

UNIVERSITÀ DEGLI STUDI DI BRESCIA



UNIVERSITÀ
DEGLI STUDI
DI BRESCIA

Dipartimento di Ingegneria dell'Informazione

Dottorato di Ricerca in Ingegneria dell'Informazione

ING-INF/05

Sistemi di elaborazione delle informazioni

Ciclo XXXVIII

AI and Deep Learning Techniques for Multi-Room Indoor Reconstruction and Exploration from Panoramic Images

PhD Candidate: Giovanni Pintore

Supervisor: Prof. Alberto Signoroni

Co-supervisor: Prof. Giorgio Vassena

Academic Year 2025/2026

Abstract

Indoor environment reconstruction from visual data is a central challenge in computer vision, with applications in digital twins, immersive visualization, virtual tours, architectural documentation, and facility management. Among the available sensing modalities, 360° panoramic imagery offers complete scene coverage from a single viewpoint and enables fast, low-cost acquisition using consumer-grade devices. However, real-world datasets are typically sparse, often providing only one panorama per room and exhibiting clutter, occlusions, reflective surfaces, and limited viewpoint overlap. These conditions invalidate the assumptions of traditional multi-view pipelines and make the reconstruction of large, multi-room environments particularly difficult.

This thesis investigates how artificial intelligence and deep learning techniques can extract metrically meaningful geometry and structural information from sparse panoramic images, and how isolated single-view predictions can be fused into coherent representations suitable for multi-room modeling and immersive exploration. The work examines how structural indoor priors, gravity alignment, and spherical imaging geometry can be leveraged to overcome the limitations imposed by sparse acquisition.

The thesis introduces three main contributions. First, it presents a structure-aware depth estimation framework that incorporates density-map consistency to improve robustness and metric fidelity under cluttered and occluded indoor conditions. Second, it proposes a method for multi-room 3D floorplan reconstruction that employs nadir-referenced projections and deformable-attention fusion to aggregate isolated per-room predictions into coherent, metrically consistent multi-room layouts. Third, it develops an immersive, VR-ready exploration pipeline that integrates depth-based geometry, denoised density maps, structured layout reasoning, and MCOP-based stereo rendering to support view-consistent navigation across complex multi-room environments.

Overall, this thesis provides a comprehensive methodology that spans single-view geometric inference, multi-room structural reconstruction, and immersive visualization, advancing the scalability and practical deployment of panoramic-based indoor modeling.

Abstract in Italiano

Abstract

La ricostruzione degli ambienti interni a partire da dati visivi costituisce una sfida centrale nella computer vision, con ricadute rilevanti in ambiti quali digital twin, visualizzazione immersiva, virtual tour, modellazione architettonica e gestione degli spazi. Le immagini panoramiche a 360° offrono una copertura completa della scena da un unico punto di vista e permettono acquisizioni rapide ed economiche mediante dispositivi di largo consumo. Tuttavia, nelle applicazioni reali tali acquisizioni risultano spesso sparse, tipicamente limitate a una sola panoramica per stanza e caratterizzate da clutter, occlusioni, superfici riflettenti e minima sovrapposizione tra punti di vista. Queste condizioni rendono inadeguati i metodi tradizionali basati su multi-view geometry e complicano la ricostruzione di ambienti multi-stanza ampi e complessi.

Questa tesi esplora come tecniche di intelligenza artificiale e deep learning possano estrarre informazioni geometriche e strutturali da panoramiche sparse e come predizioni ottenute da singole viste possano essere combinate per generare rappresentazioni coerenti a livello multi-room, adatte anche a un'esplorazione immersiva. Il lavoro analizza come priors architettonici, allineamento alla gravità e caratteristiche intrinseche della visione sferica possano compensare le limitazioni introdotte dall'acquisizione incompleta.

La tesi introduce tre contributi principali. Il primo consiste in un framework per la stima della profondità da immagini panoramiche che integra vincoli di coerenza tramite density maps, migliorando robustezza e accuratezza metrica in scenari con forte occlusione e disordine. Il secondo contributo propone un metodo di ricostruzione di floorplan 3D multi-stanza basato su proiezioni nadir e fusione deformabile, capace di aggregare predizioni locali in un layout globale coerente e metricamente consistente. Il terzo contributo sviluppa una pipeline orientata alla realtà virtuale che combina geometria derivata dalla profondità, mappe di densità denoised, reasoning strutturale e rendering stereoscopico MCOP per consentire una navigazione fluida e visivamente consistente in ambienti multi-room complessi.

Nel loro insieme, questi contributi definiscono una metodologia completa che integra inferenza geometrica single-view, ricostruzione strutturale multi-room ed esplorazione immersiva, avanzando lo stato dell'arte nella modellazione indoor basata su immagini panoramiche sparse.

Contents

Contents	1
List of Figures	4
List of Tables	9
1 Introduction	11
2 Background	16
2.1 Panoramic Imaging for Indoor Environments	16
2.1.1 Common representation	17
2.1.2 Gravity alignment	18
2.1.3 Architectural and indoor priors	19
2.2 Depth Estimation from Monocular Panoramic Images	21
2.3 Room Layout Estimation from Panoramic Images	23
2.4 Multi-Room Modeling and Floorplans	26
2.5 Immersive Visualization and VR for Indoor Scenes	29
2.6 Summary and Open Challenges	30
3 Methodological Framework	32
3.1 Data-Driven Learning for Panoramic Indoor Understanding	32
3.2 Supervised Learning Paradigm	33
3.3 Training Strategy and Evaluation Protocols	34
3.4 Datasets Related to This Thesis	34
4 Deep Depth Estimation from Panoramic Images	37
4.1 Problem Definition and Motivation	37

