A Cookbook of 3D Vision: Data, Learning Paradigms, and Application

Anonymous CVPR Workshop submission

Paper ID 7

Abstract

3D vision has rapidly evolved, driven by increasingly diverse data representations, learning paradigms, and modeling strategies. Yet the field remains fragmented across representations and benchmarks, making it difficult to develop unified perspectives on efficiency, fidelity, and scalability. This work provides a data-centric taxonomy of 3D vision that connects geometric representations, datasets, learning frameworks, and applications within a single conceptual map. We begin by surveying the principal structural representations of 3D data—point clouds, meshes, voxels, and 3D Gaussians—along with their acquisition pipelines. We then examine how dataset design, benchmark construction, and supervision regimes shape recent advances, spanning 2D-supervised 3D learning, implicit neural representations, and 4D world modeling. Through this integrative lens, we clarify the relationships among representations, learning paradigms, and downstream tasks in reconstruction, generation, and video modeling, offering a consolidated view of emerging trends toward balancing efficiency and fidelity and toward multimodal geometric grounding.

1. Introduction

3D vision has emerged as a central pillar in modern computer vision, with widespread applications in autonomous navigation [102], robotic manipulation [95], augmented reality [61, 73], and digital reconstruction [20, 68]. As sensor technologies advance and computing resources scale, from commodity RGB-D cameras and large-scale LiDAR capture to real-time neural rendering systems, 3D perception is becoming increasingly practical and ubiquitous [3, 71, 174].

Unlike 2D vision, the field of 3D vision is fundamentally more complex, both in its data structures and in its learning pipelines [49, 78, 93]. It spans a wide range of data representations, including point clouds, meshes, voxel grids, RGB-D images, multi-view images, CAD models, neural implicit fields, and 3D Gaussians, each with its own structural assumptions, learning pipelines, and computational trade-offs [68, 71, 99, 108, 129, 143, 149]. At the same time, downstream tasks range from reconstruction and segmentation to pose estimation and scene generation [20, 45, 70, 83, 107]. This diversity creates a steep learning curve for new researchers entering the domain [9, 38, 78].

While there exist many task-specific papers and tutorials, most existing reviews remain architecture-centric, representation-centric, or task-specific, rather than offering a unified and data-centric view that connects data structures, benchmark datasets, and modeling paradigms in one framework [9, 38, 78, 80, 93, 112].

In this paper, we aim to bridge this gap by providing a unified, data-centric perspective on 3D vision. Our contributions are threefold:

- We offer a high-level map of how 3D data are represented, stored, and processed in computers and machine learning systems, covering major formats such as point clouds, meshes, voxel grids, RGB-D images, CAD models, implicit fields, and 3D Gaussians within one unified view [68, 71, 99, 108, 129, 143, 149].
- We highlight how datasets and benchmarks have not only enabled fair evaluation but also actively shaped the evolution of 3D learning paradigms by defining data structures, supervision formats, and scalability constraints [19, 40, 149, 167, 177].
- We situate emerging trends, such as 2D-supervised 3D learning, neural implicit fields, and the extension of 3D vision along the temporal axis to 4D scene understanding and world modeling, within a broader narrative of efficiency, fidelity, and accessibility [2, 71, 100, 101, 107, 134, 150].

By distilling the field's complexity into a structured map, we hope to make 3D vision more approachable, interpretable, and navigable for students and practitioners entering this rapidly expanding area [49, 80].

2. Scope of the Paper

We specify the concrete scope and positioning of this survey. Our coverage spans three core axes:

- **Data Representations:** We review the major data forms in 3D vision—point clouds, meshes, voxel grids, RGB-

D and multi-view images, CAD/B-Rep models, neural implicit fields, and 3D Gaussian—and analyze their efficiency–fidelity trade-offs.

- **Datasets and Benchmarks:** We survey the dataset ecosystem across modalities and tasks, emphasizing how benchmark design both enables progress and constrains model development.
- **Modeling Paradigms:** We summarize classical geometry-based pipelines and modern neural approaches, including 2D-supervised 3D learning, implicit neural fields, and 4D video/world modeling.

Our review differs from existing reviews in both scope and perspective. Architecture-centric works [78, 93] focus on network families but not on the dataset–representation nexus. Topic-centric surveys [9, 38] provide depth on one paradigm while leaving other representations disconnected. Task-oriented overviews [49, 80, 112, 179] offer detailed taxonomies for individual applications but seldom consider supervision strategies or cross-task scalability. Finally, mechanism-focused treatments [67] analyze rendering pipelines in isolation, whereas in our survey differentiable rendering is treated only as one component of a broader supervision spectrum.

3. A Taxonomy of 3D Representations

3D vision relies on diverse data representations—voxel grids, point clouds, implicit fields, and 3D Gaussians—each tailored to specific tasks like reconstruction and recognition. This section categorizes these representations by structure and efficiency and how each data type is acquired.

3.1. RGB-D

RGB-D data integrates RGB color images with per-pixel depth maps, capturing both appearance and geometry in a structured 2.5D format. For each pixel (u, v) in the 2D image grid, the RGB value is denoted by $\mathbf{c}(u, v) \in \mathbb{R}^3$ and the corresponding depth by $d(u, v) \in \mathbb{R}$. The 3D point $\mathbf{p} = (x, y, z)$ can be recovered via:

$$\mathbf{p} = d(u, v) \cdot \mathbf{K}^{-1} \cdot [u, v, 1]^T$$

where $\mathbf{K}$ is the camera intrinsic matrix. This projection enables efficient 2D CNN processing of 3D data with a computational complexity of $O(H \times W)$, where $H \times W$ is the image resolution.

RGB-D data is typically acquired using sensors such as Microsoft Kinect [61, 174], Intel RealSense, or Structure Sensor. The depth map encodes the distance from the camera to visible surfaces in the scene, offering structured 3D geometry at the pixel level [127, 129, 159, 167]. Owing to its compactness and ease of use, RGB-D has become a widely adopted format in various 3D vision tasks, including indoor scene understanding [5, 12, 19, 45, 50], pose estimation [70, 126, 128, 136], and SLAM [20, 65, 102, 120].

3.2. Point Clouds

A point cloud is a set of discrete points in 3D space, typically captured by LiDAR, RGB-D sensors, or photogrammetry [111]. It is defined as

$$\{\mathbf{p}_i = (x_i, y_i, z_i) \in \mathbb{R}^3 \mid i = 1, \ldots, N\}$$

with optional attributes like color or normals. Processing complexity depends on the architecture: PointNet [108] operates in $O(N)$, while Transformer-based models like PointTransformer [175] scale as $O(N^2)$. State-space models, such as PointMamba [82], achieve $O(N)$ complexity by leveraging structured state transitions.

The field began with PointNet/PointNet++ [108, 110], which introduced point-wise and hierarchical feature extraction. Since then, a wide range of methods have been proposed for registration [1, 47, 114, 121, 163], classification and segmentation [44, 76, 79, 92, 115, 169, 176] using deep learning or Transformer-based architectures. Most recently, state-space models have emerged as efficient alternatives to Transformers. Oneformer3d [76], PointMamba [82], Point Transformer [154, 155, 175] and other works [86, 147, 173] significantly reduce computational cost while achieving competitive or superior performance, marking a new trend in point cloud modeling.

