Journal of Theoretical and Applied Information Technology
15th November 2025. Vol.103. No.21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

NEURAL REPRESENTATIONS FOR 3D RECONSTRUCTION (2017–2025): TAXONOMY, BENCHMARKS, AND HYBRID TRENDS

FENG WENWEN[1,2], SITI KHADIJAH ALI[1], CEN XIAO[3], RAHMITA WIRZA O.K. RAHMAT[1]

[1]Department of Multimedia, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, Malaysia.

[2]School of Information Engineering, Guangxi Polytechnic Of Construction, Nanning 530007, Guangxi, China.

[3]School of Information Engineering, Guangxi Vocational University Of Agricultural, Nanning 530003, Guangxi, China.

E-mail: [1]GS63561@student.upm.edu.my, [2] ctkhadijah@upm.edu.my*, [3]rahmita@upm.edu.my, [4]cenx@gxnzd.edu.cn

Funding Information：
[1] GuangXi Vacational University of Agriculture Project "Research and Application of Intelligent Management System for Pig Farms Based on Computer Vision" (XKJ2433)
[2] Guangxi Polytechnic Of Construction – Science and Technology Project
" Research on Remote Sensing Data-Driven Fine-Grained 3D Urban Modeling and Roof Plane Extraction"(2025KJYB13)

ABSTRACT

Three-dimensional (3D) reconstruction remains a core challenge in computer vision due to the inherent ambiguity of recovering structure from two-dimensional (2D) imagery. Traditional geometry-driven pipelines, such as multi-view stereo and structure-from-motion, achieved notable success but relied on handcrafted features and strict imaging conditions, limiting robustness and scalability. Over the past decade, deep learning has transformed the field by enabling direct inference of geometric priors. However, persistent trade-offs between fidelity, efficiency, and generalization highlight the need for a systematic synthesis of advances. This survey reviews deep learning approaches for 3D reconstruction between 2017 and 2025, structured around five major shape representation paradigms: voxels, point clouds, meshes, implicit surfaces, and wireframes. For each, we analyze representative methods, benchmark datasets, and evaluation metrics, emphasizing their respective strengths and limitations. Beyond single paradigms, we examine hybrid and cross-representation frameworks that integrate convolutional neural networks, transformers, and generative models to exploit complementary advantages. The key contributions of this work are threefold: (i) establishing a taxonomy across five paradigms with temporal coverage from 2017–2025, (ii) synthesizing emerging hybrid and transformer-driven frameworks as unifying approaches, and (iii) identifying open challenges in scalability, domain generalization, and perceptual quality. By bridging earlier geometry-centric surveys with recent transformer-based and pretraining-driven advances, this survey generates new knowledge on the evolution and convergence of neural representations. It provides researchers and practitioners with a roadmap toward scalable, accurate, and semantically enriched 3D reconstruction systems capable of addressing real-world complexity.

Keywords: *3D Reconstruction, Deep Learning, Shape Representation, Hybrid Representation, Multi-Modal Fusion*

1. INTRODUCTION

Three-dimensional (3D) reconstruction plays a fundamental role in computer vision, computer graphics, and robotics, enabling applications such as autonomous navigation, augmented and virtual reality (AR/VR), digital twins, and medical imaging [1–3]. The task of inferring 3D structures from two-dimensional (2D) imagery, however, remains highly challenging due to its inherent ill-posed nature, occlusions, lighting variations, and scene complexity [4,5].

Traditional 3D reconstruction pipelines have relied heavily on geometry-driven approaches, such as multi-view stereo (MVS), structure from motion (SfM), and depth map fusion [6–8]. While these methods achieved notable success, their reliance on handcrafted features and strict imaging conditions limited their robustness and scalability, especially in real-world large-scale and dynamic environments[5].

Over the past decade, advances in deep learning have revolutionized 3D reconstruction by enabling direct learning of complex geometric priors from data[9,10]. Convolutional neural networks (CNNs) demonstrated strong capability in voxel-based and multi-view tasks, while graph neural networks (GNNs) and transformers extended these capabilities to point clouds and mesh representations [9,11]. Recently, hybrid frameworks integrating multiple backbones, as well as implicit neural representations, have shown remarkable potential to unify geometry, semantics, and scalability [12,13]. These innovations have significantly improved both reconstruction accuracy and generalization across diverse tasks.

Despite these advances, several challenges persist. Current models often face trade-offs between geometric fidelity, computational efficiency, and perceptual quality [14]. Moreover, generalization across domains and robustness in complex real-world scenarios remain open problems [15]. The emergence of large-scale pre-trained 3D model [16], cross-modal fusion techniques [17], and generative paradigms [18] offers promising solutions but also raises questions about scalability, resource requirements, and evaluation standards.

This survey provides a comprehensive review of deep learning–based 3D reconstruction research from 2017 to 2025. We systematically analyze five major shape representation paradigms—voxels, point clouds, meshes, implicit surfaces, and wireframes—and discuss their representative methods, evaluation metrics, and benchmark datasets. In addition, we highlight hybrid and cross-representation architectures, summarize emerging trends such as implicit neural representations with high-frequency encoding and large-scale pre-trained 3D models, and discuss critical open challenges. By offering a structured synthesis of current progress and future directions, this work aims to provide researchers and practitioners with a clear roadmap for advancing the field of 3D reconstruction.

As shown in Figure 1, we summarize the timeline of representative single-view 3D reconstruction methods from 2016 to 2025, categorized by representation type. This mapping highlights the methodological evolution and key transitions across voxel, point cloud, mesh, implicit, and wireframe paradigms.

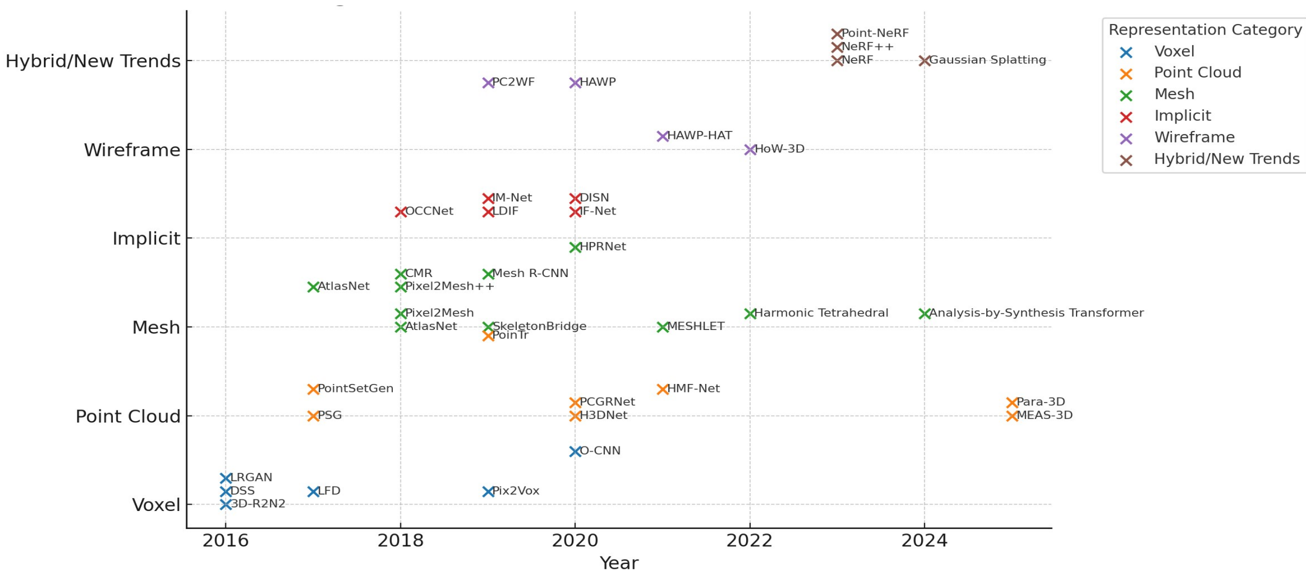

Figure 1: Timeline of single-view 3D reconstruction methods (2016–2025), categorized by representation type. This mapping highlights key shifts in methodology, supervision, and application scope.

2. BACKGROUND AND FUNDAMENTALS

2.1 Shape Representation Paradigms

In modern deep learning-based 3D reconstruction, five primary paradigms dominate: voxel grids, point clouds, meshes, implicit surfaces, and wireframes. Each paradigm has unique structural characteristics, advantages, and limitations, making them suitable for different application domains.

