisurf	✔	✘	✘
	VolSDF	✔	✘	✘
Gaussian Splatting	FSGS	✔	✔	✔
	GaussianShader	✔	✔	✔
	Gaussian Splatting	✔	✔	✔
	Scaffold-GS	✔	✔	✔
MVS	DI-MVS	✔	✔	✔
	ET-MVSNet	✔	✔	✔
	GBi-Net	✔	✔	✘
	GeoMVSNet	✔	✔	✘
	KD-MVS	✔	✔	✔
	MVSFormer	✔	✔	✔
	MVStudio	✔	✔	✔
	TransMVSNet	✔	✔	✔
MDE	Depth Anything	✔	✔	✘
	MiDaS	✔	✔	✘
	ZoeDepth	✔	✔	✘
Generative AI	One-2-3-45	✔	✔	✔
	DreamGaussian	✔	✔	✔
	Magic123	✔	✔	✔
	Zero-1-to-3	✔	✔	✔

In general, NeRF-based methods outperform other approaches in objects with small sizes and asymmetric surfaces. This suggests their suitability for application in micro-industrial object inspection, particularly those susceptible to noise interference. Learning-based MVS can output a dense and very accurate result for medium or large objects with intricate surface structures, which will not have a significant impact on the final metrics due to partial noise. The MVS methods are well suited for applications that require dense point clouds but do not demand real-time processing and visualization. They are particularly appropriate for applications in aerospace component engineering, heritage restoration, and city-level scene reconstruction. However, learning-based MVS is sensitive to transparent and highly refractive surfaces, leading to substantial errors in depth estimation and resulting in a proliferation of noise points on glass objects. For scenes involving transparent surfaces, Gaussian Splatting proved to be more suitable due to its ability to mitigate such effects. Additionally, Gaussian Splatting's explicit representation enables manual control over the number of generated points. This flexibility makes it particularly advantageous for applications requiring real-time performance, fast transmission, and storage efficiency, such as autonomous driving, AR/VR, and rapid geometry editing.

Since generative AI methods rely on the multi-view reasoning capability of foundation models, selecting a strong foundation model (such as Zero-1-to-3) and fine-tuning it with a domain-specific dataset enables the rapid generation of multiple 3D objects from single-view 2D images. This significantly reduces the cost of obtaining 3D datasets in specific industrial scenes. Additionally, in case of applications related to defect detection with learning-based methods, manually adding defects in 2D images and generating corresponding 3D geometries can potentially simplify the training procedure, offering a more efficient and cost-effective alternative to generating training data with the defects created directly in the 3D space.

6. Conclusions and Future Research Lines

This paper investigated the feasibility of employing learning-based methods to handle non-collaborative surfaces and presented a comprehensive metrological analysis using diverse types of learning-based 3D reconstruction methods. Quantitative and visual comparison tests among NeRF, MVS, Gaussian Splatting, MDE, and generative AI were performed to understand the advantages and disadvantages when dealing with non-collaborative surfaces. The research employed complex, texture-less, metallic, reflective, and transparent objects, coming from both real and virtual scenarios. The quality of the generated 3D data was assessed using various evaluation approaches and metrics, including noise level, geometric accuracy and completeness. We verified the possibility of utilizing multi-view datasets but also just one or a few images for the 3D reconstruction of non-Lambertian surfaces, even in the absence of prior camera information, paving the way for future research in this area. This study also aimed to serve as a resource of novel and relevant studies in industrial 3D reconstruction, mainly focusing on propelling continued exploration and advancement in this rapidly evolving field.

Based on the findings in this study and identified issues of performance of the investigated methods, potential future research directions in the related fields include the following:

Real-time high-fidelity rendering: Since Gaussian Splatting was proposed in 2023, it quickly became a widely popular topic for 3D reconstruction purposes due to the model's light weight, showing the potential to replace NeRF in rendering scenarios. Its explicit representational approach enables it to bypass sampling from the entire space, unlike NeRF-based methods, thereby requiring few computational resources. However, the effectiveness of Gaussian Splatting is closely tied to the quality of its initial Gaussian

points. Particularly when dealing with reflective and refractive surfaces, the quality of these initialized Gaussians emerges as a pivotal factor in achieving high-fidelity rendering. Currently, although a limited number of NeRF/Gaussian Splatting studies [22,79,203,204] have looked into enhancing performance in reflective scenes through methods like the mathematical modeling of reflections or normal estimation, there is still a need for further research in real-time, high-quality reconstruction for industrial inspections.

Few-shot 3D reconstruction: High-quality 3D reconstruction through sparse viewpoints is also a prevailing research focus in the computer vision community. Previously published learning-based reconstruction methods usually relied on dozens to hundreds of images to reconstruct a scene, which led to massive GPU memory consumption or TPUs for training purposes, particularly when higher resolution images were used. Recent studies primarily integrated techniques such as depth prior [28,62], diffusion prior [131,205,206], or geometric regularization [31,207] to achieve novel view synthesis (followed by 3D reconstruction) from a few-shot input images. However, the challenge remains unresolved and awaits further exploration.

Removing dependence on camera priors: During the training of the models, the knowledge of the camera interior and exterior parameters, as well as the redundant time loss in format transformation and coordinate system harmonization, hamper model generalization performance [208]. Previous research has utilized photometric reconstruction [209] or has incorporated undistorted monocular depth priors [210] to estimate camera parameters in typical scenarios. However, adapting and generalizing these approaches to non-collaborative surfaces remains a challenge. The recent introduction of DUSt3R [211] represents a significant achievement that enables an MVS network to eliminate its dependence on camera poses by utilizing the cross-attention mechanism of vision Transformer (ViT) [110] to perform image pair joint pose inference. Nevertheless, the experimental results presented in this paper indicate that this method temporarily falls short in surpassing other learning-based methods [101,193,212] and traditional handcrafted methods [129,213] in terms of accuracy or completeness. Therefore, the 3D reconstruction of non-collaborative surfaces without a camera prior requires further investigations.

Non-collaborative surface-related open datasets: As widely acknowledged, having sufficient data is a fundamental component in training a high-quality, learning-based model. Currently, we already have datasets such as Clearpose (Transparent objects) [214], TRansPose (Transparent objects) [143], NeRFBK (both real and synthetic transparent and shiny objects) [17], Nex (Shiny objects) [215], Trosd (Transparent objects) [216], Tom-net (Synthetic transparent objects) [217], and Industrial Metal Objects [218] at our disposal. Nevertheless, improving model generalization performance for non-collaborative surfaces still requires more than having common objects with simple structures. Challenging and complex data, captured under various lighting conditions, is necessary to reach further improvements in deep learning-based model performance.

Author Contributions: Conceptualization, Z.Y. and N.P.; methodology, Z.Y. and N.P.; software, Z.Y. and N.P.; validation, Z.Y., N.P., P.T. and E.M.F.; formal analysis, Z.Y. and N.P.; investigation, Z.Y., N.P., P.T. and F.R.; resources, F.R.; data curation, Z.Y. and F.R.; writing, original draft preparation, Z.Y. and N.P.; writing, review and editing, Z.Y., N.P., P.T., E.M.F. and F.R.; visualization, Z.Y. and N.P.; supervision, F.R.; project administration, F.R.; funding acquisition, F.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Available upon a reasonable request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Appendix A.1. Industrial_A Object

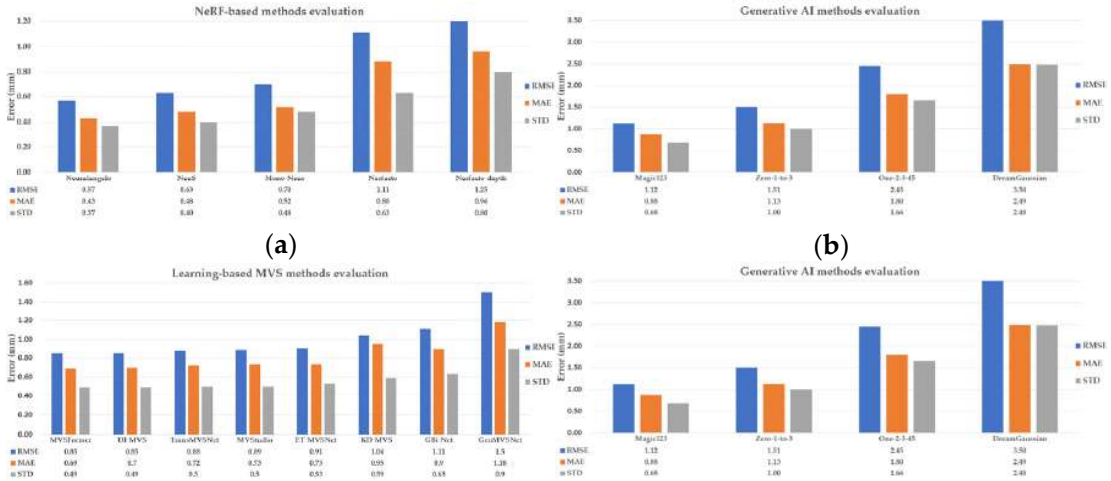


Figure A1. Metrics for the cloud-to-mesh comparisons of all tested methods applied to the Industrial_A object: (**a**) NeRF-based methods, (**b**) Gaussian Splatting methods, (**c**) learning-based MVS methods, and (**d**) generative AI methods.