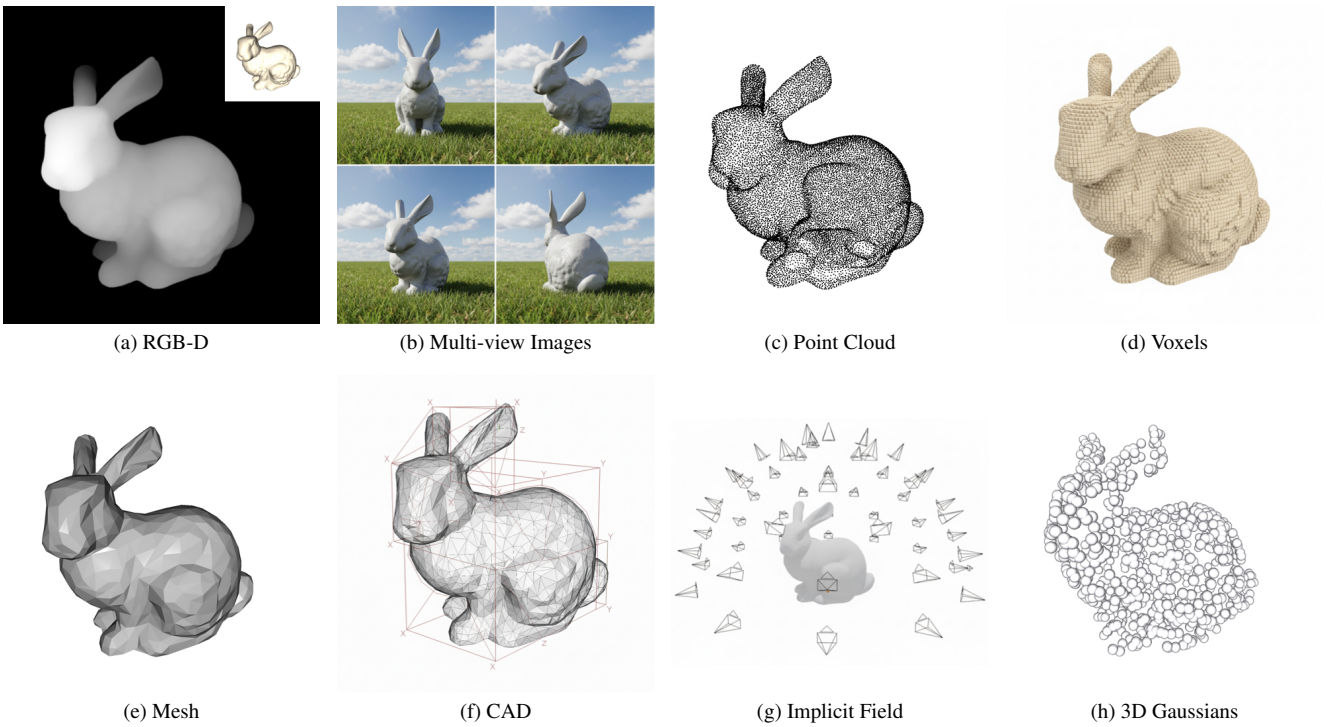
Point clouds can be acquired either directly or indirectly. Direct acquisition uses LiDAR or RGB-D sensors that measure range and back-project observations into 3D coordinates, yielding sparse outdoor scans or organized indoor point sets [3, 61, 111]. Indirect acquisition reconstructs 3D points from image collections via SfM and MVS/photogrammetry, and multi-view or multi-session captures are often merged through SLAM or global registration into a common coordinate frame [20, 122, 123]. In synthetic benchmarks, point clouds are also frequently generated by sampling surfaces from meshes or CAD models, which provides clean geometry with controllable density and annotations [7, 149].

3.3. Voxels

Voxel grids divide 3D space into uniform cells, each of which can store occupancy, color, density, or semantic information [16, 143, 156]. Their regular structure makes them naturally compatible with 3D convolutional neural networks [14, 39, 98, 118], and they are therefore widely used in volumetric reconstruction, segmentation, and object modeling [8, 21, 65, 87, 132, 168].

Voxel grids discretize a 3D volume into a grid of size $N \times N \times N$, where each voxel at position (x, y, z) is assigned a value $v(x, y, z)$. For binary occupancy, this is defined as:

$$v(x, y, z) = \begin{cases} 1, & \text{if occupied} \\ 0, & \text{otherwise} \end{cases}$$



(a) RGB-D (b) Multi-view Images (c) Point Cloud (d) Voxels

(e) Mesh (f) CAD (g) Implicit Field (h) 3D Gaussians

Figure 1. Various 3D representations of the Stanford bunny [138], including RGB-D, multi-view images, point cloud, voxels, mesh, CAD, implicit fields, and 3D Gaussians. These formats illustrate the diversity of 3D data modalities commonly used in benchmarks and learning frameworks.

For continuous attributes such as density or RGB color, the voxel value is given by $v(x, y, z) \in \mathbb{R}^k$, where k denotes the dimensionality of the attribute vector.

Voxel data are rarely sensed directly. Instead, they are typically obtained either by *voxelizing* meshes, CAD surfaces, or dense point clouds into occupancy or attribute grids, or by volumetric fusion of multi-view depth observations in TSDF/occupancy volumes from RGB-D or LiDAR scans aligned across viewpoints [7, 20, 30, 61, 148, 156]. Synthetic benchmarks often produce voxels by rasterizing clean CAD assets, whereas real-scene datasets derive them from fused sensor measurements and then optionally attach color or semantic labels.

3.4. Meshes

Meshes provide a structured surface representation for modeling 3D geometry using vertices, edges, and faces. By explicitly encoding both shape and topology, meshes are well suited for applications such as graphics rendering, CAD design, and physical simulation [68, 69, 98, 109].

Despite their expressiveness and compactness, the irregular structure of meshes makes them challenging to process using standard deep learning frameworks, which are generally optimized for grid-like data. As a result, many pipelines convert meshes to point clouds or voxels before learning [7, 98, 108, 156]. Direct mesh networks such as MeshCNN alleviate this mismatch, but remain more specialized than point- or voxel-based backbones [48].

Meshes are commonly acquired in several ways. Active 3D scanners can capture multiple range images that are aligned and stitched into polygonal surfaces [138]. RGB-D reconstruction systems instead fuse many depth frames into a volumetric field and then extract a surface mesh, as in KinectFusion, DynamicFusion, and BundleFusion [20, 61, 103]. In photogrammetry pipelines, camera poses and dense geometry are recovered from RGB images via SfM/MVS, after which a mesh is reconstructed from the resulting point cloud using surface reconstruction methods such as Poisson reconstruction [68, 69, 122, 123]. Many benchmark meshes are also obtained by tessellating CAD or artist-created assets into triangles before downstream learning [7, 149].

3.5. CAD

Computer-Aided Design (CAD) models describe 3D shapes using *smooth, mathematically defined surface patches*, most commonly through non-uniform rational B-splines (NURBS) [29, 106]. Each CAD model consists of a

Table 1. Summary of common 3D data representations.

Representation	Structure	Efficiency	Fidelity	Applications
RGB-D	2.5D Grid (RGB + Depth)	High	Medium	SLAM, indoor mapping, pose
Multi-view Images	2D Views + Poses	High	High*	SfM, MVS, NeRF input
Point Cloud	Unstructured 3D Points	High	Low–Medium	Detection, mapping, robotics
Mesh	Vertex–Edge–Face Graph	Medium	High	Modeling, animation, simulation
Voxel Grid	Dense 3D Lattice	Low	Medium	Volumetric CNN, segmentation
Implicit Field	Neural Function $f(x)$	Low	Very High	View synthesis, scene modeling
3D Gaussians	Sparse 3D Gaussian Distributions	Very High	High	Real-time NeRF-style rendering
CAD Model	Parametric Surfaces (NURBS)	Very High	Very High	CAD design, reverse engineering

* Fidelity refers to high visual fidelity (appearance); geometric structure must be inferred.

finite set of parametric patches:

$$\mathcal{M} = \bigcup_{k=1}^{K} S_k, \quad S_k : [0,1]^2 \to \mathbb{R}^3$$

with each patch parameterized by [18, 23, 106]

$$S(u,v) = \frac{\sum_{i=0}^{n}\sum_{j=0}^{m} N_{i,p}(u)\, N_{j,q}(v)\, w_{ij}\, \mathbf{P}_{ij}}{\sum_{i=0}^{n}\sum_{j=0}^{m} N_{i,p}(u)\, N_{j,q}(v)\, w_{ij}},$$

$$(u,v) \in [0,1]^2$$

where $N_{i,p}, N_{j,q}$ are B-spline basis functions, $\mathbf{P}_{ij}$ are control points, and $w_{ij} > 0$ are weights. [18, 23] This formulation enables closed-form evaluation of positions, derivatives, normals, and curvature, supporting high-fidelity rendering, exact intersections, and robust Boolean operations. [51, 96, 105, 106]

Data acquisition: CAD data are usually acquired through design workflows rather than direct sensing. In industrial practice, engineers create models interactively in CAD software, which naturally records sketches, constraints, feature histories, and final B-Rep/NURBS geometry; datasets such as Fusion 360 Gallery, SketchGraphs, and DeepCAD expose parts of this process for learning [124, 149, 152]. Large research corpora are also assembled by harvesting existing repositories and converting STEP/B-Rep assets into canonical analytic patches or sequence-like representations, as in ABC and BRep2Seq [74, 171]. When an editable model is needed for a real object, another route is scan-to-CAD retrieval and alignment, where images or reconstructed geometry are matched to a parametric template that can then be refined [43].

