A Comprehensive Study of Image-Based 3D Reconstruction Using Deep Learning

Ajit B. Aher [1,*] and R. A. Kapgate [2]

[1] Department of Mechanical Engineering, Sanjivani College of Engineering, Savitribai Phule Pune University, Maharashtra, India
[2] Department of Mechatronics, Sanjivani College of Engineering, Savitribai Phule Pune University, Maharashtra, India
Email: aajit75@gmail.com (A.B.A.); rakapgate2007@gmail.com (R.A.K.)
*Corresponding author

***Abstract*—The fast progress of Three-Dimensional (3D) reconstruction has led to the emergence of advanced Deep Learning (DL) approaches and techniques. Leveraging the technology of computers to produce realistic three-dimensional representations of objects has grown into an essential component of extensive study in a variety of domains. This review article investigates the cutting-edge methodologies, difficulties, and potential in this research field. The state-of-art study follows the development of Deep learning techniques with graphics expertise, which strengthens the requirement for good efficacy with better performance of 3D reconstruction. The research work begins by discussing classic strategies for 3D reconstruction with active and passive techniques that emphasizes their limitations with the need for cutting-edge practices. The various types of neural network architectures employed, like Convolutional Neural Networks (CNNs), autoencoders, and Generative Adversarial Networks (GANs) are explored with auxiliary information. This review aims to provide researchers and practitioners with a thorough understanding of the advancements, problems, and prospects in image-based 3D reconstruction while opting for the progressions in Deep Learning. Further, this research study presents the development in Neural Radiance Fields (NeRF) which is revolutionizing image-based rendering for efficient 3D reconstructions.**

***Keywords*—Three-Dimensional (3D) reconstruction, Deep Learning (DL), Convolutional Neural Network (CNN), Generative Adversarial Network (GAN), Neural Radiance Fields (NeRF)**

I. Introduction

Image-based Three-Dimensional (3D) reconstruction aims to create 3D model by operating the Two-Dimensional (2D) images, a process that has been revolutionized by deep learning techniques. The research trends explore the application of deep learning techniques which enables more accurate and efficient reconstruction of 3D models from 2D images. This has significant potential for numerous industrial applications, especially in design, manufacturing and maintenance. This review focuses on leveraging deep learning techniques to improve the 3D reconstruction process than the traditional approaches. Deep learning research has made considerable progress in the area of image-based 3D reconstruction, addressing many limitations presented by previous approaches [1]. 3D reconstruction from images has been a long-standing goal in computer vision, with wide-ranging applications from industrial revolutions. 3D reconstruction technological advances help to build exact digital representations by collecting the 3D geometrical information of authentic objects. It may record and store the geometry along with the design of physical equipment or mechanical parts, leading to the digital foundation for equipment maintenance.

Unlike typical manual 3D modelling, utilizing Computer-Aided Design (CAD) or Digital Content Creation (DCC) applications, the 3D reconstruction methodology starts with sensor input, such as images, point clouds, and additional data [2]. 3D reconstruction is categorized into both implicit and explicit representation strategies based on distinct methodologies, which provide a variety of viewpoints and processing methods to analyze real-world data. The term, explicit expression states to a representation method that uses explicitly defined geometric shapes and frameworks to express an object's exterior or interior geometry. The object's topology is implicitly specified by an appropriate function or mathematical equation, which is then utilized to solve the issue, where the values may be collected from spots on the surface. Training along with representing 3D models has grown into a standard survey procedure in 3D plane survey analysis. Analysing several images and then reconstructing the form and structure in three dimensions is a key goal in computer vision. Conventional multiview 3D reconstruction methods use established camera settings to extract and match important elements from images. Nevertheless, these methods are ineffective and does not completely utilize the benefits of multiview data. In the past few years, deep learning-based approaches for 3D reconstruction have grabbed the curiosity of several researchers worldwide [3]. These unique algorithms may estimate an object's or scene's 3D form intuitively through

Manuscript received June 16, 2025; revised July 4, 2025; accepted August 20, 2025; published January 15, 2026.

start to finish, avoiding the requirement for various steps including key-point detection and a successful match. Furthermore, these distinctive strategies can rebuild objects' forms given a single input view [4]. Using one or more RGB photos, Han *et al.* [5] have concentrated on deep learning approaches to estimate the 3D geometry of common items. According to their research study, the 3D geometrical structure of the multiple 2D images was determined using 3D image reconstruction [6]. Deep learning breakthroughs have transformed multiview 3D reconstruction by making end-to-end 3D shapes [7, 8]. To increase reconstruction quality and decrease processing efficiency, several representations, including volumetric, surface-based, and intermediary representations, were described. The state-of-art study helps to understand various techniques and methodology for 3D reconstruction. This paper presents a comprehensive summary of current advances in image-based 3D reconstruction. The explored approaches are analyzed from a variety of perspectives, including input kinds, model architectures, output representations, and learning strategies. The objectives of the research study are enlisted in Table I.

TABLE I. OBJECTIVES OF THE RESEARCH STUDY

Objectives	No	Description
Investigate Existing Techniques for 3D Reconstruction	(1)	Study traditional and deep learning-based methods for 3D reconstruction.
	(2)	Explore different deep learning approaches used for 3D reconstruction.
	(3)	Analyze evolution of deep learning techniques and approaches for 3D reconstruction.
Analyze Strengths and Limitations of Deep Learning Techniques	(1)	Assess the accuracy, robustness, and efficiency of different Deep Learning (DL) models.
	(2)	Discover common challenges and limitations of DL techniques and approaches for 3D reconstruction.
Explore the Future Research Trends and Enhancements	(1)	Identify gaps in current research and potential areas for improvement.
	(2)	Examine emerging trends, self-supervised learning, transformer-based architectures, and diffusion models for 3D generation.

The research work is presented into various sections, where Section II gives an overview of image-based conventional 3D reconstruction approaches and techniques. Section III outlines the deep learning techniques for 3D reconstruction. Section IV entails the state of art key studies and Section V summarizes discussion and future direction. Finally, Section VI concludes the research work.

II. IMAGE BASED 3D RECONSTRUCTION

The field of image-based 3D reconstruction involves creating three-dimensional models from two-dimensional images. It includes rebuilding and comprehending the 3D structure of objects and situations using two-dimensional images data. 3D visualization methods employ data from cameras or sensors to create a digital representation of the forms, structures, and attributes of objects in a scene. This technology has broad applications in areas such as Computer Vision, Robotics, Virtual Reality (VR), and Augmented Reality (AR). The 3D reconstructions create respective 3D models by extracting, processing, and analyzing 2D visual input. While performing 3D reconstructions, as shown in Fig. 1, it utilizes several algorithms and data collecting approaches that allow automated 3D vision models to rebuild the dimensions, outlines, and spatial coordinates of the objects in each visual environment.

In 3D reconstruction, explicit formulations are clearer and more accurate, while implicit formulations provide flexibility and efficient storage. Choosing the right representation depends on the specific application needs for 3D reconstruction in computer vision [9, 10]. The 3D models are extracted & built either by using the input from special sensors, which is referred as active data capture or by using the input from the regular cameras called as passive technique. The initial processes, like the Structure from Motion (SfM) and Multiview Stereo (MVS), rely heavily on feature matching and geometric constraints, which often result in limitations regarding robustness in complex environments.

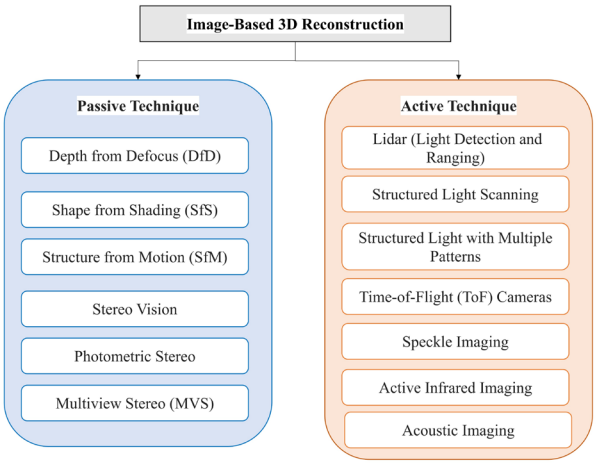


Fig. 1. Conventional image-based 3D reconstruction.

A. Passive Techniques

Passive visualization approaches for image reconstruction entail gathering detailed information from the environmental surroundings without actively transmitting any signal or light. These strategies depend upon ambient light or natural radiation. It seamlessly analyzes images or videos captured using currently available light sources. Few of the traditional passive methods which are commonly used in 3D reconstruction are:

1) Depth from Defocus (DfD)

DfD evaluates the depth or 3D structure of an object by determining the amount of blur or defocus in certain sections of an image. It operates on the concept that objects at different distances from the camera lens will display

varied degrees of defocus blur. Pentland [11] suggested a technique for detecting depth of scenes by determining the degree of defocus functions in the image being examined. The method is intriguing since it involves no correspondence [12]. The challenges with DfD include requiring multiple images with different focus settings, making it time-consuming, and difficulty in texture-less regions, increasing computational cost.

2) Shape from Shading (SfS)

SfS reconstructs an object's 3D shape from a single 2D image. This approach examines how light strikes an object with its shading patterns and how bright various parts seem with the intensity variations. It depends on the direction of the light source and the reflectance properties of the surface. It works better for reconstructing smooth surfaces. The visual data derived from a particular object's coloring can be utilized to reconstruct the contour for the observed surface [13].

3) Structure from Motion (SfM)

SfM collects the set of images of a scene from different perspectives with a single camera. The initial stage is to identify elements that are common among these images, such as corners, edges, or particular patterns. SfM then computes the location of cameras with its orientation for every image depending on the recognized features and their appearance from various perspectives as depicted in Fig. 2. Triangulation is used to establish the 3D position of the characteristics in the scene by contrasting matching features across multiple images.

Fig. 2. Various stages of SfM technique.

4) Stereo vision

Stereo vision implements the varied perception with more than two cameras placed at distinct viewpoints to capture images of the same scene. This approach works by identifying comparable spots in both images and determining their 3D coordinates using the specified camera geometry. Stereo vision techniques use inequalities, or the difference in the locations of comparable points, to determine the depth of locations throughout a scene [14]. This depth data enables precise reconstruction of complex 3D models. It mimics human binocular vision to perceive depth. Overcoming the correspondence issue associated with image pairings represents one of the key problems of stereo vision. Another problem with stereo vision is facing computational complexity and latency mostly for real-world applications.

5) Photometric stereo

Photometric stereo seizes several images of an object under divergent lighting conditions; this was presented by Robert Woodham [15]. The word "photometric" relates to the measurement of light, whereas "stereo" means the utilization of multiple images. The variations in shading are used to infer the surface normals and reconstruct the 3D shape of the object. Consider the light rays falling on a surface where, N is surface normal, L is the input light direction and V is the output light direction. Both, L and V make the respective angles with the normal N referred to as Radiance along direction L and Radiance along direction V, computed using Eq. (1).

Radiance along V = Bi-directional reflectance function (·) Radiance along L

$$L_\gamma = \rho(\theta_i, \theta_\gamma)\, L_i \cos\theta_i \tag{1}$$

6) Multiview Stereo (MVS)

Multiview Stereo (MVS) extends stereo vision to multiple viewpoints for 3D reconstruction. It represents 3D shapes using dense point clouds or surface meshes, aiding photorealistic rendering. MVS generates cohesive models using depth maps and 3D fusion while aligning images based on geometry and camera settings [16]. Advances in processing power and algorithms have improved MVS, with key contributions from Delaunoy *et al.* [17] and Seitz *et al.* [18]. The main challenge is accurately computing dense pixel correlations, as matching pixels across views remains difficult [19].

B. Active Techniques

Active 3D reconstruction techniques use any type of radiation, like sound, radio waves or light illuminating an object. It then examines the reflected light, reverberates and deformation to recreate the 3D structure of the item. These approaches use the reflection, dispersion, or absorption of the transmitted signal to obtain data regarding the surroundings. The active techniques provide better management for the imaging conditions with reliable environmental situations and applications. Typically used active strategies are as below:

1) Light Detection and Ranging (LiDAR)

Single-photon LiDAR is emerging as an effective solution for distance imaging in challenging environments. It works by emitting laser pulses and measuring the time it takes for the light to reflect off objects, generating a point cloud, which is a collection of data points used to create 3D representations of surfaces, shapes, and objects. LiDAR is capable of detecting a wide range of elements, such as physical objects, chemical substances, and even clouds. It is widely used in fields like aviation, topographic mapping, and autonomous navigation. It represents a result of extensive advancements in laser technology, optics, and remote sensing. However, LiDAR faces limitations under certain environmental conditions, like rain, fog, snow, or humidity, which can distort data. It also struggles with accurate measurements on specular surfaces, such as mirrors or glass. Despite these challenges, LiDAR continues to be crucial in sectors like self-driving vehicles, archaeology, and environmental research [20, 21].

2) Structured light scanning

This technique projects a well-planned pattern of light onto a visual scene. Grids, horizontal stripes, and more intricate designs are just a few of the various shapes that this light pattern may take. The light beams become

warped when the light pattern affects objects of various shapes and levels. To determine the 3D shape, a camera records the distorted pattern. The capacity to quickly produce accurate and high-resolution 3D models is one of the technique's main benefits. Reverse engineering, 3D printing, and CAD modeling are just a few of the uses for it.

3) Structured light with multiple patterns light scanning

Conventional 3D capture methods run a single scanning laser stripe across a target object's surface in a sequential manner. In order to acquire data using this approach, the object must stay still while many stripe photos are taken. A variety of light patterns with different spatial frequencies are projected onto the object to overcome issues like ambiguity and albedo (surface reflectivity). By analyzing the updates or highlights in the collected photos, these patterns make it possible to recreate a 3D geometry with greater accuracy [22]. This method is frequently used in industrial inspections and 3D scanning, which requires high precision. It is the outcome of years of computer vision and 3D scanning research and development by several research groups (1995–2010).

