frontiers | Frontiers in Computer Science

TYPE Review
PUBLISHED 13 June 2025
DOI 10.3389/fcomp.2025.1467103

OPEN ACCESS

EDITED BY
Wei Lang,
Sun Yat-sen University, China

REVIEWED BY
Jonathan M. Aitken,
The University of Sheffield, United Kingdom
Ismail Elezi,
Technical University of Munich, Germany
Ciprian Orhei,
Politehnica University of Timişoara, Romania

*CORRESPONDENCE
Alireza Akhavi Zadegan
✉ alireza.akhavi.zadegan@ut.ee

RECEIVED 19 July 2024
ACCEPTED 20 May 2025
PUBLISHED 13 June 2025

CITATION
Akhavi Zadegan A, Vivet D and Hadachi A (2025) Challenges and advancements in image-based 3D reconstruction of large-scale urban environments: a review of deep learning and classical methods.
Front. Comput. Sci. 7:1467103.
doi: 10.3389/fcomp.2025.1467103

Challenges and advancements in image-based 3D reconstruction of large-scale urban environments: a review of deep learning and classical methods

Alireza Akhavi Zadegan[1]*, Damien Vivet[2] and Amnir Hadachi[1]

[1]ITS Laboratory, Institute of Computer Science, University of Tartu, Tartu, Estonia, [2]ISAE-SUPAERO, Universite de Toulouse, Toulouse, France

Over the past decade, the field of image-based 3D scene reconstruction and generation has experienced a significant transformation, driven by the integration of deep learning technologies. This shift underscores a maturing discipline characterized by rapid advancements and the introduction of numerous innovative methodologies aimed at broadening research boundaries. The specific focus of this study is on image-based 3D reconstruction techniques applicable to large-scale urban environments. This focus is motivated by the growing need for advanced urban planning and infrastructure development for smart city applications and digitalization, which requires precise and scalable modeling solutions. We employ a comprehensive classification framework that distinguishes between traditional and deep learning approaches for reconstructing urban facades, districts, and entire cityscapes. Our review methodically compares these techniques, evaluates their methodologies, highlights their distinct characteristics and performance, and identifies their limitations. Additionally, we discuss commonly utilized 3D datasets for large environments and the prevailing performance metrics in this domain. The paper concludes by outlining the current challenges faced by the field and proposes directions for future research in this swiftly evolving area.

KEYWORDS

3D reconstruction, computer vision, large scale 3D urban model, 3D mapping, image based 3D modeling

1 Introduction

The rapid development of digital cameras, sensors, and computational power has led to significant progress in 3D reconstruction technology over the past few decades. The ability to generate accurate and detailed 3D models of real-world scenes has broad applications in areas such as autonomous driving, city planning, environmental monitoring, cultural heritage preservation and digital twin. In particular, image-based 3D reconstruction has become a popular technique due to its low cost and ease of use compared to other methods such as LiDAR or structured light scanning. While image-based 3D reconstruction has been successfully applied to indoor environments, the reconstruction of large outdoor scenes remains a challenging task due to the complex and dynamic nature of natural environments. Outdoor scenes are often characterized by uneven terrain, complex lighting conditions, occlusions, and varying textures and colors, making it difficult to accurately capture and process the necessary data. Despite these challenges, recent advances in camera

and computational technologies have enabled researchers to make significant progress in this field. In the domain of large-scale 3D reconstruction, primary data acquisition is dominated by two technologies: LiDAR and camera-based imagery. LiDAR, which includes airborne and terrestrial methods, utilizes Time of Flight and triangulation techniques to collect data. Airborne LiDAR provides broad coverage Vosselman and Maas (2010), while terrestrial LiDAR offers detailed, location-specific scans but may omit features like building roofs due to viewing constraints (Wang, 2013; Amberg et al., 2007; Kühner and Kümmerle, 2020). Alternatively, camera-based imagery, captured through drones (Liu and Ji, 2020), satellites (Duan and Lafarge, 2016), and handheld devices, is instrumental in generating detailed topographic maps and intricate urban models, valuable for applications such as urban planning and disaster management. The data from these technologies are represented through various methods including volumetric (Wu et al., 2015), geometric (Qi et al., 2017; Pan et al., 2018), primitive shapes (Tulsiani et al., 2017), and implicit surface representations (Xu et al., 2019), each offering unique advantages in handling complex 3D structures. In this study, we have narrowed our focus to research studies that specifically deal with large-scale outdoor 3D reconstruction using images as the input modality. Instead of conducting a comprehensive review of all related works, we have categorized the literature based on the scale of the reconstructed scene in an outdoor scenario. This categorization includes facades, districts, and cityscapes. Table 1 lists all the different outdoor scenarios and their corresponding approaches that will be covered in this paper.

TABLE 1 Large-scale 3D reconstruction studies categorized based on the scale of reconstruction.

3D reconstruction scale	Techniques	Methods
Facades	Photogrammetry and image matching	Classical
Facades	Deep Learning and hybrid methods	Deep learning
Facades	Neural rendering	Deep learning
Facades	Semantic based	Deep learning
Districts	Neural implicit based	Deep learning
Districts	Depth-based	Deep learning and classical
Districts	GAN based	Deep learning
Districts	Combinatorial strategy based	Classical
Cityscapes	Multi-procedural based	Deep learning and classical

1.1 Review methodology

The selection of papers for this review followed a systematic approach to ensure a comprehensive and unbiased evaluation of advancements in large-scale outdoor 3D reconstruction. Literature published from 2015 onwards was prioritized, as this period marked significant progress in both classical and deep learning-based approaches. Studies were included if they utilized image-based techniques for large-scale reconstruction, specifically focusing on urban facades, districts, and cityscapes. Preference was given to research incorporating deep learning, neural rendering, hybrid methods, or classical photogrammetry techniques. Conversely, studies that relied primarily on LiDAR, radar, or other non-visual sensors without integrating image-based reconstruction were excluded, as were those focusing on small-scale object reconstructions, indoor environments, or non-urban scenarios. A total of 73 papers were selected and analyzed, categorized based on the scale of reconstruction and the methodologies employed. Each study was assessed for its approach, data sources, performance evaluation, and contributions to the field, enabling a structured comparison of different techniques and their effectiveness in real-world applications. The objective of this paper is to provide a comprehensive and methodical evaluation of recent progress in large-scale outdoor 3D reconstruction, examining both conventional and deep learning methodologies. This review aims to guide readers through this rapidly evolving field, which has gained considerable attention in recent years. To the best of our knowledge, this is the first survey paper to focus exclusively on image-based 3D reconstruction for outdoor environments, specifically addressing facades, districts, and cityscapes. By extensively reviewing literature published since 2015, we present a detailed analysis of key methodologies, summarize their performance and properties, and provide a comparative overview in a structured format.

The structure of our paper is organized as follows: Section 2, delves into the categorization of existing studies on facade reconstruction, which we have divided into three primary methodologies: photogrammetry and image matching, deep learning and hybrid methods, and neural rendering techniques. This section aims to provide a foundational understanding of the current methodologies employed in facade reconstruction. Section 3 expands our exploration into the 3D reconstruction of urban districts, detailing five specific techniques: semantic-based, neural implicit-based, depth-based, GAN-based, and combinatorial-based methods. This section is designed to offer insights into the diverse approaches tailored to urban district modeling. Section 4 is dedicated to the comprehensive examination of research focused on the reconstruction of entire cityscapes. Here, we discuss the integration and scaling of reconstruction techniques to accommodate the complexity of whole-city modeling. Section 5 introduces the most widely utilized image-based datasets pertinent to large-scale 3D reconstruction. This section highlights the critical role of datasets in developing, testing, and benchmarking reconstruction algorithms. Section 6 presents common performance metrics used in image-based large-scale 3D reconstruction. This section aims to equip readers with the criteria and standards used to evaluate the effectiveness and accuracy of various reconstruction methods. Finally, section 7, serves as the culmination of our discussion, where we synthesize the prevailing trends and outline the challenges currently facing the field. Additionally, we propose potential avenues for future research within this domain, aiming to inspire continued innovation and exploration.

2 Facade reconstructions

Image-based facade reconstruction lies in its ability to provide detailed and accurate information about the facade of a building. This information can be used for a variety of purposes, such as historic preservation, architectural analysis, urban planning, maintenance and virtual tours. We have organized the extant literature on image-based facade reconstruction into three sections: photogrammetry and image matching, deep learning and hybrid methods, and neural rendering approaches.

2.1 Photogrammetry and image matching 3D reconstruction

Photogrammetry uses image matching algorithms to transform a series of overlapping two-dimensional photos into a cohesive three-dimensional point cloud. This method relies on identifying and aligning unique features visible across multiple images. By using principles of projective geometry, photogrammetry uses these feature correspondences to perform spatial triangulation, calculating the exact coordinates of the physical features in three-dimensional space. The process includes refining a mesh and applying texture mapping to create a high-fidelity 3D model that accurately captures the topology and morphology of the subject. This sophisticated technique is crucial for fields that require realistic digital representations of complex environments and objects, such as topographic surveying and the digital preservation of cultural heritage artifacts (Gruen, 2012). Building upon these foundational application, Wu et al. (2018) presented a new approach to optimize the 3D modeling of urban areas by integrating aerial oblique imagery with terrestrial imagery. The approach involves matching feature points between aerial oblique images and terrestrial images to perform combined Bundle Adjustment (BA) for the two datasets. This process yields optimal image orientation parameters that better co-register the aerial and terrestrial datasets, resulting in improved geometric accuracy. The input for the approach includes aerial and terrestrial image datasets and their respective initial image orientation parameters. The improved image orientation parameters and the sparse point clouds of the combined image block are retrieved using the BA approach. These can then be imported to generate dense point clouds and surface models with better geometric accuracy using existing scene reconstruction software. Finally, the approach optimizes the textures of building facades using image patches from terrestrial views, resulting in higher-quality 3D models. Overall, this approach provides a more accurate and comprehensive 3D model of urban areas by combining the strengths of both aerial and terrestrial imagery. Figure 1 depicts the workflow of the proposed framework.

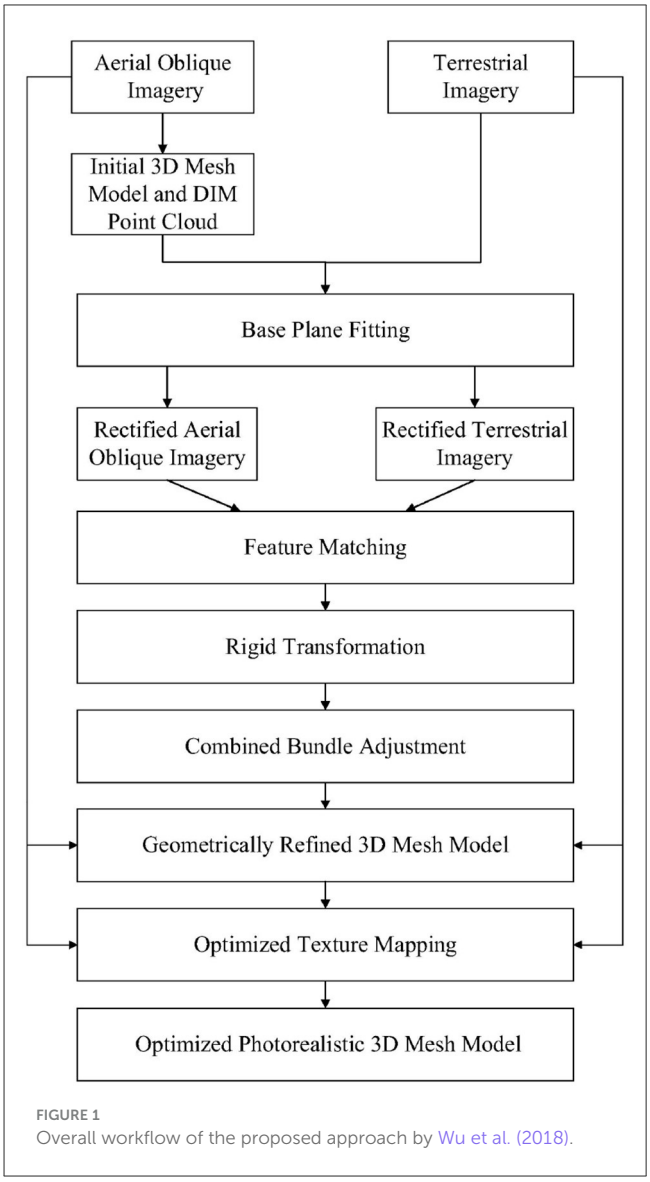

FIGURE 1
Overall workflow of the proposed approach by Wu et al. (2018).

Li et al. (2020) similarly uses a combination of aerial and terrestrial images to generate a textured 3D mesh model. The workflow of the proposed approach can be divided into three main steps: data pre-processing, combined Structure from motion (SfM), and optimal generation of a textured 3D mesh model. The first step, data pre-processing, includes optimal image selection, color equalization, and moving object removal to reduce the computational cost in subsequent steps and reduce differences between images obtained from different sources. The second step, combined SfM, involves estimating the Exterior Orientation (EO) parameters of the aerial and terrestrial images separately as initial values. These initial values are then refined through a combination of image matching and BA. In the final step, the refined point clouds are improved by detecting and modifying visibility conflicts. These refined point clouds are then used to generate a 3D mesh model, and textures are mapped using the images selected in the pre-processing step. Guo and Guo (2018) proposes a new method for optimizing the accuracy of 3D reconstruction from multi-view images in urban scenes. The proposed method incorporates 3D line information along with dense points information to refine all the existing edges in the scene. The method involves joint estimation of 3D line and dense points information which is used to remove wrong lines and incorrect points. The process includes several steps such as reconstruction of dense points and 3D line information from structure and motion result, removal of wrong lines and incorrect points, and fusion of point position information with

normal information. The proposed method was tested on several sets of real images of urban buildings and demonstrated significant improvement in reconstruction accuracy compared to the original results. Overall, the study presents a promising approach to improve the accuracy of large-scale scene reconstruction in urban areas. To address the problem of ambiguity in image matching Xu et al. (2020) proposed a method called Matching Ambiguity Reduced Multiple View Stereo (MARMVS). The ambiguity in image matching is one of the main factors that reduce the quality of the 3D model reconstructed by PatchMatch based multiple view stereo. The proposed method uses three strategies to handle the ambiguity in image matching process:

1. The matching ambiguity is measured by the differential geometry property of image surface with epipolar constraint, which is used as a critical criterion for optimal scale selection of every single pixel with corresponding neighboring images.
2. The depth of every pixel is initialized to be more close to the true depth by utilizing the depths of its surrounding sparse feature points, which yields faster convergency speed in the following PatchMatch stereo and alleviates the ambiguity introduced by self-similar structures of the image.
3. In the last propagation of the PatchMatch stereo, higher priorities are given to those planes with less ambiguity in their corresponding 2D image patches. This approach helps extend the correct reconstruction of surfaces into areas with unprocessed textures.