3.6. Gaussians Splatting

A 3D Gaussian is a continuous and compact primitive for representing spatial density, and has recently become a popular explicit representation for neural rendering [71]. Similar to point clouds, each Gaussian is defined by a position $\mu = (x, y, z)$ and a covariance matrix $\Sigma \in \mathbb{R}^{3\times3}$ that determines its shape and orientation in space. The probability density function is:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{3/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right).$$

To ensure Σ is symmetric and positive semi-definite, it is decomposed as:

$$\Sigma = RSS^TR^T,$$

where R is a rotation matrix and S is a diagonal scaling matrix. In addition to geometry, each Gaussian carries:

- **Opacity** α, which controls how transparent the Gaussian appears.
- **Spherical Harmonics (SH)** coefficients, which model view-dependent color and enable realistic shading.

Each 3D Gaussian is typically initialized from an SfM point cloud [35, 71, 122], with position $\boldsymbol{\mu}_i$, unit covariance $\Sigma_i = I$, opacity $\alpha_i = 1$, and SH color $\mathbf{c}_i$ from the RGB value. During training, parameters are optimized via gradient descent to minimize the rendering loss $\mathcal{L}_{\text{render}}$:

$$\theta_i^{(t+1)} = \theta_i^{(t)} - \eta \cdot \nabla_{\theta_i}\mathcal{L}_{\text{render}},$$

where $\theta_i \in \{\boldsymbol{\mu}_i, \Sigma_i, \alpha_i, \mathbf{c}_i\}$.

Gaussian-splatting data are typically acquired from calibrated multi-view RGB images or videos rather than from a dedicated sensor. The standard pipeline first estimates camera poses and a sparse point cloud via SfM, optionally densifies geometry with MVS or depth priors, and then optimizes Gaussian positions, covariances, opacities, and colors against photometric rendering losses [71, 122, 123]. Recent methods reduce or remove the dependence on a full SfM/COLMAP-style initialization by learning pose-free or COLMAP-free Gaussian reconstruction from unposed image collections [35, 52, 166]. In online perception, Gaussians can also be updated incrementally from streaming ob-

servations, as demonstrated in Gaussian Splatting SLAM and dynamic 3DGS variants [94, 97].

4. 3D Learning Paradigms and Applications

Modern 3D vision has increasingly shifted from explicit geometry pipelines toward learned systems that couple representation design, supervision, and practical utility [67, 145, 158]. To provide a clear conceptual map, this section is divided into two distinct parts. First, we discuss the core 3D learning and rendering paradigms that dictate how neural networks encode and supervise geometric data. Second, we explore how these fundamental paradigms are deployed across downstream applications, ranging from object reconstruction and scene generation to interactive 4D world models.

4.1. Preliminary: Differentialble Rendering

Early learning-based 3D methods often relied on direct 3D supervision, where losses such as Chamfer distance, Earth Mover's Distance, or volumetric TSDF errors were computed explicitly in 3D space [15, 27, 108, 110, 156]. Although conceptually simple, these objectives become computationally prohibitive for dense voxels or high-resolution surfaces. A pivotal transition came from differentiable rendering frameworks (e.g., Neural Mesh Renderer, Soft Rasterizer, OpenDR) [66, 89, 90]. By backpropagating through the image formation process, these methods replace explicit 3D supervision with image-plane losses on color, depth, or silhouettes:

$$\mathcal{L}_{\text{photo}} = \sum_{i=1}^{N} \left\| I_i - \mathcal{R}(\mathcal{M}_\theta, P_i) \right\|^2$$

where $\mathcal{R}$ is the differentiable rendering operator, $\mathcal{M}_\theta$ is the 3D representation, and P_i denotes the camera parameters [67]. The evolution of this rendering operator defines the computational limits of 3D learning:

- **Volume Rendering (NeRFs):** Early continuous frameworks utilized ray-marching and volumetric integration. While physically principled, the dense multi-layer perceptron (MLP) queries along each ray made end-to-end training on high-resolution data computationally prohibitive [38].
- **Tile-based Rasterization (3DGS):** The introduction of 3D Gaussian Splatting revolutionized the rendering bridge. By replacing implicit MLPs with explicit 3D Gaussians and utilizing a highly optimized, differentiable α-blending rasterizer, 3DGS reduced rendering times from seconds to milliseconds. This breakthrough directly enabled the training of massive, feed-forward 3D foundation models [71].

4.2. Learning Paradigm for End-to-End Geometric Foundation Models:

Building on image-plane supervision, image-aligned representations have emerged as a leading paradigm because they preserve dense per-pixel structure while keeping learning in the 2D domain [84, 142, 145]. Several foundational formulations define this space:

- **DUSt3R [145]:** Learns through confidence-weighted regression on image-aligned 3D outputs without explicit multi-view optimization at training time:

$$\mathcal{L}_{\text{pmap}} = \sum_i \left(\| C_i \odot (P_i - P_i^*) \| - \alpha \log C_i \right) \quad (1)$$

  where P_i and P_i^* are predicted and ground-truth 3D points, and C_i models aleatoric uncertainty.
- **VGGT [142]:** Scales the image-aligned paradigm to large multi-view sets by jointly optimizing a multi-task objective for reusable geometric backbones:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{camera}} + \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{pmap}} + \lambda \mathcal{L}_{\text{track}} \quad (2)$$

- **RayZer [62]:** Factorizes input into camera and scene representations to train entirely through 2D self-supervised reconstruction, without explicit 3D geometry:

$$\mathcal{L}_{\text{RayZer}} = \| \hat{I}_{\text{target}}(\hat{P}_{\text{target}}) - I_{\text{target}} \|_2^2 \quad (3)$$

- π^3 **[146]:** Enforces permutation-equivariant supervision over unordered image sets by optimizing local point maps (X_i) and relative poses $(T_{i\to j})$:

$$\mathcal{L}_{\pi^3} = \mathcal{L}_{\text{local}}(X_i, X_i^*) + \mathcal{L}_{\text{relative}}(T_{i\to j}, T_{i\to j}^*) \quad (4)$$

- **Depth Anything 3 [84]:** Collapses multiple geometric heads into a unified depth-plus-ray representation $R \in \mathbb{R}^{H\times W\times 6}$ (origin and direction):

$$\mathcal{L}_{\text{DA3}} = \mathcal{L}_{\text{depth}}(D, D^*) + \mathcal{L}_{\text{ray}}(R, R^*) \quad (5)$$

Optimization via Generative Priors and Structured Latents: When explicit 3D data is scarce, learning paradigms shift toward distilling priors from large-scale 2D models or utilizing structured latent spaces. Methods like DreamFusion and Magic3D optimize neural fields through Score Distillation Sampling (SDS) [83, 107]. More recently, models have moved toward **Native 3D Geometric Foundation Models**. TRELLIS learns structured 3D latents decodable into radiance fields, Gaussians, or meshes [158]. Concurrently, SAM 3D formulates learning as **Rectified Conditional Flow Matching (RCFM)**, uniquely breaking the 3D data barrier through a **Model-in-the-Loop (MITL)** data engine where generative outputs are human-vetted to create recursive supervision [10].

Table 2. Representative 3D datasets and benchmarks reviewed in this survey.

