DOI: 10.1111/cgf.70396
EUROGRAPHICS 2026
A. Bönsch and N. Umetani
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 45 (2026), Number 2
STAR – State of The Art Report

State-of-the-art in deep learning approaches for automatic single-panorama indoor modeling and exploration

G. Pintore[1], M. Agus[2], J. Schneider[2], and E. Gobbetti[1]

[1]CRS4, Italy [2]HBKU, Qatar

Abstract

A single surround-view panoramic image provides complete coverage of the environment visible from a single viewpoint and inherently supports dynamic exploration, especially when viewed through a head-mounted display. For these reasons, single or linked 360° panoramas have become a widely adopted modality for indoor scene acquisition and virtual tour creation. Despite their popularity, panoramas present inherent limitations, as they only statically represent the captured scene, do not provide explicit 3D architectural structure and geometry, and exhibit minimal parallax due to their single-viewpoint nature, which limits their application capabilities or requires significant modeling efforts to generate missing data. In this survey, we provide an up-to-date integrative overview of recent techniques designed to overcome these challenges, bringing together complementary perspectives from machine learning, computer vision, and computer graphics. After introducing a characterization of the panoramic input and the target geometric, structural, and visual outputs, we discuss the role of reconstruction priors and motivate the choice of deep learning approaches for leveraging large-scale data to infer hidden information. Next, we outline the main sub-problems involved in lifting a 360° image into a structured, explorable model and review advances in single-view pixel-wise geometric and semantic analysis, single-view indoor layout estimation, localization and multi-room reconstruction from very sparse coverage, novel view synthesis for providing parallax, and immersive model exploration. We then discuss the emergence of both general-purpose and 360°-specific vision foundation models for single-panorama indoor modeling and exploration. Finally, we highlight practical applications and identify open research directions.

Keywords: 360, panoramic images, surround-view images, omnidirectional images, indoor reconstruction, structured reconstruction, exploration, virtual reality, extended reality

CCS Concepts
• Computing methodologies → Computer graphics; Computer vision; Machine learning; Scene understanding; Reconstruction; • Human-centered computing → Mixed / augmented reality; Virtual reality;

1. Introduction

Virtually experiencing real environments improves their accessibility and enables a wide range of applications across diverse domains. Indoor settings, such as homes, offices, and other single- or multi-room environments, are of particular interest, since they constitute the primary spaces where people spend their time and carry out activities that can benefit from virtual access [GPA24]. In real estate, for example, virtual tours allow potential buyers and renters to inspect properties without the need for physical travel, eventually also evaluating different arrangements and decorations [SAB*20]. In facility management and emergency response, they support familiarization with building locations, layouts, and connections, improving task execution performance [BZG*22]. Similar benefits extend to many other domains [TJA*25].

Of the available approaches to create virtual counterparts of real spaces, automatic reconstruction from captured data is the most practical, as manual modeling is costly and time-consuming to create, hard to maintain, and rarely reflects the true as-built or as-inhabited state of an environment. Image-based methods are especially critical, since visual information is key for location recognition. Among the many image-based capture modalities, ranging from commodity RGB cameras to RGB-D devices, 360° (also called *panoramic*, *spherical*, *omnidirectional*, or *surround-view*) photography is emerging as an especially cost-effective and efficient solution. A single panoramic shot provides full coverage of the scene from the capture viewpoint, eliminating the need for multi-view setups and complex registration pipelines [YLC*20]. Thanks to the widespread availability of 360° cameras, rapid data acquisition is now possible for both expert and casual users [JOV19]. In particular, in the real estate industry, millions of users utilize consumer-grade cameras to generate and share virtual tours with minimal effort [CHL*21a, HLB*22].

Beyond acquisition efficiency, a 360° image inherently affords a more dynamic exploration than a standard image, particularly when experienced through a head-mounted display, where natural head movements enable intuitive navigation of the field of view [XLZLC20]. For this reason, omnidirectional imagery has become a central medium for general [MCE*17] and indoor [SAB*20] interactive content generation. Furthermore, panoramas are easily shared across platforms and increasingly regarded as potential building blocks of persistent digital spaces [DL22].

Despite these advantages, panoramic imagery, whether viewed individually or as linked tours, also suffers from notable limitations. In typical indoor scenarios, essential tasks include establishing a *sense of presence* through visual exploration and, for multi-room environments, gaining *location awareness* by understanding the current room's placement within the larger environment, and navigating across rooms or through a floorplan [ZXLL21]. Yet, a panorama captures only the appearance from a fixed point, omitting crucial spatial cues. The absence of stereopsis and motion parallax leads to a flattened perception of space [WGD*22], a limitation particularly noticeable in indoor scenes, where nearby walls and objects make the lack of depth more apparent [GPA24]. Furthermore, an effective navigation and localization also depends on explicit representations of the environment's architectural structure [ZXLL21]. Finally, many interactive applications, such as virtual staging, call for functionalities that go well beyond passive viewing of static panoramas, including the ability to clear spaces, refurnish interiors, or alter lighting conditions [ZCC16].

This article presents an integrative and up-to-date review of methods that address these challenges, bridging research from machine learning, computer vision, and computer graphics. We maintain a public repository of resources accompanying this survey (https://github.com/crs4/panostar).

2. Survey scope

This work examines methods for automatically transforming indoor panoramic shots into representations that support structural and geometric recovery, dynamic exploration, and basic editing. Our focus lies on techniques that operate mostly from a *single panoramic image*, covering both *analysis* (for model reconstruction) and *synthesis* (for immersive viewing, navigation, and editing).

Motivation The emphasis on monocular panoramas is driven by both practical relevance and methodological challenges. From a practical standpoint, consumer-grade 360° cameras have become widely available and have transformed industries such as real estate through crowd-sourcing techniques. Nowadays, most interiors are typically documented by asking individual agents or casual users to take a single panoramic picture in the center of each room. For example, users of the RICOH Theta series alone have already produced and shared over one hundred million panoramas [SSO*21]. Techniques capable of enriching this massive existing corpus, without requiring new captures, therefore carry enormous potential for real-world applications. From a scientific and technical standpoint, monocular analysis and exploration pose unique challenges that warrant specialized treatment. Reconstructing geometry and structure from a single image is inherently an ill-posed problem, requiring inference under strong ambiguity and relying heavily on architectural assumptions and data-driven priors, without the possibility to exploit multi-view consistency constraints or direct geometric evidence. Moreover, immersive exploration and editing of single panoramas demand specialized methods to generate parallax cues for head-mounted displays, while also opening opportunities for streamlined, image-based editing workflows.

Content In the following, after providing the relevant background (Sec. 4), we first examine automatic modeling solutions that enrich panoramas with geometric, structural, and semantic information. This includes the progression from per-pixel depth estimation to the recovery of room layouts and inter-room connectivity under sparse interior sampling (Sec. 5). We then explore how these enriched representations enable dynamic interaction, discussing view-synthesis methods for stereo parallax and motion within and across rooms, as well as techniques for scene customization, such as styling and refurnishing (Sec. 6). Next, we discuss how foundation and large pretrained models provide a promising direction for single-panorama inference by leveraging broad prior knowledge and reducing reliance on task-specific datasets (Sec. 7).. Finally, we review practical applications (Sec. 8) and identify open research challenges (Sec. 9).

Selection criteria This article surveys recent work from major graphics, vision, and machine learning journals and conference proceedings, complemented by relevant preprints. Papers were selected for importance and relevance, relying on our judgment to provide a structured overview of key advances, without claiming completeness. Rather than exhaustive coverage, we highlight representative and seminal contributions and emphasize the links between modeling, synthesis, and immersive exploration. By situating these works within a unified perspective on panoramic imaging research, we aim to clarify how diverse research directions converge to advance the understanding and use of panoramic imaging.

3. Related surveys

Several well-established surveys have covered classic geometry reasoning methods for indoor reconstruction [PMG*20], as well as more recent learning-based solutions [PPL*24, WL24a]. Their focus, however, is mostly on analysis and synthesis from point-cloud or multi-view data, also covered in other surveys dedicated specifically to RGB-D input [ZSG*18, LGW*22, ZGS*24]. We cover, instead, the specifics of monocular analysis and 360° input and restrict the target to what is needed to transform input to an explorable model. 360° imagery is also covered in other recent surveys of general 3D scene reconstruction and understanding [dSPMLJ22, GYS*22] and immersive extended reality applications [TJA*25]. The survey by Zou et al. [ZSP*21] provides a comparative study of state-of-the-art Manhattan-World layout reconstruction methods from a single 360° image, thereby covering a focused subset of our objectives. We build on their work for layout reconstruction, but our scope is broader: addressing environments beyond strict orthogonal wall assumptions and extending to research areas beyond layout reconstruction alone.

Reconstructing and exploring indoor environments from minimal panoramic input remains a challenging problem due to the limited depth and structural cues inherent in monocular imagery, the geometric and material complexity of indoor scenes, and the requirements

of exploration-oriented applications. These challenges have given rise to several distinct research directions within 3D reconstruction and omnidirectional image analysis, which have been addressed from different perspectives in existing survey literature.

Structured indoor scene reconstruction has been extensively studied in the classical survey by Pintore et al. [PMG*20], which focused on recovering approximate structured geometry linked to visual representations from sparse and incomplete observations. These observations are mainly derived from multi-view imagery or point clouds and analyzed through geometric reasoning. Beyond detailing the core challenges of structure extraction, this work highlighted the role of indoor-specific priors (especially geometric and structural ones) and critically examined the limitations of purely geometric approaches. More recent surveys have extended this analysis to more recent learning-based methods that exploit data-driven priors [PPL*24, WL24a]. However, these reviews also emphasize multi-view or point-cloud data, often supported by RGB-D sensors, whose use has been comprehensively reviewed elsewhere [ZSG*18, LGW*22, ZGS*24]. In contrast, our work focuses on monocular 360° imagery for indoor environments, as its intrinsic properties yield a distinct and well-defined range of methodological solutions. For instance, the 360° field of view enables holistic single-image reasoning and often supports more consistent reconstructions than conventional pinhole imagery, while simultaneously introducing distinctive challenges related to geometric distortion, non-uniform sampling, and large-resolution data processing.

Panoramic imagery has been examined in broader surveys on 3D scene reconstruction and understanding [GYS*22, YGH23] and in the context of immersive extended reality applications [TJA*25]. Although these studies establish the importance and distinctive properties of omnidirectional data, they do not specifically address single-view indoor reconstruction or exploration. In particular, Gao et al. [GYS*22] review panoramic imaging principles, system architectures, and applications such as autonomous driving and robotics, while Tukur et al. [TJA*25] provide a scoping review of technological frameworks, applications, and limitations of panoramic-based extended reality (XR) systems. We partially build on these surveys for general concepts and applications. Yu et al. [YGH23], instead, investigate the application of deep learning to the specific configuration of top-view omnidirectional imagery. Although this setup can support structural inference indoors, the surveyed applications primarily target ambient-assisted living and surveillance, thereby complementing the works reviewed here, emphasizing human and object detection and pose estimation rather than scene exploration.

The survey by Da Silveira et al. [dSPMLJ22] analyzed methods for 3D scene geometry recovery from one, two, or multiple panoramas across indoor and outdoor environments. Building on this foundation, our work revisits single-panorama depth and layout estimation with an updated view of recent advances and substantially broadens the scope to include multi-room reconstruction, visual data completion, and exploration-oriented methods. Layout reconstruction itself has been the focus of a comparative study by Zou et al. [ZSP*21], which systematically evaluated Manhattan-world layout recovery from a single panorama (i.e., assuming orthogonality among walls, floor, and ceiling; see Sec. 4.3). We extend beyond these assumptions to address non-Manhattan indoor scenes and also expand the scope from pure static monocular reconstruction toward dynamic exploration. Additionally, we update and unify the analyses of these two classical surveys by incorporating recent trends, including the growing prominence of generative and diffusion-based models and the emergence of foundation models.

Several recent surveys partially address these emerging directions. Meng et al. [MZZ*25] review methods for 3D indoor scene geometry recovery from omnidirectional images, with a focus on depth estimation, single-room layout reconstruction, and object recovery. Our work expands the analysis of these topics, while also addressing multi-room reconstruction, navigation in complex indoor environments, and the generation of interactive exploration experiences. Ai et al. [ACW25] provide a broad analysis of omnidirectional image processing from a representation learning perspective, covering imaging principles, learned representations, optimization strategies, and a taxonomy of deep learning methods across tasks and applications. While providing valuable context on omnidirectional vision, this survey does not specifically address single-view 3D indoor geometry or exploration. Nevertheless, its detailed analysis of representation learning concepts offers complementary insights that enrich and contextualize the scope of our survey.

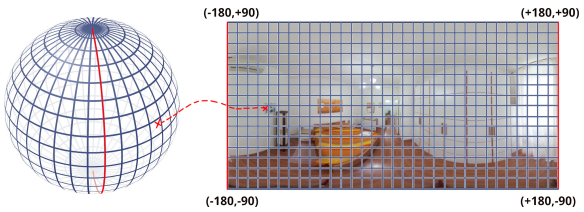

Figure 1: ***Equirectangular projection (ERP).*** *Longitude (azimuth) maps to the horizontal axis and latitude (elevation) to the vertical, with horizontal wrap-around and the poles spanning the top and bottom edges.*

4. Background

Before reviewing state-of-the-art methods, we first establish a common background. We begin by introducing the most common representations adopted for omnidirectional image capture and processing (Sec. 4.1), highlighting the dominance of the equirectangular format. Next, we motivate the reliance of many analysis and synthesis techniques on gravity-aligned images and summarize the methods used either to capture such data or to correct orientation in a preprocessing step (Sec. 4.2). We then address the challenges of extracting geometric, structural, and visual information from incomplete and noisy 360° panoramas and review the role of architectural priors, from early geometric reasoning to their integration in modern deep learning approaches (Sec. 4.3). Given the centrality of data for learning-based methods, we also provide an overview of the major publicly available datasets for training and evaluating indoor reconstruction and synthesis from 360° imagery (Sec. 4.4).

4.1. 360° image capture and processing representation

A wide range of technologies can capture 3D indoor information, from mobile laser scanners to active depth sensors [PMG*20].

Image-based solutions are particularly important, as cameras offer a practical, affordable, and widely available solution, while also providing essential visual information for tasks such as navigation, localization, and as-built reconstruction [PPG*18]. However, conventional limited field-of-view cameras fall short for full-room reconstruction, requiring multi-view acquisition that increases capture effort [GMB17] and is often hindered indoors by clutter, narrow spaces, occlusions, and textureless or reflective surfaces, which complicate feature detection and surface reconstruction [PMG*20]. For these reasons, 360° capture, providing the quickest and complete single-shot coverage, has now become the de facto standard solution [Fir16, dSPMLJ22]. While panoramic imagery can still be produced by stitching conventional photo sequences [CF14], spherical cameras have nowadays emerged as the standard solution thanks to their ease of use and reduced costs [JOV19]. With a single shutter press, commodity devices internally process multiple fisheye views to deliver a seamless full-view image, effectively capturing the entire scene in one instant. Their design guarantees an (approximately) single center of projection and consistent horizontal resolution, which are difficult to obtain through casual stitching [IHR*16, HCCJ17]. From a geometric standpoint, a spherical camera can be modeled as a unit sphere without intrinsic parameters, with image formation determined solely by extrinsic ones [dSPMLJ22].

Because the sphere cannot be mapped onto a plane without distortion, representing panoramic captures as rectangular images necessarily requires selecting and applying projection transformations. Although some devices allow access to the original unstitched fisheye views, providing maximum achievable resolution, the most common standard is only to produce a $360^\circ \times 180^\circ$ *full panoramic image* using an *equirectangular projection (ERP)* [ESLF20]. This projection maps longitude (azimuth) to the horizontal axis and latitude (elevation) to the vertical axis. As a result, meridians are represented as vertical straight lines with uniform spacing for equal angular intervals, while parallels appear as horizontal straight lines, also evenly spaced (Fig. 1). Such a mapping preserves neither areas nor angles: regions near the poles are increasingly stretched in the horizontal direction, resulting in shape distortions. Alternative projections are also employed to reduce such distortions. A notable example is the cube-map projection, which projects the sphere onto the six faces of a cube, each covering a $90^\circ \times 90^\circ$ field of view. This format is frequently used in viewing applications, including WebXR platforms and streaming services such as YouTube 360 [CDM17]. Other representations, such as tangent image projections [RAYR22a, ESLF20], have also gained traction, as reviewed in recent surveys [dSPMLJ22].

Despite some limitations, ERPs are device-independent, widely supported by cameras and HMDs, and preserve full horizontal and near-complete vertical continuity. In gravity-aligned indoor capture, resolution loss and pole distortions have limited impact, as floors and ceilings contain relatively little information (Sec. 4.2). Thus, equirectangular images dominate dataset sharing (Sec. 4.4), processing (Sec. 5), and exploration (Sec. 6).

4.2. Gravity alignment

A common assumption in both reconstruction and exploration methods is that images are captured or synthesized with the camera's up-vector aligned to the world's vertical axis. In the case of equirectangular projection, this alignment ensures that horizontal structures in the scene are aligned with image rows. Such a condition provides several advantages. First, the regions of highest distortion in equirectangular panoramas are confined to the ceiling and floor, areas that typically contain less relevant detail. Second, vertically-aligned image processing can directly exploit structural features that are also aligned with gravity [DAH20]. Finally, for rendering, this alignment maps head rotations to horizontal and vertical pixel shifts and enables stereopsis via horizontal disparities.

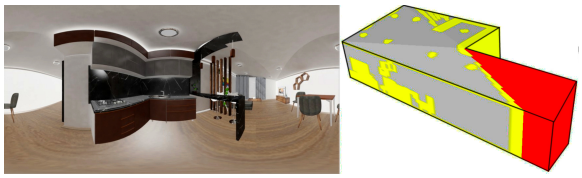

Figure 2: ***Occlusion of the architectural boundary.*** *Areas occluded by furniture are in yellow, while those occluded by other walls are in red. Image adapted from Pintore et al. [PAAG21].*

Gravity alignment is thus the standard setup for indoor capture, and nearly all publicly available 3D indoor datasets (Sec. 4.4) used for training and evaluation exhibit minimal orientation deviations [PAA*21]. In practice, alignment at capture time is easily achieved either by mounting the camera on a tripod placed on a level surface [DAH20] or, in handheld scenarios, by relying on built-in Inertial Measurement Units (IMUs). Even when this condition is not satisfied, many preprocessing techniques for image-based gravity rectification exist (e.g., [XLF*19, JLAB19, DAH20]), thereby facilitating the application of gravity-oriented methods.

4.3. Architectural and data-driven priors

A single 360° panorama captures the full appearance of a scene around a viewpoint, from which geometric, structural, and visual representations of the environment must be inferred. However, panoramas typically provide incomplete and noisy observations of the true underlying geometry, making reconstruction ill-posed. Textureless regions (e.g., bare walls), uneven angular sampling, and distortions in equirectangular images (Sec. 4.1) hinder feature detection and geometric reasoning. Transparent and reflective surfaces such as glass and mirrors introduce further complexity. Depth variation indoors is also more extreme than what is encountered in typical outdoor use cases, as images are taken in confined, narrow spaces, while depth ranges from nearby clutter to far-away walls. Most critically, furniture and objects often occlude large portions of the architectural layout, while concave room shapes cause severe self-occlusions, making much of the room boundary invisible to the viewer (Fig. 2). Consequently, depth estimation and structural recovery require not only leveraging the wide contextual cues that panoramas provide but also incorporating strong, domain-specific priors stemming from the construction of architectural spaces [PMG*20].

Beyond generic surface reconstruction assumptions such as smoothness, symmetry, or repetition [BTS*17], the indoor reconstruction literature has introduced explicit architectural priors. The

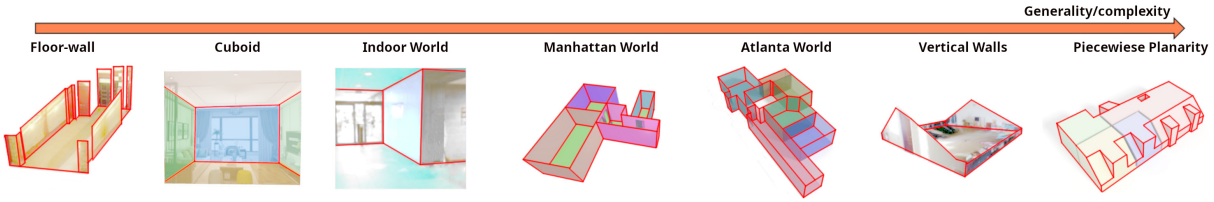

Figure 3: ***Common indoor architectural priors.*** *From left to right:* floor–wall boundaries *[DHN06],* cuboids *[HHF09], the* Indoor World Model *[LHK09], the* Manhattan world *[CY99], the* Atlanta world *[SD04], the* Vertical Walls *[PGP*18] and* piece-wise planarity *[FCSS09].*

Acronym	Name	Reference	Assumption
FW	Floor–Wall	[DHN06]	Single flat floor with unconnected vertical walls; ceiling ignored
CB	Cuboid	[HHF09]	Single cuboid-shaped room with six orthogonal faces
IW	Indoor World	[LHK09]	Single horizontal floor/ceiling, vertical walls meeting at 90°
MW	Manhattan World	[CY99]	Horizontal floors/ceilings, vertical walls meeting at 90°
AW	Atlanta World	[SD04]	Horizontal floors/ceilings, vertical walls not restricted to meeting at 90°
VW	Vertical Walls	[PGP*18]	Vertical walls, floors and ceilings may slope
PW	Piecewise Planarity	[FCSS09]	General polyhedral rooms with piecewise planar walls, floors, and ceilings

Table 1: ***Common geometric priors for indoor reconstruction.*** *See Fig. 3 for an illustration.*

most common ones are summarized in Tab. 1 and illustrated in Fig. 3. As we will see in the following sections, while early approaches relied on restrictive models (e.g., *floor–wall boundaries* [DHN06] or simple *cuboids* [HHF09]), modern solutions mostly adopt the Manhattan World prior [CY99], which aligns well with standard construction rules in residential and office buildings and, through orthogonality constraints, strongly reduces reconstruction ambiguity, offering ample spaces for method customization [ZSP*21]. Atlanta world solutions are also gaining strong traction, since they expand the applicability of the methods by relaxing right-angle restrictions, while still preserving many optimization possibilities [PAA*21].

Classical reconstruction exploited such priors via geometric reasoning, extracting image features (corners, edges, uniform regions) and interpreting them as cues under the chosen prior [PMG*20]. The success of these methods, however, is highly sensitive to the number and quality of extracted features, and these methods often struggle in real-world scenarios due to clutter, non-cooperative materials, and deviations from idealized structural assumptions.

As a result, research has shifted toward data-driven approaches [ZSP*19], particularly deep learning [ZSP*21], which can leverage large amounts of examples to learn implicit geometric and semantic cues directly from data. Deep models provide more robust solutions than purely algorithmic methods and allow relaxing many strict reconstruction assumptions. As we will see in the following sections, architectural priors, however, remain essential, often enabling lower-dimensional reasoning (e.g., extracting floor plans under vertical wall assumptions) or reducing network complexity (e.g., flattening features along the gravity axis).

4.4. Open research data

Publicly available datasets play a crucial role in advancing research on indoor reconstruction and exploration, as they not only provide standardized benchmarks for evaluating and comparing the results of different techniques but, with the current evolution towards learning methods, are also essential for providing learning material. Prior surveys on indoor reconstruction [PMG*20, PPL*24] have covered generic datasets with different contents and coming from different sources (e.g., CAD models, point clouds). In this work, we restrict our analysis to those used for training and testing methods tailored for 360° data. These datasets, in addition to including panoramic images as equirectangular images, may also include a variety of other information, such as registered depth maps, layouts, 3D models, and semantic annotations, thereby supporting supervised training and testing for different tasks. Importantly, they also reflect different capture modalities, ranging from real 360° camera acquisitions to synthetic renderings of virtual environments, with the latter providing the more precise ground-truth data. The most widely used datasets in this domain are listed in Tab. 2. We do not include in the table standalone smaller-scale datasets (<500 images) or data used for only a few specific publications (e.g., [PAG20, SAPS22]). Below, we summarize each dataset's key characteristics and intended usage.

Matterport3D [Mat17] This large-scale RGB-D dataset contains 10,800 panoramas from 90 properties, reconstructed from 194,400 RGB-D images captured with a Matterport Pro Camera. It provides raw data and annotations, including camera poses, textured meshes, floor plans, region labels, and object instances. Panoramas are sampled at human height, but spherical coverage is partial, with missing top and bottom regions often inpainted [CDF*17]. Since original panoramic poses relative to the mesh were absent, lower resolution color and depth map pairs are commonly derived from textured meshes [ZKZD18], a limitation later resolved in *360MonoDepth/M3D* [RAYR22b, RAYR22a]. The dataset supports both single-image tasks (e.g., surface normal estimation, semantic labeling) and sparse multiview tasks (e.g., keypoint matching, overlap prediction), and has spawned multiple derived datasets.

Dataset	Method	# Panos	Pano res.	Depth	Layout	Layout type	Other
Matterport3D [Mat17]	Captured, Stitched	10,800	2048×4096	(Yes)	—	(All)	S, P
MatterportLayout [ZSP*17]	Captured, Stitched	2,295	512×1024	—	Yes	Manhattan	(S, P)
Stanford2D-3D-S [SU17]	Captured, Stitched	552	2048×4096	Yes	Yes	Cuboid	N, S, P
Gibson 3D scan database [XRZH*18b]	Captured, Stiched	45,196	1024×2048	(Yes)	(Yes)	(Atlanta)	P
3DSceneGraph [AHG*19b]	Captured, Rendered	45,196	1024×2048	(Yes)	(Yes)	(Atlanta)	S, P
360MonoDepth/M3D [RAYR22b]	Captured, Stitched	9,684	1024×2048	Yes	—	(All)	P
360MonoDepth/Replica [RAYR22c]	3D Acquisition, Rendered	130	1024×2048, 2048×4096	Yes	—	(All)	P
PanoContext [ZSTX14b]	Captured, Native 360	700	512×1024	—	Yes	Cuboid	I
Zillow Indoor (ZInD) [CHL*21b]	Captured, Native 360	71,474	1024×2048	—	Yes	Atlanta	I, O, P
ZInd-Tell [DWB*24a]	Captured, Native 360	71,474	1024×2048	—	Yes	Atlanta	I, O, P, L
Structured3D [ZZL*20a]	3D Modeling, Rendered	21,000	512×1024	Yes	Yes	Atlanta	N, A, I, S, P
PNVS [XZX*21b]	3D Modeling, Rendered	34,811	512×1024	Yes	Yes	Atlanta	P
SKI360 [SUKc20]	3D Modeling, Rendered	3,550	512×1024	Yes	Yes	Atlanta	S
Pano3D [AZD*21a]	Mixed, Rendered	21,203	256×512, 512×1024	Yes	—	(All)	P, N
FutureHouse [LWH*22a]	3D Modeling, Rendered	28,579	512×1024	Yes	(Yes)	(All)	N,M
PanoSCU [KAKS25]	3D Modeling, Rendered	5,300	224×928	—	—	(All)	S,I,C

Table 2: ***Overview of most common large-scale publicly available 3D indoor datasets using panoramic images.*** *We list only datasets with substantial scale and impact for panoramic indoor scene analysis and synthesis. For each dataset, we report data type, source (real vs. synthetic), panorama generation method, number of panoramic images and resolution in equirectangular format, supported layout types, and available additional information (I = object instances, N = normals, S = semantic segmentation, O = openings/doors/windows position, P = 3D poses, A = albedo, M = material and light; L = natural language description, C = change tracking.*

MatterportLayout [ZSP*17, ZSP*21] This dataset extends the *Matterport3D* dataset with general Manhattan layout annotations for training and evaluation. It comprises 2,295 panoramas selected from the original *Matterport3D* dataset and split into training, validation, and testing sets to facilitate direct comparisons of learning models. All the annotated layouts are aligned to a camera height of 1.6m.

Stanford2D-3D-S [SU17, ASZ*16] This dataset provides large-scale indoor scans of public spaces such as offices, educational areas, lobbies, and hallways, captured with the same system as *Matterport3D*. It includes 552 equirectangular images with depths, surface normals, semantic annotations, and camera poses, as well as cuboid layouts and annotated registered meshes and point clouds.

GibsonV2 3D scan database [XRZH*18b, XRZH*18a] This dataset is similar in nature to *Matterport3D* and *Stanford2D-3D-S* but contains a greater variety of interiors (572 models, 1440 floors across households, offices, hotels, museums, hospitals, etc.). It provides reconstructed 3D meshes registered with panoramas. Several splits are provided, the largest containing 45,196 panoramas. Depth maps are not included but can be generated by ray-casting the 3D models. Originally intended for virtual simulation (e.g., robotics), it is also widely used for reconstruction and analysis tasks.

3DSceneGraph [AHG*19b, AHG*19a]. This dataset augments the *GibsonV2* scene collection with additional semantic information in the form of a 3D scene graph composed of four layers: building, room, object, and camera, further extending *GibsonV2* towards semantic analysis tasks.

360MonoDepth/M3D [RAYR22b, RAYR22a] This *Matterport3D* extension supports 360° monocular depth estimation with 9,684 panoramas (1024×2048) from 90 buildings, subdivided into training, test, and validation sets. 360° structure-from-motion [MMPM16] was employed to align captures with scene meshes, yielding pixel-accurate ground-truth depths. Camera poses that co-register images taken within the same building are also provided, enabling tasks such as navigation, view synthesis, and VR.