The proposed method is validated on public benchmarks, and the experimental results demonstrate competing performance against the state of the art. Given the adequate parameterization and discretization in the depth map computation stage, the proposed approach is exceptionally efficient even when operating on consumer-grade CPUs.

2.2 Deep learning and hybrid methods 3D reconstruction

Deep learning, when integrated with hybrid methods, is significantly advancing the field of 3D reconstruction. The essence of this advancement lies in the synergy between neural network architectures for depth estimation and image segmentation, and traditional computational geometry for model refinement. Auto-encoder networks are now being employed to infer depth from single RGB images, reducing the need for complex sensor arrays and lowering costs. Generative Adversarial Networks (GANs) enhance the segmentation of facades, leading to more accurate reconstructions. Finally, the output from deep learning models is seamlessly integrated with computational geometry techniques, resulting in automated, adaptable, and precise 3D models. This fusion of methodologies is proving to be particularly transformative in urban modeling, offering a path to detailed and structurally accurate digital representations from minimal and less complex data inputs. The aforementioned advancements in deep learning and hybrid methodologies set the stage for the work of Bacharidis et al. (2020) who sought to expand upon the robustness and applicability of their previous research Bârsan et al. (2018) in producing 3D representations of building facades. The contributions of their paper are threefold:

1. An auto-encoder neural network architecture for depth estimation using a single RGB image instead of a stereoscopic image sensor rig design. This can potentially increase the flexibility and decrease the overall cost of the framework.
2. Incorporating a deep learning-based facade segmentation stage based on GANs, enabling more scalable and robust facade element detection. This improves the accuracy and efficiency of the framework.
3. Integrating computational geometry techniques and point cloud processing algorithms to produce a detailed reconstructed 3D surface, enhancing the automation and adaptability of the suggested workflow. This makes the framework more flexible and adaptable to different scenarios.

Overall, the extension improves the 3D reconstruction framework for building facades by introducing new technologies and approaches, enhancing the accuracy, efficiency, and flexibility of the workflow compared to their previous work.

In line with learning based approaches Alidoost et al. (2020) proposed a deep learning-based framework for automatic detection, localization, and height estimation of buildings from a single aerial image. The proposed framework is based on a Y-shaped Convolutional Neural Network (Y-Net) which includes one encoder and two decoders. The input of the network is a single RGB image, and the outputs are predicted height information of buildings as well as the rooflines in three classes of eave, ridge, and hip lines. The extracted knowledge by the Y-Net is utilized for 3D reconstruction of buildings based on the third Level of Detail (LoD3). The proposed approach consists of data preparation, Convolutional Neural Network's (CNNs) training, and 3D reconstruction. For the experimental investigations, airborne data from Potsdam were used, which were provided by the International Society for Photogrammetry and Remote Sensing (ISPRS). The results shows an average Root Mean Square Error (RMSE) and a Normalized Median Absolute Deviation (NMAD) of about 3.8 m and 1.3 m, respectively, for the predicted heights. Moreover, the overall accuracy of the extracted rooflines is about 86%.

Huang et al. (2020) proposed a statistical model called "shell model" (see Figure 2). This hybrid model combines elements of Constructive Solid Geometry (CSG) and Boundary Representation (BRep) models and is designed to work with data from both terrestrial and Unmanned Aerial Vehicle (UAV) imagery. Unlike conventional surface or solid body models, the shell model consists of an outer and inner layer that defines a solid body model with a certain thickness between them, providing a more practical and suitable geometric model. The authors observe that measurement data, such as point clouds from LiDAR and image matching, only reveals the surface of the building, which is imperfect due to measurement uncertainty. They acknowledge that there are still challenges to be addressed, such as the representation of public and commercial buildings with special shapes that cannot be represented by the introduced rectangular primitives, and the modeling of superstructures on the roof and facades, as well

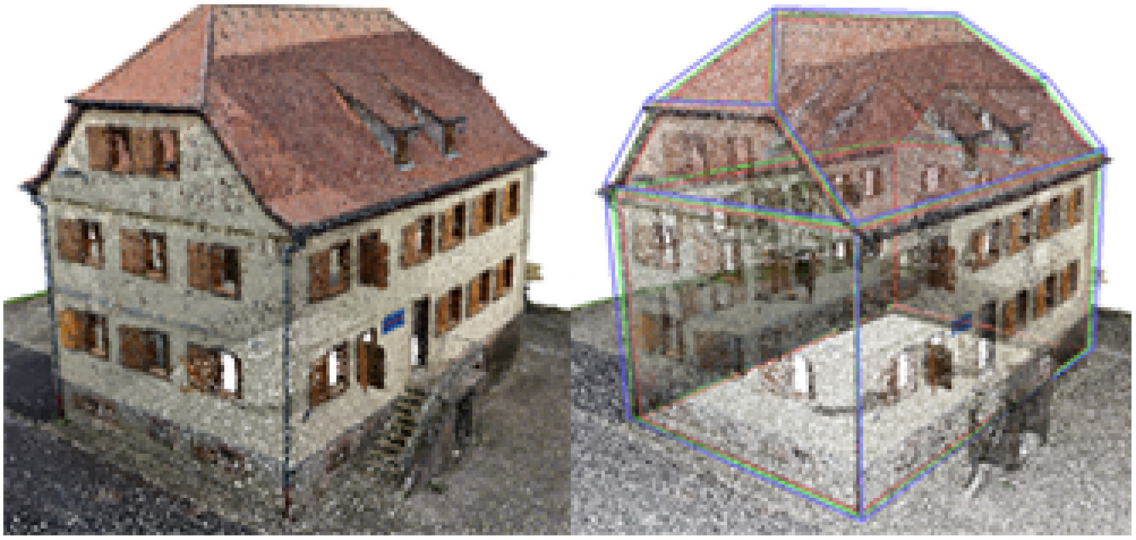

FIGURE 2
Shows the discovery of a structure with a half-hipped roof. The input point cloud is fitted to the shell model with inner (red) and outer (blue) layers (bottom). The layer (green) between them is therefore taken to be the model of best fit by Huang et al. (2020).

as annexes of the buildings. The authors suggest upgrading the library of primitives with flexible geometric shapes and specific types for superstructures and annexes. The paper also suggests the use of ConvNets for direct parsing of 3D geometry, such as the segmentation of point clouds into building parts and the detection of facade elements using both color and depth information. The study presents a promising approach to building 3D models and highlights future directions for research.

Fan et al. (2021) presents a web-based interactive platform called VGI3D, that can construct 3D building models using free images from internet users or Volunteered Geographic Information (VGI) platforms. The proposed platform is designed to address the challenges associated with creating 3D building models, which typically require significant labor and time costs, as well as expensive devices. The platform can effectively generate 3D building models from images in 30 seconds, using a user interaction module and CNN. The user interaction module provides the facade boundary for 3D building modeling, while the CNN can identify facade elements even in complex scenes with multiple architectural styles. The user interaction module is designed to be simple and user-friendly for both experts and non-experts. The paper also presents usability testing results and feedback from participants to further optimize the platform and user experience. Using VGI data reduces labor and device costs, and the CNN simplifies the process of extracting elements for 3D building modeling. In contrast, Tripodi et al. (2020) introduces an automated pipeline for creating 3D models of urban areas with Level of Detail one (LoD1) using satellite imagery. The accuracy of the model is dependent on the quality of the stereo images, which can often be affected by noise and distortion. To overcome this, the paper proposes a pipeline that combines U-net for contour extraction and optimization with computational geometry techniques for the creation of a precise digital terrain model, digital height model, and the position of building footprints. The pipeline is efficient and can work even when close-to-nadir images are not available. Experimental results demonstrate the effectiveness of the proposed pipeline in 3D building reconstruction.

Given a multi-view stereo, Romanoni et al. (2017) tries to improve the geometry and semantic labeling of a given mesh for semantic 3D reconstruction using their framework. Current methods rely on volumetric approaches to fuse RGB image data with semantic labels, but these methods are not scalable and do not produce high-resolution meshes. The proposed approach refines mesh geometry by using a variational method that optimizes a composite energy consisting of a pairwise photometric term and a single-view term that models the semantic consistency between the labels of the 3D mesh and those of the segmented images. The approach updates the semantic labeling through a Markov Random Field (MRF) formulation that considers class-specific priors estimated directly from the annotated mesh. This is the first approach to use semantics within a mesh refinement framework and it improves the robustness of noisy segmentations compared to existing methods that use handcrafted or learned priors.

Roy et al. (2022) introduces a hybrid learning based framework for generating 3D building models from 2D images. The framework is based on a parametric representation of 3D buildings, which provides human-interpretable 3D models that allow users to make edits to the model. The framework is composed of two main modules, a facade detection and frontalization module and a 2D to 3D conversion module. To train the model, the authors used a large-scale synthetic dataset generated by the hyper simulation platform, which allowed them to learn reliable models from a small amount of real data. The results show that the model is able to generate meaningful 3D models from arbitrary 2D images and can capture the structural details of the building in 2D images. The authors plan to extend their work by incorporating more parameters and more complex 3D buildings with multi-layered and non-uniform structures.

2.3 Neural rendering based 3D reconstruction

Neural Rendering based 3D Reconstruction is an advanced computational approach that merges neural network methodologies with traditional 3D rendering techniques to tackle the complexities of reconstructing large-scale scenes with variable lighting. This approach employs neural networks to create detailed surface geometries, utilizing appearance embeddings to capture lighting variations and generating meshes as direct geometric representations Tewari et al. (2022). These meshes are then integrated into standard graphics workflows, overcoming the computational intensity typically associated with volumetric radiance fields. The technique is further refined by a hybrid sampling strategy that leverages both voxel data and surface predictions, optimizing the process for environments with limited computational resources. Pioneering works in this domain not only provide more efficient and accurate reconstruction methods but also contribute specialized datasets, like Heritage-Recon, that facilitate the evaluation of these neural rendering techniques in real-world scenarios with incomplete ground truth. This field represents a significant stride in the ability to digitally capture and model our world with high fidelity, even in the face of challenging lighting and vast spatial complexity, a feat further refined by Sun et al. (2022), who introduced a proficient methodology for reconstructing the surface geometry of expansive scenes, adeptly handling fluctuating lighting conditions. Their technique is predicated on the utilization of appearance embeddings to model the variability in illumination, with an emphasis on generating mesh outputs. Meshes provide a direct representation of the scene's geometry and can be readily imported into standard graphics pipelines. To reconstruct the surface geometry, the approach leverages volume rendering methods, coupling a neural surface representation with volumetric rendering. However, integrating the surface representation with a volumetric radiance field involves huge compute demands for large-scale data collections, making it intractable in settings with limited access to high-end GPUs. To address this issue, the paper proposes a hybrid voxel- and surface-guided sampling technique. The approach uses sparse point clouds from SfM to create samples from a sparse volume. This voxel-guided method is combined with a surface-guided sampling technique that creates samples based on the current state of optimization. The paper also introduces Heritage-Recon, a new benchmark dataset derived from the public catalog of free-licensed LiDAR data available in Open Heritage 3D. The dataset is paired with Internet-derived image collections and SfM models from the MegaDepth dataset, and a carefully designed evaluation protocol suited for such large-scale scenes with incomplete ground truth is used to evaluate the approach. The results demonstrate that the proposed approach surpasses classical and neural reconstruction methods in terms of efficiency and accuracy. An interesting work by Wu et al. (2022) uses a combination of implicit neural representation and explicit multiplane images to represent 3D scenes as shown in Figure 3. Their method which is called Implicit Multiplane Images representation (ImMPI) employs a learning-based network for ImMPI initialization, which involves extracting 3D scene distribution priors. This approach accelerates and stabilizes the optimization process. Furthermore, the paper presents a new dataset for remote sensing novel view synthesis, which includes 16 real-world 3D scenes collected from Google Earth along with their multi view images. The dataset covers various terrains such as mountains, urban areas, buildings, parks, and villages.

2.4 Summary

The exploration of image-based facade reconstruction encompasses a variety of methodologies, notably photogrammetry and image matching, deep learning and hybrid methods, and neural rendering approaches. These techniques offer innovative solutions to the intricate task of digitally capturing the architectural essence of building facades. Despite their advancements and the promising outcomes they present, several potential criticisms and areas for further scrutiny emerge:

2.4.1 Dataset limitations

The effectiveness of these methods is often demonstrated on curated datasets, which, while valuable, may not encapsulate the full complexity and diversity of urban facades encountered in real-world settings. The generalizability of these techniques across varied architectural styles and environments remains a question, underscoring the need for broader dataset evaluations to ascertain their applicability.

2.4.2 Complexity and computational requirements

The reliance on advanced deep learning models, such as those employed in neural rendering and some hybrid methods, necessitates substantial computational resources. The scalability of these approaches for comprehensive urban reconstruction projects is a concern, with significant demands on processing power, memory, and data storage posing potential barriers to widespread adoption.

2.4.3 Accuracy and robustness

Despite showcasing high-quality reconstructions, the precision and reliability of these methods under less-than-ideal conditions–such as poor lighting, occlusions, and the presence of dynamic elements–warrant further investigation. The robustness of these reconstruction techniques in accurately capturing the nuances of facade geometries and textures in varied environmental conditions is crucial for their advancement.