4.2	Task-related works	39
4.3	Network architecture	40
4.4	Indoor-specific loss function and training strategy	43
4.5	Results	46
4.5.1	Datasets	46
4.5.2	Training setup and computational performance	47
4.5.3	Evaluation and comparison with the state-of-the-art	48
4.5.4	Comparison with structured-guided baselines and ablation	51
4.6	Failure cases	54
4.7	Conclusions	54
4.8	Bibliographic notes	55
5	Multi-Room 3D Floorplan	
	Reconstruction	62
5.1	Problem Definition and Motivation	62
5.2	Task-related works	64
5.3	Overview	66
5.4	Nadir shape generation	67
5.5	Nadir floorplan reconstruction	70
5.6	Results	71
5.6.1	Training and inference cost	71
5.6.2	Datasets	72
5.6.3	Nadir shape reconstruction performance	73
5.6.4	Floorplan reconstruction performance	74
5.6.5	Ablation study	75
5.7	Conclusion and future works	76
5.8	Bibliographic notes	77
6	Immersive Exploration of	
	Multi-Room Indoor Spaces	78
6.1	Problem Definition and Motivation	78
6.2	Task-related works	80
6.3	Overview	84
6.4	Geometric modeling	85
6.4.1	Depth maps prediction	85
6.4.2	Multi-room density map reconstruction	86
6.4.3	Floorplan segmentation	89
6.5	Visual modeling	90

6.5.1	Door-based graph topology	91
6.5.2	Immersive view synthesis	92
6.6	Interactive VR exploration	93
6.7	Results	94
6.7.1	Setup and computational performance	94
6.7.2	Quantitative performance analysis	96
6.7.3	Qualitative performance analysis	98
6.7.4	Failure cases	100
6.8	Conclusions	100
6.9	Bibliographic notes	101
7	Conclusions and Future Work	102
7.1	Summary of Contributions	102
7.2	Scientific and Practical Impact	104
7.3	Future Directions	104
	Publications Related to This Thesis	106
	Bibliography	109

List of Figures

2.1	Equirectangular projection of a spherical panorama. A spherical camera samples viewing directions on the unit sphere. The equirectangular projection (ERP) maps longitude to the horizontal axis and latitude to the vertical axis, resulting in a $360^\circ \times 180^\circ$ rectangular image that preserves angular spacing but introduces characteristic stretching near the poles.	17
2.2	Occlusion of the architectural boundary. Areas occluded by furniture are in yellow, while those occluded by other walls are in red. Image from Pintore et al. [1].	19
2.3	Common indoor architectural priors. From left to right: <i>floor-wall boundaries</i> [2], <i>cuboids</i> [3], the <i>Indoor World Model</i> [4], the <i>Manhattan world</i> [5], the <i>Atlanta world</i> [6], the Vertical Walls [7] and <i>piece-wise planarity</i> [8]. From Pintore et al. [9].	20
2.4	Floorplan layout and connectivity graph recovery example [10]. Using relative panorama positions and per-view depth, the system reconstructs a global floorplan enriched with semantic elements (layouts, openings, clutter) and forms a connectivity graph describing navigable structure.	28
2.5	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 [11]. . . .	30

4.1	Deep depth estimation overview. The network maps a gravity-aligned 360° image to its depth. The input image is first transformed by a ResNet block (light green) into feature maps with different depths and spatial sizes. Through a Gravity-Aligned Features (GAF) encoding block (light purple), we perform a gravity-aligned anisotropic contractive encoding to obtain latent features. Once assembled in a single sequence, latent features are processed by a single-layer multi-head self-attention scheme (light pink) to produce the final set of features whose decoding, through convolution and upsampling, produces the desired depth map. During training, we produce density maps respectively from predicted and ground truth depths and exploit them for structural loss computation. It should be noted that density maps are computed only at training time. Thus, the memory and computational costs do not affect network performance at inference time. Note that the depth image and density map colors have been visually enhanced for illustration purposes.	57
4.2	Different density maps examples. In the first row we show the input image Figure 4.2a and the equirectangular depth Figure 4.2b. In the second row we show examples of density maps recovered from the equirectangular depth. In this case we illustrate the density maps adopted by the DDD baseline [12]: Figure 4.2c and Figure 4.2d, respectively projections along the X and Y axes; Figure 4.2e along Z axis, that is the density map projected on the floorplan (i.e., vertical projection). Here we keep Figure 4.2e, which is highly representative of the room shape, and we replace the arbitrary projections Figure 4.2c and Figure 4.2d with a more general and comprehensive cylindrical projection Figure 4.2f, which cover the whole horizon (i.e., horizontal projection). As also evident visually, this projection best captures seamless structural details.	58
4.3	Depth estimation qualitative performance. We show our reconstructions and ground-truth models as depth maps and as the associated point clouds from the SKI360 dataset [13]. Thanks to structural consistency and without specific post-processing, the method effectively preserves architectural details, such as sharp edges and smooth planes.	59

4.4	Depth estimation qualitative performance and comparisons on Matterport3D dataset [14]. We illustrate our qualitative performance compared to the competitor [15] on real-world scenes. We show reconstructions and ground-truth models as depth maps and as canonical views of the associated point clouds.	59
4.5	Depth estimation qualitative performance and comparisons on Structured3D dataset [16]. We illustrate our qualitative performance compared to the competitor [15] that reported the best performance in Table 4.2. We show reconstructions and ground-truth models as depth maps and as canonical views of the associated point clouds.	60
4.6	Depth estimation qualitative comparison with structure-guided indoor depth estimation baseline. We illustrate our qualitative performance on SKI360 dataset [13], compared to a state-of-the-art solution that, as ours, exploits gravity alignment (SliceNet [15]). We show reconstructions and ground-truth models as depth maps and as canonical views of the associated point clouds.	60
4.7	Failure case illustrating the limitations of the proposed loss term.. The input RGB image (Figure 4.7a) depicts a semi-outdoor scene that violates the indoor structural assumptions of the training data. As a result, the density map, particularly on the horizontal plane (Figure 4.7c), lacks meaningful features. Despite this, the network is still able to produce a plausible pixel-wise depth prediction (Figure 4.7b). 61	
5.1	NadirFloor overview. For each input panoramic image, a Nadir shape network forward pass infers a Nadir shape and the heights of floor-ceiling planes (see section 5.4). These representations are then merged into a fixed-size Nadir floorplan P_f , exploiting the available mutual registration. The Nadir floorplan is processed through the Nadir floorplan Network (see section 5.5) to predict the 2D polygonal multi-room floorplan. Finally, a 3D floorplan is recovered by extruding the 2D plan using the room heights recovered by the Nadir shape network.	64
5.2	Nadir shape network last layers. Starting from the intermediate uncluttered depth D_e , the fully differentiable transform and projection recover the Nadir projection P_f . Finally, the last autoencoding layers return the Nadir Shape N_p	67