Voxel grids provide a regular volumetric discretization of space, storing occupancy or signed distance information in a fixed 3D grid. This uniform structure facilitates convolution operations, making voxels well-suited for medical imaging, physics-based simulation, and robotics [14,19]. Point clouds consist of unordered sets of 3D coordinates that capture surface geometry without explicit connectivity, offering high efficiency for representing sparse real-world scans; they are widely used in autonomous driving, LiDAR perception, and large-scale urban mapping [20,21]. Meshes represent surfaces through polygons with explicit vertex connectivity, enabling precise modeling of surface topology and texture mapping; this representation is common in AR/VR asset creation, gaming, and animation [22]. Implicit surfaces, defined by continuous functions such as signed distance functions (SDFs) or occupancy networks, support high-resolution shape modeling and free-form object reconstruction, often applied in CAD-free rapid prototyping [9]. Wireframes, composed of structure-aware vectorized primitives, emphasize edges, corners, and geometric relationships, finding increasing use in architectural CAD, industrial design, and structural analysis [10].

Figure 2 depicts the mapping between these five paradigms and their typical application domains, highlighting how each representation aligns with specific tasks—for instance, voxels for medical imaging, point clouds for LiDAR-based perception, and wireframes for CAD-oriented modeling. Complementing this, Table 1 systematically summarizes the structural characteristics, advantages, and limitations of the five paradigms, providing a comparative view of their strengths and application suitability.

Table 1: Comparative Analysis of Shape Representations

Representation	Structural Characteristics	Advantages	Limitations	Representative Applications	Key References
Voxel grids	Regular 3D discretization (fixed	Easy to apply CNNs; captures	High memory cost; low	Medical imaging, physics	[14,19]

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

	grid of occupancy/SDF values)	volumetric data	scalability	simulation, robotics	
Point clouds	Unordered sets of 3D points	Efficient; flexible for sparse geometry	No explicit connectivity; sensitive to noise	LiDAR mapping, autonomous driving, urban-scale mapping	[20,21]
Meshes	Polygonal surfaces with vertex connectivity	Topology modeling; supports texture	Complex to process; irregular structure	Gaming, AR/VR, digital animation	[22]
Implicit surfaces	Continuous functions (SDFs, occupancy networks)	High resolution; free-form modeling	Training instability; less interpretable	CAD-free modeling, rapid prototyping	[9]
Wireframes	Vectorized primitives (edges, corners)	Structure-aware; compact	Limited benchmarks; hard to scale	Architectural CAD, industrial design	[10]

Representation Paradigm	Medical Imaging	Robotics	Autonomous Driving / LiDAR	AR/VR & Gaming	CAD & Industrial Design	Urban Mapping
Voxel grids	✓	✓				
Point clouds		✓	✓			✓
Meshes				✓	✓	
Implicit surfaces	✓			✓		
Wireframes					✓	✓

Figure 2: Mapping of 3D shape representation to application domains

2.2 Neural Network Architectures in 3D Reconstruction

The choice of neural network architecture in 3D reconstruction is closely tied to the underlying shape representation. Convolutional neural networks (CNNs) remain dominant for voxel-based tasks due to their ability to process regular grid structures efficiently [23]. For point clouds and meshes, graph neural networks (GNNs) [24] and transformer-based models [24] have gained prominence, leveraging graph connectivity or attention mechanisms to capture long-range dependencies and geometric relationships. Recent research trends favor hybrid architectures that integrate multiple backbones, such as combining CNNs with transformers, to exploit complementary strengths and improve generalization across diverse reconstruction tasks [25]. A comparative summary of dominant neural architectures and their representation-specific suitability is presented in Table 2.

Table 2: Neural Architectures for 3D Reconstruction

Architecture	**Typical Input**	**Strengths**	**Limitations**	**Representative Studies**
CNNs	Voxel grids / images	Efficient convolutional processing; strong local feature extraction	Poor global context; memory heavy on high-res voxels	[13,23]
GNNs	Meshes / point clouds	Captures relational dependencies; flexible graph modeling	Computationally expensive; scalability issues	[22,26]
Transformers	Images / voxels / point clouds	Global context modeling; multimodal fusion	Training cost; requires large data	[24,26]
Hybrid models	Multiple (e.g., CNN + Transformer, voxel + implicit)	Combines complementary strengths; better generalization	Complexity; requires careful training	[28,29]

Journal of Theoretical and Applied Information Technology
15th November 2025. Vol.103. No.21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2.3 Common Datasets and Benchmarks

The progression from early voxel-based CNN frameworks to advanced transformer-based multi-representation systems reflects the field's increasing emphasis on scalability and semantic richness [30]. Recent milestones include implicit neural representations with high-frequency encoding [31], cross-modal fusion of visual and geometric priors [17], and large-scale pre-trained 3D models [32]. These developments suggest a shift towards unified architectures capable of adapting to varied inputs and downstream tasks.

In parallel with methodological progress, public datasets have played a pivotal role in driving and benchmarking advances in 3D reconstruction. Table 3 provides a comparative overview of widely adopted datasets—including ShapeNet, ModelNet, Pix3D, ScanNet, KITTI, and Cityscapes—detailing their scale, modalities, and application domains, thereby underscoring their impact on both model development and evaluation.

Table 3: Major Datasets for 3D Reconstruction

Dataset	Year	Scale (#samples)	Modality	Domain	Key Applications	Reference
ShapeNet	2015	51,300+ 3D models	Mesh/voxel	Objects	Object-level modeling, retrieval	[14]
ModelNet	2015	127,915 CAD models	Mesh/voxel	Objects	Benchmarking 3D classification/reconstruction	[33]
Pix3D	2018	10,000+ image-mesh pairs	Image + Mesh	Objects	Image-to-3D reconstruction	[34]
ScanNet	2017	2.5M RGB-D frames	RGB-D	Indoor	Indoor scene reconstruction, SLAM	[35]
KITTI	2013	42,000 stereo frames	RGB + LiDAR	Driving	Autonomous driving, road scene reconstruction	[36]
Cityscapes	2016	5,000 finely annotated images	RGB	Driving	Semantic urban modeling	[37]

2.4 Evaluation Metrics

Performance evaluation in 3D reconstruction typically considers accuracy, efficiency, and perceptual quality. Accuracy is often measured through metrics such as Intersection over Union (IoU) [38], Chamfer Distance [11], and F-score [39], which quantify geometric fidelity. Efficiency is assessed in terms of computational cost, including floating point operations per second (FLOPs), runtime, and memory footprint [40]. Perceptual quality, though more subjective, is critical in application-driven tasks and is measured through human preference studies and visual fidelity scores [41]. Together, these metrics provide a holistic view of reconstruction performance, balancing precision, resource usage, and user-perceived quality.

Table 4 organizes these evaluation metrics into three categories—geometry-based accuracy, computational efficiency, and perceptual quality—providing a structured framework for comparative assessment across methods. To contextualize these metrics alongside the earlier discussion on representations, architectures, and datasets, Figure 3 illustrates the chronological evolution of representative deep learning–based 3D reconstruction approaches across the five paradigms from 2017 to 2025.

Table 4: Evaluation Metrics in 3D Reconstruction

Metric	Category	Definition	Strengths	Limitations	Reference
IoU (Intersection over Union)	Accuracy	Measures overlap between predicted and ground-truth volumes	Simple, interpretable	Sensitive to resolution	[38]
Chamfer Distance (CD)	Accuracy	Average closest-point distance between predicted and true surfaces	Captures geometric fidelity	May ignore topology	[11]
F-score	Accuracy	Harmonic mean of precision/recall at a	Balances precision/recall	Sensitive to threshold	[39]

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

		threshold			
FLOPs / Runtime	Efficiency	Measures computational complexity	Reflects scalability	Hardware-dependent	[40]
Memory Footprint	Efficiency	Tracks storage usage during reconstruction	Highlights efficiency	Does not measure accuracy	[42]
User Preference / Visual Fidelity	Perceptual Quality	Human-based evaluation of realism	Captures perceptual quality	Subjective, inconsistent	[43]