2.4.4 Limitations of individual techniques

Each reconstruction approach has inherent strengths and weaknesses. For instance, photogrammetry and image matching techniques may struggle with feature ambiguity in densely textured urban scenes. Deep learning methods, while powerful, often require extensive labeled datasets for training, which can be difficult to

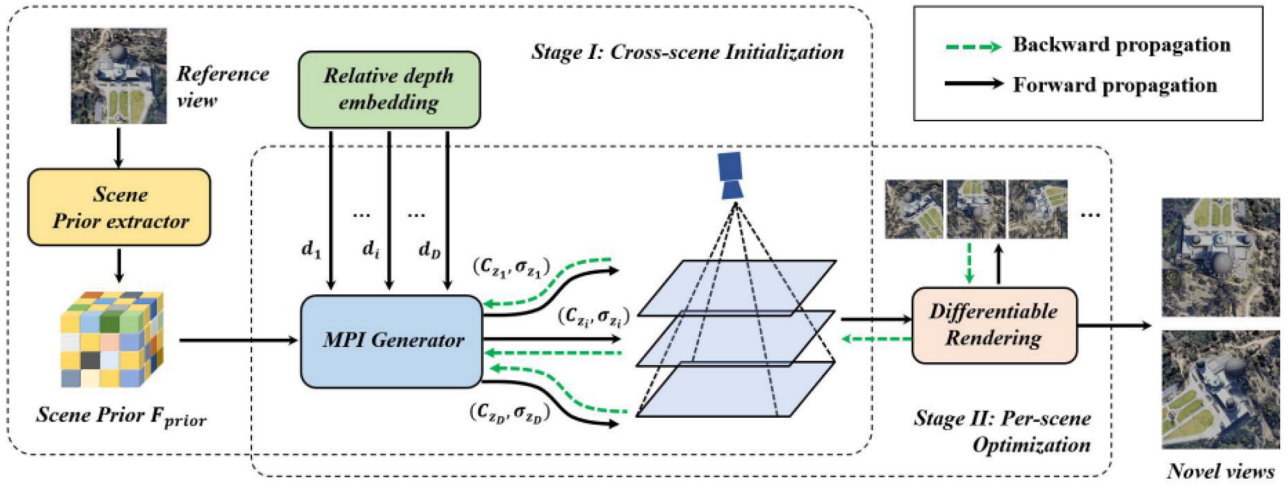

FIGURE 3
Their proposed method involves two main stages. The first stage is the cross-scene initialization, where they select a reference view and use a prior extractor encoder to generate a latent feature called F_{prior}. This feature serves as the initialization for the ImMPI. The encoder is trained on various Remote Sensing (RS) scenes using self-supervised learning. The second stage is per-scene optimization, where they take a set of input views and iteratively refine the ImMPI representation. This refined representation is then used for rendering novel views through differentiable rendering techniques Wu et al. (2022).

procure. Neural rendering approaches, though promising for their realism and detail, face challenges in computational efficiency and the need for high-quality training data.

2.4.5 Lack of comparison with traditional methods

A comprehensive evaluation that includes comparisons with traditional reconstruction methods is often missing. Such comparisons are essential to highlight the advancements these new methodologies offer over conventional techniques, providing a clearer understanding of their added value and potential limitations.

In summary, while the discussed studies introduce cutting-edge approaches to facade reconstruction, addressing the highlighted criticisms is essential for their evolution. Future research should aim to enhance dataset diversity, improve computational efficiency, and ensure the accuracy and robustness of these methods in varied real-world scenarios. Additionally, comparative analyses with traditional reconstruction techniques could further validate the effectiveness and innovation of these advanced methodologies. Tables 2–4 offers a detailed summary of the methods employed in each study, including the input data utilized and the evaluation metrics applied, alongside the year of each study.

3 District reconstructions

District reconstruction refers to the task of creating a 3D model of a entire district or urban area. This can be done using satellite imagery, aerial photographs, and street view imagery, either separately or together. The goal is to capture the spatial and structural characteristics of the district in order to create an accurate and comprehensive 3D model. We have arranged the existing research on image-based district reconstruction into five sections: semantic-based, neural implicit-based, depth-based, generative adversarial network-based, and combinatorial methods.

Semantic-based 3D reconstruction strives to simultaneously capture the geometric shape and semantic information of a 3D scene from a collection of 2D images, whether taken from a single viewpoint or multiple viewpoints. This technique involves a multi-step process encompassing tasks such as feature extraction, camera pose estimation, depth estimation, semantic labeling, and the fusion of 3D information. The essence of semantic-based 3D reconstruction lies in its ability to not only recreate the physical structure of the scene but also to associate meaningful labels with objects, enabling a richer understanding of the reconstructed environment and its elements.

Neural implicit-based 3D reconstruction is an innovative paradigm that leverages neural networks to infer complex 3D shapes from 2D images or sparse data points. This approach utilizes implicit functions, which are mathematical descriptions that can represent intricate shapes without explicitly defining them. Neural implicit-based methods excel at capturing fine details and handling diverse shapes that may be challenging for traditional geometric representations. This technique holds significant promise for advancing 3D reconstruction by harnessing the power of machine learning to derive accurate and intricate 3D models from limited visual input.

Depth-based 3D reconstruction is a foundational methodology centered on extracting three-dimensional information from images by estimating the distances to various points in a scene. This approach relies on techniques such as stereo vision, where the disparity between corresponding points in multiple images is used to calculate depth. Depth-based methods offer a straightforward and accurate means of reconstructing object shapes and spatial relationships. However, they often require well-calibrated cameras and controlled lighting conditions to yield reliable results. Despite these constraints, depth-based 3D reconstruction remains a critical technique in computer vision, particularly in

TABLE 2 Works on photogrammetry and image matching for 3D reconstruction.

References	Input data	Methods	Year	Evaluation metric
Wu et al. (2018)	Aerial oblique imagery and terrestrial imagery	Feature matching + combined bundle adjustment	2018	CMD
Li et al. (2020)	Aerial and terrestrial images	Combined SfM + optimal generation of textured 3D mesh models	2020	CMD
Guo and Guo (2018)	Multi-view images	Joint estimation of the line and dense points	2018	NA
Xu et al. (2020)	Multiple view stereo images	MARMVS	2020	R, A, HM

R, recall; A, accuracy; HM, harmonic means; CMD, cloud-mesh distance.

TABLE 3 Works on deep learning and hybrid methods for 3D reconstruction.

References	Input data	Methods	Year	Evaluation metric
Alidoost et al. (2020)	Single aerial image	Y-shaped CNN	2020	RMSE, NMAD
Huang et al. (2020)	Airborne and terrestrial imagery	Shell model	2020	NA
Fan et al. (2021)	Internet images	CNN + VGI3D	2021	P, R, I
Tripodi et al. (2020)	Stereo pairs of satellite images	CNN + optimization + computational geometry	2020	NA
Romanoni et al. (2017)	Multi-View Stereo	Variational surface evolution framework	2017	A, R, F-score
Roy et al. (2022)	2D images	Facade detection + frontalization module	2022	L1
Bacharidis et al. (2020)	Single RGB image	Auto-encoder + GAN + computational geometry techniques	2020	NA

R, recall; A, accuracy; I, Integrality; P, Precision; RMSE, Root Mean Square Error; NMAD, Normalized Median Absolute Deviation.

TABLE 4 Works on neural rendering based 3D reconstruction.

References	Input data	Methods	Year	Evaluation metric
Sun et al. (2022)	2D image collections	Neural radiance fields	2022	P, R, F1
Wu et al. (2022)	Aerial images	ImMPI	2022	PSNR, SSIM, LPIPS

R, recall; P, precision; F1, F1 Score; PSNR, Peak Signal-to-Noise Ratio; SSIM, Structural Similarity Index Measure; LPIPS, Learned Perceptual Image Patch Similarity.

applications such as robotics, augmented reality, and autonomous navigation systems.

GAN-based 3D reconstruction is an innovative approach that harnesses the power of adversarial learning to enhance the process of generating three-dimensional models from two-dimensional images. GANs consist of two neural networks, a generator, and a discriminator, engaged in a competitive learning process. The generator creates candidate 3D models from input images, while the discriminator evaluates their similarity to real 3D shapes. This interplay refines the generator's ability to create increasingly accurate 3D representations. GAN-based 3D reconstruction has shown remarkable success in producing intricate and realistic 3D models, often surpassing conventional methods. However, training GANs can be complex and resource-intensive, requiring extensive datasets and careful tuning.

3.1 Semantic based district reconstruction

The use of semantic information in 3D reconstruction has become an increasingly popular research area in recent years, with many studies focusing on the integration of object recognition and classification algorithms with traditional reconstruction methods. In particular, the reconstruction of urban districts presents a unique challenge due to the complexity and diversity of the built environment, making it an area of active research and innovation. In this regard Vineet et al. (2015) proposes an end-to-end system that can perform real-time dense stereo reconstruction and semantic segmentation of outdoor environments by outputting a per voxel probability distribution instead of a single label. The system is designed to incrementally build dense large-scale semantic outdoor maps and can handle moving objects more effectively than previous approaches by incorporating knowledge of object classes into the reconstruction process. The core of the system is a scalable fusion approach that replaces the fixed dense 3D volumetric representation of standard formulations with a hash-table-driven counterpart. The system uses stereo instead of Kinect-like cameras or LiDARs and visual odometry instead of Iterative Closest Point (ICP) camera pose estimation. The semantic segmentation pipeline extracts 2D features and evaluates unary potentials based on random forest classifier predictions. It transfers these into the 3D volume, where a densely connected Conditional Random Field (CRF) is defined to reduce the computational burden and enforce temporal consistency. The system uses online volumetric mean-field inference and a volumetric filter suitable for parallel implementation to efficiently infer the approximate Maximum Posterior Marginal (MPM) solution. The semantic labels are used to reinforce the weights in the fusion step to handle moving objects more effectively. The approach is evaluated on the KITTI dataset, and the results show high-quality dense reconstruction and labeling of several scenes.

For generating 3D maps of large-scale environments using a combination of deep learning and probabilistic graphical modeling techniques, Yang et al. (2017) introduces a method that involves

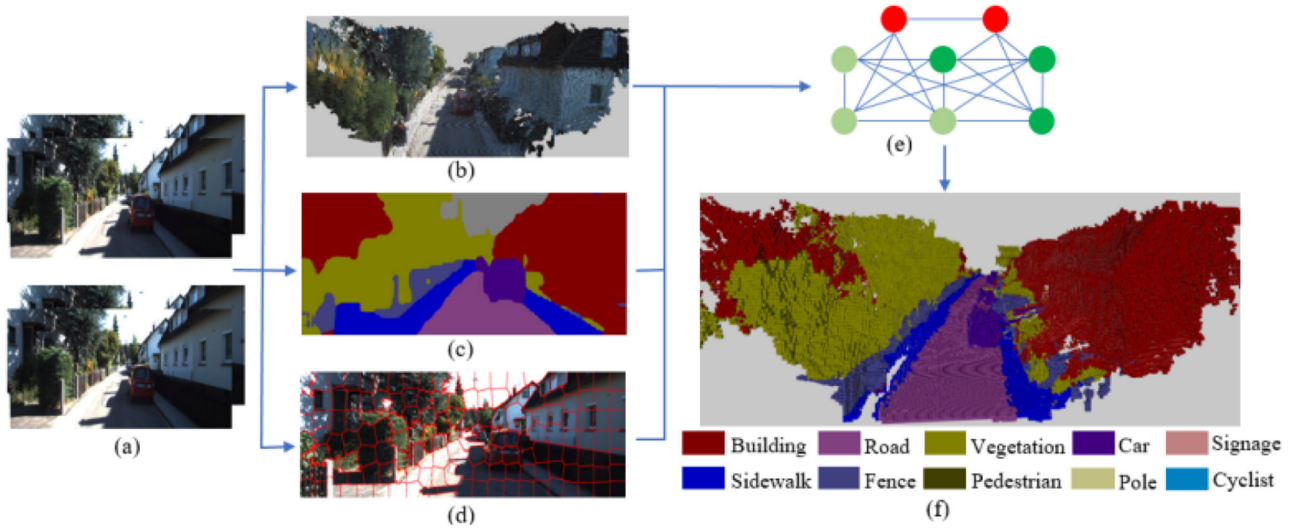

FIGURE 4
Overview of the system as originally shown in Yang et al. (2017), **(a)** Input stereo image pairs. **(b)** Geometric 3D reconstruction achieved using ORB SLAM and occupancy mapping. **(c)** 2D semantic segmentation through CNN, acting as the conditional random field (CRF) unary potential. **(d)** Generation of superpixel cliques using SLIC for ensuring label consistency within regions, which calculates the CRF high-order potential. **(e)** Introduction of the hierarchical CRF model. An efficient mean-field inference approach is developed for this hierarchical CRF in the 3D grid space. **(f)** Final semantic mapping obtained after optimization using the 3D CRF approach.

utilizing a CNN to compute pixel label distributions from 2D images and transferring this information to a 3D grid space. To enforce semantic consistency among the grids, the paper proposes a CRF model with higher-order cliques, which are generated through superpixels. An efficient filter-based mean field approximation inference is developed for this hierarchical CRF. To make the method applicable to large-scale environments and to achieve real-time computation, the paper introduces a scrolling occupancy grid that represents the world and is memory and computationally bounded. The proposed approach improves segmentation accuracy by 10% over the state-of-the-art systems on the KITTI dataset. The paper's main contributions are the proposal of a (near) real-time incremental semantic 3D mapping system for large-scale environments using a scrolling occupancy grid, the development of a filter-based mean-field inference for high-order CRFs with a robust Pn potts model by transforming it into a hierarchical pairwise model, and the improvement of segmentation accuracy over the state-of-the-art systems. Figure 4 shows the overview of the system.