Dataset	Year	Description
SAM 3D Body [164]	2025	Promptable foundation model for full-body HMR with 5M+ 3D samples
GigaHands [34]	2025	3D bimanual hand dataset with mesh and text labels
InteriorGS [130]	2025	Synthetic indoor scenes with trajectories and dense labels
HPSketch [28]	2025	A history-based parametric CAD sketch dataset
CBF [22]	2025	CAD B-rep models composed of a base plate plus three geometric features
EgoExo4D [40]	2024	Large-scale egocentric and exocentric video dataset with 3D human pose
Parametric 20000 [11]	2024	Multi-modal CAD shapes with point cloud, triangle mesh, and B-Rep file
WildRGB-D [157]	2024	Real RGB-D object videos with 360° views and masks
BRep2Seq [171]	2024	CAD dataset of B-rep solids paired with construction sequences
EgoHumans [72]	2023	Multi-view egocentric dataset for 3D human-human interaction
Aria Synthetic Environments [104]	2023	Synthetic indoor scenes with realistic device paths and labels
DL3DV-10K [85]	2023	Multi-view dataset across 65 scene types for view synthesis
PointOdyssey [178]	2023	Synthetic videos for long-term fine-grained point tracking
Aria Digital Twin [104]	2023	Egocentric dataset with 3D object & human pose
ScanNet++ [167]	2023	High-fidelity indoor scans with RGB-D and dense labels
Objaverse [24]	2023	Large 3D mesh-text pairs for multimodal learning
DIVA-360 [91]	2023	Multi-view dataset for dynamic neural fields
H3WB [182]	2022	Whole-body 3D keypoints for Human3.6M dataset
Kubric [41]	2022	Synthetic generator for scenes/objects with annotations
Amazon Berkeley Objects [17]	2021	Real-world objects with CAD, materials, and image alignment
HM3D [116]	2021	Building-scale indoor meshes with high fidelity
Fusion 360 Gallery Dataset [149]	2021	CAD dataset with meshes and assembly data
CO3Dv2 [117]	2021	Multi-view images + point clouds for 50 object categories
HyperSim [119]	2021	Photorealistic indoor scenes with dense annotations
Habitat 2.0 [133]	2021	Interactive apartments with articulated objects
StrobeNet [170]	2021	Articulated-object dataset with joints and implicit shapes
RELLIS-3D [64]	2020	Multi-sensor dataset for outdoor segmentation
Virtual KITTI 2 [4]	2020	Synthetic KITTI clones with varied conditions
FaceScape [162]	2020	High-quality textured 3D face scans with expressions
3D-FRONT [32]	2020	Synthetic furnished rooms with semantic layouts
3D-FUTURE [31]	2020	CAD furniture models with aligned textures
SketchGraphs [124]	2020	CAD sketches represented as geometric-constraint graphs
Structured3D [177]	2020	Synthetic photorealistic scenes with structure labels
Mapillaryc [26]	2020	Street-level dataset for place recognition
ScanObjectNN [139]	2019	Real-world point clouds with clutter and occlusion
ABC [75]	2019	CAD models with analytic geometry and labels
BlendedMVS [165]	2019	MVS dataset mixing rendered and real images
Replica [131]	2019	Realistic indoor reconstructions with dense labels
3DPW [140]	2018	First dataset with video and 3D ground truth from IMUs in the wild
RealEstate10K [181]	2018	YouTube real-estate videos with camera poses
MegaDepth [81]	2018	Internet photos with dense depth from SfM/MVS
DeepMVS [58]	2018	Synthetic MVS images with ground-truth matching
ScanNet [19]	2017	RGB-D scans with semantic meshes and CAD alignments
Matterport3D [6]	2017	RGB-D scans with panoramic views and segmentation
Thingi10K [180]	2016	3D printable meshes for shape analysis
Semantic3D [46]	2016	Outdoor point clouds (4B pts) with semantic labels
SceneNN [56]	2016	Indoor RGB-D reconstructions with semantic labels
A Large Dataset of Object Scans [13]	2016	Real object scans from diverse environments
Virtual KITTI [37]	2016	Synthetic KITTI sequences with full labels
ShapeNet [7]	2015	Large CAD dataset with rich annotations

This list is not exhaustive; we will maintain an updated version on our GitHub.

The Synergy of Reconstruction and Generation: Historically treated as separate domains, Geometric Foundation model now heavily couple reconstruction and generation. *Generation for Reconstruction* utilizes generative priors (e.g., RCFM or diffusion) to hallucinate missing geometry in ill-posed, sparse-view settings [10, 125]. Conversely, *Reconstruction for Generation* extracts rigid geometric scaffolding to constrain generative models to physically consistent layouts. This synergy increasingly operates within shared latent spaces, enabling a continuous *data flywheel* where synthetic generation and automated reconstruction mutually improve the training corpus [10, 63, 145].

4.3. Downstream Applications

The 3D vision field has also rapidly expanded its applicative scope by leveraging the rendering techniques, image-aligned representations, and End-to-End 3D Geometric Foundation Model.

3D Reconstruction: 3D reconstruction seeks to recover object or scene geometry from visual inputs. Classical pipelines relied on Structure-from-Motion (SfM) and multi-view stereo [36, 122], which are mathematically principled but brittle under sparse views or weak texture. Modern applications replace these bottlenecks entirely with the aforementioned image-aligned neural backbones, enabling robust, end-to-end recovery of point maps, depth, and cameras directly from uncalibrated imagery, even in zero-shot or single-view scenarios [84, 144, 145].

3D Asset and Scene-Level Generation: To circumvent the slow per-prompt optimization of SDS, modern asset generation employs feed-forward multi-view reconstruction. Multi-view diffusion models synthesize view-consistent images, which Large Reconstruction Models (LRMs) instantly map into meshes, tri-planes, or Gaussians [54, 88, 125, 135, 160]. Beyond isolated objects, applications are scaling to composition and layout. Frameworks like 3D-SceneDreamer and AnyHome target open-vocabulary generation of structured, navigable indoor environments with explicit room and object-level organization [33, 172].

3D Consistent Video Generation: Large video diffusion models (VDMs) generate visually stunning content but struggle to preserve stable geometry across time and camera motion. Applications in this domain focus on injecting 3D paradigms to regulate generation [53, 141]. *3D Geometric Preference Alignment* uses 3D consistency as a reward signal, applying Direct Preference Optimization (DPO) based on epipolar Sampson distance or distilled geometric priors from 3D Geometric Foundation Model suppresses physically implausible in videos [25, 77]. *Feature-Level Forcing* aligns latent diffusion features with depth or epipolar lines during denoising [137, 151]. Furthermore, *3D-Aware Control* conditions video synthesis on dense 3D trajectories (e.g., Diffusion as Shader), providing precise spatial manipulation over the generated motion [42].

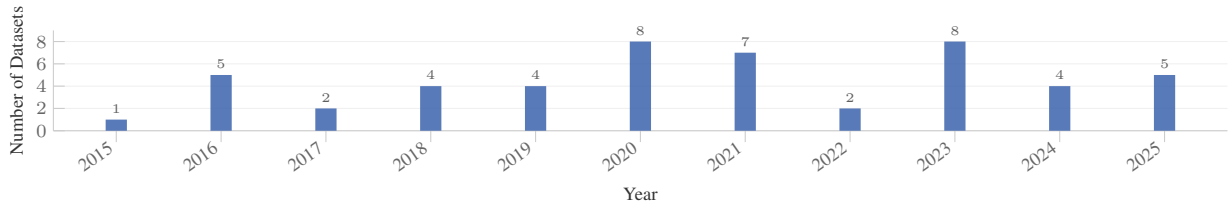
4D Rendering and 3D World Models: The application of 3D vision is expanding toward temporally persistent simulation. **4D Rendering** extends static Gaussian splatting with deformation fields, representing motion as structured 3D evolution rather than a sequence of 2D frames, enabling real-time rendering of dynamic topologies [150, 153]. Extending this concept, **3D World Models** aim to predict future states for planning. Unlike 2D sequence rollouts, models like PointWorld and ParticleFormer push the state space into persistent 3D points or particles [59, 60]. This ensures temporal consistency, strict multi-view faithfulness, and realistic physical interactions as evaluated by benchmarks like WorldSimBench [113].

Spatial Intelligence in Vision-Language-Action: The ultimate practical application of 3D world models lies in Embodied AI. Instead of mapping 2D image tokens directly to embodiment-specific motor outputs (e.g., joint torques), modern 3D-VLA systems ground perception, language, and robotic control in shared 3D representations [55, 57, 161]. By representing intent as 3D point flows or spatial trajectories, these frameworks dramatically improve viewpoint robustness, enable cross-embodiment generalization, and unlock complex spatial reasoning for physical agents [60].

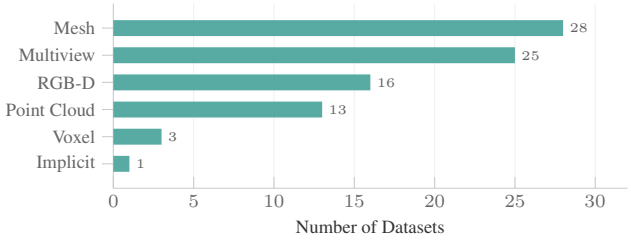
5. Dataset and Benchmark

While Sec. 3 examined the structural spectrum of 3D representations, their practical impact is ultimately mediated through benchmark datasets, which establish learning objectives, task formulations, and evaluation protocols. We categorize existing datasets along four orthogonal axes: (1) **Data modality** (RGB-D, point cloud, mesh, multi-view images, implicit fields, Gaussians); (2) **Spatial granularity** (object-level, scene-level (indoor/outdoor), human-centric (face/hand/body), or mixed); (3) **Task formulation** (segmentation, correspondence, reconstruction, generation); and (4) **Temporal dimension** (static 3D versus dynamic 4D). This lens is increasingly important because recent benchmarks no longer merely collect data; they also encode the assumptions of modern 3D pipelines, from image-aligned reconstruction to 3DGS-native learning [63, 85, 130, 167].