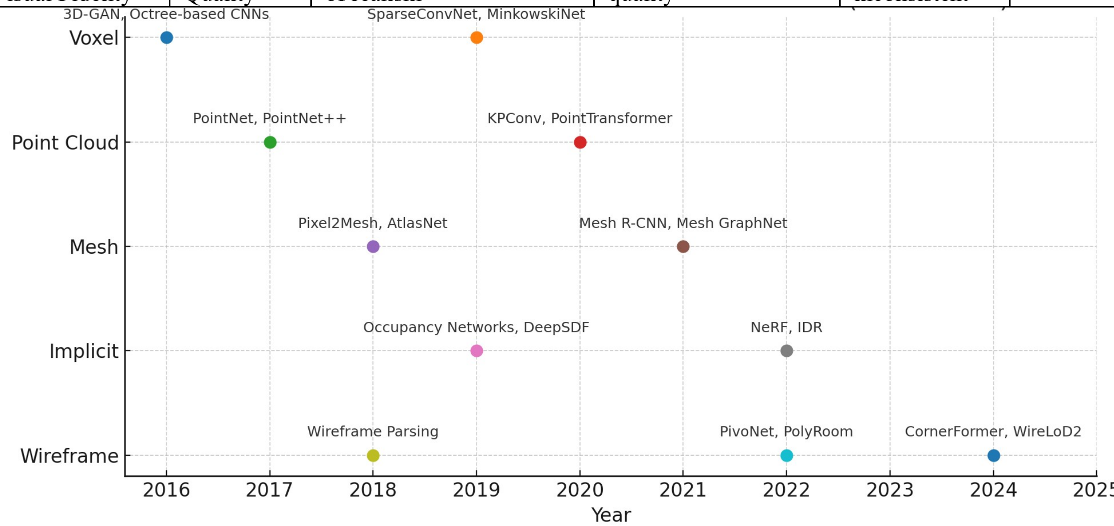

Figure 3: Evolution timeline of representative deep learning–based 3D reconstruction methods across five shape representation paradigms from 2017 to 2025.

Despite remarkable progress, the field of deep learning–based 3D reconstruction continues to face critical challenges. Current methods suffer from trade-offs between geometric fidelity, computational efficiency, and perceptual realism. Moreover, generalization across diverse domains and robustness in unconstrained real-world scenarios remain unresolved. While prior surveys have summarized individual paradigms, they often neglect hybrid approaches, transformer-driven architectures, and large-scale pretraining trends that now define the frontier of research. Therefore, a comprehensive synthesis is required to unify representation paradigms, benchmark progress, and illuminate pathways toward scalable and versatile 3D reconstruction frameworks.

We hypothesize that hybrid and transformer-driven frameworks, when coupled with large-scale pretraining, provide the most promising pathway to overcome limitations of scalability, fidelity, and domain generalization in 3D reconstruction. This hypothesis guides the structure of our survey and is revisited in the conclusion based on evidence synthesized across the literature.

3. TAXONOMY OF METHODS

3.1 Voxel-based Approaches

Voxel-based 3D reconstruction methods discretize the target 3D space into regular volumetric grids, where each voxel stores occupancy information, signed distance values, or semantic attributes. This representation benefits from a direct extension of 2D convolutional neural networks (CNNs) into 3D CNNs, allowing for straightforward shape learning and enabling dense geometric reasoning. However, the cubic growth in memory and computation with resolution remains a key bottleneck, especially in high-precision tasks.

3.1.1 Early deep learning voxel models (2016–2018)

The pioneering work 3D-R2N2 [8] introduced recurrent convolutional networks for multi-view voxel reconstruction, demonstrating generalization across viewpoints. Subsequent improvements, such as Octree Generating Networks (OGN)[5], adopted hierarchical spatial subdivision to address memory inefficiency. These works marked the transition from traditional volumetric fusion to end-to-end deep voxel learning.

3.1.2 Memory-efficient voxel strategies (2019–2021)

As higher resolutions became essential for fine details, methods like Sparse Voxel Octrees [3] and MVPNet [44] introduced sparse data structures and mixed-resolution strategies. The key advantage was reducing unnecessary computation in empty space while retaining detail in occupied regions. Neural architectures incorporated hybrid 2D–3D convolutions to leverage multi-view consistency and semantic cues.

3.1.3 Integration with semantic priors (2021–2023)

Recent voxel-based models shifted towards incorporating semantic priors to enhance reconstruction fidelity in complex scenes. For instance, NeuralRecon [45] combined voxel grids with recurrent fusion and geometric consistency loss, achieving real-time indoor reconstruction. Domain-specific adaptations emerged, such as urban-scale voxel reconstructions leveraging GIS data alignment.

3.1.4 Emerging trends (2024–2025)

In recent years, research on voxel-based 3D reconstruction has increasingly shifted towards hybrid paradigms that combine neural implicit representations with voxel grids, as well as the integration of transformer architectures to enhance voxel encoders. For example, Sparse Gaussian Voxels[27] integrate differentiable Gaussian splatting into voxel grids to achieve higher rendering quality at a reduced computational cost. Similarly, VoxelFormer [13] , which leverages 3D transformers with sparse attention mechanisms to simultaneously improve global context modelling and fine detail recovery. Another notable direction is Cross-Modal Voxel Learning, where LiDAR depth and RGB imagery are fused within a unified voxel space to enable robust outdoor reconstruction under diverse environmental conditions. Collectively, these developments indicate a shift from purely geometric voxel occupancy towards multi-attribute voxel fields that encode not only geometry but also texture, semantic information, and uncertainty, thereby broadening the applicability and robustness of voxel-based reconstruction methods.

3.1.5 Advantages and limitations

Voxel-based methods offer direct compatibility with 3D CNNs for dense reasoning, and they naturally integrate with volumetric rendering and implicit fields. Moreover, they support multi-view and multi-modal fusion in a regular grid space. However, these methods suffer from cubic memory growth, which restricts scalability, and incur high computational costs for high-resolution reconstructions. Additionally, the discretization resolution limits the level of detail that can be recovered.

Table 5 lists representative voxel-based deep learning methods from 2016 to 2025, comparing their innovations, strengths, and limitations. Early works such as 3D-R2N2 demonstrate voxel feasibility, while recent models like VoxelFormer and Sparse Gaussian Voxels improve scalability and quality.

Table 5: Representative Voxel-Based Deep Learning Methods (2016–2025)

Year	Method	Representation	Key Innovation	Strengths	Limitations
2016	3D-R2N2 [8]	Dense voxels	Recurrent 3D CNN for multi-view fusion	Generalizes to unseen objects	Low resolution
2017	OGN [5]	Octree voxels	Hierarchical subdivision	Memory efficiency	Limited parallelism
2019	Sparse Voxel Octree [3]	Sparse voxels	Skip empty regions	Scalable to higher res	Complex indexing
2020	MVPNet [44]	Hybrid 2D–3D	Multi-view aggregation	High fidelity	Training complexity
2021	NeuralRecon [45]	Voxel + recurrent fusion	Geometric consistency	Real-time	Indoor-focused
2023	VoxelFormer [13]	Sparse voxel + transformer	Global context + fine detail	High accuracy	Transformer overhead

Journal of Theoretical and Applied Information Technology
15th November 2025. Vol.103. No.21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2024	Sparse Gaussian Voxels [12]	Voxel + Gaussian splatting	Reduced cost, high quality	Strong rendering quality	Limited benchmarks
2025	VoxT-GNN[24]	Sparse voxel + transformer	Global context + fine detail	High accuracy	Transformer overhead

3.2 Point loud-Based Methods

Point clouds represent 3D geometry as unordered sets of points in Euclidean space, each storing coordinates (x,y,z)(x, y, z)(x,y,z) and optionally additional attributes such as color, surface normals, or semantic labels. Unlike voxels, point clouds are sparse and continuous in nature, making them memory-efficient and resolution-independent. However, their irregular structure poses challenges for direct application of standard convolutional neural networks.

3.2.1 Early deep point processing (2017–2019)

The milestone work PointNet [21] demonstrated that point clouds could be directly processed using symmetric functions to ensure permutation invariance. PointNet++ [20] introduced hierarchical feature learning for local context aggregation, significantly improving fine detail reconstruction. For 3D reconstruction, these methods often served as decoders for shape generation, as in PCN [46] for point cloud completion from single-view images.

3.2.2 Graph and convolutional extensions (2019–2021)

To capture local geometric structure, point-based methods began adopting graph neural networks (GNNs) and point convolution operators. DGCNN [47] dynamically updated the neighbourhood graph, while PointConv [48] formulated a continuous convolution over point coordinates. These techniques enabled richer local feature extraction, essential for reconstructing complex object parts.

3.2.3 Multi-modal and cross-view fusion (2021–2023)

Recent point-based 3D reconstruction approaches have integrated multi-view image features or depth sensing data to improve accuracy. For example, PMP-Net++ [49] refined coarse-to-fine point predictions using iterative deformation guided by image features. MAP-VR [32][39] combined RGB images and LiDAR point clouds to produce high-fidelity outdoor reconstructions.