4) Time-of-Flight (ToF) cameras

Time-of-Flight (ToF) technique measures the time taken for a light pulse to travel to an object and return. Here the distance estimation is based on the reflected light time-of-flight [23]. For each pixel in the sensor array data is captured, producing a 3D depth representation of the scene. ToF sensors give depth information for every point as compared to conventional cameras that just record color or brightness. This enables to create reconstructions of the surroundings. Common applications of ToF includes gesture recognition, industrial automation, and augmented reality. The technology development from 1930 to 2000 has focused on accurately measuring ranges by timing light signals. A significant challenge for multi-camera ToF setups is Multiple Camera Interference (MCI), requiring measures to prevent electromagnetic interference between cameras.

5) Active infrared imaging

It makes use of infrared light sources for illuminating the scene and collecting the reflected infrared light. Maiman [24] and Military-Defense Research Laboratories are significant personalities as well as organizations in the evolution of infrared technology. Active infrared imaging is the use of infrared light sources to illuminate a scene and capture the reflected light to form an image. The working of active infrared imaging yields better results in low-light or nighttime conditions; thus, it is more useful for night vision, surveillance, inspection of industry, etc.

6) Acoustic imaging

An acoustic imager uses sound like a camera uses light, detecting echoes to create images. It maps loudness with colors and is used in sonar and ultrasound imaging. Key contributors include Paul Langevin and Lewis Fry Richardson, with major advances in signal processing since 1970s. Challenges include precise geometry measurement, channel response estimation, and time synchronization.

As depicted in Fig. 3, the conventional 3D reconstruction methods are timelind with the active and passive techniques.

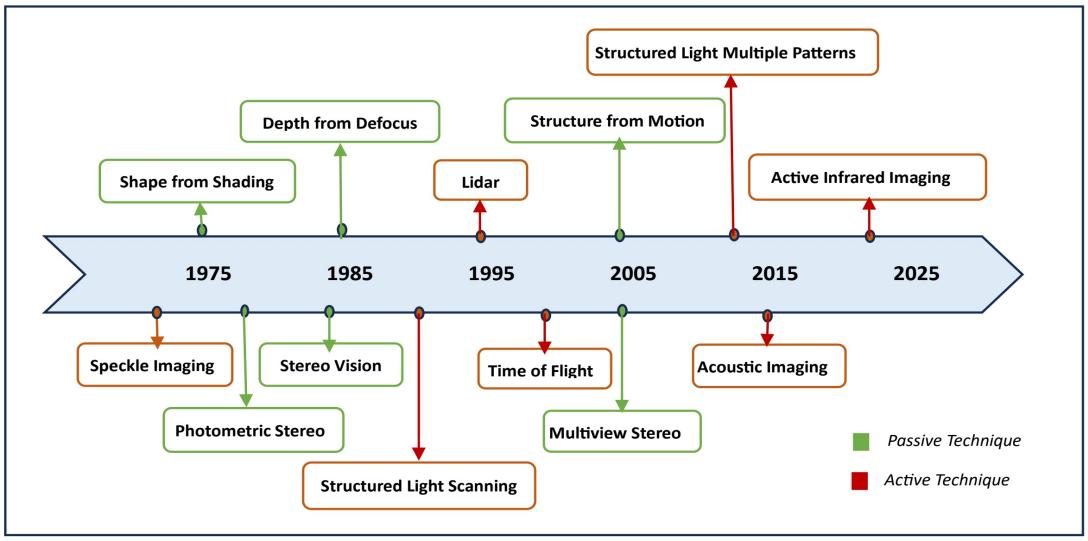


Fig. 3. Active and passive techniques for conventional 3D reconstruction methods.

III. Deep Learning Techniques for 3D Reconstruction

While the classical methods covering active and passive techniques for image-based 3D reconstruction have been effective, they often require meticulous calibration of camera pose and are highly sensitive to noise and occlusions. The emergence of deep learning introduced data-driven approaches that can learn complex mappings from images to 3D structures, offering improved robust 3D reconstruction. Deep learning techniques have significantly advanced image-based 3D reconstruction, leading to notable improvements in both reconstruction quality and robustness.

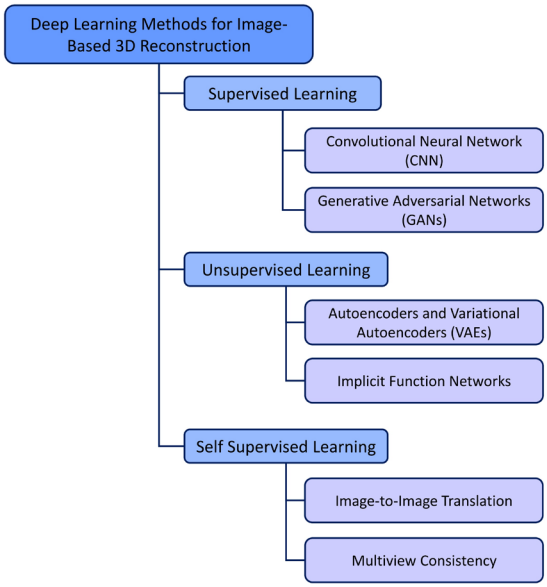


Fig. 4. Deep learning techniques for 3D reconstruction.

New horizons in Artificial Intelligence are being opened by expanding deep learning techniques. This is possible through the use of deep learning techniques, sensor emancipation and the acceptance of concurrent active and passive methodologies [25]. Conventional techniques for reconstructing a single image in three dimensions rely on specific lighting and reflectance presumptions, making them extremely vulnerable to changes in the input's reflective power, illumination, and texture. 3-dimensional forms of the objects are rebuilt with approaches using mathematical characteristics as contours, vertical-horizontal lines, and points. Other approaches use shading and repeating texture elements [26]. An overview of artificial intelligence-based techniques for 3D geometry reconstruction from a single image is presented in many studies, which investigates the possibilities of Variational Autoencoders (VA), Generative Adversarial Networks (GAN), Convolutional Neural Networks (CNN), and Zero-Shot techniques [27]. Deep learning techniques have evolved in recent past few years due to the quick growth of neural networks and the introduction of fully decentralized 3D model datasets. ShapeNet became the benchmark dataset, commonly used for evaluating 3D generative models like 3D-GAN, Occupancy networks, and DeepSDF [28]. In 3D planar survey interpreting, the training and representation of 3D models has grown into standard procedure. With the advancement of deep learning techniques since 2015, image-based 3D reconstruction using CNN has gained interest of researchers because of the impressive performance of the deep learning algorithms. Fig. 4 represents the deep learning techniques, which are classified in three categories as: (1) Supervised Learning, (2) Unsupervised Learning and (3) Self-Supervised Learning.

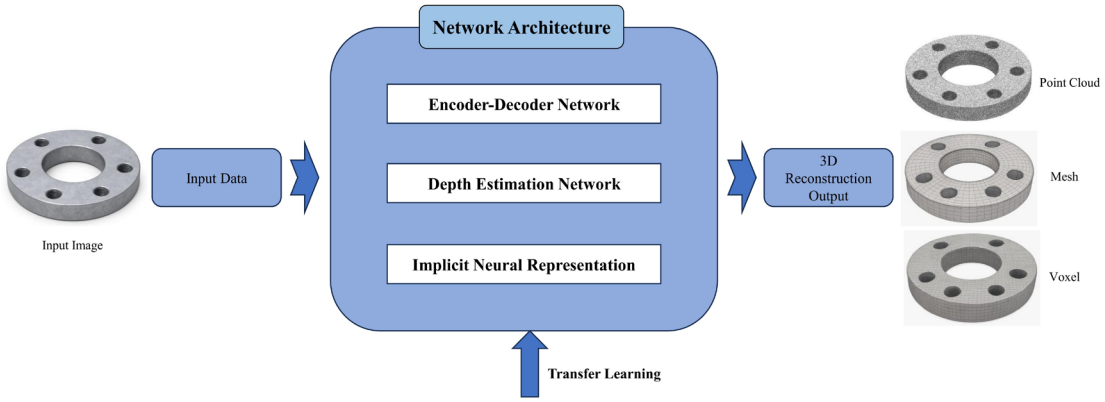


Fig. 5. Framework for 3D reconstruction using deep learning.

A. 3D Reconstruction Using Deep Learning

The various research studies are trying to solve the problem of 3D reconstruction. Convolutional neural network is one of the widely used technique for image-based 3D reconstruction for decades, and it's efficacy has been outstanding and has drawn growing attention. The three phases of the 3D reconstruction approach are as follows: first, a CNN is trained using a input image dataset to anticipate and recognize the object's features and to reliably estimate the locations from one representation in the image space. Creating the object's geometric form (mesh) is the second phase. In order to identify the surface referrals which, relate to each object polygon the third and final step of the method involves automatically translating the 3D space of the object into the 2D image space. This process yields better visual output, with respect to given inputs, including structure, expression, reflectance, and illumination [29]. As shown in Fig. 5 the framework for 3D reconstruction using deep learning consists of components as:

(1) Input Image Data;
(2) Output 3D Representations;
(3) Network Architectures;
(4) Transfer Learning.

A deep learning architecture uses an image as input and creates a 3D model output in the form of a mesh, voxel grid, and point cloud.

1) Input image data

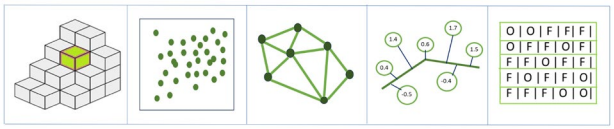
A dataset is a group of data with information unique to its category. The various types of datasets are real dataset, generated dataset, synthetic dataset, etc. with single image [30] or multiple images as data input. Typical data

inputs include subsurface depth data, LiDAR imaging point-cloud details, camera pictures, and inertial observations. RGB channel analysis is performed on camera pictures & creation of dense depth maps is facilitated by LiDAR data.

The most popular method for deep learning-based 3D reconstruction makes use of both synthetic and real-world data. ShapeNet is a vast collection of synthetic CAD models with extensive annotations. It was widely used to train, evaluate, and compare approaches, allowing for consistent assessment across algorithms, and it became a common benchmark in the literature on deep 3D reconstruction [31]. ScanNet, Matterport3D, and TUM RGB-D/S3DIS offer extensive real-world RGB-D data with semantic labelling and dense geometry for scene-level reconstruction. Usually, the parametric CAD models from the ABC-Dataset, CC3D, Fusion360 Gallery and 3D CAD model dataset are used in geometric deep learning research [32]. Emerging datasets such as Objaverse-XL push boundaries, offering millions of the annotated meshes and implicit-field representations tailored for generative and neural rendering techniques [33].

2) Output representations

Output representation is important in the selection of network architecture, it also has the impact on the quality of reconstruction and the computational efficiency. Fig. 6 depicts the commonly used representations in 3D deep learning, such as Voxel, Point Cloud, Meshes, Sign Distance Function (SDF) and Occupancy Grid.



(a) (b) (c) (d) (e)

Fig. 6. Data representation methods in 3D deep learning. (a) Voxel; (b) Point Cloud; (c) Mesh; (d) SDF; (e) Occupancy Grid.

a) Voxel grid

Voxel grids are a fundamental method for representing 3D data in Deep Learning, extending the concept of 2D pixels to 3D voxels. Each voxel, or volumetric pixel, stores information such as occupancy, probability, or material properties, and forms a grid that divides 3D space into cubic cells. Voxel grids offer a structured, regular representation of 3D space, making it easier to process with 3D convolutional neural networks, which can capture spatial patterns in the data. However, at high resolutions, voxel grids require significant memory and computational power, as the number of voxels increases cubically with resolution. At low resolutions, they may lose fine details of objects, and real-world 3D data is often sparse [34], requiring specialized data structures like octrees for efficiency.

b) Point cloud

Point clouds are the fundamental 3D representation method, which is used for efficiently representing 3D data for deep learning. It consists of (x, y, z) points with optional attributes like color or intensity. As point clouds are memory efficient compared to voxel grids, they are used in the applications like 3D object recognition, segmentation, and reconstruction. However, because of the unordered nature and sensitivity to noise it is challenging to use in the applications which requires dense information. Deep learning models like PointNet, PointNet++ [35, 36], and Dynamic Graph CNNs (DGCNN) solve these challenges with specialized architecture. PointNet gathers global features and independently processes points using Multi-Layer Perceptron (MLP). PointNet++ introduces a hierarchical structure which captures both local and global features. With edge convolution and dynamic graphs DGCNN analyzes local geometry. The needs of particular activities, such as the level of feature extraction and the handling of point density, determine which design is best. Reducing noise in the point clouds, which is particularly valuable for the condition assessment and 3D reconstruction, Emadi and Limongiello [37] presents a novel approach by integrating deep learning and clustering models to improve the quality of point clouds. Point cloud completion is a significant challenge due to incomplete or sparse data [38]. It is the task of generating a complete 3D representation of an object from a partial or incomplete point cloud input. Completing point clouds is usually not easy since they are naturally chaotic and unstructured, Point Cloud Network (PCN), Unpaired scan completion network, Morphing and sampling-based network, PF-Net, GRNet and SnowflakeNet are models used in Point Cloud completion [39].

c) Meshes

Meshes are a fundamental 3D representation used in graphics, computer vision, and deep learning. It is composed of vertices, edges, and faces which makes them ideal for high-fidelity modelling. Compared to voxel grids and point clouds, meshes are more precise but complexity and high computational demands is a challenge in using meshes [40]. These problems are addressed by specialized deep learning models, MeshCNN that uses convolutions to mesh edges, GCNs which treat meshes as graphs to capture vertex interactions, and mesh autoencoders learn compact representations for unsupervised tasks.

d) Signed Distance Functions (SDFs)

In Signed Distance Functions (SDF), 3D shapes are represented by defining a scalar field over 3D space. Value of each point indicates the shortest distance to the shape's surface. The sign of the distance shows whether the point is inside or outside the shape with negative or positive sign respectively. It provides a continuous and smooth representation, captures fine details and complex geometries. This makes SDF suitable for tasks like shape reconstruction and rendering which require high precision. They are differentiable due to their continuous nature, which helps neural network learning and optimization tasks [41]. However, because SDFs require dense grids and use a lot of memory, at high resolution its creation and processing are computationally demanding. Converting other representation, such as meshes or point cloud to SDFs can be complex and error-prone.

e) Occupancy grids

An occupancy grid is a discrete representation of 3D space, that divides it into a regular grid of cells (voxels) and labels each one as either free (empty space) or occupied (within an object). It can be probabilistic, storing the probability that a voxel will be occupied. Occupancy grids are simple, intuitive, and compatible with 3D convolutional neural networks. However, high-resolution grids require significant memory and computational resources, limiting scalability, while low-resolution grids may miss fine details. In large grids, processing sparse data effectively can be difficult and frequently calls for specialized data structures [42]. The decision between SDFs and occupancy grids depends on the particular requirements of detail, resources, and task type. Occupancy grids are well suited for applications such as robotics and 3D object detection.

f) Volumetric representations

In 3D deep learning, volumetric representation is a key approach in which 3D space is divided into a regular grid of voxels (3D pixels). A value corresponding to a 3D space attribute, like occupancy, density, or color, is stored in each voxel. As voxel is compatible with 3D convolutional neural networks, it can be used for jobs requiring in-depth 3D analysis. This allows the capture of spatial relationships and structural features. But scalability is limited by high computing resources which demand to represent high-resolution data with voxels. Accuracy may be impacted by low-resolution grids to capture minute details [31].