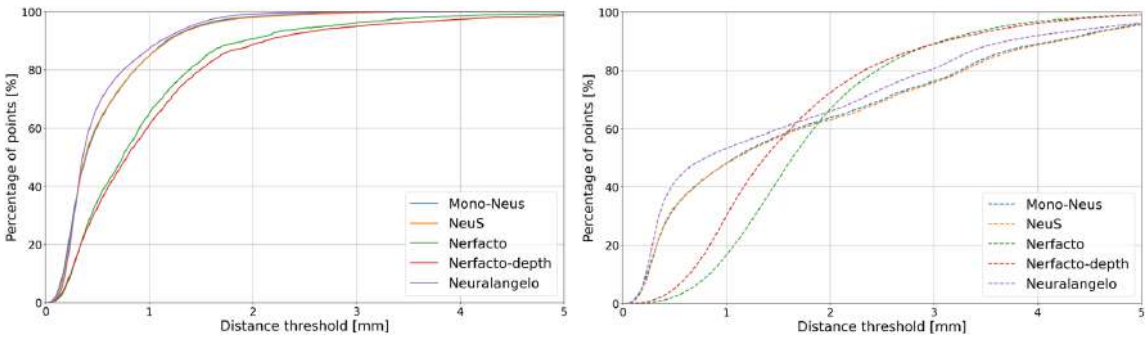


Figure A2. The estimated accuracy and completeness for NeRF-based methods using the Industrial_A object: (**a**) accuracy, (**b**) completeness.

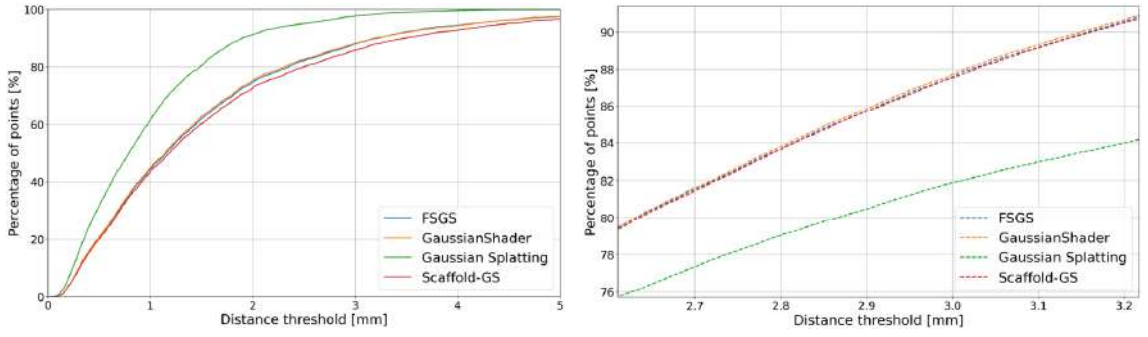
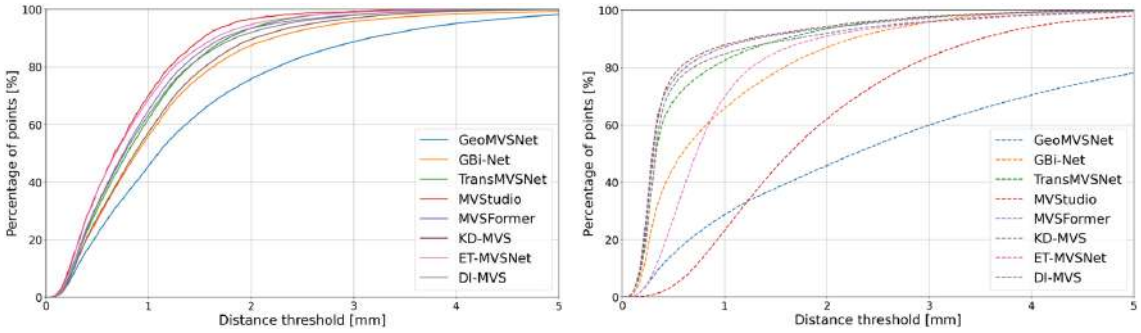
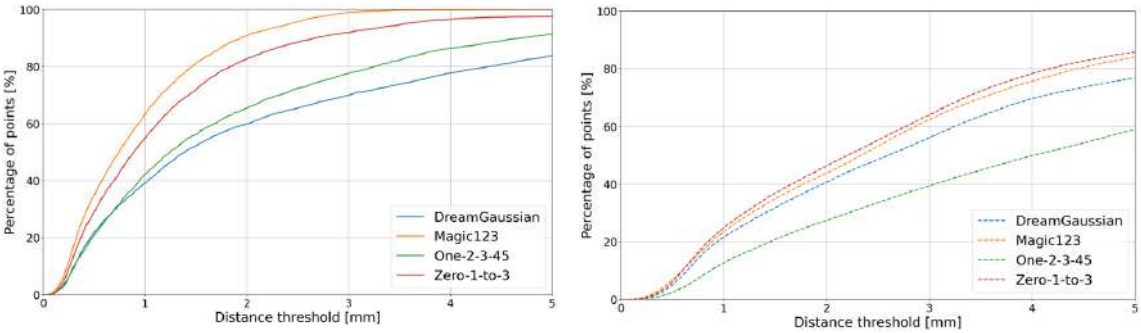


Figure A3. The estimated accuracy and completeness for Gaussian Splatting using the Industrial_A object: (**a**) accuracy, (**b**) completeness (FSGS and GaussianShader almost overlap with Scaffolds-GS).



(**a**) (**b**)

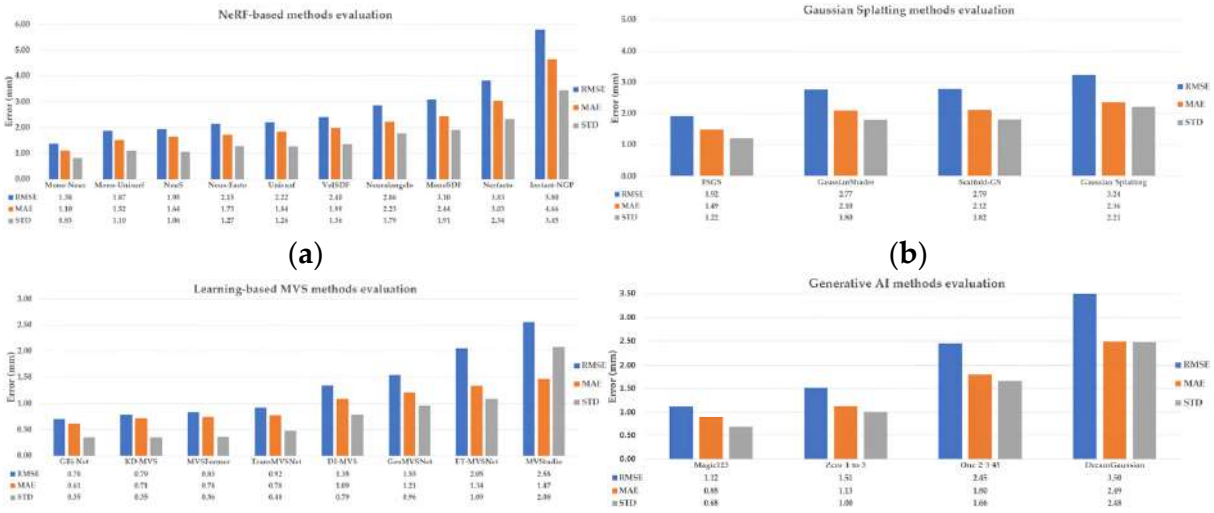
Figure A4. The estimated accuracy and completeness for MVS-based methods using the Industrial_A object: (**a**) accuracy, (**b**) completeness.



(**a**) (**b**)

Figure A5. The estimated accuracy and completeness for generative AI methods using the Industrial_A object: (**a**) accuracy, (**b**) completeness.

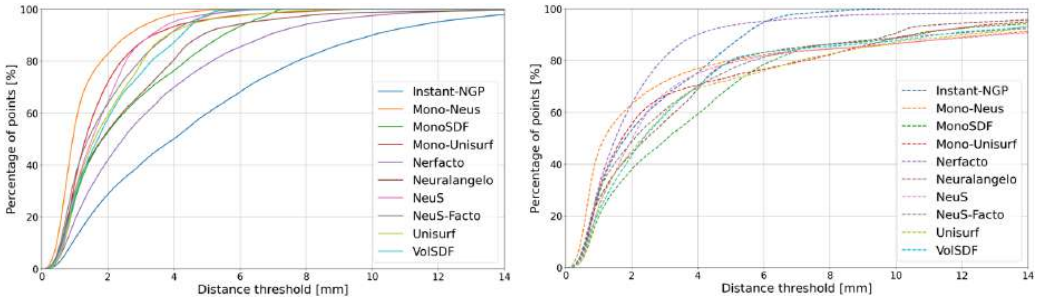
Appendix A.2. Synthetic_Metallic Objects



(**a**) (**b**)

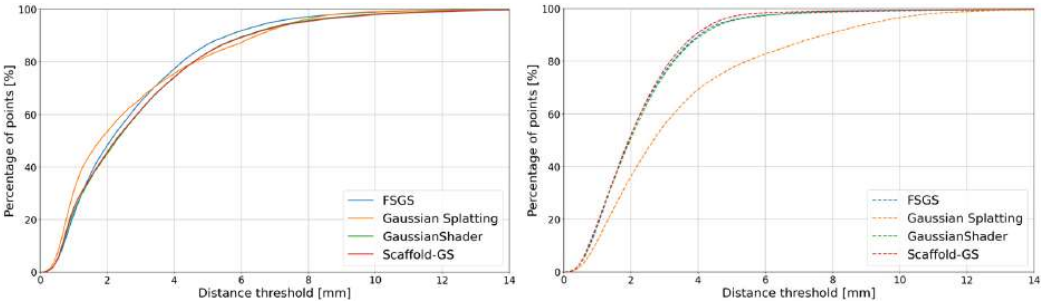
(**c**) (**d**)

Figure A6. Metrics for the cloud-to-mesh comparisons of all tested methods applied to the Synthetic_Metallic object: (**a**) NeRF-based methods, (**b**) Gaussian Splatting methods, (**c**) learning-based MVS methods, and (**d**) generative AI methods.