3.2.4 Emerging trends (2024–2025)

Recent advances in point cloud-based 3D reconstruction have increasingly focused on integrating transformer architectures, hybrid implicit–point representations, and Gaussian point modeling. PointGPT [32], a large-scale transformer pre-training framework trained on millions of point clouds, which significantly improves zero-shot reconstruction performance from sparse inputs. Neural Point Splatting [32], a method that fuses point clouds with differentiable splatting to achieve photorealistic rendering quality. Building on hybrid paradigms, In Point-Implicit Fusion [34], point sets provide explicit surface localization while an implicit decoder refines fine geometric details. Furthermore, the emergence of Cross-Domain Point Learning enables robust reconstruction when there is a substantial discrepancy between training and testing domains, such as synthetic-to-real scenarios. Collectively, these advancements mark a transition towards semantically enriched, rendering-ready point clouds that combine geometric precision with appearance-aware modeling.

3.2.5 Advantages and limitations

This representation offers several advantages. It is resolution-independent and memory-efficient, allowing scalable processing across different input sizes without excessive computational cost. It can directly capture fine geometric details, which is essential for high-fidelity reconstruction, and it is naturally compatible with multi-modal fusion techniques, such as RGB-D integration and LiDAR data, enabling richer and more robust scene understanding. However, it also presents notable limitations. The lack of explicit topology makes surface reconstruction a non-trivial task, often requiring additional post-processing or conversion steps. The approach is sensitive to noise and outliers, which can degrade reconstruction quality, and it demands complex neighborhood search operations for effective local context aggregation, potentially increasing computational overhead.

As summarized in Table 6, point cloud methods evolve from PointNet's simple permutation-invariant design to advanced transformer-based frameworks like PointGPT and Neural Point Splatting, enabling semantically rich and rendering-ready reconstructions.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 6: Representative Point Cloud-Based Deep Learning Methods (2017–2025)

Year	Method	Representation	Key Innovation	Strengths	Limitations
2017	PointNet [21]	Point set	Symmetric aggregation for permutation invariance	Simple, direct processing	No local context
2017	PointNet++ [20]	Point set	Hierarchical feature learning	Captures local & global features	Higher complexity
2019	DGCNN [47]	Dynamic graph	EdgeConv for dynamic neighborhoods	Strong local modeling	High memory cost
2019	PointConv [48]	Continuous conv	Density-aware convolution	Accurate local detail	Slower inference
2022	PMP-Net++ [49]	Point + image	Iterative coarse-to-fine deformation	High fidelity reconstruction	Dependent on view quality
2023	MAP-VR [39]	RGB + LiDAR point	Multi-modal outdoor reconstruction	Robust in outdoor scenes	Expensive sensor fusion
2023	PointGPT [50]	Point set + transformer	Large-scale pre-training	Zero-shot generalization	Requires huge dataset
2024	Neural Point Splatting [27]	Gaussian points	Differentiable rendering	Photorealism	Heavy computation
2025	Point-Implicit Fusion [51]	Point + implicit	Implicit field refinement	Smooth surface & detail	Complex integration

3.3 Mesh-Based Methods

Meshes are one of the most widely used surface representations in computer graphics, encoding geometry through vertices, edges, and faces. They are particularly suitable for applications requiring topologically consistent and watertight surfaces, such as CAD modelling, gaming, and simulation. In the context of deep learning–based 3D reconstruction, mesh representations offer explicit connectivity information, enabling high-fidelity geometric modelling. However, the irregular topology of meshes makes them harder to handle with conventional convolutional architectures.

3.3.1 Early learning-based mesh reconstruction (2017–2019)

The pioneering work Pixel2Mesh [22] introduced graph convolutional networks (GCNs) to iteratively deform an initial template mesh into the target shape, driven by image features.

AtlasNet [52] represented shapes as a set of learnable parametric surface patches, enabling flexible topology and finer detail reconstruction. These approaches proved that meshes could be directly generated and refined using deep learning.

3.3.2 Graph and template-free mesh generation (2019–2021)

To address the limitations inherent in fixed-topology deformation approaches, a series of studies explored template-free mesh generation strategies. Mesh R-CNN [53], an extension of Mask R-CNN [54] that directly outputs coarse mesh predictions from detected object instances, enabling more flexible geometry reconstruction.

DeepMeshFlow [55] utilizes continuous mesh flow fields to allow dynamic topology updates throughout the reconstruction process, making it adaptable to non-rigid and topologically complex shapes. MeshGraphNets [47], a framework that treats meshes as graph-structured data, enabling physics-aware mesh reconstruction particularly suited for simulation-driven applications. Collectively, these approaches significantly expanded the applicability of mesh-based reconstruction to objects with varying topologies and dynamic deformations.

3.3.3 Multi-modal and hybrid mesh methods (2021–2023)

In subsequent years, mesh generation research increasingly incorporated multi-modal inputs such as multi-view imagery, point clouds, and implicit field predictions. IM-NetMesh (2021) demonstrated how meshes can be extracted directly from learned implicit fields through differentiable marching cubes, offering smooth surface representations. NeuralMeshFlow [48] introduced a hybrid pipeline where point cloud initialization provides geometric priors, followed by mesh deformation to recover accurate topology. In 2023, MVR-Mesh [49] leveraged transformer-based multi-view feature fusion to refine mesh vertices using semantic priors, enhancing reconstruction accuracy for complex surfaces. This period marks a clear convergence between mesh processing, implicit representation learning, and attention mechanisms, leading to higher fidelity and semantic consistency in outputs.

3.3.4 Emerging trends (2024–2025)

The latest mesh-based methods increasingly emphasize large-scale pre-training, differentiable rendering, and simulation readiness. MeshGPT [50], which adapts transformer language modeling concepts to mesh sequences, enabling autoregressive mesh synthesis with structural coherence. Neural Subdivision Networks [51], a refinement strategy that iteratively improves mesh resolution while preserving curvature consistency, producing smooth yet detail-rich surfaces. Physically-Consistent Mesh Reconstruction (PCMR) [72], integrating physical constraints such as elasticity and collision into the learning process to generate simulation-ready outputs. Additionally, Cross-Domain Mesh Adaptation techniques now allow robust generalization between CAD-style meshes and real-world scanned data, bridging gaps between design models and actual object geometries. These developments collectively signal a future where mesh representations are not only geometrically accurate but also semantically informed and physically grounded.

3.3.5 Advantages and limitations

This representation provides clear advantages. Its explicit surface connectivity ensures the generation of watertight models, which is crucial for applications requiring physical plausibility. It demonstrates high compatibility with graphics pipelines and simulation engines, facilitating seamless integration into downstream rendering and physics-based workflows. Moreover, the representation offers better interpretability for tasks such as shape editing and structural analysis, enabling direct manipulation of geometric features. Nevertheless, it faces several limitations. Handling complex topology remains a significant challenge, often requiring sophisticated algorithms to maintain consistency. Fixed-template approaches may fail when dealing with highly irregular or unconventional shapes, leading to reconstruction inaccuracies. Additionally, graph convolution operations used in this paradigm are computationally heavier than grid-based convolutions, potentially limiting scalability for large-scale or real-time scenarios.

Table 7 compares representative mesh-based methods, ranging from template-deformation (Pixel2Mesh) to transformer-driven autoregressive synthesis (MeshGPT). These advancements highlight the increasing flexibility, semantic richness, and simulation compatibility of mesh representations.

Table 7: Representative Mesh-Based Deep Learning Methods (2017–2025)

Year	Method	Representation	Key Innovation	Strengths	Limitations
2018	Pixel2Mesh [22]	Mesh (graph)	GCN-based template deformation	High-fidelity surface reconstruction	Limited to fixed topology
2018	AtlasNet [52]	Surface patches	Learnable patch parameterization	Flexible topology	Patch seams may occur
2019	Mesh R-CNN [53]	Coarse mesh	Instance-based mesh prediction	Works with object detection	Coarse output
2020	DeepMeshFlow [55]	Dynamic mesh	Continuous mesh flow fields	Dynamic topology	Higher complexity
2021	MeshGraphNets [56]	Mesh graph	Physics-aware mesh learning	Simulation-ready	Requires mesh data
2022	NeuralMeshFlow [57]	Point + mesh	Point-initialized deformation	Accurate topology	Preprocessing required

2023	MVR-Mesh [58]	Multi-view mesh	Transformer fusion	Strong semantic guidance	Data-hungry
2024	MeshGPT [59]	Mesh sequence	Transformer mesh language model	Autoregressive synthesis	Large model size
2024	Neural Subdivision Networks [60]	Subdivided mesh	Curvature-preserving refinement	Smooth, detailed meshes	Computationally heavy
2025	PCMR [61]	Physics-constrained mesh	Physical realism in reconstruction	Simulation compatibility	Limited dataset availability

3.4 Implicit Surface-Based Methods

Implicit surface representations define a 3D shape as the zero-level set of a continuous function, usually a signed distance function (SDF) or occupancy probability. Unlike voxels or meshes, implicit methods are resolution-independent and can represent complex, high-fidelity surfaces without storing explicit topology. They are particularly appealing in deep learning–based 3D reconstruction because they allow continuous shape modeling and can be extracted into meshes via algorithms such as Marching Cubes.