360MonoDepth/Replica [RAYR22c, RAYR22a] This synthetic companion to *360MonoDepth/M3D* adds 130 photorealistic equirectangular renders from 13 *Replica* [SWM*19b, SWM*19a] rooms, with co-registered ground-truth depths at 1024×2048 and 2048×4096. Images and co-registered ground truth depth were produced by the *Replica360* renderer [ALG*20] The dataset's main purpose is to test generalization and evaluate depth estimation at the currently highest open-data resolutions.

PanoContext [ZSTX14b, ZSTX14a] The dataset contains 700 full-view panoramas for home environments from the *SUN360* database [XEOT12] (no longer available for licensing reasons), including 418 bedrooms and 282 living rooms. Cuboid layouts are provided, as well as cuboid representations of interior objects.

Zillow Indoor (ZInD) [CHL*21b, CHL*21a] The dataset includes 67,448 panoramas from 1,575 unfurnished residential homes, with primary panoramas annotated for room layouts, openings, and labels, and secondary panoramas providing denser spatial coverage via semi-automatic localization. Primary views, chosen for full-room visibility and adjacent room co-visibility enable both single-image analysis and exploration in sparsely sampled environments.

ZinD-Tell [DWB*24a, DWB*24b] This extension of the ZinD database augments the original data with floor plans and natural language descriptions of the indoor spaces. The objective of this multimodal representation is to establish semantic connections among natural language descriptions, panoramic images, rooms, and entire floors. Generative tasks are the main target.

Structured3D [ZZL*20a, ZZL*20b] This large-scale synthetic dataset contains 3.5K professionally designed 3D house models in the Atlanta Layout style. It provides over 21,000 panoramas with registered depth, normals, semantic segmentations, door/window annotations, and co-registered poses, for both empty and furnished rooms. The rich annotation sets make it a common choice for training and evaluating 360° solutions on a wide variety of tasks.

PNVS [XZX*21b, XZX*21a] This photo-realistic dataset, built on *Structured3D*, targets view synthesis with panoramas and displaced views rendered from translated cameras. It includes two subsets: small displacements (0.2–0.3 m, head movements) and large displacements (1–2 m, room-scale motion), with 13,080/1,791 and 17,661/2,279 train/validation images, respectively.

Shanghaitech-Kujiale Indoor 360° (SKI360) [SUKc20, JXZ*20] This *Structured3D*-based dataset includes 1,775 rooms captured both furnished and unfurnished. The 3,550 panoramas with color and depth are associated with architectural layouts represented by 3D corners, planes, and plane-plane intersection lines. Supported tasks include depth estimation, separation of furniture from architectural structure, empty-room synthesis, and layout recovery.

Pano3D [AZD*21a, AZD*21b] This synthetic dataset targets depth and normal estimation and cross-dataset evaluation. It re-renders 21,203 scenes from *Matterport3D* point clouds [Mat17] and *GibsonV2* meshes [XRZH*18b, XRZH*18a] into 360° panoramas in equirectangular format containing color, depth, and normal maps. Splits are designed to test robustness under distribution shifts: *Medium*, *Tiny*, and *Full* vary depth distribution and image count within residential scenes, *Fullplus* adds non-residential contexts, and *Filmic* variants of the splits simulate covariate color shifts.

FutureHouse [LWH*22a, LWH*22b] This synthetic dataset is designed for inverse rendering applications. It contains 28,579 equirectangular views from 1,752 house-scale scenes rendered with *Unreal Engine 4*. Per-pixel material annotations are provided, along with light sources and per-pixel HDR environment maps.

PanoSCU [KAKS25, KQC*25] This synthetic dataset supports change detection and understanding in single-shot panoramas. It contains 5,300 panoramas, covering only 360°×90° to match robotic acquisition with rotating standard FOV cameras, 132,222 panoramic object masks, and 8,275 panoramic captions, created by rendering 85 AI2-THOR [KMH*17] scenes. Each scenario randomly changes the position or state of 1–5 objects, categorized in 48 object types, with ground-truth change annotations, resulting in 4,150 rearrangement scenarios. It enables object detection, tracking, segmentation, and change understanding from a single panorama. The dataset is intended to support single-panorama object detection and tracking, segmentation, change understanding, and change reversal.

Several important observations arise from the analysis of the current landscape of available datasets.

Orientation All collections provide (near) gravity-aligned panoramas or transformations that can produce near gravity-alignment. When gravity alignment is not perfect, previous research has shown that deviations are very small (e.g., sub-degree deviations for *Stanford2D-3D-S*, *Matterport3D*, and derived datasets [PAA*21]). Methods can, therefore, assume gravity alignment for training and testing data, or must introduce explicit augmentations when handling non-aligned data is required.

Resolution Most collections are provided at relatively low resolutions (e.g., 512×1024), and when higher resolutions are available, they often lack comprehensive annotations. As a result, dataset choice must be tailored to the intended task. For instance, while 512×1024 panoramas captured from the center of typical rooms yield a few centimeters of spatial sampling that may suffice for floorplan or layout extraction, they may fall short for high-quality or immersive rendering [RAYR22a].

Size and variety Despite the existence of hundreds of millions of casually captured panoramas in domains such as real estate [SSO*21], curated research datasets remain much smaller in scale and typically less diverse in terms of capture conditions and environments. This limitation stems not only from the challenges of aggregating user-generated data but also from the cost and complexity of creating annotation [CHL*21a]. Synthetic datasets, either derived from re-rendering reconstructed models from dense captures [AZD*21b, RAYR22a] or from fully modeled 3D environments [ZZL*20b, JXZ*20, XZX*21a, AZD*21b], are emerging as a promising way to expand diversity and provide richer ground truth. However, even in these cases, the variety of environments, room types, and architectural styles is ultimately bounded by the available 3D sources. In this context, combining heterogeneous datasets for training and evaluation, though still uncommon (see Sec. 5 and Sec. 6), is an important area for research, particularly for testing robustness and transfer learning across domains [AZD*21b, RAYR22a]. Complementary strategies such as semi- and self-supervised learning may further enhance downstream performance by leveraging vast amounts of unlabeled data in addition to the limited annotated collections [ZLW*23, LZW*25a].

5. Automatic geometric and structural modeling

The methods surveyed here all take as sole input a single 360° image of an indoor scene or, for multi-room environments, one image per room. From this limited RGB data, the task is to infer geometric, structural, and visual properties that are not directly observable. The first group of approaches focuses on pixel-level augmentation, enriching the panorama with co-registered maps such as depth, normals, or albedo. Depth estimation, in particular, is a cornerstone task, as it provides geometric cues that underpin most subsequent reconstruction and synthesis methods (Sec. 5.1). A second group emphasizes structural abstraction, separating the architectural layout (walls, ceiling, floor) from contained objects. While some methods still annotate panoramas or floor-plan projections, many construct higher-level structured representations that also allow plausible reconstruction of occluded areas (Sec. 5.2). Finally, we examine methods for multi-room understanding under very sparse sampling, where only one panorama per room is available and traditional multi-view pipelines fail. Here, the focus lies on recovering room layouts and inter-room connections, which are crucial for navigation and exploration (Sec. 5.3).

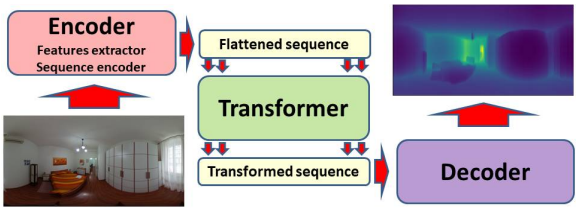

Figure 4: ***Schematic illustration of a typical encoder–transformer–decoder architecture for panoramic depth estimation.*** *The encoder extracts features directly from the equirectangular image or from patches of it. The final encoding layers shape the features as a feature sequence, ready to be processed by the transformer. The decoder recovers the predicted depth in equirectangular format.*

5.1. Depth and other pixel-wise information inference

Pixel-wise reconstruction refers to the family of tasks aimed at recovering dense per-pixel information from a single panoramic image. These representations include depth, surface normals, semantic labels, and other descriptors that collectively provide the low-level building blocks for higher-level reconstructions such as room layouts or semantic parsing. Their relevance in the panoramic setting is amplified by the nature of 360° imagery: while a single capture encodes the full visual field, perspective cues are heavily distorted or flattened, making the recovery of per-pixel geometric signals both indispensable and particularly challenging.

Within this task family, monocular depth estimation has naturally emerged as the most central and widely studied problem. Depth is not only fundamental to recovering the underlying geometry of a scene, but also closely connected to the architectural structure of the scene (Sec. 5.2) and strictly required in downstream applications such as novel view synthesis (Sec. 6.1). Depth has thus become the pivot around which most panoramic reconstruction pipelines are built. Early approaches relied on handcrafted features, geometric constraints, and multi-scale reasoning to infer depth, but these were generally brittle in real-world conditions. With the advent of deep learning, large-scale training datasets have enabled models to extract non-trivial depth cues directly from image appearance, leading to a significant improvement in robustness and generalization [PAA*21].

It is worth noting, however, that the same architectural backbones used for depth prediction can be extended to other pixel-wise signals. For example, Eigen et al. [EF15] demonstrated that depth, surface normals, and semantic segmentation can all be jointly learned from a single image. These complementary outputs enrich both the geometric and semantic understanding of the environment: normals refine the perception of local structures, while semantics provide object-level meaning. Depending on the application, these signals often act as valuable complements to depth estimation [LPLJ25]. Their role becomes especially relevant in scenarios that extend beyond pure reconstruction, such as advanced scene editing or immersive manipulation. Concrete examples, including tasks such as emptying a complete scene, selective object removal, virtual staging, or style transfer, will be discussed in Sec. 6.2.

Classic data-driven monocular depth estimation Data-driven monocular depth estimation was introduced over a decade ago for limited field-of-view imagery, with pioneering works such as Make3D [SSN09]. The emergence of deep learning and the availability of large-scale 3D datasets soon led to dramatic performance improvements. After the introduction of CNNs for regressing dense depth maps from a single image [EPF14, EF15], Laina et al. [LRB*16] proposed the standard *FCRN* encoder-decoder architecture, combining *ResNet* [HZRS16] for encoding, an up-projection module for decoding, and the reverse Huber loss [LLZ16] to improve depth prediction accuracy. Following these trends, many further solutions were introduced, including predicting depth from multiple perspective crops combined in the Fourier domain [LHKK18], using ordinal regression losses to preserve spatial ordering among neighboring depth classes [FGW*18], or exploiting Conditional Random Fields (CRFs) for refined predictions [LCG15, PXZ*15, CWS18, XWT*18], among many others. While these methods were originally designed for perspective imagery, their success inspired early attempts to extend them to the panoramic domain. A common strategy was to project 360° images into multiple perspective views, process them with pre-trained perspective networks, and then merge the results back into the equirectangular domain. Although effective as a first step, such direct adaptations did not fully exploit the unique properties of panoramic imagery, in particular, the global field of view and the associated geometric distortions, which often led to sub-optimal predictions [ZSTX14a, ZKZD18]. This limitation led to the emergence of 360°-specific approaches that explicitly model spherical geometry and global context, and now dominate research on indoor depth estimation from panoramic imagery.

Distortion-aware approaches Several solutions adapted perspective methods to 360° depth prediction by projecting panoramic images into cube maps [CCD*18] or by replacing standard convolutions with spherical/distortion-aware convolutions to cope with equirectangular distortions [SG17, TNT18, PdLGAAB18, ZKZD18, SG19a] (i.e., *distortion-aware methods* – Tab. 3). Hybrid approaches have also been proposed [WYS*20, JSZ*21, HSC*25]: for instance, Wang et al. [WYS*20] combined the two representations through a dual-branch network operating respectively on the equirectangular projection and on the cube map, coupling a distortion-aware encoder [ZKZD18] with the FCRN decoder [LRB*16].

Structure-aware approaches An important line of research targets the direct processing of equirectangular panoramas with encoders that explicitly exploit the structural regularities of indoor environments (i.e., *structured-aware methods* in Tab. 3). Representative methods include [SSC21, PAA*21, PASG24, PASG25], which leverage gravity-aligned features and geometry-aware constraints to reduce model complexity while still capturing both short- and long-range relations. An example is the *SliceNet* architecture introduced by Pintore et al. [PAA*21], which builds on the observation that man-made indoor structures are predominantly aligned with gravity, i.e., along the vertical direction. By applying a vertical compression of the equirectangular input and employing a recurrent neural network (RNN) to sequentially capture horizontal dependencies, SliceNet effectively incorporates gravity alignment into the network design. Furthermore, the feature sequence encoding

Method	Representation	Backbone / Architecture	Key Contribution
Supervised distortion-aware methods			
Su et al. [SG17] (2017)	Equirectangular	Learned spherical conv.	Distortion-aware convolution kernels
Cheng et al. [CCD*18] (2018)	Cube map	CNN encoder-decoder	Cube projection to reduce distortion
Tateno et al. [TNT18] (2018)	Equirectangular	Multi-purpose CNN	Spherical-aware convolution kernels
Zioulis et al. [ZKZD18] (2018)	Equirectangular	Distortion-aware FCRN	Geometry-aware depth estimation
Su et al. [SG19a] (2019)	Equirectangular	Kernel transformer net	Adaptive kernel learning for distortions
Wang et al. [WYS*20] (2020)	Equirect.+Cubemap	Two-branch CNN (FCRN)	Dual-branch learning (BiFuse)
Supervised structure-aware methods			
Pintore et al. [PAA*21] (2021)	Equirectangular	Gravity-aware encoder-decoder	Indoor structural priors (SliceNet)
Shen et al. [SLL*22] (2022)	Equirectangular	Panorama Transformer (PanoFormer)	ERP-transformer with spherical tangent patches
Li et al. [LGY*22] (2022)	Perspective patches	Transformer fusion	Sampled patches fused into panorama
Rey et al. [RAYR22a] (2022)	Perspective patches	CNN + fusion	Multi-resolution cubemap sampling
Shen et al. [SZL*23] (2023)	Perspective patches	Patch-based ViT	Depth fusion through attention mechanisms
Ai et al. [AW24] (2024)	Equirectangular	Transformer + bi-projection fusion	ERP-based with spherical point projection
Yan et al. [YWZ*25] (2025)	Equirect.+Spherical	ResNet + mesh encoder	SphereFusion with gated feature fusion
Weakly-/Self-supervised methods			
Wang et al. [WHC*18] (2018)	Equirectangular	Distortion-aware CNN	First self-supervised framework (360-SelfNet)
Zioulis et al. [ZKZ*19] (2019)	Equirectangular	CoordConv + view synthesis	Trinocular view synthesis
Wang et al. [WYT*22] (2022)	Equirect.+Cubemap	BiFuse++ dual-branch CNN	Self-supervised with bi-projection fusion
Wang et al. [WL24b] (2024)	Equirectangular	Distillation framework	Knowledge transfer from perspective models
Wang et al. [WHZ*25] (2025)	Equirect.+Cubemap	Dual-domain collaborative CNN	Asymmetric dual-domain collaboration
Cao et al. [CZZ*25] (2025)	Equirectangular	Weakly-supervised CNN	Scene-structural knowledge transfer

Table 3: ***Representative methods for panoramic monocular depth estimation****. Approaches are grouped by input representation and architectural design.*

and the recurrent component can be regarded as precursors to later transformer-based architectures that exploit long-range dependencies through attention mechanisms. Overall, such structure-aware methods demonstrate the benefits of embedding domain-specific priors inherent to indoor geometry.

Transformer-based approaches The advent of transformer-based architectures has marked a turning point in panoramic depth estimation, introducing the ability to model long-range dependencies and global context more effectively than traditional convolutional networks. A first example is *PanoFormer* [SLL*22], which adapts the transformer paradigm to panoramic imagery by introducing spherical tangent patches, learnable token flows, and panorama-specific metrics. More recent solutions increasingly leverage perspective sampling as a way to exploit the availability of powerful pretrained models on conventional imagery. Panoramic inputs are decomposed into a set of perspective views [LGY*22, RAYR22a], which are processed independently and then fused through transformer-based architectures. In this direction, patch-wise vision transformers have been employed to aggregate local predictions into globally consistent depth maps [SZL*23, ACC*23, AW24], showing the potential of attention mechanisms to reason across spatially disjoint observations. In parallel, hybrid strategies have been explored to overcome the limitations of individual projections. SphereFusion [YWZ*25] is an example: by combining equirectangular and spherical representations through a gated fusion module, it can balance geometric faithfulness and texture detail, while also achieving state-of-the-art inference efficiency. This demonstrates the growing interest in projection-aware fusion schemes that can exploit the complementary strengths of different panoramic representations.

Self-supervised approaches The vast majority of existing methods for panoramic depth estimation are based on supervised learning. This creates a significant limitation, since collecting large-scale ground-truth depth data for 360° imagery is particularly challenging, especially at high resolution. To address this, several attempts have been made to develop self-supervised or, at least, weakly supervised methods. While such approaches still lag behind fully supervised techniques in terms of accuracy, they represent a promising direction and are rapidly evolving (Tab. 3). Based on the cube padding strategy [CCD*18], Wang et al. [WHC*18] propose *360-SelfNet*, the first framework for self-supervised 360° depth estimation. Zioulis et al. [ZKZ*19] adopt trinocular view synthesis to provide self-supervised depth inference. Building on this line, Wang et al. [WYT*22] introduced *BiFuse++*, extending the original BiFuse dual-branch design [WYS*20] to self-supervised training with spherical padding and bi-projection fusion, showing that hybrid representations can also be effective even without ground-truth supervision. Later, Wang et al. [WL24b] introduced *Depth Anywhere*, a knowledge-distillation framework where a strong perspective depth estimator supervises the training of a panoramic model, enabling improved generalization across both domains. Complementary to this, Cao et al. [CZZ*25] tackled the challenge of high-resolution panoramic depth estimation in the absence of high-res annotations. Their weakly-supervised framework introduces a Scene Structural Knowledge Transfer (SSKT) module that injects structural cues at training time, thus allowing networks to recover fine-grained details without requiring dense ground-truth labels.

Overall architecture evolution The progression of panoramic depth estimation methods reflects a clear evolution in research focus (Tab. 3). Initial works mainly adapted solutions developed for conventional perspective imagery, concentrating on mitigating spher-

ical distortions through cube map projections or distortion-aware convolutions, largely independent of scene contents. A subsequent trend has shifted towards explicitly incorporating the structural characteristics of indoor environments, with gravity alignment, layout consistency, and domain-specific priors emerging as central design choices. Overall, a typical architecture for panoramic monocular depth estimation follows an encoder–transformer–decoder scheme, as illustrated in Fig. 4. First, an encoder extracts multi-scale features from the equirectangular input and reshapes them into a sequential representation suitable for transformer processing, mostly exploiting slicing [SSC21, PAA*21] or patching [SLL*22, LGY*22, ACC*23]. The transformer module then models long-range dependencies and global scene context across the entire panorama. Finally, a decoder maps the transformed features back to the equirectangular domain, producing dense per-pixel depth predictions. While supervised approaches still deliver the highest accuracy, there is a growing research effort in self- and weakly-supervised learning. These approaches, though less mature, are particularly promising in light of recent advances in foundation models, which may provide new pathways to reduce reliance on large-scale annotated data (Sec. 7).

Common metrics The evaluation of panoramic depth estimation still relies predominantly on standard metrics inherited from perspective-based depth regression since the introduction of FCRN, such as *AbsRel*, *RMSE*, and δ *thresholds* [EPF14, LRB*16]. Specifically, the most common metrics include:

- **Absolute Relative Error (AbsRel)**: the average of the ratio between the absolute depth error and the ground-truth depth. It emphasizes proportional errors across depth ranges.
- **Root Mean Squared Error (RMSE)**: the square root of the mean squared differences between predicted and ground-truth depth values, highlighting larger errors.
- **Mean Relative Error (MRE)**: similar to AbsRel but normalized by predicted values, providing a complementary perspective on relative error magnitudes.
- δ **thresholds**: the percentage of predicted depth values d_p that satisfy $\max\left(\frac{d_p}{d_{gt}}, \frac{d_{gt}}{d_p}\right) < \delta$, with typical thresholds $\delta = 1.25, 1.25^2, 1.25^3$. These metrics quantify prediction accuracy within multiplicative error bounds.

It is important to note that while these metrics have become standard in monocular depth evaluation, their behavior may not be equally suitable across different domains. In indoor panoramic environments, where depth ranges are relatively short and errors at small distances can dominate the evaluation, *AbsRel* tends to overweight inaccuracies in near-field geometry. Conversely, *MRE* shifts the normalization to predicted values, which can mitigate this effect but may underestimate large relative errors at greater depths. This suggests that, despite their widespread adoption, a more tailored set of metrics could be beneficial for faithfully capturing reconstruction quality in panoramic indoor scenarios.

Training objectives The effectiveness of panoramic depth estimation methods strongly depends on the choice of loss functions employed during training. Early deep learning models for monocular depth estimation typically relied on the *reverse Huber loss* (also known as the *BerHu loss*) [LRB*16], which balances sensitivity to small residuals with robustness to outliers. This formulation rapidly became a standard in indoor depth estimation, and remains at the core of several panoramic baselines such as *OmniDepth* [ZKZD18] and *360SD-Net* [WYS*20]. Subsequent works have progressively enriched the objective functions with multi-scale supervision and structural regularizers. For instance, *BiFuse* [WYS*20] and *UniFuse* [JSZ*21] combine BerHu loss with smoothness penalties that encourage spatial consistency across neighboring pixels, while *SliceNet* [PAA*21] further incorporates geometry-aware constraints aligned with indoor structural priors. With the advent of transformer-based architectures, methods such as *PanoFormer* [SLL*22] and *Elite360D* [AW24] have emphasized hybrid losses that combine point-wise regression with edge-aware or semantic-aware regularization, improving the fidelity of structural boundaries. More recently, *SphereFusion* [YWZ*25] extends this trend by adopting distance-aware objectives that adapt the penalty according to depth ranges, leading to improved accuracy in large-scale indoor scenes. Overall, the evolution of training objectives mirrors the architectural trends: from simple pixel-wise regression to more sophisticated, context- and geometry-aware formulations, designed to reduce artifacts and enhance structural consistency in panoramic depth maps. On the other hand, in the self- and weakly-supervised setting, the design of losses is even more critical, as ground-truth depth is not available. MIDAS [RLH*22] propose to perform prediction in disparity space (inverse depth up to scale and shift) through a family of scale- and shift-invariant dense losses to handle ambiguities stemming from the merging of datasets acquired in the wild. Originally developed for perspective cameras, the method has also been employed for 360° depth estimation [YKH*24b, JSL*25]. A common alternative strategy is the use of photometric reconstruction losses that compare synthesized views reconstructed from the predicted depth to the input panorama [WHC*18]. Zioulis et al. [ZKZ*19] extend this principle with a trinocular view synthesis loss, leveraging multiple synthesized perspectives to provide stronger self-supervision. Wang et al. [WYT*22] introduce *BiFuse++*, combining spherical padding and bi-projection fusion with reprojection-based losses, while more recent frameworks further refine this design. For example, Wang et al. [WHZ*25] propose an asymmetric dual-domain collaboration scheme, enforcing consistency across equirectangular and cubemap predictions, with geometry-aware self-supervised losses. Finally, weakly-supervised methods such as the one proposed by Cao et al. [CZZ*25] introduce structural knowledge-transfer modules, where geometric cues (e.g., room layout consistency) act as priors to constrain depth predictions without requiring dense annotations. Nonetheless, supervised methods remain dominant in terms of absolute accuracy, but the sophistication of loss functions in self-supervised learning is rapidly narrowing the gap.

Representative methods performances The comparative results in Tab. 4 highlight several consistent trends in 360° monocular depth estimation. The reported performance values are taken from SphereFusion [YWZ*25] experiments, while computational stats were extracted from the original papers. In terms of evaluation, nearly all methods are benchmarked with the same set of standard metrics inherited from perspective-based monocular depth estimation [EPF14, LRB*16], namely *AbsRel*, *RMSE*, and the δ thresholds. These metrics allow for comparability across different approaches, but they remain agnostic to panoramic-specific

Method	Dataset	AbsRel ↓	RMSE ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑	GFLOPs ↓	Params (M) ↓
OmniDepth [ZKZD18]	S2D3D	0.200	0.615	0.688	0.889	0.958	—	—
BiFuse [WYS*20]	S2D3D	0.121	0.414	0.866	0.958	0.986	682	253
UniFuse [JSZ*21]	S2D3D	0.111	0.369	0.871	0.966	0.988	278	52
SliceNet [PAA*21]	S2D3D	0.100	0.373	0.904	0.962	0.984	101	79
PanoFormer [SLL*22]	S2D3D	—	0.308	0.939	0.984	0.994	78	20
OmniFusion [LGY*22]	S2D3D	0.095	0.347	0.899	0.977	0.992	—	42
SphereFusion [YWZ*25]	S2D3D	**0.090**	**0.319**	**0.926**	**0.976**	**0.990**	36	25
OmniDepth [ZKZD18]	M3D	0.290	0.764	0.683	0.879	0.943	—	—
BiFuse [WYS*20]	M3D	0.205	0.626	0.845	0.932	0.963	682	253
UniFuse [JSZ*21]	M3D	0.106	0.494	0.890	0.962	0.983	278	52
SliceNet [PAA*21]	M3D	0.176	0.613	0.872	0.948	0.972	101	79
PanoFormer [SLL*22]	M3D	—	0.364	0.918	0.980	0.992	78	20
OmniFusion [LGY*22]	M3D	0.090	0.426	0.919	0.980	0.993	—	42
SphereFusion [YWZ*25]	M3D	**0.115**	**0.489**	**0.870**	**0.961**	**0.984**	36	25

Table 4: ***Accuracy and performance comparison for*** 360° ***monocular depth estimation.*** *Results are reported on Stanford-2D-3D (S2D3D) and Matterport3D (M3D). values provided by SphereFusion [YWZ*25] official comparisons. Metrics: lower is better for AbsRel, RMSE, GFLOPs; higher is better for* δ *thresholds. Computational costs (GFLOPs, Params) are taken from the original papers when available.*

distortions and structural regularities, which suggests that more tailored evaluation protocols may be beneficial. From a performance perspective, early *distortion-aware* methods such as OmniDepth [ZKZD18] and BiFuse [WYS*20] achieve reasonable accuracy but at the cost of higher residual errors, especially in challenging indoor scenarios. The introduction of *structure-aware* methods, such as SliceNet [PAA*21], brought noticeable improvements by embedding indoor geometry priors (e.g., gravity alignment), leading to better consistency and reduced computational overhead. The most recent generation of *transformer-based* approaches, such as PanoFormer [SLL*22] and SphereFusion [YWZ*25], further advances the state of the art by modeling long-range dependencies and combining multiple projection domains. These architectures improve accuracy (with SphereFusion achieving the best balance across AbsRel, RMSE, and δ scores) while reducing computational cost, as reflected in their GFLOPs and parameter counts. Overall, this trend marks a shift from purely distortion-correcting networks to scalable structure- and context-aware architectures.

5.2. Single-image room layout estimation

Depth estimation is fundamental for panoramic scene modeling, providing the dense metric basis for reconstructing indoor environments. When the viewpoint is fixed, a single panoramic depth map can support applications such as scene emptying, refurnishing, or style transfer (Sec. 6). More generally, however, it is necessary to separate permanent architectural structures (the *layout*) from movable objects or clutter [PMG*20]. Layout reconstruction thus extracts a higher-level scene representation, typically consisting of floors, ceilings, walls, and connecting doors.

Pixel-wise layout mapping In cases of small or null viewpoint motion, the separation between indoor layout and objects can be effectively performed at the pixel level, and several works in the literature adopt depth-guided segmentation masks that partition the equirectangular view into cluttered and uncluttered regions [JXZ*20, XZX*21a, GSZ*21, PAAG22, PJAG25], typically predicted using relatively simple encoder–decoder networks such as U-Net. A representative example is the work of Jin et al. [JXZ*20], where layout masks are learned jointly with depth estimation. Their training strategy leverages specialized datasets that provide paired reconstructions of the same rooms in both furnished and empty configurations (Sec. 4.4), thus enabling the network to disentangle permanent structures from movable objects (Sec. 4.4).

Indoor-structured layout In more general contexts, pixel-wise masks alone are insufficient. Full layout estimation is significantly more complex than assigning one value per pixel, as for depth estimation, because it requires extrapolating occluded or invisible surfaces and reasoning about global structure beyond the directly observable data. The difficulty is accentuated in indoor panoramic capture, where occlusions caused by furniture and self-occluding concave geometries mean that large portions of the true structure are hidden from direct observation (Fig. 2). Thus, layout reconstruction must plausibly hallucinate missing geometry, introducing a strong dependence on prior knowledge and structural constraints. To this end, a *structured scene reconstruction* is required, where the separation between permanent architectural structures (layout) and movable objects is performed in a more robust and semantically informed manner [IYF15, PMG*20]. In the specific case of panoramic indoor scenes, this process leverages the global context provided by the single panoramic observation [ZSTX14a], following pipelines such as the one illustrated in Fig. 5. Within this framework, existing panoramic works [ZSTX14a, ZCC*21, DFB*24, LPLJ25] decompose the problem into two core tasks: estimating depth and predicting the layout from the input panorama. The layout serves as the structural scaffold of the environment, enabling the placement of object instances, including openings, while depth estimation provides point clouds for modeling individual objects. Once this pipeline is established, the objects themselves can be reconstructed using standard techniques largely independent of the panoramic setting. For this reason, most research in the panoramic domain has focused on depth estimation (as discussed in the previous section) and layout estimation.