For mapping and reconstructing large-scale dynamic urban environments, Bârsan et al. (2008) presents a new algorithm that is specifically designed to separate and classify objects in the environment as either static background, moving objects, or potentially moving objects. This ensures that even objects that may not be moving at the moment but have the potential to move, like parked cars, are accurately modeled. To achieve this, the authors use a combination of instance-aware semantic segmentation and sparse scene flow. The depth maps computed from the stereo input and camera poses estimated from visual odometry are used to reconstruct both the background and (potentially) moving objects separately. The sparse scene flow helps estimate the 3D motions of the detected moving objects, resulting in more accurate reconstruction. The authors have also developed a map pruning technique to improve reconstruction accuracy and reduce memory consumption, which has led to increased scalability. The system has been thoroughly evaluated on the KITTI dataset and has shown promising results. The only limitation is the instance-aware semantic segmentation, which currently acts as the primary bottleneck, but the authors suggest that this could be addressed in future work. Furthermore, Yang et al. (2018) presented a method to create a detailed 3D map of an outdoor urban environment using binocular stereo vision. The system takes a stereo color images from a moving vehicle and uses visual odometry to estimate the camera's movement and construct a 3D space around the vehicle. At the same time, the system performs semantic segmentation using deep learning technology, which helps to verify feature matching in visual odometry. This process calculates the motion, depth, and semantic label of every pixel in the input views. To generate a 3D semantic map, the system uses a voxel CRF inference technique to fuse the semantic labels to voxel. This means that the system can accurately map the semantic labels of each voxel in the 3D space, taking into account the semantic labels of neighboring voxels. This also helps to remove moving objects and improves the accuracy of the motion segmentation. The system can generate a dense 3D semantic map of the urban environment from any length of image sequence. This means that the system can continuously update and improve the map as the vehicle moves through the environment.

A fully automated 3D reconstruction from multi-view aerial images without any additional data assistance was proposed by Yu et al. (2021) that consists of three parts: efficient dense matching and Earth surface reconstruction, reliable building footprint extraction and polygon regularization, and highly accurate height inference of building roofs and bases. The first part of the method uses a novel deep learning-based multi-view matching method

to reconstruct the Digital Surface Model (DSM) and Digital Orthophoto Map (DOM) efficiently without generating epipolarly rectified images. This is done using a convolutional neural network, gated recurrent convolutions, and a multi-scale pyramid matching structure. The second part of the method introduces a three-stage 2D building extraction method to deliver reliable and accurate building contours. Deep-learning based segmentation, assisted with DSM, is used to segment buildings from backgrounds. The generated building maps are then fused with a terrain classification algorithm to improve segmentation results. A polygon regularization algorithm and a level set algorithm are employed to transfer the binary segmentation maps to structured vector-form building polygons. Finally, a novel method is introduced to infer the height of building roofs and bases using adaptive local terrain filtering and neighborhood buffer analysis. The proposed method was tested on a large experimental area that covered 2,284 aerial images and 782 various types of buildings. The results showed that the accuracy and completeness of the reconstructed models approached that of manually delineated models to a large extent, and exceeded the results of other similar methods by at least 15% for individual 3D building models in a between-method comparison, with many of them comparable to manual delineation results.

Cheng et al. (2022) proposed a 3D semantic mapping system that can reconstruct a 3D map of the environment using only stereo images and optional sensor data such as Global Navigation Satellite System (GNSS) for global positioning and IMU measurements. The pipeline of the proposed system consists of three main modules: direct visual odometry (VO), semantic segmentation, and temporally consistent labeling. The direct VO module estimates relative camera poses and a sparse 3D reconstruction of the environment. The global map optimization is performed based on loop closure detection, which detects when the vehicle revisits a previously explored area and uses this information to improve the map's accuracy. The semantic segmentation module uses a state-of-the-art neural network to generate accurate semantic labels for each pixel of the stereo images corresponding to the keyframes defined by the VO front-end. This allows the system to understand the different types of objects in the environment, such as buildings, roads, and trees. The temporally consistent labeling module generates temporally consistent 3D point labels based on the VO outputs and the 2D semantic labels. This enables the system to create a 3D semantic map that not only shows the geometry of the environment but also the semantic information associated with it. Finally, the paper discusses how the 3D semantic mapping system can be used to create a city-scale map by stitching together the reconstructions from a fleet of vehicles.

MonoScene (Cao and de Charette, 2022a), uses a single RGB image to project 2D features along their line of sight, inspired by optics, to bridge 2D and 3D networks. This approach allows the 3D network to self-discover relevant 2D features. the proposed approach uses a pipeline that combines 2D and 3D UNets, bridged by a Features Line of Sight Projection module (FLoSP) that lifts 2D features to plausible 3D locations. This module boosts information flow and enables 2D–3D disentanglement. The 3D Context Relation Prior component (3D CRP) is inserted between the 3D encoder and decoder to capture long-range semantic context. The pipeline is guided by two complementary losses. The first is a scene-class affinity Loss that optimizes the intra-class and inter-class scene-wise metrics. The second is a frustum proportion loss that aligns the classes distribution in local frustums, which provides supervision beyond scene occlusions. The paper argues that the existing literature on Semantic scene completion (SSC) mainly relies on cross-entropy loss, which considers each voxel independently and lacks context awareness. The authors propose novel SSC losses that optimize the semantic distribution of groups of voxels, both globally and in local frustums. To further boost context understanding, the authors also designed a 3D context layer to provide the network with a global receptive field and insights about the semantic relations of the voxels. The authors extensively tested MonoScene on indoor and outdoor scenes and found that it outperformed all comparable baselines and even some 3D input baselines.

Recently Huang et al. (2023) proposed a new approach to describe 3D scenes called Tri-Perspective View (TPV) representation, which extends the Bird's Eye-View (BEV) representation by including two additional perpendicular planes. The authors argue that TPV provides a more comprehensive description of the 3D structure of a scene compared to BEV, which has difficulty representing fine-grained 3D structure with a single plane. To obtain the feature of a point in the 3D space, the authors first project it onto each of the three planes and use bilinear interpolation to obtain the feature for each projected point. They then sum the three projected features as the comprehensive feature of the 3D point. To effectively obtain the TPV features from 2D images, the authors propose a transformer-based encoder called TPVFormer. TPVFormer performs image cross-attention between TPV grid queries and the corresponding 2D image features to lift 2D information to the 3D space. Then, it performs cross-view hybrid-attention among the TPV features to enable interactions among the three planes. The authors demonstrate the superiority of TPV representation by formulating a challenging task for vision-based 3D semantic occupancy prediction, where only sparse LiDAR semantic labels are provided for training and predictions for all voxels are required for testing. They evaluate their model on two proxy tasks: LiDAR segmentation on nuScenes and 3D semantic scene completion on SemanticKITTI, both using only RGB images as inputs. Their results show that TPVFormer produces consistent semantic voxel occupancy prediction with only sparse point supervision during training, and achieves comparable performance with LiDAR-based methods on LiDAR segmentation.

An interesting work called VoxFormer Li et al. (2023b) is designed to generate complete 3D volumetric semantics from 2D images. The goal is to enable AI systems to imagine the complete 3D geometry of occluded objects and scenes, which is a vital ability for recognition and understanding. The framework has a two-stage design, where the first stage involves starting from a sparse set of visible and occupied voxel queries obtained from depth estimation. This is followed by a densification stage that generates dense 3D voxels from the sparse ones. The authors propose that starting with the featurization and prediction of the visible structures is more reliable, as the visual features on 2D images correspond only to the visible scene structures rather than the occluded or empty spaces. To propagate information to all the voxels, the authors

apply a masked autoencoder design to the sparse queries using self-attention. The framework uses Transformer-based architecture for this purpose, and they call it VoxTransformer. VoxTransformer is trained to predict dense 3D voxel occupancy and semantic labels given the sparse set of voxel queries. The authors evaluate the performance of VoxFormer on the SemanticKITTI dataset and show that it outperforms the state of the art by a relative improvement of 20.0% in geometry and 18.1% in semantics. They also report that VoxFormer reduces GPU memory during training by $\sim$ 45% to $<$16GB.

Recently, research conducted by Miao et al. (2023), suggested a method for 3D SSC that provides dense geometric and semantic scene representations. they claim accurate depth information is crucial for restoring 3D geometry, and the authors propose a stereo SSC method named OccDepth, which fully exploits implicit depth information from stereo images (or RGBD images) to help recover 3D geometric structures. To better fuse 3D depth-aware features, the authors propose the Stereo Soft Feature Assignment (Stereo-SFA) module, which implicitly learns the correlation between stereo images. Additionally, the Occupancy Aware Depth (OAD) module is used to obtain geometry-aware 3D features by knowledge distillation using pre-trained depth models. The authors also provide a reformed TartanAir benchmark, named SemanticTartanAir, for further testing their OccDepth method on SSC tasks. Extensive experiments on SemanticKITTI show that their OccDepth method achieves superior performance compared to state-of-the-art RGB-inferred SSC methods, with an improvement of +4.82% mIoU. Of this improvement, +2.49% mIoU comes from stereo images and +2.33% mIoU comes from the authors' proposed depth-aware method.

3.2 Neural implicit based district reconstruction

In the realm of district reconstruction, the use of neural implicit methods has emerged as a promising technique for accurately and intricately reconstructing complex 3D environments. This approach shows potential in overcoming the challenges associated with conventional methods, such as the need for manual intervention and limited scalability. A specific example of such a method is Block-NeRF (Tancik et al., 2022), which is a variant of Neural Radiance Fields (NeRF) Mildenhall et al. (2021) designed to represent large-scale environments with greater efficiency and accuracy. NeRF is a state-of-the-art method for rendering photorealistic 3D scenes, but it is limited in its ability to represent large-scale scenes. Thus, the authors demonstrate that by decomposing the scene into individually trained NeRFs, Block-NeRF can render city-scale scenes spanning multiple blocks, enabling rendering to scale to arbitrarily large environments and allowing per-block updates of the environment. The authors make several architectural changes to NeRF to make it robust to data captured over months under different environmental conditions. They add appearance embeddings, learned pose refinement, and controllable exposure to each individual NeRF, and introduce a procedure for aligning appearance between adjacent NeRFs so that they can be seamlessly combined, then build a grid of Block-NeRFs from 2.8 million images to create the largest neural scene representation to date, capable of rendering an entire neighborhood of San Francisco. This represents a significant advancement in the ability to represent large-scale environments using neural rendering techniques. Turki et al. (2022) describes three key challenges for using NeRFs on large-scale scenes:

1. The need to model thousands of images with varying lighting conditions, each of which captures only a small subset of the scene
2. The large model capacity required, making it infeasible to train on a single GPU
3. Significant challenges for fast rendering to enable interactive fly-throughs.

To address these challenges, the authors propose a sparse network structure that is specialized to different regions of the scene. They also introduce a geometric clustering algorithm for data parallelism, which partitions training images into different NeRF submodules that can be trained in parallel. The proposed approach is evaluated on existing datasets (Quad 6k and UrbanScene3D) as well as on drone footage collected by the authors. The results demonstrate that the proposed method improves training speed by 3x and Peak Signal-to-Noise Ratio (PSNR) by 12% compared to existing methods. The authors also evaluate recent NeRF fast renderers on top of Mega-NeRF and introduce a novel method that exploits temporal coherence. Their technique achieves a 40x speedup over conventional NeRF rendering while remaining within 0.8 db in PSNR quality, exceeding the fidelity of existing fast renderers. Overall, the paper presents a promising method for building interactive 3D environments from large-scale visual captures, addressing several challenges associated with using NeRFs on such scales.

To address the challenges of modeling 3D environments with vastly different scales, such as city scenes or landscape models Xiangli et al. (2022) presents BungeeNeRF, a progressive neural radiance field that aims to achieve level-of-detail rendering by progressively fitting distant views with a shallow base block and then appending new blocks to accommodate emerging details in increasingly closer views. The approach progressively activates high-frequency channels in NeRF's positional encoding inputs and successively unfolds more complex details as the training progresses. The paper demonstrates the superiority of BungeeNeRF in modeling diverse multi-scale scenes with varying views on multiple data sources, including city models, synthetic data, and drone-captured data. The approach supports high-quality rendering in different levels of detail, and the results show improved performance compared to traditional NeRF on multi-scale scenes (see Figure 5 for the overview of the proposed system)

A self-supervised monocular scene reconstruction method called SceneRF which only uses posed image sequences for training is presented by Cao and de Charette (2022b), the method aims to overcome the dependence on costly-acquired datasets that are typically used for 3D reconstruction from 2D images. The authors build upon recent progress in NeRF and optimize a radiance field with explicit depth optimization and a novel probabilistic sampling strategy that efficiently handles

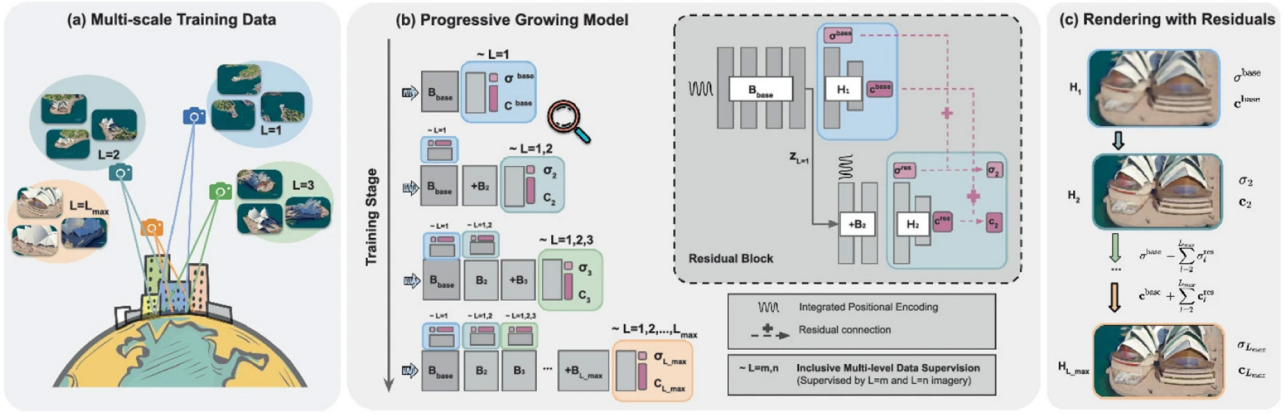

FIGURE 5
Overview of the system as originally shown in Xiangli et al. (2022), **(a)** Illustration of the multi-scale data present in city scenes. Where they use to denote scales. The model progressively evolves with the training set at each stage. **(b)** As training progresses, new residual blocks are added to the network. This addition is guided by supervision from samples encompassing the most distant scale up to the ongoing scale. The structure of a residual block is outlined in the dashed box. **(c)** The Level-of-Detail (LOD) rendering outcomes achieved at distinct residual block stages. Going from shallower to deeper levels, finer details are incrementally introduced.

large scenes. The proposed method requires only a single input image for inference and can hallucinate novel depth views that are fused together to obtain 3D scene reconstruction. Thorough experiments are conducted to demonstrate the effectiveness of the proposed method. The results show that the proposed method outperforms all recent baselines for novel depth views synthesis and scene reconstruction on both indoor BundleFusion and outdoor SemanticKITTI datasets. Overall, the proposed SceneRF method provides a promising alternative to costly-acquired datasets for 3D reconstruction from 2D images.