Techniques like 3D CNNs, VoxelNet, and 3D U-Net leverage voxel grids for various 3D tasks. 3D CNNs extend 2D convolutions to three dimensions, capturing spatial hierarchies across the grid. VoxelNet combines feature learning and 3D object detection by processing raw point clouds divided into voxels. 3D U-Net, an extension of 2D U-Net, uses an encoder-decoder structure with skip connections to capture detailed spatial information and context, making it effective for tasks like object detection, shape reconstruction, and segmentation. PointNet/PointNet++, DGCNN, CurveNet uses point cloud. Deep learning for meshes applies mesh-aware convolution and graph-based neural models. Table II shows the comparative analysis of different 3D data representation methods and suitability for practical applications.

TABLE II. COMPARISON OF 3D REPRESENTATION TECHNIQUES

Representation	Techniques	Pros	Cons	Applications
Voxel/Volumetric	Represents 3D shapes of regular cubes, 3D grid of voxel.	Easy to use with 3D Convolutional Neural Networks (CNNs), can handle any spatial structure.	High memory/computation, loses detail at low resolutions.	3D CNN tasks: segmentation, object analysis, object detection [31, 43, 44]
Point Cloud	Unordered point sets representation of points in 3D space.	Memory-efficient, directly represent the 3D shape from sensor output.	Requires special networks, sensitive to noise.	Recognition, classification, segmentation, and real-time scene understanding. [35–39, 45–51]
Mesh	Represents combinations of structured vertices, edges and faces.	High fidelity, compact surface representation.	Complex graphs, heavy computation, and Difficult for the network to learn.	Graphics, surface reconstruction, high-detail modeling. [40, 52, 53]
Sign Distance Function (SDF)	3D shape representation through a set of continuous signed distance fields.	Smooth, detailed, Strong representation ability.	Difficulty in handling complex shapes, conversion overhead issue.	High-precision reconstruction, implicit surface modeling. [41, 54, 55]
Occupancy Grid	Discrete representation of 3D space with voxel occupancy probability.	Simple, CNN-ready, supports probabilistic reasoning.	Scaling issue, difficult to represent small objects or fine details.	Robotics, 3D detection, semantic mapping. [42, 56–58]

3) Network architecture

The physical as well as logical layout of the technology, software, standards, along with the medium used in transmitting information forms a network architecture. It determines the efficiency and accuracy of the reconstruction process, influence scalability, adaptability, and computational resource requirements. A clear understanding of these frameworks helps in selecting the most suitable model for specific applications. The different architectural structures used in 3D reconstruction using Deep Learning techniques are as discussed below:

a) Convolutional Neural Networks (CNNs)

CNNs play a crucial role in 3D reconstruction tasks by processing volumetric data, point clouds, or multi-view images to extract features, learn representations, and generate accurate and detailed reconstructions of 3D shapes. These CNN architectures as represnted in Fig. 7 have significantly advanced the field of 3D reconstruction and are widely used in various applications, including computer vision, robotics, augmented reality, manufacturing industries, etc. Unlike conventional CNN methods that handle 2D data, 3D CNNs use special filters to immediately extract important characteristics from geometric data, including three-dimensional representations of objects or medical scans. This learning technique is able to interpret temporal characteristics and spatial connections as it can analyze data in three dimensions.

Consequently, 3D CNNs work well for applications such as precise segmentation of medical visuals for diagnosis, video analysis, and 3D object identification. Convolutional neural networks are widely used for feature extraction from images, which is a crucial step in image-based 3D reconstruction. CNN architectures such as ResNet, VGG, and MobileNet are commonly employed for their ability to extract hierarchical features from input images. CNNs have been instrumental in advancing 3D

reconstruction tasks, particularly in handling volumetric data and processing point clouds. CNNs include deep neural networks which employ convolution rather than matrix multiplication and train various types of layers.

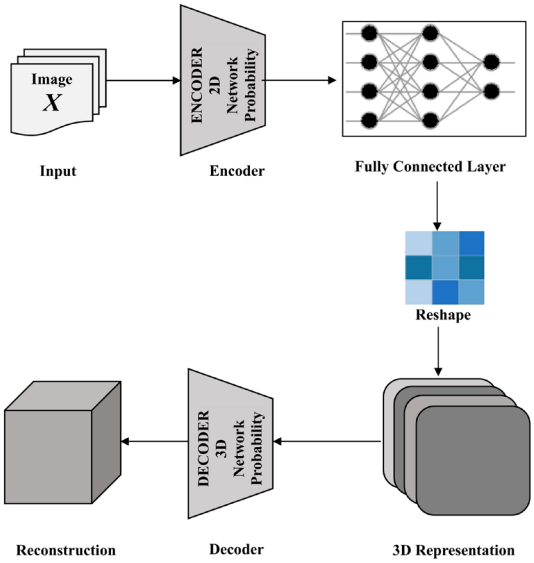


Fig. 7. CNN model for 3D image reconstruction.

b) Graph Neural Networks (GNNs)

GNNs are neural network architectures designed for processing graph-structured data, which can represent relationships between 3D points or voxels. Initially, they were used to organize building units into conventional and irregular structures through unsupervised learning. A graph G is non-Euclidean organization composed up of a collection of edges E and set of vertices V. An edge in a graph is represented by its nodes as $e_{ij} = (v_i v_j) \in E$, where v_i and $v_j \in V$. A graph is written as stated in Eq. (2).

$$G = VE \tag{2}$$

A graph can alternatively be represented like an adjacency matrix (A), as shown in Eq. (3).

$$A = \begin{matrix} \{ \text{if } e_{ij} \in \text{E then } A_{ij} = 1 \} \\ \{ \text{if } e_{ij} \notin \text{E then } A_{ij} = 0 \} \end{matrix} \tag{3}$$

Additional data could be contained in the characteristics of any node, edge, or entire graph by representing them as vectors. GNNs are increasingly being applied to 3D reconstruction tasks to exploit geometric relationships between points and improve reconstruction accuracy. A key challenge in graph development is the manual conversion of geometric primitives, meshes, floor plans, and Building Information Modeling (BIM) models into graphs or point clouds.

c) Variational Autoencoders (VAEs)

VAE models are probabilistic generative methods which facilitate latent representations of data to generate intelligent information from the learned latent space. They are crucial for generating high-dimensional data due to their capacity to integrate stochastic data representation with the efficacy of deep learning techniques [59]. In 3D shape generation, VAEs can optimise the geometry of 3D shapes to build innovative shapes by applying learned latent space knowledge. VAEs are taught to rebuild input 3D forms while minimising the disparity among the learnt latent probability and previous distribution, such as the typical normal distribution.

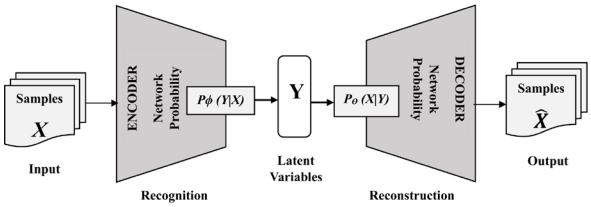


Fig. 8. VAE model for 3D image reconstruction.

It reconstructs the original data, and the metric employed assesses the disparity between the input data and the output data, as represented in Fig. 8, by calculating the Reconstruction Loss, which, in most cases, is Mean Squared Error (MSE) as depicted in Eqs. (4) and (5).

$$Loss = ||X - \hat{X}||^2 = ||X - P_\theta\left(\frac{X}{Y}\right)||^2 \tag{4}$$

$$Loss = ||X - P_\theta(P_\phi\left(\frac{Y}{X}\right))||^2 \tag{5}$$

d) Generative Adversarial Networks (GANs)

GANs are made up of two distinct artificial neural relationships: a generator and a discriminator, which are programmed to create authentic patterns. In 3D shape generation, GANs can generate new 3D shapes by training the generator to create shapes that are identical from real shapes, as determined by the discriminator. 3D-GANs and various conditional GAN architectures, as displayed in Fig. 9, have been developed for generating 3D shapes with specific properties. Generally, GANs generate better photorealistic pictures compared to VAEs. After pretraining the generator using L2 regularization, it uses phase difference as input and modifies the network design to do basic imaging tasks.

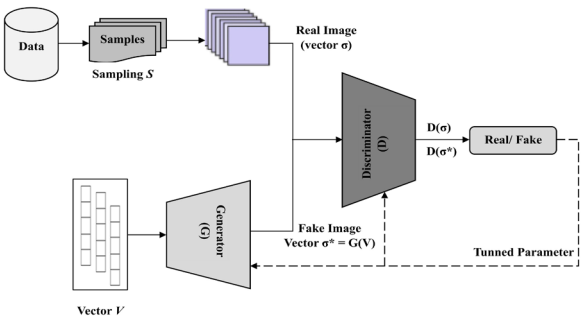


Fig. 9. GAN model for 3D image reconstruction.

Then it links it to a discriminator to create a counter network, and adds a cross verification set to track its convergence.

3D-GANs and various conditional GAN architectures as displayed in Fig. 9, have been developed for generating 3D shapes with specific properties. The primary aspect of this generator model is a maximum likelihood estimate, that enhances the possibility that the generator would

recreate conductivity using the actual distribution as implemented adapting the Eq. (6).

$$L(G,V,\sigma) = \prod_{i=1}^{M} P_{G(V)}(\sigma^{(i)};\ \theta_g) \quad (6)$$

In recent advancement of 3D deep learning, 3D CNNs are remarkably enhanced by hybrid architectures like Point-Voxel CNN (PVCNN), which has tackled the memory and computation essential to full-resolution volumetric processing. PVCNN combined with point-based sparsity with voxel-based locality has achieved up to 10× memory reduction and much higher conversion speed. GNNs provides fascinating alternative for unstructured and irregular 3D data. However, complexity and memory demands scale with graph size, transformer-based graph architectures help manage large-scale graphs efficiently. Using compact latent representations, VAEs achieve a modest balance between strong generalization through probabilistic modeling, a smaller memory footprint, and less computational load than GANs. The maximum visual realism is achieved by GANs, but at very high computational and memory costs. Hierarchical and hybrid techniques, including VAE GAN variations, assist reduce mode collapse and enhance variety.

These architectures collectively provide a complex set of trade-offs, although they need more resources, GANs push realism and quality, VAEs provide effective and generic reconstructions. CNN-based techniques can be fine-tuned for speed and resource utilization, and GNNs performs much better in flexible network topology modeling. In 3D reconstruction deep learning applications, this comparative research emphasizes the trade-offs between various neural network topologies in terms of computing demands, memory needs, and generalization capabilities. Table III depicts the key insights on the trade-offs between various network architectural methods on computational complexity, memory footprint, generalization ability, and suitability for real-world deployment.

TABLE III. TRADE-OFF BETWEEN VARIOUS NETWORK ARCHITECTURES

Network Architecture	Computational Complexity	Memory Footprint	Generalization Ability	Accuracy Indicators	Suitability for Real-World Deployment
CNNs	High: 3D CNNs are computationally intensive due to 3D convolutions, and requires optimizations to reduce latency.	Large memory use for full-resolution 3D grids; hybrid point-voxel approaches lower GPU demand by ~10×.	Effective on structured data, but performance drops on unstructured or irregular scenes; multiview CNNs help.	CNN models usually achieve high Intersection-over-Union (IoU) but only moderate Chamfer Distance, yielding decent reconstruction quality but limited surface detail.	Mature tech with real-time capabilities; optimized 3D CNNs used in robotics, medical imaging, AR/VR pipelines. [5, 39, 60–62]
Graph Neural Networks (GNNs)	Moderate to high: complexity increases with graph size; graph-transformer models improve efficiency.	Large memory requirement for storing adjacency or edge features. Optimizations in learning techniques handle large-scale graphs efficiently.	Adept at handling unstructured data, suitable for diverse 3D reconstruction tasks. Strong generalization capabilities.	Achieve low Chamfer Distance (CD), confirming superior surface accuracy and detail preservation compared to voxel CNNs and VAEs. Against Generative Adversarial Networks (GANs), GNNs often win on CD.	Emerging field; used in 3D face and mesh reconstruction with promising results, though deployment is still limited. [39, 60–62]
Variational Autoencoders (VAEs)	Moderate: VAEs are generally less computationally demanding than GANs, as encoder–decoder pairs are optimized for reconstruction tasks.	Moderate memory footprint; latent bottleneck ensures compressed representations.	Good generalization via probabilistic latent modelling; Reconstructing diverse 3D structures.	Produce smoother reconstructions; fidelity may lag behind CNNs or GANs. Achieve decent IoU but higher CD (worse) due to smoothness and less sharp features.	Stable and interpretable training; applied in robotics, compression, scene completion—well-suited for real-world. [59, 63, 64]
Generative Adversarial Networks (GANs)	Very high: Training GANs is computationally intensive because of the adversarial nature of the model; 3D GANs are resource-heavy but hierarchical training addresses the resource-intensive issue.	High memory due to dual networks; approaches like Hierarchical Amortized Training reduce demand for high-resolution volumes.	High visual realism but prone to mode collapse; blending with VAE (e.g., VAE-GAN) improves diversity.	GAN-based models yield improved visual detail and perceptual realism and can achieve low Chamfer Distance, but IoU is competitive but not consistently higher than CNNs.	Powerful results; training instability and resource demand limit deployment. Advance training strategies enable usage in medical and industrial contexts. [62–68]

4) Performance metrics for 3D reconstruction

The efficacy of 3D reconstruction method was statistically assessed utilizing various established performance metrics with State-of-Art-Analysis (STOA) as discussed below:

a) Chamfer Distance (CD)

Chamfer Distance (CD) quantifies the arithmetic mean of the nearest-neighbour distance calculations between two-point sets $\{ S_1 \subset R^3 \}$ and $\{ S_2 \subset R^3 \}$.