5.3	Floorplan qualitative performances on Structured3D (S3D) [16] dataset. Each row shows: RoomFormer (RF) [17] point cloud occupancy map (RF occ. map), our Nadir floorplan map (Our Nadir map), ground truth floorplan (GT), RoomFormer prediction (RF), our floorplan prediction just composing the Nadir shapes (our NSO - Nadir shapes only), our full floorplan prediction (Our full), our 3D prediction with predicted metric information (Our 3D).	72
5.4	Floorplan qualitative performance on ZInD dataset [18]. For each row we show: RoomFormer (RF) [17] point cloud occupancy map (RF occ. map), our Nadir floorplan map (Our Nadir map), ground truth floorplan (GT), RoomFormer prediction (RF), our floorplan prediction just composing the Nadir shapes (our NSO - Nadir shapes only), our full floorplan prediction (Our full), our 3D prediction with predicted metric information (Our 3D).	73
6.1	Geometric modeling overview. Depth maps predicted by a depth network guide both novel view synthesis and structural reconstruction. A denoising network and gated auto-encoder refine a unified density map, which is segmented to extract room layouts, openings, and clutter. The result is scaled to world dimensions for accurate floorplan modeling.	81
6.2	Visual modeling overview. Visual modeling aligns RGB-D poses to the floorplan, forming a navigable graph with immersive node views and smooth visual transitions. A view-dependent renderer and a view-synthesis network generate stereoscopic images for nodes and 360° videos for arcs, enabling real-time WebXR visualization.	84
6.3	Density denoising. Comparison between predicted (fig. 6.3c) and denoised (fig. 6.3c) density map (images enhanced for illustration purposes). Although the two maps may look the same, the denoising process recovers some of the missing structural information in the latent space as noise $E(\delta)$. To better show, such latent difference has been decoded with the autoencoder in fig. 6.3b.	87

6.4	Visual modeling main tasks. Visual modeling finalizes the geometric model by constructing a navigable graph with immersive visual content for nodes and transitions. fig. 6.4a shows the estimation of optimal navigation trajectories between panoramic viewpoints. This process accounts for structural constraints, including openings and obstacles, while minimizing visual discontinuities and occlusions during transitions. fig. 6.4b shows the baseline [19] adopted for view-synthesis, which, in our use, can output spherical slices to compose MCOP or full-view spherical images for transitions. fig. 6.4c and fig. 6.4d show HMD user interface.	87
6.5	Path optimization. At initialization, paths connect source to target poses, passing through doors (left). After optimization, paths avoid walls and clutter and minimize disocclusions (right). The green mark identifies a very difficult case, in which the detected door is invisible in one of the views.	90
6.6	Examples of model reconstructions from Structured3D [16]. For each scene, we show the denoised density map, our predicted model from denoised with openings, the ground truth, the prediction from original noisy density map, and the multi-room scene graph in world coordinates (several rotated for better fitting in the illustration), including clutter maps, computed trajectories and screenshots at primary and transition poses. Notably, our method distinguishes between doors and windows based on relative queries (section 6.4.3).	96
6.7	Transition comparison. Comparison between synthesized equirectangular views (top) and live frames (bottom) with and without our optimized trajectory. Without path optimization, the trajectory passes through walls and clutter to reach the final viewpoint.	98
6.8	Panoramic floorplan failure example. We show an example where the accuracy of the reconstruction affects the final result. Although the result appears to be a good reconstruction at first glance, with all doors present, the incorrect positioning of the wall at the top prevents access to the upper right room.	99
6.9	Examples of reconstructions on real-world captures, using our models trained on synthetic data [16] for prediction. Such experiments demonstrate the capabilities of domain transfer and zero-shot reconstruction.	99

List of Tables

4.1	Computational performance of depth inference. We show our computational performance compared to other state-of-the-art works for a 512×1024 image. We also show an example of inference time on a NVIDIA RTX 4090.	48
4.2	Depth estimation performance compared to SoA works. We present results and comparisons on both real-world and synthetic datasets. For comparison, we follow the same setup of the recent publication Elite360D [20], which provides exhaustive results about the latest state-of-the-art approaches. To support further evaluations, we train, under the same conditions [20], <i>SliceNet</i> , <i>DDD</i> , and <i>DDD++</i>	49
4.3	Comparison with structure-guided baselines. The first two rows report the results obtained with the network of Jin et al. [21] without (no SC) and with (with SC) the inclusion of geometric priors and regularizers. The third row reports the results obtained by SliceNet [15], which uses GAFs but no structural consistency terms. The last two rows report the results obtained with the DDD baseline [12] and with DDD++.	52
4.4	DDD++ Ablation study. The full model (DDD++, bottom row) combines gravity-aligned features (GAF), multi-head self-attention (MHSA), and cylindrical density loss (cyl DL), achieving top performance. Removing cyl DL (DDD row) causes a measurable drop, validating its improvement over planar density loss. Progressively disabling MHSA and GAF (upper rows) further degrades accuracy, with the plain CNN baseline (top row, no GAF/MHSA/DL) performing worst.	53

5.1	Nadir shape single room layout performance. We show our quantitative performance (in bold) compared to other representative state-of-the-art works. We compute IoU2D (pixel-wise intersection over union) and IoU3D (volumetric intersection over union) on ZInD [18] benchmark.	73
5.2	Floorplan performance. We show our quantitative performance (Our full - in bold) on Structured3D (S3D) [16] and ZInD [18], compared to other representative state-of-the-art works (RoomFormer (RF) [17]) adapted to the same 360° images input. We also include the performance of our method just spatially registering individual Nadir shapes (NS) - Nadir shapes only).	74
5.3	Ablation study. NS: spatially registering of individual Nadir shapes from cluttered depth maps; UNC+NS: NS from uncluttered depth maps; NS+NF: our full pipeline but without depth uncluttering; UNC+NS+NF: our full pipeline with uncluttered depth-shapes. . . .	76
6.1	Performance of depth inference. We show our performance compared to other state-of-the-art works for a 512×1024 image. All methods were trained and tested using the same conditions. Other results are the same as reported in recent publications [20], while ADM [19] has been retrained by us.	94
6.2	Floorplan performance and ablation. We show our quantitative performance on Structured3D (S3D) [16] (Our density denoised). For a fair comparison from predicted images, we adopt an alternative SoA depth prediction method to generate competitors' density maps (ADM [19]). We compare then with ADM [19]+Polydiffuse [22], which is currently the method that performs best for segmentation of polygonal rooms, and RoomFormer [17], which is the same baseline of PolyDiffuse but provides full floorplan segmentation (ADM [19]+RoomFormer [17]). Finally, as an ablation experiment, we compare to our pipeline but without denoising (Our plain density).	95
6.3	AE and diffusion computational performance. We show our computational performance compared to a standard solution for the same density map size of 256×256	96
6.4	View-synthesis performance. We show our quantitative performance compared to other state-of-the-art works.	97