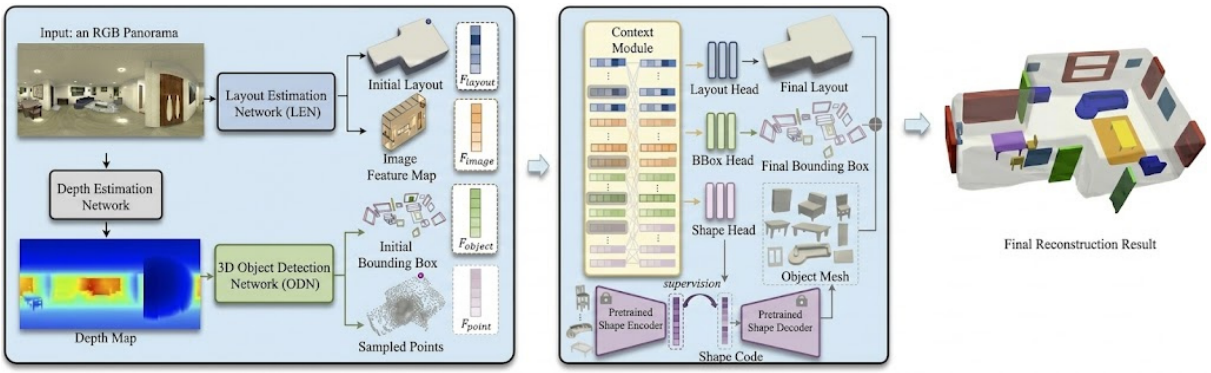

Figure 5: ***Illustration of a panoramic scene understanding pipeline*** *[DFB*24]. The diagram highlights how depth estimation and layout parsing from a single panoramic image represent the two fundamental tasks for reconstructing a complete scene. The layout provides the structural backbone of the environment, enabling the placement of object instances and openings, while depth maps supply localized point clouds that allow for the modeling of individual objects. Once this pipeline is established, each object can be reconstructed using standard modeling techniques, largely independent of the panoramic indoor context.*

Method	Input	Pre-proc.	Backbone	Post-proc.	Output
Zhang14 [ZSTX14a]	ERP	OM + GC	Geom. reasoning	Yes	Cuboid layout
LayoutNet [ZCSH18]	ERP	MW+lines	CNN (corner/bound.)	Yes	2D prob. maps
HorizonNet [SHSC19]	ERP	MW align	CNN+RNN (1D scanline)	Yes	1D vectors
DuLa-Net [YWP*19]	ERP+ceil	MW align	Dual CNN (ERP+top)	Yes	Floor-plan map
AtlantaNet [PAG20]	Ceil+floor	None	CNN proj. planes	Light	3D layout (AWM)
Deep3DLayout [PAAG21]	ERP	None	Graph NN	None	Watertight mesh
SSLayout360 [Tra21]	ERP	MW align	Semi-sup. CNN	Yes	Corners/bound.
LGT-Net [JXXZ22]	ERP	MW align	SWG-Transformer	Light	1D seq.
DMH-Net [ZWXG22]	Cubemap	MW align	CNN+Hough space	Yes	Line-based 3D
DOP-Net [SZL*23]	ERP	MW align	Plane-disent. CNN	Yes	Plane-wise feat.
PanelNet [YHJ*23]	Panels	None	Panel-wise CNN	Yes	Cont. panels
Bi-Layout [TJZ*24]	ERP	MW align	Dual embed.+guidance	Yes	2 layouts
Seg2Reg [STS*24]	ERP	None	CNN+occl.-aware render	Min.	1D depth
HUSH [LPLJ25]	ERP	None	SH-Transformer	None	Multi-task (D/N/L)

Table 5: ***Comparison of representative panoramic layout estimation methods.*** *Acronyms: ERP = Equirectangular Projection, Persp. = Perspective, Ceil = Ceiling plane projection, Floor = Floor plane projection, MW = Manhattan World, IWM = Indoor World Model, AWM = Atlanta World Model*

Geometric reasoning approaches Early approaches to indoor layout estimation exploited the strong regularities of man-made environments, relying on geometric reasoning to align image features with simplified constrained 3D models. Hedau et al. [HHF09] pioneered the use of a *cuboid prior* for pixel labeling, while Lee et al. [LHK09] exploited the Manhattan-World Model (MWM) to infer 3D structures from detected corners. Moving to panoramic imagery, Zhang et al. [ZSTX14a] were among the first to leverage 360° captures to overcome the contextual limitations of narrow field-of-view images. Their method mapped an entire panorama to a cuboid room model by combining *Orientation Maps* (OM) [LHK09] for the upper hemisphere and a *geometric context* (GC) analysis [HEH07] for the lower part. Xu et al. [XSKT17] later extended this reasoning to the IWM, while Yang et al. [YZ16] generalized to MWM layouts using partially oriented super-pixel facets and line segments. Several follow-ups [PMG*20] adopted similar strategies, though the effectiveness of these purely geometric methods strongly depends on the quantity and quality of extracted features (e.g., corners, edges, planar patches). This limitation has motivated a progressive shift towards data-driven approaches [ZSP*19].

Hybrid data-driven approaches As highlighted by Zou et al. [ZSP*21], most data-driven layout pipelines still follow a similar structure: (i) a pre-processing stage, often involving MWM alignment (e.g., based on [ZSTX14a]); (ii) prediction of layout elements in image space; and (iii) post-processing to regularize a 3D model from the predicted 2D layout elements. Representative examples include *LayoutNet* [ZCSH18], which directly predicts corner and boundary probability maps from panoramas, and *HorizonNet* [SHSC19], which encodes the layout as three 1D vectors representing, column-wise, the floor-wall, ceiling-wall, and wall-wall boundaries. The final layout is then recovered by fitting MWM segments to the predicted corners. *DuLaNet* [YWP*19] combines features from the equirectangular panorama and a ceiling-plane projection to produce a floor-plan probability map, which is regularized under MWM assumptions. Several recent works have ex-

tended HorizonNet, including *Led^{2}Net* [WYS*21], which leverages rendered horizon depth maps to improve predictions, achieving state-of-the-art performance under IWM constraints. Other methods jointly exploit depth, layout, and semantics to refine their predictions. For instance, Zeng et al. [ZKG20] estimated a layout depth map by integrating semantic segmentation and full-scene depth cues to regularize the IWM layout.

Geometry-aware data-driven approaches Despite their effectiveness, many of these methods rely on heavy pre-processing, such as vanishing line detection for Manhattan-world alignment [ZSP*19, ZSTX14a, LHK09], or on complex post-processing steps to enforce structural regularity [ZCSH18, SHSC19, YWP*19]. *AtlantaNet* [PAG20] addressed these limitations by relaxing the strict MWM/IWM assumptions: given panoramas roughly aligned with gravity, it estimates layouts under the more flexible Atlanta World Model (AWM), by projecting the spherical image onto both horizontal planes (floor and ceiling). Gravity-aligned panoramas are common across all major datasets (Sec. 4.4), and alignment can be efficiently recovered in preprocessing when missing (Sec. 4.2).

Overall, while restrictive priors such as MWM, IWM, or AWM enable simpler formulations and robust constraints, they inherently limit the expressivity of inferred layouts. Several efforts have attempted to expand the solution space. Graph-based neural networks have been proposed to directly infer watertight 3D meshes of room layouts (*Deep3Dlayout* [PAAG21]), moving beyond predefined priors toward more generalizable representations. This line of work reflects a broader shift: from rigid constraint-based formulations to flexible data-driven architectures capable of handling clutter, irregular geometries, and diverse architectural patterns. However, recovering back a structured layout (e.g., ceiling, walls, floor) from the predicted 3D mesh is not immediate, so research still follows direct regression of the layout. In parallel, *SSLayout360* [Tra21] showed that semi-supervised signals (corners/boundaries learned from mixtures of labeled and unlabeled panoramas) can approach fully supervised performance with far fewer labels—useful for domains where dense GT is costly.

Transformers with geometry awareness *LGT-Net* [JXXZ22] proposed a geometry-aware transformer (SWG-Transformer) tailored to 360 layouts, capturing long-range dependencies across the horizontal scanline while respecting Manhattan constraints. LGT-Net demonstrated that self-attention over the compressed 1D sequence is effective for both simple and complex topologies; it remains a common, high-performing reference architecture and a building block for later methods. *DMH-Net* [ZWXG22] converts panoramas to cubemaps and maps features into a *learnable Manhattan Hough space*, detecting semantic lines and assembling them into 3D layouts. Working in line/Hough space improves robustness to occlusions and long straight structures, and an optimization-based post-process enforces global consistency. The trade-off is added complexity (projection, multi-face fusion) and sensitivity to non-Manhattan cases. *DOP-Net* [SZL*23] explicitly disentangles features into orthogonal plane streams (floor/ceiling/walls), with cross-scale distortion awareness and triple attention to improve 3D reasoning. By structuring features around planes, it yields strong 3D IoU and clearer interpretability than purely 1D heads. Beyond classic horizon-depth, *PanelNet* [YHJ*23] models the 360 image as continuous horizontal panels, a representation designed to leverage the panorama's seamless azimuth continuity; while broader than layout alone, it indicates a trend toward representations that better respect spherical-image structure. To directly address ambiguous supervision/evaluation, *Bi-Layout* [TJZ*24] predicts *two* complementary layouts per image—an *enclosed* one that stops at ambiguous openings and an *extended* one that spans visible adjacent areas—via two global context embeddings and a shared guidance module that conditions the image feature on the target layout type. A "disambiguate" metric then selects the best-matching hypothesis against GT, improving both overall accuracy and particularly the hardest, ambiguity-dominated subsets. This line suggests a pragmatic path forward: rather than collapsing ambiguity into a single target, learn multiple valid hypotheses with lightweight specialization and evaluate accordingly. *Seg2Reg* [STS*24] reconciles 2D segmentation and 1D regression: while the former is easier to learn and captures occlusions but needs heavy post-processing, the latter yields better geometric consistency. Seg2Reg predicts a 2D *floor-plan density* on the ERP panorama and converts it into 1D layout-depth signals through an *occlusion-aware flattened volume rendering* module, combining dense supervision with stable training and accurate geometry. It further introduces a *3D warping augmentation* and a unified re-implementation of strong baselines, consistently outperforming LED2-Net and horizon-depth pipelines by reducing the mismatch between training (2D) and evaluation (1D/3D). More recently, *HUSH* [LPLJ25] instead aligns representation and geometry using *spherical harmonics* (SH) as basis functions on the unit sphere. Scene-adaptive SH coefficients serve as structured queries in a hierarchical attention module fused with image features, and an SH indexing module selects task-relevant bases for multi-head outputs (depth, normals, layout). This SH-driven inductive bias improves geometric consistency, achieves state-of-the-art depth, and remains competitive on normals and layout, offering a distortion-robust multi-task alternative to horizon-depth and transformer-only approaches. Multi-task loss combination is a simple static weighted sum of the individual task losses.

Overall architecture evolution As summarized in Tab. 5, existing approaches differ significantly in terms of input representation (e.g., ERP, cubemap, or hybrid views), reliance on geometric priors (MW/IWM/AWM), and network design (from CNNs to graph-based and transformer architectures). Classical methods depend heavily on pre- and post-processing to enforce structural constraints, while recent data-driven solutions such as Deep3DLayout, Seg2Reg and HUSH reduce this gap by integrating geometry-awareness directly into the learning process. Overall, the trend shows a gradual shift from rigid cuboid/MW assumptions toward more flexible, geometry-aligned representations capable of handling complex room shapes and supporting multi-task predictions.

Layout completion Once a panoramic depth map or point cloud and a consistent structural layout are available, the reconstruction of scene contents (i.e., object instances, their placement, and shapes) can be obtained by adapting existing *scene understanding* techniques originally developed for perspective images. Such pipelines typically rely on object detection, 3D box fitting, and CAD-model retrieval or implicit shape representations, and can be straightforwardly applied to the panoramic setting after suitable geometric adaptation. It is

important to stress that the recovery and detailed reconstruction of individual objects constitutes a major research topic on its own, involving challenges such as fine-grained geometry and semantics. While panoramic context provides valuable cues for these tasks, they fall beyond the scope of the present survey, which concentrates on the layout reconstruction problem as the structural backbone for subsequent scene modeling.

In this context, 3D object detection approaches [SX16, QLW*18] aimed to infer 3D bounding boxes and object poses from 2D image representations, typically through a 2D detection stage [RHGS17]. For object reconstruction, early pipelines matched CAD models from large repositories (e.g., ShapeNet) to detected 2D proposals [CFG*15, TZEM22], while more recent methods demonstrated that *implicit neural representations* outperform grid-, point-, or mesh-based encodings in capturing geometry and in learning shape priors [MON*19, SMB*20]. A second line of research has explored *joint learning* of multiple tasks to exploit contextual dependencies between layout and objects. For instance, CooP [HQZ18] introduced a cooperative training scheme for object poses and indoor layouts, while Total3D [NHHN20] was the first to address layout estimation, object detection/pose, and object reconstruction jointly from a single view. Follow-up works further improved reconstruction fidelity by coupling implicit functions with graph-based reasoning [ZCC*21, LZC*22]. Despite their progress, these methods operated exclusively on perspective images, which inherently limit contextual coverage.

The introduction of panoramic imagery enabled holistic parsing of indoor scenes. PanoContext [ZSTX14a] pioneered the use of 360° panoramas for layout estimation, while subsequent works [ZCC*21] leveraged hybrid frameworks combining image features with scene graphs to jointly estimate layouts, object shapes, and poses from a single panorama. More recently, PanoContext-Former [DFB*24] proposed to directly lift a panoramic image into a volumetric 3D representation and to employ a transformer-based context module, which models global interactions between layout and object tokens for holistic panoramic scene understanding.

Transformers have become the dominant backbone across many domains, initially in NLP and later in vision tasks with the Vision Transformer (ViT) [DBK*20]. Numerous hybrid CNN–Transformer architectures [LLC*21, TCD*21] have since been proposed to capture both local and global context efficiently. In multimodal learning, CLIP [RKH*21] jointly trained vision and text encoders, framing image classification as text retrieval. CLIP uses a contrastive learning approach that, in a pretraining base, learns a multi-modal embedding space by jointly training an image encoder and text encoder to maximize the cosine similarity of the image and text embeddings of the N correct pairs in the batch while minimizing the cosine similarity of the embeddings of the $N^2 - N$ incorrect pairings. The encoder is then specialized per task in the downstream phase. By contrast, Hu and Singh [HS21] exploited pre-trained language models [RSR*20, LLG*19] within a transformer encoder-decoder architecture [VSP*17] to encode visual and text inputs and generate label text. For both pretraining and downstream tasks, the model parameters are optimized by minimizing the negative log-likelihood of label text tokens given an input text and an image. In 3D scene understanding, transformers have been successfully applied to fuse object- and point-level features, as in Group-Free [LZC*21] and Mask3D [SEKL23] for 3D detection and semantic segmentation. PQ-Transformer [CZZZ22] and AnchorRec [DHX*23] further addressed joint 3D object and layout prediction from point clouds, though with limitations in handling complex non-Manhattan structures. Conditional object queries [WYC*22] have also been used to fuse point cloud and image features for robust 3D detection. Overall, transformers offer clear advantages in modeling contextual relationships across modalities and tasks, and represent the current state-of-the-art backbone for 360° holistic scene understanding.

Method	Input	Key Idea / Contribution
3D Detection / Reconstruction		
Song16 [SX16]	Persp.	Sliding-shape 3D boxes
Qi18 [QLW*18]	Persp.	Frustum PointNets
Chang15 [CFG*15]	Persp.	ShapeNet CAD retrieval
Tulsiani18 [TZEM22]	Persp.	Multi-view supervision
Mescheder19 [MON*19]	Persp.	Implicit occupancy networks
Sitzmann20 [SMB*20]	Persp.	SIREN implicit functions
Joint Layout + Objects		
CooP [HQZ18]	Persp.	Coop. training layout+poses
Total3D [NHHN20]	Persp.	Joint layout, boxes, reconstr.
Zhang21 [ZCC*21]	Persp.	Implicit + graph reasoning
Liu22 [LZC*22]	Persp.	Hi-fi holistic implicit recon.
Panoramic Holistic		
PanoContext [ZSTX14a]	ERP	First 360° layout parsing
DeepPanoContext [ZCC*21]	ERP	Image+scene graph joint est.
PC-Former [DFB*24]	ERP	Volumetric lift + transformer
Transformer-based		
ViT [DBK*20]	RGB	Patch transformer backbone
Swin [LLC*21]	RGB	Shifted-window transformer
CLIP [RKH*21]	RGB+Txt	Vision–language pretraining
Group-Free [LZC*21]	Pcloud	Obj/point fusion for det.
Mask3D [SEKL23]	Pcloud	Transformer 3D segmentation
PQ-Transf. [CZZZ22]	Pcloud	Joint objects + layouts
AnchorRec [DHX*23]	Pcloud	Anchor-based reconstr.
Cond.Queries [WYC*22]	RGB+Pcl	Multimodal det. w/ queries

Table 6: ***Summary of layout completion and holistic scene understanding methods for panoramic imaging.*** *Grouped into detection/reconstruction, joint layout+objects, panoramic approaches, and transformer-based frameworks.*

Unlike depth estimation or layout recovery, where benchmarks and metrics such as 2D/3D IoU or RMSE are relatively standardized, the evaluation of holistic scene understanding methods remains heterogeneous. Different works often adopt distinct datasets or emphasize different components of the task—whether layout fidelity, object detection accuracy, or reconstruction quality—making direct comparisons less straightforward. Importantly, these methods are not an end in themselves but serve as functional complements to layout estimation, enabling the reconstruction of scene contents when such detail is required.

Evaluation metrics Most panoramic layout estimation works report performance using a combination of 2D- and 3D-based measures, including:

- **2D IoU**: overlap between the predicted and ground-truth floorplan, usually in bird's-eye view.

- **3D IoU**: volumetric consistency between the predicted and ground-truth 3D room geometry.
- **Corner error**: Euclidean or angular distance between predicted and ground-truth layout corners.
- **Boundary accuracy**: pixel-level error rate of predicted structural boundaries in ERP space.

These metrics jointly capture both the global structural fidelity and the local geometric accuracy of layout predictions. In contrast to layout estimation, where evaluation is usually based on standardized geometric metrics, object-level recovery and holistic scene understanding require additional objectives tailored to detection and reconstruction. Common objective functions employed across these approaches include:

- **Cross-entropy / focal loss** – for semantic segmentation and classification tasks.
- **L1/L2 regression losses** – for depth prediction, layout parameters, and bounding box regression.
- **Chamfer Distance (CD)** – for aligning reconstructed and ground-truth 3D shapes.
- **Earth Mover's Distance (EMD)** – for fine-grained shape alignment in point-based reconstructions.
- **Adversarial losses** – to encourage realism in generative or implicit reconstructions.
- **Perceptual losses** – leveraging pretrained networks to preserve high-level structural features.

Training objectives In terms of loss design, most methods optimize combinations of pixel- or feature-level supervision signals. *Cross-entropy loss* is typically adopted for corner or boundary probability maps, while *L1/L2 regression losses* are used for continuous predictions such as scanline heights or depth values. Several methods incorporate *geometric consistency losses*, enforcing alignment between predicted corners, boundaries, or floor-plan projections and their ground-truth counterparts. Recent works also employ *auxiliary depth- or semantic-guided losses* to improve structural reasoning, or introduce *regularization terms* that enforce Manhattan/Atlanta alignment constraints. When semi-supervised or weak supervision is considered, *consistency losses* (e.g., between ERP and cubemap predictions, or across multiple layout hypotheses) are additionally used to stabilize training. On the other hand, training objectives for object-level recovery in panoramic scenes largely follow standard practices from point-cloud-based reconstruction and detection. Typical formulations combine cross-entropy or focal losses for semantic components with L1/L2 regression for depth and bounding boxes, while Chamfer Distance and Earth Mover's Distance are used to align predicted and ground-truth shapes. Adversarial or perceptual losses may be added to enhance realism. Overall, these objectives mirror established pipelines in generic 3D scene understanding and thus fall outside the specific panoramic focus of this survey, serving primarily as complementary tools for extending layout estimation into full scene modeling.

Multiple objectives and multi-task combination Most of the multi-task pipelines surveyed, as well as single-task pipelines that integrate multiple losses or regularization terms, aggregate objectives using a statically weighted sum. While fixed weights are conceptually simple, selecting effective values often relies on heuristics or exhaustive grid search [SK18]. Challenges arise not only from differences in loss magnitudes due to units or scales [KGC18], but also from heterogeneity across task types (e.g., classification vs. regression) [DFL23] and from conflicting gradient directions between objectives [LLJ*21]. When such issues are ignored, some task-specific losses may increase, harming individual task performance even as the overall loss decreases. To address these limitations, researchers have proposed dynamic reweighting, multi-objective optimization, and gradient combination strategies [CBLR18, YKG*20, LLK21], which offer promising directions for application to 360° pipelines.

Method	Backbone	2D IoU(%)	3D IoU(%)
(a) MatterportLayout test set results			
LayoutNet v2 [ZSP*21]	ResNet-34	78.73	75.82
DuLaNet v2 [ZSP*21]	ResNet-50	78.82	75.05
HorizonNet [SHSC19]	ResNet-50	81.71	79.11
AtlantaNet [PAG20]	ResNet-50	82.09	80.02
LED2-Net [WYS*21]	ResNet-50	82.61	80.14
LGT-Net [JXXZ22]	ResNet-50	83.52	81.11
Seg2Reg [STS*24]	ResNet-34	83.39	81.08
(b) ZInD test set results			
HorizonNet [SHSC19]	ResNet-50	90.44	88.59
LED2-Net [WYS*21]	ResNet-50	90.36	88.49
LGT-Net [JXXZ22]	ResNet-50	91.77	89.95
Seg2Reg [STS*24]	HRNet-18	92.50	90.73
(c) PanoContext and Stanford2D-3D-S test set results (3D IoU)			
		PanoContext	**Stanford**
LayoutNet v2 [ZSP*21]	ResNet-34	85.02	82.66
DuLaNet v2 [ZSP*21]	ResNet-50	83.77	86.60
HorizonNet [SHSC19]	ResNet-50	82.63	82.72
AtlantaNet [PAG20]	ResNet-50	–	83.94
LGT-Net [JXXZ22]	ResNet-50	85.16	86.03
Seg2Reg [STS*24]	HRNet-18	87.23	87.24

Table 7: ***Benchmark results of representative panoramic layout estimation methods on three widely used datasets.*** *(a) MatterportLayout, (b) ZInD, and (c) PanoContext and Stanford2D3D. The table reports 2D and 3D IoU scores whenever available, including only methods providing directly comparable results in the original papers. Numerical comparison from Seg2Reg [STS*24] experiments.*

Performance of representative methods Tab. 7 summarizes the performance of several representative methods for panoramic room layout estimation across multiple benchmark datasets. We report both 2D and 3D IoU scores whenever available, focusing on methods for which comparable results were published in the original papers. The results highlight the steady progression from early CNN- and RNN-based baselines (LayoutNet v2, DuLaNet v2, HorizonNet, AtlantaNet) toward more geometry-aware and transformer-based approaches (LED2-Net, LGT-Net). Seg2Reg further improves consistency by bridging segmentation and regression, achieving the best overall results across the ZInD, PanoContext, and Stanford2D3D benchmarks. While these results provide useful insights into how layout-based methods can be extended to full-scene parsing, their focus remains on generic 3D object detection protocols rather than on metrics tailored to panoramic layout estimation. For this reason, we include them here primarily as an illustrative complement, highlighting representative outcomes in a panoramic setting that, although

informative, remain tangential to the central scope of this survey. For reference, Tab. 8 shows the evaluation of object detection in panoramic scenes. It typically follows the conventions established in point-cloud benchmarks, relying on 3D IoU thresholds to compute mean Average Precision (mAP) across object categories. In this case, we summarize the evaluation presented in PC-Former [DFB*24] on

5.3. Positioning and connecting single rooms

While single-room depth and layout estimation provide essential building blocks, many practical applications require structured models that span entire apartments, offices, or floors.

Classical multi-room reconstruction pipelines rely on dense image collections with significant visual overlap to establish reliable correspondences and recover camera poses using methods such as Structure-from-Motion (SfM) [PGJG19]. These approaches are fundamentally different from the single-view setting, where only one panorama per room is typically available and overlap across rooms is minimal or even absent. In such conditions, conventional SfM completely fails [SSO*21], as feature-based matching cannot be applied. Our focus is therefore on single-view-centric approaches, where each panorama is independently reconstructed into a partial 3D model, and global consistency is enforced at a higher level. This is usually achieved by estimating the room layout, detecting openings such as doors and windows, and aligning them into a topological graph that connects adjacent panoramas. The resulting graph-based representation enables the assembly of a coherent floorplan and supports interactive navigation across multiple panoramic viewpoints, even without dense multi-view coverage.

The main challenge stems from the fact that a single panorama is inherently confined by the bounds of its capture location. Inferring the existence of multiple connected rooms, therefore, requires going beyond local observations and combining geometric cues (e.g., openings, vanishing lines, occlusion boundaries) with strong structural priors and learned patterns of indoor organization. While elements such as doors and corridors can act as weak indicators of adjacency, they are frequently occluded, cluttered, or only partially visible. Robust solutions must therefore integrate local geometric evidence with global reasoning strategies to hypothesize plausible multi-room connectivity from incomplete or noisy inputs.

Early efforts in this direction mainly relied on explicit geometric reasoning and energy-minimization frameworks to extrapolate single-room layouts into larger structures [FCSS09, SHKF12, CF14, MMBM15, IYF15]. However, these pipelines tended to be brittle, requiring clean inputs and often failing under the clutter and noise typical of real-world indoor panoramas. Recent research has therefore reformulated the problem into two complementary tasks. The first, *positioning single rooms*, focuses on estimating the relative placement of panoramas by leveraging coarse geometric priors and global alignment strategies. The second, *connecting single rooms*, goes one step further by requiring the accurate reconstruction of structural elements such as room layouts and door locations, which act as critical anchors for establishing adjacency and continuity between spaces. While positioning provides a global scaffold for arranging panoramas, connecting requires a higher level of structural reasoning to ensure that individual reconstructions can be seamlessly assembled into a coherent floor plan.

Positioning single rooms To achieve this goal, an important task when working with sparse views is spatial registration across multiple panoramas: given sparse single-view captures, methods such as Shabani et al. [SSO*21] attempt to estimate plausible relative room displacements to reconstruct a global layout of a building. Their pipeline departs from traditional SfM by explicitly addressing the lack of visual overlap between panoramas. Instead, it leverages geometric cues such as estimated single room layouts (Sec. 5.2) and detected openings, combined with semantic priors and architectural regularities, to hypothesize pairwise connections between rooms. These hypotheses are then embedded into a global optimization framework, which jointly refines room positions and adjacency to assemble a coherent floorplan (Fig. 6). More recent approaches have pushed this direction by introducing learned verification and generative refinement, in particular exploiting diffusion models [PZL24]. For instance, SALVe [LLB*22] introduces a semantic alignment verifier that exploits windows, doors, and openings (W/D/O) as robust cues for hypothesizing pairwise panorama alignment under extreme wide baselines. Candidate alignments are validated through a deep CNN operating on bird's-eye views of aligned floor and ceiling renderings, before constructing a global pose graph optimized with GTSAM. In parallel, BADGR [LBL*25] proposes a diffusion-based bundle adjustment framework that refines both camera poses and floor plan layouts from sparse panoramas. Unlike prior guided diffusion strategies, BADGR integrates a planar BA layer into the denoising process, enforcing view-consistency while injecting learned structural priors (e.g., collinearity, wall adjacency) to plausibly complete occluded parts of the layout. Together, these methods exemplify a shift from purely geometric heuristics towards hybrid pipelines where semantic cues and generative models enable both positioning and floorplan reconstruction in the challenging regime of sparse, wide-baseline panoramic captures.