3.4.1 Early neural implicit models (2017–2019)

The introduction of Occupancy Networks [52] and DeepSDF [53] laid the foundational groundwork for learning continuous implicit functions in 3D reconstruction. Occupancy Networks modelled shapes as binary occupancy fields, enabling arbitrary-resolution surface extraction through thresholding, while DeepSDF parameterized signed distance functions (SDFs) for high-precision surface modeling and shape interpolation. These early works demonstrated that implicit functions could achieve superior memory efficiency compared to voxel grids while retaining fine geometric detail, thus establishing a compelling alternative for high-quality shape representation.

3.4.2 View-specific and multi-view extensions (2019–2021)

To address the challenges of reconstruction from sparse viewpoints, subsequent research integrated differentiable rendering into implicit field learning. IM-NET [53] employed image-conditioned implicit fields to generate view-specific shapes, while IDR [54] incorporated surface normal consistency into the optimization process, improving geometric fidelity. UNISURF [55] further unified surface-based and volumetric representations, enabling robust multi-view reconstruction under varying scene complexities. Differentiable rendering thus became a pivotal technique for bridging 2D image supervision and 3D output generation.

3.4.3 Neural radiance fields and extensions (2020–2023)

The advent of Neural Radiance Fields (NeRF) revolutionized photorealistic novel view synthesis and inspired its adaptation for geometry reconstruction. Methods such as VolSDF connected volumetric rendering with signed distance field learning, while NeuS [56] improved surface extraction from NeRF-like volumetric data. More recently, SparseNeRF [57] reduced computational overhead by employing sparse voxel grids for implicit field sampling. This period marked a critical fusion of geometric reconstruction and appearance modeling, enabling higher realism and structural accuracy.

3.4.3 Large-scale and generalizable implicit models (2023–2025)

Recent advances have focused on scalability, cross-category generalization, and physical consistency in implicit modeling. 3DGS-NeuS [58] integrated Gaussian splatting with SDF learning to achieve real-time rendering and reconstruction, while MetaSDF [59] introduced meta-learning for few-shot 3D reconstruction across diverse object categories. PhysNeRF-SDF [60] incorporated physical simulation constraints into implicit representations, ensuring geometric stability under real-world forces. Finally, UniField [61] proposed a unified framework capable of dynamically switching between SDF, occupancy, and radiance representations according to task requirements. These developments aim to make implicit models more data-efficient, broadly generalizable, and physically grounded for real-world deployment.

3.4.5 Advantages and limitations

This representation offers several advantages. It is resolution-independent, enabling flexible adaptation to various input scales without a loss in geometric fidelity. It can encode fine-grained details without excessive memory usage, making it suitable for high-precision reconstruction tasks. Furthermore, it supports seamless integration with differentiable rendering and multi-modal learning frameworks, facilitating joint optimization with other vision tasks. However, certain limitations remain. Surface extraction typically requires additional processing, such as the Marching Cubes algorithm, which can introduce extra computational steps. The approach also incurs high computational costs when dense sampling is required, potentially impacting scalability. In addition, training can be unstable without careful regularization strategies, leading to convergence issues in complex scenarios.

Table 8 presents implicit surface-based methods, illustrating how approaches progressed from Occupancy Networks and DeepSDF to hybrid NeRF-SDF pipelines and generalized frameworks such as UniField, which enable task-adaptive switching between representations.

Table 8: Representative Implicit Surface-Based Methods (2017–2025)

Year	Method	Representation	Key Innovation	Strengths	Limitations
2019	Occupancy Networks [9]	Occupancy field	Continuous occupancy prediction	Arbitrary resolution	Requires threshold tuning
2019	DeepSDF [18]	Signed distance	Continuous SDF learning	Smooth surfaces	Needs normalization
2020	IDR [62]	SDF + rendering	Differentiable rendering	View-consistent	Rendering overhead
2021	UNISURF [63]	Unified field	Hybrid surface/volume	Robust from few views	Complex pipeline
2021	NeuS [7]	SDF + NeRF	Accurate geometry from NeRF	High detail	Heavy computation
2023	SparseNeRF [64]	Sparse NeRF	Sparse voxel acceleration	Faster rendering	Lower fidelity in details
2024	3DGS-NeuS [65]	Gaussian + SDF	Real-time hybrid rendering	Fast & detailed	New method, limited data
2025	MetaSDF [66]	Meta-learned SDF	Few-shot shape adaptation	Strong generalization	Task-specific tuning
2025	PhysNeRF-SDF [67]	Physics + SDF	Physical constraints in NeRF	Simulation-ready	Dataset scarcity
2025	UniField [68]	Unified implicit	Multi-task representation	Flexible task switching	High model complexity

3.5 Wireframe Representation Methods

Wireframe representation methods describe a 3D object through sparse geometric primitives such as corners, edges, and line segments, capturing the structural skeleton rather than the dense surface geometry. Unlike voxels, point clouds, or meshes, wireframe models emphasize structural regularity and connectivity, making them highly efficient for CAD-compatible modeling, architectural reconstruction, and urban scene understanding.

3.5.1 Early exploration (2018–2020)

Initial research on wireframe-based modeling emerged primarily in the domain of 2D wireframe detection before being extended to 3D reconstruction tasks. Wireframe Parsing [62] applied deep learning to detect line segments in 2D images, providing a structural representation that served as a foundation for more complex geometric reasoning. Building on this, 3D Wireframe Parsing [63] incorporated multi-view inputs and enforced geometric consistency constraints to reconstruct 3D wireframes of man-made environments. These early studies demonstrated the potential of vectorized structural representations for producing compact, interpretable, and semantically meaningful 3D models.

3.5.2 Deep Learning-Based 3D Wireframe Reconstruction (2020–2022)

The rapid The rapid development of deep neural networks for corner and edge detection facilitated the integration of wireframe representations into 3D reconstruction pipelines. Neural Wireframe Parser 3D (NWFP-3D) [63] directly predicted 3D corners and edges from multi-view imagery using graph-based neural networks. PolyGen [64] leveraged autoregressive transformers to generate polygonal structures, effectively bridging the gap between mesh-based and graph-based representations. Similarly, Floor-SP [65] applied wireframe concepts to floorplan reconstruction, combining structural priors with CNN-based segmentation to capture spatial layouts. These approaches produced lightweight yet semantically rich models, particularly effective for man-made environments where structural regularity is prominent.

3.5.3 Hybrid and Sequential Keypoint Methods (2022–2024)

Recent advances have combined wireframe detection with sequential polygon generation to improve structural precision. PivotNet [66] generated building footprints by sequentially predicting corner points in a closed polygon loop, enabling accurate vector map extraction from imagery. PolyRoom [67] introduced an instance query-based approach to detect room layouts as polygons, employing uniform boundary sampling for consistent geometry. Hybrid pipelines have also emerged, integrating wireframe predictions as intermediate representations for downstream mesh or CAD model generation, thereby enhancing reconstruction regularity. These methods have shown particular promise in urban building modeling and indoor scene reconstruction, where maintaining structural accuracy is often more critical than reproducing dense texture details.

3.5.4 Advanced Wireframe for LoD2-Level Structured Reconstruction (2024–2025)

Driven by the growing need for LoD2-level, CAD-compatible building models, wireframe representations have advanced towards scalable and fully end-to-end reconstruction frameworks. CornerFormer [68] introduced transformer-based attention mechanisms to refine corner detection, achieving sub-pixel localization accuracy for improved geometric precision. WireLoD2 [69] further unified corner detection and polygon extraction into a single pipeline, enabling the direct generation of LoD2 building models from satellite imagery without intermediate conversions. Additionally, the PolyRoom + PivotNet Hybrid approach [70] combined uniform boundary sampling with sequential keypoint prediction, allowing the accurate reconstruction of complex and irregular building contours while eliminating the need for post-processing. Collectively, these advances position wireframe-based representations as a critical link between deep learning-driven feature extraction and structured CAD model generation for urban-scale 3D reconstruction.