Here, S_1 is the Predicted point set with reference to 11,000–55,000 sampled points from reconstructed object surface and S_2 is the Ground-truth point set for the given samples. $|S_1|$ depicts the cardinality of predicted set and $|S_2|$ depicts the cardinality of ground-truth set. The 3D points of the two sets are stated as $'s' \in \{S_1\}$ and $'t' \in \{S_2\}$ respectively which implements the minimization objective function using Squared Euclidean distance measure. CD is calculated using Eq. (7).

$$CD(S_1, S_2) = \frac{1}{|S_1|}\sum_{s \in S_1} min_{(t \in S_2)}||s - t||^2 + \frac{1}{|S_2|}\sum_{t \in S_2} min_{(s \in S_1)}||t - s||^2 \quad (7)$$

Lower is the evaluated measure of $\mathrm{CD}(\mathrm{S_1}, \mathrm{S_2})$, better is the geometric fidelity with closer match between the surfaces and improved performance.

b) Intersection-over-Union (IoU)

Intersection-over-Union (IoU) measures the volume overlap metric for evaluation of voxels/meshes. Here, $V_{(pred)}$ is the Predicted voxel grid set for occupied voxels and $V_{(gt)}$ is the Ground-truth voxel grid set for occupied voxels. $|V_{(pred)} \cap V_{(gt)}|$ depicts the cardinality of voxels that are occupied in intersection set of prediction and ground-truth. $|V_{(pred)} \cup V_{(gt)}|$ depicts the cardinality of voxels that are occupied in union set of prediction and ground-truth. IoU is computed using Eq. (8).

$$IoU = \frac{|V_{(pred)} \cap V_{(gt)}|}{|V_{(pred)} \cup V_{(gt)}|} \quad (8)$$

The IoU ratio score ∈ [0, 1] where,

- Perfect reconstruction for IoU = 1.
- No overlap between S_1 as Predicted 3D reconstruction and S_2 as the Ground-truth surfaces for IoU = 0.
- Otherwise, Partial Overlap = {0 < IoU < 1}.

c) Earth Mover's Distance (EMD)

Earth Mover's Distance (EMD) is the minimum "work" to morph one point set into the other under a bijection ϕ: $S_1 \rightarrow S_2$.

d) Normal Consistency (NC)

Normal Consistency (NC) is the average cosine similarity between the normals of the Predicted surface and the Ground-truth surface that computes the smoothness and local geometric consistency. $n_i^{(pred)}$ is the normal vector at predicted surface whereas $n_i^{(gt)}$ is the normal vector at ground-truth surface. *NC* is computed using Eq. (9).

$$NC = \frac{1}{N}\sum_{i \in N} \left| n_i^{(pred)} \cdot n_i^{(gt)} \right| \quad (9)$$

Higher is the *NC* score, better is the similarity alignment amongst the predicted and true surface normals.

e) F-Score

F-Score is the harmonic mean at various distance thresholds which captures the precision and recall of reconstruction. It is the statistic measure that offers a comprehensive assessment of the completeness and precision of the reconstructed surfaces. It evaluates the accuracy of a reconstructed 3D model (predicted) against the ground-truth model.

f) Peak Signal-to-Noise Ratio (PSNR)

Peak Signal-to-Noise Ratio (PSNR) quantifies image reconstruction fidelity by comparing a Predicted image against Ground-truth image. It uses the Mean Squared Error (MSE) and the maximum possible intensity as shown in Eq. (10) while evaluating the rendered view quality (e.g., NeRF outputs).

$$PSNR = 10\log_{10}\left(\frac{MAX_I^2}{MSE}\right) \quad (10)$$

Higher PSNR is interpreted as the better reconstruction quality which is more similar to ground truth. Whereas, the low value of PSNR includes the noisy or inaccurate reconstruction.

g) Mean Squared Error (MSE)

MSE is a fundamental metric which evaluates the average squared difference between predicted 3D representation and its ground-truth for the *N* data points as shown in Eq. (11).

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(I_i^{pred} - I_i^{gt}\right)^2 \quad (11)$$

Lower MSE relates the closes of reconstructed 3D model to the ground truth. Whereas, higher MSE value fails to approximate the shape, large deviations, etc.

h) Accuracy

Accuracy computes the amount of correctly predicted elements such as voxels, points, or mesh vertices, as compared to the total number of elements.

i) Completeness

Completeness is the percentage of ground-truth points that are successfully reconstructed, or the percentage of ground-truth points that are located within a given distance threshold from any point in the reconstructed output. This is frequently used in conjunction with Accuracy. These two together make up the F-Score.

Table IV depicts the most commonly used performance metrics which evaluates the various 3D reconstruction models.

TABLE IV. COMMONLY USED PERFORMANCE METRICS

Model Type	Models	Performance Metrics
Voxel-based	Pix2Vox, 3D-R2N2	IoU, F-Score
Point Cloud-based	PCN, PointNet, PointNet++, Dynamic Graph CNNs (DGCNN)	CD, Earth Mover's Distance (EMD), F-Score
Mesh-based	Pixel2Mesh, Pixel2Mesh++, MeshCNN	CD, EMD, Normal Consistency (NC)
Multi-View Stereo (MVS)	MVSNet, DeepMVS	Accuracy, Completeness
Implicit SDF	DeepSDF	CD, EMD
Neural Rendering	Neural Radiance Fields (NeRF)	Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Root Mean Square Error (RMSE)

5) Transfer learning for 3D reconstruction

Compatible conversion computation, implicit representation synthesis from raw 3D data, and data-driven training for spatial coherence learning are some of the processes involved in 3D reconstruction. The pre-trained learning techniques can be used to train deep networks for 3D reconstruction without requiring explicit supervision. By leveraging geometric constraints or image correspondences, the training methods can learn to reconstruct 3D geometry directly from unlabelled or weakly labelled data. These DL techniques have remarkably improved image-based 3D reconstruction, permitting more accurate, scalable, and versatile approaches for producing 3D models. With the influence of deep learning, researchers are discovering various 3D reconstruction approaches in various domains such as robotics, augmented reality, virtual reality, etc [30].

3D multimedia relies on photorealistic models, but creating high-quality 3D designs is time-consuming and costly. This drives research into automated methods for generating textured 3D models from multiple viewpoints. Transfer learning and fine-tuning strategies can significantly enhance the performance of image-based 3D reconstruction. These techniques are as follows:

a) Transfer learning

Transfer learning starts with a pre-trained CNN model which is trained using an appropriate dataset, such as ImageNet. They learn to have knowledge with generic features that can be beneficial for multiple imaging tasks.

b) Feature extraction

The pre-trained CNN is utilised as a feature extractor. Remove the fully connected layers of the CNN and use result of previous convolutional layers as feature representations for input images. These features capture high-level semantic information relevant to 3D reconstruction tasks.

c) Domain adaptation

It includes the fine-tuning of pre-trained CNN on lesser dataset specific for 3D reconstruction task. This process adapts the generic features learned from the foundation sets to the target domain with 3D reconstruction while improving the model's performance on the target task.

d) Model architecture adaptation

It adjusts the architecture of pre-trained CNN to better suit requirements of the 3D reconstruction task. This may involve modifying the network's depth, width, or adding specialized layers to handle data.

B. Evolution of DL in 3D Reconstruction

During the last ten years, deep learning algorithms have advanced significantly in 3D reconstruction which has revolutionized the creation of 3D models from a variety of data sources. Starting with traditional 3D object recognition approaches, the research expedition has advanced to complex reconstruction techniques driven by deep learning that incorporate generative models, multiview learning, and point cloud processing.

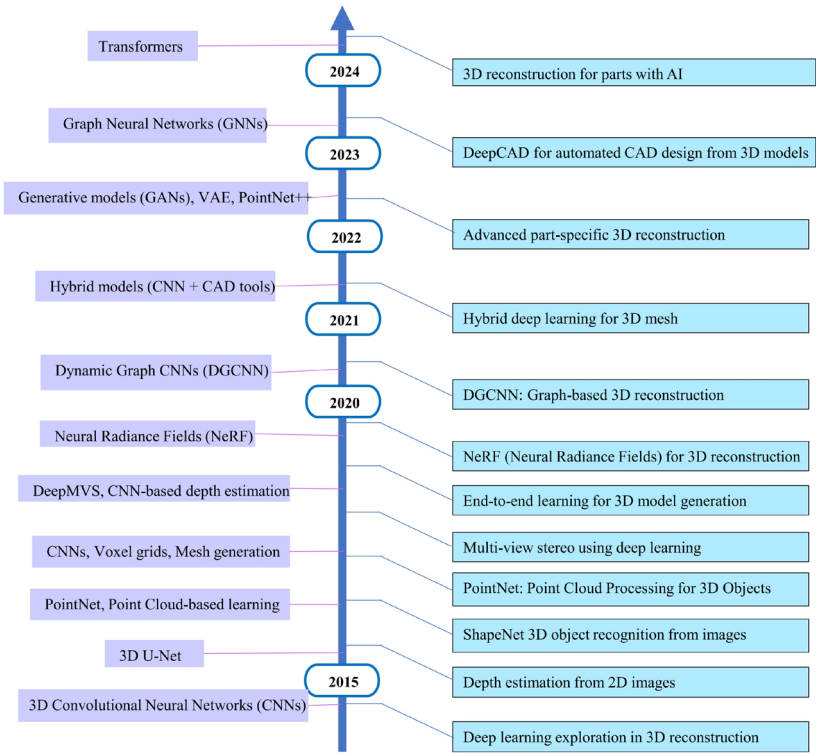


Fig. 10. Evolution of deep learning techniques and approaches for 3D reconstruction.

Early in 2015–2019, ShapeNet for 3D object recognition and PointNet, which enabled point cloud-based processing, were introduced, laying the foundation of deep learning for 3D reconstruction. These techniques made it possible to extract features from sparse 3D data points. CNNs based on voxel grids began to

appear around the same time, processing three-dimensional data [31]. With the introduction of DGCNN, the ability to model spatial relationships within point clouds has enhanced significantly. MVS methods combined with deep learning has shown significant improved in depth estimation from 2D images, which enabled generation of more accurate 3D structures. Generalizable Reconstruction (GenRe) aimed at improving single-image 3D shape reconstruction by making it more class-agnostic [30].

By 2020, with the further development in neural architectures, deep learning techniques like Neural Radiance Fields (NeRF) [69] has revolutionized the area. It offered a novel method of using implicit neural modeling to create continuous 3D representations from sparse 2D images. The incorporation of GNNs further refined 3D shape understanding, capturing geometric dependencies between points and edges. Hybrid models integrating CNNs and CAD tools allowed for DeepCAD, enabling automated design generation from 3D models. Generative models [70] such as GANs and VAEs were introduced to reconstruct missing or occluded parts of 3D objects, greatly enhancing realism in 3D reconstruction [71]. With latest advancements and focus on transformers for 3D learning, which leverage self-attention mechanism to capture long-range dependencies in 3D data. Hybrid deep learning models, integrating multiple techniques such as NeRF with GANs and GNN-enhanced depth estimation, have further pushed the boundaries of 3D reconstruction accuracy. In 2023–2024, research has moved towards part-specific 3D reconstruction, AI-assisted automated CAD modeling, and real-time 3D scene reconstruction using multi-modal data fusion. Deep learning-based 3D mesh reconstruction has evolved, improving object surface generation with minimal artifacts. Fig. 10 shows an overview of the evolution of deep learning techniques and approaches for 3D reconstruction. It traces the field's journey from early voxel-based CNNs and point-cloud encoders to Multiview stereo networks like MVSNet and hybrid point-voxel models, then onto implicit representation methods such as NeRF and DeepSDF.

IV. State-of-Art Advances and Key Studies of 3D Reconstruction Using Deep Learning

3D object reconstruction is a fundamental problem in computer vision and graphics, enabling applications in augmented reality, robotics, medical imaging, and cultural heritage preservation. The field has witnessed significant advancements over the years, with deep learning playing a crucial role in improving reconstruction accuracy and efficiency. 3D reconstruction from images has seen significant advancements due to deep learning techniques, allowing for more accurate and high-resolution reconstructions. Traditional approaches such as SfM and MVS have been supplemented by CNN, GANs, and GNNs. This survey explores key contributions in 3D reconstruction using deep learning which include approaches, techniques, methodologies, significance and limitations in this field as portrayed in Table V.

