Survey on 3D Reconstruction Techniques: Large-Scale Urban City Reconstruction and Requirements

Andreas Christodoulides, Gary K. L. Tam, James Clarke, Richard Smith, Jon Horgan, Nicholas Micallef, Jeremy Morley, Nelly Villamizar, and Sean Walton

(Survey Paper)

***Abstract*—3D representations of large-scale and urban scenes are crucial across various industries, including autonomous driving, urban planning, natural resource supervision and many more. Large-scale industrial reconstructions are inherently complex and multifaceted. Many existing surveys primarily focus on academic progressions and often neglect the intricate and diverse needs of industry. This survey aims to bridge this gap by providing a comprehensive analysis of 3D reconstruction methods, with a focus on industrial requirements such as scalability and integration of human interaction. Our approach involves utilizing Affinity Diagramming to systematically analyze qualitative data gathered from industrial partners. This methodology enables us to gain deep insights into how recent literature addresses these specific industrial needs. The survey encompasses various aspects, including input and reconstruction modalities, architectural models, datasets, evaluation metrics, and the incorporation of prior knowledge. We further discuss practical implications derived from our analysis, highlighting key considerations for future advancements in 3D reconstruction methods tailored for large-scale applications.**

***Index Terms*—3D reconstruction, large-scale urban reconstruction, industrial requirements, KJ method, human-in-the-loop.**

I. Introduction

3D RECONSTRUCTION is rapidly expanding within computer graphics and vision. Its diverse applications include urban pollution modeling [40], natural resource supervision [45], autonomous vehicle navigation [11], virtual and augmented reality [75], virtual asset creation [77], telepresence [26], creativity [184], and digital twins [134], to name a few.

3D reconstruction methods create 3D scene representations from input data like images, point clouds, and depth maps. These methods exploit data patterns, requiring compatible models and architectures. Data collection varies between outdoor and indoor settings, with outdoor methods using equipment on land or aerial vehicles. Indoor methods can use in-the-wild imagery [39], scans without camera pose information [11], or camera rigs with known parameters [189]. Single object reconstruction often uses semantic prior knowledge for shape completion [94] or interpolation [184]. Geometric priors can also guide optimization, enrich data, or impose constraints.

TABLE I
Comparison of Recent 3D Reconstruction Surveys

Ref	Non-Learn. Methods	Focus	Prior Break.	Human-in-the-Loop	Indust. Needs
[29]	×	Deep Learning	×	×	×
[117]	✓	Deep Learning	×	×	×
[2]	×	3D Generative AI	×	×	×
[23]	×	3D Human Avatars	×	×	×
[121]	×	Single Object Reconstruction	×	×	×
[34]	×	Single Object Reconstruction	✓	×	×
[27]	×	Single-View Reconstruction	✓	×	×
[116]	×	Single Object Reconstruction	×	×	×
[51]	✓	Underwater Reconstruction	×	×	×
[64]	×	VR Teleoperation	×	×	×
[30]	×	Infrastructure Reconstruction	×	×	×
Ours	✓	Large-scale and Urban Areas	✓	✓	✓

This survey identifies common challenges in 3D reconstruction, such as modeling thin structures, specular or transparent objects, sensitivity to occluded views, and complex illumination conditions. The most notable challenge for city-scale 3D reconstruction is scalability. Deep learning methods are computationally expensive, and costs rise significantly with highly complex scenes like city-scale environments.

Several surveys have addressed 3D reconstruction (Table I). Frashian et al. [29] explore input data, acquisition methods, datasets, evaluation metrics, and deep learning-based methods. Samavati et al. [117] examine diverse 3D reconstructions, including input types, model structures, output representations, and training strategies. Other surveys focus on specific aspects: [2], [23], [88] on human tissue or figures, [121] on single object reconstructions, and [51] on underwater scenes. Finally, [30], [64] investigate industrial applications but do not incorporate industrially driven requirements in their analysis.

We are not aware of any literature survey on 3D reconstruction that thoroughly examines how State-of-the-Art (SOTA) research

Received 25 July 2024; revised 17 December 2024; accepted 1 February 2025. Date of publication 19 February 2025; date of current version 5 September 2025. The work of Gary K. L. Tam was supported by a Royal Society fund under Grant IEC/NSFC/211159. This work was supported in part by EPSRC under Grant EP/S021892/1 and in part by Ordnance Survey. Recommended for acceptance by L. Gao. *(Corresponding author: Andreas Christodoulides.)*

Andreas Christodoulides, Gary K. L. Tam, Richard Smith, Nicholas Micallef, Nelly Villamizar, and Sean Walton are with Swansea University, SA2 8PP Swansea, U.K. (e-mail: a.christodoulides@swansea.ac.uk).

James Clarke, Jon Horgan, and Jeremy Morley are with Ordnance Survey, SO16 0AS Southampton, U.K.

Digital Object Identifier 10.1109/TVCG.2025.3540669

aligns with industrial demands or explores the integration of human interaction within large-scale 3D city reconstruction workflows. In this survey, we address the following research questions: (1) What models, architectures, input data, and reconstruction modalities are best suited for creating methods prioritizing speed, scalability, accuracy, and fidelity of the reconstructions, tailored to industrial requirements for large-scale 3D products? (2) How are human factors integrated within existing SOTA methods? (3)How does current SOTA research align with industrial requirements, and to what extent are these industrial needs being met by the developments in the field?

Large-scale industrial reconstructions are complex and multifaceted, often not fully addressed by academic developments. To guide our survey, we used the organizational exercise Affinity Diagramming (AD), also known as the Jiro Kawakita (KJ) method.

By applying thematic analysis on qualitative data from our partner organization, we derive themes driving our survey. The industry prioritizes scalability, semantic capabilities, well-defined surfaces, speed, and efficient expert collaboration. While State-of-the-Art (SOTA) research investigates most of these requirements, we identify potential shortcomings in producing reconstructions compliant with industrial needs, including strategies for large-scale and complex scenes, watertight and manifold representations, and human interaction integration. We address these discrepancies and provide implications for future work, including data collection practices to reflect real-world scenes, overcoming scalability issues for national-scale reconstructions, and incorporating human-in-the-loop initiatives.

Our main contributions are as follows:

- We present a comprehensive survey of 3D reconstruction, covering input and reconstruction modalities, models and architectures, datasets and evaluation metrics, semantic and geometric priors, and human-in-the-loop initiatives. Unlike other surveys, we incorporate traditional methods, analyze prior information usage, evaluate human interaction in 3D reconstruction pipelines, and discuss industrial needs for large-scale products. Table I provides a comparative summary of our survey against others in the field.
- We introduce a people-centered methodology to elicit requirements for large-scale 3D reconstruction. To our knowledge, this is the first study to prioritize the human element in this context. Using the KJ method, we systematically map qualitative data from interactions with our industrial partner to inform our survey.
- To our knowledge, we are pioneering the comparison of industrial needs with academic research in this field. Our discussion critically analyzes the evaluated research and its implications for future work in large-scale 3D reconstruction, aiming to bridge the gap between academic findings and industrial requirements.

Our survey is structured as follows: Section II presents our scope, methodology and rationale behind the literature collection. Section III presents our chronological diagram mapping the history of the 3D reconstruction field. Section IV presents our Affinity Diagram methodology and analysis. Section V investigates input and reconstruction modalities for 3D reconstruction. Section VI demonstrates different reconstruction methodologies undertaken by the literature while Section VII analyzes the different learning-based frameworks used. Section VIII explores the datasets utilized in 3D reconstruction, followed by Section IX which delves into evaluation metrics used. Section X scrutinizes prior knowledge utilized by 3D reconstruction methods, and Section XI investigates the incorporation of Human-in-the-Loop initiatives within current SOTA. Section XII explores privacy and security mitigation strategies that can be undertaken by 3D reconstruction methods. Our discussion and recommendations for future work is provided in Section XIII, and we conclude in Section XIV.

TABLE II
SOURCES OF THE SURVEY

Journal/Conference	Publications	Journal/Conference	Publications
CVPR	26	ICCV	18
TVCG	10	SIGGRAPH	8
TPAMI	7	SIGGRAPH ASIA	6
arXiv	2	ECCV	1
IJGIS	1	MDPI	1
WAVC	1	N/A	

II. METHODOLOGY AND SCOPE

3D reconstruction has a long history, expanding rapidly with advancements in sensing technologies. While optimization methods were once prevalent, deep learning is now the primary approach. 3D reconstruction creates 3D models from sensed data, potentially from various modalities. Unlike shape completion, which reconstructs scenes or objects from partial inputs, 3D reconstruction generates a complete 3D model of the input scene. Including color reconstruction results in a 360-degree view synthesis.

Our survey addresses industrial needs and large-scale reconstruction with two main criteria: focusing on rigid bodies (excluding organic tissues, human faces, bodies, fabrics, etc.) and 3D reconstruction modalities (point clouds, volumes, depths, radiance fields), excluding 2D layouts, building footprints, and pseudo-3D images. This ensures relevance to large-scale 3D reconstruction of rigid bodies like buildings and city furniture.

We collected and analyzed papers in 3D reconstruction using keywords like ‘3D reconstruction,’ ‘3D surface reconstruction,’ ‘3D object reconstruction,’ ‘large-scale 3D reconstruction,’ and ‘city-scale 3D reconstruction.’ We searched academic databases (Google Scholar, IEEE, Scopus) and recent conference repositories (CVPR, SIGGRAPH/ASIA, ICCV), yielding 188 papers. Older methodologies are covered by [29], [117], so our analysis focuses on recent methods since 2020. Our final selection includes 81 papers based on our research objectives and criteria. The source repositories can be reviewed in Table II

III. HISTORY OF 3D RECONSTRUCTION

In this section, we provide a historical overview of the 3D reconstruction landscape, discussing key methods that have shaped the field. We start with small-scale traditional methods based on optimization principles and show how these concepts expanded to large-scale scenes. We then cover deep learning

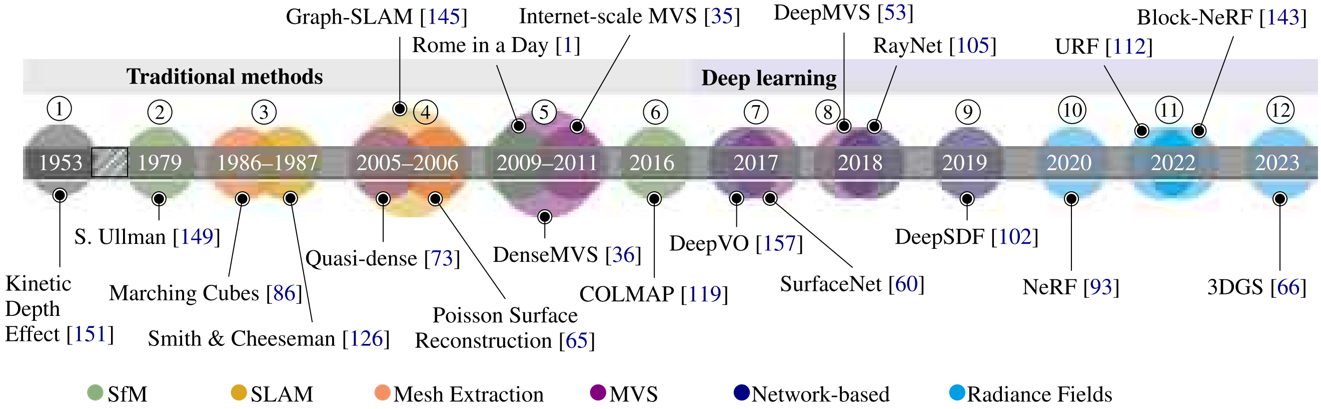

Fig. 1. Chronological diagram mapping monumental methods in 3D reconstruction. We aggregate methods with respect to their publication date and number these aggregations. We use the numbering system to refer to these methods in-text.

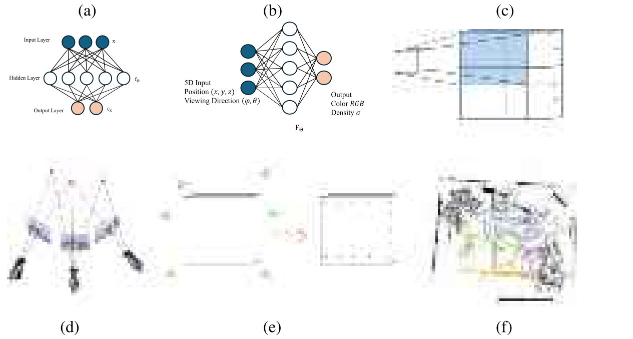

Fig. 2. 3D reconstruction fundamental methods: (a) Multi-layer perceptron (MLP), (b) Neural radiance field (NeRF), (c) Convolutional neural network (CNN), (d) Structure-from-motion (SfM), (e) Multi-view stereo (MVS), (f) Simultaneous localization and mapping (SLAM). (Image courtesy of : [150]).

methods, beginning with initial concepts of popular methodologies and highlighting key concepts that extend learning-based methods to large-scale and complex scenes. We use the numbering system in Fig. 1 to cross-reference these monumental methods, making it easier for our readers to follow the discussion. In addition, in Fig. 2 we visualize fundamental methods in 3D reconstruction.

Small-scale Traditional Methods: Computer vision and 3D graphics began with psychological observations. O'Connell ① [151] noted that humans infer movement and depth from moving images. S. Ullman ② [149] laid the foundations of modern Structure-from-Motion (SfM) algorithms, using three orthographic projections to compute 3D structures. Today, SfM reconstructs 3D structures from 2D images by estimating camera motion [100]. This involves: (1) extracting features to identify keypoints, (2) estimating camera motion and poses, and (3) triangulating or adjusting bundles to compute 3D keypoint positions and sparsely reconstruct scene geometry.

Multi-View Stereo (MVS) often complements SfM, using input images, sparse point clouds, and camera parameters from SfM to reconstruct detailed scene geometry. K.N. Kutalos [72] introduced one of the first MVS frameworks, optimizing points to represent the scene. This voxel-based work laid the foundation for future point cloud contributions. Lhuillier and Quan ④ [73] proposed a patch-based point cloud densification algorithm for MVS. Furukawa and Ponce ⑤ [36], inspired by Lhuillier and Quan, defined the three major MVS steps: matching, expanding, and filtering. Their method improves processing of complex surfaces and outlier removal by iterating between expansion and filtering, unlike the greedy expansion in [73].

However, early SfM and MVS advancements suffered from long computing times and lack of generalizability. Schonben and Frahm proposed the first general-purpose SfM system with COLMAP ⑥ [119], addressing key limitations like robustness, accuracy, completeness, and scalability. They also included a general-purpose MVS implementation. Despite its popularity, COLMAP is sensitive to image quality and requires high-performance GPUs for reasonable computation times.

While primarily focused on robotic navigation, Simultaneous Localization and Mapping (SLAM) is heavily used in 3D reconstruction. Smith and Cheeseman ③ [126] pioneered modern SLAM, introducing techniques to estimate spatial relationships and expected errors between coordinate frames based on object locations. Their approach models sensor uncertainties, allowing a robot to refine its knowledge of its location and surrounding objects through sensor measurements. However, they assumed perfect sensor models, which do not capture real-world conditions. Modern SLAM is a specialized SfM application [100], estimating depth while tracking a moving agent's position.

Explicit Mesh Extraction: The Marching Cubes algorithm by Lorensen and Cline ③ [86] extracts explicit 3D geometries from implicit representations, initially used for medical imaging like CT and MR. Today, it's applied for mesh extraction from implicit volumetric representations. Similarly, the Poisson Surface Reconstruction ④ [65] method reconstructs meshes from oriented point clouds, casting oriented points as a spatial Poisson problem. It uses locally supported basis functions to reduce the solution to a well-conditioned, sparse linear system. Meshes' explicit and continuous properties allow for detailed surface representation and enable tasks like physics simulations.

Deep Learning: With the rise of machine learning, deep learning has become the main approach for 3D reconstruction. Early implementations extended traditional algorithms like

MVS and SLAM into learning-based frameworks using Convolutional Neural Networks (CNNs). The convolution process extracts feature maps to infer depth, shape, or texture. DeepVO ⑦ [157], focusing on SLAM's localization component, shows how feature extraction, matching, and motion estimation can be done end-to-end with Recurrent Convolutional Neural Networks (RCNNs). SurfaceNet ⑦ [60] uses 3D CNNs end-to-end by encoding camera parameters with input images.

Multilayer Perceptrons (MLPs) are simple deep neural networks with an input layer, hidden layers, and an output layer. Their fully connected nature captures intricate details and learns properties of 3D scenes like occupancy [61], surface normals, and implicit representations. DeepSDF ⑨ [102] generates Signed Distance Function (SDF) representations, a continuous volumetric field of the scene. This MLP-based approach improves fidelity, efficiency, and model compression over classical SDF methods.

Mildenhall et al. ⑩ [93] proposed Neural Radiance Fields (NeRF), an implicit representation that receives a continuous 5D vector (spatial location (x, y, z) and viewing direction (ϕ, θ)) and outputs RGB and density for any novel view. NeRF provides unparalleled appearance modeling and realistic illumination but suffers from extended training times and slow novel view synthesis at inference. Kerb et al. ⑫ [66] proposed 3D Gaussian Splatting (3DGS) for radiance field rendering, optimizing a set of 3D Gaussians propagated by SfM. These elliptical primitives with appearance distributions produce a radiance field with much shorter training times than NeRF and instantaneous inference.