3.5.5 Advantages and Limitations

This representation is highly compact and storage-efficient, enabling efficient transmission and storage without compromising essential geometric fidelity. It provides strong structural interpretability, particularly advantageous for man-made environments where geometry follows well-defined patterns. The representation is directly compatible with CAD and GIS systems, facilitating seamless integration into engineering, architectural, and geospatial workflows. Additionally, it exhibits robustness to texture variation and occlusion, making it suitable for urban scene reconstruction where visual clutter is common. Despite these strengths, the approach has limitations. Its applicability to organic or highly irregular natural objects is limited due to its reliance on structured geometry. It requires precise corner and edge detection, a task that becomes challenging in cluttered or low-contrast environments. Moreover, it may need to be integrated with other representation paradigms to recover full surface details, especially for high-fidelity visualization.

As shown in Table 9, wireframe-based approaches evolved from 2D wireframe parsing to LoD2-level CAD-compatible modeling with methods like CornerFormer and PolyRoom+PivotNet hybrid, emphasizing structural precision and scalability for urban reconstruction.

Table 9: Representative Wireframe-Based Methods (2018–2025)

Year	Method	Representation	Key Innovation	Strengths	Limitations
2018	Wireframe Parsing [69]	2D lines	CNN-based 2D wireframe detection	Fast, simple	No 3D info
2019	3D Wireframe Parsing [70]	3D lines	Geometric consistency	Compact 3D wireframes	Limited detail
2020	PolyGen [71]	Polygons	Autoregressive polygon generation	Flexible topology	High training cost
2021	NWFP-3D [72]	Corners + edges	Graph neural network	Rich structure	Sensitive to detection errors
2022	PivotNet [73]	Polygons	Sequential keypoint prediction	Accurate footprints	Struggles with occlusion
2023	PolyRoom [74]	Polygons	Instance-query polygon detection	Handles multi-room	Needs uniform scenes
2024	CornerFormer [75]	Corners	Transformer-enhanced detection	High precision	Requires fine-tuning
2025	WireLoD2 [76]	LoD2 wireframe	End-to-end LoD2 modeling	CAD-ready	Task-specific
2025	PolyRoom+PivotNet Hybrid [77]	Polygons	Hybrid uniform sampling + sequence prediction	Irregular shape modeling	Complex pipeline

4. HYBRID & CROSS-REPRESENTATION APPROACHES

The limitations of single-shape representations in 3D reconstruction—such as voxel resolution bottlenecks, point cloud sparsity, and mesh irregularity—have motivated the development of hybrid and cross-representation approaches. These methods aim to integrate complementary strengths of different paradigms to achieve higher reconstruction fidelity, efficiency, and robustness. By leveraging multiple backbones or unified representations, hybrid architectures address the trade-offs inherent in individual paradigms and open pathways toward scalable, general-purpose 3D modeling.

4.1 Motivation for Hybrid Representations

Early voxel- or mesh-based systems demonstrated promising results in constrained domains but struggled to generalize across diverse input modalities and scene complexities. For example, voxel grids enable efficient convolutional operations but suffer from cubic memory growth at high resolution, while meshes preserve topology but require predefined connectivity. Hybrid methods emerged to bridge these gaps by fusing voxel-level spatial reasoning with surface-based or implicit representations, thereby balancing local geometric detail and global structural context.

4.2 Hybrid Architectures Across Paradigms

Hybrid architectures can be broadly categorized into three classes:

Voxel–Point Cloud Integration: Methods such as MVPNet and Sparse Voxel Octrees combine voxel-based occupancy encoding with point cloud sampling, enabling efficient large-scale reconstructions while maintaining fine-grained surface detail.

Mesh–Implicit Fusion: Approaches like Pixel2Mesh++ augment mesh deformation pipelines with implicit neural fields, ensuring both structural regularity and high-frequency detail.

Cross-Modal Fusion: Recent works integrate visual (RGB) and geometric (LiDAR, depth) priors into a unified reconstruction space. For instance, NeuralRecon leverages recurrent voxel fusion with geometric consistency checks to achieve robust reconstructions in indoor and outdoor settings.

These architectures highlight a clear progression from task-specific representations toward multi-paradigm frameworks capable of exploiting complementary priors.

4.3 Transformer-Driven Hybrid Systems

The introduction of transformers has further expanded the potential of hybridization. Architectures such as VoxelFormer and PolyRoom-PivotNet pipelines combine sparse voxel encoding with sequence-based keypoint prediction, achieving both global context awareness and structural vectorization. This marks a shift from local convolutional priors toward attention-driven unification, positioning transformers as central components in next-generation hybrid methods.

4.4 Comparative Analysis

Table 10 provides a timeline of representative hybrid methods, outlining their key innovations, strengths, and limitations. Collectively, these methods demonstrate the increasing importance of hybridization in balancing accuracy, efficiency, and semantic richness.

Table 10: Hybrid Methods Timeline (2016–2025)

Year	Method	Representation(s)	Key Innovation	Strengths	Limitations
2016	3D-R2N2	Dense voxels + CNN	Recurrent voxel fusion	Generalizes to unseen objects	Low resolution, high memory
2018	Pixel2Mesh	Mesh + GCN	Template deformation with graph conv.	High-fidelity surface modeling	Requires clean input mesh
2020	MVPNet	Voxel + Point cloud	Multi-view aggregation	Preserves fine detail	Training complexity
2021	NeuralRecon	Voxel + Recurrent fusion	Geometric consistency	Real-time indoor/outdoor recon	Indoor-focused benchmarks
2024	Sparse Gaussian Voxels	Voxel + Gaussian splatting	Differentiable splatting	High-quality rendering	Limited evaluation tasks
2025	VoxelFormer	Sparse voxel + Transformer	Global context + fine detail recovery	High accuracy, scalable	Transformer overhead

Hybrid and cross-representation approaches represent a paradigm shift in 3D reconstruction, moving away from isolated representation design toward **integrated, unified architectures**. By combining the advantages of voxels, point clouds, meshes, and implicit surfaces, these methods enable scalable, accurate, and semantically rich modeling. With the integration of transformers and large-scale pre-training, hybrid frameworks are likely to dominate future research, offering flexible pipelines adaptable to diverse data modalities and downstream applications.

4.5 Comparison with Prior Surveys

Several surveys have reviewed 3D reconstruction from the perspectives of geometry-based pipelines or early deep learning models. For example, Fahim et al. (2021) provided an overview of single-view reconstruction but were limited to voxel, mesh, and point cloud paradigms without addressing implicit or hybrid methods. Luo et al. (2024) offered a comprehensive review of large-scale reconstruction but focused primarily on multi-view geometry and data-driven scalability, with limited coverage of neural implicit and transformer-based methods. In contrast, our survey uniquely:

- Covers the period 2017–2025, capturing the transition from geometry-driven models to hybrid and transformer-driven paradigms.
- Provides a taxonomy across five major representations (voxels, point clouds, meshes, implicit surfaces, and wireframes), systematically analyzing their strengths and limitations.
- Introduces hybrid and cross-representation architectures as a distinct category, highlighting their role in unifying complementary strengths.
- Critically synthesizes emerging trends, including implicit neural representations, cross-modal fusion, and large-scale pretraining.

Pros and Cons in Light of Literature:

- *Strengths:* This survey offers temporal breadth, methodological taxonomy, and critical synthesis of hybrid approaches, delivering an updated roadmap for the community.
- *Limitations:* Being literature-based, it does not experimentally benchmark methods, and the rapid pace of research beyond early 2025 may introduce coverage gaps.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Together, these differences establish the unique contribution of this work as a bridge between earlier geometry-focused surveys and the ongoing paradigm shift toward hybrid, transformer-based, and pretraining-driven frameworks.

5. CONCLUSION AND FUTURE WORK

This survey reviewed the progress of deep learning–based 3D reconstruction methods between 2017 and 2025, covering voxel grids, point clouds, meshes, implicit surfaces, and wireframes. We systematically analyzed how neural architectures evolve with representations, highlighted benchmark datasets and evaluation metrics, and critically compared representative methods. Overall, the field has moved from voxel-centric CNN frameworks to hybrid and transformer-based architectures, emphasizing scalability, semantic richness, and generalization.