Connecting single rooms To reliably connect multiple panoramas, it is not sufficient to estimate their relative placement: some form of floorplan reconstruction, even in a simplified representation, is required to provide the structural scaffold that links individual rooms together. Floor plan reconstruction is a more complex problem than just relative localization. State-of-the-art multi-room reconstructors usually require the availability of reliable density maps for their analyses [CQF22, YKSE23]. These maps, accumulating the occurrence of 3D points projected onto the floorplan, are built from dense point clouds that are hard to generate with sufficient precision using single-view inference from purely visual data [PSAG25]. In the context of dense coverage (i.e., density maps from point clouds), recent hybrid approaches leverage deep networks to detect low-level primitives (e.g., corners, edges, or room regions) and assemble them into structured floor plans through optimization. Floor-SP [CLWF19] and Nauata et al. [NF20] employ Mask R-CNN [HGDG17] to segment room regions and reconstruct polygons via shortest-path optimization, while MonteFloor [SRFL21] explores multiple room arrangements with Monte-Carlo Tree Search. Similarly, FloorNet [LWF18] detects corners and generates wall segments using integer programming to yield coherent layouts. End-to-end solutions further extend this trend: HEAT [CQF22] predicts corners and edges with a transformer-based backbone, RoomFormer [YKSE23] generates polygonal sequences from sparse 3D cues, SLIBO-Net [STP*23] integrates Manhattan World priors to

Method	cab.	door	chair	curtain	lamp	rug	sofa	table	trash	TV	mAP↑
DeepPanoContext [ZCC*21]	35.33	6.78	47.04	13.60	12.15	4.49	26.87	73.34	39.59	4.86	26.41
DeepPanoContext-3D [ZCC*21]	52.49	11.42	70.39	32.38	20.02	9.10	30.13	82.24	63.22	12.19	38.36
Group-Free [LZC*21]	59.56	42.21	52.83	34.07	19.65	32.90	80.59	51.47	44.64	52.76	47.07
PC-Former [DFB*24]	63.69	46.74	54.02	30.41	20.04	48.53	80.96	46.42	51.53	47.82	49.02

Table 8: ***Object detection results*** *on the 360MonoDepth/Replica [RAYR22c] dataset, where mAP is a 3D IoU threshold of 0.15 over 10 representative categories, reported by PC-Former [DFB*24]. DeepPanoContext [ZCC*21] predicts 2D object layouts from panoramic images, while DeepPanoContext-3D extends the pipeline by lifting predictions into a volumetric 3D space for direct object reconstruction. Unlike layout estimation, these benchmarks follow point-cloud-based detection protocols; we include them here as illustrative examples of object evaluation in the panoramic context, which remains ancillary to the main focus of this survey.*

Method	Input	Architecture	Output
Extreme SfM [SSO*21]	Sparse panoramas without visual overlap	Single-image layout + Geometric optimization	Displacements of rooms; Coarse global floorplan
SALVe [LLB*22]	Sparse panoramas with minimal overlap	Learned structural priors + optimization	Panorama positioning and layout alignment
BADGR [LBL*25]	Wide-baseline panoramas	Diffusion-based bundle adjustment	Accurate relative localization of rooms
NadirFloorNet [PSAG25]	Single panorama per room + positions	Single-image layout + deformable attention	Metrically coherent 3D floorplan
PanoFloor [PJAG25]	Sparse panoramas with relative positions	Diffused density map + deformable attention	Floor plan layout, doors, windows, clutter; connectivity graph for VR-ready model

Table 9: ***Comparison of representative methods for sparse multi-room panoramic reconstruction.*** *Each method is characterized by its inputs, core architecture, and produced outputs.*

refine geometry, and PolyDiffuse [CDF24] applies conditional diffusion to infer plausible multi-room structures from density maps. In the context of extreme multi-room reconstruction from sparse panoramas, *NadirFloorNet* [PSAG25] extends the transformer-based *RoomFormer* [YKSE23] to single-image-per-room settings. The method first predicts a clutter-free nadir depth projection to recover room geometry, and then fuses these representations into a consistent global floorplan using deformable attention, given the image positions [SSO*21, LLB*22, LBL*25]. By jointly estimating ceiling heights and 2D polygons, it achieves metrically coherent 3D layouts from extremely sparse inputs. *PanoFloor* [PJAG25], instead, takes as input the relative positions and individually predicted depths of multiple panoramas (Sec. 5.1) and reconstructs a global floor plan enriched with semantic elements. Specifically, the method estimates room layouts, places doors and windows, and accounts for clutter, while also building a connectivity graph that describes the navigable structure of the environment. This representation enables the creation of a complete, VR-explorable model that integrates geometry, semantics, and accessibility (Sec. 6.1.3). In summary, as shown in Tab. 5.2, these approaches span from geometry-based positioning methods to transformer- and diffusion-based pipelines that jointly recover localization and structured floorplans, progressively moving towards metrically consistent, semantically enriched, and VR-ready reconstructions. These pipelines further highlight how depth and layout estimation remain fundamental tasks, as they provide the geometric and structural backbone required for reasoning about room connectivity and assembling coherent multi-room models under such sparse conditions.

6. Novel view synthesis and immersive model exploration

While geometric and structural descriptions provide a strong foundation, they are not enough for interactive use. We review techniques for exploration from a single enriched panorama, including methods that recover stereo cues and motion parallax and support advanced navigation in sparsely-sampled multi-room environments (Sec. 6.1), and approaches that go beyond plain navigation (Sec. 6.2).

6.1. View synthesis from a single panorama

Although a single-shot panorama offers an attractive way to capture and replicate real environments, it inherently limits content to what is visible from a fixed viewpoint [WGD*22]. This restriction reduces degrees of freedom to pure rotation about the panorama center, leading to artifacts and perceptual limitations. In particular, binocular disparity and motion parallax, which are two key components of immersive VR, are absent, causing indoor panoramas to appear flat. Moreover, clutter and occlusions often prompt users not only to change their viewing angle but also to attempt translational movements to peek around objects or structural elements, but such data is not present in the original panorama [MCE*17]. To achieve convincing immersion, therefore, systems must respond to viewpoint translation as well as rotation, making novel-view synthesis from a single panorama a central research challenge [WGD*22].

This involves combining occlusion-aware reprojection to present to the viewer the originally captured surfaces at the correct image location, with the completion of the image in the disoccluded areas. While augmenting panoramas with depth, using one of the techniques illustrated in Sec. 5.1, significantly helps reprojection, novel solutions must be introduced for completion. The problem is com-

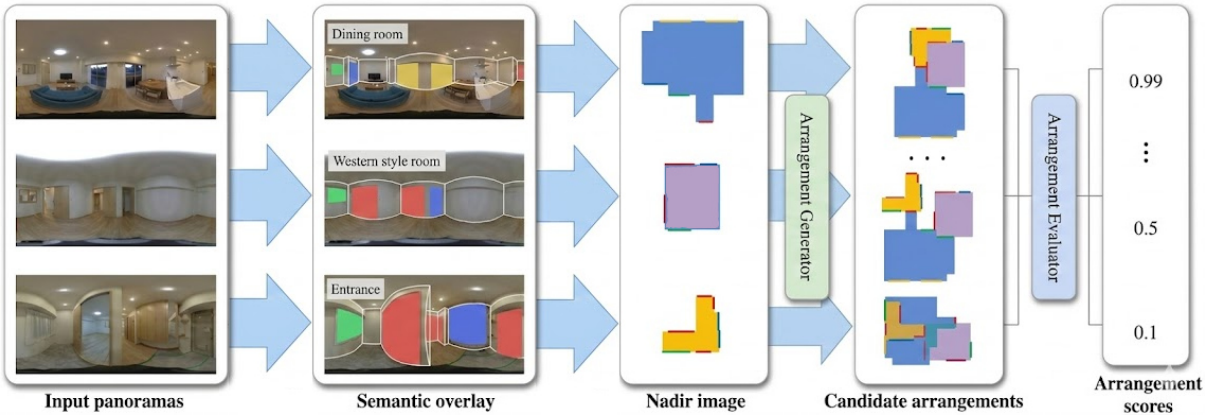

Figure 6: ***Positioning single room layouts example*** *[SSO*21]. The pipeline leverages geometric cues, such as estimated single-room layouts (Sec. 5.2) and detected openings, and combines them with semantic priors and architectural regularities to hypothesize pairwise room connections. These hypotheses are then integrated into a global optimization framework that jointly refines both room positions and adjacencies, ultimately assembling a coherent and metrically consistent floorplan.*

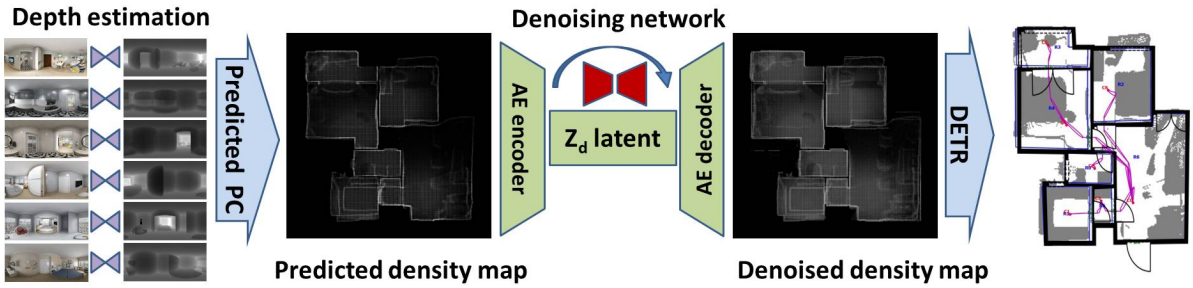

Figure 7: ***Floorplan layout and connectivity graph recovery example*** *[PJAG25]. The pipeline takes as input the relative positions and individually predicted depths of multiple panoramas to reconstruct a global floor plan enriched with semantic elements (layouts, doors, windows, and clutter), also building a connectivity graph that describes the navigable structure of the environment.*

plex, as not only must the newly generated content be plausible and consistent with the originally sampled data, but it must be consistent throughout time during interactive exploration. In particular, the appropriate strategies are fundamentally determined by the extent of the newly exposed regions induced by viewpoint motion, and thus by the nature and complexity of the inference required to plausibly fill the missing content.

For *small-to-medium viewpoint displacements*, such as those induced by stereo or head motion for limited parallax, most of the visible content remains close to what was originally observed. Disoccluded regions are typically narrow and localized, often appearing along object boundaries or depth discontinuities. As a result, infilling can rely heavily on nearby visual evidence, geometric consistency, and depth-aware reprojection, requiring relatively limited semantic or structural inference. Methods in this category, thus, have to primarily focus on accurate reprojection and lightweight completion to maintain visual coherence while minimizing artifacts during interactive viewing. Moreover, the constrained motion amount also opens the door to specific optimization meant to produce compact and efficient restricted 3D representations. We provide an overview of this class of methods in Sec. 6.1.1.

In contrast, *large viewpoint displacements* introduce extensive disocclusion, revealing substantial portions of the scene that were entirely unseen in the original panorama. In this regime, infilling cannot be guided solely by local image evidence, as large areas of missing content must be synthesized without direct observations. Consequently, these methods must make stronger assumptions and rely on richer scene representations, such as explicit 3D reconstructions, learned priors over indoor geometry, or generative models capable of hallucinating plausible structure and appearance. The increased reliance on inference and semantic understanding makes large-motion approaches inherently more complex, not only in terms of methods and generated representations, but also necessary to support unrestricted navigation and room-scale exploration. We provide an overview of this class of methods in Sec. 6.1.2.

Building on these techniques, specialized methods have also been introduced to handle *room-to-room navigation*, where viewpoint changes can span multiple connected spaces. These approaches must not only synthesize previously unseen content within individual

rooms but also ensure spatial and visual consistency across room boundaries, often requiring integration of global scene layout to generate plausible and navigable transitions between rooms. We summarize this class of approaches in Sec. 6.1.3.

6.1.1. Handling limited disocclusions under small-to-medium viewpoint motion

A first class of approaches focuses on supporting restricted viewpoint changes, such as stereo generation or limited head motion. Stereo can be modeled as eye positions moving along a circular arc when the head rotates around the capture point, while small parallax corresponds to translations within a compact volume. Prior studies on novel-view synthesis [SKC*19] have reported that head displacements during interactive VR exploration generally remain within a range of approximately 35 cm. In such cases, specialized methods can be designed to manage modest disocclusions and perspective changes, enabling compact, fast-rendering representations tailored to the capture environment.

Depth estimation and reprojection A general approach is to pair a panorama with an inferred view-independent depth map (Sec. 5.1) and render the resulting textured spherical meshes to achieve parallax [ZWF*13, HCCJ17, WZS*18, TPG*23]. These methods are efficient, since heavy computation is offline, but suffer from artifacts in disoccluded regions due to view-independent interpolation.

Indoor-specific inpainting and synthesis techniques Quality can be improved by rendering the depth image as a point cloud from the target view and using inpainting to fill disoccluded areas, as done in perspective images with classic [SSK19] and learning-based [YLY*19, ZFCG19] methods. Structural priors and geometric constraints have proven useful for guiding novel color generation [WGSJ20]. For indoor panoramas, end-to-end view synthesis networks exploiting domain-specific priors have been proposed to generate shifted views at runtime [XZX*21a, PBAG23]. In these methods, a first offline stage extracts from the input panorama a dense feature map, depths, and room layout information (Sec. 5.2). In the second online stage, the information is reprojected via forward splatting and layout transformation, and the final complete view is produced from this partial view. Fig. 8 illustrates the pipeline of Xu et al. [XZX*21a]. Training uses synthetic databases (e.g., PNVS [XZX*21b, XZX*21a]) with source-target pairs. While enabling free-form motion, computational demands restrict the use of these methods on HMDs to remote rendering with upscaling [PBAG23].

Precomputed representations for small displacements To reduce per-frame synthesis costs, researchers have proposed a number of precomputed intermediate representations suitable for accelerating subsequent rendering. Classical methods were designed on multi-view input (e.g., Neural Reflectance Fields NeRF [MST*21], 3D Gaussian Splatting (3DGS) [KKLD23], depth- or flow-based proxies [BCR19, BYLR20, LXRY18, HCCJ17]), with 360°-specific variants using layered images [HASK17, HK18], multi-depth panoramas [LXM*20], layered meshes [BFO*20], or panoramic Gaussian splats [BHG*25]. Recent work has adapted these multi-view representations to single-view input using synthesis approaches previously described to generate the required multi-view input. Multi-plane images (MPI) [ZTF*18], later adapted to panoramas as multi-spherical (MSI) [ALG*20] or multi-cylinder images (MCI) [WGD*22] are representations in which scenes are represented through geometrically simple slices taken at fixed distances. Their regular structure makes them suitable as a target for inference. Tucker and Snavely [TS20] introduced a method to infer an MPI from a single perspective image, and *PanoSynthVR* [WGD*22] extended this approach to MCI representations. While effective for small motions, they blur at disocclusions and degrade rapidly with viewpoint shifts if the number of layers is maintained low, which is necessary to meet memory and rendering speed constraints. In their original presentation, no indoor-specific information is used for inference (e.g., layouts, as in indoor-specific view synthesis [XZX*21a, PBAG23]). Recently, interest has shifted from layered representations to neural representations. A prominent example is the 3603DPhotos [RAR25] approach. Depth is first estimated from the input 360° panorama using methods from Sec. 5.1. The resulting RGB-D panorama is then projected into 20 perspective RGB-D images, centered at the original capture position and uniformly distributed on the sphere. These images initialize a set of 3D Gaussian Splats, forming a first approximation of the 3D scene explaining the original 360° input. To support head movements, novel virtual views are sampled by shifting the origin within a small sphere (50 cm radius) from the original capture positions, and disoccluded regions in these views are inpainted with variants of the methods described earlier. The inpainted content is added as new Gaussians, and the representation is optimized to align with both the input and the inpainted images. After iterative refinement, the final result is a full implicit neural representation valid for all views within the defined head movement volume. The representation is reasonably fast to render and much more expressive than layered ones, but has a much higher memory occupation since, in 3DGS, each 3D Gaussian requires 48 parameters for view-dependent color representation via spherical harmonics and 7 parameters encoding anisotropic scale and rotation. A similar approach, for NeRFs rather than 3DGs, is provided by PERF [WWC*24], which is much more compact, but also much slower to render, especially on embedded devices. Recently introduced 3DGS compression solutions [BKL*25] might be beneficial, but have not yet been adapted to 3DGSs created from single panoramas.

Precomputed representations for omnidirectional stereo Since stereo cues are especially critical indoors, due to the closeby objects, researchers have strived to create optimized representations for that case [WGD*22]. The solution space for novel view synthesis is constrained by the fact that during head rotations in HMDs, the eyes trace a small circular path around the capture point, with a radius of $\approx 10cm$. Each view can thus be parameterized as a function of the angle. PanoVerse [PJH*23] takes the straightforward approach of synthesizing a set of equally spaced views along the path. At display time, separately for each eye, the two images angularly closer to the current viewing direction are selected and cross-blended to provide a continuous experience. Storage constraints restrict the number of images, causing ghosting in transition regions. More compact solutions use omnidirectional stereo projection, a multiperspective approach [RB98, PBE99] that encodes stereo in two equirectangular images with per-column centers of projection, displayed like conven-

Family	Core idea	Representation	Range	Characteristics	Represent. methods
Depth estimation, reprojection, and infilling	Infer per-pixel depth from RGB panorama; reproject to target view; fill holes	Offline depth estimation, online computation from RGB-D	Small	Low memory overhead; fast rendering; per-frame infilling valid only for small disocclusions; requires temporal smoothing	[ZWF*13, HCCJ17, TPG*23]
Layout/semantic-guided IBR	View synthesis from reprojected RGB-D	Infer per-pixel depth from RGB panorama; reproject and synthesize target-view image	Small	Low memory overhead; fast rendering; data-driven completion consistent with room structure; valid only for small disocclusions; requires temporal smoothing	[XZX*21a, PBAG23]
Omnidirectional stereo	Synthesize views on the eye's path, merge them into a pair of MCOP panoramas	MCOP panorama	Stereo with head rotation	Extremely compact, very fast rendering, but stereo-only; quality drops slightly in the peripheral areas and when the view direction converges towards the poles	[Bou10, PJVH*24]
Layered representation (MPI/MSI/MCI)	Discretize scene into fronto-parallel (or spherical/cylindrical) layers; alpha compositing for parallax	Discretized near-volumetric representation	Small–mediun	Low memory overhead only with few layers; quality degrades quickly far from original viewpoint unless many layers; good temporal continuity	[ZTF*18, TS20, ALG*20, WGD*22]
Explicit 3D model extraction	Lift to explicit 3D model from single panorama via layout estimation and object recognition	3D model	Small–large	Full renderable model with semantics available; works mostly in limited-vocabulary contexts	[XSKT17, YJL*18, ZCC*21, DFB*24]
Neural representation from pseudo-multiviews	Hallucinate RGB-D images for novel viewpoints, merge them into a consistent 3D neural models	Neural (3DGS, NeRF)	Small–large (method-dependent)	Learned hallucination for disoccluded content mostly using foundation models; reconciling views proves hard	[KYS23, WWC*24, PZL24, ZCY*24, RAR25]

Table 10: ***Technique families used for*** **novel-view synthesis from a single panorama**. *We separate approaches by representation and typical motion ranges, separating between small-to-medium (Sec. 6.1.1) and larger motions (Sec. 6.1.2).*

tional stereo panoramas [Bou10]. Pintore et al. [PJVH*24] introduce a deep system that creates the representation by blending panoramic slices inferred from a single view with a custom view-synthesis network (Fig. 10). The method requires very limited rendering resources, but, as for all omnidirectional stereo methods [MP21], quality drops slightly in the peripheral areas and when the view direction converges towards the poles. Since depth is available, future work may leverage runtime, gaze-dependent adaptations [MP21].

6.1.2. Novel content generation for large viewpoint displacements

When the user translates significantly from the capture location, the panorama must be lifted into a full 3D renderable model either by reconstructing a renderable geometric model using semantic priors or by synthesizing novel views with generative techniques and fusing them into a consistent model.

Layout reconstruction with object recognition One strategy exploits semantic priors: methods such as Pano2CAD [XSKT17], Auto3DIndoor [YJL*18], DeepPanoContext [ZCC*21], and PanoContextFormer [DFB*24] combine layout estimation with object recognition to reconstruct complete rooms. The general concept behind this class of objects is to apply variations of the methods presented in Sec. 5.2 to determine the architectural boundary of the room, and create a textured model of the emptied space. Then, individual objects recognized in the panoramic image are extracted and replaced with textured meshes. These approaches produce a very compelling representation, close to a building information model (BIM) with a structural, semantic, and visual model that can then be used for a variety of downstream applications, and not only for rendering. However, real environments often contain clutter and objects with unknown semantics, limiting reconstruction quality. For this reason, the methods tend not to work well in an open-vocabulary context [PZL24].

Neural 3D renderable representations from a single panorama Learning-based renderable representations such as Neural Radiance Fields (NeRFs) [MST*21] and 3D Gaussian Splatting (3DGS) [KKLD23] offer an alternative to explicit shape and material models, producing photorealistic novel views but, in their original formulation, typically require many input images to optimize their parametric representation. Recent extensions address single-panorama input by synthesizing auxiliary, but possibly low-quality, views with variations of the methods described in previous sections, then training NeRF or 3DGS models to reconcile them. Examples include DietNeRF [JTA21], Pix2NeRF [CODVG22], SinNeRF [XJW*22], NerfDiff [GTL*23], NerDi [DJQ*23], 360FusionNerf [KYS23], PERF [WWC*24], PixelSplat [CLTS24], HoloDreamer [ZCY*24], Pano2Room [PZL24], 3603DPhotos [RAR25]. State-of-the-art pipelines [KYS23, WWC*24, PZL24, ZCY*24, RAR25] typically generate RGB-D images for novel viewpoints and apply collaborative inpainting to fill missing regions before consolidating them into consistent 3D neural models. Fig. 9 illustrates the concept behind the iterative creation of a 3D Gaussian Splat representation from a single panorama. While the method is mostly designed for small-to-medium motion ranges, it can also potentially

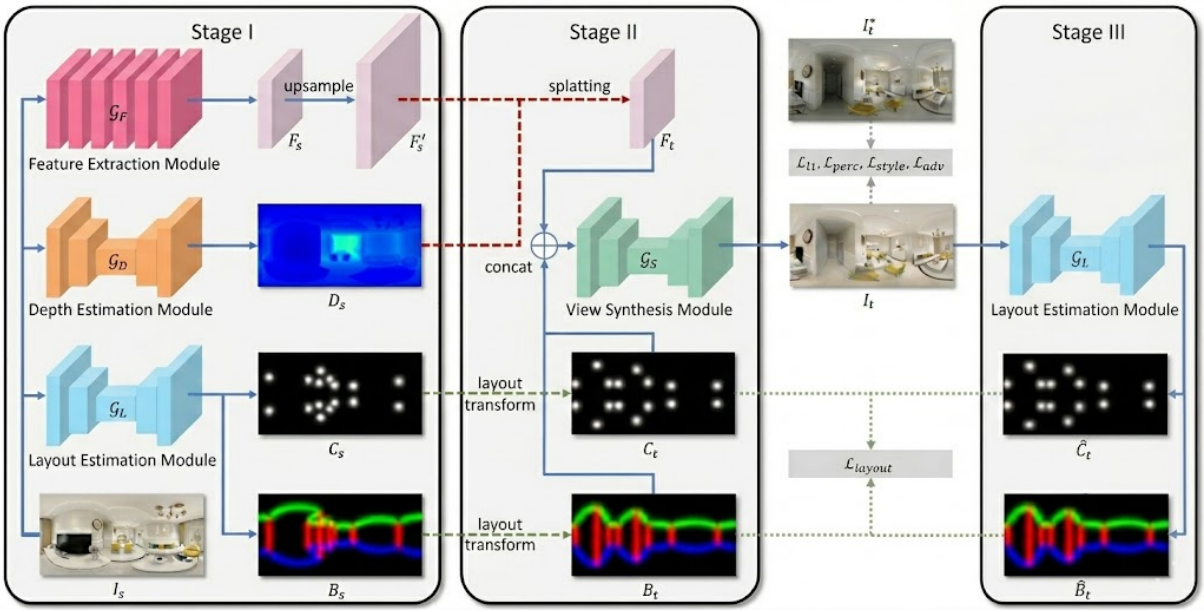

Figure 8: ***Layout-guided view synthesis*** *The method introduced by Xu et al. [XZX*21a] shows the close connection between depth estimation, layout estimation, and novel view synthesis.*

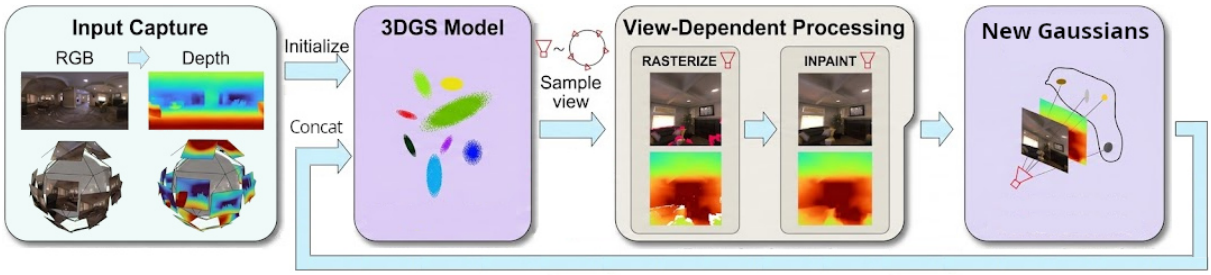

Figure 9: ***Synthesis of a 3D Gaussian Splat representation from a single panorama*** *[RAR25]. Depth is first estimated from the input* 360° *panorama to form an RGB-D representation, which is projected into multiple uniformly distributed perspective views to initialize a 3D Gaussian Splat scene. To enable limited head motion, novel nearby viewpoints are synthesized, disocclusions are inpainted, and the newly generated content is incorporated into the Gaussian representation through joint optimization with the original and synthesized views.*

be used for larger motion ranges by using pretrained generative models [GCC*24] for the inpainting process [KYS23,RAR25]. However, sequential inpainting may constrain virtual camera trajectories and produce ghost geometries, while inconsistent novel views can introduce blurring. Pano2Room [PZL24] improves robustness under large motion ranges by first estimating a mesh, iteratively refining it with panoramic RGB-D inpainting, and finally converting the result into a 3DGS optimized with pseudo-novel views (Fig. 11). Starting from a panorama augmented with measured or predicted depth, the method synthesizes an initial mesh. It then iteratively selects camera poses with the lowest view completeness, reprojects the current mesh from these views, completes the panoramas with an RGB-D inpainter, and merges the recovered geometry while discarding conflicting regions. The new, consistent geometry is used for the subsequent view planning and iterative improvement generation. The refined mesh is ultimately transformed into a 3DGS and trained using the collected novel pseudo views. Since inpainting is performed on reasonably large areas using a pre-trained generative diffusion model for standard images [RBL*22], this strategy enables the insertion of novel content to support large motions, but also introduces challenges: diffusion-based inpainting may generate inconsistent geometry, and the incremental refinement loop can make maintaining global consistency difficult.

6.1.3. Connecting multiple rooms

Generative view-synthesis methods, such as those summarized in Sec. 6.1.2, are, in principle, capable of handling arbitrary displacements and could be applied to multi-room navigation. For instance, *360Roam* [HCZY22] jointly extracts a floor plan and a NeRF representation of the environment from multi-view spherical input. The approach is, however, hard to apply to multi-room environments with just one image per room and little or no overlap.

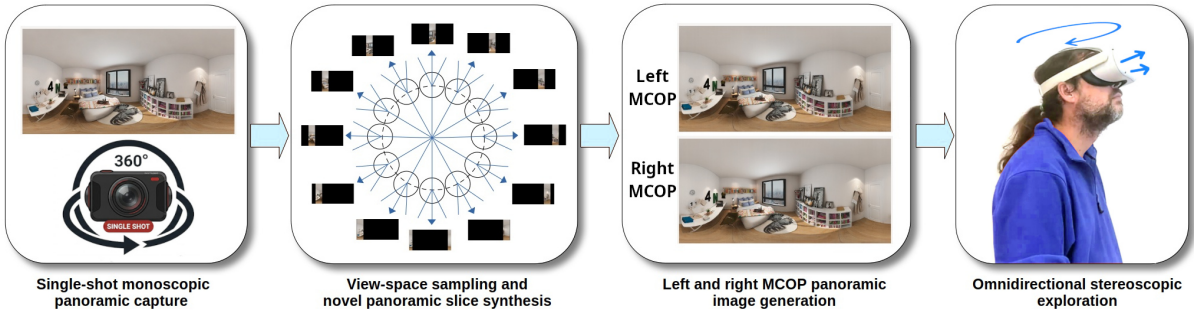

Figure 10: ***Synthesis of omnidirectional stereo pairs from a single panorama*** *Slices generated by a view-synthesis network are composed in a single multiple-center-of-projection equirectangular image per eye [PJVH*24].*

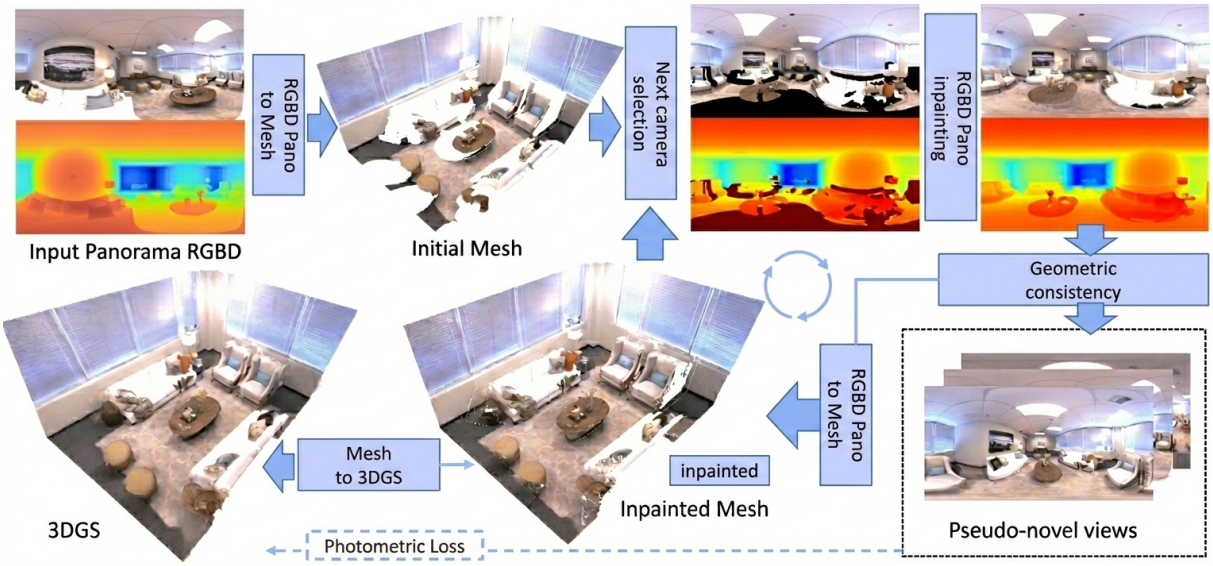

Figure 11: ***Using geometric consistency for synthesizing a 3D Gaussian Splat representation from a single panorama*** *[PZL24]. With a panorama as input, augmented with measured or predicted depth, the method synthesizes an initial mesh. Next, it iteratively searches for a nearby camera pose with the lowest view completeness, reprojects the current mesh from the selected views, completes the panoramic view with an RGBD-D inpainter, and merges the new geometry with the current geometry, excluding portions that create conflicts. Finally, the inpainted mesh is converted to a 3DGS and trained with collected pseudo novel views.*

A full precomputation supporting arbitrary motion in such environments would be very costly to store and would be hard to generate while maintaining consistency everywhere. At the same time, generative methods have a high generation cost that hinders their run-time application to low-latency and high-frequency rendering in response to user motion. For this reason, the typical approach is to use them as building blocks for precomputing videos along paths that connect the original sampling locations.