1 Introduction

The automatic 3D reconstruction and modeling of indoor environments has gained substantial traction in recent years, becoming a central research topic in both computer graphics and computer vision. The demand for accurate and scalable indoor models is driven by a wide range of applications—including architecture, engineering, facility management, cultural heritage preservation, immersive visualization, real estate, and digital twins—where virtual replicas of real environments must integrate geometry, appearance, and semantics in a unified, interactive representation.

Traditional indoor acquisition pipelines rely on dense measurements obtained through RGB-D sensors, laser scanners, or multi-view images. While these technologies can deliver accurate reconstructions, they often require specialized equipment, significant capture time, and controlled acquisition conditions. In contrast, *360° panoramic imagery* provides a uniquely efficient alternative: a single panorama captures the full visual content surrounding a viewpoint, supports rapid and low-cost acquisition with consumer devices, and is natively compatible with immersive visualization. These advantages have led to the widespread adoption of panoramic imaging in industrial applications such as real-estate virtual tours, inspection workflows, and cultural-heritage digitization.

However, panoramic imagery also introduces fundamental challenges. The equirectangular projection exhibits characteristic distortions; large indoor spaces often feature irregular geometry, clutter, and self-occlusions; and reflective or transparent surfaces introduce ambiguities that complicate geometric inference. Moreover, real-world panoramic datasets are typically *sparse*: collections such as Zillow Indoor or Matterport frequently include only a single panorama per room, sometimes with minimal overlap between viewpoints across rooms. As discussed in Chapter 2, this sparse distribution violates the assumptions underlying standard multi-view reconstruction, where geometric correspondences or depth-from-motion cues are essential.

For these reasons, indoor reconstruction from panoramic imagery requires methods that can maximize the structural information extracted from *each individual*

panorama, while also enabling reasoning that spans multiple sparsely distributed views. This thesis addresses this challenge by studying how modern AI and deep learning techniques can be leveraged to infer depth, structure, and spatial layout from single panoramas, and how these local predictions can be fused into coherent *multi-room* reconstructions that support immersive exploration.

A central insight guiding this research is that indoor environments exhibit strong structural regularities: gravity-aligned architecture, vertical walls, floor–ceiling parallelism, piecewise planarity, and Manhattan or Atlanta world tendencies. When combined with the global field of view provided by panoramic imagery, these priors supply powerful constraints that can compensate for the lack of multi-view redundancy. The background discussion in Chapter 2 highlights how these geometric cues form the basis for both single-view reasoning (e.g., depth and layout prediction) and multi-room structural modeling.

Within this context, two tasks become fundamental for understanding the indoor scene from a single spherical observation:

- **Depth estimation**, which provides the essential metric 3D structure required for interpreting geometry, identifying openings, and supporting view synthesis and VR rendering;
- **Layout prediction**, which abstracts the architectural scaffold of a room, separating permanent structural elements (walls, floors, ceilings) from clutter, and serving as the basis for floorplan inference and multi-room connectivity.

Together, these tasks enable rich single-view understanding, but extending them to multi-room settings requires additional mechanisms for global reasoning. Sparse panoramas must be fused into a consistent structural representation that captures room adjacencies, wall continuity, openings, and overall topology—despite the lack of viewpoint overlap. This challenge motivates the development of multi-room floorplan models, nadir-shape fusion strategies, and connectivity graphs discussed later in this thesis.

Beyond reconstruction, real-world applications increasingly require immersive exploration of indoor spaces. To support VR navigation, reconstructed models must satisfy stronger geometric and perceptual constraints: consistent parallax, smooth view transitions, stereo cues, and efficient rendering representations. Chapter 2 reviews the evolution of immersive visualization techniques based on panoramic images, including depth-based warping, spherical novel-view synthesis, and multi-center-of-projection rendering. These challenges directly motivate the VR-oriented pipeline developed in Chapter 6.

Altogether, the aims of this thesis can be summarized as designing a principled and scalable methodology for reconstructing and exploring multi-room indoor scenes from *sparse panoramic data*, by combining single-view geometric inference with global structural reasoning and immersive visualization.

Research Problem and Objectives

The central research problem addressed by this thesis is:

How can large, complex, multi-room indoor environments be reconstructed and made immersively explorable using only sparse panoramic images—often with a single view per room?

This question gives rise to three specific objectives:

- **Depth estimation under indoor structural constraints:** designing models that exploit gravity alignment, architectural priors, and structure-aware supervision to improve robustness in cluttered and occluded scenes;
- **Sparse multi-room floorplan reasoning:** developing representations and learning models capable of aggregating isolated per-room predictions into a globally coherent, metrically meaningful multi-room layout;
- **Immersive exploration and VR synthesis:** generating view-consistent, perceptually stable renderings that support stereo cues, parallax, and navigation across multiple rooms, despite sparse geometric input.

Addressing these objectives requires bridging single-view inference with global multi-room reasoning, while ensuring that the final representations remain compatible with real-time, VR-ready visualization.

Research Questions

The thesis is structured around the following key research questions:

1. **Single-view geometry:** *How can depth from a single panoramic image be improved by exploiting indoor structural priors, geometric consistency, and the global context of spherical imagery?*
2. **Sparse multi-room modeling:** *How can individually reconstructed rooms be fused into a coherent, metrically consistent multi-room floorplan when panoramas are extremely sparse and non-overlapping?*

3. **Immersive multi-room exploration:** *How can the reconstructed geometry and floorplan be transformed into a VR-ready representation that supports realistic parallax, stereo rendering, and smooth navigation across multiple rooms?*

These questions directly correspond to the three core contributions of the thesis.