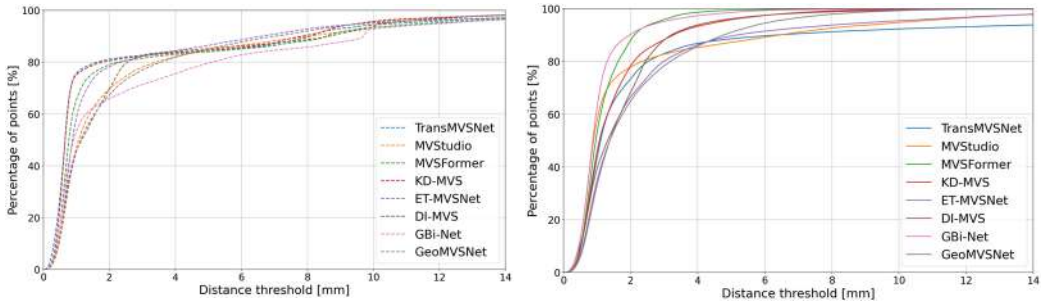
(**a**) (**b**)

Figure A7. The estimated accuracy and completeness for NeRF-based methods using the Synthetic_Metallic object: (**a**) accuracy, (**b**) completeness.



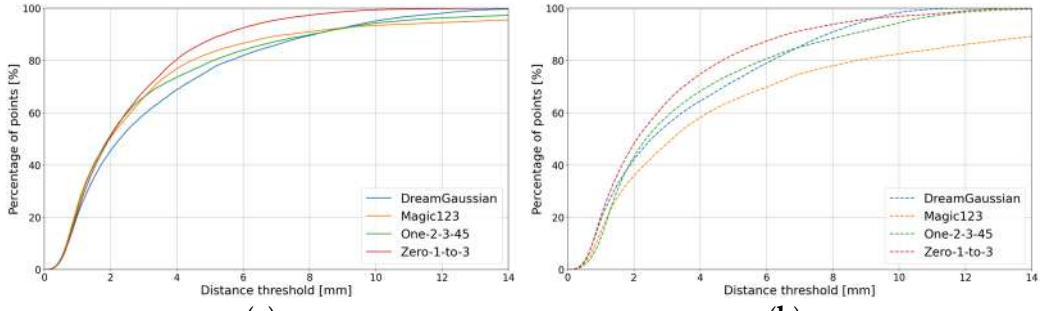
(**a**) (**b**)

Figure A8. The estimated accuracy and completeness for Gaussian Splatting using the Synthetic_Metallic object: (**a**) accuracy, (**b**) completeness.



(**a**) (**b**)

Figure A9. The estimated accuracy and completeness for MVS-based methods using the Synthetic_Metallic object: (**a**) accuracy, (**b**) completeness.

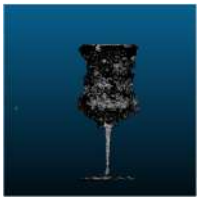
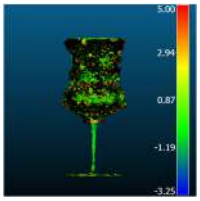
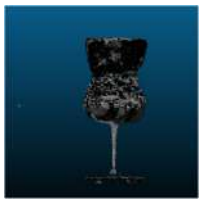
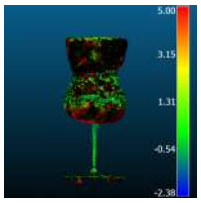
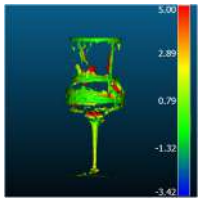


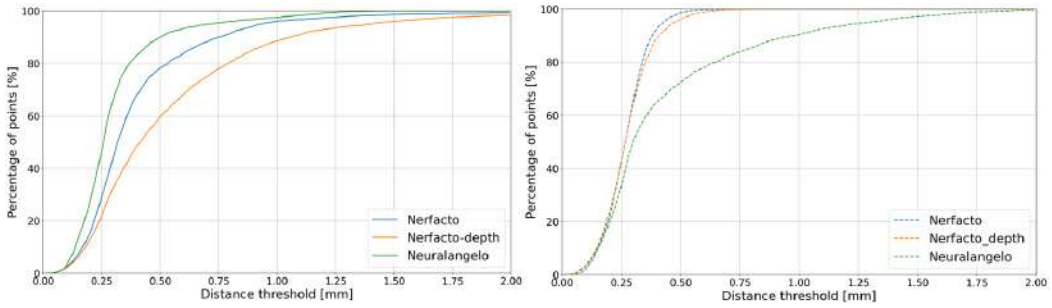
(**a**) (**b**)

Figure A10. The estimated accuracy and completeness for generative AI methods using the Synthetic_Metallic object: (**a**) accuracy, (**b**) completeness.

Appendix A.3. Synthetic_Glass Objects

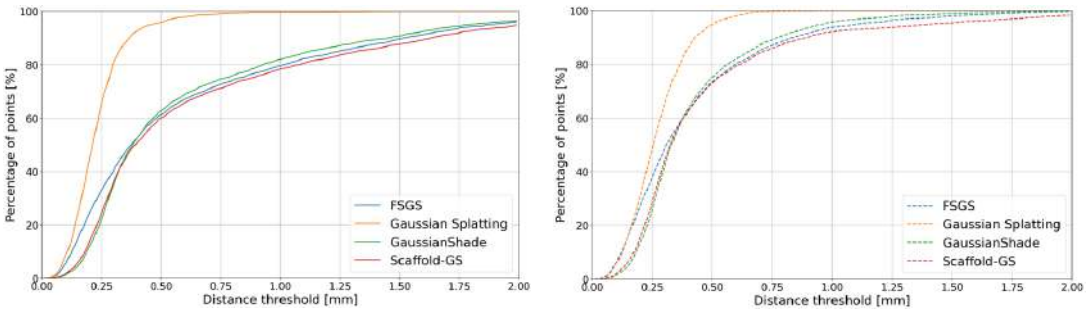
Table A1. Visuals and metrics for the cloud-to-mesh comparisons of the tested NeRF-based methods applied to the Synthetic_Glass object.

Method	3D Geometry	Comparison Result [mm]	Metric [mm]			
			RMSD	MAE	STD	Mean_E
Nerfacto			3.2	2.43	2.08	1.98
Nerfacto-depth			5.03	3.76	3.33	3.40
Neuralangelo			**2.29**	**1.72**	**1.51**	**1.19**



(a) **(b)**

Figure A11. The estimated accuracy and completeness for NeRF-based methods using the Synthetic_Glass object: (**a**) accuracy, (**b**) completeness.



(**a**) (**b**)

Figure A12. The estimated accuracy and completeness for Gaussian Splatting using the Synthetic_Glass object: (**a**) accuracy, (**b**) completeness.

Table A2. Visuals and metrics for the cloud-to-mesh comparisons of the tested Gaussian Splatting methods applied to the Synthetic_Glass object.

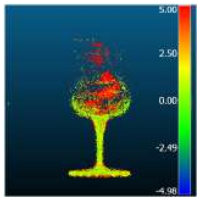
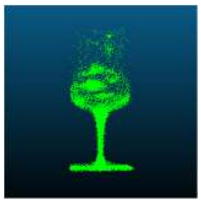
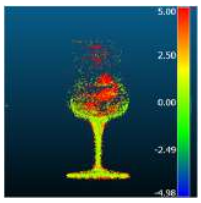
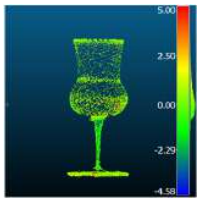
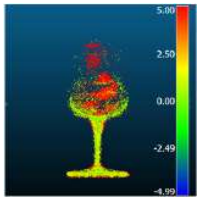
Method	3D Geometry	Comparison Result [mm]	Metric [mm]			
			RMSD	MAE	STD	Mean_E
FSGS			5.78	3.63	4.5	2.9
GaussianShader			5.39	3.4	4.18	2.63
Gaussian Splatting			**1.54**	**1.22**	**0.93**	**0.44**
Scaffold-GS			6.16	3.84	4.81	3.13

Table A3. Metrics for the cloud-to-mesh comparisons of the tested learning-based MVS methods applied to the Synthetic_Glass object.

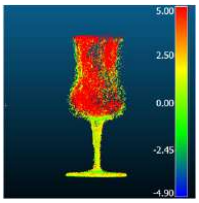
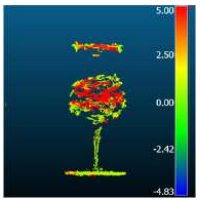
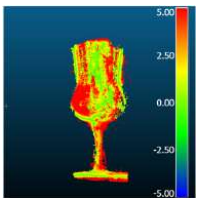
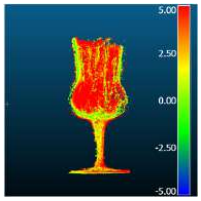
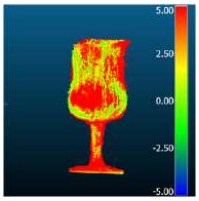
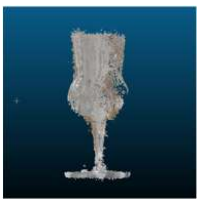
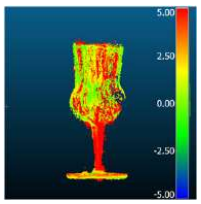
Method	3D Geometry	Comparison Result [mm]	Metric [mm]			
			RMSD	MAE	STD	Mean_E
ET-MVSNet			9.49	6.31	7.08	5.82
MVStudio			**3.14**	**1.69**	**1.43**	**0.93**
TransMVSNet			8.17	5.16	6.33	4.61

Table A3. *Cont.*

Method	3D Geometry	Comparison Result [mm]	Metric [mm]			
			RMSD	MAE	STD	Mean_E
DI-MVS			6.06	3.44	4.98	2.72
KD-MVS			3.33	2.49	2.20	1.62
MVSFormer			5.82	3.99	4.24	3.62

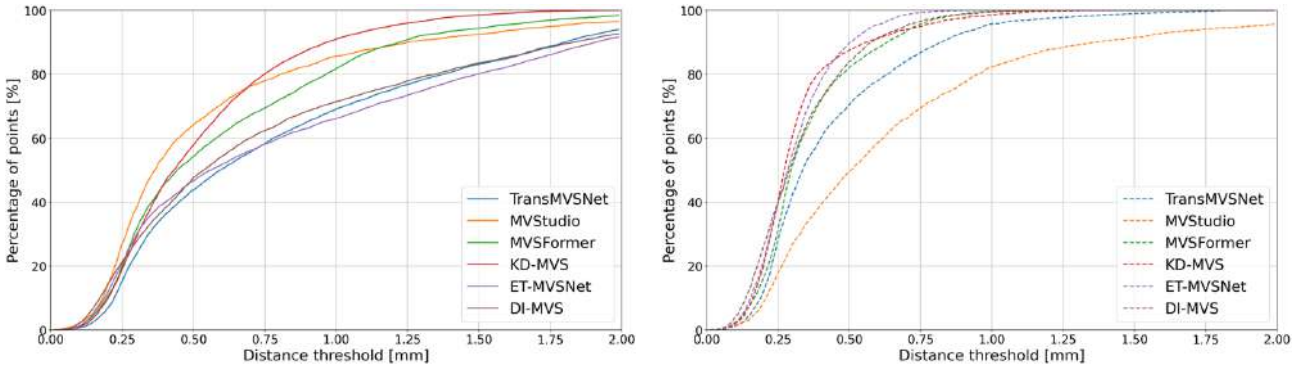


Figure A13. The estimated accuracy and completeness for MVS-based methods using the Synthetic_Glass object: (**a**) accuracy, (**b**) completeness.