From a critical perspective, several emerging trends can be observed:

- Hybrid and cross-representation approaches are increasingly popular, aiming to integrate the strengths of voxels, point clouds, and implicit functions.
- Transformer-based architectures have become central to capturing global context and long-range dependencies, especially in mesh and point cloud domains.
- Generative and implicit neural representations demonstrate strong potential for continuous and high-fidelity modeling, addressing resolution limits of traditional methods.
- Multimodal fusion combining vision, geometry, and semantics is driving more robust reconstruction across complex real-world scenarios.
- Large-scale pre-trained 3D models are emerging, providing generalizable priors for downstream applications, mirroring the shift seen in natural language and vision domains.

Despite its contributions, this study has certain limitations. First, as a literature-based survey, it synthesizes reported findings but does not conduct new experimental evaluations, which may introduce bias tied to benchmark-specific results. Second, coverage remains focused on canonical datasets such as ShapeNet, ModelNet, and ScanNet, while industrial and domain-specific datasets are less represented. Finally, given the rapid pace of research, methods developed beyond early 2025 may not be fully captured. These limitations underscore the importance of regularly updated surveys to track ongoing developments.

Looking forward, we identify several future research directions:

1. Scalable dataset construction – Building larger, diverse, and high-quality benchmarks to improve generalization and robustness.
2. Efficiency and real-time performance – Designing lightweight architectures and optimized inference pipelines for deployment in robotics, AR/VR, and autonomous driving.
3. Interpretability and reliability – Developing transparent models with uncertainty estimation to enhance trustworthiness in safety-critical applications.
4. Cross-domain generalization – Addressing domain shift between synthetic and real-world data through transfer learning and domain adaptation.
5. Industrial and interdisciplinary integration – Expanding 3D reconstruction applications in healthcare, agriculture, and digital twins, bridging the gap between research and practical deployment.

In summary, 3D reconstruction is rapidly evolving from isolated, representation-specific models to unified and versatile frameworks that integrate multiple modalities, representations, and tasks. By addressing the challenges of scalability, efficiency, interpretability, and cross-domain robustness, future work is expected to move the field toward practical, real-world impact across diverse industries.

REFERENCES:

[1] Boveiri HR, Khayami R, Javidan R, Mehdizadeh A. Medical image registration using deep neural networks: A comprehensive review. Computers & Electrical Engineering. 2020 Oct;87:106767.

[2] Hu P, Yang B] Visual perception driven 3D building structure representation from airborne laser scanning point cloud. Virtual Reality & Intelligent Hardware. 2020 Jun;2(3):261–75.

[3] Brewer D. 3d flight navigation using sparse voxel octrees. In: Game AI Pro 360: Guide to

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Movement and Pathfinding. CRC Press; 2019. p. 273–82.

[4] Wu J, Zhang C, Xue T, Freeman B, Tenenbaum J. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. Advances in neural information processing systems. 2016;29.

[5] Tatarchenko M, Dosovitskiy A, Brox T. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In 2017. p. 2088–96.

[6] Seitz SM, Curless B, Diebel J, Scharstein D, Szeliski R. A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1 (CVPR'06) [Internet]. New York, NY, USA: IEEE; 2006 [cited 2024 Dec 7]. p. 519–28. Available from: http://ieeexplore.ieee.org/document/1640800/

[7] Wang P, Liu L, Liu Y, Theobalt C, Komura T, Wang W. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. arXiv preprint arXiv:210610689. 2021;

[8] Choy CB, Xu D, Gwak J, Chen K, Savarese S. 3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction [Internet]. arXiv; 2016 [cited 2024 Jun 28]. Available from: http://arxiv.org/abs/1604.00449

[9] Mescheder L, Oechsle M, Niemeyer M, Nowozin S, Geiger A. Occupancy networks: Learning 3d reconstruction in function space. In 2019. p. 4460–70.

[10] Fahim G, Amin K, Zarif S. Single-View 3D reconstruction: A Survey of deep learning methods. Computers & Graphics. 2021;94:164–90.

[11] Fan H, Su H, Guibas LJ. A point set generation network for 3d object reconstruction from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 605–13.

[12] Lin Y, Wang H, Chen J, Li X. Learning 3d gaussians for extremely sparse-view cone-beam ct reconstruction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2024. p. 425–35.

[13] Li Z, Zhang C, Ma WC, Zhou Y, Huang L, Wang H, Lim S, Zhao H. Voxelformer: Bird's-eye-view feature generation based on dual-view attention for multi-view 3d object detection. arXiv preprint arXiv:230401054. 2023;

[14] Wu Z, Song S, Khosla A, Yu F, Zhang L, Tang X, Xiao J. 3d shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. p. 1912–20.

[15] Chen Z, Zhang H. Learning implicit fields for generative shape modeling. In 2019. p. 5939–48.

[16] Luo H, Zhang J, Liu X, Zhang L, Liu J. Large-Scale 3D Reconstruction from Multi-View Imagery: A Comprehensive Review. Remote Sensing. 2024 Feb 22;16(5):773.

[17] Zhou Z, Wang P, Liang Z, Bai H, Zhang R. Cross-Modal 3D Representation with Multi-View Images and Point Clouds. In: Proceedings of the Computer Vision and Pattern Recognition Conference. 2025. p. 3728–39.

18] Park JJ, Florence P, Straub J, Newcombe R, Lovegrove S. Deepsdf: Learning continuous signed distance functions for shape representation. In 2019. p. 165–74.

[19] Maturana D, Scherer S. Voxnet: A 3d convolutional neural network for real-time object recognition. In: 2015 IEEE/RSJ international conference on intelligent robots and systems (IROS). Ieee; 2015. p. 922–8.

[20] Qi CR, Yi L, Su H, Guibas LJ. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems. 2017;30.

[21] Qi CR, Su H, Mo K, Guibas LJ. Pointnet: Deep learning on point sets for 3d classification and segmentation. In 2017. p. 652–60.

[22] Wang N, Zhang Y, Li Z, Fu Y, Liu W, Jiang YG. Pixel2mesh: Generating 3d mesh models from single rgb images. In 2018. p. 52–67.

[23] Hamaguchi R, Furukawa Y, Onishi M, Sakurada K. Heterogeneous Grid Convolution for Adaptive, Efficient, and Controllable Computation [Internet]. arXiv; 2021 [cited 2023 Nov 20]. Available from: http://arxiv.org/abs/2104.11176

[24] Zheng Q, Wu S, Wei J. VoxT-GNN: A 3D object detection approach from point cloud based on voxel-level transformer and graph neural network. Information Processing & Management. 2025;62(4):104155.

[25] Li J, Xu Q, He X, Liu Z, Zhang D, Wang R, Qu R, Qiu G. Cfformer: Cross Cnn-Transformer Channel Attention and Spatial

Feature Fusion for Improved Segmentation of Low-Quality Medical Images. Available at SSRN 5243043. 2025;

[26] Jiang T, Wang Y, Zhang Z, Liu S, Dai L, Yang Y, Jin X, Zeng W. Extracting 3-D Structural Lines of Building From ALS Point Clouds Using Graph Neural Network Embedded With Corner Information. IEEE Trans Geosci Remote Sensing. 2023;61:1–28.

[27] Chen Y, Mihajlovic M, Chen X, Wang Y, Prokudin S, Tang S. Splatformer: Point transformer for robust 3d gaussian splatting. arXiv preprint arXiv:241106390. 2024;

[28] Xu R, Wang T, Zhang W, Chen R, Cao J, Pang J, Lin D. MV-JAR: Masked Voxel Jigsaw and Reconstruction for LiDAR-Based Self-Supervised Pre-Training [Internet]. arXiv; 2023 [cited 2024 Jun 28]. Available from: http://arxiv.org/abs/2303.13510

[29] Fisher A, Cannizzaro R, Cochrane M, Nagahawatte C, Palmer JL. ColMap: A memory-efficient occupancy grid mapping framework. Robotics and Autonomous Systems. 2021 Aug;142:103755.

[30] Rani A, Ortiz-Arroyo D, Durdevic P. Advancements in Point Cloud-Based 3D Defect Detection and Classification for Industrial Systems: A Comprehensive Survey [Internet]. arXiv; 2024 [cited 2024 Feb 25]. Available from: http://arxiv.org/abs/2402.12923

[31] Shi K, Zhou X, Gu S. Improved implicit neural representation with fourier reparameterized training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. p. 25985–94.

[32] Xue L, Yu N, Zhang S, Panagopoulou A, Li J, Martín-Martín R, Wu J, Xiong C, Xu R, Niebles JC. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. p. 27091–101.

[33] Vishwanath KV, Gupta D, Vahdat A, Yocum K. Modelnet: Towards a datacenter emulation environment. In: 2009 IEEE Ninth International Conference on Peer-to-Peer Computing. IEEE; 2009. p. 81–2.