Original Contributions

This thesis investigates **AI and deep learning techniques for multi-room indoor reconstruction and immersive exploration from sparse panoramic images**. Its contributions are structured around the three major stages of the reconstruction pipeline, each developed within a dedicated chapter of the thesis.

1. **Deep Depth Estimation from Panoramic Images (Chapter 4).** This chapter introduces a novel structure-aware framework for monocular depth estimation from 360° panoramas. The method leverages *density-map consistency* to enforce geometric regularity and improve metric accuracy under indoor clutter and occlusions. It demonstrates how gravity alignment, structural priors, and panoramic context can be exploited to significantly enhance depth predictions in challenging real-world environments.
2. **Multi-Room 3D Floorplan Reconstruction (Chapter 5).** This chapter presents a transformer-based approach for reconstructing coherent, metrically meaningful multi-room floorplans from sparse panoramic inputs. The method uses nadir-referenced geometric projections and deformable-attention fusion to combine isolated per-room observations into globally consistent 2D and 3D layouts, addressing the limitations of sparse, non-overlapping panoramic acquisitions.
3. **Immersive Exploration of Multi-Room Indoor Spaces (Chapter 6).** This chapter introduces a complete VR-oriented pipeline that transforms the reconstructed geometric and structural information into a view-consistent, navigable multi-room representation. It integrates density-map denoising, structured layout reasoning, stereo MCOP rendering, and optimized navigation trajectories to enable realistic VR exploration directly from sparse panoramic data.

Together, these three components form a unified framework that spans the entire processing chain—from single-view geometric inference, to multi-room structural

modeling, to immersive VR visualization—establishing a principled methodology for reconstructing complex indoor environments from sparse 360° imagery.

Thesis Organization

Chapter 2 provides a comprehensive background on panoramic imaging, gravity alignment, structural indoor priors, monocular depth estimation, layout prediction, multi-room reconstruction, and immersive visualization. Chapter 3 describes the methodological assumptions, learning paradigms, and evaluation protocols adopted throughout the thesis. Chapter 4 introduces *DDD++*, the proposed structure-aware depth estimator. Chapter 5 presents *NadirFloorNet*, the transformer-based framework for multi-room floorplan reconstruction. Chapter 6 describes *PanoFloor*, the immersive multi-room exploration pipeline for VR-ready rendering. Finally, Chapter 7 summarizes the findings and outlines directions for future research.

2 Background

This chapter reviews the main concepts and building blocks underlying panoramic indoor reconstruction and immersive exploration. We first introduce 360° imaging for indoor environments, discussing acquisition setups, equirectangular representations, gravity alignment, and the architectural priors that make single-panorama analysis feasible. We then focus on two fundamental tasks that operate at the single-room level: monocular depth estimation from panoramic images, which provides the metric 3D structure required by most downstream applications, and room layout estimation, which captures the architectural scaffold separating permanent structures from clutter. Building on these elements, we review methods for extending single-room analysis to multi-room modeling and floorplan reconstruction in the realistic regime of sparse panoramic coverage, where only one or a few views per room are available. Finally, we survey immersive visualization and VR pipelines that turn structured geometric and semantic reconstructions into view-consistent, navigable experiences, enabling stereo rendering, parallax, and multi-room navigation from sparse panoramic inputs.

2.1 Panoramic Imaging for Indoor Environments

Indoor environments can be documented through a variety of sensing technologies, from terrestrial laser scanners to active depth cameras [9]. Among these, photographic solutions remain extremely relevant, as they provide a practical and low-cost means of acquiring visual information while capturing appearance cues that are indispensable for tasks such as navigation, localization, and as-built modeling [23]. However, standard perspective cameras, constrained by their narrow field of view, are not ideal for full-scene coverage and typically require multi-view acquisition [24]. In indoor settings, this need for multiple viewpoints is further complicated by clutter, limited spatial accessibility, occluding structures, and surfaces with little texture or strong reflectance, all of which impair feature matching and surface reconstruction [9]. For these reasons, capturing a complete 360° image from a single location has become the preferred solution, as it enables fast and comprehensive acquisition [25, 26].

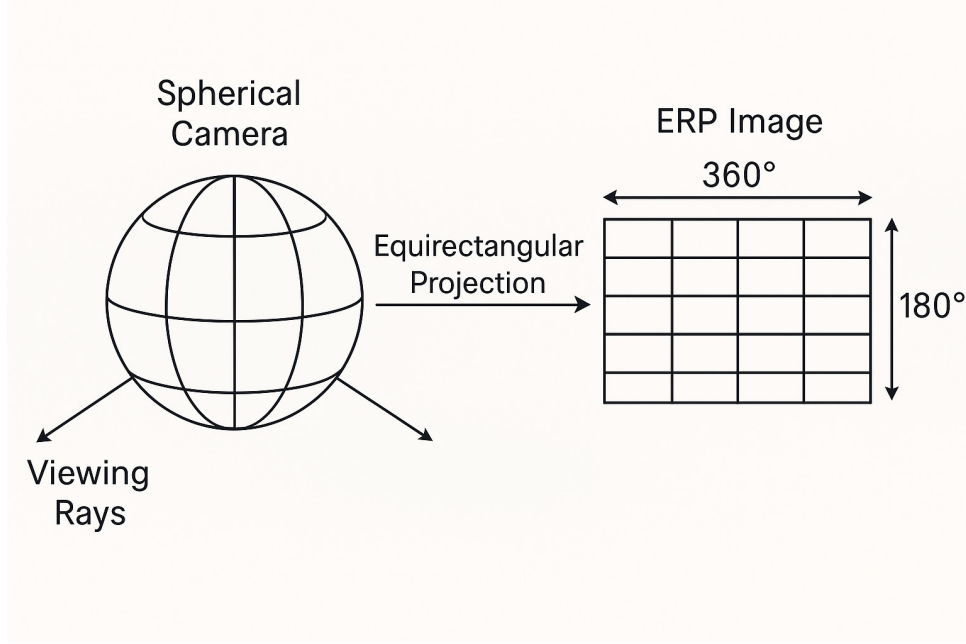


Figure 2.1: **Equirectangular projection of a spherical panorama.** A spherical camera samples viewing directions on the unit sphere. The equirectangular projection (ERP) maps longitude to the horizontal axis and latitude to the vertical axis, resulting in a $360^\circ \times 180^\circ$ rectangular image that preserves angular spacing but introduces characteristic stretching near the poles.