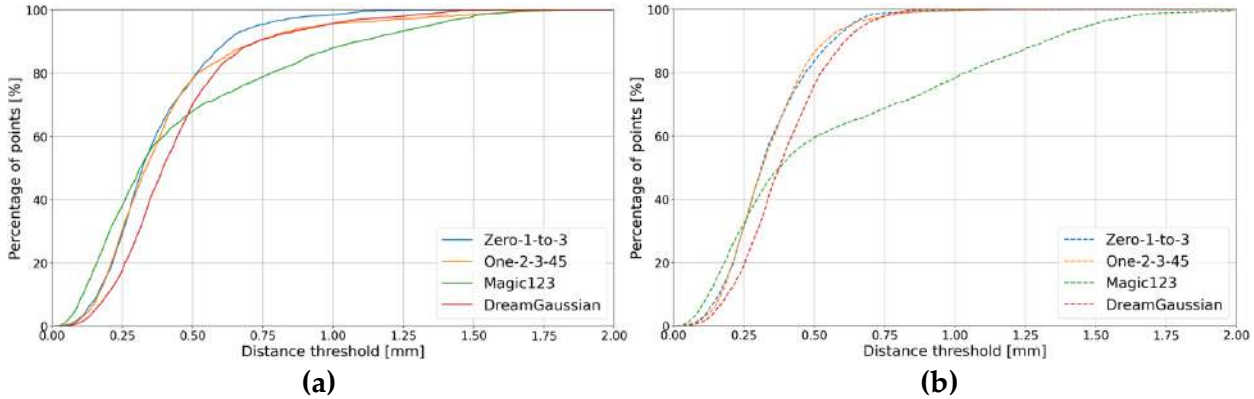
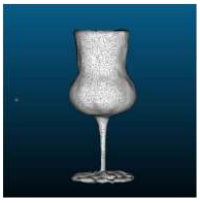
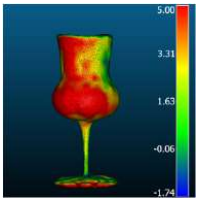
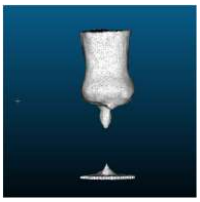
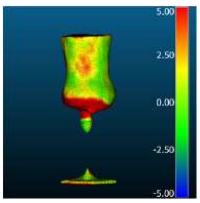
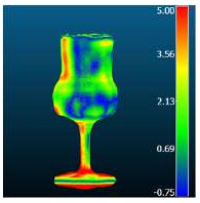
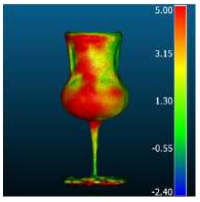


Figure A14. The estimated accuracy and completeness for generative AI methods using the Synthetic_Glass object: (**a**) accuracy (**b**) completeness.

Table A4. Metrics for the cloud-to-mesh comparisons of the tested generative AI methods applied to the Synthetic_Glass object.

Method	3D Geometry	Comparison Result [mm]	Metric [mm]			
			RMSD	MAE	STD	Mean_E
DreamGaussian			6.26	4.59	4.25	4.17
Magic1233			4.12	3.42	2.30	3.16
One-2-3-45			3.52	2.79	2.15	2.48
Zero-1-to-3			**3.32**	**2.76**	**1.85**	**2.39**

References

1. Li, Q.; Huang, H.; Yu, W.; Jiang, S. Optimized views photogrammetry: Precision analysis and a large-scale case study in qingdao. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 1144–1159.
2. Hosseininaveh, A.; Yazdan, R.; Karami, A.; Moradi, M.; Ghorbani, F. A low-cost and portable system for 3D reconstruction of texture-less objects. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2015**, *40*, 327–332.
3. Ahmadabadian, A.H.; Karami, A.; Yazdan, R. An automatic 3D reconstruction system for texture-less objects. *Robot. Auton. Syst.* **2019**, *117*, 29–39. [CrossRef]
4. Menna, F.; Nocerino, E.; Morabito, D.; Farella, E.M.; Perini, M.; Remondino, F. An open source low-cost automatic system for image-based 3D digitization. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *42*, 155–162.
5. Hu, Y.; Wang, S.; Cheng, X.; Xu, C.; Hao, Q. Dynamic deformation measurement of specular surface with deflectometry and speckle digital image correlation. *Sensors* **2020**, *20*, 1278. [CrossRef]
6. Parras-Burgos, D.; Fernández-Pacheco, D.G.; Cavas-Martínez, F.; Nieto, J.; Cañavate, F.J. Initiation to Reverse Engineering by Using Activities Based on Photogrammetry as New Teaching Method in University Technical Studies. In Proceedings of the 13th UAHCI, Orlando, FL, USA, 26 July 2019; Volume 21, pp. 159–176.
7. Huang, S.; Xu, K.; Li, M.; Wu, M. Improved visual inspection through 3D image reconstruction of defects based on the photometric stereo technique. *Sensors* **2019**, *19*, 4970. [CrossRef] [PubMed]
8. Karami, A.; Menna, F.; Remondino, F. Investigating 3D reconstruction of non-collaborative surfaces through photogrammetry and photometric stereo. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2021**, *43*, 519–526. [CrossRef]
9. Lu, Z.; Cai, L. Accurate three-dimensional measurement for small objects based on the thin-lens model. *Appl. Opt.* **2020**, *59*, 6600–6611. [CrossRef]
10. Anciukevičius, T.; Xu, Z.; Fisher, M.; Henderson, P.; Bilen, H.; Mitra, N.J.; Guerrero, P. Renderdiffusion: Image diffusion for 3D reconstruction, inpainting and generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 12608–12618.

11. Hafeez, J.; Lee, J.; Kwon, S.; Ha, S.; Hur, G.; Lee, S. Evaluating feature extraction methods with synthetic noise patterns for image-based modelling of texture-less objects. *Remote Sens.* **2020**, *12*, 3886. [CrossRef]
12. Morelli, L.; Karami, A.; Menna, F.; Remondino, F. Orientation of images with low contrast textures and transparent objects. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci* **2022**, *48*, 77–84.
13. Gao, K.; Gao, Y.; He, H.; Lu, D.; Xu, L.; Li, J. Nerf: Neural radiance field in 3D vision, a comprehensive review. *arXiv* **2022**, arXiv:2210.00379.
14. Remondino, F.; Karami, A.; Yan, Z.; Mazzacca, G.; Rigon, S.; Qin, R. A critical analysis of nerf-based 3D reconstruction. *Remote Sens.* **2023**, *15*, 3585.
15. Stathopoulou, E.K.; Remondino, F. A survey on conventional and learning-based methods for multi-view stereo. *Photogramm. Rec.* **2023**, *38*, 374–407.
16. Yin, W.; Zhang, C.; Chen, H.; Cai, Z.; Yu, G.; Wang, K.; Chen, X.; Shen, C. Metric3D: Towards zero-shot metric 3D prediction from a single image. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Vancouver, BC, Canada, 1–6 October 2023; pp. 9043–9053.
17. Yan, Z.; Mazzacca, G.; Rigon, S.; Farella, E.M.; Trybala, P.; Remondino, F. NeRFBK: A holistic dataset for benchmarking NeRF-based 3D reconstruction. *ISPRS Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2023**, *48*, 219–226.
18. Cline, H.E.; Dumoulin, C.L.; Hart, H.R., Jr.; Lorensen, W.E.; Ludke, S. 3D reconstruction of the brain from magnetic resonance images using a connectivity algorithm. *Magn. Reson. Imaging* **1987**, *5*, 345–352.
19. Mildnhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* **2021**, *65*, 99–106.
20. Guo, Y.C.; Kang, D.; Bao, L.; He, Y.; Zhang, S.H. Nerfren: Neural radiance fields with reflections. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 17–24 June 2022; pp. 18409–18418.
21. Ichnowski, J.; Avigal, Y.; Kerr, J.; Goldberg, K. Dex-NeRF: Using a neural radiance field to grasp transparent objects. In Proceedings of the 5th Conference on Robot Learning, Baltimore, MD, USA, 18–24 June 2022; Volume 164, pp. 526–536.
22. Verbin, D.; Hedman, P.; Mildenhall, B.; Zickler, T.; Barron, J.T.; Srinivasan, P.P. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5481–5490.
23. Choi, C.; Kim, J.; Kim, Y.M. IBL-NeRF: Image-Based Lighting Formulation of Neural Radiance Fields. *arXiv* **2022**, arXiv:2210.08202.
24. Yu, Z.; Peng, S.; Niemeyer, M.; Sattler, T.; Geiger, A. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Proc. NeurIPS* **2022**, *35*, 25018–25032.
25. Li, Z.; Müller, T.; Evans, A.; Taylor, R.H.; Unberath, M.; Liu, M.Y.; Lin, C.H. Neuralangelo: High-Fidelity Neural Surface Reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 8456–8465.
26. Müller, T.; Evans, A.; Schied, C.; Keller, A. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.* **2022**, *41*, 1–15.
27. Deng, K.; Liu, A.; Zhu, J.Y.; Ramanan, D. Depth-supervised nerf: Fewer views and faster training for free. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12882–12891.
28. Wang, G.; Chen, Z.; Loy, C.C.; Liu, Z. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. *arXiv* **2023**, arXiv:2303.16196.
29. Jain, A.; Tancik, M.; Abbeel, P. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 5885–5894.
30. Niemeyer, M.; Barron, J.T.; Mildenhall, B.; Sajjadi, M.S.; Geiger, A.; Radwan, N. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5480–5490.
31. Kwak, M.S.; Song, J.; Kim, S. Geconerf: Few-shot neural radiance fields via geometric consistency. *arXiv* **2023**, arXiv:2301.10941.
32. Roessle, B.; Barron, J.T.; Mildenhall, B.; Srinivasan, P.P.; Nießner, M. Dense depth priors for neural radiance fields from sparse input views. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12892–12901.
33. Seo, S.; Han, D.; Chang, Y.; Kwak, N. MixNeRF: Modeling a Ray with Mixture Density for Novel View Synthesis from Sparse Inputs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 20659–20668.
34. Yang, Z.; Yang, H.; Pan, Z.; Zhu, X.; Zhang, L. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. *arXiv* **2023**, arXiv:2310.10642.
35. Somraj, N.; Soundararajan, R. ViP-NeRF: Visibility Prior for Sparse Input Neural Radiance Fields. *arXiv* **2023**, arXiv:2305.00041.