TABLE V. Summary of Key Studies on Deep Learning for 3D Reconstruction

Article	Approach	Technique	Significance/ Impact	Limitations
Image based 3D Reconstruction [5]	Surveys DL techniques for 3D approximation	-	(i) Examines GANs and VAEs in improving 3D shape synthesis. (ii) Discusses NeRFs for view synthesis and transformer-based models for better global shape understanding.	(i) Single-view reconstruction is ambiguous, lacking full 3D shape information. (ii) Multiview reconstruction depends on accurate camera poses; errors reduce quality.
3D GAN [72]	Introduced GAN-based 3D model generation	3DGAN	(i) Easy to implement. (ii) Learns realistic 3D structures without explicit supervision. (iii) Improving the sharpness and realism of generated 3D objects.	(i) Memory-intensive, limiting high-resolution shape generation. (ii) Struggles to generate fine-grained geometric details, no smooth details. (iii) Difficult to model thin structures or highly detailed surfaces.
3D Reconstruction of Industrial Parts from a Single Image [32]	Hybrid network for improved 3D object reconstruction,	CAD-ClassNet & CAD-ReconNet	(i) A dataset of 2D images of industrial parts is created. (ii) Deep Learning infers 3D shapes from a single image. (iii) Handles occlusions and missing details for better accuracy. (iv) Integrates geometric and semantic features.	(i) Single-image Computer-Aided Design (CAD) reconstruction struggles with complex geometries. (ii) Lacks parametric constraints, requiring manual edits. (iii) Occlusions and ambiguities can cause errors.
Pixel2Mesh: 3D Mesh Models [40]	Graph-based CNNs to correct 3D shape	Pixel2-Mesh	(i) Learns to predict a 3D mesh from a single image, efficient for real-time graphics. (ii) GCNs refine mesh vertices and faces, enabling better 3D shape representation. (iii) Silhouette, perceptual, and shape consistency losses improve reconstruction quality.	(i) A predefined ellipsoid limits representing complex topologies. (ii) Limited generalization to unseen or out-of-distribution objects. (iii) Self-occlusions or ambiguous views cause errors due to single-point input.
Asteroid-NeRF: 3D Surface Reconstruction [50]	Neural implicit representation for surface modeling based on NeRF	SDF with Multiview	(i) Illumination adaptation for appearance embedding. (ii) Enables continuous 3D surface reconstruction from sparse views. (iii) Better than Stereo-Photogrammetry (SPG) and Stereo-Photoclinometry (SPC). (iv) Multiview photometric consistency optimization.	(i) Availability of images under different illumination conditions. (ii) Computationally expensive due to SDF and appearance modeling. (iii) May underperform on low-texture or shadowed surfaces. (iv) Suffers from uncertainties related to surface reflectance & albedo.

DL-Based Monocular 3D Reconstruction Pipeline. [57]	Monocular RGB, U-Net++ model trained on NYU Depth V2 for depth prediction	Customized U-Net++ depth network	(i) Lightweight pipeline with depth estimation generates reliable 3D reconstruction (ii) Outperforms or matches heavier models (like GLPN) in both accuracy and speed (iii) Improving the sharpness and realism of generated 3D objects.	(i) Less consistent. (ii) Reliability challenges with top-tier models. (ii) Scope limited by training data.
Boosting MVS with Depth Foundation Models [68]	Leverages depth priors from a foundation model for generating pseudo-labels.	Pseudo-supervised MVS	(i) Enables high-quality MVS training without ground truth. (ii) Excels on DTU and Tanks & Temples. (iii) Error correction.	(i) Performance may depend on quality of pretrained depth foundation models. (ii) Training stability may vary.
NeRF for Continuous Scene Representation [69]	Implicit neural representations for continuous volumetric rendering.	NeRF	(i) Pioneered novel view synthesis and 3D scene representation. (ii) Implicit 3D representation allows infinite resolution rendering. (iii) NeRF offers a more flexible, detailed 3D scene representation. (iv) Reducing reliance on explicit modeling (v) Overcomes voxel and point cloud resolution limitations.	(i) NeRF struggles with sparse inputs, needing multiple views for accuracy. (ii) Requires known camera poses; errors degrade reconstruction quality. (iii) NeRF is limited to static scenes, failing on moving objects. (iv) NeRF lacks direct 3D surfaces; extracting meshes is computationally costly.
Shape Inpainting using 3D-GAN [73]	Uses 3D Encoder-Decoder GANs for shape completion	3D-ED-GAN, LRCN	(i) High-fidelity shape completion. (ii) Hybrid model improves shape inpainting, GANs generate realistic 3D shapes, while RCNs refine the structure. (iii) Volumetric representation allows better feature learning for generating & refining 3Dshapes.	(i) Suffer from high memory usage and low resolution, limits scalability. (ii) Struggle when large portions of a shape are missing, leading to unrealistic reconstructions and generating ambiguous or inconsistent structures.
MVSNet: Multiview Stereo Network [74]	End-to-end learning for depth inference in multiview stereo	MVSNet	(i) Deep Learning-based depth estimation outperforms traditional MVS method. (ii) Uses a 3D CNN to process cost volumes, enhancing depth estimation and generalization. (iii) Outperforms classical MVS methods such as COLMAP in certain scenarios.	(i) Uniform depth sampling may be suboptimal for varying depth scenes. (ii) Fixed depth planes lead to suboptimal depth resolution and accuracy. (iii) Lacks domain adaptability, struggles with large-scale outdoor scene w/ depth variations.
ShapeNet: Large-Scale 3D Model Dataset [75]	dataset for 3D deep learning applications	ShapeNet	(i) First large-scale 3D model dataset for DL. (ii) Enabled diverse 3D learning, advancing 3D-GANs, PointNet, NeRF, and implicit fields. (iii) ShapeNet is the benchmark for 3D generative model evaluation.	(i) ShapeNet's synthetic Computer-Aided Design (CAD) models lack real-world variations, limiting generalization. (ii) Lacks detailed textures, limits usefulness for realistic rendering. (iii) Bias in shape complexity hinders generalization.
DL for CAD Model Reconstruction [76]	Encoder–decoder architecture for robust 3D CAD model generation.	Encoder–decoder network	(i) Deep Learning automates CAD reconstruction, reducing manual work. (ii) Structured CAD models outperform raw meshes. (iii) Automatically extracts machining features.	(i) Deep Learning automates CAD but struggles with constraints and precision. (ii) Lacks parametric relations, making AI-generated CAD less editable. (iii) High-resolution, large-scale CAD remains challenging.
MonoLI: Precise Monocular 3-D Object Detection [77]	Location-aware attention mechanism and importance-aware detection head	CNN, DLA-34 (Deep Layer Aggregation)	(i) High-precision 3D detection using only a single camera. (ii) Superior performance on KITTI BEV and 3D metrics. (iii) Reweighted feature map. (iv) Lightweight Partial convolutional blocks.	(i) Limitations in diverse real-world conditions (e.g. occlusion, weather). (ii) Unknown performance on larger datasets. (ii) Complex LiDAR based detectors.
Learning a Model of Shape & Appearance from a Single View. [78]	Implicit neural surface reconstruction using SDF and differentiable rendering	SDF, Gradient-based normals	(i) Learns geometry and appearance from posed RGB images without ground-truth 3D. (ii) Delivers high-fidelity reconstruction. (iii) Foundational for future NeRF-SDF hybrids. (iv) Differentiable ray marching.	(i) Requires accurate camera positions. (ii) Scene-specific, slow ray-marching. (iii) Limited scalability.

The three-Dimensional Generative Adversarial Networks (3DGAN) are introduced which are pioneering the use of GANs for generating 3D models from 2D images [72]. Wang *et al.* [73] extended this approach with Shape Inpainting using 3D-GAN, leveraging a 3D encoder-decoder GAN to complete missing structures in partially observed 3D models. This approach effectively improved shape completion for occluded objects but struggled with high computational costs and artifacts in fine structures.

Yao *et al.* [74] proposed MVSNet, a multiview stereo approach that integrates depth inference with a deep learning framework. Unlike traditional MVS methods, MVSNet optimizes depth maps directly from multiple views, reducing errors in disparity estimation. The key challenge with MVSNet is its requirement for high computational power and limited performance in textureless or reflective surfaces. Pixel2Mesh introduced a graph-based CNN that represents 3D meshes as graphs and progressively deforms an initial ellipsoid into the target shape [40]. This approach enables efficient mesh-based reconstruction, preserving fine details compared to voxel-based methods. However, its primary drawback lies in handling topological changes, which are crucial for

reconstructing highly complex objects. The model demonstrated the ability to synthesize high-quality volumetric shapes, making it a cornerstone for subsequent developments. However, the primary limitation was the difficulty in generating fine-grained details, especially for complex objects with intricate geometries.

Han *et al.* [5] provided a state-of-the-art review of deep learning approaches for 3D reconstruction, analyzing CNN-based depth estimation, GAN-based shape synthesis, and hybrid approaches. The paper highlighted the lack of large, high-quality 3D datasets as a fundamental bottleneck, limiting generalization across diverse object categories. Voxel-based methods represent 3D objects as discrete volumetric grids, making them intuitive for deep learning models [9]. However, they often suffer from high memory consumption and computational inefficiency at higher resolutions. Xie *et al.* [79] introduced Pix2Vox, a context-aware approach to single-view and multiview 3D reconstruction, significantly improving reconstruction accuracy [8, 80]. The model consists of an encoder-decoder framework that progressively refines the 3D reconstruction. Key contributions of Pix2Vox include: (1) Hierarchical feature fusion- integrating features from different levels of the neural network for improved reconstruction. (2) Multiview consistency-handling multiple views to ensure consistency in 3D object generation. (3) Adaptive merging- reducing redundancy with high-quality voxel-based reconstructions. Applications of Pix2Vox include real-time 3D modeling, augmented reality applications, and object recognition tasks that require volumetric representation. Neural implicit representations use continuous functions to represent 3D structures, allowing for high-resolution reconstructions without memory constraints.

Mildenhall *et al.* [69] introduced a novel approach, NeRF to view synthesis by representing a scene as a fully connected neural network. Unlike traditional methods, NeRF optimizes its representation using only a sparse set of 2D images with known camera poses. Instead of discretized voxels, NeRF models scenes using a 5D function that maps spatial locations and viewing directions to color and density values. The model uses volume rendering techniques to synthesize novel views by integrating information along camera rays. NeRF does not require explicit 3D supervision, making it highly effective for scenarios where 3D ground truth data is unavailable. NeRF has found applications in high-fidelity 3D scene reconstruction, visual effects, and virtual reality, revolutionizing the way 3D models are generated from 2D images.

Dataset-driven approaches have played a pivotal role in advancing 3D reconstruction by leveraging large-scale labelled datasets for model training and evaluation. ShapeNet which is a large-scale 3D model repository that has been instrumental in training deep learning models for 3D object recognition and reconstruction. It provides the highly annotated 3D models dataset for supervised learning, that serves as a standardized benchmark 3D reconstruction. ShapeNet has facilitated numerous advancements in CAD model reconstruction that introduces a method for machining features with a deep learning-based encoder-decoder network. It generates features through parametric modelling and converts the CAD models into voxel representations for deep learning training [75]. This approach enables multiple machining features to be reconstructed efficiently, supporting industrial and mechanical design applications. In CAD models without historical data, design history reconstruction entails determining features, parameters, and their order. Although earlier methods were not very successful, new developments in deep learning provide hopeful answers along with potential research questions [81].

Xu *et al.* [32] proposed an image-based approach for reconstructing 3D industrial parts from 2D images. This study includes developing algorithms to extract 3D parameters and pose information from images, enabling reconstruction of parts such as bolts, gears, and roller bearings. The use of CAD-ClassNet and CAD-ReconNet in this study highlights the importance of deep learning in reconstructing industrial components accurately. Wang *et al.* [35] suggests a neural implicit 3D reconstruction technique that uses sparse convolutions and concentrates calculations solely on grid points close to the surface in order to increase efficiency and preserve detail. A 3D residual UNet improves robustness to noise while maintaining fine features [36]. Zheng *et al.* [82] study on a generative machine learning model for 3D reconstruction of material microstructure, used U-Net architectures and GANs to recreate material microstructure in three dimensions. GAN-based realism and efficient reconstruction of minute microstructural features are made possible by U-Net's encoder-decoder structure with multiscale feature extraction. Enhancement guarantees that artificial structures closely mimic actual microstructures. These developments highlight the expanding use of generative models in 3D reconstruction by enabling precise digital twins for production and replacement parts.

Asteroid-NeRF is a specific advancement in 3D reconstruction for planetary research, providing geometrically reliable and illumination-resistant reconstructions. Its utilization of a global SDF and visual embeddings distinguishes it from previous local or volumetric approaches such as Sparse Point Geometry (SPG) and Sparse Point Clouds (SPC) that rely on local fusion. However, its computational expense, dependency on camera postures, and lack of physically oriented lighting models are opportunities for development. The study contributes robust and generalizable NeRF applications in planetary exploration [50].

Pananthula and Sebastian [57] proposed a lightweight pipeline using monocular RGB images for 3D reconstruction, in their study used a custom U-Net++ trained on NYU Depth V2 dataset for depth prediction, followed by Open3D point-cloud generation and Poisson mesh reconstruction. The model achieves competitive accuracy with superior efficiency, making it suitable for real-time and domain-constrained problems. However, it trails Global-Local Path Networks (GLPN) in consistency

and would benefit from enhanced generalization via multi-view training and uncertainty modeling. Hong *et al.* [83] used multiview UAV pictures and the deep learning-based MVS model to rebuild the 3D model of the structures after an earthquake to help with the process of assessing the damage to the buildings. The study analysed different MVS models for 3D reconstruction of UAV images and their applicability.

DFM-MVS method leverages the depth foundation model and offers an intriguing improvement to MVS learning by substituting depth priors from large foundation models for supervision. It offers state-of-the-art findings and a workable solution to dataset-label scarcity, making it a significant advancement for extensible and label-free 3D scene reconstruction [84, 85]. The transformer model approach advances the field by enabling multiview Transformer reconstruction across hundreds of photos at once, providing significant gains in reconstruction accuracy and speed [86].