Panoramic images can still be created by stitching sequences of conventional photographs [27], but dedicated spherical cameras [28] have now become the standard option. Modern dual-fisheye devices internally align and blend multiple views, delivering a seamless panorama with a single shutter press. Their design ensures an approximately unique projection center and consistent sampling along the horizontal direction, advantages that are difficult to achieve through manual stitching [29, 30]. Such cameras can be modeled as sampling incoming rays on a unit sphere without intrinsic calibration parameters, with image formation depending only on the camera orientation [26].

2.1.1 Common representation

Because the sphere cannot be flattened without distortion, panoramic imagery must adopt a projection model. Although some cameras expose raw fisheye images for maximum resolution, the most widespread representation remains the $360^\circ \times 180^\circ$ *equirectangular projection (ERP)* [31]. ERP maps longitude to the horizontal axis and latitude to the vertical one, producing a rectangular layout where meridians and parallels appear as straight, uniformly spaced lines. This mapping is illustrated in Figure 2.1. While simple and device-independent, ERP introduces horizontal

stretching near the poles because neither area nor angles are preserved.

Alternative projections have been explored to mitigate such distortions. The cube-map projection distributes the spherical view across six faces of a cube with $90^\circ \times 90^\circ$ coverage each, and is widely used in visualization systems such as WebXR and YouTube 360 [32]. Tangent-image representations [33, 31] have also gained popularity, and comprehensive comparisons can be found in recent surveys [26].

Despite the existence of such alternatives, equirectangular images remain the predominant choice in indoor datasets and processing pipelines because they maintain full horizontal continuity, are supported across all major acquisition devices, and introduce distortions primarily in ceiling and floor regions, which carry limited structural detail in typical gravity-aligned captures. For these reasons, ERP continues to be the principal representation for the acquisition and analysis of 360° panoramic indoor imagery.

2.1.2 Gravity alignment

A widely accepted premise in both panoramic reconstruction pipelines and immersive exploration systems is that the acquired (or rendered) images are produced with the camera’s up-vector aligned to the global gravity direction. When this condition holds under an equirectangular projection, horizontal planar structures in the scene map neatly to horizontal rows in the panoramic image, while vertical elements follow consistent directions across the panorama. This apparently simple constraint has meaningful consequences for many processing stages. The most pronounced distortions of the ERP model occur near the poles of the spherical domain, which correspond to the ceiling and floor in typical indoor scenes; these areas generally contain fewer semantically or geometrically important details, making the distortion less detrimental for reconstruction and learning tasks. At the same time, an upright orientation ensures that many dominant spatial cues—walls, vertical edges, columnar structures—are naturally aligned with the pixel grid, allowing algorithms to leverage gravity-consistent features and leading to more stable and interpretable processing [34]. From a rendering perspective, such alignment produces an intuitive mapping between head rotations and pixel displacements: yaw and pitch motions correspond directly to horizontal and vertical shifts in the equirectangular frame, and stereoscopic visualization can be efficiently implemented using horizontal disparities, as illustrated in Figure 2.1. Because of these benefits, acquiring panoramic imagery under gravity alignment has become standard practice in indoor environments. Most publicly released 360° indoor datasets show negligible deviations from upright orientation [15], reflecting both the practical ease and methodological importance of maintaining

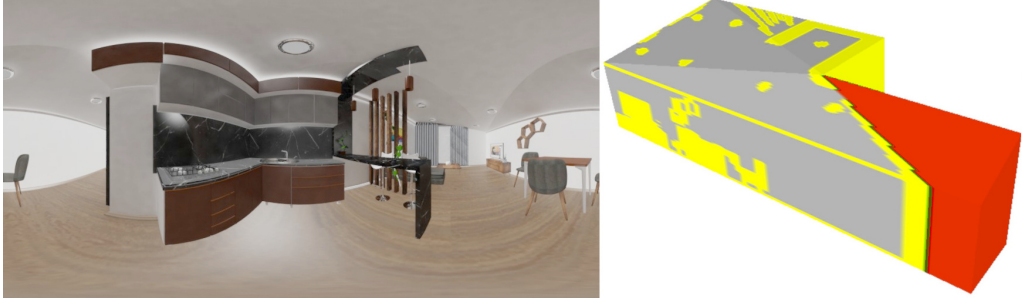


Figure 2.2: **Occlusion of the architectural boundary.** Areas occluded by furniture are in yellow, while those occluded by other walls are in red. Image from Pintore et al. [1].

this constraint. Achieving alignment during capture is generally straightforward: a spherical camera mounted on a leveled tripod naturally satisfies the requirement [34], while handheld devices usually rely on embedded inertial measurement units (IMUs), to measure and stabilize orientation in real time. Even in cases where images are taken without explicit attention to leveling, several image-based gravity estimation and rectification techniques can recover the dominant vertical direction from visual cues (e.g., [35, 36, 34]). These methods analyze structural regularities, vanishing directions, or the distribution of geometric primitives to infer the orientation of gravity, making it possible to normalize panoramas before applying gravity-dependent algorithms.

Overall, maintaining an upright camera configuration—either at acquisition time or through preprocessing—significantly simplifies the interpretation and processing of panoramic indoor imagery. It enables stronger geometric priors, reduces the impact of projection distortions, and provides stable reference axes that are exploited by both classical geometry-based pipelines and modern learning-based approaches.