36. Somraj, N.; Karanayil, A.; Soundararajan, R. SimpleNeRF: Regularizing Sparse Input Neural Radiance Fields with Simpler Solutions. In Proceedings of the SIGGRAPH Asia 2023 Conference Papers, Sydney, NSW, Australia, 12–15 December 2023; pp. 1–11.
37. Yu, A.; Ye, V.; Tancik, M.; Kanazawa, A. pixelnerf: Neural radiance fields from one or few images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 4578–4587.
38. Wynn, J.; Turmukhambetov, D. Diffusionerf: Regularizing neural radiance fields with denoising diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 20–25 June 2023; pp. 4180–4189.
39. Wu, R.; Mildenhall, B.; Henzler, P.; Park, K.; Gao, R.; Watson, D.; Holynski, A. ReconFusion: 3D Reconstruction with Diffusion Priors. *arXiv* **2023**, arXiv:2312.02981.
40. Cheng, K.; Long, X.; Yin, W.; Wang, J.; Wu, Z.; Ma, Y.; Wang, K.; Chen, X.; Chen, X. UC-NERF: Neural Radiance Field for under-calibrated multi-view cameras. *arXiv* **2023**, arXiv:2311.16945.
41. Westoby, M.J.; Brasington, J.; Glasser, N.F.; Hambrey, M.J.; Reynolds, J.M. Structure-from-Motion'photogrammetry: A low-cost, effective tool for geoscience applications. *Geomorphology* **2012**, *179*, 300–314.
42. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sutskever, I. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning (PMLR), Virtually, 18–24 April 2021; pp. 8748–8763.
43. Sofiiuk, K.; Petrov, I.; Barinova, O.; Konushin, A. f-brs: Rethinking backpropagating refinement for interactive segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8623–8632.
44. Zwicker, M.; Pfister, H.; Van Baar, J.; Gross, M. EWA volume splatting. In Proceedings of the Visualization, San Diego, CA, USA, 19–24 October 2001; VIS'01. pp. 29–538.
45. Kerbl, B.; Kopanas, G.; Leimkühler, T.; Drettakis, G. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.* **2023**, *42*, 139.
46. Chen, G.; Wang, W. A Survey on 3D Gaussian Splatting. *arXiv* **2024**, arXiv:2401.03890.
47. Guédon, A.; Lepetit, V. SuGaR: Surface-Aligned Gaussian Splatting for Efficient 3D Mesh Reconstruction and High-Quality Mesh Rendering. *arXiv* **2023**, arXiv:2311.12775.
48. Fei, B.; Xu, J.; Zhang, R.; Zhou, Q.; Yang, W.; He, Y. 3D Gaussian as a New Vision Era: A Survey. *arXiv* **2024**, arXiv:2402.07181.
49. Bottou, L. Large-scale machine learning with stochastic gradient descent. In Proceedings of the COMPSTAT'2010: 19th International Conference on Computational Statistics, Paris, France, 22–27 August 2010; pp. 177–186.
50. Lee, B.; Lee, H.; Sun, X.; Ali, U.; Park, E. Deblurring 3D Gaussian Splatting. *arXiv* **2024**, arXiv:2401.00834.
51. Huang, L.; Bai, J.; Guo, J.; Guo, Y. GS++: Error Analyzing and Optimal Gaussian Splatting. *arXiv* **2024**, arXiv:2402.00752.
52. Feng, Y.; Feng, X.; Shang, Y.; Jiang, Y.; Yu, C.; Zong, Z.; Shao, T.; Wu, H.; Zhou, K.; Jiang, C.; et al. Gaussian Splashing: Dynamic Fluid Synthesis with Gaussian Splatting. *arXiv* **2024**, arXiv:2401.15318.
53. Fu, Y.; Liu, S.; Kulkarni, A.; Kautz, J.; Efros, A.A.; Wang, X. COLMAP-Free 3D Gaussian Splatting. *arXiv* **2023**, arXiv:2312.07504.
54. Qin, M.; Li, W.; Zhou, J.; Wang, H.; Pfister, H. LangSplat: 3D Language Gaussian Splatting. *arXiv* **2023**, arXiv:2312.16084.
55. Li, M.; Liu, S.; Zhou, H. SGS-SLAM: Semantic Gaussian Splatting For Neural Dense SLAM. *arXiv* **2024**, arXiv:2402.03246.
56. Zuo, X.; Samangouei, P.; Zhou, Y.; Di, Y.; Li, M. FMGS: Foundation Model Embedded 3D Gaussian Splatting for Holistic 3D Scene Understanding. *arXiv* **2024**, arXiv:2401.01970.
57. Gao, L.; Yang, J.; Zhang, B.T.; Sun, J.M.; Yuan, Y.J.; Fu, H.; Lai, Y.K. Mesh-based Gaussian Splatting for Real-time Large-scale Deformation. *arXiv* **2024**, arXiv:2402.04796.
58. Cheng, K.; Long, X.; Yang, K.; Yao, Y.; Yin, W.; Ma, Y.; Wang, W.; Chen, X. GaussianPro: 3D Gaussian Splatting with Progressive Propagation. *arXiv* **2024**, arXiv:2402.14650.
59. Yan, Z.; Low, W.F.; Chen, Y.; Lee, G.H. Multi-Scale 3D Gaussian Splatting for Anti-Aliased Rendering. *arXiv* **2023**, arXiv:2311.17089.
60. Chung, J.; Oh, J.; Lee, K.M. Depth-regularized optimization for 3D gaussian splatting in few-shot images. *arXiv* **2023**, arXiv:2311.13398.
61. Xiong, H.; Muttukuru, S.; Upadhyay, R.; Chari, P.; Kadambi, A. SparseGS: Real-Time 360 {\deg} Sparse View Synthesis using Gaussian Splatting. *arXiv* **2023**, arXiv:2312.00206.
62. Zhu, Z.; Fan, Z.; Jiang, Y.; Wang, Z. FSGS: Real-Time Few-shot View Synthesis using Gaussian Splatting. *arXiv* **2023**, arXiv:2312.00451.
63. Tang, J.; Ren, J.; Zhou, H.; Liu, Z.; Zeng, G. Dreamgaussian: Generative gaussian splatting for efficient 3D content creation. *arXiv* **2023**, arXiv:2309.16653.
64. Yan, Z.; Dong, W.; Shao, Y.; Lu, Y.; Haiyang, L.; Liu, J.; Wang, H.; Wang, Z.; Wang, Y.; Remondino, F.; et al. Renderworld: World model with self-supervised 3D label. *arXiv* **2024**, arXiv:2409.11356.