The problem of generating smooth video transitions between panoramas has been extensively studied in the literature. Early approaches typically relied on detecting and matching feature points between the source and target panoramas, and then constraining the interpolation through warping combined with cross-blending [ZWF*13]. While effective in simple cases, this strategy is prone to ghosting and structural deformations when the underlying geometries do not align well, an issue that becomes particularly severe for wide-baseline transitions, especially indoors, due to the presence of very nearby objects. To mitigate these artifacts, later work has integrated depth estimation and view synthesis, enabling more natural intermediate frames and thus perceptually smoother interpolations between panoramas (e.g., [LZH*22, CCG*23, SZ25]). Still, these methods require a good amount of overlap between panoramas, which is often hard to obtain with just one panorama per room. Moreover, the confined nature of architectural spaces, the presence of occluding furniture, and the non-convexity of room layouts require more sophisticated trajectory design than simple generation of intermediate frames along linear paths.

To address this, classical methods (e.g., [DBGBR*14]) have proposed computing curved transition paths that respect collision-avoidance constraints and smoothness objectives, with later refine-

ments optimizing viewpoint selection for visual quality and user immersion [FSG09, BKH*23]. These techniques, however, typically assume access to detailed geometric models, resources that may not always be reliable in sparse capture settings. To address this, Pintore et al. [PJAG25] generate collision-free multi-room paths that enforce door traversal and clutter avoidance while minimizing the number of disoccluded pixels to preserve visual quality.

Building on purely visual input, a number of authors have explored the use of learned aesthetic metrics to guide viewpoint and trajectory selection. In particular, neural aesthetic predictors have been optimized directly on synthesized views [WZS*18], allowing systems to automatically identify visually pleasing camera positions [ALB21, UKE25] and to plan navigation paths that maximize aesthetic quality [XHS*23]. These methods represent an important step toward integrating perceptual considerations into navigation design, but they are typically restricted to relatively simple scenarios, such as navigation within a single room. As a result, they do not address the structural complexity and occlusion challenges that arise in multi-room indoor environments.

6.2. Dynamic modifications

The components surveyed in Sec. 5 deliver pixel-wise depth/normal/semantic estimates, room layout abstractions, and graph connectivity across spaces. These reconstructions are not only useful for navigation (Sec. 6.1), but they are essential to support editing with limited or no user input, as geometry proxies guide occlusion reasoning and re-projection, semantics provide category-level control, and coarse illumination estimates support photometric consistency. We discuss three common editing tasks enabled by this panorama-first pipeline:

1. *Diminished reality/clutter removal* aims to remove objects or categories (e.g., furniture) and plausibly synthesize the newly revealed background at the original viewpoint.
2. *Virtual staging & inverse rendering* insert, rearrange, or restyle furniture while maintaining physical plausibility by estimating intrinsic scene factors (geometry, reflectance, and illumination).
3. *Style transfer* re-styles the overall scene (appearance, materials, or tone) while preserving the spherical geometry and semantics of the panorama.

Diminished reality and emptying Early panorama-specific works demonstrated that category-aware removal combined with layout priors can *empty* cluttered rooms with minimal supervision. Pintore *et al.* perform instant automatic emptying by fusing semantics with layout-consistent inpainting and geometry-aware view constraints, producing background completions that respect walls, floors, and large structural boundaries [PAAG22]. Complementarily, *PanoDR* targets spherical diminished reality for indoor scenes by operating directly in equirectangular space and enforcing spherical consistency during removal and completion [GSZ*21]. More recently, Slavcheva et al. [SGC*24] exploited diffusion models for automatic defurnishing. Dolhasz et al. [DMG*25] extended this approach by integrating a mesh representation to enhance the controllability of the defurnishing diffusion model. These approaches illustrate a key theme of this survey: the same reconstruction signals that make navigation robust (depth, normals, layout) also reduce ambiguity during removal and completion, allowing largely automatic pipelines with only coarse category selection (or none at all).

Virtual staging via inverse rendering and object insertion Virtual staging benefits from disentangling appearance into geometry, reflectance, and light so that inserted assets "sit" correctly, cast/shadow realistically, and respond to the scene illumination. In practice, inverse rendering from *one* 360° image is severely under-constrained: to explain the observed radiance, the system must *jointly* estimate multiple dense signals—geometry (depth, normals), intrinsic layers (albedo and shading), material/BRDF cues, semantics, and HDR illumination—while remaining consistent with spherical projection. Multi-task dense-prediction frameworks designed for panoramas address this need by producing several of these cues in one pass [ZCB*22, LWH*22b, STA*24]. Such multi-signal estimates stabilize contact and occlusion reasoning, improve shadow and reflection plausibility, and supply strong priors for subsequent optimization. Exploiting these ideas, SSAD learns a semantically supervised appearance decomposition from a *single* panorama, enabling realistic staging without multi-view capture [ZCB*22]. *PhyIR* brings explicit physics into the loop, recovering scene intrinsics and environment illumination for panoramic indoor images, which in turn improves relighting and consistent object shading [LWH*22b]. Finally, *VISPI* integrates multi-task learning (geometry/semantics/illumination) [STA*24] with inverse rendering to support turnkey virtual staging from one panorama, including insertion and material harmonization, and contact reasoning [SJT*25]. Recent diffusion-based approaches further boost photorealism and reduce user effort by coupling generative priors with physically grounded factors [PZL24]. Diffusion-guided inverse rendering optimizes geometry/BRDF/lighting to match a powerful image prior, enabling high-fidelity object insertion with correct material response and shadows [LGND*24]. Latent-intrinsic formulations jointly reason about intrinsics and appearance in a diffusion backbone to achieve robust indoor relighting and material edits [XGK*25], while diffusion-driven single-image illumination estimation produces HDR environment lighting suitable for staging and re-rendering [SBXX25]. Operating directly in an intrinsic-image latent space, *IntrinsicEdit* provides precise, local manipulations (materials, global relighting, insertion/removal) with improved identity preservation and consistent global illumination [LDHG*25]. Across these methods, the proxies from Sec. 5.1 and Sec. 5.2 (depth, normals, floor/wall estimates) are pivotal for contact estimation, collision checking, and occlusion ordering, while Sec. 5.3 adds spatial context for multi-room staging scenarios.

Panoramic style transfer Panoramic style transfer must respect spherical distortions and large field-of-view coherence. *PanoStyle* couples semantic and geometry cues to achieve shading-independent photorealistic transfer in indoor panoramas, preserving structural boundaries and material identity [TUP*23, TJAH*25]. *PAST* introduces deformable distortion constraints tailored to panoramas, improving the alignment of stylistic features over long geodesics [YWL*25]. When multiple panoramas are available (e.g., per-room captures connected in Sec. 6.1.3), *PIST* targets cross-view consistency via multi-scale attention and global feature sharing to avoid view-to-view style drift [WQTX25]. Together, these techniques com-

plement staging by harmonizing inserted assets with existing decor and enabling global appearance changes without full re-rendering.

Discussion Diminished reality, virtual staging with inverse rendering, and style transfer are three faces of the same capability: *editability from a single capture*. The reconstruction signals established earlier provide the geometric and photometric scaffolding that turns one-shot panoramas into editable scenes. As a result, modern systems increasingly offer end-to-end, low-touch workflows—*empty* first to declutter, *stage* with inverse-rendered consistency, and optionally *re-style*—all while remaining compatible with the interactive view-synthesis tools of Sec. 6.1. Concurrently, latest research trends increasingly aim to generate *entire 3D scenes* from scratch or with minimal user input, such as a textual prompt. In the context of *panoramic* scene generation, the last year has seen a surge of diffusion-based methods that operate directly in the spherical domain or in equirectangular projection, producing 360° RGB images (and sometimes approximate geometry/semantics) from text prompts. Representative examples include spherical/equirectangular text-to-360 diffusion models [ZFX*25, YJC*24], controllable room synthesis that accepts layout or category constraints [FDL*25, ZWG*24], layered or geometry-aware generators that target 3D-consistent outputs from a single panorama [YTZ*25], and systems or datasets that scale to multi-room and world-like environments [HHY*25, YDH*25]. While training objectives and architectural details vary, these approaches share a goal: *photorealistic, prompt-aligned 360 content with enough structural coherence to support downstream use.*

7. Emerging pre-trained and foundation models

As discussed in the previous sections, extracting geometric, structural, and semantic information from a single panorama (Sec. 5) and augmenting these representations to support dynamic exploration and interaction (Sec. 6) involve solving highly ill-posed problems, which necessitate the use of strong priors. Learning such priors by leveraging massive, predominantly annotated datasets to extract latent information relevant to a given task has proven highly effective, but it also raises substantial practical challenges. Learning such priors from massive, task-specific datasets has proven highly effective, but it also raises substantial practical challenges. Suitable datasets are often limited to specific environments or unavailable for the targeted tasks, and creating them is costly, particularly when aiming to cover the wide variability of indoor environments. Moreover, even when appropriate data exists, training specialized models from scratch remains computationally expensive and time-consuming.

In many deep learning domains, recent years have seen significant progress in developing general-purpose pre-trained, or *foundation*, models trained on large-scale, diverse data, often exploiting weakly or self-supervised approaches to extract general, reusable representations. Once trained, these models can be adapted or fine-tuned as reusable building blocks for a wide range of downstream tasks, exploiting the knowledge extracted from the original training data. Notably, the rapid advances in language-related tasks, particularly through large language models (LLMs), are largely attributable to the massive scaling of both training data and model capacity [KMH*20, BMR*20, NKQ*25]. Because images represent a higher-dimensional, noisier, and more redundant modality than text, the computer vision community has early on adopted large-scale unsupervised and self-supervised pre-training strategies [RKH*21], leading to widely used foundation models for many vision tasks, such as segmentation [KMR*23] and object detection [XZH*21]. Building on these developments, pre-trained vision-language models such as CLIP [CRC*20] combine both trends and have demonstrated strong zero-shot performance across a variety of downstream vision tasks, including image classification and object detection.

In the context of single-image 360° inference, pre-trained or foundation models are leveraged in two main ways: by repurposing standard computer vision models trained on conventional imaging to support specific sub-tasks and combining them to address omnidirectional problems (Sec. 7.1), and by developing dedicated 360° models that explicitly exploit the holistic and intrinsic properties of omnidirectional images or videos (Sec. 7.2).

7.1. Transfer and composition of conventional vision foundation models for 360° tasks

Leveraging standard vision foundation models for omnidirectional tasks offers a practical means to extend their applicability to more diverse environments and higher-resolution settings. Many existing models, whether designed for omnidirectional imagery or not, operate at relatively low spatial resolutions (e.g., 1024×512 pixels, approximately 0.5 MP), which are significantly below the requirements of modern virtual reality head-mounted displays (e.g., HTC Vive at 2448×2448 pixels per eye) and current omnidirectional imaging sensors, which can reach 8K resolution (8192×4096) or even 60 MP (e.g., Ricoh Theta X at 11008×5504). Simple post-hoc upsampling of holistic predictions only partially alleviates this limitation: fine geometric details, thin structures, small or distant objects, and high-frequency textures are often lost when depth or geometry is estimated at low resolution and upsampled afterward, thereby constraining downstream tasks such as depth estimation, SLAM, surface meshing, or neural rendering [ACW25]. To address these issues, composing partial inferences obtained from multiple limited–field-of-view perspective views or spherical tiles is emerging as an effective strategy for preserving spatial detail, while also addressing the limitations due to the scarcity of large-scale, high-resolution panoramic training datasets by transferring knowledge learned from much broader sets of examples.

For instance, pre-trained foundation depth models have recently led to substantial advances in zero-shot monocular depth estimation, notably improving cross-domain robustness and metric consistency (e.g., *Depth Anything* [YKH*24a, YKH*24b]). In the context of panoramic imagery, several strategies have emerged to adapt such foundation models to omnidirectional depth estimation.

A first line of work follows a *distillation or transfer* paradigm, in which depth predictions from perspective models are used to supervise networks operating directly on equirectangular panoramas [WL24b, JSL*25]. This approach alleviates the scarcity of annotated panoramic data, but the resulting models often face scalability issues, as operating directly on high-resolution panoramas remains computationally demanding on standard hardware.

A second family of methods adopts a *projection-and-fuse* strategy, applying pre-trained depth models independently to overlapping

perspective patches sampled from the panorama and subsequently reprojecting the predictions into the panoramic domain using seam-aware fusion [LGY*22, RAYR22a, SZL*23, ACC*23]. While this approach preserves high-frequency details, the independently generated depth maps are frequently inconsistent and difficult to reconcile, often requiring substantial post-processing to obtain a coherent global result [RAYR22a, SZL*23]. To improve inter-patch consistency, ST^{2}360D [CZA*25] reformulates the problem by converting panoramas into perspective video sequences and predicting depth using foundation video models [LHS*20, CGZ*25]. Temporal continuity across frames helps enforce smoother transitions between adjacent views. However, the lack of a truly holistic representation can still hinder seamless global consistency. Peng and Zhang [PZ23] further address this issue by aligning perspective-based predictions to a low-resolution panoramic depth estimate, improving global registration while remaining dependent on the quality of the individual perspective inferences. Moreover, the subsequent composition and fusion stages may still introduce discontinuities and blurring near patch boundaries.

Overall, this line of research reveals the emergence of a multi-resolution solving paradigm, in which a coarse, low-resolution full-panorama prediction may serve not only as global structural guidance for conditioning high-resolution local inferences from perspective foundation models, but also as an explicit scaffold to be exploited during the merging and fusion stages. By anchoring independent local predictions to a shared global representation, this approach would facilitate more consistent aggregation into a coherent solution and is likely to generalize to a wide range of omnidirectional inference problems. Conditioning local predictions from general limited-view foundation models, often also employing a globally constrained merging step, is, for instance, exploited by many of the view-synthesis approaches surveyed in Sec. 6.1. In particular, conditioned open diffusion models [RBL*22, GCC*24] are exploited for inpainting in novel view synthesis [KYS23, RAR25, PZL24] (Sec. 6.1).

7.2. Native 360° foundation models

Although the development of foundation models explicitly tailored to the intrinsic properties of 360° imagery is a potentially promising research direction, it remains relatively underexplored compared to foundation models designed for conventional perspective images.

On the one hand, training foundation models is computationally and economically very expensive. To provide a reference, training a frontier foundation model can nowadays exceed the hundreds of millions of Euros [MFP*25], and costs are projected to increase [CRF*25]. Models trained on massive collections of web-scale perspective images or synthetic data have already demonstrated strong performance on a range of 3D tasks that can be transferred to omnidirectional image analysis and synthesis [LZW*25b]. In particular, recent work has shown that unified, generalist generative foundation models for videos can address a wide spectrum of analysis and synthesis problems, exhibiting emerging zero-shot perception and reasoning capabilities [WLV*25]. As a result, both academic and industrial efforts have largely concentrated on leveraging and adapting such general-purpose models rather than developing omnidirectional-specific foundations from scratch.

On the other hand, despite the notable successes of models trained on synthetic imagery or conventional videos, the scarcity of large-scale, high-quality real-world 3D data remains a fundamental challenge for 360 and 3D tasks. Synthetic data can provide dense and accurate annotations for supervised learning, but its scale and diversity are inherently limited. Approaches such as PanDA [CZZ*25], that aim to construct a foundation 360 variant of Depth Anything, tackle the challenge with a weakly supervised approach. First, a teacher model is obtained by fine-tuning Depth Anything [YKH*24a] on synthetic indoor and outdoor panoramic data. Then, a student model is trained on large-scale unlabeled data using teacher-generated pseudo-labels. DA360 [JSL*25], instead, fine-tunes the DAv2 zero-shot single pinhole depth estimation model [YKH*24b] with synthetic panorama depth datasets to produce scale-invariant globally consistent panoramic depth maps. The key innovation involves learning a shift parameter from the vision transform backbone, transforming the model's scale- and shift-invariant output into a scale-invariant estimate that directly yields well-formed 3D point clouds.

Videos constitute a particularly promising alternative, as structural cues can be inferred from inter-frame correspondences, enabling self-supervised learning. However, acquiring diverse yet corresponding views of the same scene content at scale is difficult, and standard videos are captured from fixed viewpoints, restricting access to a broader range of informative perspectives. Large-scale 360° videos naturally alleviate these limitations by enabling scalable extraction of corresponding views from multiple, diverse viewpoints. Initial efforts in this direction are emerging. For example, Wallingford et al. [WBK*24] trained a diffusion-based model on one million 360° YouTube videos and demonstrated applications to novel view synthesis as well as geometry and layout extraction. Further progress in this direction would be beneficial not only for video generation but for all tasks requiring holistic understanding.

7.3. Opportunities and limitations

Overall, foundation models are emerging as a powerful paradigm for single-view 360° inference, as they enable the reuse of knowledge distilled from large and diverse datasets, improved cross-domain generalization, and, in some cases, zero-shot or few-shot capabilities in settings with limited panoramic supervision. This trend is currently more prominent in adapting general vision foundation models to omnidirectional tasks than in creating foundation models specifically designed for 360° imagery. However, the development and deployment of foundation models incur substantial costs in terms of data, computation, and energy, placing them beyond the practical reach of many researchers and raising concerns about transparency, direction, and the concentration of power [MFP*25]. Open models may offer only a partial solution [BKK*24]. Furthermore, for many application scenarios, task-specific lightweight models remain significantly more efficient to train, execute, and deploy, particularly on constrained hardware. For these reasons, we anticipate continued research on integrating foundation models as reusable building blocks for omnidirectional problems and on devising lean, task-specific solutions that prioritize efficiency and accessibility.

Domain	Typical Tasks	Pipeline pieces to enable	Key KPIs / Constraints
AEC/BIM & Facility Mgmt	As-built documentation; floorplan bootstrapping; deviation/issue views; lighting "what-if"	Depth/normals (Sec. 5.1); layout & portals/graphs (Sec. 5.2, Sec. 6.1.3); short-baseline NVS (Sec. 6.1); inverse rendering for relighting/staging (Sec. 6.2)	Layout IoU/corner error; door/portal F1; room area/offset/length/elevation errors; time-to-answer; photometric plausibility
Real estate / Interior / Retail	Defurnishing (emptying); virtual staging; style harmonization; guided tours (web/HMD)	Diminished reality; staging + relighting (inverse rendering); panorama-aware style transfer (Sec. 6.2); short-baseline NVS for parallax	Engagement/session length; preference scores; shadow/contact realism; decision latency; web load time
Cultural heritage & Museums	Restricted-space tours; thematic overlays; alternative epochs/looks; gap-filling when capture is limited	Layout/semantic priors (Sec. 5.2); controlled NVS (Sec. 6.1); style transfer; text-to-360 for missing areas)	Curatorial ratings; visitor comprehension; provenance tracking; policy constraints (no tripod/flash; limited dwell)
Construction education, safety training & upskilling	Multiuser panoramic site visits; hazard drills; instrument callouts; scenario variation (day/night)	Semantics for callouts (Sec. 5.1); NVS in HMD (Sec. 6.1); room graphs for procedure flow (Sec. 6.1.3); diffusion-guided relighting (Sec. 6.2)	Task time/accuracy; presence/immersion; SSQ; pre/post learning gains
Progress monitoring, quality control & robotics	Waypoint capture (human/robot); coverage and change detection; BIM-aware patrols; coarse point clouds	Depth/layout for coverage/change; semantics for route planning; portals/graphs for pathing; short-baseline NVS for line-of-sight; normalization via inverse-rendering relighting	Capture time/cost; detection precision/recall; inter-rater agreement; robot path success; repeatability
Collaboration, remote walkthroughs & RFIs	Remote inspections; RFIs with anchored views; design reviews; "as-designed" vs "as-built" alternates	Context-aware annotations snapped to planes/assets (semantics + layout); guided tours via room graphs; on-the-fly emptying/staging (Sec. 6.2)	RFI turnaround; comment resolution rate; reviewer agreement; navigation clarity; bandwidth/latency
Content creation & datasets	Domain randomization; supervised views from one shot; synthetic scenes for training	Text-to-360 generation; staging/transfer (Sec. 6.2); short-baseline NVS for multi-view supervision	Downstream model gains; domain gap; diversity/coverage; licensing & provenance

Table 11: ***Application matrix according to the current usage of panoramic imagery****: what users do, which single-panorama pipeline components enable it, and how success can be measured.*

8. Applications

Panorama-first workflows turn a single capture per room into structured, explorable, and editable indoor scenes with minimal user effort. Building on pixel-wise geometry and semantics (Sec. 5.1), room layout recovery (Sec. 5.2), multi-room connectivity (Sec. 5.2), interactive view synthesis (Sec. 6.1), and editing/generation tools (Sec. 6.2), we outline where these components already see practical use. A common characteristic of these applications is the need for *plausible*, rather than *accurate/measurable*, models, where ease of capture, exploration, and modification of approximate models is at a premium. For a broader discussion of potential applications of immersive panoramic content, not limited to single-image pipelines, we refer the reader to the recent survey from Tukur et al. [TJA*25].

AEC/BIM, facility management, and digital delivery While measurable models are typically acquired with active measurement devices, on active projects, sparse panoramic capture is routinely used for as-built documentation and progress reports, often aligned with BIM models for issue tracking and deviation visualization [SLK*23]. Single-image 360° layout reconstruction has found applications in construction management, ranging from automatic progress assessment of interior sites [FLW*23] to detecting functional elements such as lights and outlets [PPG*18].

Construction education, safety training, and workforce upskilling In classrooms and training centers, 360° imagery has become a practical substitute for costly synthetic VR scenes: it is fast to capture, easy to deploy on browsers or commodity HMDs, and preserves job-site realism [KLL19, EGE20]. Multiuser panoramic visits support collaborative problem solving with shared pointers and voice [EWG22], while recent case studies show that immersive storytelling in 360° increases engagement for safety training [IEAB24].

Cultural heritage, museums, and public dissemination Heritage institutions adopt web-based thematic tours built on 360° imagery to balance accessibility, curatorial storytelling, and management needs, often interlinking panoramas with maps, metadata, and archival documents [DFBF22]. For fragile or conflict-affected sites, interactive 360° media has been used to document and communicate cultural value with minimal on-site footprint [DAW24].

Progress monitoring, quality control, and robotics Routine progress capture often follows predefined waypoints with panoramic imagery. Comparative pilots show that panoramic photogrammetry provides a favorable time–cost trade-off for coarse point clouds, while laser scanning remains the choice for tight tolerances [SG19b]. However, single-panorama analysis is emerging as a viable medium to support object detection and tracking, segmentation, change understanding, and change reversal [KQC*25].

Collaboration, remote walkthroughs, and telepresence Teams rely on multiuser 360° walkthroughs to conduct remote inspections, Requests for Information (RFIs), and design reviews without dense 3D models. Empirical work reports a higher presence and attention

relative to 2D baselines and, in some cases, comparable learning outcomes to synthetic VR [KLL19, EGE20, EWG22]. The reconstruction signals surveyed here make these sessions more actionable: annotations can snap to planes or assets via semantics and layout; room graphs constrain guided tours to safe, comprehensible paths; and on-the-fly staging or emptying can present "as-designed" versus "as-built" alternatives without another site visit.

Real estate, interior design, and retail Virtual tours built from 360° panoramas are now mainstream for property marketing and in-store/showroom experiences. Beyond basic retouching, practitioners increasingly combine defurnishing, virtual staging, and style harmonization to accelerate client decisions. The editing stack surveyed in Sec. 6.2 – diminished reality to declutter (e.g., [PAAG22]), inverse-rendered insertion for contacts and shadows [ZCB*22, LWH*22b, SJT*25], and panorama-aware style transfer [TUP*23, YWL*25, WQTX25] – supports rapid, low-touch iterations that read well on the web and in HMDs.

Generative 360° and hybrid capture Text- or image-conditioned 360° generation [ZFX*25, XWY*25, YTZ*25] and diffusion-guided relighting/editing (Sec. 6.2) are beginning to complement capture. In practice, teams adopt hybrid flows: generate an empty, layout-conforming shell to explore options, then conform the chosen look to a real panorama; or propose lighting and material variants via diffusion models before committing to staged assets. The reconstruction signals from earlier sections act as constraints, keeping the generative steps structurally coherent and editable.

Operational considerations and evaluation Across domains, privacy and safety require default blurring of faces/PII and careful handling of mirrors and glass. Web delivery benefits from progressive, ERP-aware streaming that balances resolution and bandwidth. For AEC and monitoring, useful structural metrics include layout IoU, corner error, door/portal F1, and room-wise area/offset/length/elevation errors reported in recent panoramic-SLAM studies [WLX*25]; for education and telepresence, presence/immersion, pre/post learning gains, and scenario completion times are standard [KLL19, EGE20, EWG22, IEAB24]; for staging and retail, practitioner studies emphasize contact/shadow realism, preference tests, and decision latency.

9. Research trends and open challenges

As highlighted in the preceding sections, deep learning has driven major progress across both analysis and synthesis, and methods are reaching real-world applications. Nevertheless, many challenges remain open. Discussed throughout the survey and reflected in the cited literature, they outline the current frontier of the field. Below, we summarize several key trends and future research directions.

Training data limitations While enormous amounts of image data are available, available 360° imagery of interiors is much less abundant. Moreover, even though millions of indoor panoramas may exist in the wild and could, in theory, be harvested, curated research datasets remain small, less diverse, and costly to annotate. Synthetic datasets help expand scale and provide ground truth, but their diversity is still ultimately constrained by available 3D sources, which are currently limited and costly to generate. Key challenges include improving training data size and diversity through better annotation of real sources and faster generation of synthetic models, as well as, and most importantly, developing methods that work better in low-data regimes, exploiting labeled and unlabeled panoramas through semi- or self-supervised learning. An interesting related research direction is to use the more abundant regular perspective images for training 360° models through a combination of on-the-fly conversion to omnidirectional formats, augmentation, and masking of valid pixels. This approach has shown good potential for depth estimation [GGM*25, LZH*26], and may be extended to other tasks.

Annotation accuracy on real-world data In addition to its scarcity, real-world panoramic image annotation suffers from several limitations. While ground truth depth can be measured through instruments (i.e., depth sensors), room layouts and other structures generally come from human annotation and rely on per-dataset representation. It is not uncommon to have a mismatch between annotations and image features, and there is no universal way to use annotations from multiple datasets. Several efforts have strived to define and apply annotation workflows in different contexts (e.g., [CHL*21a, DWB*24b]).

Generalization challenges Due to a variety of reasons, including the scarcity of datasets, many of the solutions surveyed in this report appear to be tested and trained using a single (or coherent) dataset, and leave out testing across domains, scene types, and capture conditions. In real cases, especially when scarce or synthetic data is used for training, data seen at inference time may often be very different from the one in the training set. Recent datasets have started to appear with training and testing sets coming from very different distributions (e.g., *360MonoDepth* [RAYR22a] and Pano3D [AZD*21b], that mix synthetic and real data for cross-generalization purposes). While indoor pipelines do not require full open-domain generalization, enlarging the variety of supported interiors is an important goal for future research.

Resolution Most of the solutions presented for analysis (e.g., depth inference) and synthesis (e.g., renderable model generation) still work at relatively low resolutions (e.g., 512×1024). This is due both to the lack of high-resolution training data available and to the complexity limitations imposed on the networks, especially when they need to work at high frequency on embedded devices. Current cameras and HMDs, however, have a significantly higher resolution (i.e., 4K), indicating a space for improvement of current methods.

From single-task learning to holistic models Very significant progress has been made on each of the individual low-level tasks, such as depth, layout, and semantics inference and object detection. At the same time, there is a clear trend to address them jointly, moving towards holistic scene understanding rather than combining individual task solutions. How to perform this combination is, however, still an open problem, both in terms of efficient network design (i.e., common or parallel branches, which architecture, ...) and loss combination (since losses may have different scales and conflicting gradients).

Cooperation between 360° and regular image processing The wide context available in 360° images is essential for the under-

standing of indoor environments, with 360° solutions for depth and layout inference surpassing traditional constrained-FOV solutions. At the same time, many pre-trained solution exists to solve specific tasks using regular FOV images, from object detection to diffusion-based image generation. In the last few years, several combinations have been attempted, from diffusion-based floor plan generation to the conditional creation of pseudo-novel views for view synthesis. Given the increasing availability of such powerful models and the great cost for training them, we expect that the trend of combining panoramic and restricted-view processing pipelines will continue.

360° foundation models foundation models are being created for regular images and videos. As for what happened with large language models (LLMs), research has shown that large, generative models trained on web-scale data have characteristics that allow them to be used not as task-specific models, but as unified, generalist foundation models that can solve a variety of analysis and synthesis tasks [WLV*25]. Some initial works are in progress for 360° imagery (Sec. 7), and further progress in this direction would be beneficial for all tasks requiring holistic views. Training and inference efficiency would be a concern, especially at high resolutions. Current works seem to be limited to relatively low resolution. The current prominence of very few industrial actors raises concerns about transparency, direction, and the concentration of power [MFP*25], and open models need coordination efforts [BKK*24].