Zakharov et al. (2021) proposed a new method that represents movable objects separately from the static background and recovers a full 3D model of each distinct object as well as their spatial relations in the scene using only a single image. The method leverages transformer-based detectors and neural implicit 3D representations and builds a Scene Decomposition Network (SDN) that reconstructs the scene in 3D. The authors also demonstrate that the 3D reconstruction can be used in an analysis-by-synthesis setting via differentiable rendering. The method is trained only on simulated road scenes but generalizes well to real data in the same class without any adaptation thanks to its strong inductive priors. Experiments on two synthetic-real dataset pairs (PD-DDAD and VKITTI-KITTI) show that the method can robustly recover scene geometry and appearance, as well as reconstruct and re-render the scene from novel viewpoints. Overall, the method provides a new approach to scene reconstruction that can capture the 3D geometry and appearance of a road scene from a single image.

3.3 Depth based district reconstruction

Depth sensors are capable of capturing geometric and spatial information from regions with homogeneous and poor textures, which are common in indoor situations. However, these sensors have limitations when it comes to outdoor reconstruction due to their restricted range and vulnerability to sunlight interference with the patterns used for depth estimation. Considering these issues and the prevalence of cameras in modern mobile devices, it becomes practical to explore entirely vision-based solutions for creating detailed 3D models of outdoor areas. This is precisely the approach followed by Schöps et al. (2017). The approach uses plane sweep stereo to build depth maps with a fisheye camera and GPU. The study describes a set of filtering procedures for recognizing and rejecting unreliable depth measurements. Following that, the retained depth map sections are combined into a volumetric representation of the environment using a truncated signed distance function. Notably, this technique enables real-time reconstruction of large outdoor sceneries utilizing mobile devices, which was previously unfeasible. The paper extensively evaluates the proposed method and demonstrates the benefit of rigorously filtering depth maps. Overall, the approach is significant because it enables real-time reconstruction of large-scale outdoor scenes on mobile devices. Figure 6 shows the overall systems pipeline. In a recent work by Yin et al. (2022) a two-stage pipeline for 3D shape estimation from single images, consisting of a depth recovery module and a point cloud module has been proposed. The depth recovery module takes a single image as input and outputs a depth map. If sparse depth points are available, they are also used as input to the module to output a metric depth map. The point cloud module takes the predicted depth map and an initial estimate of the focal length as input and outputs shift adjustments to the depth map and the focal length to improve the geometry of the reconstructed 3D scene shape. The two modules are trained separately on different data sources and combined at inference time. The depth recovery module is trained on multiple sources of data, including high-quality LiDAR sensor data, medium-quality calibrated stereo data, and low-quality web stereo data, using a mix of heterogeneous losses depending on the quality of the data source. The point cloud module is trained on synthetic data generated from 3D models and real-world data captured using a LiDAR sensor. The proposed pipeline addresses the limitations of previous depth completion methods by improving robustness to diverse scenes,

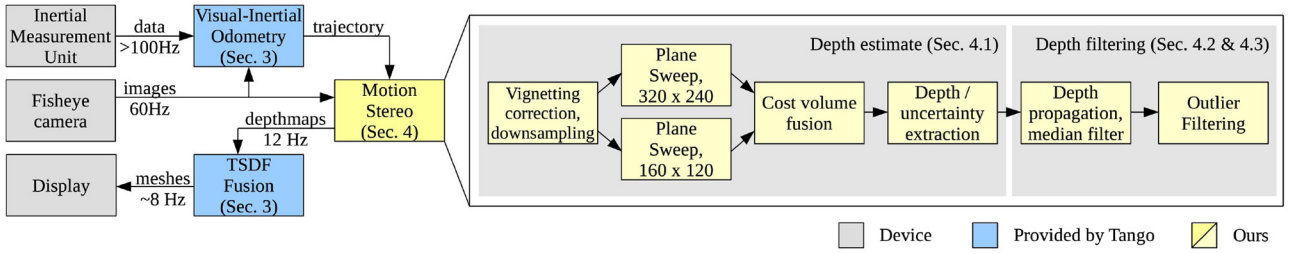

FIGURE 6
The pipeline of the mobile reconstruction system as originally shown in Schöps et al. (2017).

various sparsity patterns, and noisy inputs, and showing promising metric reconstruction results with very sparse depth points and noisy inputs.

3.4 GAN based district reconstruction

GANs have emerged as a promising tool for generating realistic and high-quality 3D models from various data sources, including images, point clouds, and voxels. By learning the underlying distribution of the input data, GANs can produce new samples that are visually and structurally consistent with the original data, while also introducing novel and diverse variations. One interesting study in this realm is Rodríguez-Santiago et al. (2021) that proposed a configured deep learning architecture as an autoencoder, but with differences in the encoder stage which is set as a residual net with four residual blocks, and the decoder stage, which is set as a generative adversarial network called a GAN-Decoder. The network takes a sequence of 2D aerial images as input, with the encoder stage extracting feature maps from the images, and the GAN-Decoder generating a point cloud based on the information obtained. The experiments show that the proposed system is capable of performing three-dimensional reconstructions of an area flown over by a drone using the point cloud generated with the deep architecture. The proposed system is compared with other works and commercial software, and the results show that it can generate reconstructions in less processing time, with less overlapping percentage between 2D images and is invariant to the type of flight path. Details of the suggested model's general configuration are shown in Figure 7.

3.5 Combinatorial based district reconstruction

Combinatorial optimization has been widely used in urban planning and design to solve complex spatial problems, such as facility location, transportation network design, and land-use allocation. Recently, researchers have explored the use of combinatorial optimization techniques for district reconstruction, which involves generating 3D models of urban districts from various data sources, such as satellite imagery, LiDAR data, and street-level photos. Combinatorial-based approaches can handle large-scale datasets and complex urban environments by partitioning the input data into smaller sub-regions and optimizing the arrangement of building blocks or other elements within each sub-region. In this regard Romanoni and Matteucci (2015) proposes a method for performing incremental urban reconstruction from a monocular video captured by a surveying vehicle, using a delaunay triangulation of edge-points to capture the sharp edges of the urban landscape. The proposed method allows online incremental mapping for tasks such as traversability analysis or obstacle avoidance. The delaunay triangulation of edge-points constrains the edges of the 3D delaunay triangulation to real-world edges, improving the accuracy of the reconstruction. The paper also introduces the inverse cone heuristic, which preemptively avoids the creation of artifacts in the reconstructed manifold surface. This is important because a manifold surface allows the application of computer graphics or photometric refinement algorithms to the output mesh, improving the visual quality of the final 3D model. The proposed approach was evaluated on four real sequences of the public available KITTI dataset by comparing the incremental reconstruction against Velodyne measurements. The results show that the proposed method achieves comparable or better results than existing methods while being computationally efficient and able to handle dynamic scenes. Another related work is by Piazza et al. (2018) that proposes a real-time incremental manifold reconstruction algorithm that runs on a single CPU and can handle large scale scenarios. Figure 8 shows an example of the reconstructed mesh using the algorithm. The proposed algorithm speeds up the run time of existing algorithms while improving the accuracy of the reconstruction. The authors achieved these results by redesigning some of the classical manifold reconstruction steps, proposing a novel shrinking method and a novel ray tracing approach that leverage on hashing and caching strategies. One of the significant advantages of the proposed algorithm is that it is also able to manage moving points without using any approximate heuristics, resulting in negligible overheads. This is in contrast to existing algorithms that require the use of approximate heuristics to handle moving points. As a possible future development, the authors plan to manage loop closures and update the map shape accordingly, which means handling the change of mesh genus when needed. They are also working on improving the accuracy of the reconstruction by incorporating shape prior, such as planes, through constrained delaunay triangulation while still preserving real-time execution.

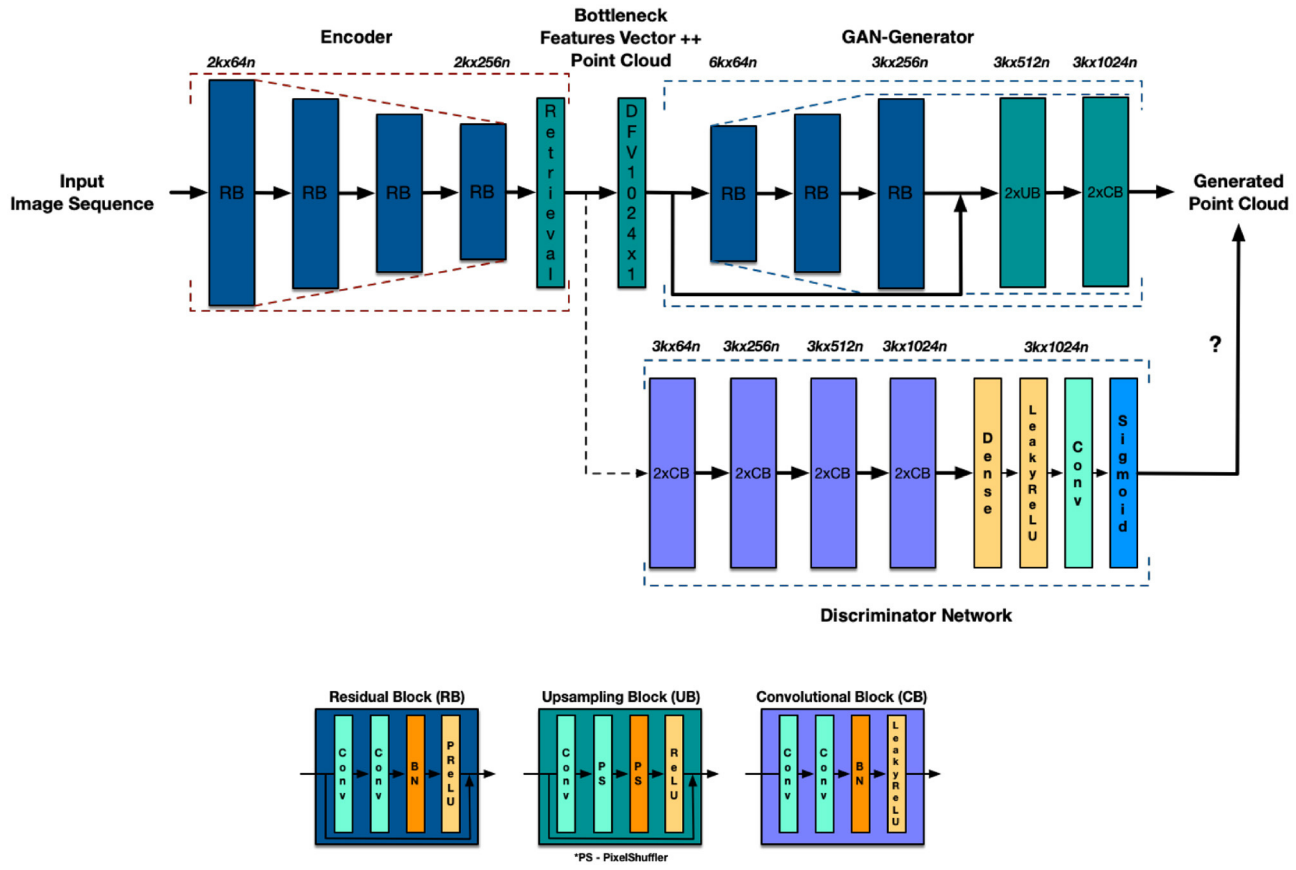

FIGURE 7
The internal layer configuration for each stage of the architecture, encompassing both the Encoder and the Decoder, is now referred to as GAN-Decoder. Specific blocks utilized in the model, including the Residual Block (RB), Upsampling Block (UB), and Convolutional Block (CB), are presented in detail at the bottom, as originally shown in Rodríguez-Santiago et al. (2021).

3.6 Summary

Overall, the mentioned studies focus on different techniques and methods for image-based district reconstruction, specifically semantic-based, neural implicit-based, depth-based, GAN-based, and combinatorial methods. While these studies propose innovative approaches and achieve promising results, there are some potential criticisms to consider:

3.6.1 Dataset limitations

Most of the studies evaluate their proposed methods on specific datasets such as KITTI, nuScenes, or SemanticKITTI. While these datasets provide valuable benchmarks, they may not fully represent the complexity and diversity of real-world urban environments. The generalizability of the proposed methods to different contexts and datasets should be further explored.

3.6.2 Complexity and computational requirements

Several studies rely on deep learning techniques, such as CNNs or GANs, which often require significant computational resources and training data. The practicality and scalability of these methods for large-scale district reconstruction need to be carefully considered, especially in terms of memory consumption, processing time, and hardware requirements.

3.6.3 Accuracy and robustness

While the aforementioned studies demonstrate high-quality reconstruction results, there may still be limitations in terms of accuracy and robustness. Factors such as occlusions, lighting variations, and dynamic objects pose challenges for accurate 3D reconstruction. Further investigation is needed to evaluate the methods' performance under various real-world scenarios and challenging conditions.