Large-scale Traditional Methods: Large-scale scenes are more complex due to many elements present in a scene. GraphSLAM ④ [145] extends traditional SLAM to handle large-scale environments with millions of features. It transforms the SLAM posterior into a graphical network and reduces the graph through variable elimination, resulting in lower-dimensional problems than other methods at the time.

Agarwal et al. ⑤ [1] implemented a scalable SfM using parallel distributed matching and reconstruction algorithms, enabling city-scale reconstruction from thousands of internet images within a day, using a cluster of 500 CPUs. Inspired by this, Furukawa et al.⑤ [35] developed an MVS implementation that uses internet images to produce large-scale, high-density point clouds by clustering overlapping images and processing each cluster independently and in parallel.

Large-scale Deep Learning: Deep learning methods are computationally intensive, requiring high memory, storage, and compute power. DeepMVS ⑧ [53] uses a pre-trained network for per-pixel feature extraction, performing feature aggregation through a U-Net network. However, it suffers from computational intensity. RayNet ⑧ [105] fuses a CNN with Markov Random Fields (MRF) for volumetric reconstruction, encoding the physics of perspective projection and occlusion to estimate surface probabilities along viewing rays.

Few NeRF-based approaches tackle large-scale scenes. URF ⑪ [112] extends NeRF by fusing asynchronously captured LiDAR with images, handling exposure variations, and using predicted image segmentations to supervise ray density towards the sky. This manages the complexity of large-scale scenes and distant sky elements. Block-NeRF ⑪ [143] reconstructs city-scale scenes by training multiple NeRF "blocks" in parallel, each representing different areas. This allows for spatio-temporal updates of local regions without affecting others but requires significant computational power, using 32 TPUs, each storing an individual Block-NeRF model.

IV. Affinity Diagram (AD)

The KJ method, or Affinity Diagramming, developed by anthropologist Jiro Kawakita, helps participants map qualitative data based on natural relationships, aiding decision-making. It minimizes biases by focusing solely on the "affinities" between the data.

The KJ method can also serve as a quality control method, fostering creativity in analyzing unstructured qualitative data [62]. Affinity Diagramming has been used in various industries, including user experience design [50], interactive prototype evaluation [87], customer-centered product design [57], and logistics optimization [95].

A. Affinity Diagram Methodology

Qualitative data collection involved engaging with our industrial partner and directly observing their practices at their headquarters. The lead researcher recorded data throughout the engagement.

We applied the KJ method [120] to create an AD mapping of our partner's product requirements, following these steps: (1) Label making, (2) Label grouping, (3) Chart making, and (4) Written/verbal explanation. Raw notes were transcribed into simplified labeled statements. Relevant labels were grouped based on their "affinities" through iterative adjustments. Preliminary titles were assigned to groups, and statements were re-grouped until no further changes could be made.

B. Affinity Diagram Results

Fig. 3 shows the final affinity diagram. Key themes identified include stakeholder requirements, data availability, industrial standards, and Human-in-the-loop initiatives. The stakeholder requirements theme has three sub-themes: scene understanding capabilities, core functionalities of a 3D reconstruction pipeline, and reconstruction requirements, representing the technical functionalities desired by our industrial partners.

The first sub-theme relates to scene understanding capabilities, including semantic understanding in the 3D reconstruction pipeline, object identification and differentiation within urban scenes, and real-time scene understanding. Required features include noise treatment, removal of unwanted classes, and classification by Level of Detail (LoD).

The Core sub-theme relates to the desired core functionalities of the 3D pipeline. Key aspects include optional class handling with system deletion (identifying and temporarily removing a class of objects) and terrain model generation. Additionally, there is a preference for 3D representations over renderings and incorporating human control within the reconstruction methodology.

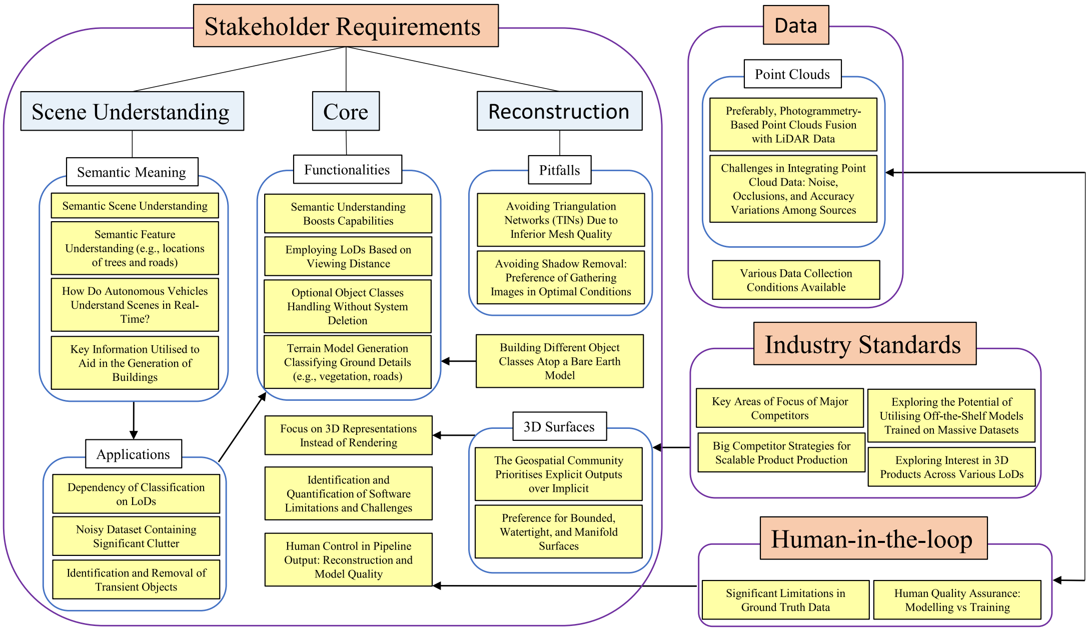
Stakeholder Requirements
Scene Understanding
Core
Reconstruction
Semantic Meaning
Semantic Scene Understanding
Semantic Feature Understanding (e.g., locations of trees and roads)
How Do Autonomous Vehicles Understand Scenes in Real-Time?
Key Information Utilised to Aid in the Generation of Buildings
Applications
Dependency of Classification on LoDs
Noisy Dataset Containing Significant Clutter
Identification and Removal of Transient Objects
Functionalities
Semantic Understanding Boosts Capabilities
Employing LoDs Based on Viewing Distance
Optional Object Classes Handling Without System Deletion
Terrain Model Generation Classifying Ground Details (e.g., vegetation, roads)
Focus on 3D Representations Instead of Rendering
Identification and Quantification of Software Limitations and Challenges
Human Control in Pipeline Output: Reconstruction and Model Quality
Pitfalls
Avoiding Triangulation Networks (TINs) Due to Inferior Mesh Quality
Avoiding Shadow Removal: Preference of Gathering Images in Optimal Conditions
Building Different Object Classes Atop a Bare Earth Model
3D Surfaces
The Geospatial Community Prioritises Explicit Outputs over Implicit
Preference for Bounded, Watertight, and Manifold Surfaces
Data
Point Clouds
Preferably, Photogrammetry-Based Point Clouds Fusion with LiDAR Data
Challenges in Integrating Point Cloud Data: Noise, Occlusions, and Accuracy Variations Among Sources
Various Data Collection Conditions Available
Industry Standards
Key Areas of Focus of Major Competitors
Exploring the Potential of Utilising Off-the-Shelf Models Trained on Massive Datasets
Big Competitor Strategies for Scalable Product Production
Exploring Interest in 3D Products Across Various LoDs
Human-in-the-loop
Significant Limitations in Ground Truth Data
Human Quality Assurance: Modelling vs Training

Fig. 3. Finalized affinity diagram.

The final sub-theme focuses on specific 3D reconstruction requirements. Challenges identified by our industry partner include avoiding Triangulation Networks (TINs) [84] due to inferior mesh quality and skipping shadow removal to optimize cost-efficiency and resource allocation. Additionally, there's an emphasis on ensuring reconstructed surfaces are bounded, watertight, and manifold for seamless integration in engineering applications like simulation.

The second prominent theme is data, focusing on the modalities provided by our partner. Point clouds are preferred for their ease of generation but face challenges like noise, occlusions, and accuracy variations. Other modalities can support or replace point cloud data in reconstruction workflows.

The final two themes are industrial standards and the human-in-the-loop initiative. The Industry Standards theme emphasizes researching key tech organizations with GIS 3D products, their focus areas, and strategies for scalable 3D products. It's also essential to investigate how these entities use 3D products across different levels of detail (LoDs). The Human-in-the-loop theme addresses the lack of ground truth in our partner organization's data. It identifies where people can engage with the 3D pipeline to accommodate this gap, specifically during quality assurance in modeling or training stages.

C. *Affinity Diagram Guiding Survey Research*

We use our AD to guide the analysis of SOTA research, identifying themes like Inputs and Outputs, Datasets, Evaluation Metrics, 3D Reconstruction Strategies, and Learning-based strategies, similar to previous surveys [29], [117], [121]. Our partner's preference for photogrammetry-based point clouds led to a broad analysis of traditional and learning-based methods. The Scene Understanding sub-theme has driven the Prior Knowledge for Reconstruction Section X where we provide with a low-level analysis of prior information used by 3D reconstruction methods, whether they are semantic or geometric. The Data theme has led us to explore multi-modal representations and focus our analysis on large/city-scale reconstructions. Finally, the Human-in-the-Loop theme guided the analysis in Section XI, emphasizing human involvement in the reconstruction process.

V. Input and Output Modalities in 3D Reconstruction

Input data availability often determines the models and architectures used in reconstruction methods. Inputs are categorized as unimodal (single modality, e.g., image sets or point clouds) or multimodal (multiple modalities, e.g., images with depth). Additional modalities can enhance editing capabilities and generate richer 3D features. We categorize inputs and outputs to help researchers understand how different modalities relate to their work.

3D reconstruction outputs can be explicit or implicit. Explicit representations are defined by quantifiable geometrical parameters, with surfaces being parametric (deformations of predefined surfaces) or triangulated (points forming triangles). Implicit representations rely on functions where isosurfaces approximate input data shapes.

A. *Inputs*

Traditionally, 3D reconstruction methods receive an input modality and produce a 3D representation through optimization. With deep learning, multimodal inputs are increasingly used, employing multiple data formats within pipelines. We distinguish inputs as unimodal or multimodal, as shown in Table III.

1) Unimodal Inputs: Most 3D reconstruction literature uses unimodal inputs like images or point clouds. Camera calibration parameters are not considered an additional modality as they

TABLE III
INPUT DATA MODALITIES IMPLEMENTED BY RECENT LITERATURE

Input	Input Type	Research Paper
Unimodal Inputs	Single Image	[19], [21], [56], [83], [89], [94], [163], [179], [187]
	Sequential Images	[31], [39], [48], [78], [133], [158], [159], [188]
	Multi-View Images	[10], [14], [16], [17], [42], [46], [47], [61], [66], [74], [108], [109], [110], [113], [118], [128], [129], [130], [136], [143], [147], [152], [158], [162], [168], [169], [174], [181], [182], [185], [191], [193], [194], [195], [196], [197]
	Point Cloud	[6], [52], [54], [66], [71], [94], [123], [161], [172], [184]
	GNSS Signal	[3]
Multimodal Inputs	RGB-D	[9], [32], [43], [55], [63], [85], [153], [167], [177], [189]
	Language + Images	[77]
	Point Cloud + Images	[112], [148], [166]
	Semantic Point Cloud	[11]

spatially enrich the image rather than representing different information types.

a) Images: In the reconstruction community, images are crucial due to their ease of capture and extensive existing work. Typically represented by the RGB channel, image quality is determined by pixel count. Approaches using this modality can utilize single images, multi-view images, or sequential images (videos).

Multi-view Images: 3D reconstruction methods vary in generating models. Multi-view image techniques capture spatial and geometric data from different angles. Early optimization methods use image redundancy for robust reconstruction [100], establishing frame correspondences and triangulating 3D positions. NeRF architectures [10], [14], [42], [66], [74], [110], [128], [129], [143], [147], [169], [176], [195] utilize multi-view images, needing accurate camera positions for high-fidelity reconstruction. City-scale reconstructions [74], [143], [169] require thousands of multi-view images.

Sequential Images: Sequential images reconstruct dynamic scenes by considering spatial and temporal information. These methods estimate camera poses and use depth recovery techniques like Multi-view Stereo (MVS) [133], Structure-from-Motion (SfM) [66], or Simultaneous Localization and Mapping (SLAM) [9]. Sequential reconstructions are scalable, leveraging known or easily computed camera poses.

Few-shot Images: Few-shot or sparse image inputs often rely on depth estimates [128], [129], [153]. SparseNeRF [153] utilizes coarse depth observations generated from either consumer-level sensors or pre-trained depth models. Sparse depth estimates from sparse views are used to supervise the RF output in [128], [129]. VIP-NeRF [129] generates a visibility prior to guide NeRF supervision based on these sparse depth observations. Both methods use sparse depth estimates from sparse views to supervise their RF outputs. SimpleNeRF [128] employs an augmented model that rejects unreliable depth estimates.

Single Image: Research on generating 3D reconstructions from a single image is increasing. These techniques often use pre-trained deep learning models, enhanced with 2D features from text embeddings. Methods include task-specific convolutional backbones [21], [56], [94], [118], [163], [165], [193] and generative models like diffusion models [19], [89], [184] for single object 3D reconstructions. Lin et al. [83] use a Vision Transformer (ViT) pre-trained backbone to learn global features.

b) Point Clouds: Point clouds are valuable for analyzing 3D spaces, represented as points in a Cartesian coordinate system, and easily generated from sensors like LiDAR [169]. Approaches vary based on point density or sparsity. Point normals, describing surface orientation, impact processing [71], [161]. Some methods [66] generate point clouds through SfM instead of sensors.

c) 3D Mesh: 3D meshes are rarely used as input in 3D reconstructions since they are software-generated and no sensors can reliably produce them automatically. Xu et al. [171] use meshes from oblique photography for reconstructions, mapping them onto a point cloud.

d) Global Navigation Satellite System (GNSS) Data: GNSS data, including GPS and GLONASS signals, offers a cost-effective alternative to laser scanning or aerial photogrammetry [3]. Its main advantage is smartphone compatibility, enabling easy global data collection. However, challenges like antenna directionality, polarization sensitivity, signal strength, and lack of precise ground truth labels limit its use in 3D reconstruction. GNSS data is typically part of a multimodal system with sensors like cameras, LiDAR, or IMUs, enhancing 3D model accuracy. By tracking device movement, algorithms infer spatial layouts. Recently, GNSS-based skyplots transformed into composite images via epipolar geometry have emerged [3], but currently offer low-detail reconstructions.

2) Multimodal Inputs: Multimodal input, though less common than unimodal data, has great potential for enhancing 3D reconstruction. Integrating multiple modalities captures a broader spectrum of information, providing valuable insights into 3D complexity. We distinguish between methodologies using multimodal inputs and those relying solely on unimodal inputs.

a) RGB-D: Techniques using depth information often incorporate it alongside the RGB channel for sequential images [32], [63], [189]. These methods leverage RGB-D streams often alongside SLAM algorithms to estimate camera poses [32], [189]. Depth data enables real-time reconstruction by providing additional geometric information, enhancing accuracy and robustness. SparseNeRF [153] reconstructs scenes from a single image, considering inaccurate depth maps from consumer sensors or pre-trained models.

b) Images With Language: Language has gained attention as a valuable modality for 3D reconstruction. Combining visual and textual information, methods leverage richer modalities with multi-modal models like CLIP [19], [56], [184]. Text in reconstruction pipelines offers enhanced guidance and intuitive interactivity.

c) Images With Point Cloud: Large-scale 3D reconstruction methods often use image sequences combined with LiDAR-derived point clouds [112], [148], [166]. This multimodal approach, first proposed for large-scale RF representation, eliminates the need for dense multi-view images and improves 3D surface extraction accuracy in NeRF-based methods.

d) Segmented Point Cloud: Semantically enriched point clouds improve reconstruction quality. ScanBot [11] enriches point clouds from laser scanning in real-time via a SLAM module. These semantics enhance the efficiency of the

TABLE IV
OUTPUT MODALITIES OF RECENT 3D RECONSTRUCTION METHODS

Output	Output Type	Research Paper
Implicit Repr.	Signed Distance Field (SDF)	[14], [19], [56], [61], [71], [77], [78], [109], [123], [136], [158], [159], [161], [168], [176], [189], [191], [193], [194], [196], [197]
	Truncated SDF	[31], [48], [55], [63], [85], [94], [133], [177], [187]
	Radiance Field	[10], [14], [16], [17], [39], [42], [74], [83], [110], [128], [129], [136], [152], [153], [162], [169], [174], [181], [182], [185], [188], [195]
	Voxelised/Volume	[11], [21], [43], [47], [159], [177]
	Occupancy Field	[61]
Explicit Repr.	Point Cloud	[6], [9], [54], [89], [113], [118], [130], [167], [171], [179]
	3D Mesh	[10], [31], [43], [46], [47], [52], [56], [61], [63], [71], [77], [78], [85], [113], [133], [159], [161], [172], [176], [182], [184], [187], [189], [193], [195], [196], [197]
	Depth Map	[3], [118], [123], [167]
	Graph Repr.	[108]

vehicle's automated scanning and the detail of the volumetric representation.

B. Outputs

The 3D representations generated by modern reconstruction methods can be either explicit or implicit. The Marching Cubes algorithm [86] is commonly used to extract 3D meshes from implicit volumes. Interestingly, there are methods reconstructing hybrid representations that can leverage different components of their explicit and implicit representations to allow for their innovations. We differentiate between methods based on whether their reconstructed outputs are explicit or implicit surfaces, as seen in Table IV.