Balancing specialization and general-purpose solutions Large, pre-trained generic models are enabling a wide range of applications, but they are costly, energy-intensive, and demanding to both train and deploy. As their size and generality grow, such models are increasingly feasible only through substantial investments, typically accessible only to national organizations or corporations with massive computational resources. For tasks requiring broad reasoning or the generation of entirely novel data, these models may be indispensable. However, many indoor analysis and synthesis tasks operate within a narrower solution space, where leaner, task-specific models remain viable. Developing efficient, specialized approaches, ideally lightweight enough to run on embedded devices, and finding the right balance between pre-training and specialization are therefore important research goals.

Uncertainty and explainability All machine learning–based solutions discussed here produce plausible rather than strictly accurate models, inferring missing information from training data priors. As a result, assessing the uncertainty of predictions—whether depth, layout, or novel views—and explaining the rationale behind a given output, including possible alternatives, remains challenging. Recent work has begun to address these issues in the context of generative and neural scene representations (see, e.g., [KMKS24, AVM25]). Extending these approaches to 360° indoor imagery is promising but introduces new challenges in uncertainty visualization, such as identifying multiple plausible multi-room layouts, linking them to supporting evidence, and visualizing the alternative hypotheses.

10. Conclusion

Surround-view panoramic imaging enables rapid and comprehensive scene capture and is increasingly supported by both consumer and professional devices. In this work, we have reviewed recent advances in deep learning methods for transforming indoor panoramas into representations that support structural and geometric recovery, dynamic exploration, and editing, with particular emphasis on single-image approaches, a use case having vast practical importance and posing great visual computing challenges.

Since the field is evolving at a rapid pace, a complete survey is beyond reach, and our goal has been to organize current and emerging directions, thereby providing reasoned and accessible entry points for both researchers and practitioners. The report highlights that significant progress has already been achieved across analysis and synthesis, with many solutions transitioning into real-world applications. Nonetheless, substantial challenges remain, and we hope this overview will serve as a foundation and catalyst for further innovation in panoramic visual computing.

Acknowledgments This publication was supported by NPRP-S 14th Cycle grant 0403-210132 AIN2 from the Qatar National Research Fund (a member of Qatar Foundation). GP and EG also acknowledge support from Sardinian Regional Authorities under the XDATA project (Art9 LR 20/2015). The findings herein reflect the work and are solely the responsibility of the authors. ChatGPT was used to improve the language, grammar, and flow of the manuscript.

References

[ACC*23] AI H., CAO Z., CAO Y.-P., SHAN Y., WANG L.: HRDFuse: Monocular 360° depth estimation by collaboratively learning holistic-with-regional depth distributions. In *Proc. CVPR* (2023), pp. 13273–13282. doi:10.1109/CVPR52729.2023.01275. 9, 10, 25

[ACW25] AI H., CAO Z., WANG L.: A survey of representation learning, optimization strategies, and applications for omnidirectional vision. *Int J Comput Vis 133* (2025), 4973–5012. doi:10.1007/s11263-025-02391-w. 3, 24

[AHG*19a] ARMENI I., HE Z.-Y., GWAK J., ZAMIR A. R., FISCHER M., MALIK J., SAVARESE S.: 3D Scene Graph: A structure for unified semantics, 3D space, and camera. In *Proc. CVPR* (2019), pp. 5664–5673. doi:10.1109/ICCV.2019.00576. 6

[AHG*19b] ARMENI I., HE Z.-Y., GWAK J., ZAMIR A. R., FISCHER M., MALIK J., SAVARESE S.: The 3DSceneGraph Dataset. Public dataset, 2019. [Online; accessed 2025-09-17]. URL: https://github.com/StanfordVL/3DSceneGraph/. 6

[ALB21] ALZAYER H., LIN H., BALA K.: Autophoto: Aesthetic photo capture using reinforcement learning. In *Proc. IROS* (2021), pp. 944–951. doi:10.1109/IROS51168.2021.9636788. 23

[ALG*20] ATTAL B., LING S., GOKASLAN A., RICHARDT C., TOMPKIN J.: MatryODShka: Real-time 6DoF video view synthesis using multi-sphere images. In *Proc. ECCV* (2020), pp. 441–459. doi:10.1007/978-3-030-58452-8_26. 6, 19, 20

[ASZ*16] ARMENI I., SENER O., ZAMIR A. R., JIANG H., BRILAKIS I., FISCHER M., SAVARESE S.: 3D semantic parsing of large-scale indoor spaces. In *Proc. CVPR* (2016), pp. 1534–1543. doi:10.1109/CVPR.2016.170. 6

[AVM25] AIRA L. S., VALSESIA D., MAGLI E.: Modeling uncertainty for gaussian splatting. *IEEE Transactions on Neural Networks and Learning Systems 36*, 6 (2025), 11657–11663. doi:10.1109/TNNLS.2025.3553582. 28

[AW24] AI H., WANG L.: Elite360D: Towards efficient 360 depth estimation via semantic-and distance-aware bi-projection fusion. In *Proc. CVPR* (2024), pp. 9926–9935. doi:10.1109/CVPR52733.2024.00947. 9, 10

[AZD*21a] ALBANIS G., ZIOULIS N., DRAKOULIS P., GKITSAS V., STERZENTSENKO V., ALVAREZ F., ZARPALAS D., DARAS P.: Pano3D. Public dataset, 2021. [Online; accessed 2025-09-17]. URL: https://vcl3d.github.io/Pano3D/download/. 6, 7

[AZD*21b] ALBANIS G., ZIOULIS N., DRAKOULIS P., GKITSAS V., STERZENTSENKO V., ALVAREZ F., ZARPALAS D., DARAS P.: Pano3D: A holistic benchmark and a solid baseline for 360° depth estimation. In *Proc. CVPR Workshops* (2021), pp. 3727–3737. doi:10.1109/CVPRW53098.2021.00413. 7, 27

[BCR19] BERTEL T., CAMPBELL N. D., RICHARDT C.: MegaParallax: Casual 360° panoramas with motion parallax. *IEEE TVCG 25*, 5 (2019), 1828–1835. doi:10.1109/TVCG.2019.2898799. 19

[BFO*20] BROXTON M., FLYNN J., OVERBECK R., ERICKSON D., HEDMAN P., DUVALL M., DOURGARIAN J., BUSCH J., WHALEN M., DEBEVEC P.: Immersive light field video with a layered mesh representation. *ACM TOG 39*, 4 (2020), 86:1–86:15. doi:10.1145/3386569.3392485. 19

[BHG*25] BAI J., HUANG L., GUO J., GONG W., LI Y., GUO Y.: 360-GS: Layout-guided panoramic gaussian splatting for indoor roaming. In *Proc. 3DV* (2025), pp. 1042–1053. doi:10.1109/3DV66043.2025.00101. 19

[BKH*23] BOORBOOR S., KIM Y., HU P., MOSES J. M., COLLE B. A., KAUFMAN A. E.: Submerse: Visualizing storm surge flooding simulations in immersive display ecologies. *IEEE TVCG 30*, 9 (2023), 6365–6377. doi:10.1109/TVCG.2023.3332511. 23

[BKK*24] BOMMASANI R., KAPOOR S., KLYMAN K., LONGPRE S., RAMASWAMI A., ZHANG D., SCHAAKE M., HO D. E., NARAYANAN A., LIANG P.: Considerations for governing open foundation models. *Science 386*, 6718 (2024), 151–153. doi:10.1126/science.adp1848. 25, 28

[BKL*25] BAGDASARIAN M. T., KNOLL P., LI Y., BARTHEL F., HILSMANN A., EISERT P., MORGENSTERN W.: 3DGS.zip: A survey on 3d gaussian splatting compression methods. *Computer Graphics Forum 44*, 2 (2025), e70078. doi:10.1111/cgf.70078. 19

[BMR*20] BROWN T. B., MANN B., RYDER N., SUBBIAH M., KAPLAN J., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A., AGARWAL S., HERBERT-VOSS A., KRUEGER G., HENIGHAN T., CHILD R., RAMESH A., ZIEGLER D. M., WU J., WINTER C., HESSE C., CHEN M., SIGLER E., LITWIN M., GRAY S., CHESS B., CLARK J., BERNER C., MCCANDLISH S., RADFORD A., SUTSKEVER I., AMODEI D.: Language models are few-shot learners. In *Proc. NeurIPS* (2020), vol. 33, pp. 1877–1901. doi:10.48550/arXiv.2005.14165. 24

[Bou10] BOURKE P.: Capturing omni-directional stereoscopic spherical projections with a single camera. In *Proc. IEEE VSMM* (2010), pp. 179–183. doi:10.1109/VSMM.2010.5665988. 20

[BTS*17] BERGER M., TAGLIASACCHI A., SEVERSKY L. M., ALLIEZ P., GUENNEBAUD G., LEVINE J. A., SHARF A., SILVA C. T.: A survey of surface reconstruction from point clouds. *Computer Graphics Forum 36*, 1 (2017), 301–329. doi:10.1111/cgf.12802. 4

[BYLR20] BERTEL T., YUAN M., LINDROOS R., RICHARDT C.: OmniPhotos: Casual 360° VR photography. *ACM TOG 39*, 6 (2020), 266:1–266:12. doi:10.1145/3414685.3417770. 19

[BZG*22] BOGUSLAWSKI P., ZLATANOVA S., GOTLIB D., WYSZOMIRSKI M., GNAT M., GRZEMPOWSKI P.: 3D building interior modelling for navigation in emergency response applications. *nt. J. Appl. Earth Obs. Geoinf 114* (2022), 103066. doi:10.1016/j.jag.2022.103066. 1

[CBLR18] CHEN Z., BADRINARAYANAN V., LEE C.-Y., RABINOVICH A.: GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proc. ICML* (2018), PMLR, pp. 794–803. doi:10.48550/arXiv.1711.02257. 15

[CCD*18] CHENG H., CHAO C., DONG J., WEN H., LIU T., SUN M.: Cube padding for weakly-supervised saliency prediction in 360 videos. In *Proc. CVPR* (2018), pp. 1420–1429. doi:10.1109/CVPR.2018.00154. 8, 9

[CCG*23] CHEN Z., CAO Y.-P., GUO Y.-C., WANG C., SHAN Y., ZHANG S.-H.: PanoGRF: generalizable spherical radiance fields for wide-baseline panoramas. In *Proc. NeurIPS* (2023). doi:10.48550/arXiv.2306.01531. 22

[CDF*17] CHANG A., DAI A., FUNKHOUSER T., HALBER M., NIESSNER M., SAVVA M., SONG S., ZENG A., ZHANG Y.: Matterport3D: Learning from RGB-D data in indoor environments. In *Proc. 3DV* (2017), pp. 667–676. doi:10.1109/3DV.2017.00081. 5

[CDF24] CHEN J., DENG R., FURUKAWA Y.: Polydiffuse: Polygonal shape reconstruction via guided set diffusion models. *NeurIPS 36* (2024), 1863–1888. URL: https://dl.acm.org/doi/abs/10.5555/3666122.3666212. 17

[CDM17] CHAMPEL M.-L., DORÉ R., MOLLET N.: Key factors for a high-quality VR experience. In *Applications of Digital Image Processing XL* (2017), vol. 10396, SPIE, pp. 183–194. doi:10.1117/12.2274336. 4

[CF14] CABRAL R., FURUKAWA Y.: Piecewise planar and compact floorplan reconstruction from images. In *Proc. CVPR* (2014), pp. 628–635. doi:10.1109/CVPR.2014.546. 4, 16

[CFG*15] CHANG A. X., FUNKHOUSER T., GUIBAS L., HANRAHAN P., HUANG Q., LI Z., SAVARESE S., SAVVA M., SONG S., SU H., XIAO J., YI L., YU F.: ShapeNet: An information-rich 3D model repository. In *arXiv preprint arXiv:1512.03012* (2015). doi:10.48550/arXiv.1512.03012. 14

[CGZ*25] CHEN S., GUO H., ZHU S., ZHANG F., HUANG Z., FENG J., KANG B.: Video Depth Anything: Consistent depth estimation for super-long videos. In *Proc. CVPR* (2025), pp. 22831–22840. doi:10.1109/CVPR52734.2025.02126. 25

[CHL*21a] CRUZ S., HUTCHCROFT W., LI Y., KHOSRAVAN N., BOYADZHIEV I., KANG S. B.: Zillow indoor dataset: Annotated floor plans with 360° panoramas and 3D room layouts. In *Proc. CVPR* (2021), pp. 2133–2143. doi:10.1109/CVPR46437.2021.00217. 1, 6, 7, 27

[CHL*21b] CRUZ S., HUTCHCROFT W., LI Y., KHOSRAVAN N., BOYADZHIEV I., KANG S. B.: Zillow Indoor Dataset (ZInD). Public dataset, 2021. [Online; accessed: 2025-09-16]. URL: https://github.com/zillow/zind. 6

[CLTS24] CHARATAN D., LI S. L., TAGLIASACCHI A., SITZMANN V.: PixelSplat: 3D Gaussian Splats from image pairs for scalable generalizable 3D reconstruction. In *Proc. CVPR* (2024), pp. 19457–19467. doi:10.1109/CVPR52733.2024.01840. 20

[CLWF19] CHEN J., LIU C., WU J., FURUKAWA Y.: Floor-SP: Inverse CAD for floorplans by sequential room-wise shortest path. In *Proc. CVPR* (2019), pp. 2661–2670. doi:10.1109/ICCV.2019.00275. 16

[CODVG22] CAI S., OBUKHOV A., DAI D., VAN GOOL L.: Pix2nerf: Unsupervised conditional P-GAN for single image to neural radiance fields translation. In *Proc. CVPR* (2022), pp. 3981–3990. doi:10.1109/CVPR52688.2022.00395. 20

[CQF22] CHEN J., QIAN Y., FURUKAWA Y.: HEAT: Holistic edge attention transformer for structured reconstruction. In *Proc. CVPR* (2022), pp. 3866–3875. doi:10.1109/CVPR52688.2022.00384. 16

[CRC*20] CHEN M., RADFORD A., CHILD R., WU J., JUN H., LUAN D., SUTSKEVER I.: Generative pretraining from pixels. In *Proc. PMLR* (2020), pp. 1691–1703. URL: https://proceedings.mlr.press/v119/chen20s.html. 24

[CRF*25] COTTIER B., RAHMAN R., FATTORINI L., MASLEJ N., BESIROGLU T., OWEN D.: The rising costs of training frontier ai models. *arXiv preprint arXiv:2405.21015* (2025). doi:10.48550/arXiv.2405.21015. 25

[CWS18] CAO Y., WU Z., SHEN C.: Estimating depth from monocular images as classification using deep fully convolutional residual networks. *IEEE TCSVT 28*, 11 (2018), 3174–3182. doi:10.1109/TCSVT.2017.2740321. 8

[CY99] COUGHLAN J. M., YUILLE A. L.: Manhattan World: Compass direction from a single image by Bayesian inference. In *Proc. ICCV* (1999), vol. 2, pp. 941–947. doi:10.1109/ICCV.1999.790349. 5

[CZA*25] CAO Z., ZHU J., AI H., JIANG L., LYU Y., XIONG H.: ST^{2}360D: Spatial-to-temporal consistency for training-free 360 monocular depth estimation. In *Proc. NeurIPS* (2025). URL: https://neurips.cc/virtual/2025/loc/san-diego/poster/117039. 25

[CZZ*25] CAO Z., ZHU J., ZHANG W., AI H., BAI H., ZHAO H., WANG L.: PanDA: Towards panoramic Depth Anything with unlabeled panoramas and mobius spatial augmentation. In *Proc. CVPR* (2025), pp. 982–992. doi:10.1109/CVPR52734.2025.00100. 9, 10, 25

[CZZZ22] CHEN X., ZHAO H., ZHOU G., ZHANG Y.-Q.: PQ-Transformer: Jointly parsing 3D objects and layouts from point clouds. In *Proc. CVPR* (2022). doi:10.1109/LRA.2022.3143224. 14

[DAH20] DAVIDSON B., ALVI M. S., HENRIQUES J. F. H.: 360 camera alignment via segmentation. In *Proc. ECCV* (2020), pp. 579–595. doi:10.1007/978-3-030-58604-1_35. 4

[DAW24] DIEB R., ALSALLOUM A., WEBB N.: Interactive 360° media for the dissemination of endangered world heritage sites: the ancient city of palmyra in syria. *Built Heritage 8*, 1 (2024), 18. doi:10.1186/s43238-024-00126-3. 26

[DBGBR*14] DI BENEDETTO M., GANOVELLI F., BALSA RODRIGUEZ M., JASPE VILLANUEVA A., SCOPIGNO R., GOBBETTI E.: ExploreMaps: Efficient construction and ubiquitous exploration of panoramic view graphs of complex 3D environments. *Comput. Graph. Forum 33*, 2 (2014), 459–468. doi:10.1111/cgf.12334. 22

[DBK*20] DOSOVITSKIY A., BEYER L., KOLESNIKOV A., WEISSENBORN D., ZHAI X., UNTERTHINER T., DEHGHANI M., MINDERER M., HEIGOLD G., GELLY S., USZKOREIT J., HOULSBY N.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020). doi:10.48550/arXiv.2010.11929. 14

[DFB*24] DONG Y., FANG C., BO L., DONG Z., TAN P.: PanoContext-Former: Panoramic total scene understanding with a transformer. In *Proc. CVPR* (2024), pp. 28087–28097. doi:10.1109/CVPR52733.2024.02653. 11, 12, 14, 16, 17, 20

[DFBF22] DE FINO M., BRUNO S., FATIGUSO F.: Dissemination, assessment and management of historic buildings by thematic virtual tours and 3D models. *Virtual Archaeology Review 13*, 26 (2022), 88–102. doi:10.4995/var.2022.15426. 26

[DFL23] DAI Y., FEI N., LU Z.: Improvable gap balancing for multi-task learning. In *Uncertainty in Artificial Intelligence* (2023), PMLR, pp. 496–506. doi:10.48550/arXiv.2307.15429. 15

[DHN06] DELAGE E., HONGLAK LEE, NG A. Y.: A dynamic Bayesian network model for autonomous 3D reconstruction from a single indoor image. In *Proc. CVPR* (2006), vol. 2, pp. 2418–2428. doi:10.1109/CVPR.2006.23. 5

[DHX*23] DONG M., HUAN L., XIONG H., SHEN S., ZHENG X.: Shape anchor guided holistic indoor scene understanding. In *Proc. ICCV* (2023), pp. 21916–21926. doi:10.1109/ICCV51070.2023.02003. 14

[DJQ*23] DENG C., JIANG C., QI C. R., YAN X., ZHOU Y., GUIBAS L., ANGUELOV D.: NeRDI: Single-view NeRF synthesis with language-guided diffusion as general image priors. In *Proc. CVPR* (2023), pp. 20637–20647. doi:10.1109/CVPR52729.2023.01977. 20

[DL22] DONG H., LEE J. S. A.: The metaverse from a multimedia communications perspective. *IEEE MultiMedia 29*, 4 (2022), 123–127. doi:10.1109/MMUL.2022.3217627. 2

[DMG*25] DOLHASZ A., MA C., GAUSEBECK D., CHEN K., MILLER G., HAYNE L., HOVDEN G., SABIK A., BRANDT O., SLAVCHEVA M.: Defurnishing with X-Ray vision: Joint removal of furniture from panoramas and mesh. In *Proc. CVPR* (2025), pp. 6280–6290. doi:10.1109/CVPRW67362.2025.00624. 23

[dSPMLJ22] DA SILVEIRA T. L., PINTO P. G., MURRUGARRA-LLERENA J., JUNG C. R.: 3D scene geometry estimation from 360° imagery: A survey. *ACM Computing Surveys 55*, 4 (2022), 1–39. doi:10.1145/3519021. 2, 3, 4

[DWB*24a] DEB T., WANG L., BESSINGER Z., KHOSRAVAN N., PENNER E., KANG S. B.: ZInD-Tell Dataset. Public dataset, 2024. [Online; accessed: 2025-09-18]. URL: https://github.com/zillow/zindtell. 6

[DWB*24b] DEB T., WANG L., BESSINGER Z., KHOSRAVAN N., PENNER E., KANG S. B.: ZInD-Tell: Towards translating indoor panoramas into descriptions. In *Proc. CVPR Workshops* (2024), pp. 2050–2059. doi:10.1109/CVPRW63382.2024.00210. 6, 27

[EF15] EIGEN D., FERGUS R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proc. ICCV* (2015), pp. 2650–2658. doi:10.1109/ICCV.2015.304. 8

[EGE20] EIRIS R., GHEISARI M., ESMAEILI B.: Desktop-based safety training using 360-degree panorama and static virtual reality techniques: A comparative experimental study. *Automation in construction 109* (2020), 102969. doi:10.1016/j.autcon.2019.102969. 26, 27

[EPF14] EIGEN D., PUHRSCH C., FERGUS R.: Depth map prediction from a single image using a multi-scale deep network. In *Proc. NeurIPS* (2014), pp. 2366–2374. URL: https://dl.acm.org/doi/abs/10.5555/2969033.2969091. 8, 10

[ESLF20] EDER M., SHVETS M., LIM J., FRAHM J.-M.: Tangent images for mitigating spherical distortion. In *Proc. CVPR* (2020). doi:10.1109/CVPR42600.2020.01244. 4

[EWG22] EIRIS R., WEN J., GHEISARI M.: ivisit-collaborate: Collaborative problem-solving in multiuser 360-degree panoramic site visits. *Computers & Education 177* (2022), 104365. doi:10.1016/j.compedu.2021.104365. 26, 27

[FCSS09] FURUKAWA Y., CURLESS B., SEITZ S. M., SZELISKI R.: Reconstructing building interiors from images. In *Proc. ICCV* (2009), pp. 80–87. doi:10.1109/ICCV.2009.5459145. 5, 16

[FDL*25] FANG C., DONG Y., LUO K., HU X., SHRESTHA R., TAN P.: Ctrl-Room: Controllable Text-to-3D room meshes generation with layout constraints. In *Proc. 3DV* (2025), pp. 692–701. doi:10.1109/3DV66043.2025.00069. 24

[FGW*18] FU H., GONG M., WANG C., BATMANGHELICH K., TAO D.: Deep ordinal regression network for monocular depth estimation. In *Proc. CVPR* (2018). doi:10.1109/CVPR.2018.00214. 8

[Fir16] FIRMAN M.: RGBD datasets: Past, present and future. In *Proc. CVPR* (2016), pp. 19–31. doi:10.1109/CVPRW.2016.88. 4

[FLW*23] FANG X., LI H., WU H., FAN L., KONG T., WU Y.: A fast end-to-end method for automatic interior progress evaluation using panoramic images. *Engineering Applications of Artificial Intelligence 126* (2023), 106733. doi:10.1016/j.engappai.2023.106733. 26

[FSG09] FEIXAS M., SBERT M., GONZÁLEZ F.: A unified information-theoretic framework for viewpoint selection and mesh saliency. *ACM Trans. Appl. Percept. 6*, 1 (2009), 1:1–1:23. doi:10.1145/1462055.1462056. 23

[GCC*24] GOKASLAN A., COOPER A. F., COLLINS J., SEGUIN L., JACOBSON A., PATEL M., FRANKLE J., STEPHENSON C., KULESHOV V.: CommonCanvas: Open diffusion models trained on creative-commons images. In *Proc. CVPR* (2024), pp. 8250–8260. doi:10.1109/CVPR52733.2024.00788. 21, 25

[GGM*25] GUO Y., GARG S., MIANGOLEH S. M. H., HUANG X., REN L.: Depth Any Camera: Zero-shot metric depth estimation from any camera. In *Proc. CVPR* (2025), pp. 26996–27006. doi:10.1109/CVPR52734.2025.02514. 27

[GMB17] GODARD C., MAC AODHA O., BROSTOW G. J.: Unsupervised monocular depth estimation with left-right consistency. In *Proc. CVPR* (2017). doi:10.1109/CVPR.2017.699. 4

[GPA24] GOBBETTI E., PINTORE G., AGUS M.: Automatic 3D modeling and exploration of indoor structures from panoramic imagery. In *SIGGRAPH Asia Courses* (2024), pp. 1:1–1:9. doi:10.1145/3680532.3689580. 1, 2

[GSZ*21] GKITSAS V., STERZENTSENKO V., ZIOULIS N., ALBANIS G., ZARPALAS D.: PanoDR: Spherical panorama diminished reality for indoor scenes. In *Proc. CVPR Workshops* (2021), pp. 3716–3726. doi:10.1109/CVPRW53098.2021.00412. 11, 23

[GTL*23] GU J., TREVITHICK A., LIN K.-E., SUSSKIND J. M., THEOBALT C., LIU L., RAMAMOORTHI R.: NeRFDiff: Single-image view synthesis with NeRF-guided distillation from 3D-aware diffusion. In *Proc. ICML* (2023), pp. 11808–11826. URL: https://dl.acm.org/doi/abs/10.5555/3618408.3618881. 20

[GYS*22] GAO S., YANG K., SHI H., WANG K., BAI J.: Review on panoramic imaging and its applications in scene understanding. *IEEE TIM 71* (2022), 1–34. doi:10.1109/TIM.2022.3216675. 2, 3

[HASK17] HEDMAN P., ALSISAN S., SZELISKI R., KOPF J.: Casual 3D photography. *ACM Transactions on Graphics (TOG) 36*, 6 (2017), 1–15. doi:10.1145/3130800.3130828. 19

[HCCJ17] HUANG J., CHEN Z., CEYLAN D., JIN H.: 6-DOF VR videos with a single 360-camera. In *Proc. IEEE VR* (2017), pp. 37–44. doi:10.1109/VR.2017.7892229. 4, 19, 20

[HCZY22] HUANG H., CHEN Y., ZHANG T., YEUNG S.-K.: 360Roam: Real-time indoor roaming using geometry-aware 360° radiance fields. *arXiv preprint arXiv:2208.02705* (2022). doi:10.48550/arXiv.2208.02705. 21

[HEH07] HOIEM D., EFROS A. A., HEBERT M.: Recovering surface layout from an image. *International Journal of Computer Vision 75*, 1 (2007), 151–172. doi:10.1007/s11263-006-0031-y. 12

[HGDG17] HE K., GKIOXARI G., DOLLÁAR P., GIRSHICK R.: Mask R-CNN. In *Proc. ICCV* (2017), pp. 2961–2969. doi:10.1109/ICCV.2017.322. 16

[HHF09] HEDAU V., HOIEM D., FORSYTH D.: Recovering the spatial layout of cluttered rooms. In *Proc. ICCV* (2009), pp. 1849–1856. doi:10.1109/ICCV.2009.5459411. 5, 12

[HHY*25] HUANG Z., HE J., YE J., JIANG L., LI W., CHEN Y., HAN T.: Scene4U: Hierarchical layered 3D scene reconstruction from single panoramic image for your immerse exploration. In *Proc. CVPR* (2025), pp. 26723–26733. doi:10.1109/CVPR52734.2025.02489. 24

[HK18] HEDMAN P., KOPF J.: Instant 3D photography. *ACM TOG. 37*, 4 (2018), 101:1–101:12. doi:10.1145/3197517.3201384. 19

[HLB*22] HUTCHCROFT W., LI Y., BOYADZHIEV I., WAN Z., WANG H., KANG S. B.: CoVisPose: Co-visibility pose transformer for wide-baseline relative pose estimation in 360° indoor panoramas. In *Proc. ECCV* (2022), pp. 615–633. doi:10.1007/978-3-031-19824-3_36. 1

[HQZ18] HUANG S., QI S., ZHU Y.: Cooperative holistic scene understanding: Unifying 3D object, layout, and camera pose estimation. In *Proc. NeurIPS* (2018), pp. 294–305. URL: https://dl.acm.org/doi/abs/10.5555/3326943.3326963. 14

[HS21] HU R., SINGH A.: Uniting vision-and-language tasks via text generation. In *Proc. ICML* (2021). doi:10.48550/arXiv.2102.02779. 14

[HSC*25] HUANG C., SHAO F., CHEN H., MU B., XU L.: Gadfnet: Geometric priors assisted dual-projection fusion network for monocular panoramic depth estimation. *IEEE Trans. Circuits and Systems for Video Technology 35*, 9 (2025), 9060–9074. doi:10.1109/TCSVT.2025.3553472. 8

[HZRS16] HE K., ZHANG X., REN S., SUN J.: Deep residual learning for image recognition. In *Proc. CVPR* (2016), pp. 770–778. doi:10.1109/CVPR.2016.90. 8

[IEAB24] ISINGIZWE J., EIRIS R., AL-BAYATI A. J.: Enhancing safety training engagement through immersive storytelling: A case study in the residential construction. *Safety Science 179* (2024), 106631. doi:10.1016/j.ssci.2024.106631. 26, 27

[IHR*16] IM S., HA H., RAMEAU F., JEON H.-G., CHOE G., KWEON I. S.: All-around depth from small motion with a spherical panoramic camera. In *Proc. ECCV* (2016), pp. 156–172. doi:10.1007/978-3-319-46487-9_10. 4

[IYF15] IKEHATA S., YANG H., FURUKAWA Y.: Structured indoor modeling. In *Proc. ICCV* (2015), pp. 1323–1331. doi:10.1109/ICCV.2015.156. 11, 16

[JLAB19] JUNG R., LEE A. S. J., ASHTARI A., BAZIN J.: Deep360Up: A deep learning-based approach for automatic VR image upright adjustment. In *Proc. IEEE VR* (2019), pp. 1–8. doi:10.1109/VR.2019.8798326. 4