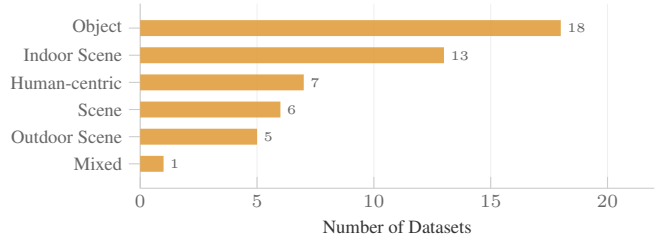
As illustrated in Figure 2, dataset releases have surged over the past decade, reflecting both advances in sensor technology and growing demand for 3D benchmarks. The updated counts from the appendix tables show two especially active release windows in 2020 and 2023, suggesting that benchmark growth is driven not by a steady linear trend but by bursts tied to new sensing pipelines and model families. Recent examples already show three distinct scaling directions: high-fidelity real capture in curated settings (e.g., ScanNet++ [167]), in-the-wild object-centric RGB-D acquisition (e.g., WildRGB-D [157]), and large synthetic or semi-synthetic corpora for long-range correspondences and scene reconstruction, such as PointOdyssey [178] and DL3DV-10K [85]. Modality coverage also remains highly uneven: mesh-backed datasets (28/50) and multi-view benchmarks (25/50) are much more common than voxel



(a) Number of datasets released each year, derived from Table 2.



(b) Dataset counts per modality from Table 3. Modalities overlap across datasets, so the bars are not mutually exclusive.



(c) Dataset counts by spatial granularity from Table 3.

Figure 2. Summary statistics for the 50 representative datasets listed in Tables 3 and 2. The release timeline in (a) shows two recent peaks in 2020 and 2023. The modality chart in (b) replaces the previous pie chart because benchmark modalities are multi-label rather than mutually exclusive. The granularity chart in (c) shows that object-centric and indoor-scene benchmarks currently dominate the landscape.

(3/50) or implicit-field (1/50) datasets. Spatially, object-centric (18) and indoor-scene (13) datasets dominate, while mixed and outdoor settings remain comparatively scarce. We provide a comprehensive breakdown of these statistics in Table 3 in the Appendix, and Table 2 further underscores this fragmentation.

Another recent shift is that benchmark construction itself is becoming model-aware. MegaSynth uses synthesized scenes to scale pretraining for scene reconstruction, while InteriorGS provides semantically labeled indoor scenes directly in the 3D Gaussian Splatting regime rather than only in meshes or point clouds [63, 130]. At the evaluation level, suites such as WorldSimBench suggest that future 3D/4D benchmarks must assess not only reconstruction fidelity but also whether generative models behave like usable simulators under long-horizon, physically grounded tasks [113].

Despite rapid progress, these trends expose fundamental gaps. Current benchmarks still lack large-scale, multi-modal coverage that simultaneously supports heterogeneous representations (e.g., points, meshes, splats, and images), temporal consistency, and open-world generalization. Scene datasets such as ScanNet++ [167] and DL3DV-10K [85] emphasize geometry and view diversity, object datasets such as WildRGB-D [157] emphasize real-world capture, and synthetic datasets such as PointOdyssey [178], MegaSynth [63], and InteriorGS [130] emphasize controllable scale or representation alignment; few benchmarks combine all of these attributes within one unified protocol. Bridging these gaps will require datasets that balance scale with diversity, minimize annotation overhead, and support both synthetic and in-the-wild scenarios—providing the foundation for robust and generalizable 3D/4D learning.

6. Conclusion

We offer a data-centric view of 3D vision, unifying *representations, datasets, and learning paradigms* into a coherent framework. By tracing the trade-offs among different data representations, we clarify how efficiency, fidelity, and scalability jointly shape representation design. We further mapped the benchmark landscape and reviewed the evolution from geometry-based methods to neural implicit fields and 2D-supervised pipelines, highlighting how supervision regimes co-evolve with data availability.

Despite remarkable progress, key challenges remain: fragmented datasets hinder fair comparison, voxel- and mesh-based approaches struggle with scalability, and generalization beyond curated domains is still limited. At the same time, emerging areas—such as 4D spatiotemporal reasoning, physics-aware modeling, and world-consistent video generation—call for tighter integration of 3D priors with multimodal and physical signals.

Looking ahead, we see three promising directions: (i) unified benchmarks and evaluation protocols that span objects, scenes, and dynamics; (ii) cross-modal and 2D-supervised learning strategies that exploit large-scale image data while preserving geometric grounding; and (iii) scalable, real-time representations, from Gaussian splats to parametric CAD, that balance efficiency with fidelity.

References

[1] Yasuhiro Aoki, Hunter Goforth, Rangaprasad Arun Srivatsan, and Simon Lucey. Pointnetlk: Robust & efficient

point cloud registration using pointnet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[2] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models, 2025. 1

[3] J Behley, M Garbade, A Milioto, J Quenzel, S Behnke, C Stachniss, J Gall, and Semantickitti. A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9297–9307. 1, 2

[4] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2, 2020. 6

[5] Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion, 2022. 2

[6] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments, 2017. 6

[7] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository, 2015. 2, 3, 6

[8] Anpei Chen, Zexiang Xu, Matthew Tancik, Jingyi Xu, Xiuming Zhang, Hiroharu Kato, and Jingyi Yu. Tensorf: Tensorial radiance fields. In *ECCV*, pages 333–350, 2022. 2

[9] Guikun Chen and Wenguan Wang. A survey on 3d gaussian splatting, 2025. 1, 2

[10] Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. Sam 3d: 3dfy anything in images. *arXiv preprint arXiv:2511.16624*, 2025. 5, 6

[11] Xi Cheng. Parametric 20000. Mendeley Data, V1, 2024. 6

[12] Sungjoon Choi, Qian-Yi Zhou, and Vladlen Koltun. Robust reconstruction of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5556–5565, 2015. 2

[13] Sungjoon Choi, Qian-Yi Zhou, Stephen Miller, and Vladlen Koltun. A large dataset of object scans. *arXiv:1602.02481*, 2016. 6

[14] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, pages 3075–3084, 2019. 2

[15] Christopher B. Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction, 2016. 5

[16] Ozgun Cicek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. In *MICCAI*, pages 424–432, 2016. 2

[17] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik. Abo: Dataset and benchmarks for real-world 3d object understanding. *CVPR*, 2022. 6

[18] M. G. Cox. The numerical evaluation of b-splines. *IMA Journal of Applied Mathematics*, 10(2):134–149, 1972. 4

[19] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2432–2443, 2017. 1, 2, 6

[20] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics (TOG)*, 36(4):1–18, 2017. 1, 2, 3

[21] Angela Dai, Maximilian Dahnert, and Matthias Nießner. Scancomplete: Large-scale scene completion and semantic segmentation for 3d scans. In *CVPR*, pages 4578–4587, 2018. 2

[22] Yongkang Dai, Xiaoshui Huang, Yunpeng Bai, Hao Guo, Hongping Gan, Ling Yang, and Yilei Shi. Brepformer: Transformer-based b-rep geometric feature recognition. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, page 155–163, New York, NY, USA, 2025. Association for Computing Machinery. 6

[23] Carl de Boor. *A Practical Guide to Splines*. Springer, New York, revised 2001 edition, 1978. 4

[24] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663*, 2023. 6

[25] Hongyang Du, Junjie Ye, Xiaoyan Cong, Runhao Li, Jingcheng Ni, Aman Agarwal, Zeqi Zhou, Zekun Li, Randall Balestriero, and Yue Wang. Videogpa: Distilling geometry priors for 3d-consistent video generation, 2026. 7

[26] Christian Ertler, Jerneja Mislej, Tobias Ollmann, Lorenzo Porzi, Gerhard Neuhold, and Yubin Kuang. The mapillary traffic sign dataset for detection and classification on a global scale, 2020. 6

[27] Haoqiang Fan, Hao Su, and Leonidas Guibas. A point set generation network for 3d object reconstruction from a single image, 2016. 5