1) Implicit Representations:

a) Signed Distance Field (SDF): SDFs are widely used in implicit surface-based reconstructions. They position sampling points in a continuous 3D field [102]. The distance between each point and the nearest surface is measured and optimized by a network predicting the SDF. Minimizing a loss function adjusts the SDF to represent the scene accurately. The predicted SDF is compared with image inputs on a pixel-wise basis [196]. Another approach compares the SDF with other 3D modalities, like depth maps [189].

The continuous nature of SDFs allows for smoother representations and more detailed recovery than explicit surfaces like 3D meshes. Their inherent gradient information makes SDF networks robust and efficient, popularizing them in deep learning-based reconstruction methods.

Truncated Signed Distance Field (TSDF): TSDFs are similar to SDFs but use a distance threshold to eliminate points, improving efficiency [133]. This efficiency reduces reconstruction quality, making TSDFs mainly used for real-time reconstruction [31], [63], [94], [133].

b) Occupancy Field: Occupancy fields, like SDFs and TSDFs, determine if a point is inside or outside a surface in 3D space but differ in representation. They use binary values or probabilistic models for occupancy probability [133]. More computationally efficient than TSDFs, they are scalable and easier to train [90]. Using grids or voxels, they are less precise but provide a concise, efficient way to represent occupancy and enable uncertainty modeling

c) Radiance Field (RF): While RFs are mainly for rendering, they connect to 3D reconstruction through volume density learned from multi-view images. NeRF-based methods produce high-quality renderings but require significant computational resources and training data, with limitations in capturing fine details. More details on NeRF methods are in Section III. Zhang et al. [188] propose Nerflets, a set of local radiance fields representing a scene. VQ-NeRF [191] uses Vector Quantization for 3D scene editing. NeRF-XL [74] optimizes multi-GPU scalability with distributed training and rendering, maintaining equivalence to single GPU methods. For a deep overview of NeRF methods and radiance fields, see [37].

2) Explicit Representations:

a) Point Cloud: Point clouds are fundamental in 3D reconstruction, providing spatial details crucial for understanding 3D relationships. Optimized, well-oriented normals in point clouds facilitate detailed 3D mesh extraction [65]. Traditional methods like SfM, MVS, SLAM, and registration techniques generate point clouds as a byproduct [100], [114]. Other methods [54] extend NeRF approaches for novel inputs like LiDAR, enabling LiDAR-based novel-view synthesis.

Point clouds offer several advantages as a 3D representation. They provide a direct depiction of scene geometry, preserving fine details and enabling flexible processing and analysis. Point clouds are ideal for applications needing precise spatial information, like autonomous driving and environmental mapping. However, they can be memory-intensive and computationally demanding due to their large size. Other representations, like meshes or volumetric grids, may be more compact but typically sacrifice some detail at the individual point level.

b) 3D Mesh: Meshes are a common output modality for 3D reconstruction, composed of vertices, edges, and faces. The size of the polygonal faces determines the reconstruction's detail level. Algorithms like Marching Cubes [86] extract meshes from volumetric implicit representations, enabling straightforward and interactive visualizations. Guo et al. [46] uniquely reconstruct a manifold surface mesh from a 3D line cloud, comprised of matched 2D segments from images.

3) Hybrid Representations:

a) Meshes With Implicit Representations: Recently, researchers have developed hybrid mesh/implicit representations. Dreameditor [195] uses a hybrid RF/mesh representation, allowing mesh edits via text prompts to reflect in the RF. Yuan et al. [182] construct meshes corresponding to the radiance field, using ARAP [131] mesh deformation for edits. BakedSDF [176] provides a neural volume-surface representation, baking it into a high-quality, editable mesh with real-time rendering. VMesh [47] leverages a mesh/volume hybrid modality, overcoming volume rendering complexity for fast, high-resolution, storage-efficient renderings capable of modeling complex geometries.

b) Voxels With Implicit Representation: LoD-NeUS [196] combines voxels with SDFs for a hybrid representation. To reduce the memory demands of voxel-based representations, they use a novel tri-plane position encoding. They address voxel aliasing issues with multi-convolved features to approximate ray cone sampling and estimate the SDFs of the samples. An error-guided sampling strategy directs SDF growth, recovering fine details like thin structures.

c) Point Cloud With Implicit Representation: URF [112] combines radiance fields and point clouds for a hybrid representation. It produces accurate point clouds by leveraging ray parameters and depth estimates. However, while the point cloud corresponds to the RF in location, it is computed using trained NeRF parameters and cannot be inversely used to influence the RF output.

4) Inputs and Outputs Discussion: 3D reconstruction methods use diverse modalities for various purposes. For image inputs, different architectures suit different capture conditions and image densities. Multi-view images are computationally expensive, used for implicit scene representations like SDFs [14], [61], [136], [158], [168], [196], [197] and RFs [10], [14], [42], [66], [110], [128], [129], [176], [195]. They lack scalability, needing large datasets for panoramic views. NeRF-based methods are costly at inference due to multiple network evaluations per pixel but offer accurate scene reconstruction with realistic lighting. Sequential images are faster, enabling real-time reconstructions [31], [63], [133], and can be used unimodally or multimodally, often with depth [94]. Unimodal methods like SfM [66] and MVS [133] recover depth or generate point clouds. Multimodal approaches use SLAM for camera parameter estimation [32], [189] and geometric priors [189]. Temporal data methods often use TSDF-based reconstructions for real-time results, with Occupancy Fields also used for real-time inference. Generally, faster or cheaper methods tend to be less robust.

Integrating language with image-based inputs enhances 3D reconstruction models, enabling richer semantic feature extraction and improved accuracy. This multimodal approach allows interactive, text-guided generation. While point-based shape completion is beyond our survey's scope, many methods use pre-trained backbones to generate enhanced features and enable single-image reconstruction, reducing the need for extensive dataset annotation.

Point clouds, derived from sensor data, are widely used in real-world scenarios. Despite being memory-intensive and costly to process, recent advancements have mitigated these challenges. Some methods focus on high-fidelity reconstructions without normal information. 3DGS [66] demonstrates an RF output method using unoriented point clouds for fast, scalable reconstructions. POCO [6] showcases a point feature convolution strategy for scalable large scenes through SDF outputs.

VI. 3D Reconstruction Strategies

Most recent 3D reconstruction methods use deep learning, surpassing traditional methods like SfM, MVS, and standard optimization. With increased computational power, deep learning excels across all metrics. These methods can be supervised (using paired data), unsupervised (learning without labels), self-learning (learning independently), or weakly-supervised (learning with partial data). Fig. 4 shows our 3D reconstruction categorizations. Table V summarizes learning-based strategies discussed in the literature. Methods relating to *Large-Scale Reconstruction* are grouped together and clearly identified. We consider both city-scale methods and smaller outdoor scenes.

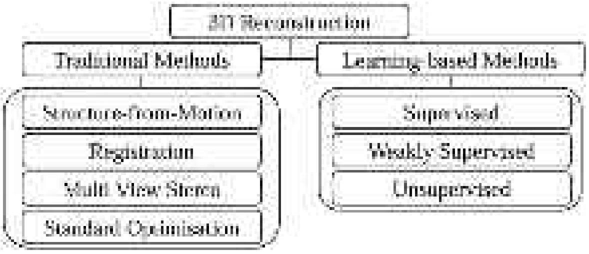

Fig. 4. 3D reconstruction techniques categorization.

A. Traditional/Optimization-Based Methods

We summarize recent improvements and applications to traditional techniques. For more details on optimization methods in 3D reconstruction, see the review by Poh Lim and Haron [82].

1) Structure-From-Motion (SfM): Some recent techniques aim to enhance SfM algorithms. NeRF-based methods often use SfM when camera positions are unknown. CamP [104] introduces a preconditioning method to address imperfectly generated camera positions in SfM. It samples points from a camera frustum and uses their projections as a proxy problem, computing the covariance matrix of these projections to establish a per-camera precondition. Instead of directly optimizing camera parameters, CamP optimizes preconditioned parameters and their covariance matrix to obtain suitable camera parameters for NeRF.

3DGS [66] uses multi-view images and their corresponding camera calibrations from SfM as input. They generate a sparse point cloud from these inputs to create their sets of 3D Gaussians.

Large-Scale Reconstruction: As mentioned in Section III, even modern SfM approaches such as COLMAP [119] struggle with large-scale and complex scenes due to the unavoidable dynamic elements present.

2) Simultaneous Localization and Mapping (SLAM): Recent literature employs SLAM with sequential data sources [9], [11], [189]. It is straightforward to envision replacing the robotic or vehicle agent with a camera position to generate camera parameters in reconstruction methodologies. ScanBot [11] uses a robotic agent to optimize both the reconstructed scenes and the robot's trajectory. GO-SLAM [189] continuously aligns camera poses across frames from sequential image inputs using the SLAM algorithm. PlanarFusion [43] employs a submap-based planar SLAM system, estimating relative poses between submaps through global optimization to ensure a consistent and dense 3D reconstruction.

Large-Scale Reconstruction: While SLAM has been used before for outdoor scene reconstruction, we observe that recent

TABLE V
ARCHITECTURE TYPES AND LEARNING STRATEGY IMPLEMENTED BY RECENT LITERATURE

Architecture Type	Learning Methods			
	Supervised	Self-supervised	Weakly-Supervised	Unsupervised
Multilayer Perceptron	[16], [31], [42], [47], [54], [56], [61], [71], [74], [77], [78], [109], [110], [113], [129], [152], [158], [169], [174], [182], [184], [188], [189], [191], [195]	[10], [17], [181]	[128], [153], [194]	[6], [14], [32], [136], [161], [168], [176], [185], [196], [197]
Convolutional Network	[19], [31], [56], [63], [94], [130], [167], [184], [187], [193]	[9], [10], [11], [85], [118], [179]	[55]	[6], [108], [161], [168], [196]
Gated Recurrent Units	[39]	[11], [118]		
Generative models	[19], [89], [94], [184], [193], [195]			[108], [172]
Transformer	[63], [94], [113], [187], [193]		[153]	
Traditional Machine Learning	[171]			

methods are primarily focused with indoor settings. Recently, some methods [58], [101] have been proposed that have the ability to handle large-scale and outdoor scenes by combining LiDAR with NeRF to generate neural implicit map representations.

3) Multi-View Stereo (MVS): MVS attempts to receive SfM inputs and outputs (source images, sparse points clouds, camera poses) and generate a dense representation of the scene [114], [132]. Details on MVS history and inner workings can be found in Section III.

Deep learning methods often use MVS for depth prior generation. For instance, NerfingMVS [162] uses an MVS-based point cloud from COLMAP [119] to generate depth priors for training a monocular depth network. Similarly, FineRecon [133] employs MVS depth estimates to inject a multi-view depth prior into monocular video data. C2F2NeUS [168] uses MVS for cross-view fusion, enhancing reconstruction accuracy with per-view depth estimates.

4) Standard Optimization: Under Standard Optimization, methods apply distinct principles from SfM or MVS techniques. GaussianFusion [15] integrates features from multiple views using geodesic curves between Gaussian measurements, optimized with a network simplex algorithm. Guo et al. [46] construct a 3D line cloud from 2D lines extracted from images. They use a multi-resolution line detector to extract 2D segments, cluster them with segments from neighboring views, and use a probabilistic Bayesian plane prediction to generate candidate planes for 3D reconstruction. Their 3D model is generated using PolyFit [98]. ManhattanFusion [177] employs Manhattan keyframes, aligning with dominant orthogonal directions in structured environments, as geometric constraints for local pose optimization, but is limited to indoor scenes.

Large-Scale Reconstruction: City3D [52] introduces an optimization framework for large-scale building reconstructions. Guided by building footprint data, they segment point clouds from aerial LiDAR data for detailed reconstruction. To handle missing vertical walls, they incorporate an energy term favoring roof structures and enforce constraints for topological correctness, based on the PolyFit [98] method. Poliarnyi [106] proposes a method for generalizing over Depth maps, RGB-D images, and Terrestrial LiDAR point clouds with known sensor poses in city-scale contexts, using adaptive octrees optimized with Total General Variation (TGV) minimization.

B. Learning-Based Frameworks

Deep learning methodologies are the primary approach for 3D reconstruction in recent literature, varying based on data quality and availability. In Section VI-B1, we summarize supervised methods based on the type of data used for supervision. In Section VI-B2, we cover methodologies that use weakly-supervised or self-supervised strategies. Finally, in Section VI-B3, we discuss methodologies that do not use additional data modalities for supervision.

1) Supervised: Supervised techniques in 3D reconstruction use annotated data to guide learning, enabling intricate relationships between input and output. Common data pairs include image-depth, image-3D, and image-language, with depth and 3D modalities directly associating images with dimensional data. Much research uses NeRF supervision, measuring differences between predicted and input pixel values. These approaches are crucial for accuracy and reliability in computer vision and 3D reconstruction.

a) Occupancy: Some techniques use labeled data to address occupancy in 3D modeling and reconstruction [6], [71], [171]. Occupancy refers to determining whether points in a 3D space are inside or outside a defined object or volume. POCO [6] is a supervised method that uses labels to indicate whether points in a point cloud are within or outside a shape. It encodes input points into latent vectors and learns an occupancy function mapping each query point to a probabilistic occupancy. Hewa Koneputugodage et al. [71] use labeling to supervise the initial training of their INR model to obtain an SDF, which involves understanding the occupancy of points in 3D space.

b) NeRF-Based Supervision: Methods [16], [78], [110], [158], [188] use a supervised approach due to their direct implementation of the NeRF [93] method. Neuralangelo [78] and NeUS2 [158] employ a color loss between input and rendered images, while MERF [110] also considers density. Other methods [39], [47], [128], [129] have their own supervision strategies. VMesh [47] uses a supervised appearance optimization strategy, blending volume-rendered and rasterized colors based on mesh opacity, then comparing this blend with ground truth pixels. SurfelNeRF [39] produces intermediate and local rendering volumes for direct comparison to ground truth images. ViP-NeRF [129] estimates a visibility prior map from primary and secondary views to supervise pixel visibility in their NeRF model. Nerflets [188] performs 2D supervision of their

MLPs using segmented and instance images from their semantic backbone. NeRF-Art [152] applies CLIP-based supervision to constrain similarity between rendered views and text in the embedding space. Jiang et al. [61] reformulate the volume-rendering equation to replace densities with an occupancy grid or SDF, allowing 2D supervision by comparing ground truth color at the pixel emitting ray. Lin et al. [83] use a supervised methodology with a pre-trained backbone learning specific objects from the ShapeNet dataset. NFL [54] extends the NeRF architecture to LiDAR representations, optimizing synthetic views to match input LiDAR characteristics.

Prior work shows that methods advancing NeRF-based architectures primarily use direct comparison with ground truth images for supervision. This strategy is suitable for NeRF supervision, as the main output is the RF itself. However, the RF is a pseudo-3D reconstruction, composed of new images covering all possible viewpoints. RF-based methods assess image quality and fidelity but cannot assess 3D volume density, which is often too noisy to extract a high-quality mesh. NeUS [155] partially alleviates this by reformulating the volume rendering equation to produce an SDF with a coloring field. For further reading on NeRF advancements, refer to a NeRF-focused survey [37].

Large-Scale Reconstruction: While [112], [166], [188] do not provide with a city-scale methods, their methods schemes tackle large scale and outdoor reconstruction. URF [112] segments sky pixels to address NeRF limitations with infinitely distant pixels, like those pointing to the sky. S-NeRF [166] extends URFs [112] by being able to receive sparse and real LiDAR measurements. Nerflets [188] use their novel representation to perform 2D semantic supervision for novel-view synthesis. Finally, the city-scale Block-NeRF [143], computes a visibility prediction parameter to supervise their density prediction network.

c) Depth-Based Supervision: Many authors implementing supervised methods rely on depth supervision [9], [63], [167], [194]. Zhu et al. [194] use depth and normal priors to supervise the network for handling shape-radiance ambiguities. Cai et al. [9] use image sequences and raw depth data to reconstruct depth by prediction, specializing in transparent objects. DG-Recon [63] employs a depth network trained in a fully supervised manner, with depth sensor readings as the target. FrozenRecon [167] leverages pre-trained depth estimation to optimize geometric and photometric consistencies, based on depth estimates from the affine-invariant depth model's predictions.

Most depth supervision methods use sequential data, with the depth channel from RGB-D input supervising the reconstruction. However, alternative strategies, such as [194], compute depth as a prior. These methods rarely rely on raw depth readings and usually optimize the depth before using it for reconstruction.

Large-Scale Reconstruction: Depth plays a key role in supervising large-scale NeRF-based methods [112], [148], [166]. URF [112] uses LiDAR scans to supervise the expected depth from the volumetric rendering process, matching the LiDAR measurements. SUDS [148] uses additional data for supervision, including sparse LiDAR depth measurements, DINO descriptors for semantic manipulation, and 2D optical flow predictions to model scene dynamics. S-NERF [166] uses a depth/flow-based strategy, constructing a confidence map to address LiDAR measurement imperfections and ensuring geometric consistency across different views. Both methods [148], [166] base their supervision strategy to [112] aiming to train on sparse and imperfect LiDAR scans.

d) Language-Based Supervision: Language can be used to supervise networks. Uni3D [187] relies on the image-language pre-trained CLIP model, aligning CLIP features with those from a point cloud. ShapeClipper [56] learns a single-view 3D reconstruction model without 3D viewpoints or multi-view supervision, using Semantic-based Consistency to pull instances with similar CLIP embeddings together. NeRF-Art [152] developed a CLIP-based contrastive loss to align the solution with the target style and distance it from other styles, combining a directional constraint and a global-local contrastive loss.

e) TSDF Supervision: FineRecon [133] and CVRecon [31] use ground-truth TSDF volumes to supervise their 3D geometry predictions. It integrates depth information into an initial TSDF reconstruction, then fuses this TSDF with a global volume using a 3D CNN, supervised by initial TSDF values where ground truth data is available. Liu et al. [85] use a multi-resolution supervision strategy, relying on pairs of incomplete-complete TSDFs at low and high resolutions.