[JOV19] JOKELA T., OJALA J., VÄÄNÄNEN K.: How people use 360-degree cameras. In *Proc. International Conference on Mobile and Ubiquitous Multimedia* (2019), pp. 1–10. doi:10.1145/3365610.3365645. 1, 4

[JSL*25] JIANG H., SONG Z., LOU Z., XU R., TAN M.: Depth Anything in 360°: Towards scale invariance in the wild. *arXiv preprint arXiv:2512.22819* (2025). doi:10.48550/arXiv.2512.22819. 10, 24, 25

[JSZ*21] JIANG H., SHENG Z., ZHU S., DONG Z., HUANG R.: Unifuse: Unidirectional fusion for 360 panorama depth estimation. *IEEE Robotics and Automation Letters 6*, 2 (2021), 1519–1526. doi:10.1109/LRA.2021.3058957. 8, 10, 11

[JTA21] JAIN A., TANCIK M., ABBEEL P.: Putting NeRF on a diet: Semantically consistent few-shot view synthesis. In *Proc. ICCV* (2021), pp. 5885–5894. doi:10.1109/ICCV48922.2021.00583. 20

[JXXZ22] JIANG Z., XIANG Z., XU J., ZHAO M.: LGT-Net: Indoor panoramic room layout estimation with geometry-aware transformer network. In *Proc. CVPR* (2022), pp. 1654–1663. doi:10.1109/CVPR52688.2022.00170. 12, 13, 15

[JXZ*20] JIN L., XU Y., ZHENG J., ZHANG J., TANG R., XU S., YU J., GAO S.: Geometric structure based and regularized depth estimation from 360 indoor imagery. In *Proc. CVPR* (2020), pp. 889–898. doi:10.1109/CVPR42600.2020.00097. 7, 11

[KAKS25] KHAN M., ABU-KHALAF J., SUTER D.: PanoSCU Dataset. Public dataset, 2025. [Online; accessed: 2025-04-16]. URL: https://doi.org/10.25958/4p6j-z657. 6, 7

[KGC18] KENDALL A., GAL Y., CIPOLLA R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proc. CVPR* (2018), pp. 7482–7491. doi:10.1109/CVPR.2018.00781. 15

[KKLD23] KERBL B., KOPANAS G., LEIMKÜHLER T., DRETTAKIS G.: 3D gaussian splatting for real-time radiance field rendering. *ACM TOG 42*, 4 (2023), 139:1–139:14. doi:10.1145/3592433. 19, 20

[KLL19] KIM J. S., LEATHEM T., LIU J.: Comparing virtual reality modalities and 360 photography in a construction management classroom. In *55th ASC Annual International Conference Proceedings* (2019), pp. 221–228. URL: http://ascpro0.ascweb.org/archives/cd/2019/paper/CERT284002019.pdf. 26, 27

[KMH*17] KOLVE E., MOTTAGHI R., HAN W., VANDERBILT E., WEIHS L., HERRASTI A., DEITKE M., EHSANI K., GORDON D., ZHU Y., KEMBHAVI A., GUPTA A., FARHADI A.: AI2-THOR: An interactive 3D environment for visual AI. *arXiv preprint arXiv:1712.05474* (2017). doi:10.48550/arXiv.1712.05474. 7

[KMH*20] KAPLAN J., MCCANDLISH S., HENIGHAN T., BROWN T. B., CHESS B., CHILD R., GRAY S., RADFORD A., WU J., AMODEI D.: Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020). doi:10.48550/arXiv.2001.08361. 24

[KMKS24] KLASSON M., MEREU R., KANNALA J., SOLIN A.: Sources of uncertainty in 3D scene reconstruction. In *Proc. ECCV Workshops* (2024), Springer, pp. 271–289. doi:10.1007/978-3-031-91585-7_17. 28

[KMR*23] KIRILLOV A., MINTUN E., RAVI N., MAO H., ROLLAND C., GUSTAFSON L., XIAO T., WHITEHEAD S., BERG A. C., LO W.-Y., DOLLÁR P., GIRSHICK R.: Segment anything. In *Proc. ICCV* (2023), pp. 3992–4003. doi:10.1109/ICCV51070.2023.00371. 24

[KQC*25] KHAN M., QIU Y., CONG Y., ABU-KHALAF J., SUTER D., ROSENHAHN B.: PanoSCU: A simulation-based dataset for panoramic indoor scene understanding. *IEEE Access 13* (2025), 72456–72476. doi:10.1109/ACCESS.2025.3561055. 7, 26

[KYS23] KULKARNI S., YIN P., SCHERER S.: 360FusionNeRF: Panoramic neural radiance fields with joint guidance. In *Proc. IROS* (2023), pp. 7202–7209. doi:10.1109/IROS55552.2023.10341346. 20, 21, 25

[LBL*25] LI Y., BOYADZHIEV I., LIU Z., SHAPIRO L., COLBURN A.: Badgr: Bundle adjustment diffusion conditioned by gradients for wide-baseline floor plan reconstruction. In *Proc. CVPR* (2025), pp. 16785–16795. doi:10.1109/CVPR52734.2025.01564. 16, 17

[LCG15] LIU F., CHUNHUA SHEN, GUOSHENG LIN: Deep convolutional neural fields for depth estimation from a single image. In *Proc. CVPR* (2015), pp. 5162–5170. doi:10.1109/CVPR.2015.7299152. 8

[LDHG*25] LYU L., DESCHAINTRE V., HOLD-GEOFFROY Y., HASAN M., YOON J. S., LEIMKÜHLER T., THEOBALT C., GEORGIEV I.: IntrinsicEdit: Precise generative image manipulation in intrinsic space. *ACM Trans. Graph. 44*, 4 (2025). doi:10.1145/3731173. 23

[LGND*24] LIANG R., GOJCIC Z., NIMIER-DAVID M., ACUNA D., VIJAYKUMAR N., FIDLER S., WANG Z.: Photorealistic object insertion with diffusion-guided inverse rendering. In *Proc. ECCV* (2024), pp. 446–465. doi:10.1007/978-3-031-73030-6_25. 23

[LGW*22] LI J., GAO W., WU Y., LIU Y., SHEN Y.: High-quality indoor scene 3D reconstruction with rgb-d cameras: A brief review. *Computational Visual Media 8*, 3 (2022), 369–393. doi:10.1007/s41095-021-0250-8. 2, 3

[LGY*22] LI Y., GUO Y., YAN Z., HUANG X., DUAN Y., REN L.: Omnifusion: 360 monocular depth estimation via geometry-aware fusion. In *Proc. CVPR* (2022), pp. 2801–2810. doi:10.1109/CVPR52688.2022.00282. 9, 10, 11, 25

[LHK09] LEE D. C., HEBERT M., KANADE T.: Geometric reasoning for single image structure recovery. In *Proc. CVPR* (2009), pp. 2136–2143. doi:10.1109/CVPR.2009.5206872. 5, 12, 13

[LHKK18] LEE J., HEO M., KIM K., KIM C.: Single-image depth estimation based on fourier domain analysis. In *Proc. CVPR* (2018), pp. 330–339. doi:10.1109/CVPR.2018.00042. 8

[LHS*20] LUO X., HUANG J.-B., SZELISKI R., MATZEN K., KOPF J.: Consistent video depth estimation. *ACM Trans. Graph. 39*, 4 (2020). doi:10.1145/3386569.3392377. 25

[LLB*22] LAMBERT J., LI Y., BOYADZHIEV I., WIXSON L., NARAYANA M., HUTCHCROFT W., HAYS J., DELLAERT F., KANG S. B.: Salve: Semantic alignment verification for floorplan reconstruction from sparse panoramas. In *Proc. ECCV* (2022), Springer, pp. 647–664. doi:10.1007/978-3-031-19821-2_37. 16, 17

[LLC*21] LIU Z., LIN Y., CAO Y., HU H., WEI Y., ZHANG Z., LIN S., GUO B.: Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. ICCV* (2021), pp. 10012–10022. doi:10.1109/ICCV48922.2021.00986. 14

[LLG*19] LEWIS M., LIU Y., GOYAL N., GHAZVININEJAD M., MOHAMED A., LEVY O., STOYANOV V., ZETTLEMOYER L.: BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019). doi:10.48550/arXiv.1910.13461. 14

[LLJ*21] LIU B., LIU X., JIN X., STONE P., LIU Q.: Conflict-averse gradient descent for multi-task learning. *Proc. NeurIPS 34* (2021), 18878–18890. doi:10.48550/arXiv.2110.14048. 15

[LLK21] LEE J.-H., LEE C., KIM C.-S.: Learning multiple pixelwise tasks based on loss scale balancing. In *Proc. ICCV* (2021), pp. 5107–5116. 15

[LLZ16] LAMBERT-LACROIX S., ZWALD L.: The adaptive BerHu penalty in robust regression. *Journal of Nonparametric Statistics 28* (2016), 1–28. doi:10.1080/10485252.2016.1190359. 8

[LPLJ25] LEE J., PARK H., LEE B.-U., JOO K.: HUSH: Holistic panoramic 3D scene understanding using spherical harmonics. In *Proc. CVPR* (2025), pp. 16599–16608. doi:10.1109/CVPR52734.2025.01547. 8, 11, 12, 13

[LRB*16] LAINA I., RUPPRECHT C., BELAGIANNIS V., TOMBARI F., NAVAB N.: Deeper depth prediction with fully convolutional residual networks. In *Proc. 3DV* (2016), pp. 239–248. doi:10.1109/3DV.2016.32. 8, 10

[LWF18] LIU C., WU J., FURUKAWA Y.: FloorNet: A unified framework for floorplan reconstruction from 3D scans. In *Proc. ECCV* (2018), pp. 201–217. doi:10.1007/978-3-030-01231-1_13. 16

[LWH*22a] LI Z., WANG L., HUANG X., PAN C., YANG J.: FutureHouse Dataset. Public dataset, 2022. [Online; accessed: 2025-10-09]. URL: https://github.com/LZleejean/FutureHouse. 6, 7

[LWH*22b] LI Z., WANG L., HUANG X., PAN C., YANG J.: PhyIR: Physics-based inverse rendering for panoramic indoor images. In *Proc. CVPR* (2022), pp. 12703–12713. doi:10.1109/CVPR52688.2022.01238. 7, 23, 27

[LXM*20] LIN K.-E., XU Z., MILDENHALL B., SRINIVASAN P. P., HOLD-GEOFFROY Y., DIVERDI S., SUN Q., SUNKAVALLI K., RAMAMOORTHI R.: Deep multi depth panoramas for view synthesis. In *Proc. ECCV* (2020), pp. 328–344. doi:10.1007/978-3-030-58601-0_20. 19

[LXRY18] LUO B., XU F., RICHARDT C., YONG J.-H.: Parallax360: Stereoscopic 360° scene representation for head-motion parallax. *IEEE TVCG 24*, 4 (2018), 1545–1553. doi:10.1109/TVCG.2018.2794071. 19

[LZC*21] LIU Z., ZHANG Z., CAO Y., HU H., TONG X., DAI B., LIN S.: Group-free 3D object detection via transformers. In *Proc. ICCV* (2021), pp. 2949–2958. doi:10.1109/ICCV48922.2021.00294. 14, 17

[LZC*22] LIU H., ZHENG Y., CHEN G., CUI S., HAN X.: Towards high-fidelity single-view holistic reconstruction of indoor scenes. In *Proc. ECCV* (2022), pp. 429–446. doi:10.1007/978-3-031-19769-7_25. 14

[LZH*22] LI D., ZHANG Y., HÄNE C., TANG D., VARSHNEY A., DU R.: OmniSyn: Synthesizing 360 videos with wide-baseline panoramas. In *Proc. VRW* (2022), pp. 670–671. doi:10.48550/arXiv.2202.08752. 22

[LZH*26] LI H., ZHENG W., HE J., LIU Y., LIN X., YANG X., CHEN Y.-C., GUO C.: DA^2: Depth Anything in any direction. In *Proc. ICLR* (2026). URL: https://depth-any-in-any-dir.github.io/. 27

[LZW*25a] LI T., ZHANG Z., WANG Y., CUI Y., LI Y., ZHOU D., YIN B., YANG X.: Self-supervised indoor scene point cloud completion from a single panorama. *The Visual Computer 41*, 3 (2025), 1891–1905. doi:10.1007/s00371-024-03509-w. 7

[LZW*25b] LIU X., ZHOU T., WANG C., WANG Y., WANG Y., CAO Q., DU W., YANG Y., HE J., QIAO Y., ET AL.: Toward the unification of generative and discriminative visual foundation model: a survey. *The Visual Computer 41*, 5 (2025), 3371–3412. doi:10.1007/s00371-024-03608-8. 25

[Mat17] MATTERPORT: Matterport3D. Public dataset, 2017. [Online; accessed: 2025-09-16]. URL: https://github.com/niessner/Matterport. 5, 6, 7

[MCE*17] MATZEN K., COHEN M. F., EVANS B., KOPF J., SZELISKI R.: Low-cost 360 stereo photography and video capture. *ACM TOG 36*, 4 (2017), 148:1–148:12. doi:10.1145/3072959.3073645. 2, 17

[MFP*25] MASLEJ N., FATTORINI L., PERRAULT R., GIL Y., PARLI V., KARIUKI N., CAPSTICK E., REUEL A., BRYNJOLFSSON E., ETCHEMENDY J., ET AL.: Artificial intelligence index report 2025. *arXiv preprint arXiv:2504.07139* (2025). doi:10.48550/arXiv.2310.03715. 25, 28

[MMBM15] MONSZPART A., MELLADO N., BROSTOW G. J., MITRA N. J.: Rapter: rebuilding man-made scenes with regular arrangements of planes. *ACM TOG 34*, 4 (2015), 103–1. doi:10.1145/2766995. 16

[MMPM16] MOULON P., MONASSE P., PERROT R., MARLET R.: OpenMVG: Open multiple view geometry. In *International Workshop on Reproducible Research in Pattern Recognition* (2016), Springer, pp. 60–74. doi:10.1007/978-3-319-56414-2_5. 6

[MON*19] MESCHEDER L., OECHSLE M., NIEMEYER M., NOWOZIN S., GEIGER A.: Occupancy networks: Learning 3D reconstruction in function space. In *Proc. CVPR* (2019), pp. 4460–4470. doi:10.1109/CVPR.2019.00459. 14

[MP21] MARRINAN T., PAPKA M. E.: Real-time omnidirectional stereo rendering: generating 360° surround-view panoramic images for comfortable immersive viewing. *IEEE TVCG 27*, 5 (2021), 2587–2596. doi:10.1109/TVCG.2021.3067780. 20

[MST*21] MILDENHALL B., SRINIVASAN P. P., TANCIK M., BARRON J. T., RAMAMOORTHI R., NG R.: NeRF: representing scenes as neural radiance fields for view synthesis. *Commun. ACM 65*, 1 (2021), 99—-106. doi:10.1145/3503250. 19, 20

[MZZ*25] MENG M., ZHU Y., ZHAO Y., LI Z., ZHU Z.: 3D indoor scene geometry estimation from a single omnidirectional image: A comprehensive survey. *Computational Visual Media 11*, 3 (2025), 431–464. doi:10.26599/CVM.2025.9450438. 3

[NF20] NAUATA N., FURUKAWA Y.: Vectorizing world buildings: Planar graph reconstruction by primitive detection and relationship inference. In *Proc. ECCV* (2020), pp. 711–726. doi:10.1007/978-3-030-58598-3_42. 16

[NHHN20] NIE Y., HOU J., HAN X., NIESSNER M.: Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *Proc. CVPR* (2020), pp. 55–64. doi:10.1109/CVPR42600.2020.00013. 14

[NKQ*25] NAVEED H., KHAN A. U., QIU S., SAQIB M., ANWAR S., USMAN M., AKHTAR N., BARNES N., MIAN A.: A comprehensive overview of large language models. *ACM Trans. Intell. Syst. Technol. 16*, 5 (2025). doi:10.1145/3744746. 24

[PAA*21] PINTORE G., AGUS M., ALMANSA E., SCHNEIDER J., GOBBETTI E.: SliceNet: deep dense depth estimation from a single indoor panorama using a slice-based representation. In *Proc. CVPR* (2021), pp. 11536–11545. doi:10.1109/CVPR46437.2021.01137. 4, 5, 7, 8, 9, 10, 11

[PAAG21] PINTORE G., ALMANSA E., AGUS M., GOBBETTI E.: Deep3DLayout: 3D reconstruction of an indoor layout from a spherical panoramic image. *ACM TOG 40*, 6 (2021), 250:1–250:12. doi:10.1145/3478513.3480480. 4, 12, 13

[PAAG22] PINTORE G., AGUS M., ALMANSA E., GOBBETTI E.: Instant automatic emptying of panoramic indoor scenes. *IEEE TVCG 28*, 11 (2022), 3629–3639. doi:10.1109/TVCG.2022.3202999. 11, 23, 27

[PAG20] PINTORE G., AGUS M., GOBBETTI E.: AtlantaNet: Inferring the 3D indoor layout from a single 360 image beyond the Manhattan World assumption. In *Proc. ECCV* (2020), pp. 432–448. doi:10.1007/978-3-030-58598-3_26. 5, 12, 13, 15

[PASG24] PINTORE G., AGUS M., SIGNORONI A., GOBBETTI E.: DDD: Deep indoor panoramic depth estimation with density maps consistency. In *STAG: Smart Tools and Applications in Graphics* (2024). doi:10.2312/stag.20241336. 8

[PASG25] PINTORE G., AGUS M., SIGNORONI A., GOBBETTI E.: DDD++: Exploiting density map consistency for deep depth estimation in indoor environments. *Graphical Models 140* (2025), 101281. doi:10.1016/j.gmod.2025.101281. 8

[PBAG23] PINTORE G., BETTIO F., AGUS M., GOBBETTI E.: Deep scene synthesis of Atlanta-world interiors from a single omnidirectional image. *IEEE TVCG 29*, 11 (2023), 4708–4718. doi:10.1109/TVCG.2023.3320219. 19, 20

[PBE99] PELEG S., BEN-EZRA M.: Stereo panorama with a single camera. In *Proc. CVPR* (1999), pp. 395–401. doi:10.1109/CVPR.1999.786969. 19

[PdLGAAB18] PAYEN DE LA GARANDERIE G., ATAPOUR ABARGHOUEI A., BRECKON T. P.: Eliminating the blind spot: Adapting 3D object detection and monocular depth estimation to 360 panoramic imagery. In *Proc. ECCV* (2018), pp. 812–830. doi:10.1007/978-3-030-01261-8_48. 8

[PGJG19] PINTORE G., GANOVELLI F., JASPE VILLANUEVA A., GOBBETTI E.: Automatic modeling of cluttered floorplans from panoramic images. *Computer Graphics Forum 38*, 7 (2019), 347–358. 16

[PGP*18] PINTORE G., GANOVELLI F., PINTUS R., SCOPIGNO R., GOBBETTI E.: 3D floor plan recovery from overlapping spherical images. *Computational Visual Media 4*, 4 (2018), 367–383. doi:10.1007/s41095-018-0125-9. 5

[PJAG25] PINTORE G., JASHARI S., AGUS M., GOBBETTI E.: PanoFloor: Reconstruction and immersive exploration of large multi-room scenes from a minimal set of registered panoramic images using denoised density maps. In *Proc. ISMAR* (2025), pp. 414–424. doi:10.1109/ISMAR67309.2025.00052. 11, 17, 18, 23

[PJH*23] PINTORE G., JASPE VILLANUEVA A., HADWIGET M., GOBBETTI E., SCHNEIDER J., AGUS M.: PanoVerse: automatic generation of stereoscopic environments from single indoor panoramic images for metaverse applications. In *Proc. ACM Web3D* (2023), pp. 2:1–2:10. doi:10.1145/3611314.3615914. 19

[PJVH*24] PINTORE G., JASPE-VILLANUEVA A., HADWIGER M., SCHNEIDER J., AGUS M., MARTON F., BETTIO F., GOBBETTI E.: Deep synthesis and exploration of omnidirectional stereoscopic environments from a single surround-view panoramic image. *Computers & Graphics 119* (2024), 103907. doi:10.1016/j.cag.2024.103907. 20, 22

[PMG*20] PINTORE G., MURA C., GANOVELLI F., FUENTES-PEREZ L., PAJAROLA R., GOBBETTI E.: State-of-the-art in automatic 3D reconstruction of structured indoor environments. *Comput. Graph. Forum 39*, 2 (2020), 667–699. doi:10.1111/cgf.14021. 2, 3, 4, 5, 11, 12

[PPG*18] PINTORE G., PINTUS R., GANOVELLI F., SCOPIGNO R., GOBBETTI E.: Recovering 3D existing-conditions of indoor structures from spherical images. *Computers & Graphics 77* (2018), 16–29. doi:10.1016/j.cag.2018.09.013. 4, 26

[PPL*24] PATIL A. G., PATIL S. G., LI M., FISHER M., SAVVA M., ZHANG H.: Advances in data-driven analysis and synthesis of 3D indoor scenes. *Computer Graphics Forum 43*, 1 (2024), e14927. doi:10.1111/cgf.14927. 2, 3, 5

[PSAG25] PINTORE G., SHAH U., AGUS M., GOBBETTI E.: NadirFloorNet: reconstructing multi-room floorplans from a small set of registered panoramic images. In *Proc. CVPR* (2025), pp. 1985–1994. doi:10.1109/CVPRW67362.2025.00186. 16, 17

[PXZ*15] PENG WANG, XIAOHUI SHEN, ZHE LIN, COHEN S., PRICE B., YUILLE A.: Towards unified depth and semantic prediction from a single image. In *Proc. CVPR* (2015), pp. 2800–2809. doi:10.1109/CVPR.2015.7298897. 8

[PZ23] PENG C.-H., ZHANG J.: High-resolution depth estimation for 360° panoramas through perspective and panoramic depth images registration. In *Proc. WACV* (2023), pp. 3115–3124. doi:10.1109/WACV56688.2023.00313. 25

[PZL24] PU G., ZHAO Y., LIAN Z.: Pano2Room: Novel view synthesis from a single indoor panorama. In *Proc. SIGGRAPH Asia Conference Papers* (2024). doi:10.1145/3680528.3687616. 16, 20, 21, 22, 23, 25

[QLW*18] QI C. R., LIU W., WU C., SU H., GUIBAS L. J.: Frustum pointnets for 3D object detection from RGB-D data. In *Proc. CVPR* (2018), pp. 918–927. doi:10.1109/CVPR.2018.00102. 14

[RAR25] REY-AREA M., RICHARDT C.: 360° 3D photos from a single 360° input image. *IEEE TVCG 31*, 5 (2025), 2426–2434. doi:10.1109/TVCG.2025.3549538. 19, 20, 21, 25

[RAYR22a] REY-AREA M., YUAN M., RICHARDT C.: 360MonoDepth: High-resolution 360° monocular depth estimation. In *Proc. CVPR* (2022), pp. 3762–3772. doi:10.1109/CVPR52688.2022.00374. 4, 5, 6, 7, 9, 25, 27

[RAYR22b] REY-AREA M., YUAN M., RICHARDT C.: 360MonoDepth Matterport3D. Public dataset, 2022. [Online; accessed 2025-09-17]. URL: https://researchdata.bath.ac.uk/1126/. 5, 6

[RAYR22c] REY-AREA M., YUAN M., RICHARDT C.: 360MonoDepth Replica. Public dataset, 2022. [Online; accessed 2025-09-17]. URL: https://manurare.github.io/360monodepth/. 6, 17

[RB98] RADEMACHER P., BISHOP G.: Multiple-center-of-projection images. In *Proc. SIGGRAPH* (1998), pp. 199–206. doi:10.1145/280814.280871. 19

[RBL*22] ROMBACH R., BLATTMANN A., LORENZ D., ESSER P., OMMER B.: High-resolution image synthesis with latent diffusion models. In *Proc. CVPR* (2022), pp. 10684–10695. doi:10.1109/CVPR52688.2022.01042. 21, 25

[RHGS17] REN S., HE K., GIRSHICK R., SUN J.: Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence 39*, 6 (2017), 1137–1149. doi:10.1109/TPAMI.2016.2577031. 14

[RKH*21] RADFORD A., KIM J. W., HALLACY C., RAMESH A., GOH G., AGARWAL S., SASTRY G., ASKELL A., MISHKIN P., CLARK J., KRUEGER G., SUTSKEVER I.: Learning transferable visual models from natural language supervision. In *Proc. ICML* (2021). URL: https://icml.cc/virtual/2021/oral/9194. 14, 24

[RLH*22] RANFTL R., LASINGER K., HAFNER D., SCHINDLER K., KOLTUN V.: Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer . *IEEE Transactions on Pattern Analysis & Machine Intelligence 44*, 3 (2022), 1623–1637. doi:10.1109/TPAMI.2020.3019967. 10

[RSR*20] RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W., LIU P. J.: Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res. 21*, 1 (2020). 14

[SAB*20] SULAIMAN M. Z., AZIZ M. N. A., BAKAR M. H. A., HALILI N. A., AZUDDIN M. A.: Matterport: virtual tour as a new marketing approach in real estate business during pandemic COVID-19. In *Proc. IMDES* (2020), pp. 221–226. doi:10.2991/assehr.k.201202.079. 1, 2

[SAPS22] SÁNCHEZ ALCÁZAR A. A., PINTORE G., SGRENZAROLI M.: Indoor3Dmapping dataset. Public dataset, 2022. [Online; accessed: 2025-09-17]. URL: https://doi.org/10.5281/zenodo.6367381. 5

[SBXX25] SHEN S., BAO Z., XU W., XIAO C.: IllumiDiff: Indoor illumination estimation from a single image with diffusion model. *IEEE TVCG 31*, 10 (2025), 7752–7768. doi:10.1109/TVCG.2025.3553853. 23

[SD04] SCHINDLER G., DELLAERT F.: Atlanta world: an expectation maximization framework for simultaneous low-level edge grouping and camera calibration in complex man-made environments. In *Proc. CVPR* (2004), vol. 1, pp. I–I. doi:10.1109/CVPR.2004.1315033. 5

[SEKL23] SCHULT J., ENGELMANN F., KONTOGIANNI T., LEIBE B.: Mask3D: Mask transformer for 3D semantic instance segmentation. In *Proc. ICCV* (2023). doi:10.1109/ICRA48891.2023.10160590. 14

[SG17] SU Y.-C., GRAUMAN K.: Learning spherical convolution for fast features from 360 imagery. In *Proc. NeurIPS* (2017), pp. 529–539. URL: https://dl.acm.org/doi/abs/10.5555/3294771.3294822. 8, 9

[SG19a] SU Y., GRAUMAN K.: Kernel transformer networks for compact spherical convolution. In *Proc. CVPR* (2019), pp. 9434–9443. doi:10.1109/CVPR.2019.00967. 8, 9

[SG19b] SUBRAMANIAN P., GHEISARI M.: Using 360-degree panoramic photogrammetry and laser scanning techniques to create point cloud data: A comparative pilot study. In *Proc. Associated Schools of Construction Annual Conference* (2019), pp. 743–750. URL: http://ascpro0.ascweb.org/archives/cd/2019/paper/CPRT305002019.pdf. 26

[SGC*24] SLAVCHEVA M., GAUSEBECK D., CHEN K., BUCHHOFER D., SABIK A., MA C., DHILLON S., BRANDT O., DOLHASZ A.: An empty room is all we want: Automatic defurnishing of indoor panoramas. In *Proc. CVPRW* (2024), pp. 7384–7394. doi:10.1109/CVPRW63382.2024.00734. 23

[SHKF12] SILBERMAN N., HOIEM D., KOHLI P., FERGUS R.: Indoor segmentation and support inference from rgbd images. In *Proc. ECCV* (2012), Springer, pp. 746–760. doi:10.1007/978-3-642-33715-4_54. 16

[SHSC19] SUN C., HSIAO C.-W., SUN M., CHEN H.-T.: HorizonNet: Learning room layout with 1D representation and pano stretch data augmentation. In *Proc. CVPR* (2019), pp. 1047–1056. doi:10.1109/CVPR.2019.00114. 12, 13, 15

[SJT*25] SHAH U., JASHARI S., TUKUR M., HOUSEH M., SCHNEIDER J., PINTORE G., GOBBETTI E., AGUS M.: Virtual staging of indoor panoramic images via multi-task learning and inverse rendering. *IEEE CG&A* (2025), 1–14. doi:10.1109/MCG.2025.3605806. 23, 27

[SK18] SENER O., KOLTUN V.: Multi-task learning as multi-objective optimization. In *Proc. NeurIPS* (2018), p. 525–536. doi:10.48550/arXiv.1810.04650. 15

[SKC*19] SERRANO A., KIM I., CHEN Z., DIVERDI S., GUTIERREZ D., HERTZMANN A., MASIA B.: Motion parallax for 360 rgbd video. *IEEE TVCG 25*, 5 (2019), 1817–1827. doi:10.1109/TVCG.2019.2898757. 19

[SLK*23] SHINDE Y., LEE K., KIPER B., SIMPSON M., HASANZADEH S.: A systematic literature review on 360° panoramic applications in architecture, engineering, and construction (AEC) industry. *J. ITcon 28*, 21 (2023), 405–437. 26

[SLL*22] SHEN Z., LIN C., LIAO K., NIE L., ZHENG Z., ZHAO Y.: PanoFormer: Panorama transformer for indoor 360 depth estimation. In *Proc. ECCV* (2022), Springer, pp. 195–211. doi:10.1007/978-3-031-19769-7_12. 9, 10, 11

[SMB*20] SITZMANN V., MARTEL J., BERGMAN A., LINDELL D., WETZSTEIN G.: Implicit neural representations with periodic activation functions. In *Proc. NeurIPS* (2020). doi:10.48550/arXiv.2006.09661. 14