[28] Rubin Fan, Fazhi He, Yuxin Liu, and Jing Lin. A history-based parametric cad sketch dataset with advanced engineering commands. *Computer-Aided Design*, 182:103848, 2025. 6

[29] Gerald Farin. *Curves and Surfaces for CAGD: A Practical Guide*. Morgan Kaufmann, San Diego, 5 edition, 2002. 3

[30] Alex Fisher, Ricardo Cannizzaro, Madeleine Cochrane, Chatura Nagahawatte, and Jennifer L. Palmer. Colmap: A memory-efficient occupancy grid mapping framework. *Robotics and Autonomous Systems*, 142:103755, 2021. 3

[31] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture, 2020. 6

[32] Huan Fu, Bowen Cai, Lin Gao, Lingxiao Zhang, Jiaming Wang Cao Li, Zengqi Xun, Chengyue Sun, Rongfei Jia,

Binqiang Zhao, and Hao Zhang. 3d-front: 3d furnished rooms with layouts and semantics, 2021. 6

[33] Rao Fu, Zehao Wen, Zichen Liu, and Srinath Sridhar. Anyhome: Open-vocabulary generation of structured and textured 3d homes, 2024. 7

[34] Rao Fu, Dingxi Zhang, Alex Jiang, Wanjia Fu, Austin Fund, Daniel Ritchie, and Srinath Sridhar. Gigahands: A massive annotated dataset of bimanual hand activities. 2025. 6

[35] Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A. Efros, and Xiaolong Wang. Colmap-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20796–20805, 2024. 4

[36] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multi-view stereopsis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(8):1362–1376, 2010. 7

[37] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis, 2016. 6

[38] Kyle Gao, Yina Gao, Hongjie He, Dening Lu, Linlin Xu, and Jonathan Li. Nerf: Neural radiance field in 3d vision: A comprehensive review (updated post-gaussian splatting), 2025. 1, 2, 5

[39] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. Submanifold sparse convolutional networks. In *CVPR*, pages 9224–9232, 2018. 2

[40] Kristen Grauman, Andrew Westbury, Eugene Patterson, Tsung-Yi Fu, Gijsbert Halbertsma, Lijun Zhao, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23533–23545, 2024. 1, 6

[41] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: a scalable dataset generator. 2022. 6

[42] Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, Wenping Wang, and Yuan Liu. Diffusion as shader: 3d-aware video diffusion for versatile video generation control, 2025. 7

[43] Can Gumeli, Angela Dai, and Matthias Niebner. Roca: Robust cad model retrieval and alignment from a single image. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 4012–4021. IEEE, 2022. 4

[44] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R. Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational Visual Media*, 7(2):187–199, 2021. 2

[45] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from rgb-d images for object detection and segmentation. In *European Conference on Computer Vision (ECCV)*, pages 345–360, 2014. 1, 2

[46] Timo Hackel, N. Savinov, L. Ladicky, Jan D. Wegner, K. Schindler, and M. Pollefeys. SEMANTIC3D.NET: A new large-scale point cloud classification benchmark. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pages 91–98, 2017. 6

[47] Xian-Feng Han, Yi-Fei Jin, Hui-Xian Cheng, and Guo-Qiang Xiao. Dual transformer for point cloud analysis. *IEEE Transactions on Multimedia*, 25:5638–5648, 2023. 2

[48] Rana Hanocka, Amir Hertz, Noa Fish, Raja Giryes, Shachar Fleishman, and Daniel Cohen-Or. Meshcnn: A network with an edge. In *ACM Transactions on Graphics (TOG)*, pages 1–12, 2019. 3

[49] Yong He, Hongshan Yu, Xiaoyan Liu, Zhengeng Yang, Wei Sun, Saeed Anwar, and Ajmal Mian. Deep learning based 3d segmentation: A survey, 2024. 1, 2

[50] Peter Henry, Michael Krainin, Evan Herbst, Xiaofeng Ren, and Dieter Fox. Rgb-d mapping: Using depth cameras for dense 3d modeling of indoor environments. *The International Journal of Robotics Research*, 31(5):647–663, 2012. 2

[51] Christoph M. Hoffmann. *Geometric and Solid Modeling: An Introduction*. Morgan Kaufmann, San Mateo, CA, 1989. 4

[52] Sunghwan Hong, Jaewoo Jung, Heeseong Shin, Jisang Han, Jiaolong Yang, Chong Luo, and Seungryong Kim. Pf3plat: Pose-free feed-forward 3d gaussian splatting, 2025. 4

[53] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 7

[54] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 7

[55] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36, 2023. 7

[56] Binh-Son Hua, Quang-Hieu Pham, Duc Thanh Nguyen, Minh-Khoi Tran, Lap-Fai Yu, and Sai-Kit Yeung. Scenenn: A scene meshes dataset with annotations. In *International Conference on 3D Vision (3DV)*, 2016. 6

[57] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. In *Proceedings of the 41st International Conference on Machine Learning*. JMLR.org, 2024. 7

[58] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6

[59] Suning Huang, Qianzhong Chen, Xiaohan Zhang, Jiankai Sun, and Mac Schwager. Particleformer: A 3d point cloud

world model for multi-object, multi-material robotic manipulation, 2025. 7

[60] Wenlong Huang, Yu-Wei Chao, Arsalan Mousavian, Ming-Yu Liu, Dieter Fox, Kaichun Mo, and Li Fei-Fei. Pointworld: Scaling 3d world models for in-the-wild robotic manipulation, 2026. 7

[61] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Daniel Freeman, Andrew Davison, et al. Kinectfusion: Real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pages 559–568, 2011. 1, 2, 3

[62] Hanwen Jiang, Hao Tan, Peng Wang, Haian Jin, Yue Zhao, Sai Bi, Kai Zhang, Fujun Luan, Kalyan Sunkavalli, Qixing Huang, and Georgios Pavlakos. Rayzer: A self-supervised large view synthesis model, 2025. 5

[63] Hanwen Jiang, Zexiang Xu, Desai Xie, Ziwen Chen, Haian Jin, Fujun Luan, Zhixin Shu, Kai Zhang, Sai Bi, Xin Sun, Jiuxiang Gu, Qixing Huang, Georgios Pavlakos, and Hao Tan. Megasynth: Scaling up 3d scene reconstruction with synthesized data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16441–16452, 2025. 6, 7, 8

[64] Peng Jiang, Philip Osteen, Maggie Wigness, and Srikanth Saripalli. Rellis-3d dataset: Data, benchmarks and analysis, 2020. 6

[65] Haian Jin, Isabella Liu, Peijia Xu, Xiaoshuai Zhang, Songfang Han, Sai Bi, Xiaowei Zhou, Zexiang Xu, and Hao Su. Tensoir: Tensorial inverse rendering, 2024. 2

[66] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer, 2017. 5

[67] Hiroharu Kato, Deniz Beker, Mihai Morariu, Takahiro Ando, Toru Matsuoka, Wadim Kehl, and Adrien Gaidon. Differentiable rendering: A survey, 2020. 2, 5

[68] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Trans. Graph.*, 32(3), 2013. 1, 3

[69] Michael Kazhdan, Michael Bolitho, and Hugues Hoppe. Poisson surface reconstruction. *Proceedings of the Fourth Eurographics Symposium on Geometry Processing*, 7:61–70, 2006. 3

[70] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2938–2946, 2015. 1, 2

[71] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering, 2023. 1, 4, 5

[72] Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani. Egohumans: An egocentric 3d multi-human benchmark, 2023. 6

[73] Georg Klein and David Murray. Parallel tracking and mapping for small ar workspaces. In *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, pages 225–234, 2007. 1

[74] Sebastian Koch, Albert Matveev, Zhongshi Jiang, Francis Williams, Alexey Artemov, Evgeny Burnaev, Marc Alexa, Denis Zorin, and Daniele Panozzo. Abc: A big cad model dataset for geometric deep learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9593–9603, 2019. 4

[75] Sebastian Koch, Albert Matveev, Zhongshi Jiang, Francis Williams, Alexey Artemov, Evgeny Burnaev, Marc Alexa, Denis Zorin, and Daniele Panozzo. Abc: A big cad model dataset for geometric deep learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[76] Maxim Kolodiazhnyi, Anna Vorontsova, Anton Konushin, and Danila Rukhovich. Oneformer3d: One transformer for unified point cloud segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20943–20953, 2024. 2

[77] Orest Kupyn, Fabian Manhardt, Federico Tombari, and Christian Rupprecht. Epipolar geometry improves video generation models, 2025. 7