2) Semi-Supervised: Semi-supervised strategies in 3D reconstruction overcome the need for abundant, well-structured datasets. Recent methods use either weakly-supervised or self-supervised procedures. Weakly-supervised methods use sparsely labeled data for supervision. Self-supervised methods eliminate the need for ground-truth data by using depth information from the input data to refine the 3D representation.

a) Weakly-Supervised: We observed limited work on weakly-supervised methods for 3D reconstruction. Zhu et al. [194] train material and emissions fields by re-rendering results instead of relying on strong material priors. SimpleNeRF [128] evaluates depth accuracy from its augmented models and uses accurate estimates to supervise their NeRF model, enabling few-shot training. SparseNeRF [153] infers an RF from a single image using a weakly supervised method with relative depth information from coarse depth maps. When depth data is lacking, the DPT model serves as a pre-trained backbone, ensuring consistency between the NeRF-generated and backbone depth maps through a spatial continuity constraint.

b) Self-Supervised: Almost every self-supervised method employs depth in their methodologies. RICO [77] uses image inputs and semantic masks. They developed a self-supervised strategy to eliminate artifacts in occluded regions by computing normals and depths to create a semantic patch mask. These masks regularize occluded background regions, forcing smoother surfaces. The semantic logits from these masks are used in the cross-entropy loss to facilitate SDF growth. SceneRF [10] optimizes depth without ground-truth data using a NeRF model for depth synthesis and applies photometric projection loss to optimize this depth, then fuses a TSDF onto the optimized depth.

Large-Scale Reconstruction: R3D3 [118] reconstructs large-scale scenes by training depth on raw, real-world sequences in a self-supervised manner, minimizing view synthesis loss and eliminating reliance on synthetic data. They use a co-visibility graph to accumulate point clouds over time, representing the underlying scenes. GridNeRF [169] uses a self-supervised two-branch network. A shallow MLP renderer, pre-trained to predict radiance values of coarsely sampled rays, generates multi-scale feature maps. These maps guide the NeRF branch's sampling focus, crucial for city-scale representation.

3) *Unsupervised:* Unsupervised techniques do not use ground-truth data for supervision. Instead, they discover underlying patterns in the input data to infer 3D shapes, such as generative models that learn 3D representations from the latent space. Our analysis reveals no unsupervised methods targeting large-scale or city-scale reconstruction.

a) Normal Free Methods: Unsupervised 3D reconstruction methods use various strategies. 3DGS [66] models geometries using 3D Gaussians without point normal information. Wang et al. [161] handle unoriented point clouds by incorporating a singular-Hessian term into their implicit function, progressively reducing its influence and using a coarse-to-fine procedure for high-fidelity reconstruction.

b) SDF Guidance: C2F2NeUS [168] produces Signed Distance Functions (SDFs) for their smooth geometries, training their SDF network unsupervised via volume rendering. BakedSDF [176] identifies and removes erroneous mesh regions without supervising its SDF model. LoD-NeuS [196] uses error-guided sampling for SDF refinement. SOL-NeRF [136] computes lighting-SDF representations unsupervised, using Spherical Gaussians and Spherical Harmonics for lighting estimation, predicting SDFs with NeUS.

c) Unsupervised Data Generation: Generative network-based methods [19], [94], [184] use unsupervised learning. LION [184] denoises data through diffusion, learning shape latents from random noise. AutoSDF [94] extracts low-dimensional 3D shape representations with a patch-wise encoder and uses a transformer-based autoregressive model. SDFusion [19] compresses high-resolution 3D shapes into a compact latent space using a diffusion model.

VII. Learning-Based Frameworks

In this Section we describe the technical frameworks used by deep learning methods. Where appropriate, we use the same notation to Section VI to highlight how these frameworks are used when reconstructing outdoor and large-scale scenes.

A. Multi-Layer Perceptron (MLP)

MLPs are key for producing implicit representations, useful for computing fields like radiance [10], [14], [39], [42], [74], [83], [110], [112], [128], [129], [143], [153], [169], [176], [195], simple coloring [32], [129], [158], [168], [189], signed distance [14], [47], [56], [136], [158], [161], [168], [183], [189], [194], [196], volume [39], [77], [129], light transport [183], reflectance [191], and material [194]. VQ-NeRF [191] uniquely transforms continuous reflectance fields into discrete ones for material decomposition.

Large-Scale Reconstruction: The following methods address large-scale scenes [74], [136], [143], [169], [188]. BlockNeRF [143] and NeRF-XL [74] manage city-scale scenes with multiple GPUs, each storing a NeRF model. GridNeRF [169] achieves city-scale reconstructing by using feature grid representations for computational efficiency. Nerflets [188] employ compact, localized neural radiance fields for efficient memory use. SOL-NeRF [136] uses a two-branch approach for light decomposition in large-scale scenes. Both [136], [188] handle large-scale and outdoor scenes but lack city-scale reconstruction capabilities.

B. Convolutional Neural Networks (CNN)

Many methods use convolutional backbones for feature extraction and transfer learning. ResNet-based encoders are common for extracting image features [21], [56], [94], [118], [163], [165], [187], [193]. DG-Recon [63] and SurfelNeRF [39] use the MnasNet backbone [140]. Other methods [6], [10], [133], [188] use various convolutional backbones [7], [18], [49], [141], [142]. Nerflets [188] use Panoptic-deeplab [18] for semantic and instance images. POCO [6] uses FKAConv [7] for point-based convolutions. CVRecon [31] uses a monocular depth estimation backbone and a 3D CNN for TSDF volume transformation. Lin et al. [83] pair a ViT-based encoder with a convolutional decoder, generating multi-level features, which a 2D CNN then combines into a hybrid feature representation.

CNNs are robust for processing images, but recent reconstruction methods prefer MLPs to learn complex spatial relationships. These relationships can be signed distances for SDFs or 5D vectors for RFs. Methods using CNNs often employ a backbone to enrich reconstruction with semantics or geometric information.

Large-Scale Reconstruction: Methods [118], [163] use CNNs for large-scale scenes but do not consider city-scale scenes. R3D3 [118] compresses multi-view images with a U-Net module to predict depth in complex scenes. Wimbauer et al. [163] predict accurate depth from a single image but haven not shown how to fuse sequential depth predictions into a complete scene.

C. Transformer

Transformer models use attention mechanisms to manage input-output relationships, excelling at capturing long-range dependencies and contextual information. Initially for natural language processing, transformers have been extended to vision applications like the Vision Transformer (ViT) [25]. In 3D reconstruction, they model complex spatial relationships, handle multi-modal data, and perform data fusion, resulting in more accurate and detailed 3D models.

AutoSDF [94] uses a Transformer-based autoregressive prior over a discretized latent space to reduce complexity. Volrecon [113] employs two transformers with linear attention: a view self-attention transformer for multi-view feature aggregation and a ray-based transformer for computing SRDF values. Zins et al. [197] use an attention-based encoder for multi-scale, multi-view feature combination. DG-Recon [63] uses four transformer decoder layers with cross-attention for feature fusion. Uni-3D [187] introduces a transformer for 2D depth-aware panoptic segmentation, providing strong 3D shape priors. Zhou et al. [193] introduce the Epipolar Feature Transformer (EFT) for low-resolution feature grids and RGB values. SparseNeRF [153] uses the DPT ViT as a pre-trained backbone on diverse depth datasets. Lin et al. [83] adopt the ViT-b16 model as a transformer encoder, generating latent features for further processing by convolutional components and performing self-attention on images.

Transformers face criticism for computational inefficiency due to quadratic complexity. Our literature survey found no methods using Transformers for city-scale scene reconstruction, suggesting this complexity deters their use at such scales.

While transformers aggregate feature maps, GRUs (Section VII-D) are used for similar tasks in SLAM-based 3D reconstruction. GRUs offer better scalability, resulting in faster and more efficient processing.

D. Gated Recurrent Units (GRU)

Gated Recurrent Units (GRU) [20] are recurrent neural networks that effectively extract features from sequential data. They excel in capturing long-term dependencies, crucial for maintaining temporal or spatial coherence in 3D reconstruction, and enable efficient processing and continuous integration.

In SLAM-related tasks, SurfelNeRF [39] uses a GRU module to fuse surfel data with global representations, enhancing spatial data refinement. R3D3 [118] employs a convolutional GRU to improve depth and pose estimation by refining flow corrections. ScanBot [11] uses a GRU layer for memory in reinforcement learning, enhancing decision-making in navigation and mapping.

E. Generative Models

Generative models are crucial in 3D reconstruction, synthesizing realistic and detailed structures from input data. They use various techniques to generate data distributions that capture the complexity and richness of real-world objects.

1) Diffusion Models (DM): Diffusion Models (DM), proposed by Sohl-Dickstein et al. [127], are inspired by statistical nonequilibrium thermodynamics. They introduce noise into data through an iterative forward process and learn the reverse diffusion process to restore original structures. This generative method is known for its flexibility and tractability.

Methods using diffusion models (DMs) enhance interactivity. LION [184] employs two DM models on hierarchical latent spaces for conditional synthesis and shape interpolation. SDFusion [19] trains a DM on compact latent representations of 3D shapes, allowing user control over reconstruction. DreamEditor [195] fine-tunes a Stable Diffusion model with sampled views, creating 3D editing regions from 2D masks. Melas-Kyriazi et al. [89] use a gradual diffusion process to project local image features onto a partially denoised point cloud, achieving high-resolution sparse geometries. These DM approaches are limited to single object or small-scale scene reconstructions.

2) Variational Autoencoder (VAE): Variational Autoencoders (VAEs) are used in 3D reconstruction, mapping input data into a latent space and decoding it back. Kingma and Welling [68] employ VAEs to maximize the likelihood of the original data distribution, generating diverse and realistic 3D models from sparse or incomplete data. This framework computes posterior distributions over latent variables, effectively modeling complex datasets.

Diffusion-based architectures like Lion [184] and SDFusion [19] integrate within a VAE framework to leverage compact latent representations, enhancing generative performance. AutoSDF [94] uses a Vector-Quantized VAE (VQ-VAE) to learn discrete latent representations, reducing the computational complexity of their Transformer model. Xu et al. [172] use the continuous latent space of VAEs to transform search and retrieval from a combinatorial challenge into a continuous optimization task.

3) GAN: Generative Adversarial Networks (GANs) are used in 3D reconstruction, with a generator creating new data samples and a discriminator distinguishing between real and generated data [44]. GANs generate realistic shapes from limited or noisy input, enhancing model fidelity. RoofGAN [108] produces structured roof models as graphs, with primitive geometry encoded as raster images at each node and inter-primitive relationships at the edges. Discriminators assess the quality and realism of the generated roof models, evaluating geometry composition and inter-primitive relationships.

VIII. Datasets

The availability of suitable datasets drives the development of data-driven methods. A good dataset contains large volumes of data that reflect the real world. We categorize the datasets analyzed by their scale: Object level VIII-A, Indoor Scenes VIII-B, Outdoor Scenes VIII-C, and a mixture of Indoor and Outdoor Scenes VIII-D.

A. Object Level

The DTU dataset: contains real single-object scenes with high-resolution images and point clouds generated by MVS [59]. Each scene includes 49 or 64 images and ground truth point clouds.

CO3D: contains 1.5 million frames from about 19,000 videos, including objects from 50 classes [111]. The dataset comprises multi-view images of real-world objects annotated with camera poses and ground truth 3D point clouds.

B. Indoor Scenes

The TUM: dataset is a small-scale indoor dataset with thirty-nine sequences of color and depth imagery captured in an office and an industrial hall [135]. It also provides ground truth camera poses from a motion capture system.

ScanNet: is an indoor scene dataset with 1513 RGB-D scans over 707 unique scenes, including camera parameters, dense surface reconstructions, textured meshes, CAD models, and semantically segmented scenes [24]. Data is acquired through RGB-D sensors and reconstructed as volumetric representations.

ScanNet++: enhances ScanNet with sub-millimeter resolution scans, registered megapixel images, and RGB-D streams from iPhones [178]. The 3D reconstructions include long-tail and label-ambiguous annotations.

7-Scenes: includes seven real indoor scenes captured by RGB-D sensors, with sequences ranging from 500 to 1,000 frames [92], [122]. KinectFusion generates TSDF volumes and camera positions.

The ARKitScenes: dataset includes about 5000 captures of 1661 unique scenes. It was the first to use the Apple LiDAR scanner for RGB-D data. The dataset provides ground truth data with registered RGB-D frames and oriented bounding boxes for various furniture types.

The NYU: dataset is large-scale, containing 1449 RGB-D images of 464 complex indoor scenes [124]. It provides per-pixel manual annotations for scene segmentation.

The ToyDesk: dataset includes two small-scale scenes, each with multi-view images capturing the entire scene. Toys often occlude each other depending on the view [173]. Using SfM [119], MVS [170], and 3D mesh generation [65], they generate camera poses and meshes. Target objects are manually labeled, and 2D instance segmentation is created from 3D labels. The manual annotations provide highly accurate and consistent ground truth.

Matterplot 3D: includes ninety large-scale building interiors, with 10800 panoramic views from 194400 RGB-D images, surface reconstructions, camera poses, and 2D/3D semantic segmentations [12]. RGB-D data are captured using a rig with three color and three depth cameras. 3D surfaces are generated using Poisson and voxel hashing methods [65], [99]. Camera poses are estimated via global bundle adjustment, and semantic labels are manually generated.

3D-Front: is a large-scale indoor dataset with 18968 rooms, each containing 3D objects. They use a 3D convolutional method for mesh generation [154] and TM-NET for texturing [38]. Semantic segmentations are generated using Structured3D [190].

SUN3D: is a large-scale dataset of RGB-D sequences of indoor scenes, including camera poses and object labels for each frame. Camera poses are estimated using SfM [164], with LabelMe-style annotations [115].

C. Outdoor Scenes

BlendedMVS: tackles the shortage of large-scale scenes with ground truth data for learning-based methods [175]. Using Altizure, they create high-quality 3D meshes and camera poses. These 3D models are rendered to each camera pose to produce images and depth maps.

EuRoC: dataset includes synchronized stereo images, IMU measurements, and ground truth point clouds [8]. Sensors, including a stereo camera, IMU, and laser scanner, are mounted on a hex-rotor vehicle.

The Waymo: dataset offers high-resolution synchronized data from cameras and LiDAR in various driving environments, including urban and suburban areas [137]. It features over 12 million 3D and 2D labeled bounding boxes, 3D road graph data, and semantic annotations for 23 common driving environment classes.

MegaDepth: uses multi-view Internet photograph collections to generate training data via SfM and MVS methods [80]. Dense MVS point clouds are re-projected to create depth maps for general-case training.

MatrixCity: uses the Unreal Engine 5 City Sample project to create a high-quality city-scale dataset [76]. It includes about 67 k aerial and 452 k street views with ground-truth camera poses, depth maps, and surface normals.

The UrbanScene3D: dataset includes 16 large-scale real and synthetic urban scenes with 128 k high-resolution images. It features high-precision LiDAR scans, and buildings have unique instance labels.

The Mill 19: dataset, created by Mega-NeRF authors for benchmarking [147], includes two large-scale scenes covering 0.125 km^2. “Building” features an industrial building exterior and environment, while “Rubble” contains a construction site.

Quad 6K: is a large-scale SfM dataset from Arts Quad, Cornell University, with 6514 images and ground truth camera poses. Around 5000 images have consumer GPS geotags, and 348 images have precise survey-quality GPS coordinates.

KITTI: is an outdoor large-scale dataset with six hours of recorded traffic scenarios. It includes high-resolution color and grayscale stereo imagery, point clouds, GPU/IMU data, and 3D tracklet annotations [41]. Sequences are captured using road vehicles equipped with two color and two grayscale cameras, a 3D laser scanner, and a GPU/IMU system. *KITTI-360* enhances this with 300K images, 80 k point cloud scans, geo-registered camera poses, and 3D annotations [81]. Images are captured using 180° fisheye cameras on the sides and a 90° perspective stereo camera at the front. Point clouds are generated with a pushbroom laser scanner on the vehicle roof. An online annotation tool based on WebGL is used for 3D annotations. *SemanticKITTI* [4] adds semantic point clouds, temporal sequences, GPS/IMU data, sensor poses, and range images to the KITTI Odometry Benchmark.

D. Indoor and Outdoor Scenes

Tanks and Temples: is a large-scale reconstruction dataset with both indoor and outdoor scenes [69]. It includes high-resolution video sequences from high-end cameras and ground truth point clouds from an industrial laser scanner.

The RealEstate10K: dataset includes about 80000 video clips from 10000 YouTube videos. SLAM and bundle adjustment algorithms compute camera positions and orientations along a trajectory [192].

The mip-NeRF 360: dataset includes nine large-scale, challenging scenes (five outdoor and four indoor). Each scene features a central object and detailed background, aiming to provide 360° views of an object. Camera poses are generated by COLMAP [119].