[SRFL21] STEKOVIC S., RAD M., FRAUNDORFER F., LEPETIT V.: Montefloor: Extending MCTS for reconstructing accurate large-scale floor plans. In *Proc. CVPR* (2021), pp. 16034–16043. doi:10.1109/ICCV48922.2021.01573. 16

[SSC21] SUN C., SUN M., CHEN H.-T.: HoHoNet: 360° indoor holistic understanding with latent horizontal features. In *Proc. CVPR* (2021), pp. 2573–2582. doi:10.1109/CVPR46437.2021.00260. 8, 10

[SSK19] SRIDEVI G., SRINIVAS KUMAR S.: Image inpainting based on fractional-order nonlinear diffusion for image reconstruction. *Circuits, Systems, and Signal Processing 38*, 8 (2019), 3802–3817. doi:10.1007/s00034-019-01029-w. 19

[SSN09] SAXENA A., SUN M., NG A. Y.: Make3D: Learning 3D scene structure from a single still image. *IEEE TPAMI 31*, 5 (2009), 824–840. doi:10.1109/TPAMI.2008.132. 8

[SSO*21] SHABANI M. A., SONG W., ODAMAKI M., FUJIKI H., FURUKAWA Y.: Extreme structure from motion for indoor panoramas without visual overlaps. In *Proc. ICCV* (2021), pp. 5683–5691. doi:10.1109/ICCV48922.2021.00565. 2, 7, 16, 17, 18

[STA*24] SHAH U., TUKUR M., ALZUBAIDI M., PINTORE G., GOBBETTI E., HOUSEH M., SCHNEIDER J., AGUS M.: MultiPanoWise: holistic deep architecture for multi-task dense prediction from a single panoramic image. In *Proc. CVPRW - OmniCV* (2024), pp. 1311–1321. doi:10.1109/CVPRW63382.2024.00138. 23

[STP*23] SU J.-W., TUNG K.-Y., PENG C.-H., WONKA P., CHU H.-K.: SLIBO-Net: Floorplan reconstruction via slicing box representation with local geometry regularization. In *Proc. NeurIPS* (2023), pp. 1–12. URL: https://dl.acm.org/doi/abs/10.5555/3666122.3668240. 16

[STS*24] SUN C., TAI W.-E., SHIH Y.-L., CHEN K.-W., SYU Y.-J., WANG Y.-C. F., CHEN H.-T.: Seg2Reg: Differentiable 2D segmentation to 1D regression rendering for 360° room layout reconstruction. In *Proc. CVPR* (2024), pp. 10435–10445. doi:10.1109/CVPR52733.2024.00993. 12, 13, 15

[SU17] STANFORD-UNIVERSITY: BuildingParser Dataset. Public dataset, 2017. [Online; accessed: 2025-09-16]. URL: https://sdss.redivis.com/datasets/9q3m-9w5pa1a2h. 6

[SUKc20] SHANGHAITECH-UNIVERSITY, KUJIALE-COM: Shanghaitech-Kujiale Indoor 360° (SKI360) dataset. Public dataset, 2020. [Online; accessed 2025-09-16]. URL: https://svip-lab.github.io/dataset/indoor_360.html. 6, 7

[SWM*19a] STRAUB J., WHELAN T., MA L., CHEN Y., WIJMANS E., GREEN S., ENGEL J. J., MUR-ARTAL R., REN C., VERMA S., CLARKSON A., YAN M., BUDGE B., YAN Y., PAN X., YON J., ZOU Y., LEON K., CARTER N., BRIALES J., GILLINGHAM T., MUEGGLER E., PESQUEIRA L., SAVVA M., BATRA D., STRASDAT H. M., NARDI R. D., GOESELE M., LOVEGROVE S., NEWCOMBE R.: The Replica dataset: A digital replica of indoor spaces. *ArXiv e-print arXiv:1906.05797* (2019), 1–10. doi:10.48550/arXiv.1906.05797. 6

[SWM*19b] STRAUB J., WHELAN T., MA L., CHEN Y., WIJMANS E., GREEN S., ENGEL J. J., MUR-ARTAL R., REN C., VERMA S., CLARKSON A., YAN M., BUDGE B., YAN Y., PAN X., YON J., ZOU Y., LEON K., CARTER N., BRIALES J., GILLINGHAM T., MUEGGLER E., PESQUEIRA L., SAVVA M., BATRA D., STRASDAT H. M., NARDI R. D., GOESELE M., LOVEGROVE S., NEWCOMBE R.: The Replica Dataset. Public dataset, 2019. [Online; accessed 2025-09-17]. URL: https://github.com/facebookresearch/Replica-Dataset. 6

[SX16] SONG S., XIAO J.: Deep sliding shapes for amodal 3D object detection in RGB-D images. In *Proc. CVPR* (2016), pp. 808–816. doi:10.1109/CVPR.2016.94. 14

[SZ25] SONG K., ZHANG L.: Novel view synthesis with wide-baseline stereo pairs based on local–global information. *Computers & Graphics 126* (2025), 104139. doi:10.1016/j.cag.2024.104139. 22

[SZL*23] SHEN Z., ZHENG Z., LIN C., NIE L., LIAO K., ZHENG S., ZHAO Y.: Disentangling orthogonal planes for indoor panoramic room layout estimation with cross-scale distortion awareness. In *Proc. CVPR* (2023), pp. 17337–17345. doi:10.1109/CVPR52729.2023.01663. 9, 12, 13, 25

[TCD*21] TOUVRON H., CORD M., DOUZE M., MASSA F., SABLAYROLLES A., JÉGOU H.: Training data-efficient image transformers & distillation through attention. In *Proc. ICML* (2021). URL: https://proceedings.mlr.press/v139/touvron21a. 14

[TJA*25] TUKUR M., JASHARI S., ALZUBAIDI M., ABIODUN SALAMI B., BORAEY Y., YONG S., SALEH D., PINTORE G., GOBBETTI E., SCHNEIDER J., FETAIS N., AGUS M.: Panoramic imaging in immersive extended reality: A scoping review of technologies, applications, perceptual studies, and user experience challenges. *Frontiers in Virtual Reality 6* (2025). doi:10.3389/frvir.2025.1622605. 1, 2, 3, 26

[TJAH*25] TUKUR M., JASHARI S., ABOU HASSANAIN D., BETTIO F., SCHNEIDER J., PINTORE G., GOBBETTI E., AGUS M.: PanoStyleVR: style-based similarity metrics for web-based immersive panoramic style transfer. In *Proceedings of the 30th International Conference on 3D Web Technology* (2025), pp. 1–10. doi:10.1145/3746237.3746287. 23

[TJZ*24] TSAI Y.-J., JHANG J.-C., ZHENG J., WANG W., CHEN A. Y., SUN M., KUO C.-H., YANG M.-H.: No more ambiguity in 360° room layout via bi-layout estimation. In *Proc. CVPR* (2024), pp. 28056–28065. doi:10.1109/CVPR52733.2024.02650. 12, 13

[TNT18] TATENO K., NAVAB N., TOMBARI F.: Distortion-aware convolutional filters for dense prediction in panoramic images. In *Proc. ECCV* (2018), pp. 732–750. doi:10.1007/978-3-030-01270-0_43. 8, 9

[TPG*23] TUKUR M., PINTORE G., GOBBETTI E., SCHNEIDER J., AGUS M.: SPIDER: A framework for processing, editing and presenting immersive high-resolution spherical indoor scenes. *Graphical Models 128* (2023), 101182:1–101182:11. doi:10.1016/j.gmod.2023.101182. 19, 20

[Tra21] TRAN P. V.: SSLayout360: Semi-supervised indoor layout estimation from 360deg panorama. In *Proc. CVPR* (2021), pp. 15353–15362. doi:10.1109/CVPR46437.2021.01510. 12, 13

[TS20] TUCKER R., SNAVELY N.: Single-view view synthesis with multiplane images. In *Proc. CVPR* (2020), pp. 551–560. doi:10.1109/CVPR42600.2020.00063. 19, 20

[TUP*23] TUKUR M., UR REHMAN A., PINTORE G., GOBBETTI E., SCHNEIDER J., AGUS M.: PanoStyle: Semantic, geometry-aware and shading independent photorealistic style transfer for indoor panoramic scenes. In *Proc. ICCVW* (2023), pp. 1553–1564. doi:10.1109/ICCVW60793.2023.00170. 23, 27

[TZEM22] TULSIANI S., ZHOU T., EFROS A. A., MALIK J.: Multi-view supervision for single-view reconstruction via differentiable ray consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence 44*, 12 (2022), 8754–8765. doi:10.1109/TPAMI.2019.2898859. 14

[UKE25] UCHIDA T., KANAMORI Y., ENDO Y.: 3D view optimization for improving image aesthetics. In *Proc. ICASSP* (2025), pp. 1–5. doi:10.1109/ICASSP49660.2025.10888966. 23

[VSP*17] VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L., POLOSUKHIN I.: Attention is all you need. In *Proc. NeurIPS* (2017), pp. 6000–60010. doi:10.48550/arXiv.1706.03762. 14

[WBK*24] WALLINGFORD M., BHATTAD A., KUSUPATI A., RAMANUJAN V., DEITKE M., KEMBHAVI A., MOTTAGHI R., MA W.-C., FARHADI A.: From an image to a scene: Learning to imagine the world from a million 360 videos. *Proc. NeurIPS 37* (2024), 17743–17760. doi:10.48550/arXiv.2412.07770. 25

[WGD*22] WAIDHOFER J., GADGIL R., DICKSON A., ZOLLMANN S., VENTURA J.: PanoSynthVR: Toward light-weight 360-degree view synthesis from a single panoramic input. In *Proc. ISMAR* (2022), pp. 584–592. doi:10.1109/ISMAR55827.2022.00075. 2, 17, 19, 20

[WGSJ20] WILES O., GKIOXARI G., SZELISKI R., JOHNSON J.: Synsin: End-to-end view synthesis from a single image. In *Proc. CVPR* (2020), pp. 7467–7477. doi:10.1109/CVPR42600.2020.00749. 19

[WHC*18] WANG F.-E., HU H.-N., CHENG H.-T., LIN J.-T., YANG S.-T., SHIH M.-L., CHU H.-K., SUN M.: Self-supervised learning of depth and camera motion from 360 videos. In *Proc. ACCV* (2018), pp. 53–68. doi:10.1007/978-3-030-20873-8_4. 9, 10

[WHZ*25] WANG X., HE Z., ZHANG Q., YANG Y., ZHAO T., JIANG J.: Geometry-aware self-supervised indoor 360° depth estimation via asymmetric dual-domain collaborative learning. *IEEE Trans. Multimedia 27* (2025), 3224–3237. doi:10.1109/TMM.2025.3535340. 9, 10

[WL24a] WANG H., LI M.: A new era of indoor scene reconstruction: A survey. *IEEE Access* (2024). doi:10.1109/ACCESS.2024.3440260. 2, 3

[WL24b] WANG N.-H. A., LIU Y.-L.: Depth anywhere: Enhancing 360 monocular depth estimation via perspective distillation and unlabeled data augmentation. *Proc. NeurIPS 37* (2024), 127739–127764. URL: https://dl.acm.org/doi/10.5555/3737916.3741972. 9, 24

[WLV*25] WIEDEMER T., LI Y., VICOL P., GU S. S., MATARESE N., SWERSKY K., KIM B., JAINI P., GEIRHOS R.: Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328* (2025). doi:10.48550/arXiv.2509.20328. 25, 28

[WLX*25] WANG S., LU Y., XIA Z., CHEN Z., WANG Y., ZHONG L., ZHONG T.: Integrating visual-SLAM and multi-view panoramas for efficient indoor 3D layout reconstruction in building projects. *Advanced Engineering Informatics 68* (2025), 103629. doi:10.1016/j.aei.2025.103629. 27

[WQTX25] WANG W., QING C., TAN J., XU X.: Multi-view panoramic image style transfer with multi-scale attention and global sharing. *ACM Trans. Multimedia Comput. Commun. Appl. 21*, 7 (2025). doi:10.1145/3735137. 23, 27

[WWC*24] WANG G., WANG P., CHEN Z., WANG W., LOY C. C., LIU Z.: PERF: Panoramic Neural Radiance Field from a single panorama. *IEEE TPAMI* (2024), 1–15. doi:10.1109/TPAMI.2024.3387307. 19, 20

[WYC*22] WANG Y., YE T., CAO L., HUANG W., SUN F., HE F., TAO D.: Bridged transformer for vision and point cloud 3D object detection. In *Proc. CVPR* (2022), pp. 12114–12123. doi:10.1109/CVPR52688.2022.01180. 14

[WYS*20] WANG F.-E., YEH Y.-H., SUN M., CHIU W.-C., TSAI Y.-H.: BiFuse: Monocular 360 depth estimation via bi-projection fusion. In *Proc. CVPR* (2020), pp. 462–471. doi:10.1109/CVPR42600.2020.00054. 8, 9, 10, 11

[WYS*21] WANG F.-E., YEH Y.-H., SUN M., CHIU W.-C., TSAI Y.-H.: LED2-Net: Monocular 360 layout estimation via differentiable depth rendering. In *Proc. CVPR* (2021), pp. 12956–12965. doi:10.1109/CVPR46437.2021.01276. 13, 15

[WYT*22] WANG F.-E., YEH Y.-H., TSAI Y.-H., CHIU W.-C., SUN M.: Bifuse++: Self-supervised and efficient bi-projection fusion for 360 depth estimation. *IEEE TPAMI 45*, 5 (2022), 5448–5460. doi:10.1109/TPAMI.2022.3203516. 9, 10

[WZS*18] WEI Z., ZHANG J., SHEN X., LIN Z., MECH R., HOAI M., SAMARAS D.: Good view hunting: Learning photo composition from dense view pairs. In *Proc. CVPR* (2018), pp. 5437–5446. doi:10.1109/CVPR.2018.00570. 19, 23

[XEOT12] XIAO J., EHINGER K. A., OLIVA A., TORRALBA A.: Recognizing scene viewpoint using panoramic place representation. In *Proc. CVPR* (2012), pp. 2695–2702. doi:10.1109/CVPR.2012.6247991. 6

[XGK*25] XING X., GROH K., KARAOGLU S., GEVERS T., BHATTAD A.: Luminet: Latent intrinsics meets diffusion models for indoor scene relighting. In *Proc. CVPR* (2025), pp. 442–452. doi:10.1109/CVPR52734.2025.00050. 23

[XHS*23] XIE D., HU P., SUN X., PIRK S., ZHANG J., MECH R., KAUFMAN A. E.: GAIT: Generating aesthetic indoor tours with deep reinforcement learning. In *Proc. ICCV* (2023), pp. 7409–7419. doi:10.1109/ICCV51070.2023.00681. 23

[XJW*22] XU D., JIANG Y., WANG P., FAN Z., SHI H., WANG Z.: SinNeRF: Training neural radiance fields on complex scenes from a single image. In *Proc. ECCV* (2022), pp. 736–753. doi:10.1007/978-3-031-20047-2_42. 20

[XLF*19] XIAN W., LI Z., FISHER M., EISENMANN J., SHECHTMAN E., SNAVELY N.: UprightNet: geometry-aware camera orientation estimation from single images. In *Proc. ICCV* (2019), pp. 9974–9983. doi:10.1109/ICCV.2019.01007. 4

[XLZLC20] XU M., LI C., ZHANG S., LE CALLET P.: State-of-the-art in 360° video/image processing: Perception, assessment and compression. *IEEE Journal of Selected Topics in Signal Processing 14*, 1 (2020), 5–26. doi:10.1109/JSTSP.2020.2966864. 2

[XRZH*18a] XIA F., R. ZAMIR A., HE Z., SAX A., MALIK J., SAVARESE S.: Gibson Env: real-world perception for embodied agents. In *Proc. CVPR* (2018). doi:10.1109/CVPR.2018.00945. 6, 7

[XRZH*18b] XIA F., R. ZAMIR A., HE Z., SAX A., MALIK J., SAVARESE S.: GibsonV2. Public dataset, 2018. [Online; accessed 2025-09-17]. URL: https://github.com/StanfordVL/GibsonEnv/tree/master/gibson/data. 6, 7

[XSKT17] XU J., STENGER B., KEROLA T., TUNG T.: Pano2CAD: Room layout from a single panorama image. In *Proc. WACV* (2017), pp. 354–362. doi:10.1109/WACV.2017.46. 12, 20

[XWT*18] XU D., WANG W., TANG H., LIU H., SEBE N., RICCI E.: Structured attention guided convolutional neural fields for monocular depth estimation. In *Proc. CVPR* (2018), pp. 3917–3925. doi:10.1109/CVPR.2018.00412. 8

[XWY*25] XIA Y., WENG S., YANG S., LIU J., ZHU C., TENG M., JIA Z., JIANG H., SHI B.: PanoWan: Lifting diffusion video generation models to 360° with latitude/longitude-aware mechanisms. *arXiv preprint arXiv:2505.22016* (2025). doi:10.48550/arXiv.2505.22016. 27

[XZH*21] XU M., ZHANG Z., HU H., WANG J., WANG L., WEI F., BAI X., LIU Z.: End-to-end semi-supervised object detection with soft teacher. In *Proc. ICCV* (2021), pp. 3040–3049. doi:10.1109/ICCV48922.2021.00305. 24

[XZX*21a] XU J., ZHENG J., XU Y., TANG R., GAO S.: Layout-guided novel view synthesis from a single indoor panorama. In *Proc. CVPR* (2021), pp. 16438–16447. doi:10.1109/CVPR46437.2021.01617. 7, 11, 19, 20, 21

[XZX*21b] XU J., ZHENG J., XU Y., TANG R., GAO S.: PNVS Dataset. Public dataset, 2021. [Online; accessed: 2025-09-16]. URL: https://github.com/bluestyle97/PNVS. 6, 7, 19

[YDH*25] YU H.-X., DUAN H., HERRMANN C., FREEMAN W. T., WU J.: WonderWorld: Interactive 3D scene generation from a single image. In *Proc. CVPR* (2025), pp. 5916–5926. doi:10.1109/CVPR52734.2025.00555. 24

[YGH23] YU J., GRASSI A. C. P., HIRTZ G.: Applications of deep learning for top-view omnidirectional imaging: A survey. In *Proc. CVPRW* (2023), pp. 6421–6433. doi:10.1109/CVPRW59228.2023.00683. 3

[YHJ*23] YU H., HE L., JIAN B., FENG W., LIU S.: PanelNet: Understanding 360 indoor environment via panel representation. In *Proc. CVPR* (2023), pp. 878–887. doi:10.1109/CVPR52729.2023.00091. 12, 13

[YJC*24] YE W., JI C., CHEN Z., GAO J., HUANG X., ZHANG S.-H., OUYANG W., HE T., ZHAO C., ZHANG G.: DiffPano: Scalable and consistent text to panorama generation with spherical epipolar-aware diffusion. In *Proc. NeurIPS* (2024), vol. 37, pp. 1304–1332. doi:10.48550/arXiv.2410.24203. 24

[YJL*18] YANG Y., JIN S., LIU R., , YU J.: Automatic 3D indoor scene modeling from single panorama. In *Proc. CVPR* (2018), pp. 3926–3934. doi:10.1109/CVPR.2018.00413. 20

[YKG*20] YU T., KUMAR S., GUPTA A., LEVINE S., HAUSMAN K., FINN C.: Gradient surgery for multi-task learning. *Proc. NeurIPS 33* (2020), 5824–5836. doi:10.48550/arXiv.2001.06782. 15

[YKH*24a] YANG L., KANG B., HUANG Z., XU X., FENG J., ZHAO H.: Depth Anything: Unleashing the power of large-scale unlabeled data. In *Proc. CVPR* (2024), pp. 10371–10381. doi:10.1109/CVPR52733.2024.00987. 24, 25

[YKH*24b] YANG L., KANG B., HUANG Z., ZHAO Z., XU X., FENG J., ZHAO H.: Depth Anything V2. *Proc. NeurIPS 37* (2024), 21875–21911. doi:10.52202/079017-0688. 10, 24, 25

[YKSE23] YUE Y., KONTOGIANNI T., SCHINDLER K., ENGELMANN F.: Connecting the Dots: Floorplan Reconstruction Using Two-Level Queries. In *Proc. CVPR* (2023). doi:10.1109/CVPR52729.2023.00088. 16, 17

[YLC*20] YANG S., LI B., CAO Y.-P., FU H., LAI Y.-K., KOBBELT L., HU S.-M.: Noise-resilient reconstruction of panoramas and 3D scenes using robot-mounted unsynchronized commodity RGB-D cameras. *ACM TOG 39*, 5 (2020), 152:1–152:15. doi:10.1145/3389412. 1

[YLY*19] YU J., LIN Z., YANG J., SHEN X., LU X., HUANG T. S.: Free-form image inpainting with gated convolution. In *Proc. ICCV* (2019), pp. 4471–4480. doi:10.1109/ICCV.2019.00457. 19

[YTZ*25] YANG S., TAN J., ZHANG M., WU T., WETZSTEIN G., LIU Z., LIN D.: LayerPano3D: Layered 3D panorama for hyper-immersive scene generation. In *Proc. SIGGRAPH Conference Papers* (2025). doi:10.1145/3721238.3730643. 24, 27

[YWL*25] YE W., WANG Y., LIU Y., LIN W., XIANG X.: Panoramic arbitrary style transfer with deformable distortion constraints. *Journal of Visual Communication and Image Representation 106* (2025), 104344. doi:https://doi.org/10.1016/j.jvcir.2024.104344. 23, 27

[YWP*19] YANG S.-T., WANG F.-E., PENG C.-H., WONKA P., SUN M., CHU H.-K.: DuLa-Net: A dual-projection network for estimating room layouts from a single RGB panorama. In *Proc. CVPR* (2019). doi:10.1109/CVPR.2019.00348. 12, 13

[YWZ*25] YAN Q., WANG Q., ZHAO K., CHEN J., LI B., CHU X., DENG F.: Spherefusion: Efficient panorama depth estimation via gated fusion. In *Proc. 3DV* (2025), pp. 855–865. doi:10.1109/3DV66043.2025.00084. 9, 10, 11

[YZ16] YANG H., ZHANG H.: Efficient 3D room shape recovery from a single panorama. In *Proc. CVPR* (2016), pp. 5422–5430. doi:10.1109/CVPR.2016.585. 12

[ZCB*22] ZHI T., CHEN B., BOYADZHIEV I., KANG S. B., HEBERT M., NARASIMHAN S. G.: Semantically supervised appearance decomposition for virtual staging from a single panorama. *ACM TOG 41*, 4 (2022). doi:10.1145/3528223.3530148. 23, 27

[ZCC16] ZHANG E., COHEN M. F., CURLESS B.: Emptying, refurnishing, and relighting indoor spaces. *ACM TOG 35*, 6 (2016), 174:1–174:14. doi:10.1145/2980179.2982432. 2

[ZCC*21] ZHANG C., CUI Z., CHEN C., LIU S., ZENG B., BAO H., ZHANG Y.: DeepPanoContext: Panoramic 3D scene understanding with holistic scene context graph and relation-based optimization. In *Proc. ICCV* (2021), pp. 12632–12641. doi:10.1109/ICCV48922.2021.01240. 11, 14, 17, 20

[ZCSH18] ZOU C., COLBURN A., SHAN Q., HOIEM D.: LayoutNet: Reconstructing the 3D room layout from a single RGB image. In *Proc. CVPR* (2018), pp. 2051–2059. doi:10.1109/CVPR.2018.00219. 12, 13

[ZCY*24] ZHOU H., CHENG X., YU W., TIAN Y., YUAN L.: HoloDreamer: Holistic 3D panoramic world generation from text descriptions. *arXiv preprint arXiv:2407.15187* (2024). doi:10.48550/arXiv.2407.15187. 20

[ZFCG19] ZENG Y., FU J., CHAO H., GUO B.: Learning pyramid-context encoder network for high-quality image inpainting. In *Proc. CVPR* (2019), pp. 1486–1494. doi:10.1109/CVPR.2019.00158. 19

[ZFX*25] ZHOU S., FAN Z., XU D., CHANG H., CHARI P., BHARADWAJ T., YOU S., WANG Z., KADAMBI A.: DreamScene360: Unconstrained text-to-3d scene generation with panoramic gaussian splatting. In *Proc. ECCV* (2025), pp. 324–342. doi:10.1007/978-3-031-72658-3_19. 24, 27

[ZGS*24] ZHU J., GAO C., SUN Q., WANG M., DENG Z.: A survey of indoor 3D reconstruction based on RGB-D cameras. *IEEE Access* (2024). doi:10.1109/ACCESS.2024.3443065. 2, 3

[ZKG20] ZENG W., KARAOGLU S., GEVERS T.: Joint 3D layout and depth prediction from a single indoor panorama image. In *Proc. ECCV* (2020), pp. 666–682. doi:10.1007/978-3-030-58517-4_39. 13

[ZKZ*19] ZIOULIS N., KARAKOTTAS A., ZARPALAS D., ALVAREZ F., DARAS P.: Spherical view synthesis for self-supervised 360° depth estimation. In *Proc. 3DV* (2019). doi:10.1109/3DV.2019.00081. 9, 10

[ZKZD18] ZIOULIS N., KARAKOTTAS A., ZARPALAS D., DARAS P.: OmniDepth: Dense depth estimation for indoors spherical panoramas. In *Proc. ECCV* (2018), pp. 453–471. doi:10.1007/978-3-030-01231-1_28. 5, 8, 9, 10, 11

[ZLW*23] ZHUANG C., LU Z., WANG Y., XIAO J., WANG Y.: SPDET: Edge-aware self-supervised panoramic depth estimation transformer with spherical geometry. *IEEE TPAMI 45*, 10 (2023), 12474–12489. doi:10.1109/TPAMI.2023.3272949. 7

[ZSG*18] ZOLLHÖFER M., STOTKO P., GÖRLITZ A., THEOBALT C., NIESSNER M., KLEIN R., KOLB A.: State-of-the-art on 3D reconstruction with RGB-D cameras. *Computer graphics forum 37*, 2 (2018), 625–652. doi:10.1111/cgf.13386. 2, 3

[ZSP*17] ZOU C., SU J.-W., PENG C.-H., COLBURN A., SHAN Q., WONKA P., CHU H.-K., HOIEM D.: MatterportLayout. Public dataset, 2017. [Online; accessed: 2025-09-16]. URL: https://github.com/ericsujw/Matterport3DLayoutAnnotation. 6

[ZSP*19] ZOU C., SU J.-W., PENG C.-H., COLBURN A., SHAN Q., WONKA P., CHU H.-K., HOIEM D.: 3D Manhattan room layout reconstruction from a single 360 image. *ArXiv e-print arXiv:1910.04099* (2019). doi:10.48550/arXiv.1910.04099. 5, 12, 13

[ZSP*21] ZOU C., SU J. W., PENG C. H., COLBURN A., SHAN Q., WONKA P., CHU H. K., HOIEM D.: Manhattan room layout reconstruction from a single 360 image: A comparative study of state-of-the-art methods. *International Journal of Computer Vision 129* (2021), 1410–1431. doi:10.1007/s11263-020-01426-8. 2, 3, 5, 6, 12, 15

[ZSTX14a] ZHANG Y., SONG S., TAN P., XIAO J.: PanoContext: A whole-room 3D context model for panoramic scene understanding. In *Proc. ECCV* (2014), pp. 668–686. doi:10.1007/978-3-319-10599-4_43. 6, 8, 11, 12, 13, 14

[ZSTX14b] ZHANG Y., SONG S., TAN P., XIAO J.: PanoContext dataset. Public dataset, 2014. [Online; accessed: 2025-09-16]. URL: https://panocontext.cs.princeton.edu/. 6

[ZTF*18] ZHOU T., TUCKER R., FLYNN J., FYFFE G., SNAVELY N.: Stereo magnification: Learning view synthesis using multiplane images. *ACM TOG 37*, 4 (2018), 68:1–68:12. doi:10.1145/3197517.3201323. 19, 20

[ZWF*13] ZHAO Q., WAN L., FENG W., ZHANG J., WONG T.-T.: Cube2Video: Navigate between cubic panoramas in real-time. *IEEE Trans. Multimedia 15*, 8 (2013), 1745–1754. doi:10.1109/TMM.2013.2280249. 19, 20, 22

[ZWG*24] ZHANG C., WU Q., GAMBARDELLA C. C., HUANG X., PHUNG D., OUYANG W., CAI J.: Taming stable diffusion for text to 360° panorama image generation. In *Proc. CVPR* (2024), pp. 6347–6357. doi:10.1109/CVPR52733.2024.00607. 24

[ZWXG22] ZHAO Y., WEN C., XUE Z., GAO Y.: 3D room layout estimation from a cubemap of panorama image via deep Manhattan hough transform. In *Proc. ECCV* (2022), Springer, pp. 637–654. doi:10.1007/978-3-031-19769-7_37. 12, 13

[ZXLL21] ZHANG J., XIA X., LIU R., LI N.: Enhancing human indoor cognitive map development and wayfinding performance with immersive augmented reality-based navigation systems. *Advanced Engineering Informatics 50* (2021), 101432. doi:10.1016/j.aei.2021.101432. 2

[ZZL*20a] ZHENG J., ZHANG J., LI J., TANG R., GAO S., ZHOU Z.: Indoor3Dmapping dataset. Public dataset, 2020. [Online; accessed: 2025-09-16]. URL: https://structured3d-dataset.org/. 6, 7

[ZZL*20b] ZHENG J., ZHANG J., LI J., TANG R., GAO S., ZHOU Z.: Structured3D: A large photo-realistic dataset for structured 3D modeling. In *Proc. ECCV* (2020), pp. 519–535. doi:10.1007/978-3-030-58545-7_30. 7