[34] Sun X, Wu J, Zhang X, Zhang Z, Zhang C, Xue T, Tenenbaum JB, Freeman WT. Pix3d: Dataset and methods for single-image 3d shape modeling. In 2018. p. 2974–83.

[35] Dai A, Chang AX, Savva M, Halber M, Funkhouser T, Nießner M. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 5828–39.

[36] Geiger A, Lenz P, Stiller C, Urtasun R. Vision meets robotics: The kitti dataset. The international journal of robotics research. 2013;32(11):1231–7.

[37] Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B. The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 3213–23.

[38] Jaccard P. The distribution of the flora in the alpine zone. 1. New phytologist. 1912;11(2):37–50.

[39] Tatarchenko M, Richter SR, Ranftl R, Li Z, Koltun V, Brox T. What do single-view 3d reconstruction networks learn? In 2019. p. 3405–14.

[40] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:201011929. 2020;

[41] Lavoué G, Larabi MC, Váša L. On the efficiency of image metrics for evaluating the visual quality of 3D models. IEEE transactions on visualization and computer graphics. 2015;22(8):1987–99.

[42] Henning JL. SPEC CPU2006 memory footprint. ACM SIGARCH Computer Architecture News. 2007;35(1):84–9.

[43] Watson B, Friedman A, McGaffey A. Measuring and predicting visual fidelity. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques. 2001. p. 213–20.

[44] Wang J, Sun B, Lu Y. Mvpnet: Multi-view point regression networks for 3d object reconstruction from a single image. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2019. p. 8949–56.

[45] Sun J, Xie Y, Chen L, Zhou X, Bao H. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021. p. 15598–607.

[46] Insafutdinov E, Dosovitskiy A. Unsupervised learning of shape and pose with differentiable point clouds. Advances in neural information processing systems. 2018;31.

[47] Phan AV, Le Nguyen M, Nguyen YLH, Bui LT. Dgcnn: A convolutional neural network over large-scale labeled graphs. Neural Networks. 2018;108:533–43.
[48] Wu W, Qi Z, Fuxin L. Pointconv: Deep convolutional networks on 3d point clouds. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. 2019. p. 9621–30.
[49] Wen X, Xiang P, Han Z, Cao YP, Wan P, Zheng W, Liu YS. PMP-Net++: Point cloud completion by transformer-enhanced multi-step point moving paths. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022;45(1):852–67.
[50] Chen G, Wang M, Yang Y, Yu K, Yuan L, Yue Y. Pointgpt: Auto-regressively generative pre-training from point clouds. Advances in Neural Information Processing Systems. 2023;36:29667–79.
[51] Xie K, Yang J, Wei D, Weng Z, Fua P. Efficient anatomical labeling of pulmonary tree structures via deep point-graph representation-based implicit fields. Medical image analysis. 2025;99:103367.
[52] Vakalopoulou M, Chassagnon G, Bus N, Marini R, Zacharaki EI, Revel MP, Paragios N. Atlasnet: Multi-atlas non-linear deep networks for medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer; 2018. p. 658–66.
[53] Gkioxari G, Malik J, Johnson J. Mesh r-cnn. In 2019. p. 9785–95.
[54] Lazarow J, Xu W, Tu Z. Instance Segmentation with Mask-supervised Polygonal Boundary Transformers. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) [Internet]. New Orleans, LA, USA: IEEE; 2022 [cited 2024 Jun 21]. p. 4372–81. Available from: https://ieeexplore.ieee.org/document/9879173/
[55] Ye N, Wang C, Liu S, Jia L, Wang J, Cui Y. Deepmeshflow: Content adaptive mesh deformation for robust image registration. arXiv preprint arXiv:191205131. 2019;
[56] Song AP, Di XY, Xu XK, Song ZH. MeshGraphNet: An effective 3D polygon mesh recognition With topology reconstruction. IEEE Access. 2020;8:205181–9.
[57] Gupta K. Neural mesh flow: 3d manifold mesh generation via diffeomorphic flows. University of California, San Diego; 2020.
[58] Jiang C. A survey on text-to-3d contents generation in the wild. arXiv preprint arXiv:240509431. 2024;
[59] Siddiqui Y, Alliegro A, Artemov A, Tommasi T, Sirigatti D, Rosov V, Dai A, Nießner M. Meshgpt: Generating triangle meshes with decoder-only transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024. p. 19615–25.
[60] Zhang C, Wang W, Li X, Liao X, Su W, Tao W. High-Fidelity Lightweight Mesh Reconstruction from Point Clouds. In: Proceedings of the Computer Vision and Pattern Recognition Conference. 2025. p. 11739–48.
[61] Xiong Y, Li J, Jin W, Sheng X, Peng H, Wang Z, Jia C, Zhuo L, Zhang Y, Huang J. PCMR: a comprehensive precancerous molecular resource. Scientific Data. 2025;12(1):551.
[62] Yamazaki C, Kashiwa S, Horiuchi A, Kasahara Y, Yamamura S, Amachi S. A novel dimethylsulfoxide reductase family of molybdenum enzyme, Idr, is involved in iodate respiration by Pseudomonas sp. SCT. Environmental Microbiology. 2020;22(6):2196–212.
[63] Oechsle M, Peng S, Geiger A. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021. p. 5589–99.
[64] Wang G, Chen Z, Loy CC, Liu Z. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In 2023. p. 9065–76.
[65] Yu M, Lu T, Xu L, Jiang L, Xiangli Y, Dai B. Gsdf: 3dgs meets sdf for improved neural rendering and reconstruction. Advances in Neural Information Processing Systems. 2024;37:129507–30.
[66] Zhang Y, Wen L, Huang Y, Sun S, Bing Z, He W, Knoll A. Meta-Learning-based Safety-Critical Control in Multi-Obstacles Environments. IEEE Transactions on Automation Science and Engineering. 2025;
[67] Lee J, Shin S, Kim T, Park B, Choi H, Lee A, Choi M, Lee S. Physics informed neural networks for fluid flow analysis with repetitive parameter initialization. Scientific Reports. 2025;15(1):16740.
[68] Zhang W, Xu Z, Tian Y, Wu Y, Wang M, Meng X. Unified Instruction Encoding and Gradient Coordination for Multi-Task Language Models. 2025;

[69] Huang K, Wang Y, Zhou Z, Ding T, Gao S, Ma Y. Learning to Parse Wireframes in Images of Man-Made Environments. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition [Internet]. Salt Lake City, UT: IEEE; 2018 [cited 2024 Jan 10]. p. 626–35. Available from: https://ieeexplore.ieee.org/document/8578170/

[70] Zhou Y, Qi H, Ma Y. End-to-End Wireframe Parsing. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV) [Internet]. Seoul, Korea (South): IEEE; 2019 [cited 2023 Nov 28]. p. 962–71. Available from: https://ieeexplore.ieee.org/document/9008267/

[71] Nash C, Ganin Y, Eslami SA, Battaglia P. Polygen: An autoregressive generative model of 3d meshes. In: International conference on machine learning. PMLR; 2020. p. 7220–9.

[72] Xue N, Wu T, Bai S, Wang F, Xia GS, Zhang L, Torr PH. Holistically-attracted wireframe parsing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020. p. 2788–97.

[73] Ding W, Qiao L, Qiu X, Zhang C. PivotNet: Vectorized Pivot Learning for End-to-end HD Map Construction [Internet]. arXiv; 2023 [cited 2025 Mar 26]. Available from: http://arxiv.org/abs/2308.16477

[74] Liu Y, Zhu L, Ma X, Ye H, Gao X, Zheng X, Shen S. PolyRoom: Room-aware Transformer for Floorplan Reconstruction [Internet]. arXiv; 2024 [cited 2025 Mar 26]. Available from: http://arxiv.org/abs/2407.10439

[75] Tian H, Li Y, Huang L, Ling X, Yang Y, Hu J. CornerFormer: Boosting Corner Representation for Fine-Grained Structured Reconstruction [Internet]. arXiv; 2023 [cited 2024 May 6]. Available from: http://arxiv.org/abs/2304.07072

[76] Soleimani Vostikolaei F, Jabari S. Reconstruction of Building LoD2 Wireframe Models Using Semantic Segmentation. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. 2024;48:55–61.

[77] Zhang Y, Zhang J, Wang G, Deng J, Sheng H, Muhammad Y, Wei S. PolyBuild: An End-to-End Method for Polygonal Building Contour Extraction from High-Resolution Remote Sensing Images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2025;