Synthetic NeRF: dataset was introduced with the NeRF reconstruction technique [93], using a physically-based renderer. It includes pathtraced images of eight objects with complex geometries and non-Lambertian materials, providing 300 views per scene.

IX. Evaluation Metrics

Evaluation metrics for 3D reconstruction quality are similar across methods and are classified into two categories: 3D and 2D. 3D metrics apply only to 3D representations, while 2D metrics, often used in image reconstruction, assess the quality of RFs and depth.

A. 3D Evaluation Metrics

Chamfer Distance (CD): measures distances between two point sets [28]. For each point in one set, it finds the nearest neighbor in the other and sums their squared distances.

Earth Mover's Distance (EMD): solves the assignment problem between two point sets [28]. For pairs of equal size, it computes the optimal bijection to minimize distances, making EMD differentiable. Due to its computational expense, EMD is approximated using a specialized relaxation scheme [5].

Normal Consistency (NC): measures the normal similarity between two meshes. It is defined as the mean absolute dot product of the normals in one mesh and the normals at the corresponding nearest neighbors in the other mesh [90].

Intersection over Union (IoU): is a well-known metric, also called the Jaccard Index, initially used for 2D object segmentation to measure overlap between prediction and ground truth regions. In 3D reconstruction, it measures similarity between two 3D shapes represented by binary occupancy maps. IoU is the volume of intersection over the volume of the union of these shapes, where intersection is the overlap and union includes all voxels in either shape [144].

B. 2D Evaluation Metrics

2D metrics, primarily used by NeRF-based methods, measure the quality of novel view synthesis relative to the original input images. They can also assess depth maps.

1) NeRF Related Metrics: Peak Signal-to-Noise Ratio (PSNR): calculates the logarithm of the ratio of the squared maximum pixel value to the Mean Squared Error (MSE) between images [37]. A higher PSNR indicates greater similarity to the reference image.

Structural Similarity Index Measure (SSIM): evaluates structural information in an image by normalizing luminance and contrast [160]. It compares normalized local pixel intensity patterns, separating similarity into luminance, contrast, and structure comparisons.

Learned Perceptual Image Patch Similarity (LPIPS): captures nuances of human perception [186]. It uses a pre-trained convolutional backbone to extract deep features from reference and novel images, assessing their perceptual similarity.

2) Depth Related Metrics: The following metrics assess depth accuracy and sometimes other representations like points [107]:

Absolute Relative Error (Abs Rel): measures the average relative difference between estimated and ground truth depths, considering the total number of camera positions [118].

Squared Relative Error (Sqr Rel): measures the squared relative difference between estimated and ground truth depths, leading to larger error values [118].

Root Mean Squared Error (RMSE): computes the square root of the averaged difference between estimated and ground truth depths, also considering the number of camera positions [118].

TABLE VI
Breakdown of Prior Knowledge in Recent 3D Methods

Prior Knowledge	**Model Awareness**	**Research Paper**
Semantic Priors	Semantics Extracted Using Backbones	[21], [56], [94], [187]
	Other Semantic Priors	[11], [171]
Geometry-based Priors	Depth Priors	[17], [63], [128], [133], [167]
	Depth and Normal Priors	[77], [194]
	Spatial and Temporal Priors	[9], [10], [16], [39], [42], [47], [66], [110], [118], [128], [158], [168], [196], [197]
	Smoothness Priors	[47], [71], [78]
	Other Geometric Priors	[52], [56], [94], [113], [136]

X. Prior Knowledge for Reconstruction

Guided by our Affinity Diagram, we recognize the need to explore the types of knowledge used in reconstruction. Prior knowledge often guides the process or imposes constraints, falling into two main categories: semantic and geometric. Semantic priors provide contextual information and spatial relationships, while geometric priors focus on shape characteristics like smoothness and symmetry. Table VI categorizes these semantic and geometric priors identified in the literature.

A. Semantic Priors

1) Semantics Extracted Using Backbones: Panoptic Semantics: 3D reconstruction can enhance its precision and detail by using insights from panoptic segmentation. This method creates a comprehensive segmentation map that includes both ’stuff’ (amorphous regions like grass, sky, road) and ’things’ (countable objects like cars, people, animals). Integrating a panoptic segmentation backbone transfers knowledge, improving 3D reconstruction.

BUOL [21] uses a three-branch decoder on its semantic backbone to produce panoptic semantic heads, estimate instance centers, and a three-plane occupancy. The 2D information is then elevated to 3D voxels, refined, and grouped based on the estimated instance centers. Uni-3D [187] inputs back-projected multi-scale features into a two-branch transformer for 2.5D depth estimation and 2D panoptic segmentation. A 3D U-Net converts these features into 3D segments with query embeddings, leveraging robust 2D priors. Nerflets [188] compares its backbone predictions with volume-rendered semantic logit pixels using a softmax-cross-entropy loss function. These semantic logits describe the object's category and assign an instance label for real-world identification. Huang et al. [55] generate semantic priors through a pre-trained backbone, organizing the reconstructed scene into semantic submaps and constructing a global pose graph with semantic links for consistent 3D reconstruction.

Other Segmentation Semantics: Xu et al. [172] use labelled meshes for urban areas, including buildings, ground, trees, and other elements, generating segmented point clouds. ScanBot [11] employs a semantic SLAM module to segment and reconstruct scenes, combining semantic priors with action decision modules to create strong shape priors. SUDS [148] performs unsupervised instance segmentation by inferring 3D representations for static and dynamic elements, obtaining 3D centroids via k-means and rendering instance predictions.

Language Semantics: The 3D reconstruction process, which captures the shape and appearance of real-world objects, can be enhanced by incorporating language priors. These priors, provided by language models, offer semantic embeddings that capture the context and meaning of words and phrases. This understanding guides the 3D reconstruction, resulting in more accurate and contextually relevant models.

ShapeClipper [56] introduces an example learning approach for semantically supervising 3D shape reconstruction by computing five semantic neighbors based on CLIP embeddings, providing rich context to guide the reconstruction process. NeRF-Art [152] uses a contrastive loss to strengthen stylization capabilities, aligning results closer to the target style while distinguishing them from others. AutoSDF [94] uses language priors to generate conditional distributions with their BERT module, providing instance-specific guidance for reconstruction. However, this strategy may not be optimal for large-scale, task-specific data, as it only approximates the joint distribution, highlighting challenges in applying language priors in 3D reconstruction.

B. Geometry-Based Priors

1) Depth: FineRecon [133] uses depth guidance, utilizing multi-view depth priors estimated by MVS to enhance reconstruction. DG-Recon [63] integrates depth priors from its monocular depth backbone to guide feature back projection to 3D space. These priors also compute occupancy predictions for 3D volume sparsification and provide auxiliary features to their cross-view fusion module. FrozenRecon [167] leverages a pre-trained monocular depth model to obtain scene geometry priors, using an affine-invariant depth model and optimizing sparse parameters for correcting depth maps, camera poses, and intrinsic parameters. SimpleNeRF [128] estimates depth through augmented models to supervise NeRF models. SparseNeRF [153] uses depth priors from real-world observations, distilling local depth ranking priors from a pre-trained model and employing a spatial continuity constraint to ensure spatial consistency with the pre-trained model.

2) Depth and Normal Priors: I^2-SDF [194] uses depth and normal priors for multi-view input images, allowing joint disassembly of shape, radiance, and material of an indoor scene. Depth information is recovered by combining each point's alpha value with accumulated transmittance, and normal values are estimated by computing the gradient of the SDF at any point. RICO [77] samples patches in an image and points along the rays of these patches. For occluded regions, it applies regularization to ensure smoothness of the rendered depth and normal of the background surface.

3) Spatial and Temporal Priors: Spatial priors involve learned information about a scene's spatial arrangements, used by methods receiving multi-view image inputs or existing 3D modalities like point clouds. Temporal priors relate to methods using image sequences, ensuring consistency between frames.

Cai et al. [9] generate depth maps using spatial consistency of video inputs with their plane sweep stereo module. R3D3 [118] extracts spatial and spatial-temporal edges using three priors: calibration camera proximity, presumed forward motion, and observer motion constraints. Zins et al. [197] propose an energy function to evaluate photo-consistency along viewing lines across multiple depth maps.

4) Camera Priors: Camera priors are crucial for most NeRF-based methods [10], [16], [39], [42], [47], [74], [110], [128], [143], [158], [168], [196]. The CamP [104] method starts with camera poses estimated by COLMAP. 3DGS [66] use SfM to generate sparse point clouds and calibrated camera parameters but only operate on unoriented side-product point clouds. ViP-NeRF [129] uses a plane sweep to compute a visibility prior, regularizing NeRF training with sparse inputs and reformulating the NeRF MLP to output visibility.

5) Smoothness Priors: VMesh [47] computes a smoothness prior by predicting normal values close to the analytical normal value. Hewa Koneputugodag et al. [71] minimize a smoothness constraint to handle missing regions effectively. Neuralangelo [78] introduces smoothness by regularizing the mean curvature of the reconstructed SDF. SUDS [148] ensures smoothness in flow across space and time by regularizing sampled 3D points along a ray, promoting piecewise linear trajectories [79].

6) Other Geometric Priors: Manhattan-World Assumption: The Manhattan-World assumption imposes perpendicularity between three planes in a scene. SOL-NeRF [136] enforces a relaxed version of this assumption, allowing some plane inclination for more accurate results, along with Sunlight color priors. ManhattanFusion [177] uses the Manhattan-World assumption to generate keyframes for planar re-alignment and guide depth registration. These keyframes also optimize camera poses and minimize local drift errors, fusing each reconstructed segment into a global volume.

Planar Prior: To reduce reliance on the Manhattan-world assumption for indoor scene reconstruction, Gong et al. [43] propose the planar prior. This prior relates to plane orientation in each frame and is generated through a plane detection algorithm.

Symmetry Constraint: Besides its semantic neighbor strategy, ShapeClipper [56] uses the symmetry constraint to produce uniform view priors. These priors regularize viewpoint learning, preventing the degeneration of rotation estimation.

Autoregressive Prior: AutoSDF [94] proposes a novel non-sequential autoregressive prior over a compressed discrete representation. This prior is learned by a Transformer model that receives a discretized latent, representing the high-dimensional 3D shape in lower dimensions to reduce complexity. This implementation overcomes the need for temporal information.

Global Shape Priors: VolRecon [113] concatenates multi-view image-based projection features with volume features interpolated from a global feature volume. This results in a combined feature that adds a global shape prior.

Face Prior: Huang et al. with City3D [52] employ the face prior constraint, ensuring the highest confidence planar surface is always selected as a prior for the rest of the building's reconstruction.

Line-of-sight Priors: The line-of-sight assumption states that for any point observed by a LiDAR sensor, there is a corresponding location on a non-transparent surface where the atmospheric media does not affect the color measure with respect to a LiDAR ray [112].

Geometry Confidence: Geometry confidence measures the consistency of depth and flow across different views by projecting depth pixels into target views [166].

Reprojection Confidence: Reprojection confidence measures the accuracy of depth maps and identifies outliers. Similar to the geometry confidence constraint, pixels from the source image are reprojected to target views to measure projection confidence at the pixel, patch, and feature levels [166].

XI. Human-in-the-Loop and Interactivity

A key requirement for future AI representations is the Human-in-the-Loop initiative, as shown in our Affinity Diagram. Our exploration includes both methods that investigate different user inputs and interactivity mechanisms with 3D representations and methods that offer means for interactivity within their reconstruction technique.

A. Interactive Mechanisms for Engaging With 3D Representations

Eye tracking is crucial for 3D representation interactivity. Kollert et al. [70] map eye fixations onto a 3D point cloud to visualize user gaze in urban scenes. Singh et al. [125] gather gaze data in real-world scenes but do not optimize 3D representations with it. Min Yoon et al. [180] use eye movements for orientation changes and zooming, though sustained blinking can be tiring in industrial settings. For more, see the survey [138].

Haptic interactivity is another crucial interactivity mechanism. Mulumba et al. [97] use the Geomagic Touch device for real-time interaction with reconstructed objects, providing haptic feedback based on on-screen selections. Tian et al. [146] use a haptic device to deform a 3D point cloud for large-scale reconstructions.

Human-AI interaction can still use standard point-and-click navigation. Park [103] allows users to select editable 3D primitives to improve scene geometry, helping escape local minima for better solutions. Specifically for urban scenes, Kim and Han [67] offer a semi-automated 3D reconstruction pipeline for buildings, requiring minimal user input to identify building footprints.

B. Interactivity Within 3D Reconstruction Methods

Few studies integrate human subjects into reconstruction workflows. Some, like [56], [94], [184], [195], use natural language for interactivity. In [56], [94], [184], language-guided 3D generation is used, with [184] and [56] incorporating mesh texturization via Text2Mesh [91]. LION [184] offers shape interpolations, SDFusion [19] focuses on text-guided shape completion, and DreamEditor [195] uses text to identify and edit specific regions.

While language-based interactivity is common in 3D reconstruction, other methods also enhance user engagement. Neuralangelo [78] uses commercial software for loading point clouds and selecting regions for reconstruction, offering a user-friendly interface. Yuan et al. [182] enable mesh deformation with box abstractions as deformation handles, using ARAP [131] for control points to edit the mesh.

Another approach, as described in [71], is algorithmic interactivity, where users modify octree labels to guide reconstruction, offering fine-grained control. VQ-NeRF [191] provides an interface for selecting regions and new textures from a database. SeamlessNeRF [42] stitches 3D objects represented by radiance fields into a homogeneous scene.

XII. Privacy and Security in 3D Representations

In this section, we examine methods authors use to address privacy and security concerns in urban scene data, focusing on anonymization techniques to protect personal and spatial privacy. Privacy and security issues in 3D representations include removing sensitive infrastructure for national security and anonymizing sensitive information in urban scenes, like faces, building numbers, and vehicle plates. This subsection discusses literature addressing these concerns.

Chelani et al. [13] explore point cloud to line cloud transformations for privacy preservation. Despite line clouds being unintelligible to humans, they show that the original point cloud can be approximated. Suzuki and Teppei [139] introduce a Federated Learning (FL) with 3DGS method. While FL aims to create a lighter, scalable 3DGS model, it also enhances 3D reconstruction privacy by storing data locally and sharing only model weights globally.

Frome et al. [33] developed an early method to remove private information from large-scale streetview data, including face blurring and vehicle license plates. This was an initial step in Google Street View's privacy measures. Today, users can also request the deletion of their private domicile's 360-view.

XIII. Discussion and Future Work

A. Data Availability

Images: Data availability and quality are crucial for model architecture design. Many methods use multi-view images for 3D scene representations. NeRF-based approaches need extensive multi-view image datasets for high-quality radiance-based reconstructions, despite being computationally intensive. MERF [110] uses multi-resolution hash encodings by Müller et al. [96] for real-time inference and faster training. SimpleNeRF [128] employs few-shot learning to reduce dataset size but adds computational overhead, affecting scalability. ViP-NeRF [129] uses sparse images and MLP for visibility, reducing training times. Some methods use pre-trained backbones [19], [56], [187], needing only one image for reconstructions.

Usually, NeRF-based methods use synthetic or idealized datasets. Researchers should develop datasets that reflect real-world sensing challenges to enable industrial applications.

Sequential images are often used to reconstruct TSDF-based surfaces [31], [63], [94], [133]. TSDFs are lighter than other implicit representations like Radial Basis Functions (RBFs) and SDFs, making them suitable for real-time reconstruction. Sequential images mimic an agent's path, ideal for methods where autonomous vehicles or robots collect data over time.

Noise in sequential image datasets harms reconstruction accuracy, especially due to dynamic object artefacts. Spatial and temporal priors are crucial for frame consistency and robust reconstruction. For large-scale scenes, multi-modal input methods enforce spatial consistency without heavily relying on temporal elements [112].

Depth: SLAM methods use depth information to improve reconstruction. They fall into two categories: sensing and optimizing depth [189], and directly estimating depth [9]. However, solely relying on sensed depth without refinement is unreliable.

Depth is a crucial geometric prior, recognized for its efficiency and cost-effectiveness in supervision. Strategies for direct depth supervision include iterative optimization of depth maps, as seen in SLAM methods, and focusing on regions with high depth accuracy, like SimpleNeRF [128]. Depth is popular in semi-supervised methods [10], [77], [118]. Some methods [10], [118] optimize depth directly, while others [77] combine depth with point normals to smooth occluded regions.

Depth is used as both a primary and intermediate modality in supervised and semi-supervised methods. However, methods relying on depth are susceptible to imperfect data. SureflNeRF [39] faces geometry discontinuities, and FineRecon [133] misses local fine structures. RICO [77] addresses this with a regularization process for smoothness and continuity in occluded regions. SimpleNeRF [128] uses ternary masks to select accurate depth estimates when using synthetic depth for their weakly-supervised approach.

Future work using depth to learn 3D structures should address missing or occluded regions better. Depth-based methods often struggle with thin structures and intricate details in complex scenes. While they excel in real-time reconstructions from sequential data, overcoming these limitations without multi-view sets is crucial, as they may not always be available. Incorporating semantics in a depth-based pipeline shows promise, but large occluded areas remain problematic [187]. Exploring generative methods to reconstruct these occluded areas while maintaining high fidelity would be particularly interesting.

Point Clouds: Point clouds significantly enhance volumetric representations in reconstruction pipelines due to their 3D expressiveness. Some methods focus on unoriented point clouds. Hewa Koneputugodage et al. [71] use a leave-labelling strategy with human input, while [161] employs a vanishing gradient alleviation strategy. 3DGS [66] use SfM-generated points to create 3D Gaussians, approximating volumetric properties and reducing inference and training times.