A volumetric scene is 3D depiction and representation of a physical world that divides the interior of the space into volume elements, or voxels. These voxels store data with various attributes such as vibrance, density, surface geometry, semantic labels, etc. A novel explicit volumetric rendering technique called LinPrim substitutes linear primitives, specifically tetrahedra and octahedra, for dense voxels or neural fields. Differentiable rendering is achieved by rasterization of primitives on GPU and linear interpolation of color/opacity inside each cell built on gradient, allowing the method to be trained like a NeRF. Volumetric scene includes voxel-based volumetric reconstructions using multiple calibrated camera views. Silhouette techniques are implemented for binary images, whereas voxel coloring uses the volumetric warping tag. It implements photo-consistency and visibility constraints across arbitrarily multiple views. Dense 3D reconstruction enabled without requiring sparse feature matches. Domain-warping supports large-scale or infinite scene modeling [87, 88]. The challenges posed during reconstructions are the computationally intensive operations wherein, voxel grids scale poorly. Precise camera calibration and controlled illumination are required for appropriate processing. The technique struggles in untextured regions or shadows. Implicit Differentiable Renderer (IDR) enables surface reconstruction through differentiable rendering by introducing an implicit neural representation that can concurrently describe 3D geometry and appearance from multiview images. The SDF gradients are used to compute precise surface normals for shading. Learnable appearance function models the view-dependent appearance for photorealistic rendering [77].

In recent years NeRF revolutionized image-based rendering; Table VI shows the developments in the field of NeRF.

TABLE VI. DEVELOPMENTS IN THE FIELD OF NERF

Model/Method	Methodology	Algorithmic Techniques	Significance	Challenges
Vanilla NeRF [69]	Per-scene Multi-Layer Perceptron (MLP) optimized via volume rendering from multiview calibrated images.	Positional encoding; hierarchical sampling; gradient-based optimization.	Detailed and view-dependent scene representations without explicit geometry; NeRF can be optimized directly from 2D images.	Extremely slow convergence, impractical for real-time applications; designed for static scenes and requires retraining for each new scene, limiting generalization.
Mip-NeRF [89]	Extensions of NeRF addressing aliasing, pose refinement, and speed (7% faster than NeRF).	Anti-aliasing via Gaussian frustums; multiscale positional encoding.	Scales NeRF to multiscale sampling using conical frustum representation; encodes a whole Gaussian region—preserving high-frequency details.	While Mip-NeRF improves upon NeRF, rendering scenes with high fidelity still requires significant computational resources. Applying to diverse, real-world datasets can be challenging due to variations in scene complexity.
PlenOctrees / SNeRG [90, 91]	Bake trained NeRF into fast lookup structures: octree or sparse grids.	Octree-based spherical radiance lookup (PlenOctrees); sparse voxel grids with residual MLP (SNeRG).	Enabling real-time rendering 3K× faster.	Pre-processing heavy, large memory; inflexible after baking.
FastNeRF [92]	Factorizes radiance field for extremely high-fps rendering.	Precompute deep radiance maps; directional query via lookup tables; graphics-inspired factorization.	Strong result (~3000× faster than NeRF); enables Real-Time and Interactive Applications.	Limited support for dynamic or unbounded scenes; training dataset and preprocessing overhead.
Instant-NGP [93]	Hash-based multiresolution grid encoding allowing minute-scale per-scene training.	Spatial hash encoding; tiny MLP; multires grid.	Massive Speed-Up via Compact Encoding; Instant-NGP sparked wide adoption across fields like NeRF acceleration, neural image/volume representations, signed-distance functions.	High GPU memory requirements; needs empirical hyperparameter tuning.
SSDNeRF [94]	Single-stage NeRF with effective generative modelling	Single-stage latent diffusion + NeRF fusion.	Strong performance under sparse input.	Diffusion model adds training complexity and runtime.
PixelNeRF [95]	Feed-forward encoder conditioned NeRF enabling generalization across scenes.	CNN (ResNet) extracts per-view features; conditions MLP weights or embeddings; no per-scene optimization.	Generate plausible novel view synthesis from very few input images without test-time optimization.	Lower fidelity than per-scene NeRF; blurry in occluded areas, limited to training distribution.
D-NeRF [96]	Dynamic NeRF modeling motion via time-conditioned volume rendering.	Two MLPs: canonical scene + deformation network; time as input; joint learning.	Extends NeRF to dynamic, non-Rigid scenes.	Requires dense temporal data, handles only single-object motion.

Block-NeRF / Mega-NeRF [97]	Decomposes large scenes into per-block NeRFs with appearance alignment.	Learned pose refinement; per-block appearance embedding; exposure control; scene tiling.	Enables large-scale scene reconstruction by representing the environment using multiple compacts NeRFs that each fit into memory.	Complex pipeline, heavy data and computing.
Mixer-NeRF [98]	Hybrid spatial feature module before MLP to boost detail in real multiview scenes.	MLCA feature mixing; squeeze-and-excitation module before MLP.	This method improves the learning of perceptual image block similarity by more than 30%.	Architecturally complex, limited real-scene evaluation.
MIS-NeRF [99]	Adapted to surgical endoscopy images for intraoperative 3D reconstruction.	Camera-centre input and response modeling; specular-aware loss; depth smoothing; ICP alignment.	Novel medical adaptation.	Long runtime, designed for specific anatomy; lighting/specular dependency.
Compressed Instant-NGP (CNC) [100]	Context-based compression of Instant-NGP representation for storage reduction.	Context models over hash grids; entropy modeling; hash collisions and occupancy priors.	CNC can significantly compress multi-resolution Instant-NGP-based NeRFs and achieve SOTA performance.	Focus on storage, not reconstruction speed; adds modeling complexity.
PhysicsNeRF [101]	Sparse-view 3D reconstruction depending on fundamental physical condition/states.	Depth ranking, consistency and sparsity priors integrated into NeRF.	Enables reconstruction with just 8 input views with ~21.4 dB PSNR with good generalization.	Needs carefully tuned priors, performance deteriorates on highly novel scenes.
RA-NeRF [102]	Pose refinement with parameter tunning and mapping in dynamic 3D scenes.	Joint photometric + flow-based pose filtering within NeRF.	Robust while handling the noisy or missing data; useful in real-time environments.	Complex system with real-time tasks and computations.
FA-BARF (NeRF Convergence) [103]	The convergence is speeded within the pose uncertainty.	Frequency-adapted spatial filtering replaces coarse-to-fine scheduling.	Speeds up NeRF convergence and enables robust reconstruction even with noisy camera inputs.	Performance degrades in large, wide-area scenes; but works for object-centric settings.
Snake-NeRF (Tile and Slide) [104]	Scaling NeRF to large/satellite scenes or volumetric representations.	Out-of-core tiling & sampling with GPU-memory-aware scheduling.	Combines NeRF reconstruction and generation, while effectively handles the sparse views also.	Difficult to handle visual tasks between tiles.
PC-NeRF [105]	Large-scale scene reconstruction in autonomous driving	Hierarchical Parent–Child NeRF using sparse LiDAR + camera views.	Enables NeRF on massive topographic data that demands more memory requirements.	Assumes availability of LiDAR; may not generalize to LiDAR-free settings.

Since its inception in 2020, NeRF revolutionized image-based rendering by learning a continuous 5D function, mapping spatial coordinates and viewing direction to volume density with emitted radiance, for multiview photographs [69]. With large computational complexities, NeRF exhibits limitations in aliasing and scale variation handling. Mip-NeRF addressed this issue by conical frustums instead of point-based rays while integrating the positional encoding over these regions, reducing rendering artifacts, with faster inferences [89]. Baking NeRF precomputed NeRF content into sparse neural voxel grids or light field caches with residual MLPs to enable very efficient real-time rendering [90]. PlenOctrees packed the NeRF style data into octrees with precomputed spherical radiance and interpolation, providing interactive frame rates suitable for real-time applications [91]. FastNeRF focused on ultra-fast inference by factoring the network into spatial and directional components while caching intermediate radiance maps and enabling high-fidelity rendering at over 200 FPS [92]. Instant Neural Graphics Primitives (Instant-NGP) dramatically shortened training in seconds and rendering in milliseconds by using a multiresolution hash encoding of spatial features paired with a small MLP [93]. SSDNeRF [94] is a single-stage NeRF framework that fuses latent diffusion with neural radiance fields to enable strong performance from sparse input views, although the added diffusion component increases both training complexity and runtime. PixelNeRF learned with the convolutional NeRF that prior conditioned on one or few input images, eliminating the need for per-scene optimization and generalizing to unseen scenes [95]. NeRF introduced time as an additional input and trained deformation fields to handle non-rigid temporal motion, to model dynamic content [96]. Block-NeRF, decomposed large scenes into spatial blocks with individual NeRFs, coupled with appearance alignment and pose refinement [97]. Mixer NeRF proposed a hybrid spatial-feature mixing architecture to improve 3D reconstruction efficiency by combining features across scales and spatial regions [98]. MIS-NeRF reconstructed volumetric anatomy from limited visual inputs in complex operative environments [99].

A compression-focused extension of Instant-NGP investigated the lossy encodings and pruning strategies to reduce model size while retaining reconstruction fidelity [100]. PhysicsNeRF incorporated physical illumination priors into training, enhancing the reconstruction accuracy from sparse views under unknown lighting by embedding physics-guided constraints [101]. RA-NeRF proposed robust camera pose estimation under complex motion trajectories, jointly optimizing pose and scene radiance to reconstruct challenging trajectory datasets [102]. FA-BARF replaced cyclic frequency annealing in BARF with spatial frequency adaptation, accelerating convergence and improving joint pose-scene

optimization robustness [103]. Single-stage diffusion NeRF unified generative diffusion modeling and NeRF-based reconstruction into a single-stage network for generation and view synthesis simultaneously [104]. Tile-and-slide extended NeRF representations globally through tiling and stitching mechanisms, enabling earth-scale reconstruction by adapting local NeRF blocks to global scales [105]. Lastly, Difix3D+ improved reconstruction quality by incorporating diffusion-based priors to refine geometry and view synthesis from single-step diffusion models in NVIDIA's 2025 release.

A critical investigation of NeRFs in image-based 3D reconstruction is presented by Remondino *et al.* [106]. Using a variety of criteria, including noise level, geometric accuracy, and the number of necessary images, this study impartially assesses the advantages and disadvantages of NeRFs and offers insights into their suitability for various real-world situations as well as the caliber of the resulting 3D reconstruction.

V. Summary and Discussion

The revolution in image-based 3D reconstruction is marked by 3D GAN deep learning techniques. The current research emphasizes the hybrid architectures with self-supervised training which enhances the performance of model. The industrial applications for real-time automation includes CAD designing with building information modeling. 3D GANs impact the CAD workflows in manufacturing industries. GANs with 3D deep learning generate the CAD models from images using voxels, meshes, etc. with varied geometric representations, which has its description in Table VII. 3D GAN methods, including voxel-based, point cloud, mesh-based, and implicit function GANs, enable efficient and high-quality 3D CAD reconstruction. These models, such as 3D-GAN, pointgrow, and DeepSDF, cater to different reconstruction scenarios, from basic shape generation to high-resolution modeling.

TABLE VII. 3D GAN Models and Representations Methods

Representation	Models	Advantages	Limitations
Voxel	3D-GAN, VoxGAN	Easy to process with CNNs, Structured grid	High memory usage, Low resolution
Point Cloud	PointGrow, Tree GAN	Efficient, Compact, Suitable for sparse data	No surface connectivity, Requires post-processing
Mesh	MeshGAN, GraphGAN	Explicit surfaces, Suitable for rendering	Irregular topology, Difficult optimization
Implicit Function	Occupancy Networks, DeepSDF	High-resolution details, Memory efficient	Requires optimization for sampling, No explicit geometry

Their selection depends on criteria like structured training, point-based efficiency, and rendering capability. It is expected that in future, 3D reconstruction innovation will incorporate the benefits of deep learning and classical approaches to further investigate and improve the ability to produce more effective and precise 3D reconstruction. It might be broadly employed in multimodal information fusion, modest terminology with good data synthesis, generative machine learning techniques or algorithms, and various additional contexts with robust adaptability in real-time environments.

Fig. 11 dipicts applications of the 3D GAN techniques. deep learning based 3D reconstruction has evolved with robust advancements in various research domains and its real-time applications. Still few challenges need to be addressed where the recent techniques struggle to stabilize high geometric and texture reliability at the cost of heavy computations with less scalability. Techniques such as NeRF demand large GPU memory for processing graphical data with long training sessions and yet, lack generalization across varied conditions. Dynamic scenes and moving objects pose additional challenges for most deep learning pipelines. Several approaches of deep learning implemented for 3D reconstruction, like volumetric representations, implicit data, point-based networks, mesh networks, etc. have a persistent problem since they rely heavily on large, view-diverse, annotated training sets, limiting their utility across real-time data. Also, it may be challenging to implement models such as occupancy networks or DeepSDF in real-world situations as they need substantial preprocessing and they frequently lose fine geometric detail because of pooling or global representations. 3D GANs face problems while dealing with 3D object reconstructions from limited views or noisy depth maps. Recent 3D GANs frequently produce coarse, low-resolution meshes or voxel grids, that is affected from training instability due to high-dimensional outputs.

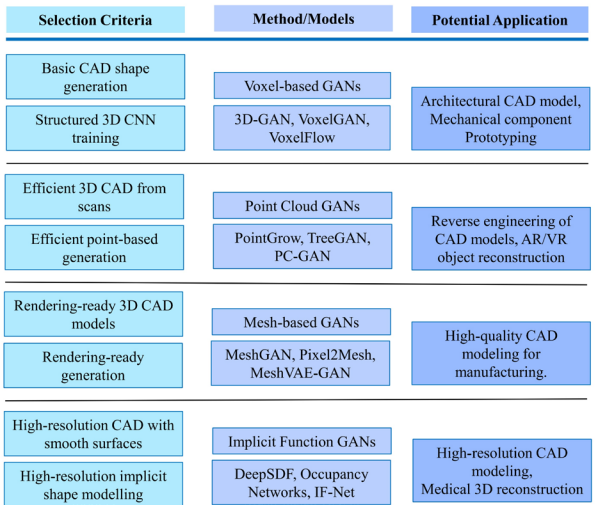


Fig. 11. Potential applications of 3D GAN techniques.