2.1.3 Architectural and indoor priors

A single 360° panorama offers a holistic visual record of everything surrounding a viewpoint, from which one must reconstruct the geometry, structure, and appearance of the environment. Nevertheless, the information contained in such panoramas is often partial and corrupted by noise, making the underlying reconstruction problem highly underconstrained. Indoor panoramas frequently include large areas with scarce texture—such as plain walls—and the non-uniform angular sampling combined with the intrinsic distortions of the equirectangular format hampers the extraction of reliable features and consistent geometric cues (Figure 2.1). The situation becomes even more complex in the presence of reflective or transparent elements like mirrors

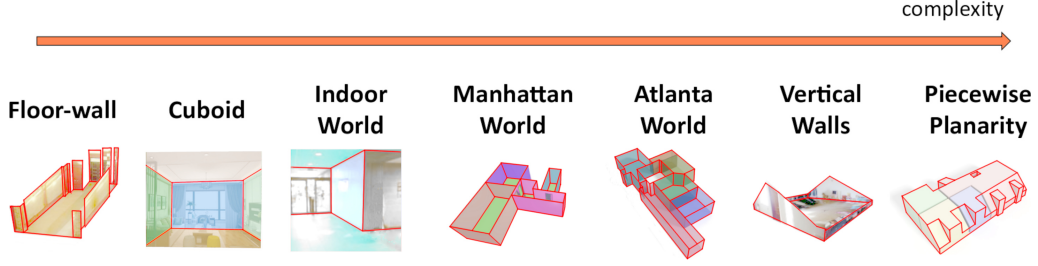


Figure 2.3: **Common indoor architectural priors.** From left to right: *floor-wall boundaries* [2], *cuboids* [3], the *Indoor World Model* [4], the *Manhattan world* [5], the *Atlanta world* [6], the *Vertical Walls* [7] and *piece-wise planarity* [8]. From Pintore et al. [9].

and glass, which generate ambiguous or deceptive observations. Furthermore, indoor depth ranges are typically much more irregular than in many outdoor scenarios: narrow, cluttered spaces introduce very close objects, while distant walls or openings extend several meters away. A more severe challenge arises from the fact that furniture and everyday objects occlude extensive portions of the architectural envelope, and concave room configurations create strong self-occlusions that hide large parts of the room boundary from the camera’s viewpoint (Figure 2.2).

For these reasons, achieving accurate depth estimation or structural recovery from a single panorama requires leveraging not only the global contextual information present in the image but also a set of robust, domain-specific assumptions derived from common architectural patterns [9]. Although general-purpose priors such as smoothness, symmetry, or repetition [37] can provide useful guidance, indoor reconstruction methods have traditionally relied on more explicit architectural knowledge. The most widely used priors are visualized in Figure 2.3—serve to constrain the problem and reduce ambiguity. Early techniques tended to adopt simple and highly constrained scene models, such as floor-wall boundary representations [2] or basic cuboidal abstractions [3]. With time, the Manhattan World assumption [5] has become a dominant reference model, as its orthogonality constraints align well with typical residential and office construction conventions and markedly limit the range of possible structural configurations [38]. More flexible Atlanta World formulations are also increasingly adopted, allowing horizontal directions to vary while preserving many optimization advantages [15].

Classical geometric reconstruction pipelines exploited these priors by detecting and interpreting low-level visual features—corners, straight lines, edge clusters, or homogeneous regions—according to the expected architectural structure [9]. The reliability of such approaches, however, is tightly linked to the quality of the extracted features, which often deteriorates in cluttered, occluded, or visually complex

environments, limiting their robustness in realistic conditions.

Consequently, the field has progressively transitioned toward data-driven methods [39], especially those based on deep neural networks [38]. By learning from extensive annotated datasets, deep models can capture latent geometric and semantic regularities that are difficult to encode manually and generally deliver more stable performance than purely geometric pipelines. Despite this shift, architectural priors remain a crucial component of modern methods: they narrow the solution space, impose meaningful structural constraints, and often allow reductions in problem dimensionality—for instance by exploiting vertical-wall hypotheses to derive floor plans—or simplify network architectures by organizing features along gravity-aligned directions.

2.2 Depth Estimation from Monocular Panoramic Images

Monocular depth estimation is a fundamental task in both computer graphics and computer vision, as it serves as the primary gateway to recovering the 3D structure of a scene. In indoor environments, depth provides the essential geometric foundation required for a wide range of downstream applications, including room layout reconstruction, navigation, interactive rendering, and novel view synthesis. Its importance is even more pronounced in the context of 360° panoramic imagery: while a single panorama captures the full field of view, it does so with significant geometric distortions, making metrically consistent depth estimation both crucial and challenging.

Early monocular depth estimation approaches relied on handcrafted cues, geometric reasoning, or multi-scale priors, but these solutions were often brittle in real-world indoor conditions. The emergence of deep learning, coupled with large-scale training datasets, enabled models to extract non-trivial depth cues directly from image appearance, dramatically improving robustness and generalization [15]. Moreover, the same architectural backbones used for depth estimation can be extended to predict additional signals such as surface normals or semantic labels. Eigen et al. [40] showed that these outputs can be jointly learned, enriching both geometric and semantic scene understanding. Depending on the application, such complementary signals may support tasks beyond reconstruction—e.g., selective editing, virtual staging, or object-level manipulation.

Classic data-driven monocular depth estimation Data-driven monocular depth estimation was introduced over a decade ago for limited field-of-view imagery, as in Make3D [41]. Deep CNN-based regressors [42, 40] soon replaced hand-crafted methods, and the influential Fully Convolutional Residual Network (*FCRN*) model [43] combined a *ResNet* [44] encoder with an up-projection decoder and the reverse Huber loss [45]. Numerous improvements followed, including Fourier-domain fusion of perspective crops [46], ordinal regression losses [47], and CRF-based refinement [48, 49, 50, 51]. Although originally designed for perspective imagery, these methods inspired initial attempts to handle panoramic content by sampling multiple perspective views from an equirectangular image and merging predictions back into the spherical domain. However, the approach did not fully exploit the characteristics of 360° imagery—its global context and inherent distortions—which often limited accuracy [52, 53].

Distortion-aware approaches To address the limitations of naive adaptations, distortion-aware methods were developed. Some solutions relied on cube-map projections [54], while others introduced spherical or distortion-aware convolutions to better model equirectangular geometry [55, 56, 57, 53, 58]. Hybrid designs also emerged, such as the dual-branch network of Wang et al. [59], which processes the panorama both in equirectangular format and as cube-map faces, combining a distortion-aware encoder [53] with a FCRN decoder [43].

Structure-aware approaches Another important direction exploits the structural regularities of indoor scenes (i.e., *structure-aware models*). These approaches [60, 15, 61] leverage gravity alignment and geometry-aware constraints to simplify the representation while still capturing long-range dependencies. SliceNet [15], for example, vertically compresses the equirectangular input and processes it with an RNN to learn horizontal structure, embedding indoor-specific priors directly into the architecture. This sequential encoding can be regarded as an early precursor to later transformer-based formulations that operate on structured panoramic tokens.