65. Yan, Z.; Li, L.; Shao, Y.; Chen, S.; Kai, W.; Hwang, J.N.; Zhao, H.; Remondino, F. 3DSceneEditor: Controllable 3D scene editing with gaussian splatting. *arXiv* **2024**, arXiv:2412.01583.
66. Yi, T.; Fang, J.; Wu, G.; Xie, L.; Zhang, X.; Liu, W.; Wang, X. Gaussiandreamer: Fast generation from text to 3D gaussian splatting with point cloud priors. *arXiv* **2023**, arXiv:2310.08529.
67. Yang, Z.; Gao, X.; Zhou, W.; Jiao, S.; Zhang, Y.; Jin, X. Deformable 3D gaussians for high-fidelity monocular dynamic scene reconstruction. *arXiv* **2023**, arXiv:2309.13101.
68. Huang, Y.; Cui, B.; Bai, L.; Guo, Z.; Xu, M.; Ren, H. Endo-4dgs: Distilling depth ranking for endoscopic monocular scene reconstruction with 4d gaussian splatting. *arXiv* **2024**, arXiv:2401.16416.
69. Wu, G.; Yi, T.; Fang, J.; Xie, L.; Zhang, X.; Wei, W.; Liu, W.; Tian, Q.; Wang, X. 4d gaussian splatting for real-time dynamic scene rendering. *arXiv* **2023**, arXiv:2310.08528.
70. Moreau, A.; Song, J.; Dhamo, H.; Shaw, R.; Zhou, Y.; Pérez-Pellitero, E. Human Gaussian Splatting: Real-time Rendering of Animatable Avatars. *arXiv* **2023**, arXiv:2311.17113.
71. Zielonka, W.; Bagautdinov, T.; Saito, S.; Zollhöfer, M.; Thies, J.; Romero, J. Drivable 3D gaussian avatars. *arXiv* **2023**, arXiv:2311.08581.
72. Qian, Z.; Wang, S.; Mihajlovic, M.; Geiger, A.; Tang, S. 3DGS-Avatar: Animatable Avatars via Deformable 3D Gaussian Splatting. *arXiv* **2023**, arXiv:2312.09228.
73. Dhamo, H.; Nie, Y.; Moreau, A.; Song, J.; Shaw, R.; Zhou, Y.; Pérez-Pellitero, E. HeadGaS: Real-Time Animatable Head Avatars via 3D Gaussian Splatting. *arXiv* **2023**, arXiv:2312.02902.
74. Chen, Y.; Wang, L.; Li, Q.; Xiao, H.; Zhang, S.; Yao, H.; Liu, Y. Monogaussianavatar: Monocular gaussian point-based head avatar. *arXiv* **2023**, arXiv:2312.04558.
75. Liu, Y.; Li, C.; Yang, C.; Yuan, Y. EndoGaussian: Gaussian Splatting for Deformable Surgical Scene Reconstruction. *arXiv* **2024**, arXiv:2401.12561.
76. Yan, C.; Qu, D.; Wang, D.; Xu, D.; Wang, Z.; Zhao, B.; Li, X. GS-SLAM: Dense Visual SLAM with 3D Gaussian Splatting. *arXiv* **2023**, arXiv:2311.11700.
77. Yugay, V.; Li, Y.; Gevers, T.; Oswald, M.R. Gaussian-SLAM: Photo-realistic Dense SLAM with Gaussian Splatting. *arXiv* **2023**, arXiv:2312.10070.
78. Lu, T.; Yu, M.; Xu, L.; Xiangli, Y.; Wang, L.; Lin, D.; Dai, B. Scaffold-GS: Structured 3D Gaussians for View-Adaptive Rendering. *arXiv* **2023**, arXiv:2312.00109.
79. Jiang, Y.; Tu, J.; Liu, Y.; Gao, X.; Long, X.; Wang, W.; Ma, Y. GaussianShader: 3D Gaussian Splatting with Shading Functions for Reflective Surfaces. *arXiv* **2023**, arXiv:2311.17977.
80. Tancik, M.; Weber, E.; Ng, E.; Li, R.; Yi, B.; Wang, T.; Kristoffersen, A.; Austin, J.; Salahi, K.; Ahuja, A.; et al. Nerfstudio: A modular framework for neural radiance field development. In Proceedings of the ACM SIGGRAPH 2023 Conference, Los Angeles, CA, USA, 6–10 August 2023; pp. 1–12.
81. Hisham, M.B.; Yaakob, S.N.; Raof, R.A.A.; Nazren, A.A.; Wafi, N.M. Template matching using sum of squared difference and normalized cross correlation. In Proceedings of the 2015 IEEE Student Conference on Research and Development, Kuala, Malaysia, 13–14 December 2015; pp. 100–104.
82. Wang, Y.; Luo, K.; Chen, Z.; Ju, L.; Guan, T. DeepFusion: A simple way to improve traditional multi-view stereo methods using deep learning. *Knowl. Based Syst.* **2021**, *221*, 106968.
83. Rong, F.; Xie, D.; Zhu, W.; Shang, H.; Song, L. A survey of multi view stereo. In Proceedings of the 2021 International Conference on Networking Systems of AI, Shanghai, China, 19–20 November 2021; pp. 129–135.
84. Ji, M.; Gall, J.; Zheng, H.; Liu, Y.; Fang, L. Surfacenet: An end-to-end 3D neural network for multiview stereopsis. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2307–2315.
85. Yao, Y.; Luo, Z.; Li, S.; Fang, T.; Quan, L. Mvsnet: Depth inference for unstructured multi-view stereo. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 767–783.
86. Wei, Z.; Zhu, Q.; Min, C.; Chen, Y.; Wang, G. Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 6187–6196.
87. Chen, R.; Han, S.; Xu, J.; Su, H. Point-based multi-view stereo network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1538–1547.
88. Luo, K.; Guan, T.; Ju, L.; Huang, H.; Luo, Y. P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 10452–10461.
89. Xue, Y.; Chen, J.; Wan, W.; Huang, Y.; Yu, C.; Li, T.; Bao, J. Mvscrf: Learning multi-view stereo with conditional random fields. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4312–4321.

90. Yang, J.; Mao, W.; Alvarez, J.M.; Liu, M. Cost volume pyramid based depth inference for multi-view stereo. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4877–4886.
91. Yu, Z.; Gao, S. Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1949–1958.
92. Cheng, S.; Xu, Z.; Zhu, S.; Li, Z.; Li, L.E.; Ramamoorthi, R.; Su, H. Deep stereo using adaptive thin volume representation with uncertainty awareness. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2524–2534.
93. Yi, H.; Wei, Z.; Ding, M.; Zhang, R.; Chen, Y.; Wang, G.; Tai, Y.W. Pyramid multi-view stereo net with self-adaptive view aggregation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 October 2020; pp. 766–782.
94. Yan, J.; Wei, Z.; Yi, H.; Ding, M.; Zhang, R.; Chen, Y.; Tai, Y.W. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 October 2020; pp. 674–689.
95. Zhang, J.; Li, S.; Luo, Z.; Fang, T.; Yao, Y. Vis-mvsnet: Visibility-aware multi-view stereo network. *Int. J. Comput. Vis.* **2023**, *131*, 199–214.
96. Ma, X.; Gong, Y.; Wang, Q.; Huang, J.; Chen, L.; Yu, F. Epp-mvsnet: Epipolar-assembling based depth prediction for multi-view stereo. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 5732–5740.
97. Kar, A.; Häne, C.; Malik, J. Learning a multi-view stereo machine. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 365–376.
98. Chen, R.; Han, S.; Xu, J.; Su, H. Visibility-aware point-based multi-view stereo network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3695–3708.
99. Wang, Y.; Guan, T.; Chen, Z.; Luo, Y.; Luo, K.; Ju, L. Mesh-guided multi-view stereo with pyramid architecture. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2039–2048.
100. Wang, F.; Galliani, S.; Vogel, C.; Speciale, P.; Pollefeys, M. Patchmatchnet: Learned multi-view patchmatch stereo. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 October 2021; pp. 14194–14203.
101. Xu, Q.; Tao, W. Multi-scale geometric consistency guided multi-view stereo. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 October 2019; pp. 5483–5492.
102. Chen, P.H.; Yang, H.C.; Chen, K.W.; Chen, Y.S. MVSNet++: Learning depth-based attention pyramid features for multi-view stereo. *IEEE Trans. Image Process.* **2020**, *29*, 7261–7273.
103. Rich, A.; Stier, N.; Sen, P.; Höllerer, T. 3Dvnet: Multi-view depth prediction and volumetric refinement. In Proceedings of the 2021 International Conference on 3D Vision (3DV), Virtually, 1–3 December 2021; pp. 700–709.
104. Mi, Z.; Di, C.; Xu, D. Generalized binary search network for highly-efficient multi-view stereo. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12991–13000.
105. Wang, X.; Zhu, Z.; Huang, G.; Qin, F.; Ye, Y.; He, Y.; Chi, X.; Wang, X. MVSTER: Epipolar transformer for efficient multi-view stereo. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 573–591.
106. Zhang, Z.; Peng, R.; Hu, Y.; Wang, R. GeoMVSNet: Learning Multi-View Stereo With Geometry Perception. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 21508–21518.
107. Eigen, D.; Puhrsch, C.; Fergus, R. Depth map prediction from a single image using a multi-scale deep network. *Adv. Neural Inf. Process. Syst.* **2014**, *27*.
108. Laina, I.; Rupprecht, C.; Belagiannis, V.; Tombari, F.; Navab, N. Deeper depth prediction with fully convolutional residual networks. In Proceedings of the 2016 Fourth international conference on 3D vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 239–248.
109. Yang, G.; Tang, H.; Ding, M.; Sebe, N.; Ricci, E. Transformer-based attention networks for continuous pixel-wise prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 16269–16279.
110. Ranftl, R.; Bochkovskiy, A.; Koltun, V. Vision transformers for dense prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 12179–12188.
111. Yuan, W.; Gu, X.; Dai, Z.; Zhu, S.; Tan, P. Neural window fully-connected crfs for monocular depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 3916–3925.
112. Shao, S.; Pei, Z.; Wu, X.; Liu, Z.; Chen, W.; Li, Z. IEBins: Iterative elastic bins for monocular depth estimation. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 53025–53037.
113. Ming, Y.; Meng, X.; Fan, C.; Yu, H. Deep learning for monocular depth estimation: A review. *Neurocomputing* **2021**, *438*, 14–33.