Table VIII shows the comparative analysis of performance metrics; it helps in understanding the strengths and limitations of various deep learning methods for 3D reconstruction. Voxel-based models such as Pix2Vox/Pix2Vox++ and 3D-R2N2 show improved IoU and F-Score with multi-view inputs, though they lag behind point-based methods in fine detail preservation. Mesh-based approaches like Pixel2Mesh/Pixel2Mesh++ achieve competitive F-Scores and lower CD, indicating

better geometric fidelity. Point-based architectures including PointNet, PointCNN, PointNet++, DGCNN, and PCN consistently demonstrate high IoU values, underlining their robustness in capturing local and global structures. Implicit representations such as DeepSDF achieve very low CD and EMD, signifying superior shape accuracy. Multi-view stereo methods like MVSNet balance accuracy and completeness, though performance depends heavily on input quality. Finally, volumetric radiance field models like NeRF excel in perceptual quality with high PSNR and SSIM, making them effective for photorealistic reconstruction. Overall, the trade-offs between accuracy, completeness, and perceptual quality across representations suggest complementary strengths depending on application requirements.

The future research need focus on the research-gaps identified by performing the state-of-art study which includes: (1) Efficient, scalable architectures with hybrid pipelines that integrate fast depth fusion with selective high-detail refinement; (2) Data-efficient models that generalizes with self-supervision, domain adaptation, few-shot/ meta-learning; (3) Dynamic scene modeling that handles motion and occlusion jointly with semantic constraints; (4) Unified training that includes surface extraction at resolution beyond coarse grids; and (5) Robust evaluation frameworks and datasets, especially for complex real-world domains like underwater.

TABLE VIII. PERFORMANCE METRICS

Model	Output	Dataset	Performance Metrics	Ref. No.
Pix2Vox / Pix2Vox++	Voxel	ShapeNet	IoU: ~0.670 F-Score: ~0.436 (Single view) F-Score: ~0.452–0.462 IoU: ~0.695–0.719 (Multi views)	[43]
3D-R2N2	Voxel	ShapeNet	IoU: ~0.560 F-Score: ~0.351 (Single view) IoU: ~0.603–0.636 F-Score: ~0.368–0.383 (Multi views)	[43]
Pixel2Mesh	Mesh	ShapeNet	CD: ~ 0.591 EMD: ~1.380 F-Score: 0.5972	[40]
Pixel2Mesh++	Mesh	ShapeNet	CD: ~ 0.486 F-Score: 0.6648	[52, 53]
PointNet	Point Cloud	ShapeNet S3DIS ScanNet v2	IoU: 0.837 IoU: 0.411 IoU: 0.557	[38]
PointCNN	Point Cloud	ShapeNet S3DIS ScanNet v2	IoU: 0.851 IoU: 0.572 IoU: 0.484	[38, 107]
PointNet++	Point Cloud	ShapeNet	IoU: 0.85	[107]
DGCNN	Point Cloud	ShapeNet	IoU: 0.852	[102]
PCN	Point Cloud	ShapeNet	IoU: 0.851	[108]
DeepSDF	Implicit SDF	ShapeNet	CD: ~0.006 EMD: ~0.07	[28]
MVSNet	Depth volumes	DTU	Accuracy: 0.396 Completeness: 0.527 Overall: 0.462 Accuracy: 0.375; Completeness: 0.283; Overall: 0.329 (Distance metrics)	[74, 84]
NeRF	Volumetric radiance field	Realistic Synthetic 360◦	PSNR = 31.01, SSIM = 0.947	[69]

VI. CONCLUSION

With the developing technology and expanding the possibilities of connection between the real and virtual worlds, 3D reconstruction is a highly evolving discipline for research and practical use. These developments have the power to completely transform a number of industries, including manufacturing, medicine, automation and the protection of historical assets. These advancements will raise living standards by opening up new avenues for research and creativity. There is a need to discover how technological advances are changing the world around us. This research study seeks to identify the potentials and difficulties for enhancing the use of image-based 3D building through an analysis of the current literature. The learnings of this research provide insights for various methods of reconstructing 3D models from images and highlight the great potential of reconstructing 3D models for industrial real-world applications. The neural network architectures such as CNN, GNN, autoencoders, GAN are investigated with supplementary research understandings.

Thus, the state-of-art research study depicts the most advanced state of evolution in deep learning for 3D reconstruction. 3D CAD model produces an improved cost-effective final quality product by locating and getting rid of inefficiencies. Use of advanced deep learning techniques like 3D GAN, deep SDF with different

representations have found increasing trends. As transformer models are becoming more popular for 3D reconstruction with better control over manufacturing and increased workload capacity to channelise the process; deep learning techniques have revolutionized image-based 3D reconstruction by enabling more accurate, efficient, and scalable methods for generating 3D models from 2D images. NeRF research has evolved significantly in recent 5 years, focusing on improving rendering speed, scalability, and applicability to diverse scenarios. Further exploration of dynamic scenes, more efficient representations, and integration with other computer vision tasks will be intersert for further reserch.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

ABA and RAK conducted the research survey; ABA wrote the original draft; RAK provided insights on the methodology; Finally, reviewing and editing was handled by ABA and RAK; all authors had approved the final version.

REFERENCES

[1] L. Zhou, G. Wu, Y. Zuo *et al.*, "A comprehensive review of vision-based 3D reconstruction methods," *Sensors (Basel)*, vol. 24, no. 7, 2314, 2024.

[2] Y. You, M. A. Uy, J. Han *et al*., "Img2CAD: Reverse engineering 3D CAD models from images through VLM-assisted conditional factorization," arXiv preprint, arXiv: 2408.01437, 2024.

[3] Q. C. Xu, T. J. Mu, and Y. L. Yang, "A survey of deep learning-based 3D shape generation," *Comp. Visual Media,* vol. 9, no. 3, pp. 407–442, 2023.

[4] C. B. Choy, D. Xu, J. Y. Gwak *et al*., "3D-R2N2: A unified approach for single and multi-view 3D object reconstruction," in *Proc. European Conf. on Computer Vision,* 2016, pp. 628–644.

[5] X. F. Han, H. Laga, and M. Bennamoun, "Image-based 3D object reconstruction: State-of-the-art and trends in the deep learning era," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 5, pp. 1578–1604, 2019.

[6] Y. Bai, L. H. Wong, and T. Y. Twan, "Survey on fundamental deep learning 3D reconstruction techniques," arXiv preprint, arXiv: 2407.08137, 2024.

[7] L. Ge, "Research on 3D reconstruction methods based on deep learning," *Transactions on Computer Science and Intelligent Systems Research*, vol 5, pp. 678–684, 2024.

[8] S. Qi, X. Ning, G. Yang *et al*., "Review of multi-view 3D object recognition methods based on deep learning," *Displays*, vol. 69, 102053, 2021.

[9] Y. Niu, L. Liu, F. Huang *et al*., "Overview of image-based 3D reconstruction technology," *J. Eur. Opt. Society-Rapid Publ*., vol. 20, no. 1, 18, 2024.

[10] X. Zeng, Y. Jing, Q Tang *et al*., "Multi-view 3D reconstruction based on speckle correlation," in *Proc. the VIII International Conf. on Speckle Metrology*, 2024.

[11] A. P. Pentland, "A new sense for depth of field," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-9, no. 4, pp. 523–531, 1987.

[12] T. Hwang, J. J. Clark, and A. L. Yuille, "A depth recovery algorithm using defocus information," in *Proc. CVPR '89: IEEE Computer Society Conf. on Computer Vision and Pattern Recognition,* San Diego, 1989, pp. 476–482.

[13] H. Berthold, "Obtaining shape from shading information," *Shape from Shading*, pp. 123–171, 1989.

[14] S. N. Sinha, "Multiview Stereo," in *Computer Vision: A Reference Guide*, Springer, Ikeuchi, K. (eds), Boston, 2014, pp. 516–522.

[15] R. J. Woodham, "Photometric method for determining surface orientation from multiple images," *Optical Engineering,* vol. 19, no. 1, pp. 139–144, 1980.

[16] Y. Furukawa and C. Hernández, "Multi-view stereo: A tutorial", *Foundations and Trends® in Computer Graphics and Vision*, vol. 9, no. 1–2, pp. 1–148, 2015.

[17] A. Delaunoy, E. Prados, and P. N. Belhumeur, "Towards full 3D Helmholtz stereovision algorithms," in *Proc. 10th Asian Conf. on Computer Vision*, Berlin, 2010, pp. 39–52.

[18] S. M. Seitz, B. Curless, J. Diebel *et al*., "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *Proc. 2006 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR'06),* New York, 2006, pp. 519–528.

[19] B. Lu, L. Sun, L. Yu *et al*., "An improved graph cut algorithm in stereo matching," *Displays*, vol. 69, 102052, 2021.

[20] J. Tacheles, Y. Altmann, S. McLaughlin *et al*., "3D reconstruction using single-photon LiDAR data exploiting the widths of the returns," in *Proc. 2019 IEEE International Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, 2019, pp. 7815–7819.

[21] J. Tachella, Y. Altmann, N. Mellado *et al.,* "Real-time 3D reconstruction from single-photon LiDAR data using plug-and-play point cloud denoisers," *Nat Commun.*, vol. 10, no. 1, 4984, 2019.

[22] T. Petković and T. Pribanić, "Multiprojector multicamera structured light surface scanner," *IEEE Access*, vol. 10, pp. 90321–90337, 2022.

[23] L. Li, "Time-of-flight camera–An introduction," *Technical White Paper*, 2014.

[24] T. H. Maiman, "Stimulated optical radiation in ruby," *Nature*, vol. 187, pp. 493–494, 1960.

[25] J. Butime, D. Gutierrez, L. G. Corzo *et al*., "3D reconstruction methods, A survey," in *Proc. the First International Conf. on Computer Vision Theory and Applications,* 2016, pp. 57–463.

[26] Y. Jin, D. Jiang, and M. Cai, "3D reconstruction using deep learning: A survey," *Communications in Information and Systems,* vol. 20, no. 4, pp. 389–413, 2020.

[27] R. Furferi, "Deep learning approaches for 3D model generation from 2D artworks to aid blind people with tactile exploration," *Heritage*, vol. 8, no. 1, 12, 2025.

[28] M. Comi, Y. Lin, A. Church *et al*., "TouchSDF: A DeepSDF Approach for 3D shape reconstruction using vision-based tactile sensing," *IEEE Robotics and Automation Letters*, vol. 9, no. 6. pp. 5719–5726, 2024.

[29] A. Yuniarti and N. Suciati, "A review of deep learning techniques for 3D reconstruction of 2D images," in *Proc. 12th International Conf. on Information & Communication Technology and System (ICTS)*, Surabaya, 2019, pp. 327–331.

[30] X. Zhang, Z. Zhang, C. Zhang *et al*., "Learning to reconstruct Shapes from unseen classes," *Advances in Neural Information Processing Systems*, 2018, pp. 2257–2268.

[31] Z. Wu, S. Song, A. Khosla *et al*., "3D Shapenets: A deep representation for volumetric shapes," in *Proc. the IEEE Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 1912–1920.

[32] Z. Xu, A. Wang, F. Hou *et al*., "Three-dimensional reconstruction of industrial parts from a single image," *Visual Computing for Industry, Biomedicine, and Art (VCIBA),* vol. 7, no. 1, 7, 2024.

[33] M. Deitke, R. Liu, M. Wallingford *et al*., "Objaverse-XL: A universe of 10M+ 3D objects," in *Proc. the 37th International Conf. on Neural Information Processing Systems (NIPS '23).* NY, 2023, pp. 35799–35813.

[34] K. Fu, J. Peng, Q. He *et al*, "Single image 3D object reconstruction based on deep learning: A review," *Multimedia Tools Applications*, vol. 80, no. 1, pp. 463–498, 2021.

[35] T. Wang, J. Wu, Z. Ji *et al*., "Sparse convolutional networks for surface reconstruction from noisy point clouds," in *Proc. 2024 IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV)*, Waikoloa, 2024, pp. 3200–3209.

[36] H. Remmach, R. Mouachi, M. Sadga *et al*., "Reg-PointNet++: A CNN network based on PointNet++ architecture for 3D reconstruction of 3D objects modeled by super shapes," *Journal of Image and Graphics*, vol. 11, no. 4, pp. 405–413, 2023.

[37] S. Emadi and M. Limongiello, "Optimizing 3D point cloud reconstruction through integrating deep learning and clustering models," *Electronics*, vol. 14, 399, 2025.

[38] Y. Sun, X. Zhang, and Y. Miao, "A review of point cloud segmentation for understanding 3D indoor scenes," *Vis. Intell.*, vol.