3.6.4 Limitations of individual techniques

Each category of reconstruction methods has its own strengths and limitations. For example, semantic-based methods heavily rely on object recognition and classification algorithms, which can be affected by occlusions and variations in appearance. Neural implicit-based methods may face challenges in terms of training data availability and the interpretability of learned

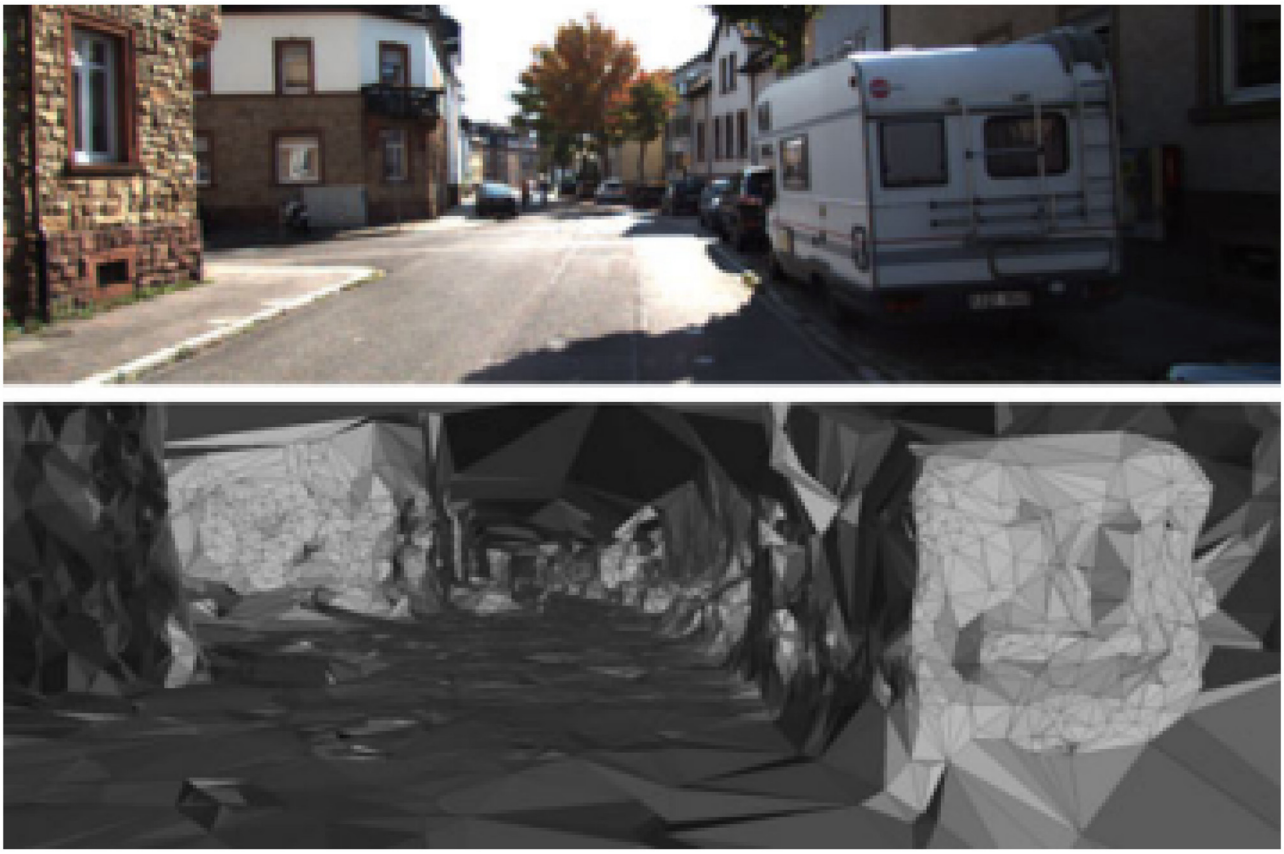

FIGURE 8
Real-time mesh reconstruction using the suggested algorithm, as originally shown in Piazza et al. (2018).

representations. Depth-based methods depend on accurate depth estimation, which can be challenging in certain scenarios. GAN-based methods may struggle with complex and diverse 3D shapes and require careful training to avoid generating unrealistic or distorted reconstructions. Combinatorial methods that combine different techniques may face integration challenges and potential trade-offs between accuracy and computational efficiency.

3.6.5 Lack of comparison with traditional methods

While the studies highlight the improvements over previous approaches, it would be beneficial to compare the proposed methods with traditional techniques, such as manual reconstruction or simpler algorithms, to assess the added value and potential limitations of the proposed approaches.

In conclusion, while the mentioned studies present innovative approaches and achieve promising results in image-based district reconstruction, further research and evaluation are needed to address the limitations and challenges associated with these methods. Additionally, considering a broader range of datasets and comparison with traditional techniques would provide a more comprehensive understanding of the strengths and weaknesses of the proposed approaches. Additionally, the following tables (Tables 5–9) offer detailed summaries of the methodologies employed in each research endeavor, including the specific types of input data utilized, the evaluation metrics applied, and the years of publication for each study.

4 City-scape reconstructions

Virtual city modeling is a huge challenge with applications in many sectors, including gaming, film, and civil engineering, but it takes a large amount of time and work, even for professionals. There have been earlier attempts to produce 3D city models using two methodologies: procedural modeling (Bulbul, 2023; Parish and Müller, 2001; Aliaga et al., 2008; Vanegas et al., 2012; Emilien et al., 2015; Kelly and McCabe, 2007) and image-based modeling (Vezhnevets et al., 2007; Barinova et al., 2018). Procedural modeling involves generating city components based on a set of rules or grammar, whereas image-based modeling creates components based on the existing street-level or aerial images. Both strategies try to recreate a genuine city, either using grammatical rules or input images, and 3D components are built only when the input image has a corresponding component. Kim et al. (2020) proposes a system for generating 3D models of cities from street-level images. The system consists of four stages: scene parsing, generation of city property vectors, generation of terrain and height map, and 3D model construction. In the first stage, the input query

TABLE 5 Works on semantic-based district 3D reconstructions.

Ref	Input data	Methods	Year	Evaluation metric
Vineet et al. (2015)	Stereo images	Hash-based technique + mean-field inference	2015	A
Yang et al. (2017)	Multi-view stereo images	CNN + CRF	2017	A, IoU
Bârsan et al. (2008)	Stereo images	Hybrid approach	2018	Semantic-aware Evaluation
Yang et al. (2018)	Stereo images	CNN + voxel CRF	2018	Average IoU
Yu et al. (2021)	Multi-view aerial images	Mask R-CNN + MA-FCN	2021	IoU
Cheng et al. (2022)	Stereo images	VO + semantic segmentation + temporally consistent labeling	2022	IoU, mIoU
Cao and de Charette (2022a)	Single RGB image	MonoScene	2022	IoU, mIoU
Huang et al. (2023)	RGB images	TPVFormer	2023	IoU, mIoU
Li et al. (2023b)	RGB images	VoxFormer	2023	IoU, mIoU
Miao et al. (2023)	Stereo images	OccDepth	2023	IoU, mIoU

A, Accuracy; IoU, Intersection over Union; mIoU, Mean Intersection over Union.

TABLE 6 Works on neural implicit-based district 3D reconstructions.

Ref	Input data	Methods	Year	Evaluation metric
Tancik et al. (2022)	RGB images	Block-NeRF	2022	PSNR, SSIM, LPIPS
Turki et al. (2022)	RGB images	Mega-NeRF	2022	PSNR, SSIM, LPIPS
Xiangli et al. (2022)	Multi-view stereo images	BungeeNeRF	2022	PSNR, SSIM, LPIPS, Mean PSNR
Cao and de Charette (2022b)	Multi-view stereo images	SceneRF	2022	IoU, P, R
Zakharov et al. (2021)	Multi-view stereo images	Scene Decomposition Network	2021	PSNR, SSIM, LPIPS

R, Recall; P, Precision; PSNR, Peak Signal-to-Noise Ratio; SSIM, Structural Similarity Index Measure; LPIPS, Learned Perceptual Image Patch Similarity; IoU, Intersection over Union.

TABLE 7 Works on depth-based district 3D reconstructions.

Ref	Input data	Methods	Year	Evaluation metric
Schöps et al. (2017)	Stereo images	Monocular motion stereo	2017	A, Completeness
Yin et al. (2022)	RGB images	Affine-invariant depth map + locally weighted linear regression	2022	LSIV

A, Accuracy; LSIV, Locally Scale-Invariant RMSE.

TABLE 8 Works on GAN-based district 3D reconstructions.

Ref	Input data	Methods	Year	Evaluation metric
Rodríguez-Santiago et al. (2021)	Sequences of 2D aerial images	Modified autoencoder architecture	2021	CD, EMD, time

CD, Chamfer Distance; EMD, Earth Mover's Distance; time, processing time.

TABLE 9 Works on combinatorial-based district 3D reconstructions.

Ref	Input data	Methods	Year	Evaluation metric
Romanoni and Matteucci (2015)	Monocular video	Delaunay triangulation of edge-points	2015	RE
Piazza et al. (2018)	Stereo image sequences	Incremental manifold reconstruction	2018	MAE

RE, Reconstruction Error; MAE, Mean Absolute Error.

image is parsed, and the city component vector is extracted by filtering segmentation labels associated with city modeling. In the second stage, a CNN model takes the original image as input to obtain a city property vector. In the third stage, the city component vector is passed through a trained GAN model to obtain terrain and height maps. Finally, the 3D city model is synthesized based on the obtained terrain and height maps by applying parameters collected from the city property vector. The paper uses paired training data of CNN and GAN models. The CNN model is trained using pairs of street-level city images and city property vectors. The segmented city images are converted to city component vectors by filtering labels associated with city modeling. Pairs of city component vectors and terrain and height maps are used to train the GAN model. The proposed system enables the generation of 3D models of cities from street-level images. The system is trained using paired data of CNN and GAN models, which enables the system to learn the relationship between the input query image and the desired 3D model output. Figure 9 illustrates the overview of the system.

A complex and multi stage pipeline for generating virtual 3D city models is presented by Singla and Padia (2021) that uses open-source libraries and in-house routines. The inputs required for generating the models are high-resolution satellite imagery, high-resolution Digital Elevation Model (DEM), and vector shape files from OpenStreetMap (OSM). The paper highlights that virtual 3D city models are used for various applications, including smart city, virtual reality, disaster management, education, tourism, and real estate services. The proposed approach involves mosaicking preprocessed DEM scenes and images, resampling datasets using the cubic-resampling algorithm, registering the image, DEM, and vector layers using phase correlation algorithm, and extracting height information from the DEM and shape files. The vector layers are registered using latitude and longitude and overlaid on the

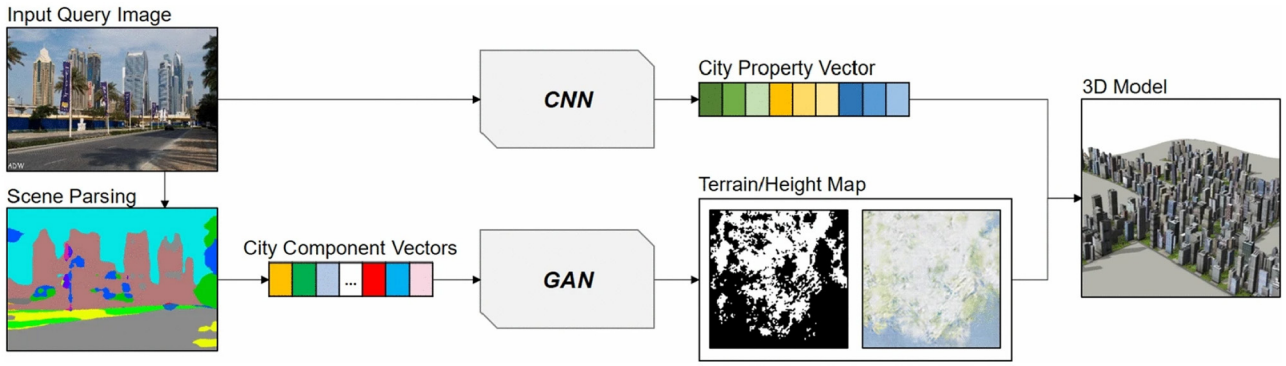

FIGURE 9
The system's overview outlines its functionality into two primary sub-tasks: (1) employing a CNN to produce a city property vector using the input query image, and (2) utilizing a GAN to generate terrain and height maps based on the segmentation outcome of the input query image. The final 3D city model is constructed by merging the two outcomes from the learning model, as originally shown in Kim et al. (2020).

3D city model to provide information about places, roads, and water bodies. To make the virtual city model more realistic, virtual trees are placed in open spaces, and sun position, day/night effects, ambient condition effects, and building texture are incorporated. For efficient visualization and rendering of the model, the paper uses the OpenSceneGraph library and a tiling-based approach. The tiles are replaced in the main memory by their sub-tiles based on the viewer's position. The proposed approach is cost-effective and can be used to develop virtual 3D city models of any given area, provided the required inputs are available. The paper primarily focuses on the approach and the advantages of this new cost-effective approach to develop virtual 3D city models using Indian Remote Sensing datasets.

Another multi stage work is presented by Cheng et al. (2022) that includes direct visual odometry, global map optimization, semantic segmentation, and temporally consistent labeling. The direct VO stage estimates the relative camera poses and a sparse 3D reconstruction of the environment. The global map optimization stage improves the accuracy of the map by detecting loop closures and optimizing the map based on this information. The semantic segmentation stage uses a state-of-the-art neural network to generate accurate semantic labels for each pixel of the stereo images corresponding to the keyframes defined by the visual odometry front-end. The temporally consistent labeling stage generates temporally consistent 3D point labels based on the visual odometry outputs and the 2D semantic labels. The paper demonstrates that the proposed pipeline can be used to create city-scale maps based on a fleet of vehicles. This is a significant achievement since it shows that purely vision-based mapping systems can generate accurate and detailed maps of the environment. The paper also suggests that the pipeline can be extended to extract information like lane markings, which could be used for autonomous driving applications.

4.1 Summary

The studies mentioned have several limitations that warrant critical consideration. Firstly, Kim et al. (2020) rely heavily on paired training data, which can be labor-intensive and resource-demanding to obtain. The lack of comprehensive discussion on the accuracy and realism of the generated 3D city models raises concerns about their reliability for various applications. Similarly, Singla and Padia (2021) primarily emphasize the advantages and cost-effectiveness of their approach without thoroughly addressing the fidelity and accuracy of the resulting models. The proposed method's applicability may also be restricted to specific regions or datasets, limiting its generalizability. Furthermore, Cheng et al. (2022) primarily focus on generating accurate maps rather than explicitly addressing the quality and realism of the 3D city models. The scalability of their pipeline and potential computational challenges remain unexplored. Collectively, these limitations highlight the need for further research to address issues related to data availability, model accuracy, realism, and computational scalability to enhance the overall utility and reliability of virtual 3D city modeling approaches. Table 10 provides an overview of each study, detailing the various types of input data used, the methodologies implemented, the year of publication, and the evaluation metrics applied.