[78] Jean Lahoud, Jiale Cao, Fahad Shahbaz Khan, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Ming-Hsuan Yang. 3d vision with transformers: A survey, 2022. 1, 2

[79] Xin Lai, Jianhui Liu, Li Jiang, Liwei Wang, Hengshuang Zhao, Shu Liu, Xiaojuan Qi, and Jiaya Jia. Stratified transformer for 3d point cloud segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8500–8509, 2022. 2

[80] Xiaoyu Li, Qi Zhang, Di Kang, Weihao Cheng, Yiming Gao, Jingbo Zhang, Zhihao Liang, Jing Liao, Yan-Pei Cao, and Ying Shan. Advances in 3d generation: A survey, 2024. 1, 2

[81] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 6

[82] Dingkang Liang, Xin Zhou, Wei Xu, Xingkui Zhu, Zhikang Zou, Xiaoqing Ye, Xiao Tan, and Xiang Bai. Pointmamba: A simple state space model for point cloud analysis, 2024. 2

[83] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 5

[84] Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025. 5, 7

[85] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision, 2023. 6, 7, 8

[86] Jiuming Liu, Ruiji Yu, Yian Wang, Yu Zheng, Tianchen Deng, Weicai Ye, and Hesheng Wang. Point mamba: A novel point cloud backbone based on state space model with octree-based ordering strategy, 2024. 2

[87] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Lin Bao, Jan Kautz, and Christian Theobalt. Neural sparse voxel fields. In *NeurIPS*, pages 15651–15663, 2020. 2

[88] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 7

[89] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning, 2019. 5

[90] Matthew M. Loper and Michael J. Black. Opendr: An approximate differentiable renderer. In *Computer Vision – ECCV 2014*, pages 154–169, Cham, 2014. Springer International Publishing. 5

[91] Cheng-You Lu, Peisen Zhou, Angela Xing, Chandradeep Pokhariya, Arnab Dey, Ishaan Nikhil Shah, Rugved Mavidipalli, Dylan Hu, Andrew I. Comport, Kefan Chen, and Srinath Sridhar. Diva-360: The dynamic visual dataset for immersive neural fields. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 22466–22476. IEEE, 2024. 6

[92] Dening Lu, Qian Xie, Kyle Gao, Linlin Xu, and Jonathan Li. 3dctn: 3d convolution-transformer network for point cloud classification. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):24854–24865, 2022. 2

[93] Dening Lu, Qian Xie, Mingqiang Wei, Kyle Gao, Linlin Xu, and Jonathan Li. Transformers in 3d point clouds: A survey, 2022. 1, 2

[94] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis, 2023. 5

[95] Jeffrey Mahler, Jacky Liang, Siddhartha Niyaz, Michael Laskey, Richard Doan, Xue Bin Liu, Jose A. Ojea, and Ken Goldberg. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. In *Robotics: Science and Systems (RSS)*, 2017. 1

[96] Martti Mäntylä. *An Introduction to Solid Modeling*. Computer Science Press, Rockville, MD, 1988. 4

[97] Hidenobu Matsuki, Riku Murai, Paul H. J. Kelly, and Andrew J. Davison. Gaussian splatting slam, 2024. 5

[98] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *IROS*, pages 922–928, 2015. 2, 3

[99] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4460–4470, 2019. 1

[100] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12663–12673, 2023. 1

[101] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65(1):99–106, 2021. 1

[102] Raul Mur-Artal and Juan D Tardos. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017. 1, 2

[103] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 343–352, 2015. 3

[104] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Carl Yuheng Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception, 2023. 6

[105] Nicholas M. Patrikalakis and Takashi Maekawa. *Shape Interrogation for Computer Aided Design and Manufacturing*. Springer, Berlin, 2002. 4

[106] Les Piegl and Wayne Tiller. *The NURBS Book*. Springer, Berlin, 2 edition, 1997. 3, 4

[107] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 1, 5

[108] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 652–660, 2017. 1, 2, 3, 5

[109] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation, 2017. 3

[110] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5099–5108, 2017. 2, 5

[111] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 918–927, 2018. 2

[112] Rui Qian, Xin Lai, and Xirong Li. 3d object detection for autonomous driving: A survey. *Pattern Recognition*, 130: 108796, 2022. 1, 2

[113] Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, Lei Bai, Wanli Ouyang, and Ruimao Zhang. Worldsimbench: Towards video generation models as world simulators, 2024. 7, 8

[114] Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, and Kai Xu. Geometric transformer for fast and robust point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11143–11152, 2022. 2

[115] Shi Qiu, Saeed Anwar, and Nick Barnes. Pu-transformer: Point cloud upsampling transformer. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 2475–2493, 2022. 2

[116] Santhosh K. Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X. Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai, 2021. 6

[117] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *International Conference on Computer Vision*, 2021. 6

[118] Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. Octnet: Learning deep 3d representations at high resolutions. In *CVPR*, pages 3577–3586, 2017. 2

[119] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding, 2021. 6

[120] Nicholas Runz and Lourdes Agapito. Cofusion: Real-time segmentation, tracking and fusion of multiple objects. *IEEE Transactions on Visualization and Computer Graphics*, 24 (11):2957–2968, 2018. 2

[121] Vinit Sarode, Xueqian Li, Hunter Goforth, Yasuhiro Aoki, Rangaprasad Arun Srivatsan, Simon Lucey, and Howie Choset. Pcrnet: Point cloud registration network using pointnet encoding, 2019. 2

[122] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 3, 4, 7

[123] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 2, 3, 4

[124] Ari Seff, Yaniv Ovadia, Wenda Zhou, and Ryan P. Adams. Sketchgraphs: A large-scale dataset for modeling relational geometry in computer-aided design, 2020. 4, 6

[125] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation, 2024. 6, 7

[126] Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from a single depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1297–1304, 2011. 2

[127] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgb-d images. In *European Conference on Computer Vision (ECCV)*, pages 746–760. Springer, 2012. 2

[128] Shuran Song and Jianxiong Xiao. Deep sliding shapes for amodal 3d object detection in rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 808–816, 2016. 2

[129] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 567–576, 2015. 1, 2

[130] Manycore Tech Inc. SpatialVerse Research Team. Interiorgs: A 3d gaussian splatting dataset of semantically labeled indoor scenes. https://huggingface.co/datasets/spatialverse/InteriorGS, 2025. 6, 7, 8

[131] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 6

[132] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction, 2022. 2

[133] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat, 2022. 6

[134] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. *arXiv preprint arXiv:2303.14184*, 2023. 1

[135] Tencent Hunyuan3D Team. Hunyuan3d 2.5: Towards high-fidelity 3d assets generation with ultimate details, 2025. 7

[136] Zachary Teed and Jia Deng. Deepv2d: Video to depth with differentiable structure from motion. In *International Conference on Learning Representations (ICLR)*, 2020. 2

[137] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models, 2023. 7

[138] Greg Turk and Marc Levoy. Zippered polygon meshes from range images. In *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques*, page 311–318, New York, NY, USA, 1994. Association for Computing Machinery. 3

[139] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Duc Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data, 2019. 6

[140] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving

camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018. 6

[141] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 7

[142] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. 2025. 5

[143] Peng-Shuai Wang, Yang Liu, Yueshan Guo, Chun-Yu Sun, and Xiao Tong. O-cnn: Octree-based convolutional neural networks for 3d shape analysis. *ACM TOG*, 36(4):1–11, 2017. 1, 2

[144] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details, 2025. 7

[145] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697–20709, 2024. 5, 6, 7

[146] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Scalable permutation-equivariant visual geometry learning, 2025. 5

[147] Zicheng Wang, Zhenghao Chen, Yiming Wu, Zhen Zhao, Luping Zhou, and Dong Xu. Pointramba: A hybrid transformer-mamba framework for point cloud analysis, 2024. 2

[148] Diana Werner, Ayoub Al-Hamadi, and Philipp Werner. Truncated signed distance function: Experiments on voxel size. In *Image Analysis and Recognition*, pages 357–364, Cham, 2014. Springer International Publishing. 3

[149] Karl D. D. Willis, Yewen Pu, Jieliang Luo, Hang Chu, Tao Du, Joseph G. Lambourne, Armando Solar-Lezama, and Wojciech Matusik. Fusion 360 gallery: A dataset and environment for programmatic cad construction from human design sequences. *ACM Transactions on Graphics (TOG)*, 40(4), 2021. 1, 2, 3, 4, 6