2, no. 14, 2024.
[39] L. Liu, C. Wang, C. Feng *et al*., "Incremental SFM 3D reconstruction based on deep learning," *Electronic*, vol. 13, 2850, 2025.
[40] N. Wang, Y. Zhang, Z. Li *et al*., "Pixel2Mesh: Generating 3D mesh models from single RGB images," in *Proc. the European Conf. on Computer Vision (ECCV)*, Berlin, 2018, pp. 55–71.
[41] J. Zhang, Y. Yao, and L. Quan, "Learning signed distance field for multi-view surface reconstruction," in *Proc. the IEEE/CVF International Conf. on Computer Vision*, 2021, pp. 6525–6534.
[42] H. Jang, T. Kim, K. Ahn *et al*., "Dynamic occupancy grid map with semantic information using deep learning-based BEVFusion method with camera and LiDAR fusion," *Sensors*, vol. 24, 2828, 2024.
[43] H. Xie, H. Yao, S. Zhang *et al*.," Pix2Vox++: Multi-scale context-aware 3D object reconstruction from single and multiple images," *International Journal of Computer Vision*, vol. 128, pp. 2919–2935, 2020.
[44] H. Gao, Y. Sun, J. Xiao *et al*., "Toward effective 3D object detection via multimodal fusion to automatic driving for industrial cyber-physical systems," *IEEE Transactions on Industrial Cyber-Physical Systems*, vol. 2, pp. 281–291, 2024.
[45] J. Wu, O. Wyman, Y. Tang *et al*., "Multi-view 3D reconstruction based on deep learning: A survey and comparison of methods," *Neurocomputing*, vol. 582, 2024.
[46] C. J. Si, Z. B. Yin, Z. Q. Fan *et al.,*" Point cloud completion network for 3D shapes with morphologically diverse structures," *Complex Intell. Syst*., vol. 10, no. 3, pp. 3389–3409, 2024.
[47] C. R. Qi, H. Su, K. Mo *et al*., "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proc. the IEEE Conf. on Computer Vision and Pattern Recognition*, 2017, pp. 652–660.
[48] H. Zhang, C. Wang, S. Tian *et al*., "Deep learning-based 3D point cloud classification: A systematic survey and outlook," *Displays*, vol. 79, 2023.
[49] C. Wang, M. A. Reza, V. Vats *et al*., "Deep learning-based 3D reconstruction from multiple images: A survey," *Neurocomputing*, vol. 597, 2024.
[50] S. Chen, B. Wu, H. Li *et al*., "Asteroid-NeRF: A deep learning method for 3D surface reconstruction of asteroids," *Astronomy & Astrophysics,* vol. 687, A278, 2024.
[51] L. Davies, B. Li, M. Saada *et al*., "Transformation & translation occupancy grid mapping: 2-dimensional deep learning refined SLAM," arXiv preprint, arXiv: 2504.19654, 2025.
[52] C. Wen, Y. Zhang, Z. Li *et al*., "Pixel2mesh++: Multi-view 3d mesh generation via deformation," in *Proc. the IEEE/CVF International Conf. on Computer Vision*, 2019, pp. 1042–1051.
[53] R. Chen, X. Yin, Y. Yang *et al.,* "Multi-view Pixel2Mesh++: 3D reconstruction via Pixel2Mesh with more images," *Vis Comput*., vol. 39, no. 10, pp. 5153–5166, 2023.
[54] S. Liu, Y. Zhang, S. Peng *et al*., "Dist: Rendering deep implicit signed distance function with differentiable sphere tracing," in *Proc. the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2020, pp. 2019–2028.
[55] R. Shrestha, Z. Fan, Q. Su *et al*., "MeshMVS: multi-view stereo guided mesh reconstruction," in *Proc. the International Conf. on 3D Vision (3DV)*, 2022, pp. 1290–1300.
[56] B. Ma, Y. Liu, M. Zwicker *et al*., "Inferring 3D occupancy fields through implicit reasoning on silhouette images," in *Proc. the 32nd ACM International Conf. on Multimedia (MM'24). Association for Computing Machinery*, New York, 2024, pp. 10248–10257.
[57] M. K. B. Pananthula and S. Sebastian, "3D image reconstruction from single 2D image using deep learning," *International Journal of Scientific Research in Engineering and Management*, vol. 9, no. 4, 2025.
[58] L. Mescheder, M. Oechsle, M. Niemeyer *et al*., "Occupancy networks: Learning 3D reconstruction in function space," in *Proc. the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4460–4470.
[59] S. Molnár and L. Tamás, "Variational autoencoders for 3D data processing," *Artificial Intelligence Rev*., vol. 57, no. 2, 42, 2024.
[60] D. Li, C. Lu, Z. Chen *et al*., "Graph neural networks in point clouds: A survey," *Remote Sensing*, vol. 16, no. 14, 2518, 2024.
[61] A. S. Gezawa, Y. Zhang, Q. Wang *et al*., "A review on deep learning approaches for 3D data representations in retrieval and classifications," *IEEE Access,* vol. 8, pp. 57566–57593, 2020.
[62] L. Sun, J. Chen, Y. Xu *et al*., "Hierarchical amortized GAN for 3D high resolution medical image synthesis," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 3966–3975, 2022.
[63] E. A. Ajayi, K. M. Lim, S. C. Chong *et al*., "3D shape generation via variational autoencoder with signed distance function relativistic average generative adversarial network," *Applied Sciences*., vol. 13, no. 10, 5925, 2023.
[64] M. S. Yu, T. W. Jung, D. Y. Yun *et al*., "A variational autoencoder cascade generative adversarial network for scalable 3D object generation and reconstruction," *Sensors (Basel),* vol. 24, no. 3, 751, 2024.
[65] Z. Liu, H. Tang, Y. Lin *et al*., "Point-voxel CNN for efficient 3D deep learning," in *Proc. the 33rd International Conf. on Neural Information Processing Systems*, NY, 87, 2019, pp. 965–975.
[66] S. P. Tata and S. Mishra, "3D GANs and latent space: A comprehensive survey," arXiv preprint, arXiv: 2304.03932, 2023.
[67] M. Fathallah, S. Eletriby, M. Alsabaan *et al*., "Advanced 3D face reconstruction from single 2D images using enhanced adversarial neural networks and graph neural networks," *Sensors* (Basel), vol. 24, no. 19, 6280, 2024.
[68] J. Zhu, B. Peng, Z. Zhang *et al*., "Boosting multi-view stereo with depth foundation model in the absence of real-world labels," arXiv preprint, arXiv: 2504.11845, 2025.
[69] B. Mildenhall, P. P. Srinivasan, M. Tancik *et al*., "NeRF: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM,* vol. 65, no. 1, pp. 99–106, 2020.
[70] M. Malah, R. Agaba, and F. Abbas, "Generating 3D reconstructions using generative models," in *Applications of Generative AI,* Springer, Cham, 2024, pp. 403–419.
[71] K. Berahmand, F. Daneshfar, E. S. Salehi *et al*., "Autoencoders and their applications in machine learning: A survey," *Artificial Intelligence Review*, vol. 57, no. 2, 28, 2024.
[72] J. Wu, C. Zhang, T. Xue *et al*., "Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling," in *Proc. Neural Information Processing Systems*, *Computer Vision and Pattern Recognition*, 2017.
[73] W. Wang, Q. Huang, S. You *et al*., "Shape inpainting using 3D generative adversarial network and recurrent convolutional networks," in *Proc. 2017 IEEE International Conf. on Computer Vision (ICCV),* 2017, pp. 2317–2325.
[74] Y. Yao, Z. Luo, S. Li *et al*., "MVSNet: Depth inference for unstructured multi-view stereo," in *Proc. the European Conf. on Computer Vision (ECCV),* vol. 11212, 2018, pp. 785–801.
[75] A. X. Chang, T. Funkhouser, L. Guibas *et al*., "ShapeNet: An information-rich 3D model," arXiv preprint, arXiv:1512.03012, 2015.
[76] H. Lee, J. Lee, H. Kim *et al*., "Dataset and method for deep learning-based reconstruction of 3D CAD models containing machining features for mechanical parts," *Journal of Computational Design & Engineering*, vol. 9, no. 1, pp. 114–127, 2022.
[77] H. Gao, X. Yu, Y. Xu *et al*., "MonoLI: Precise monocular 3-D object detection for next-generation consumer electronics for autonomous electric vehicles," *IEEE Transactions on Consumer Electronics,* vol. 70, no. 1, pp. 3475–3486, 2024.
[78] A. Gropp, L. Yariv, N. Haim *et al*., "Implicit geometric regularization for learning shapes," in *Proc. the 37th International Conf. on Machine Learning,* vol. 119, 2020, pp. 3789–3799.
[79] H. Xie, H. Yao, X. Sun *et al*., "Pix2Vox: Context-aware 3D reconstruction from single and multi-view images," in *Proc. IEEE/CVF International Conf. on Computer Vision (ICCV)*, Seoul, 2019, pp. 2690–2698.
[80] P. K. Vinodkumar, D. Karabulut, E. Avots *et al*., "Deep learning for 3D reconstruction, augmentation, and registration: A review paper," *Entropy*, vol. 26, no. 3, 235, 2024.
[81] B. C. Kim, "Reconstruction of design history of 3D CAD models using deep learning: research trends," *JMST Advances,* vol. 5, no. 4, pp.113–119, 2023.
[82] Y. Zheng, Z. Li, and Z. Song, "A generative machine learning model for the 3D reconstruction of material microstructure and performance evaluation," *Computer Methods in Applied Mechanics and Engineering*, vol. 430, 117224, 2024.
[83] Z. Hong, Y. Yang, J. Liu *et al*., "Enhancing 3D reconstruction model by deep learning and its application in building damage assessment after earthquake," *Appl. Sci.*, vol. 12, no. 19, 9790, 2022.

[84] X. Wang, C. Wang, B. Liu *et al*., "Multi-view stereo in the deep learning era: A comprehensive review," *Displays,* vol. 70, 102102, 2021.
[85] K. Xiong, R. Peng, Z. Zhang *et al*., "CL-MVSNet: Unsupervised multi-view stereo with dual-level contrastive learning," in *Proc. the IEEE/CVF International Conf. on Computer Vision*, 2025, pp. 3769–3780.
[86] J. Yang, A. Sax, K. J. Liang *et al*., "Fast3R: Towards 3D reconstruction of 1000+ images in one forward pass," in *Proc. the Computer Vision and Pattern Recognition Conf.*, 2025, pp. 21924–21935.
[87] C. R. Dyer, "*Volumetric Scene Reconstruction from Multiple Views*," in *Foundations of Image Understanding,* L.S. Davis, eds. MA: Springer US, vol. 628. 2001, pp. 469–489.
[88] N. Lützow von and M. Nießner, "LinPrim: Linear primitives for differentiable volumetric rendering," arXiv preprint, arXiv: 2501.16312, 2025.
[89] J. T. Barron, B. Mildenhall, M. Tancik *et al*., "Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields," in *Proc. International Conf. on Computer Vision*, Montreal, 2021, pp. 5855–5864.
[90] P. Hedman, P. P. Srinivasan, B. Mildenhall *et al*., "Baking neural radiance fields for real-time view synthesis," *IEEE Transaction Pattern Analysis Machine Intelligence,* vol. 47, no. 5, pp. 3310–3321, 2025.
[91] A. Yu, R. Li, M. Tancik *et al*., "PlenOctrees for real-time rendering of neural radiance fields," in *Proc. IEEE/CVF International Conf. on Computer Vision (ICCV)*, Montreal, 2022, pp. 5732–5741.
[92] S. J. Garbin, M. Kowalski, M. Johnson *et al*., "FastNeRF: High-fidelity neural rendering at 200FPS," in *Proc. International Conf. on Computer Vision*, Montreal, 2021, pp. 14326–14335.
[93] T. Müller, A. Evans, C. Schied *et al*., "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Transactions Graph*, vol. 41, no. 4, pp. 1–15, 2022.
[94] H. Chen, J. Gu, A. Chen *et al*., "Single-stage diffusion NeRF: A unified approach to 3D generation and reconstruction," in *Proc. the IEEE/CVF International Conf. on Computer Vision*, Paris, 2023, pp. 2416–2425.
[95] A. Yu, V. Ye, M. Tancik *et al*., "PixelNeRF: Neural radiance fields from one or few images," in *Proc. the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2021, pp. 4578–4587.
[96] A. Pumarola, E. Corona, G. Pons-Moll *et al*., "D-NeRF: Neural radiance fields for dynamic scenes," in *Proc. the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Nashville, 2021, pp. 10313–10322.
[97] M. Tancik, V. Casser, X. Yan *et al*., "Block-NeRF: Scalable large scene neural view synthesis," in *Proc. the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 8238–8248.
[98] L. Jiang, R. Che, L. Hu *et al*., "Mixer-NeRF: Research on 3D reconstruction methods of neural radiance fields based on hybrid spatial feature information," *Journal of Computing and Electronic Information Management*, vol. 15, no. 3, pp. 149–155, 2024.
[99] S. B. Khojasteh, D. Fuentes-Jimenez, D. Pizarro *et al*., "MIS-NeRF: Neural radiance fields in minimally-invasive surgery," *International Journal Computer Assisted Radiology & Surgery,* vol. 20, pp. 1481–1490, 2025.
[100] Y. Chen, Q. Wu, M. Harandi *et al*., "How far can we compress instant-NGP-based NeRF?" in *Proc. the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Seattle, 2024, pp. 20321–20330.
[101] M. Barhdadi, H. Kurban, and H. Alnuweiri, "PhysicsNeRF: Physics-guided 3D reconstruction from sparse views," arXiv preprint, arXiv: 2505.23481, 2025.
[102] Q. Yan, Q. Wang, K. Zhao *et al*., "RA-NeRF: Robust neural radiance field reconstruction with accurate camera pose estimation under complex trajectories," arXiv preprint, arXiv:2506.15242, 2025.
[103] R. Qian, C. Zhang, Y. Di *et al*., "FA-BARF: Frequency adapted bundle-adjusting neural radiance fields," arXiv preprint, arXiv: 2503.12086, 2025.
[104] C. Billouard, D. Derksen, A. Constantin *et al*., "Tile and slide: A new framework for scaling NeRF from local to global 3D earth observation," arXiv preprint, arXiv:2507.01631, 2025.
[105] J. Z. Wu, Y. Zhang, H. Turki *et al*., "Difix3D+: Improving 3D reconstructions with single-step diffusion models," in *Proc. the Computer Vision and Pattern Recognition Conf.,* 2025, pp. 26024–26035.
[106] F. Remondino, A. Karami, Z. Yan *et al*., "A critical analysis of NeRF-based 3D reconstruction," *Remote Sensing*, vol. 15, no. 14, 3585, 2023.
[107] C. Qi, L. Yi, H. Su *et al*., "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in Neural Information Processing Systems*, pp. 5099–5108, 2017.
[108] Y. Wang, Y. Sun, Z. Liu *et al*., "Dynamic graph cnn for learning on point clouds." arXiv preprint, arXiv:1801.07829, 2018.