5 Common datasets

This section curates a collection of prominent datasets employed for large-scale 3D reconstruction from images. We explore their diverse applications and impact across various fields.

5.1 Outdoor scene reconstruction

BigSfM by Cornell University: This project offers a rich repository of Structure-from-Motion (SfM) datasets, including Quad 6K (Crandall et al., 2012), Dubrovnik6K (Li et al., 2010), and Rome16K (Li et al., 2010). Primarily sourced from public platforms like Flickr and Google, these datasets target reconstructing outdoor scenes of city landmarks.

TABLE 10 Works on city-scape reconstruction.

Ref	Input data	Methods	Year	Evaluation metric
Kim et al. (2020)	Street-level images	CNN + GAN	2020	NA
Singla and Padia (2021)	Satellite imagery, DEM, Vector shape files	Open-source libraries and in-house routines	2021	NA
Cheng et al. (2022)	Stereo images	Direct VO + global map optimization + semantic segmentation + temporally consistent labeling	2022	IoU, mIoU

IoU, Intersection over Union; mIoU, Mean Intersection over Union.

5.2 Detailed disparity maps

WHU-Stereo (Li et al., 2023a): this dataset leverages LiDAR data alongside imagery to create detailed disparity maps. While offering high-quality depth information, it presents challenges due to seasonal variations. US3D (Bosch et al., 2019): this collection utilizes WorldView-3 observation satellite imagery to generate disparity maps, facilitating 3D reconstruction, but with potential limitations arising from seasonal changes.

5.3 Datasets for robotics and perception

KITTI-360 (Liao et al., 2022): designed for autonomous driving and mobile robotics research, this comprehensive dataset provides a suite of sensor data, including LiDAR and camera feeds. Its diverse data streams prove invaluable for various computer vision tasks related to object detection, localization, and mapping.

5.4 Photogrammetry and remote sensing

ISPRS Benchmark Datasets (Nex et al., 2015): catering to the photogrammetry and remote sensing community, these datasets offer high-resolution aerial imagery with precise ground truth information. This enables complex analyses like 3D building reconstruction and semantic segmentation.

5.5 3D modeling resources

Microsoft's Bing Maps Streetside (Pendleton, 2010): this dataset offers panoramic street-level imagery, a valuable resource for tasks related to urban planning and infrastructure management. Google Earth: This platform enables users to create custom 3D reconstructions by leveraging its extensive collection of satellite imagery and aerial photography.

5.6 Object detection challenges

SpaceNet (Weir et al., 2019) Multi-View Overhead Imagery: this dataset features multi-angle, annotated imagery, specifically designed to challenge and advance object detection algorithms in satellite and aerial imagery.

5.7 Diverse environments for 3D modeling

ETH3D (Schops et al., 2017), Tanks and Temples (Knapitsch et al., 2017): these datasets encompass high-definition camera footage and scenes captured by Unmanned Aerial Vehicles (UAVs). They provide diverse environments for detailed 3D modeling tasks.

5.8 Extensive image collections

GL3D (Yao et al., 2020; Luo et al., 2018) and UrbanScene3D (Lin et al., 2022): these datasets offer large collections of high-resolution images, well-suited for both SfM and Multi-View Stereo (MVS) techniques. They play a crucial role in in-depth studies of urban scene reconstruction. In conclusion, these prominent datasets empower a wide range of applications in 3D reconstruction. They continuously push the boundaries of accuracy and complexity in the models we can create.

6 Common performance metrics in large-scale image-based 3D reconstruction

This section introduces common performance metrics utilized in the evaluation of large-scale image-based 3D reconstructions, grouped by their specific application and measurement focus.

6.1 Geometric accuracy

Geometric accuracy metrics assess the precision of the reconstructed model by measuring the closeness of reconstructed points to their true positions in the ground truth. These metrics are crucial for evaluating the fidelity of the 3D shapes and surfaces generated by reconstruction algorithms.

- **Accuracy (A):** measures the average proximity of the reconstructed points to their true positions, indicating geometric precision.

$$A = \frac{1}{|P|} \sum_{p \in P} \min_{p' \in P'} \|p - p'\| \quad (1)$$

- **Chamfer distance (CD):** evaluates the bidirectional nearest-point distances between models and ground truth.

$$CD(P, P') = \frac{1}{|P|} \sum_{p \in P} \min_{p' \in P'} \|p - p'\| + \frac{1}{|P'|} \sum_{p' \in P'} \min_{p \in P} \|p' - p\| \quad (2)$$

- **Cloud-mesh distance (CMD):** measures the average distance from points in a cloud to the nearest mesh surface.

$$CMD = \frac{1}{|P|} \sum_{p \in P} \min_{m \in M} \|p - m\| \tag{3}$$

6.2 Error measurement

Error measurement metrics quantify the overall accuracy and reliability of the reconstruction process by calculating the deviations between the reconstructed and actual data points. These metrics provide a comprehensive view of error distribution and magnitude, highlighting potential areas for algorithmic improvement.

- **Root Mean Square Error (RMSE):** indicates the square root of the mean of squared differences between estimated and actual values.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \tag{4}$$

- **Mean Absolute Error (MAE):** represents the average magnitude of errors between reconstructed and actual values.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{5}$$

- **Reconstruction Error (RE):** provides a general assessment of discrepancy between the reconstructed model and the ground truth.

$$RE = \frac{1}{|P|} \sum_{p \in P} \min_{p' \in P'} \|p - p'\| \tag{6}$$

6.3 Distribution and density

Distribution and density metrics evaluate how well the reconstruction captures the overall spread and density of the true data points. These metrics are essential for assessing the completeness of the reconstruction and its effectiveness in representing the entire dataset.

- **Earth Mover's Distance (EMD):** quantifies the minimum cost to transform one point cloud into another.

$$EMD(P, P') = \min_{\phi : P \to P'} \sum_{p \in P} \|p - \phi(p)\| \tag{7}$$

- **Completeness (C):** measures the proportion of the ground truth accurately captured by the reconstruction.

$$C = \frac{|\{p' \in P' : \min_{p \in P} \|p - p'\| < \delta\}|}{|P'|} \tag{8}$$

6.4 Advanced statistical metrics

Advanced statistical metrics apply sophisticated statistical methods to further analyze the accuracy and robustness of the reconstruction. These metrics are designed to provide deeper insights into the performance of 3D reconstruction methods under various conditions and scales.

- **Locally Scale-Invariant RMSE (LSIV):** adjusts RMSE to be invariant to local scaling in depth maps.

$$LSIV = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\log D_i - \log D_i' - \mu)^2} \tag{9}$$

where $\mu = \frac{1}{N} \sum_{i=1}^{N} (\log D_i - \log D_i')$

- **Normalized Median Absolute Deviation (NMAD):** provides a robust measure of error spread, less sensitive to outliers.

$$NMAD = 1.4826 \times \text{median}(|y_i - \text{median}(y)|) \tag{10}$$

TABLE 11 Summary of open problems and challenges in large-scale 3D reconstruction.

Challenge	Description
Standardization	Lack of a definitive standard for 3D shape representation across different scales (facade, district, city-scale), impacting consistency and comparability.
Training Details	The training details of 3D reconstruction networks, particularly hybrid training architectures, require further exploration to enhance model performance and efficiency.
Environmental Factors	Handling occlusions, dynamic scenes, textureless surfaces, and illumination changes presents significant challenges in accurately capturing real-world complexity.
Technical Issues	Camera calibration errors and computational complexity are significant hurdles, affecting the accuracy and feasibility of 3D reconstructions.
Scale and Diversity	The vast scale and diversity of large-scale scenes pose challenges in maintaining geometric and textural accuracy across varied environments.
Input Image Quality	High-quality input images are crucial, especially in outdoor settings, where varying conditions can impede data collection and model fidelity.
Generalizability	Limited adaptability of models trained on restricted datasets to new and diverse settings hinders the application of 3D reconstruction techniques.
Computational Demands	The computational demands of complex algorithms, including time and resource requirements, limit the scalability and practicality of current approaches.
Future Directions	Panoptic neural rendering, as a future direction, offers advantages in holistic scene synthesis and dynamic object removal, promising more accurate and contextually rich reconstructions.

6.5 Evaluation of feature detection and matching

Metrics in this category focus on the effectiveness of feature detection and matching within the reconstructed model. They measure the accuracy and quality of object detection, segmentation, and the alignment of features between the reconstructed model and the ground truth, critical for applications involving detailed and complex scene reconstructions.

- **Precision (P), Recall (R), Intersection over Union (I):** evaluate the accuracy and quality of object detection and segmentation.

$$P = \frac{TP}{TP + FP} \tag{11}$$

$$I = \frac{|A \cap B|}{|A \cup B|} \tag{12}$$

- **Harmonic Mean (HM) of precision and recall:** provides a balanced average of precision and recall rates.

$$HM = 2 \times \frac{P \times R}{P + R} \tag{13}$$

7 Discussion and conclusion

This survey has explored recent advances in image-based 3D reconstruction for large-scale outdoor environments, analyzing methodologies at three different scales: facades, districts, and cityscapes. By categorizing these approaches and summarizing their methodologies and performance, we provide a structured comparison of techniques that highlights both their strengths and limitations. The evolution of 3D reconstruction has been driven by technological advancements, shifting from classical geometric-based methods to data-driven deep learning approaches. Traditional techniques, such as photogrammetry and structure-from-motion (SfM), have long been the foundation of 3D reconstruction, relying on multi-view geometry, feature matching, and optimization techniques. While these methods achieved high geometric accuracy, they struggled with challenges such as occlusions, textureless surfaces, varying lighting conditions, and large-scale scene complexity. The emergence of deep learning has transformed the field, introducing powerful methods capable of learning features directly from data rather than relying on hand-crafted descriptors. Convolutional Neural Networks (CNNs) have improved depth estimation and object recognition, while hybrid methods combining classical techniques with neural networks have enhanced efficiency and robustness. More recently, neural implicit representations, such as Neural Radiance Fields (NeRF), have further advanced the field by offering continuous and high-fidelity 3D scene representations. Despite these advancements, there is still no definitive standard for 3D shape representation, and scalability remains a significant challenge. Large-scale 3D reconstruction continues to face unresolved issues, including the impact of environmental factors (e.g., occlusions, dynamic objects, and illumination changes), computational demands, and the need for high-quality input data. Additionally, generalizability remains a key hurdle, as many methods are trained on restricted datasets, limiting their adaptability to new and diverse environments. These challenges are summarized in Table 11, which outlines key obstacles and their implications for the field.

One promising direction for future research is panoptic neural rendering (Kundu et al., 2022; Zhang et al., 2023), which extends traditional neural rendering by incorporating comprehensive scene understanding. Unlike standard rendering techniques that focus on generating realistic views of objects, panoptic neural rendering enables the synthesis of entire scenes, considering multiple objects, lighting conditions, and interactions. This approach not only enhances the realism of 3D reconstructions but also facilitates dynamic object removal, a critical capability for applications such as autonomous navigation and digital twin environments.

By addressing these challenges and leveraging advancements in deep learning and neural rendering, the field of large-scale 3D reconstruction is poised to achieve greater accuracy, efficiency, and adaptability. We hope this survey provides valuable insights into the evolution of the field and serves as a foundation for future research, guiding efforts toward more robust and scalable reconstruction techniques.

Author contributions

AZ: Conceptualization, Visualization, Writing – original draft, Writing – review & editing. DV: Writing – review & editing. AH: Supervision, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Aliaga, D. G., Vanegas, C. A., and Benes, B. (2008). Interactive example-based urban layout synthesis. *ACM Trans. Graph* 27, 1–10. doi: 10.1145/1409060.1409113

Alidoost, F., Arefi, H., and Hahn, M. (2020). Y-shaped convolutional neural network for 3d roof elements extraction to reconstruct building models from a single aerial image. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.* 2, 321–328. doi: 10.5194/isprs-annals-V-2-2020-321-2020

Amberg, B., Romdhani, S., and Vetter, T. (2007). "Optimal step nonrigid icp algorithms for surface registration," in *2007 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE), 1–8. doi: 10.1109/CVPR.2007.383165

Bacharidis, K., Sarri, F., Paravolidakis, V., Ragia, L., and Zervakis, M. (2018). Fusing georeferenced and stereoscopic image data for 3D building façade reconstruction. *ISPRS Int. J. Geo-Inf.* 7:151. doi: 10.3390/ijgi7040151

Bacharidis, K., Sarri, F., and Ragia, L. (2020). 3D building façade reconstruction using deep learning. *ISPRS Int. J. Geo-Inf.* 9:322. doi: 10.3390/ijgi9050322

Barinova, O., Konushin, V., Yakubenko, A., Lee, K., Lim, H., and Konushin, A. (2008). "Fast automatic single-view 3-d reconstruction of urban scenes," in *Computer Vision-ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part II 10* (Springer: New York), 100–113. doi: 10.1007/978-3-540-88688-4_8

Bârsan, I. A., Liu, P., Pollefeys, M., and Geiger, A. (2018). "Robust dense mapping for large-scale dynamic environments," in *2018 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE), 7510–7517. doi: 10.1109/ICRA.2018.8462974

Bosch, M., Foster, K., Christie, G., Wang, S., Hager, G. D., and Brown, M. (2019). "Semantic stereo for incidental satellite images," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)* (IEEE), 1524–1532. doi: 10.1109/WACV.2019.00167

Bulbul, A. (2023). Procedural generation of semantically plausible small-scale towns. *Graph. Models* 126:101170. doi: 10.1016/j.gmod.2023.101170

Cao, A.-Q., and de Charette, R. (2022a). "Monoscene: monocular 3D semantic scene completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3991–4001. doi: 10.1109/CVPR52688.2022.00396

Cao, A.-Q., and de Charette, R. (2022b). Scenerf: Self-supervised monocular 3D scene reconstruction with radiance fields. *arXiv preprint arXiv:2212.02501*. doi: 10.1109/ICCV51070.2023.00861