Point cloud approaches generally show scalability [6], [11], [66], high reconstruction quality [66], and fidelity [161]. However, none address all these aspects simultaneously, which is crucial for industrial applications. POCO [6] offers high reconstruction quality at competitive speeds and supports domain shifting, learning objects, and operating in diverse scenes. Extending methods like POCO is essential for handling large/city-scale 3D environments.

While point clouds are beneficial for 3D reconstruction due to their structure, their storage-intensive nature can limit efficiency in large-scale scenarios. Like depth, point clouds can be generated using Commercial-off-the-Shelf (COTS) sensors or reconstruction methods. Exploring point clouds as a complementary modality for neural reconstruction methods holds promise [112], [148], [166]. NeRF-based [112], [166] have shown similar fusion frameworks for supervising NeRF through LiDAR scans for urban and outdoor scenes to accommodate poorly observed regions. In addition S-NERF [166] has achieved the same performance while using sparse and imperfect scans.

A future direction involves integrating point clouds with neural reconstruction architectures for large-scale scenes. For city-scale methods we observe a reliance on multiple GPUs to achieve fast NeRF inference. The computationally efficient 3DGS approach [66] presents a particularly compelling direction for neural reconstruction, as it naturally relies on sparse point clouds for 3D Gaussian propagation. Furthermore, given its inherently lightweight design compared to NeRF, 3DGS has the potential to serve as a valuable framework to address the multi-GPU requirements that are a major challenge in city-scale neural methods.

Multi-Modal Inputs: Future research should focus on developing methods to handle multi-modal and multi-scale data. Industrial organizations have vast repositories of diverse data, such as street-level views, aerial drone views, and satellite views. Effective methods that learn from these various scales and modalities are crucial for enhancing 3D reconstruction efficiency in industrial workflows. Integrating diverse data sources (e.g., street-level point clouds, urban annotations, images, traffic, demographic, and environmental data) through data fusion-centric methodologies [198] can streamline operations and enhance data completeness and coherence in industrial settings.

B. Common Issues in 3D Reconstruction

Reconstructed surfaces often have qualitative imperfections, such as handling thin structures and recovering fine details [19], [118], [133], [176]. Challenges also include accurately representing specular and transparent objects [9], [47], [128], [129], [136], [176], effectively handling occluded regions [9], [66], [161], [187], and realistically modeling illumination effects [42], [110], [136], [195].

Thin structure handling: is challenging for methods like R3D3 [118] and FineRecon [133] due to training limitations and deep network downsampling. LoD-NeUS [196] addresses these issues with tri-plane positional encodings and error-guided sampling, capturing multiscale features and preserving surfaces. VMesh [47] excels in capturing thin structures using volumetric primitives for robust detail recovery, unlike traditional mesh representations.

Specular/Transparent objects: LoD-NeUS [196] addresses specular and transparent object challenges with error-guided sampling. VMESH [47] uses a hybrid volume and mesh representation but may inaccurately represent transparent and furry objects, embedding them into the mesh instead of the volumetric representation. ViP-NeRF [129] might miss significant color changes on specular surfaces, especially with drastic viewpoint variations.

Occluded Regions: SimpleNeRF [128] struggles with high reprojection errors in complex regions. SOL-NeRF [136] focuses on non-specular surfaces, while methods like [9] face occlusions from multiple transparent objects. 3DGS [66] introduce artifacts in poorly observed regions, and both Uni3D [187] and Wang et al. [161] have issues with missing or occluded large areas. RICO [77] recovers occluded backgrounds in indoor scenes using regularization strategies that exploit foreground-background relationships and promote smoothness in unseen regions. Image-based methods with extensive datasets are generally less susceptible to these challenges, as they capture detailed scene information effectively.

Illumination Modelling: DreamEditor [195] and Seamless-NeRF [42] lack control over lighting conditions, relying on matching source and target lighting. SOL-NeRF [136] struggles with light emitters like streetlights.

The challenges in 3D reconstruction methods are diverse and complex. Many issues can be mitigated by using rich datasets with diverse scene views. Image-based methods with extensive datasets are generally less susceptible to these difficulties, though such datasets may not always be available. Future research could explore adapting approaches like RICO [77], which uses semantic relationships to regularize less observed regions, for outdoor settings. Hybrid strategies like VMesh [47] show promise in addressing geometric challenges like thin structure modeling and editing implicit representations.

C. Camera Pose Estimation

While all data-driven methods need high-quality data, NeRF-based approaches often credit their success to accurate and robust camera parameters. However, these parameters are not always available, requiring some methods to use SfM or other estimation techniques. Additionally, sequential methods using depth information [32], [133] also rely on precise camera parameters for effective reconstruction.

To enhance future datasets for NeRF-based methods, we recommend prioritizing providing ground-truth camera poses and parameters. This ensures accuracy, reduces computational costs for parameter estimation, and allows researchers to focus on inference optimization, training efficiency, or maintaining quality for large-scale urban applications. CamP [104] introduces a preconditioning method to optimize camera parameters within the NeRF framework, significantly improving 3D reconstruction quality. However, it still relies on SfM-based COLMAP [119] for initializing unknown camera parameters.

Generating and optimizing camera positions is crucial for neural reconstruction. While traditional SfM is commonly used, other strategies can be incorporated into novel 3D reconstruction methods. PlaneFusion [43] optimizes camera poses between submaps through a global optimization scheme, while ManhattanFusion [177] uses a local optimization strategy, registering depth keyframes to Manhattan keyframes. Future research could explore incorporating these optimization methodologies within deep learning techniques to enhance the accuracy of neural 3D representations.

D. Issues in Large-Scale Reconstruction

Scalability concerns are common in 3D reconstruction pipelines, hindering industrial adoption. These issues include memory constraints, handling large datasets, and training efficiency limitations.

Most large-scale outdoor methods lack city-scale capabilities [54], [78], [112], [118], [163], [166], [188]. Although metrics like model capacity and training times are often omitted, methods using point cloud representation with multi-view images, such as [112], [166], can adapt to a BlockNeRF [143] scheme, utilizing multiple GPUs, each with its own NeRF model, for city-scale scenes. NeRF-XL [74] extends this by disjoining individual NeRF blocks, training different spatial locations independently on each GPU to maximize efficiency and eliminate blending errors. SUDS [148] and GridNeRF [169] create city-scale RF representations using multi-scale features instead of multi-GPU schemes. SUDS [148] uses a multi-resolution hash table architecture to distinguish static from dynamic elements, leveraging rich data like RGB imagery, LiDAR, and optical flow. GridNeRF [169] employs a two-branch NeRF approach, with one branch capturing the scene coarsely and the other storing high-frequency information. However, they face long training and inference times and may need to use BlockNeRF for faster inference in very large scenes.

We highlight the significant storage required by city-scale methods. NeRF-XL [74] uses 258K images from the Matrix-City [76] dataset to reconstruct 25 square kilometers. Grid-NeRF [169] captures over 5K drone images for a 2.7 square kilometer scene. SUDS [148] uses 1.2 million frames from 1.7 k videos, requiring 20 TB of storage after compression. Current city-scale reconstruction methods need excessive storage for both input data and 3D models, in addition to high-end and often numerous GPUs required to process and learn from these extensive datasets.

Optimization schemes like [52], [106] achieve city-scale reconstructions efficiently on consumer hardware but have limitations compared to learning-based methods. City3D [52] provides low LoD reconstructions and is limited to identifiable buildings. In contrast, [106] captures fine details but requires very high-resolution inputs, with reconstruction quality dependent on input resolution. Both methods lack appearance modeling, a significant advantage of neural methods.

While effective, GPU cluster-based city-scale reconstruction methods are hardware-limited, cost-ineffective, and compute-intensive, especially for nationwide geospatial needs. Future research could leverage transfer learning for city-scale+ scenes. Techniques like knowledge distillation [148] and fine-tuning enable smaller models to approximate larger ones, enhancing neural inference capabilities. Additional data or priors could improve accuracy in poorly reconstructed areas. Many methods lack resource utilization metrics when evaluating 3D reconstruction techniques. Standardizing metrics like GPU utilization, memory usage, and computational efficiency in future publications would enhance transparency and facilitate better comparisons.

E. Human-in-the-Loop

Few studies address human interactivity in 3D reconstruction. Textual input methods often struggle with industrial demands and scalability in large or complex scenes [19], [56]. DreamEditor [195] allows region selection and editing based on human inputs but can be inefficient for complex scenes. Hewa Koneputugodage et al. [71] introduce point cloud interaction, but it's limited to very small scenes due to intricate editing. While these methods lay the groundwork for future interactions, they are not ready for large-scale deployment. Future research should explore mixed-interactivity modes for efficient and intuitive human interaction with large-scale 3D representations.

Human roles in quality assurance are notably absent in the literature, especially for large-scale reconstruction methods. These methods could benefit from algorithms learning through human guidance or integrating humans into the pipeline to reflect industrial workflows. This highlights the challenges in effectively integrating human interaction within 3D reconstruction methodologies.

Future research should identify effective human integration points in reconstruction pipelines to enhance accuracy, fidelity, quality, and scalability. Ethnographic evaluations of 3D experts' interactions with data modalities and reconstruction methods are recommended to find seamless integration points for human behavior in workflows. These studies would also reveal which models or architectures are most compatible with human interaction.

This research could develop human-in-the-loop methods leveraging 3D experts' experiences for intuitive interactions in current workflows. Exploring reinforcement learning methodologies, like hybrid approaches such as VMesh [47], could integrate human agents into the reconstruction process, enabling learned corrections and improvements. Future research should investigate semantics in 3D reconstruction workflows, such as how defining a building's era (e.g., Victorian) might influence accuracy, even with modern renovations. The intricacies of language semantics combined with interactive modes like region selection should be explored for large-scale applications.

F. Industry versus Academia

Preference for Explicit Representations: The industry prefers explicit representations for their quantifiable nature and ease of editing. Two approaches [176], [195] show that hybrid representations can enhance the editability of non-editable implicit representations. However, BakedSDF [176] is limited to editable appearance and illumination. Methods like [47], [196] demonstrate how hybrid representations can reduce computational complexity and improve the recovery of challenging details like thin structures. This raises questions: How can hybrid representations leverage the strengths of both explicit and implicit representations? What are the most reliable methods for establishing robust correspondence between implicit and explicit representations? Which methods offer the most efficient and intuitive interaction for editability? These questions highlight the challenges and opportunities in developing hybrid representations that combine the benefits of both explicit and implicit approaches in 3D reconstruction and editing workflows.

Terrain Models: There is a gap in the literature on generating terrain models and integrating reconstructed surfaces onto terrain. Existing research often focuses on aspects like glacial erosion [22] or considers structures and city furniture as part of the terrain [156]. Industry stakeholders emphasize the importance of terrain model generation and integrating reconstructed objects onto these models for applications such as urban planning, environmental modeling, infrastructure development, simulations, and accurate placement of reconstructed objects in their geographical context.

Terrain modeling is crucial for multi-modal large-scale urban reconstructions, addressing challenges like reducing noise from artifacts, thin structures, transparent or specular objects, and sensor interference. Starting with terrain identification and modeling, a systematic approach can manage street objects and other noisy sources layer by layer, considering their spatial and semantic relationship with the terrain. This process can yield precise, detailed, and layered 3D reconstructions for industrial applications.

Semantic and Geometric-based Priors: Incorporating semantic understanding and scene comprehension is crucial for unlocking numerous applications of 3D assets. Several methods use pre-trained backbones to grasp semantic features within their scenes [21], [56], [94], [187]. Leveraging these backbones enriches features and enhances scene understanding.

The ShapeClipper method [56] constructs a semantic neighborhood around objects to guide generative reconstruction effectively. This approach is valuable for tasks like removing unwanted urban scene elements (e.g., transparent objects) and predicting the structure beneath them. Additionally, Nerflets [188] introduces a method that stores semantic information and tracks objects within scenes, relying on SOTA encoders for success.

Recent literature often employs depth as a modality for geometric exploitation. Depth serves well as an input and guides learning-based methods. It is also used diversely as a prior, such as in SparseNeRF [153], which uses inaccurate depth maps to generate effective depth priors, overcoming data scarcity and sensor limitations. Another notable method [71] imposes smoothness priors as constraints on occluded regions. Beyond depth, the AutoSDF [94] method is intriguing, transforming autoregressive priors for use in multi-view scenarios.

The Manhattan-World Assumption may have limited utility in urban scenes lacking strict 3-plane perpendicularity. Manhattan's grid layout was planned for rapid growth, accommodating horse-drawn carriages and later motor vehicles. In contrast, cities like London, developed over centuries with roads initially designed for horse carriages, have narrower, winding streets and a less uniform layout. This historical and country-wise context influences how cities adapt their infrastructure over time, alongside diverse architectural styles and urban asymmetry. Addressing these complexities in future constraints should prioritize semantic-based approaches that consider historical context to better capture their diversity.

Employing multi-scale methods enhances industrial functionality. Multi-scale reconstructions offer scalability, with lower detail levels being lighter and sufficient for some applications. However, a low LoD reconstruction may be insufficient.

LoD-NeUS [196] aggregates features from multiple LoDs, focusing on high-quality reconstructions and addressing aliasing issues. Currently, no methods can simultaneously reconstruct scenes at various LoDs. While such a pipeline may seem cost-prohibitive, insights from our Affinity Diagram suggest industrial applications could benefit from varying LoD reconstructions. This approach allows finer details to adjust based on visibility or smoothness criteria, integrating well with Human-in-the-loop initiatives proposed for industrial pipelines. One strategy involves human assessment of a coarse model's quality, informing an upscale to a high-detail representation.

XIV. Conclusion

Our survey assessed state-of-the-art literature on 3D reconstruction, focusing on its relevance to industrial needs for large-scale urban environments. Guided by the KJ method, we organized qualitative data from industrial partners into an Affinity Diagram. Besides standard topics like input modalities, reconstruction methods, and evaluation metrics, our Affinity Diagram provided valuable insights.

We expanded our analysis to include discussions on prior information guiding reconstruction methods, human-in-the-loop initiatives, and interaction modes with 3D urban products. We also explored implications for future research on large-scale urban environments. Key themes include strategies for future dataset creation, data priors or constraints for large-scale urban 3D fabrication, and opportunities for integrating human-centered reconstruction methods into urban planning and development workflows.

Limitations of Our Study: Our survey has benefited significantly from collaboration with Ordnance Survey, a leading organization with extensive experience in managing large-scale geospatial data and addressing national-level requirements for city- and country-scale applications. Their expertise provided us with invaluable insights onto the challenges and advancements of 3D reconstruction tailored to urban environments, grounding our analysis in real-world, high-impact applications. However, a limitation of this survey is that the qualitative data informing our AD were primarily drawn from Ordnance Survey. While this provided a rich and focused perspective, it may limit the breadth of applicability of our findings to other organizations or contexts with different operational priorities. To address this, we aim to extend our study in the future by engaging with a wider range of companies and stakeholders, thereby capturing more diverse perspectives and enhancing the comprehensiveness of our survey.

Acknowledgment

For the purpose of open access the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

References

[1] S. Agarwal et al., "Building Rome in a day," *Commun. ACM*, vol. 54, no. 10, pp. 105–112, 2011.

[2] S. Bai and J. Li, "Progress and prospects in 3D generative AI: A technical overview including 3D human," 2024, *arXiv:2401.02620*.

[3] A. Basiri, T. Lines, and M. F. Pereira, "Scalable 3D mapping of cities using computer vision and signals of opportunity," *Int. J. Geographical Inf. Sci.*, vol. 37, no. 7, pp. 1470–1495, 2023.

[4] J. Behley et al., "SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9297–9307.

[5] D. P. Bertsekas, "A distributed asynchronous relaxation algorithm for the assignment problem," in *Proc. 24th IEEE Conf. Decis. Control*, 1985, pp. 1703–1704.

[6] A. Boulch and R. Marlet, "POCO: Point convolution for surface reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 6302–6314.

[7] A. Boulch, G. Puy, and R. Marlet, "FKAConv: Feature-kernel alignment for point cloud convolution," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 381–399.

[8] M. Burri et al., "The EuRoC micro aerial vehicle datasets," *Int. J. Robot. Res.*, vol. 35, no. 10, pp. 1157–1163, 2016.

[9] Y. Cai, Y. Zhu, H. Zhang, and B. Ren, "Consistent depth prediction for transparent object reconstruction from RGB-D camera," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 3459–3468.

[10] A.-Q. Cao and R. de Charette, "SceneRF: Self-supervised monocular 3D scene reconstruction with radiance fields," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 9387–9398.

[11] H. Cao, X. Xia, G. Wu, R. Hu, and L. Liu, "ScanBot: Autonomous reconstruction via deep reinforcement learning," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 157–1, 2023.

[12] A. Chang et al., "Matterport3D: Learning from RGB-D data in indoor environments," 2017, *arXiv: 1709.06158*.

[13] K. Chelani, F. Kahl, and T. Sattler, "How privacy-preserving are line clouds? Recovering scene details from 3D lines," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15668–15678.

[14] A. Chen, Z. Xu, X. Wei, S. Tang, H. Su, and A. Geiger, "Dictionary fields: Learning a neural basis decomposition," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 1–12, 2023.

[15] D. Chen, Z. Tang, Z. Xu, Y. Zheng, and Y. Liu, "Gaussian fusion: Accurate 3D reconstruction via geometry-guided displacement interpolation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 5916–5925.