[150] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering, 2024. 1, 7

[151] Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling, 2025. 7

[152] Rundi Wu, Chang Xiao, and Changxi Zheng. Deepcad: A deep generative network for computer-aided design models. In *ICCV*, pages 6762–6772, 2021. 4

[153] Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T. Barron, and Aleksander Holynski. Cat4d: Create anything in 4d with multi-view video diffusion models, 2024. 7

[154] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling, 2022. 2

[155] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4840–4851, 2024. 2

[156] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *CVPR*, pages 1912–1920, 2015. 2, 3, 5

[157] Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos, 2024. 6, 7, 8

[158] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024. 5

[159] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1625–1632, 2013. 2

[160] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 7

[161] Runsen Xu, Xiaojuan Wang, Tai Wang, Kai Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. *arXiv preprint arXiv:2308.16966*, 2023. 7

[162] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: A large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6

[163] Heng Yang, Jingnan Shi, and Luca Carlone. Teaser: Fast and certifiable point cloud registration. *IEEE Transactions on Robotics*, 37(2):314–333, 2021. 2

[164] Xitong Yang, Devansh Kukreja, Don Pinkus, Anushka Sagar, Taosha Fan, Jinhyung Park, Soyong Shin, Jinkun Cao, Jiawei Liu, Nicolas Ugrinovic, Matt Feiszli, Jitendra Malik, Piotr Dollar, and Kris Kitani. Sam 3d body: Robust full-body human mesh recovery. *arXiv preprint*, 2025. 6

[165] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks, 2020. 6

[166] Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images, 2024. 4

[167] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes, 2023. 1, 2, 6, 7, 8

[168] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields, 2021. 2

[169] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19291–19300, 2022. 2

[170] Ge Zhang, Or Litany, Srinath Sridhar, and Leonidas Guibas. Strobenet: Category-level multiview reconstruction of articulated objects, 2021. 6

[171] Shuming Zhang, Zhidong Guan, Hao Jiang, Tao Ning, Xiaodong Wang, and Pingan Tan. Brep2seq: a dataset and hierarchical deep learning network for reconstruction and generation of computer-aided design models. *Journal of Computational Design and Engineering*, 11(1):110–134, 2024. 4, 6

[172] Songchun Zhang, Yibo Zhang, Quan Zheng, Rui Ma, Wei Hua, Hujun Bao, Weiwei Xu, and Changqing Zou. 3d-scenedreamer: Text-driven 3d-consistent scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10170–10180, 2024. 7

[173] Tao Zhang, Haobo Yuan, Lu Qi, Jiangning Zhang, Qianyu Zhou, Shunping Ji, Shuicheng Yan, and Xiangtai Li. Point cloud mamba: Point cloud learning via state space model, 2024. 2

[174] Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE Multimedia*, 19(2):4–10, 2012. 1, 2

[175] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip Torr, and Vladlen Koltun. Point transformer, 2021. 2

[176] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H.S. Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16259–16268, 2021. 2

[177] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling, 2020. 1, 6

[178] Yang Zheng, Adam W. Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J. Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking, 2023. 6, 7, 8

[179] Linglong Zhou, Guoxin Wu, Yunbo Zuo, Xuanyu Chen, and Hongle Hu. A comprehensive review of vision-based 3d reconstruction methods. *Sensors*, 24(7), 2024. 2

[180] Qingnan Zhou and Alec Jacobson. Thingi10k: A dataset of 10,000 3d-printing models. *arXiv preprint arXiv:1605.04797*, 2016. 6

[181] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images, 2018. 6

[182] Yue Zhu, Nermin Samet, and David Picard. H3wb: Human3.6m 3d wholebody dataset and benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20166–20177, 2023. 6

A. Extended Dataset Modalities

Table 3. Modalities of 3D datasets. ✓ denotes availability of the modality, ✗ denotes absence. This table provides a detailed breakdown of data forms across major benchmarks.

Dataset	Granularity (Size†)	RGB-D	Point Cloud	Mesh*	Multiview	Voxel	Implicit
SAM 3D Body	Human (1M+)	✗	✗	✓	✓	✗	✗
GigaHands	Human Hand (14K)	✗	✗	✓	✓	✗	✗
InteriorGS	Indoor Scene (100K)	✓	✗	✗	✓	✗	✗
HPSketch	Object (151.9K)	✗	✗	✗	✗	✗	✗
CBF	Object (20K)	✗	✗	✗	✗	✗	✗
EgoExo4D	Human (1.3k hrs)	✗	✓	✓	✓	✗	✗
Parametric 20000	Object (20K)	✗	✓	✓*	✗	✗	✗
WildRGB-D	Object (8.5K)	✓	✓	✗	✓	✗	✗
BRep2Seq	Object (1M)	✗	✗	✓*	✗	✗	✗
EgoHumans	Multi-human (125K)	✗	✗	✓	✓	✗	✗
Aria Synthetic Environments	Indoor Scene (100K)	✓	✗	✗	✗	✗	✗
DL3DV-10K	Scene (10K)	✗	✗	✗	✓	✗	✗
PointOdyssey	Scene (104)	✗	✗	✓	✓	✗	✗
Aria Digital Twin	Indoor Scene (400)	✓	✗	✗	✓	✗	✗
ScanNet++	Indoor Scene (1K)	✓	✓	✓	✓	✗	✗
Objaverse	Object (800K)	✗	✗	✓	✗	✗	✗
DIVA-360	Object (50)	✗	✗	✗	✓	✗	✗
H3WB	Human (100K)	✗	✗	✗	✓	✗	✗
Kubric	Mixed (N/A)	✓	✓	✓	✓	✗	✗
Amazon Berkeley Objects	Object (8K)	✗	✗	✓*	✓	✗	✗
HM3D	Indoor Scene (1K)	✗	✗	✓	✗	✗	✗
Fusion 360 Gallery Dataset	Object (8K)	✗	✗	✓*	✗	✗	✗
CO3Dv2	Object (19K)	✗	✓	✗	✓	✗	✗
HyperSim	Indoor Scene (461)	✓	✗	✗	✓	✗	✗
Habitat 2.0	Indoor Scene (111)	✗	✗	✓	✗	✗	✗
StrobeNet	Object (120K)	✗	✓	✗	✓	✗	✓
RELLIS-3D	Outdoor Scene (13K)	✗	✓	✗	✗	✗	✗
Virtual KITTI 2	Outdoor Scene (5)	✓	✗	✗	✗	✗	✗
FaceScape	Human Face (18K)	✗	✗	✓	✓	✗	✗
3D-FRONT	Indoor Scene (18K)	✗	✗	✓*	✗	✗	✗
3D-FUTURE	Object (10K)	✗	✗	✓*	✗	✗	✗
SketchGraphs	Object (15M)	✗	✗	✗	✗	✗	✗
Structured3D	Indoor Scene (3.5K)	✓	✗	✓	✓	✗	✗
Mapillary	Outdoor Scene (1.6M)	✗	✗	✗	✓	✗	✗
ScanObjectNN	Object (700)	✗	✓	✗	✗	✗	✗
ABC	Object (1M)	✗	✗	✓*	✗	✓	✗
BlendedMVS	Scene (113)	✓	✗	✓	✓	✗	✗
Replica	Indoor Scene (18)	✗	✗	✓	✗	✗	✗
3DPW	Human (51K)	✗	✗	✓	✗	✗	✗
RealEstate10K	Scene (10K)	✗	✗	✗	✓	✗	✗
MegaDepth	Scene (200)	✓	✗	✗	✓	✗	✗
DeepMVS	Scene (120)	✓	✗	✗	✗	✗	✗
ScanNet	Indoor Scene (1.5K)	✓	✓	✓	✓	✗	✗
Matterport3D	Indoor Scene (90)	✓	✗	✓	✗	✗	✗
Thingi10K	Object (300)	✗	✗	✓*	✗	✓	✗
Semantic3D	Outdoor Scene (30)	✗	✓	✗	✗	✗	✗
SceneNN	Indoor Scene (100)	✓	✓	✓	✓	✗	✗
A Large Dataset of Object Scans	Object (10K)	✗	✓	✓	✗	✗	✗
Virtual KITTI	Outdoor Scene (35)	✓	✗	✗	✗	✗	✗
ShapeNet	Object (300M)	✗	✗	✓*	✗	✓	✗

† Size refers to the number of distinct objects, scenes, or human instances covered by the dataset (rather than the total number of raw data samples).
* Dataset contains CAD meshes.