Cheng, Q., Zeller, N., and Cremers, D. (2022). "Vision-based large-scale 3D semantic mapping for autonomous driving applications," in *2022 International Conference on Robotics and Automation (ICRA)* (IEEE), 9235–9242. doi: 10.1109/ICRA46639.2022.9811368

Crandall, D. J., Owens, A., Snavely, N., and Huttenlocher, D. P. (2012). SfM with MRFs: discrete-continuous optimization for large-scale structure from motion. *IEEE Trans. Pattern Anal. Mach. Intell.* 35:2841–2853. doi: 10.1109/TPAMI.2012.218

Duan, L., and Lafarge, F. (2016). "Towards large-scale city reconstruction from satellites," in *Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14* (Springer: New York), 89–104. doi: 10.1007/978-3-319-46454-1_6

Emilien, A., Vimont, U., Cani, M.-P., Poulin, P., and Benes, B. (2015). Worldbrush: interactive example-based synthesis of procedural virtual worlds. *ACM Trans. Graph.* 34:106–101. doi: 10.1145/2766975

Fan, H., Kong, G., and Zhang, C. (2021). An interactive platform for low-cost 3D building modeling from vgi data using convolutional neural network. *Big Earth Data* 5, 49–65. doi: 10.1080/20964471.2021.1886391

Gruen, A. (2012). Development and status of image matching in photogrammetry. *Photogrammetric Rec.* 27, 36–57. doi: 10.1111/j.1477-9730.2011.00671.x

Guo, H., and Guo, F. (2018). "Urban scene 3D reconstruction optimization leveraged by line information," in *Proceedings of the 2nd International Conference on Innovation in Artificial Intelligence*, 92–96. doi: 10.1145/3194206.3194212

Huang, H., Michelini, M., Schmitz, M., Roth, L., and Mayer, H. (2020). LOD3 building reconstruction from multi-source images. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43. doi: 10.5194/isprs-archives-XLIII-B2-2020-427-2020

Huang, Y., Zheng, W., Zhang, Y., Zhou, J., and Lu, J. (2023). Tri-perspective view for vision-based 3D semantic occupancy prediction. *arXiv preprint arXiv:2302.07817*. doi: 10.1109/CVPR52729.2023.00890

Kelly, G., and McCabe, H. (2007). "Citygen: an interactive system for procedural city generation," in *Fifth International Conference on Game Design and Technology*, 8–16.

Kim, S., Kim, D., and Choi, S. (2020). Citycraft: 3D virtual city creation from a single image. *Vis. Comput.* 36, 911–924. doi: 10.1007/s00371-019-01701-x

Knapitsch, A., Park, J., Zhou, Q.-Y., and Koltun, V. (2017). Tanks and temples: benchmarking large-scale scene reconstruction. *ACM Trans. Graph* 36, 1–13. doi: 10.1145/3072959.3073599

Kühner, T., and Kümmerle, J. (2020). "Large-scale volumetric scene reconstruction using lidar," in *2020 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE), 6261–6267. doi: 10.1109/ICRA40945.2020.9197388

Kundu, A., Genova, K., Yin, X., Fathi, A., Pantofaru, C., Guibas, L. J., et al. (2022). "Panoptic neural fields: a semantic object-aware neural scene representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12871–12881. doi: 10.1109/CVPR52688.2022.01253

Li, S., He, S., Jiang, S., Jiang, W., and Zhang, L. (2023a). Whu-stereo: a challenging benchmark for stereo matching of high-resolution satellite images. *IEEE Trans. Geosci. Remote Sens.* 61, 1–14. doi: 10.1109/TGRS.2023.3245205

Li, Y., Snavely, N., and Huttenlocher, D. P. (2010). "Location recognition using prioritized feature matching," in *Computer Vision-ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part II 11* (Springer: New York), 791–804. doi: 10.1007/978-3-642-15552-9_57

Li, Y., Yu, Z., Choy, C., Xiao, C., Alvarez, J. M., Fidler, S., et al. (2023b). Voxformer: sparse voxel transformer for camera-based 3D semantic scene completion. *arXiv preprint arXiv:2302.12251*. doi: 10.1109/CVPR52729.2023.00877

Li, Z., Wu, B., and Li, Y. (2020). Integration of aerial, mms, and backpack images for seamless 3D mapping in urban areas. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 43, 443–449. doi: 10.5194/isprs-archives-XLIII-B2-2020-443-2020

Liao, Y., Xie, J., and Geiger, A. (2022). Kitti-360: a novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 3292–3310. doi: 10.1109/TPAMI.2022.3179507

Lin, L., Liu, Y., Hu, Y., Yan, X., Xie, K., and Huang, H. (2022). "Capturing, reconstructing, and simulating: the urbanscene3D dataset," in *European Conference on Computer Vision* (Springer: New York), 93–109. doi: 10.1007/978-3-031-20074-8_6

Liu, J., and Ji, S. (2020). "A novel recurrent encoder-decoder structure for large-scale multi-view stereo reconstruction from an open aerial dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6050–6059. doi: 10.1109/CVPR42600.2020.00609

Luo, Z., Shen, T., Zhou, L., Zhu, S., Zhang, R., Yao, Y., et al. (2018). "Geodesc: learning local descriptors by integrating geometry constraints," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 168–183. doi: 10.1007/978-3-030-01240-3_11

Miao, R., Liu, W., Chen, M., Gong, Z., Xu, W., Hu, C., et al. (2023). Occdepth: a depth-aware method for 3D semantic scene completion. *arXiv preprint arXiv:2302.13540*. doi: 10.48550/arXiv.2302.13540

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. (2021). Nerf: representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 99–106. doi: 10.1145/3503250

Nex, F., Gerke, M., Remondino, F., Przybilla, H.-J., Bäumker, M., and Zurhorst, A. (2015). ISPRS benchmark for multi-platform photogrammetry. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.* 2, 135–142. doi: 10.5194/isprsannals-II-3-W4-135-2015

Pan, J., Li, J., Han, X., and Jia, K. (2018). "Residual meshnet: learning to deform meshes for single-view 3D reconstruction," in *2018 International Conference on 3D Vision* (3DV) (IEEE), 719–727. doi: 10.1109/3DV.2018.00087

Parish, Y. I., and Müller, P. (2001). "Procedural modeling of cities," in *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, 301–308. doi: 10.1145/383259.383292

Pendleton, C. (2010). The world according to bing. *IEEE Comput. Graph. Appl.* 30, 15–17. doi: 10.1109/MCG.2010.77

Piazza, E., Romanoni, A., and Matteucci, M. (2018). Real-time CPU-based large-scale three-dimensional mesh reconstruction. *IEEE Robot. Autom. Lett.* 3, 1584–1591. doi: 10.1109/LRA.2018.2800104

Qi, C. R., Su, H., Mo, K., and Guibas, L. J. (2017). "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 652–660.

Rodríguez-Santiago, A. L., Arias-Aguilar, J. A., Takemura, H., and Petrilli-Barceló, A. E. (2021). A deep learning architecture for 3D mapping urban landscapes. *Appl. Sci.* 11:11551. doi: 10.3390/app112311551

Romanoni, A., Ciccone, M., Visin, F., and Matteucci, M. (2017). "Multi-view stereo with single-view semantic mesh refinement," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 706–715. doi: 10.1109/ICCVW.2017.89

Romanoni, A., and Matteucci, M. (2015). "Incremental reconstruction of urban environments by edge-points delaunay triangulation," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE), 4473–4479. doi: 10.1109/IROS.2015.7354012

Roy, A., Kim, S., Yin, M., Yeh, E., Nakabayashi, T., Campbell, M., et al. (2022). "A learning-based framework for generating 3D building models from 2d images," in *2022 IEEE Workshop on Design Automation for CPS and IoT (DESTION)* (IEEE), 45–49. doi: 10.1109/DESTION56136.2022.00014

Schöps, T., Sattler, T., Häne, C., and Pollefeys, M. (2017). Large-scale outdoor 3D reconstruction on a mobile device. *Comput. Vis. Image Underst.* 157, 151–166. doi: 10.1016/j.cviu.2016.09.007

Schops, T., Schonberger, J. L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., et al. (2017). "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3260–3269. doi: 10.1109/CVPR.2017.272

Singla, J. G., and Padia, K. (2021). A novel approach for generation and visualization of virtual 3D city model using open source libraries. *J. Indian Soc. Remote Sens.* 49, 1239–1244. doi: 10.1007/s12524-020-01191-8

Sun, J., Chen, X., Wang, Q., Li, Z., Averbuch-Elor, H., Zhou, X., et al. (2022). "Neural 3D reconstruction in the wild," in *ACM SIGGRAPH 2022 Conference Proceedings*, 1–9. doi: 10.1145/3528233.3530718

Tancik, M., Casser, V., Yan, X., Pradhan, S., Mildenhall, B., Srinivasan, P. P., et al. (2022). "Block-nerf: scalable large scene neural view synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8248–8258. doi: 10.1109/CVPR52688.2022.00807

Tewari, A., Thies, J., Mildenhall, B., Srinivasan, P., Tretschk, E., Yifan, W., et al. (2022). Advances in neural rendering. *Comput. Graph. Forum* 41, 703–735. doi: 10.1111/cgf.14507

Tripodi, S., Duan, L., Poujade, V., Trastour, F., Bauchet, J.-P., Laurore, L., et al. (2020). "Operational pipeline for large-scale 3D reconstruction of buildings from satellite images," in *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium* (IEEE), 445–448. doi: 10.1109/IGARSS39084.2020.9324213

Tulsiani, S., Su, H., Guibas, L. J., Efros, A. A., and Malik, J. (2017). "Learning shape abstractions by assembling volumetric primitives," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2635–2643. doi: 10.1109/CVPR.2017.160

Turki, H., Ramanan, D., and Satyanarayanan, M. (2022). "Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12922–12931. doi: 10.1109/CVPR52688.2022.01258

Vanegas, C. A., Garcia-Dorado, I., Aliaga, D. G., Benes, B., and Waddell, P. (2012). Inverse design of urban procedural models. *ACM Trans. Graph* 31, 1–11. doi: 10.1145/2366145.2366187

Vezhnevets, V., Konushin, A., and Ignatenko, A. (2007). "Interactive image-based urban modeling," in *Proceedings of PIA*, 63–68.

Vineet, V., Miksik, O., Lidegaard, M., Nießner, M., Golodetz, S., Prisacariu, V. A., et al. (2015). "Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction," in *2015 IEEE international conference on robotics and automation (ICRA)* (IEEE), 75–82. doi: 10.1109/ICRA.2015.7138983

Vosselman, G., and Maas, H.-G. (2010). *Airborne and Terrestrial Laser Scanning*. CRC Press (Taylor and Francis).

Wang, R. (2013). 3D building modeling using images and lidar: a review. *Int. J. Image Data Fusion* 4, 273–292. doi: 10.1080/19479832.2013.811124

Weir, N., Lindenbaum, D., Bastidas, A., Etten, A. V., McPherson, S., Shermeyer, J., et al. (2019). "Spacenet mvoi: a multi-view overhead imagery dataset," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 992–1001. doi: 10.1109/ICCV.2019.00108

Wu, B., Xie, L., Hu, H., Zhu, Q., and Yau, E. (2018). Integration of aerial oblique imagery and terrestrial imagery for optimized 3D modeling in urban areas. *ISPRS J. Photogramm. Remote Sens.* 139, 119–132. doi: 10.1016/j.isprsjprs.2018.03.004

Wu, Y., Zou, Z., and Shi, Z. (2022). Remote sensing novel view synthesis with implicit multiplane representations. *IEEE Trans. Geosci. Remote Sens.* 60, 1–13. doi: 10.1109/TGRS.2022.3197409

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., et al. (2015). "3D shapenets: a deep representation for volumetric shapes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1912–1920.

Xiangli, Y., Xu, L., Pan, X., Zhao, N., Rao, A., Theobalt, C., et al. (2022). "Bungeenerf: progressive neural radiance field for extreme multi-scale scene rendering," in *Computer Vision-ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXII* (Springer: New York), 106–122. doi: 10.1007/978-3-031-19824-3_7

Xu, Q., Wang, W., Ceylan, D., Mech, R., and Neumann, U. (2019). Disn: Deep implicit surface network for high-quality single-view 3D reconstruction. *Adv. Neural Inf. Process. Syst.* 32.

Xu, Z., Liu, Y., Shi, X., Wang, Y., and Zheng, Y. (2020). "MARMVS: matching ambiguity reduced multiple view stereo for efficient large scale scene reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5981–5990. doi: 10.1109/CVPR42600.2020.00602

Yang, S., Huang, Y., and Scherer, S. (2017). "Semantic 3D occupancy mapping through efficient high order CRFS," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE), 590–597. doi: 10.1109/IROS.2017.8202212

Yang, Y., Qiu, F., Li, H., Zhang, L., Wang, M.-L., and Fu, M.-Y. (2018). Large-scale 3D semantic mapping using stereo vision. *Int. J. Autom. Comput.* 15, 194–206. doi: 10.1007/s11633-018-1118-y

Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., et al. (2020). "BlendedMVS: a large-scale dataset for generalized multi-view stereo networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1790–1799. doi: 10.1109/CVPR42600.2020.00186

Yin, W., Zhang, J., Wang, O., Niklaus, S., Chen, S., Liu, Y., et al. (2022). Towards accurate reconstruction of 3D scene shape from a single monocular image. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 6480–6494. doi: 10.1109/TPAMI.2022.3209968

Yu, D., Ji, S., Liu, J., and Wei, S. (2021). Automatic 3D building reconstruction from multi-view aerial images with deep learning. *ISPRS J. Photogramm. Remote Sens.* 171, 155–170. doi: 10.1016/j.isprsjprs.2020.11.011

Zakharov, S., Ambrus, R. A., Guizilini, V. C., Park, D., Kehl, W., Durand, F., et al. (2021). "Single-shot scene reconstruction," in *5th Annual Conference on Robot Learning*.

Zhang, X., Kundu, A., Funkhouser, T., Guibas, L., Su, H., and Genova, K. (2023). "Nerflets: Local radiance fields for efficient structure-aware 3D scene representation from 2d supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8274–8284. doi: 10.1109/CVPR52729.2023.00800