[16] Y. Chen and G. H. Lee, "SCALAR-NeRF: Scalable large-scale neural radiance fields for scene reconstruction," 2023, *arXiv:2311.16657*.

[17] Z. Chen, C. Wang, Y.-C. Guo, and S.-H. Zhang, "StructNeRF: Neural radiance fields for indoor scenes with structural hints," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 15694–15705, Dec. 2023.

[18] B. Cheng et al., "Panoptic-DeepLab: A simple, strong, and fast baseline for bottom-up panoptic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12475–12485.

[19] Y.-C. Cheng, H.-Y. Lee, S. Tulyakov, A. G. Schwing, and L.-Y. Gui, "SDFusion: Multimodal 3D shape completion, reconstruction, and generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 4456–4465.

[20] K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014, *arXiv:1406.1078*.

[21] T. Chu, P. Zhang, Q. Liu, and J. Wang, "BUOL: A bottom-up framework with occupancy-aware lifting for panoptic 3D scene reconstruction from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 4937–4946.

[22] G. Cordonnier et al., "Forming terrains by glacial erosion," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 1–14, 2023.

[23] H. A. Correia and J. H. Brito, "3D reconstruction of human bodies from single-view and multi-view images: A systematic review," *Comput. Methods Programs Biomed.*, vol. 239, 2023, Art. no. 107620.

[24] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3D reconstructions of indoor scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5828–5839.

[25] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv: 2010.11929*.

[26] F. E. Fadzli, A. W. Ismail, and S. Abd Karim Ishigaki, "A systematic literature review: Real-time 3D reconstruction method for telepresence system," *PLoS One*, vol. 18, no. 11, 2023, Art. no. e0287155.

[27] G. Fahim, K. Amin, and S. Zarif, "Single-view 3D reconstruction: A survey of deep learning methods," *Comput. Graph.*, vol. 94, pp. 164–190, 2021.

[28] H. Fan, H. Su, and L. J. Guibas, "A point set generation network for 3D object reconstruction from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 605–613.

[29] A. Farshian et al., "Deep-learning-based 3-D surface reconstruction—A survey," *Proc. IEEE*, vol. 111, no. 11, pp. 1464–1501, Nov. 2023.

[30] H. Fathi, F. Dai, and M. Lourakis, "Automated as-built 3D reconstruction of civil infrastructure using computer vision: Achievements, opportunities, and challenges," *Adv. Eng. Inform.*, vol. 29, no. 2, pp. 149–161, 2015.

[31] Z. Feng, L. Yang, P. Guo, and B. Li, "CVRecon: Rethinking 3D geometric feature learning for neural reconstruction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 17750–17760.

[32] L. Fink, D. Rückert, L. Franke, J. Keinert, and M. Stamminger, "LiveNVS: Neural view synthesis on live RGB-D streams," in *Proc. SIGGRAPH Asia Conf. Papers*, 2023, pp. 1–11.

[33] A. Frome et al., "Large-scale privacy protection in Google street view," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2009, pp. 2373–2380.

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

[35] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski, "Towards internet-scale multi-view stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 1434–1441.

[36] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1362–1376, Aug. 2010.

[37] K. Gao, Y. Gao, H. He, D. Lu, L. Xu, and J. Li, "NeRF: Neural radiance field in 3D vision, a comprehensive review," 2022, *arXiv:2210.00379*.

[38] L. Gao, T. Wu, Y.-J. Yuan, M.-X. Lin, Y.-K. Lai, and H. Zhang, "TM-NET: Deep generative networks for textured meshes," *ACM Trans. Graph.*, vol. 40, no. 6, pp. 1–15, 2021.

[39] Y. Gao, Y.-P. Cao, and Y. Shan, "SurfelNeRF: Neural surfel radiance fields for online photorealistic reconstruction of indoor scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 108–118.

[40] X. J. Ge, P. Livesey, J. Wang, S. Huang, X. He, and C. Zhang, "Deconstruction waste management through 3D reconstruction and BIM: A case study," *Vis. Eng.*, vol. 5, no. 1, pp. 1–15, 2017.

[41] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.

[42] B. Gong, Y. Wang, X. Han, and Q. Dou, "SeamlessNeRF: Stitching part NeRFs with gradient propagation," in *Proc. SIGGRAPH Asia Conf. Papers*, 2023, pp. 1–10.

[43] B. Gong, Z. Zhu, C. Yan, Z. Shi, and F. Xu, "PlaneFusion: Real-time indoor scene reconstruction with planar prior," *IEEE Trans. Vis. Comput. Graph.*, vol. 28, no. 12, pp. 4671–4684, Dec. 2022.

[44] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[45] S. Grunwald and P. Barak, "3D geographic reconstruction and visualization techniques applied to land resource management," *Trans. GIS*, vol. 7, no. 2, pp. 231–241, 2003.

[46] J. Guo, Y. Liu, X. Song, H. Liu, X. Zhang, and Z. Cheng, "Line-based 3D building abstraction and polygonal surface reconstruction from images," *IEEE Trans. Vis. Comput. Graph.*, vol. 30, no. 7, pp. 3283–3297, Jul. 2024.

[47] Y.-C. Guo, Y.-P. Cao, C. Wang, Y. He, Y. Shan, and S.-H. Zhang, "VMesh: Hybrid volume-mesh representation for efficient view synthesis," in *Proc. SIGGRAPH Asia Conf. Papers*, 2023, pp. 1–11.

[48] L. Han, S. Gu, D. Zhong, S. Quan, and L. Fang, "Real-time globally consistent dense 3D reconstruction with online texturing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1519–1533, Mar. 2022.

[49] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," in *Proc. Asian Conf. Comput. Vis.*, Springer, 2017, pp. 213–228.

[50] Z. He and H. Peng, "Research on user experience design based on affinity diagram assisting user modeling–taking music software as an example," in *Proc. Int. Conf. Hum. Comput. Interact.*, Springer, 2023, pp. 516–526.

[51] K. Hu et al., "Overview of underwater 3D reconstruction technology based on optical images," *J. Mar. Sci. Eng.*, vol. 11, no. 5, 2023, Art. no. 949.

[52] J. Huang, J. Stoter, R. Peters, and L. Nan, "City3D: Large-scale building reconstruction from airborne LiDAR point clouds," *Remote Sens.*, vol. 14, no. 9, 2022, Art. no. 2254.

[53] P.-H. Huang, K. Matzen, J. Kopf, N. Ahuja, and J.-B. Huang, "DeepMVS: Learning multi-view stereopsis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2821–2830.

[54] S. Huang et al., "Neural LiDAR fields for novel view synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 18236–18246.

[55] S.-S. Huang, H. Chen, J. Huang, H. Fu, and S.-M. Hu, "Real-time globally consistent 3D reconstruction with semantic priors," *IEEE Trans. Vis. Comput. Graph.*, vol. 29, no. 4, pp. 1977–1991, Apr. 2023.

[56] Z. Huang, V. Jampani, A. Thai, Y. Li, S. Stojanov, and J. M. Rehg, "ShapeClipper: Scalable 3D shape learning from single-view images via geometric and clip-based consistency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 12912–12922.

[57] C. Institute for Manufacturing, "Affinity charting," 2023. Accessed: May 19, 2024. [Online]. Available: https://www.ifm.eng.cam.ac.uk/research/decision-support-tools/affinity-charting/

[58] S. Isaacson, P.-C. Kung, M. Ramanagopal, R. Vasudevan, and K. A. Skinner, "LONER: LiDAR only neural representations for real-time SLAM," *IEEE Trans. Robot. Autom.*, vol. 8, no. 12, pp. 8042–8049, Dec. 2023.

[59] R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, and H. Aanæs, "Large scale multi-view stereopsis evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 406–413.

[60] M. Ji, J. Gall, H. Zheng, Y. Liu, and L. Fang, "SurfaceNet: An end-to-end 3D neural network for multiview stereopsis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017, pp. 2307–2315.

[61] S. Jiang, J. Hua, and Z. Han, "Coordinate quantized neural implicit representations for multi-view reconstruction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 18358–18369.

[62] T. Jokela and A. Lucero, "MixedNotes: A digital tool to prepare physical notes for affinity diagramming," in *Proc. 18th Int. Academic MindTrek Conf. Media Bus. Manage. Content Serv.*, 2014, pp. 3–6.

[63] J. Ju, C. W. Tseng, O. Bailo, G. Dikov, and M. Ghafoorian, "DG-Recon: Depth-guided neural 3D scene reconstruction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 18184–18194.

[64] A. Kamran-Pishhesari, A. Moniri-Morad, and J. Sattarvand, "Applications of 3D reconstruction in virtual reality-based teleoperation: A review in the mining industry," *Technologies*, vol. 12, no. 3, 2024, Art. no. 40.

[65] M. Kazhdan, M. Bolitho, and H. Hoppe, "Poisson surface reconstruction," in *Proc. Symp. Geometry Process.*, 2006, pp. 61–70, doi: 10.2312/SGP/SGP06/061-070.

[66] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian splatting for real-time radiance field rendering," *ACM Trans. Graph.*, vol. 42, no. 4, 2023, Art. no. 139.

[67] H. Kim and S. Han, "Interactive 3D building modeling method using panoramic image sequences and digital map," *Multimedia Tools Appl.*, vol. 77, no. 20, pp. 27387–27404, 2018.

[68] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[69] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–13, 2017.

[70] A. Kollert, M. Rutzinger, M. Bremer, K. Kaufmann, and T. Bork-Hüffer, "Mapping of 3D eye-tracking in urban outdoor environments," *ISPRS Ann. Photogrammetry Remote Sens. Spatial Inf. Sci.*, vol. 4, pp. 201–208, 2021.

[71] C. H. Koneputugodage, Y. Ben-Shabat, and S. Gould, "Octree guided unoriented surface reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 16717–16726.

[72] K. N. Kutulakos and S. M. Seitz, "A theory of shape by space carving," *Int. J. Comput. Vis.*, vol. 38, pp. 199–218, 2000.

[73] M. Lhuillier and L. Quan, "A quasi-dense approach to surface reconstruction from uncalibrated images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 418–433, Mar. 2005.

[74] R. Li, S. Fidler, A. Kanazawa, and F. Williams, “NeRF-XL: Scaling NeRFs with multiple GPUs,” 2024, *arXiv:2404.16221*.

[75] S. Li et al., “Instant-3D: Instant neural radiance field training towards on-device AR/VR 3D reconstruction,” in *Proc. 50th Annu. Int. Symp. Comput. Archit.*, 2023, pp. 1–13.

[76] Y. Li et al., “MatrixCity: A large-scale city dataset for city-scale neural rendering and beyond,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 3205–3215.

[77] Z. Li, X. Lyu, Y. Ding, M. Wang, Y. Liao, and Y. Liu, “RICO: Regularizing the unobservable for indoor compositional reconstruction,” 2023, *arXiv:2303.08605*.

[78] Z. Li et al., “Neuralangelo: High-fidelity neural surface reconstruction,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 8456–8465.

[79] Z. Li, S. Niklaus, N. Snavely, and O. Wang, “Neural scene flow fields for space-time view synthesis of dynamic scenes,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6498–6508.

[80] Z. Li and N. Snavely, “MegaDepth: Learning single-view depth prediction from internet photos,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2041–2050.

[81] Y. Liao, J. Xie, and A. Geiger, “KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3292–3310, Mar. 2023.

[82] S. P. Lim and H. Haron, “Surface reconstruction techniques: A review,” *Artif. Intell. Rev.*, vol. 42, pp. 59–78, 2014.

[83] K.-E. Lin, Y.-C. Lin, W.-S. Lai, T.-Y. Lin, Y.-C. Shih, and R. Ramamoorthi, “Vision transformer for NeRF-based view synthesis from a single input image,” in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 806–815.

[84] H. Liu and C. Wu, “Developing a scene-based triangulated irregular network (TIN) technique for individual tree crown reconstruction with LiDAR data,” *Forests*, vol. 11, no. 1, 2019, Art. no. 28.

[85] Z.-N. Liu, Y.-P. Cao, Z.-F. Kuang, L. Kobbelt, and S.-M. Hu, “High-quality textured 3D shape reconstruction with cascaded fully convolutional networks,” *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 1, pp. 83–97, Jan. 2021.

[86] W. E. Lorensen and H. E. Cline, “Marching cubes: A high resolution 3D surface construction algorithm,” in *Seminal Graphics: Pioneering Efforts that Shaped the Field*. New York, NY, USA: ACM, 1998, pp. 347–353.

[87] A. Lucero, “Using affinity diagrams to evaluate interactive prototypes,” in *Proc. IFIP Conf. Hum. Comput. Interact.*, Springer, 2015, pp. 231–248.

[88] P. Maken and A. Gupta, “2D-to-3D: A review for computational 3D image reconstruction from X-ray images,” *Arch. Comput. Methods Eng.*, vol. 30, no. 1, pp. 85–114, 2023.

[89] L. Melas-Kyriazi, C. Rupprecht, and A. Vedaldi, “PC2: Projection-conditioned point cloud diffusion for single-image 3D reconstruction,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 12923–12932.

[90] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, “Occupancy networks: Learning 3D reconstruction in function space,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4460–4470.

[91] O. Michel, R. Bar-On, R. Liu, S. Benaim, and R. Hanocka, “Text2Mesh: Text-driven neural stylization for meshes,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 13492–13502.

[92] Microsoft Research, “RGB-D dataset 7-Scenes,” Jan. 2013. Accessed: Feb. 01, 2024. [Online]. Available: https://www.microsoft.com/en-us/research/project/rgb-d-dataset-7-scenes/

[93] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, “NeRF: Representing scenes as neural radiance fields for view synthesis,” *Commun. ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[94] P. Mittal, Y.-C. Cheng, M. Singh, and S. Tulsiani, “AutoSDF: Shape priors for 3D completion, reconstruction and generation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 306–315.

[95] D. Mohamedally and P. Zaphiris, “Categorization constructionist assessment with software-based affinity diagramming,” *Int. J. Hum. Comput. Interact.*, vol. 25, no. 1, pp. 22–48, 2009.

[96] T. Müller, A. Evans, C. Schied, and A. Keller, “Instant neural graphics primitives with a multiresolution hash encoding,” *ACM Trans. Graph.*, vol. 41, no. 4, pp. 1–15, 2022.

[97] T. Mulumba, M. Eid, and A. Dhabi, “Real-time 3D reconstruction for haptic interaction,” in *Proc. IEEE Int. Conf. Comput. Intell. Virtual Environ. Meas. Syst. Appl.*, 2017.

[98] L. Nan and P. Wonka, “PolyFit: Polygonal surface reconstruction from point clouds,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017, pp. 2353–2361.

[99] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger, “Real-time 3D reconstruction at scale using voxel hashing,” *ACM Trans. Graph.*, vol. 32, no. 6, pp. 1–11, 2013.

[100] O. Özyeşil, V. Voroninski, R. Basri, and A. Singer, “A survey of structure from motion*,” *Acta Numerica*, vol. 26, pp. 305–364, 2017.

[101] Y. Pan, X. Zhong, L. Wiesmann, T. Posewsky, J. Behley, and C. Stachniss, “PIN-SLAM: LiDAR SLAM using a point-based implicit neural representation for achieving global map consistency,” 2024, *arXiv:2401.09101*.

[102] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, “DeepSDF: Learning continuous signed distance functions for shape representation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 165–174.

[103] J.-S. Park, “Interactive 3D reconstruction from multiple images: A primitive-based approach,” *Pattern Recognit. Lett.*, vol. 26, no. 16, pp. 2558–2571, 2005.

[104] K. Park, P. Henzler, B. Mildenhall, J. T. Barron, and R. Martin-Brualla, “CamP: Camera preconditioning for neural radiance fields,” *ACM Trans. Graph.*, vol. 42, no. 6, pp. 1–11, 2023.

[105] D. Paschalidou, O. Ulusoy, C. Schmitt, L. Van Gool, and A. Geiger, “RayNet: Learning volumetric 3D reconstruction with ray potentials,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3897–3906.

[106] N. Poliarnyi, “Out-of-core surface reconstruction via global TGV minimization,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 5641–5650.

[107] L. Puig and K. Daniilidis, “Monocular 3D tracking of deformable surfaces,” in *Proc. IEEE Int. Conf. Robot. Automat.*, 2016, pp. 580–586.

[108] Y. Qian, H. Zhang, and Y. Furukawa, “Roof-GAN: Learning to generate roof geometry and relations for residential houses,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2796–2805.

[109] J. Qiu, Z.-X. Yin, M.-M. Cheng, and B. Ren, “NeRC: Rendering planar caustics by learning implicit neural representations,” *IEEE Trans. Vis. Comput. Graph.*, vol. 30, no. 7, pp. 4339–4348, Jul. 2024.

[110] C. Reiser et al., “MERF: Memory-efficient radiance fields for real-time view synthesis in unbounded scenes,” *ACM Trans. Graph.*, vol. 42, no. 4, pp. 1–12, 2023.

[111] J. Reizenstein, R. Shapovalov, P. Henzler, L. Sbordone, P. Labatut, and D. Novotny, “Common objects in 3D: Large-scale learning and evaluation of real-life 3D category reconstruction,” in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 10881–10891.

[112] K. Rematas et al., “Urban radiance fields,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12932–12942.

[113] Y. Ren, T. Zhang, M. Pollefeys, S. Süsstrunk, and F. WOneang, “VolRecon: Volume rendering of signed ray distance functions for generalizable multi-view reconstruction,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 16685–16695.