114. Fu, H.; Gong, M.; Wang, C.; Batmanghelich, K.; Tao, D. Deep ordinal regression network for monocular depth estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2002–2011.
115. Facil, J.M.; Ummenhofer, B.; Zhou, H.; Montesano, L.; Brox, T.; Civera, J. CAM-Convs: Camera-aware multi-scale convolutions for single-view depth. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 13–19 June 2019; pp. 11826–11835.
116. Wofk, D.; Ma, F.; Yang, T.J.; Karaman, S.; Sze, V. Fastdepth: Fast monocular depth estimation on embedded systems. In Proceedings of the 2019 International Conference on Robotics and Automation, Montreal, QC, Canada, 15–20 May 2019; pp. 6101–6108.
117. Zhao, S.; Fu, H.; Gong, M.; Tao, D. Geometry-aware symmetric domain adaptation for monocular depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9788–9798.
118. Wu, Z.; Wu, X.; Zhang, X.; Wang, S.; Ju, L. Spatial correspondence with generative adversarial network: Learning depth from monocular videos. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7494–7504.
119. Wimbauer, F.; Yang, N.; Von Stumberg, L.; Zeller, N.; Cremers, D. MonoRec: Semi-supervised dense reconstruction in dynamic environments from a single moving camera. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6112–6122.
120. Zhang, N.; Nex, F.; Vosselman, G.; Kerle, N. Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 18537–18546.
121. Bhoi, A. Monocular depth estimation: A survey. *arXiv* **2019**, arXiv:1901.09402.
122. Garg, R.; Bg, V.K.; Carneiro, G.; Reid, I. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Part VIII 14. pp. 740–756.
123. Godard, C.; Mac Aodha, O.; Brostow, G.J. Unsupervised monocular depth estimation with left-right consistency. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 June 2017; pp. 270–279.
124. Zhou, T.; Brown, M.; Snavely, N.; Lowe, D.G. Unsupervised learning of depth and ego-motion from video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 June 2017; pp. 1851–1858.
125. Flynn, J.; Neulander, I.; Philbin, J.; Snavely, N. Deepstereo: Learning to predict new views from the world's imagery. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5515–5524.
126. Bhat, S.F.; Birkl, R.; Wofk, D.; Wonka, P.; Müller, M. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv* **2023**, arXiv:2302.1228.
127. Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor segmentation and support inference from RGBd images. In Proceedings of the Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; Volume 12, pp. 746–760.
128. Menze, M.; Geiger, A. Object scene flow for autonomous vehicles. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3061–3070.
129. Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; Koltun, V. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 1623–1637.
130. Birkl, R.; Wofk, D.; Müller, M. MiDaS v3. 1--A Model Zoo for Robust Monocular Relative Depth Estimation. *arXiv* **2023**, arXiv:2307.14460.
131. Yang, L.; Kang, B.; Huang, Z.; Xu, X.; Feng, J.; Zhao, H. Depth anything: Unleashing the power of large-scale unlabeled data. *arXiv* **2024**, arXiv:2401.10891.
132. Zhang, L.; Rao, A.; Agrawala, M. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 3836–3847.
133. Gu, S.; Chen, D.; Bao, J.; Wen, F.; Zhang, B.; Chen, D.; Lu, Y.; Guo, B. Vector quantized diffusion model for text-to-image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10696–10706.
134. Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 22500–22510.
135. Xu, Z.; Xing, S.; Sangineto, E.; Sebe, N. SpectralCLIP: Preventing Artifacts in Text-Guided Style Transfer from a Spectral Perspective. In Proceedings of the Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 February 2024; pp. 5121–5130.

136. Liao, M.; Dong, H.B.; Wang, X.; Yan, Z.; Shao, Y. GM-MoE: Low-Light Enhancement with Gated-Mechanism Mixture-of-Experts. *arXiv* **2025**, arXiv:2503.07417.
137. Esser, P.; Chiu, J.; Atighehchian, P.; Granskog, J.; Germanidis, A. Structure and content-guided video synthesis with diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 7346–7356.
138. Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S.W.; Fidler, S.; Kreis, K. Align your latents: High-resolution video synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 22563–22575.
139. Khachatryan, L.; Movsisyan, A.; Tadevosyan, V.; Henschel, R.; Wang, Z.; Navasardyan, S.; Shi, H. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv* **2023**, arXiv:2303.13439.
140. Ge, S.; Nah, S.; Liu, G.; Poon, T.; Tao, A.; Catanzaro, B.; Jacobs, D.; Huang, J.-B.; Liu, M.-Y.; Balaji, Y. Preserve your own correlation: A noise prior for video diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 4–6 October 2023; pp. 22930–22941.
141. Luo, Z.; Chen, D.; Zhang, Y.; Huang, Y.; Wang, L.; Shen, Y.; Tan, T. VideoFusion: Decomposed Diffusion Models for High-Quality Video Generation. *arXiv* **2023**, arXiv:2303.08320.
142. Shao, Y.; Lin, D.; Zeng, F.; Yan, M.; Zhang, M.; Chen, S.; Fan, Y.; Yan, Z.; Wang, H.; Guo, J.; et al. TR-DQ: Time-Rotation Diffusion Quantization. *arXiv* **2025**, arXiv:2503.06564.
143. Kim, G.; Chun, S.Y. Datid-3D: Diversity-preserved domain adaptation using text-to-image diffusion for 3D generative model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 14203–14213.
144. Ding, L.; Dong, S.; Huang, Z.; Wang, Z.; Zhang, Y.; Gong, K.; Xu, D.; Xue, T. Text-to-3D Generation with Bidirectional Diffusion using both 2D and 3D priors. *arXiv* **2023**, arXiv:2312.04963.
145. Poole, B.; Jain, A.; Barron, J.T.; Mildenhall, B. Dreamfusion: Text-to-3D using 2d diffusion. *arXiv* **2022**, arXiv:2209.14988.
146. Fang, C.; Hu, X.; Luo, K.; Tan, P. Ctrl-Room: Controllable Text-to-3D Room Meshes Generation with Layout Constraints. *arXiv* **2023**, arXiv:2310.03602.
147. Liu, M.; Shi, R.; Chen, L.; Zhang, Z.; Xu, C.; Wei, X.; Chen, H.; Zeng, C.; Gu, J.; Su, H. One-2-3-45++: Fast single image to 3D objects with consistent multi-view generation and 3D diffusion. *arXiv* **2023**, arXiv:2311.07885.
148. Long, X.; Guo, Y.C.; Lin, C.; Liu, Y.; Dou, Z.; Liu, L.; Ma, Y.; Zhang, S.-H.; Habermann, M.; Theobalt, C.; et al. Wonder3D: Single image to 3D using cross-domain diffusion. *arXiv* **2023**, arXiv:2310.15008.
149. Qian, G.; Mai, J.; Hamdi, A.; Ren, J.; Siarohin, A.; Li, B.; Ghanem, B. Magic123: One image to high-quality 3D object generation using both 2d and 3D diffusion priors. *arXiv* **2023**, arXiv:2306.17843.
150. He, L.; Yan, H.; Luo, M.; Luo, K.; Wang, W.; Du, W.; Chen, H.; Yang, H.; Zhang, Y. Iterative reconstruction based on latent diffusion model for sparse data reconstruction. *arXiv* **2023**, arXiv:2307.12070.
151. Yu, C.; Zhou, Q.; Li, J.; Zhang, Z.; Wang, Z.; Wang, F. Points-to-3D: Bridging the gap between sparse points and shape-controllable text-to-3D generation. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November; pp. 6841–6850.
152. Liu, M.; Shi, R.; Kuang, K.; Zhu, Y.; Li, X.; Han, S.; Su, H. Openshape: Scaling up 3D shape representation towards open-world understanding. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 44860–44879.
153. Chang, A.X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; et al. Shapenet: An information-rich 3D model repository. *arXiv* **2015**, arXiv:1512.03012.
154. Liu, M.; Sheng, L.; Yang, S.; Shao, J.; Hu, S.M. Morphing and sampling network for dense point cloud completion. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–8 February 2020; Volume 34, pp. 11596–11603.
155. Liu, M.; Sung, M.; Mech, R.; Su, H. Deepmetahandles: Learning deformation meta-handles of 3D meshes with biharmonic coordinates. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12–21.
156. Huang, Z.; Stojanov, S.; Thai, A.; Jampani, V.; Rehg, J.M. Planes vs. chairs: Category-guided 3D shape learning without any 3D cues. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer Nature: Cham, Switzerland, 2022; pp. 727–744.
157. Xu, J.; Wang, X.; Cheng, W.; Cao, Y.P.; Shan, Y.; Qie, X.; Gao, S. Dream3D: Zero-shot text-to-3D synthesis using 3D shape prior and text-to-image diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 20908–20918.
158. Vincent, P.; Larochelle, H.; Bengio, Y.; Manzagol, P.A. Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 5–9 July 2008; pp. 1096–1103.
159. Hyvärinen, A.; Dayan, P. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.* **2005**, *6*, 695–709.
160. Norris, J.R. *Markov Chains*; Cambridge University Press: Cambridge, UK, 1998.

161. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
162. Du, Y.; Mordatch, I. Implicit generation and modeling with energy based models. In Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS): Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, Vancouver, BC, Canada, 6–12 December 2020; pp. 3603–3613.
163. Liu, R.; Wu, R.; Van Hoorick, B.; Tokmakov, P.; Zakharov, S.; Vondrick, C. Zero-1-to-3: Zero-shot one image to 3D object. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 9298–9309.
164. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10684–10695.
165. Shi, Y.; Wang, P.; Ye, J.; Long, M.; Li, K.; Yang, X. Mvdream: Multi-view diffusion for 3D generation. *arXiv* **2023**, arXiv:2308.16512.
166. Liu, Y.; Lin, C.; Zeng, Z.; Long, X.; Liu, L.; Komura, T.; Wang, W. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv* **2023**, arXiv:2309.03453.
167. Ye, J.; Wang, P.; Li, K.; Shi, Y.; Wang, H. Consistent-1-to-3: Consistent image to 3D view synthesis via geometry-aware diffusion models. *arXiv* **2023**, arXiv:2310.03020.
168. Shi, R.; Chen, H.; Zhang, Z.; Liu, M.; Xu, C.; Wei, X.; Chen, L.; Zeng, C.; Su, H. Zero123++: A single image to consistent multi-view diffusion base model. *arXiv* **2023**, arXiv:2310.15110.
169. Wang, P.; Shi, Y. ImageDream: Image-Prompt Multi-view Diffusion for 3D Generation. *arXiv* **2023**, arXiv:2312.02201.
170. Melas-Kyriazi, L.; Rupprecht, C.; Vedaldi, A. PC2: Projection-Conditioned Point Cloud Diffusion for Single-Image 3D Reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 12923–12932.
171. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
172. Reizenstein, J.; Shapovalov, R.; Henzler, P.; Sbordone, L.; Labatut, P.; Novotny, D. Common objects in 3D: Large-scale learning and evaluation of real-life 3D category reconstruction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 10901–10911.
173. Shao, Y.; Liang, S.; Ling, Z.; Yan, M.; Liu, H.; Chen, S.; Yan, Z.; Zhang, C.; Qin, H.; Magno, M.; et al. GWQ: Gradient-Aware Weight Quantization for Large Language Models. *arXiv* **2024**, arXiv:2411.00850.
174. Shao, Y.; Xu, Y.; Long, X.; Chen, S.; Yan, Z.; Yang, Y.; Liu, H.; Wang, Y.; Tang, H.; Lei, Z. AccidentBlip: Agent of Accident Warning based on MA-former. *arXiv* **2024**, arXiv:2404.12149.
175. Shao, Y.; Yan, M.; Liu, Y.; Chen, S.; Chen, W.; Long, X.; Yan, Z.; Li, L.; Zhang, C.; Sebe, N.; et al. In-Context Meta LoRA Generation. *arXiv* **2025**, arXiv:2501.17635.
176. Liu, T.; Hu, Y.; Wu, W.; Wang, Y.; Xu, K.; Yin, Q. Dap: Domain-aware prompt learning for vision-and-language navigation. *arXiv* **2023**, arXiv:2311.17812.
177. Yu, Z.; Dou, Z.; Long, X.; Lin, C.; Li, Z.; Liu, Y.; Wang, W. Surf-D: High-Quality Surface Generation for Arbitrary Topologies using Diffusion Models. *arXiv* **2023**, arXiv:2311.17050.
178. Chen, Y.; Pan, Y.; Li, Y.; Yao, T.; Mei, T. Control3D: Towards controllable text-to-3D generation. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 1148–1156.
179. Mercier, A.; Nakhli, R.; Reddy, M.; Yasarla, R.; Cai, H.; Porikli, F.; Berger, G. HexaGen3D: StableDiffusion is just one step away from Fast and Diverse Text-to-3D Generation. *arXiv* **2024**, arXiv:2401.07727.
180. Jiang, Z.; Lu, G.; Liang, X.; Zhu, J.; Zhang, W.; Chang, X.; Xu, H. 3D-togo: Towards text-guided cross-category 3D object generation. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 1051–1059.
181. Li, W.; Chen, R.; Chen, X.; Tan, P. Sweetdreamer: Aligning geometric priors in 2d diffusion for consistent text-to-3D. *arXiv* **2023**, arXiv:2310.02596.
182. Park, J.; Kwon, G.; Ye, J.C. ED-NeRF: Efficient Text-Guided Editing of 3D Scene using Latent Space NeRF. *arXiv* **2023**, arXiv:2310.02712.
183. Yang, H.; Chen, Y.; Pan, Y.; Yao, T.; Chen, Z.; Mei, T. 3Dstyle-diffusion: Pursuing fine-grained text-driven 3D stylization with 2d diffusion models. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 6860–6868.
184. Yu, Y.; Zhu, S.; Qin, H.; Li, H. BoostDream: Efficient Refining for High-Quality Text-to-3D Generation from Multi-View Diffusion. *arXiv* **2024**, arXiv:2401.16764.
185. Chen, Z.; Wang, F.; Liu, H. Text-to-3D using gaussian splatting. *arXiv* **2023**, arXiv:2309.16585.
186. Li, X.; Wang, H.; Tseng, K.K. GaussianDiffusion: 3D Gaussian Splatting for Denoising Diffusion Probabilistic Models with Structured Noise. *arXiv* **2023**, arXiv:2311.11221.

187. Vilesov, A.; Chari, P.; Kadambi, A. Cg3D: Compositional generation for text-to-3D via gaussian splatting. *arXiv* **2023**, arXiv:2311.17907.
188. Schoenberger, J.L. Robust Methods for Accurate and Efficient 3D Modeling from Unstructured Imagery. Ph.D. Thesis, ETH Zurich, Zürich, Switzerland, 2018.
189. Besl, P.J.; McKay, N.D. Method for registration of 3-D shapes. In *Sensor Fusion IV: Control Paradigms and Data Structures*; SPIE: Bellingham, WA, USA, 1992; Volume 1611, pp. 586–606.
190. Knapitsch, A.; Park, J.; Zhou, Q.Y.; Koltun, V. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Trans. Graph.* **2017**, *36*, 1–13.
191. Nocerino, E.; Stathopoulou, E.K.; Rigon, S.; Remondino, F. Surface reconstruction assessment in photogrammetric applications. *Sensors* **2020**, *20*, 5863. [CrossRef]
192. Yu, Z.; Chen, A.; Antic, B.; Peng, S.P.; Bhattacharyya, A.; Niemeyer, M.; Tang, S.; Sattler, T.; Geiger, A. SDFStudio: A Unified Framework for Surface Reconstruction. 2022. Available online: https://github.com/autonomousvision/sdfstudio (accessed on 25 June 2023).
193. Wang, P.; Liu, L.; Liu, Y.; Theobalt, C.; Komura, T.; Wang, W. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv* **2021**, arXiv:2106.10689.
194. Oechsle, M.; Peng, S.; Geiger, A. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 5589–5599.
195. Yariv, L.; Gu, J.; Kasten, Y.; Lipman, Y. Volume rendering of neural implicit surfaces. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 4805–4815.
196. Jiang, J.; Cao, M.; Yi, J.; Li, C. DI-MVS: Learning Efficient Multi-View Stereo With Depth-Aware Iterations. In Proceedings of the ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing, Seoul, Republic of Korea, 14–19 April 2024; pp. 3180–3184.
197. Liu, T.; Ye, X.; Zhao, W.; Pan, Z.; Shi, M.; Cao, Z. When epipolar constraint meets non-local operators in multi-view stereo. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 4–6 October 2023; pp. 18088–18097.
198. Ding, Y.; Zhu, Q.; Liu, X.; Yuan, W.; Zhang, H.; Zhang, C. Kd-mvs: Knowledge distillation based self-supervised learning for multi-view stereo. In *European Conference on Computer Vision*; Springer Nature: Cham, Switzerland, 2022; pp. 630–646.
199. Cao, C.; Ren, X.; Fu, Y. MVSFormer: Multi-View Stereo by Learning Robust Image Features and Temperature-based Depth. *arXiv* **2022**, arXiv:2208.02541.
200. Nan, L. 2018. Available online: https://github.com/LiangliangNan/MVStudio?tab=readme-ov-file (accessed on 1 November 2023).
201. Ding, Y.; Yuan, W.; Zhu, Q.; Zhang, H.; Liu, X.; Wang, Y.; Liu, X. Transmvsnet: Global context-aware multi-view stereo network with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8585–8594.
202. Zhou, Q.Y.; Park, J.; Koltun, V. Open3D: A modern library for 3D data processing. *arXiv* **2018**, arXiv:1801.09847.
203. Li, Z.; Yeh, Y.Y.; Chandraker, M. Through the looking glass: Neural 3D reconstruction of transparent shapes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1262–1271.
204. Liu, Y.; Wang, P.; Lin, C.; Long, X.; Wang, J.; Liu, L.; Kumora, T.; Wang, W. Nero: Neural geometry and brdf reconstruction of reflective objects from multiview images. *ACM Trans. Graph.* **2023**, *42*, 1–22.
205. Deng, C.; Jiang, C.; Qi, C.R.; Yan, X.; Zhou, Y.; Guibas, L.; Anguelov, D. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 20637–20647.
206. Gu, J.; Trevithick, A.; Lin, K.E.; Susskind, J.M.; Theobalt, C.; Liu, L.; Ramamoorthi, R. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3D-aware diffusion. In Proceedings of the International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; pp. 11808–11826.
207. Song, L.; Li, Z.; Gong, X.; Chen, L.; Chen, Z.; Xu, Y.; Yuan, J. Harnessing low-frequency neural fields for few-shot view synthesis. *arXiv* **2023**, arXiv:2303.08370.
208. Parameshwara, C.M.; Hari, G.; Fermüller, C.; Sanket, N.J.; Aloimonos, Y. Diffposenet: Direct differentiable camera pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 6845–6854.
209. Wang, Z.; Wu, S.; Xie, W.; Chen, M.; Prisacariu, V.A. NeRF--: Neural radiance fields without known camera parameters. *arXiv* **2021**, arXiv:2102.07064.
210. Bian, W.; Wang, Z.; Li, K.; Bian, J.W.; Prisacariu, V.A. Nope-nerf: Optimising neural radiance field with no pose prior. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 4160–4169.

211. Wang, S.; Leroy, V.; Cabon, Y.; Chidlovskii, B.; Revaud, J. DUSt3R: Geometric 3D Vision Made Easy. *arXiv* **2023**, arXiv:2312.14132.
212. Ma, Z.; Teed, Z.; Deng, J. Multiview stereo with cascaded epipolar raft. In Proceedings of the European Conference on Computer Vision, Tel Aviv-Yafo, Israel, 23–27 October 2022; Springer Nature: Cham, Switzerland, 2022; pp. 734–750.
213. Stereopsis, R.M. Accurate, dense, and robust multiview stereopsis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1362–1376.
214. Chen, X.; Zhang, H.; Yu, Z.; Opipari, A.; Chadwicke Jenkins, O. Clearpose: Large-scale transparent object dataset and benchmark. In Proceedings of the Computer Vision—ECCV 2022, Tel Aviv-Yafo, Israel, 23–27 October 2022; pp. 381–396.
215. Wizadwongsa, S.; Phongthawee, P.; Yenphraphai, J.; Suwajanakorn, S. Nex: Real-time view synthesis with neural basis expansion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8534–8543.
216. Sun, T.; Zhang, G.; Yang, W.; Xue, J.H.; Wang, G. TROSD: A new RBG-d dataset for transparent and reflective object segmentation in practice. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 5721–5733.
217. Chen, G.; Han, K.; Wong, K.Y.K. Tom-net: Learning transparent object matting from a single image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 9233–9241.
218. De Roovere, P.; Moonen, S.; Michiels, N.; Wyffels, F. Dataset of industrial metal objects. *arXiv* **2022**, arXiv:2208.04052.