[114] F. Rong, D. Xie, W. Zhu, H. Shang, and L. Song, “A survey of multi view stereo,” in *Proc. 2021 Int. Conf. Netw. Syst. AI*, 2021, pp. 129–135.

[115] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, “LabelMe: A database and web-based tool for image annotation,” *Int. J. Comput. Vis.*, vol. 77, pp. 157–173, 2008.

[116] A. Salvi, N. Gavenski, E. Pooch, F. Tasoniero, and R. Barros, “Attention-based 3D object reconstruction from a single image,” in *Proc. 2020 Int. Joint Conf. Neural Netw.*, 2020, pp. 1–8.

[117] T. Samavati and M. Soryani, “Deep learning-based 3D reconstruction: A survey,” *Artif. Intell. Rev.*, vol. 56, pp. 9175–9219, 2023.

[118] A. Schmied, T. Fischer, M. Danelljan, M. Pollefeys, and F. Yu, “R3D3: Dense 3D reconstruction of dynamic scenes from multiple cameras,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 3216–3226.

[119] J. L. Schonberger and J.-M. Frahm, “Structure-from-motion revisited,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4104–4113.

[120] R. Scupin, “The KJ method: A technique for analyzing data derived from Japanese ethnology,” *Hum. Org.*, vol. 56, no. 2, pp. 233–237, 1997.

[121] H. Shalma and P. Selvaraj, “A review on 3D image reconstruction on specific and generic objects,” *Mater. Today: Proc.*, vol. 80, pp. 2400–2405, 2023.

[122] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon, “Scene coordinate regression forests for camera relocalization in RGB-D images,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2930–2937.

[123] Y. Siddiqui, J. Thies, F. Ma, Q. Shan, M. Nießner, and A. Dai, "RetrievalFuse: Neural 3D scene reconstruction with a database," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 12568–12577.

[124] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. 12th Eur. Conf. Comput. Vis.*, Springer, 2012, pp. 746–760.

[125] K. Singh, M. Kalash, and N. Bruce, "Capturing real-world gaze behaviour: Live and unplugged," in *Proc. 2018 ACM Symp. Eye Tracking Res. Appl.*, 2018, pp. 1–9.

[126] R. Smith, M. Self, and P. Cheeseman, "Estimating uncertain spatial relationships in robotics," *Mach. Intell. Pattern Recognit.*, vol. 5, pp. 435–461, 1988.

[127] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2256–2265.

[128] N. Somraj, A. Karanayil, and R. Soundararajan, "SimpleNeRF: Regularizing sparse input neural radiance fields with simpler solutions," in *Proc. SIGGRAPH Asia Conf. Papers*, 2023, pp. 1–11.

[129] N. Somraj and R. Soundararajan, "ViP-NeRF: Visibility prior for sparse input neural radiance fields," 2023, *arXiv:2305.00041*.

[130] Z. Song, X. Wang, H. Zhu, G. Zhou, and Q. Wang, "Learning reliable gradients from undersampled circular light field for 3D reconstruction," *IEEE Trans. Vis. Comput. Graph.*, vol. 29, no. 12, pp. 5194–5207, Dec. 2023.

[131] O. Sorkine and M. Alexa, "As-rigid-as-possible surface modeling," in *Proc. Symp. Geometry Process.*, 2007, pp. 109–116.

[132] E. K. Stathopoulou and F. Remondino, "A survey on conventional and learning-based methods for multi-view stereo," *Photogrammetric Rec.*, vol. 38, no. 183, pp. 374–407, 2023.

[133] N. Stier et al., "FineRecon: Depth-aware feed-forward network for detailed 3D reconstruction," 2023, *arXiv:2304.01480*.

[134] J. Stjepandić, M. Sommer, and B. Denkena, *DigiTwin: An Approach for Production Process Optimization in a Built Environment*. Berlin, Germany: Springer, 2022.

[135] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2012, pp. 573–580.

[136] J.-M. Sun, T. Wu, Y.-L. Yang, Y.-K. Lai, and L. Gao, "SOL-NeRF: Sunlight modeling for outdoor scene decomposition and relighting," in *Proc. SIGGRAPH Asia Conf. Papers*, 2023, pp. 1–11.

[137] P. Sun et al., "Scalability in perception for autonomous driving: Waymo open dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2446–2454.

[138] V. Sundstedt and V. Garro, "A systematic review of visualization techniques and analysis tools for eye-tracking in 3D environments," *Front. Neuroergonomics*, vol. 3, 2022, Art. no. 910019.

[139] T. Suzuki, "Fed3DGS: Scalable 3D Gaussian splatting with federated learning," 2024, *arXiv:2403.11460*.

[140] M. Tan et al., "MnasNet: Platform-aware neural architecture search for mobile," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2820–2828.

[141] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[142] M. Tan and Q. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10096–10106.

[143] M. Tancik et al., "Block-NeRF: Scalable large scene neural view synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8248–8258.

[144] M. Tatarchenko, S. R. Richter, R. Ranftl, Z. Li, V. Koltun, and T. Brox, "What do single-view 3D reconstruction networks learn?," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3405–3414.

[145] S. Thrun and M. Montemerlo, "The graph SLAM algorithm with applications to large-scale mapping of urban structures," *Int. J. Robot. Res.*, vol. 25, no. 5/6, pp. 403–429, 2006.

[146] Y. Tian, C. Li, X. Guo, and B. Prabhakaran, "Real time stable haptic rendering of 3D deformable streaming surface," in *Proc. 8th ACM Multimedia Syst. Conf.*, 2017, pp. 136–146.

[147] H. Turki, D. Ramanan, and M. Satyanarayanan, "Mega-NeRF: Scalable construction of large-scale NeRFs for virtual fly-throughs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12922–12931.

[148] H. Turki, J. Y. Zhang, F. Ferroni, and D. Ramanan, "SUDS: Scalable urban dynamic scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 12375–12385.

[149] S. Ullman, *The Interpretation of Visual Motion*. Cambridge, MA, USA: MIT Press, 1979.

[150] J. Ventura, C. Arth, G. Reitmayr, and D. Schmalstieg, "Global localization from monocular SLAM on a mobile phone," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 4, pp. 531–539, Apr. 2014.

[151] H. Wallach and D. O'connell, "The kinetic depth effect," *J. Exp. Psychol.*, vol. 45, no. 4, 1953, Art. no. 205.

[152] C. Wang, R. Jiang, M. Chai, M. He, D. Chen, and J. Liao, "NeRF-Art: Text-driven neural radiance fields stylization," *IEEE Trans. Vis. Comput. Graph.*, vol. 30, no. 8, pp. 4983–4996, Aug. 2024.

[153] G. Wang, Z. Chen, C. C. Loy, and Z. Liu, "SparseNeRF: Distilling depth ranking for few-shot novel view synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 9065–9076.

[154] K. Wang, M. Savva, A. X. Chang, and D. Ritchie, "Deep convolutional priors for indoor scene synthesis," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–14, 2018.

[155] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, "NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction," 2021, *arXiv:2106.10689*.

[156] Q. Wang, "Towards real-time 3D terrain reconstruction from aerial imagery," *Geographies*, vol. 4, no. 1, pp. 66–82, 2024.

[157] S. Wang, R. Clark, H. Wen, and N. Trigoni, "DeepVO: Towards end-to-end visual odometry with deep recurrent convolutional neural networks," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2017, pp. 2043–2050.

[158] Y. Wang, Q. Han, M. Habermann, K. Daniilidis, C. Theobalt, and L. Liu, "NeuS2: Fast learning of neural implicit surfaces for multi-view reconstruction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 3295–3306.

[159] Y. Wang, X. He, S. Peng, H. Lin, H. Bao, and X. Zhou, "AutoRecon: Automated 3D object discovery and reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 21382–21391.

[160] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[161] Z. Wang et al., "Neural-singular-hessian: Implicit neural representation of unoriented point clouds by enforcing singular hessian," *ACM Trans. Graph.*, vol. 42, no. 6, pp. 1–14, 2023.

[162] Y. Wei, S. Liu, Y. Rao, W. Zhao, J. Lu, and J. Zhou, "NerfingMVS: Guided optimization of neural radiance fields for indoor multi-view stereo," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 5610–5619.

[163] F. Wimbauer, N. Yang, C. Rupprecht, and D. Cremers, "Behind the scenes: Density fields for single view reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 9076–9086.

[164] J. Xiao, A. Owens, and A. Torralba, "SUN3D: A database of big spaces reconstructed using SFM and object labels," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1625–1632.

[165] Y. Xie, M. Gadelha, F. Yang, X. Zhou, and H. Jiang, "PlanarRecon: Real-time 3D plane detection and reconstruction from posed monocular videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 6219–6228.

[166] Z. Xie, J. Zhang, W. Li, F. Zhang, and L. Zhang, "S-NeRF: Neural radiance fields for street views," 2023, *arXiv:2303.00749*.

[167] G. Xu, W. Yin, H. Chen, C. Shen, K. Cheng, and F. Zhao, "FrozenRecon: Pose-free 3D scene reconstruction with frozen depth models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 9276–9286.

[168] L. Xu et al., "C2F2NeUS: Cascade cost frustum fusion for high fidelity and generalizable neural surface reconstruction," 2023, *arXiv:2306.10003*.

[169] L. Xu et al., "Grid-guided neural radiance fields for large urban scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 8296–8306.

[170] Q. Xu and W. Tao, "Multi-scale geometric consistency guided multi-view stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5483–5492.

[171] W. Xu, Y. Zeng, and C. Yin, "3D city reconstruction: A novel method for semantic segmentation and building monomer construction using oblique photography," *Appl. Sci.*, vol. 13, no. 15, 2023, Art. no. 8795.

[172] X. Xu, P. Guerrero, M. Fisher, S. Chaudhuri, and D. Ritchie, "Unsupervised 3D shape reconstruction by part retrieval and assembly," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 8559–8567.

[173] B. Yang et al., "Learning object-compositional neural radiance field for editable scene rendering," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 13779–13788.

[174] G.-W. Yang, W.-Y. Zhou, H.-Y. Peng, D. Liang, T.-J. Mu, and S.-M. Hu, "Recursive-NeRF: An efficient and dynamically growing NeRF," *IEEE Trans. Vis. Comput. Graph.*, vol. 29, no. 12, pp. 5124–5136, Dec. 2023.

[175] Y. Yao et al., "BlendedMVS: A large-scale dataset for generalized multi-view stereo networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1790–1799.

[176] L. Yariv et al., "BakedSDF: Meshing neural SDFs for real-time view synthesis," 2023, *arXiv:2302.14859*.

[177] M. Yazdanpour, G. Fan, and W. Sheng, "ManhattanFusion: Online dense reconstruction of indoor scenes from depth sequences," *IEEE Trans. Vis. Comput. Graph.*, vol. 28, no. 7, pp. 2668–2681, Jul. 2022.

[178] C. Yeshwanth, Y.-C. Liu, M. Nießner, and A. Dai, "ScanNet++: A high-fidelity dataset of 3D indoor scenes," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 12–22.

[179] W. Yin et al., "Towards accurate reconstruction of 3D scene shape from a single monocular image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 6480–6494, May 2023.

[180] S. M. Yoon and H. Graf, "Eye tracking based interaction with 3D reconstructed objects," in *Proc. 16th ACM Int. Conf. Multimedia*, 2008, pp. 841–844.

[181] Y.-J. Yuan, Y.-K. Lai, Y.-H. Huang, L. Kobbelt, and L. Gao, "Neural radiance fields from sparse RGB-D images for high-quality view synthesis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 8713–8728, Jul. 2023.

[182] Y.-J. Yuan et al., "Interactive NeRF geometry editing with shape priors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 14821–14837, Dec. 2023.

[183] C. Zeng, G. Chen, Y. Dong, P. Peers, H. Wu, and X. Tong, "Relighting neural radiance fields with shadow and highlight hints," in *Proc. ACM SIGGRAPH Conf. Proc.*, 2023, pp. 1–11.

[184] X. Zeng et al., "LION: Latent point diffusion models for 3D shape generation," 2022, *arXiv:2210.06978*.

[185] J. Zhang, M. Ji, G. Wang, Z. Xue, S. Wang, and L. Fang, "SurRF: Unsupervised multi-view stereopsis by learning surface radiance field," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7912–7927, Nov. 2022.

[186] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.

[187] X. Zhang, Z. Chen, F. Wei, and Z. Tu, "Uni-3D: A universal model for panoptic 3D scene reconstruction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 9256–9266.

[188] X. Zhang, A. Kundu, T. Funkhouser, L. Guibas, H. Su, and K. Genova, "Nerflets: Local radiance fields for efficient structure-aware 3D scene representation from 2D supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 8274–8284.

[189] Y. Zhang, F. Tosi, S. Mattoccia, and M. Poggi, "GO-SLAM: Global optimization for consistent 3D instant reconstruction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 3727–3737.

[190] J. Zheng, J. Zhang, J. Li, R. Tang, S. Gao, and Z. Zhou, "Structured3D: A large photo-realistic dataset for structured 3D modeling," in *Proc. 16th Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 519–535.

[191] H. Zhong, J. Zhang, and J. Liao, "VQ-NeRF: Neural reflectance decomposition and editing with vector quantization," *IEEE Trans. Vis. Comput. Graph.*, vol. 30, no. 9, pp. 6247–6260, Sep. 2024.

[192] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely, "Stereo magnification: Learning view synthesis using multiplane images," *ACM Trans. Graph.*, vol. 37, 2018, Art. no. 65.

[193] Z. Zhou and S. Tulsiani, "SparseFusion: Distilling view-conditioned diffusion for 3D reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 12588–12597.

[194] J. Zhu et al., "I2-SDF: Intrinsic indoor scene reconstruction and editing via raytracing in neural SDFs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 12489–12498.

[195] J. Zhuang, C. Wang, L. Lin, L. Liu, and G. Li, "DreamEditor: Text-driven 3D scene editing with neural fields," in *Proc. SIGGRAPH Asia Conf. Papers*, 2023, pp. 1–10.

[196] Y. Zhuang et al., "Anti-aliased neural implicit surfaces with encoding level of detail," in *Proc. SIGGRAPH Asia Conf. Papers*, 2023, pp. 1–10.

[197] P. Zins, Y. Xu, E. Boyer, S. Wuhrer, and T. Tung, "Multi-view reconstruction using signed ray distance functions (SRDF)," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 16696–16706.

[198] X. Zou et al., "Deep learning for cross-domain data fusion in urban computing: Taxonomy, advances, and outlook," *Inf. Fusion*, vol. 113, 2025, Art. no. 102606.

Andreas Christodoulides received the BEng and MSc degrees in aerospace engineering, and the MSc degree in computer science. He is a PhD researcher and a member of the EPIC CDT, Swansea University, funded by Ordnance Survey. His research relates to computer graphics, deep learning, and interactive systems.

Gary K. L. Tam received the MPhil degree from the City University of Hong Kong, and the PhD degree from Durham University. He is a senior lecturer with the Department of Computer Science, Swansea University. His recent research interests focus on 3D geometry and computer vision. He has served as a guest editor for special issues of the *International Journal of Computer Vision* (2017), *Computers* (2018, 2019), and *Applied Sciences* (2023, 2024). He is also an associate editor of the *AI Communications* since 2024.

James Clarke is a research software engineer with Ordnance Survey, the U.K.'s national mapping agency. His work focuses on ensuring research outputs are reusable, reproducible, and maintainable. His research specialty is point cloud capture and processing, and he explores the role of point clouds in the future of 3D mapping.

Richard Smith is associate professor in human geography with Swansea University in the U.K. Research focuses on postmodernism, global cities, the business of cities, the future of cities, and the application of poststructuralist theory for understanding the contemporary world. More than 100 publications having written on social theory and urban space, world cities, global cities, urban networks, globalization, corporate services, structuralism, poststructuralism, postmodernism, human geography, and Jean Baudrillard.

Jon Horgan received the honours degree in GIS and remote sensing, in 2001. He has worked with Ordnance Survey since 1987 where he started his career as a field surveyor. In 2004, he attended the Institute of Geomatics, Castelldefels, Catalonia to study for a masters in airborne photogrammetry and remote sensing. Working in data geomatics research since 2003, he has an excellent knowledge of planning, extracting, and exploiting 3D data.

Nicholas Micallef received the PhD degree from Glasgow Caledonian University, U.K. He is a lecturer with Swansea University. His current research focuses on misinformation, both from a data-driven and human aspects perspective. In his previous role at New York University Abu Dhabi, he was involved in various projects that studied the countering of misinformation and the spreading of cross-platform multimodal misinformation on various platforms, including messaging services such as WhatsApp.

Jeremy Morley has been chief geospatial scientist with Ordnance Survey since 2015. At OS, he leads the Research team, focusing on commissioning, planning and executing research projects with universities, promoting active knowledge transfer and horizon scanning to identify new business opportunities and emerging research. Previously, he was an academic with University College London and the University of Nottingham.

Nelly Villamizar received the PhD degree in mathematics from the University of Oslo, Norway, in 2013. She is an associate professor with the Department of Mathematics, Swansea University. Her research interests include applied and computational algebraic geometry, with a focus on developing algebraic tools to solve problems arising in geometric modeling and Computer Aided Geometric Design (CAGD).

Sean Walton is a computer science lecturer with Swansea University, Sêr Cymru II fellow, and founding director of Pill Bug Interactive (https://www.pillbug.zone/), a BAFTA Cymru nominated games development studio. His research background is primarily in using evolutionary optimisation techniques in engineering. Since joining the Computational Foundry at Swansea, he has expanded his research interests into educational technology and the procedural generation of video game content.