Building on this line of research, **DDD++** [61]—presented as one of the main contributions of this thesis in Chapter 4—extends structure-aware depth estimation by incorporating *density-map consistency* as an additional supervisory signal. Instead of relying solely on appearance cues or axis-aligned priors, DDD++ enforces geometric coherence across the predicted depth map by modeling the local statistics of 3D point distributions projected onto the panorama. This strategy preserves the lightweight structure-aware design while significantly improving robustness to clutter, occlusions, and challenging indoor geometries. As a result, DDD++ demonstrates how domain-

specific structural constraints and learned geometric regularities can be jointly exploited to advance panoramic depth estimation beyond existing structure-aware approaches.

Transformer-based approaches Transformers introduced the ability to capture global context and long-range relations more effectively than CNNs. *PanoFormer* [62] adapts transformer mechanisms using spherical tangent patches and specialized tokenization. Other approaches decompose the panorama into perspective views [63, 33] and fuse them using transformer-based reasoning. Patch-wise vision transformers [64, 65, 20] further demonstrate the benefits of attention for aggregating spatially disjoint cues. Hybrid multi-projection systems such as SphereFusion [66] combine spherical and equirectangular representations through gated fusion, balancing geometric fidelity and texture detail.

Self-supervised approaches Because high-quality 360° depth annotations are difficult to obtain, self- and weakly supervised approaches have attracted increasing attention. Early work includes *360-SelfNet* [67], based on cube padding, and the trinocular-synthesis framework of Zioulis et al. [68]. BiFuse++ [69] extended dual-branch designs [59] to self-supervision with spherical padding. More recent methods include distillation-based training, as in Depth Anywhere [70], and high-resolution weakly supervised strategies such as the Self-Supervised Knowledge Transfer (SSKT) framework by Cao et al. [71]. Although supervised solutions still offer the highest accuracy, these emerging techniques highlight promising advancements toward reducing reliance on large annotated datasets.

Overall, the progression from perspective-based regressors to spherical CNNs, structure-aware encoders, transformer-based approaches, and emerging self-supervised frameworks reflects a sustained research effort toward achieving robust, metrically consistent depth estimation from monocular panoramic images—especially at high resolutions and in complex indoor environments.

2.3 Room Layout Estimation from Panoramic Images

Room layout estimation is a fundamental component of panoramic indoor modeling, as it provides the structural scaffold of the environment—typically consisting of walls, floor, ceiling, and openings—separating permanent architectural elements from movable objects or clutter [9]. A single panoramic observation captures an entire

room from a fixed viewpoint, making layout prediction particularly attractive for compact indoor spaces where a single depth or layout map can support tasks such as scene emptying, refurnishing, or virtual staging. In more general scenarios, however, recovering a clean architectural layout requires disentangling structural geometry from occlusions and self-occluding furniture, which is especially challenging in indoor environments.

Pixel-wise layout mapping In settings with limited or no viewpoint displacement, several methods estimate layout masks directly at the pixel level, typically using depth-guided segmentation to partition the equirectangular image into cluttered and uncluttered regions [21, 72, 73, 74, 10]. These networks often follow a simple encoder–decoder design, such as U-Net. Jin et al. [21], for example, learn layout masks jointly with depth estimation by leveraging datasets containing paired furnished and empty room captures. Although effective for clean, single-room scenes, pixel-wise approaches generally struggle with heavy occlusions and cannot infer unobserved surfaces.

Indoor-structured layout Full layout estimation goes beyond per-pixel prediction and requires reasoning about occluded, non-visible surfaces and long-range structural relationships. Panoramic scenes commonly contain significant occlusions from furniture or concave shapes, leaving large regions of the true structure unseen. Consequently, layout estimation must incorporate strong priors and contextual reasoning. Structured indoor reconstruction methods [75, 9] exploit the global context of a panorama [52], where depth estimation and layout prediction form the two core steps. The layout provides the room’s structural scaffold, while depth contributes local object geometry. Most panoramic layout research thus focuses primarily on these two tasks.

Geometric reasoning approaches Early panoramic layout methods relied on strong geometric priors about man-made indoor structures. Hedau et al. [3] introduced a *cuboid prior* for room labeling, and Lee et al. [4] leveraged Manhattan-World constraints using detected corners. Building on this, Zhang et al. [52] were among the first to exploit the global context of full panoramas, mapping entire rooms to cuboid models by combining Orientation Maps (OM) for the upper hemisphere and geometric context (GC) [76] for the lower part. Later, Xu et al. [77] generalized these ideas to the Indoor World Model (IWM), and Yang et al. [78] extended them using partially oriented super-pixels. While geometric approaches are interpretable and efficient, they depend heavily on reliable extraction of low-level features (e.g.,

vanishing lines, corners), which is brittle in cluttered or irregular rooms. Their assumptions—predominantly Manhattan or Atlanta Worlds—also restrict generalization and inherently limit scalability to complex or multi-room layouts.

Hybrid data-driven approaches As summarized by Zou et al. [38], many learning-based layout pipelines retain the overall structure of geometric methods: (i) Manhattan-world alignment, often following [52]; (ii) prediction of layout elements (corners, boundary maps, or horizon curves); and (iii) post-processing to enforce structural regularity. Representative methods include *LayoutNet* [79], which predicts corner and boundary maps directly from panoramas, and *HorizonNet* [80], which models the layout as three 1D vectors—the column-wise floor-wall, ceiling-wall, and wall-wall boundaries. *DuLaNet* [81] combines features from the equirectangular domain and a ceiling-plane projection to generate a floorplan probability map, regularized under MWM constraints. Recent extensions, such as *Led²Net* [82], incorporate horizon depth cues for improved accuracy, and works like Zeng et al. [83] integrate semantics and depth to stabilize predictions. These methods often achieve excellent performance in single-room, low-clutter scenarios but remain limited by strong Manhattan/IWM assumptions and reliance on post-processing heuristics, which make them less suitable for multi-room or highly irregular spaces.

Geometry-aware data-driven approaches To overcome the rigidity of Manhattan-world priors, more flexible geometric formulations have been proposed. *AtlantaNet* [84] relaxes strict MWM/IWM constraints by adopting the Atlanta World Model (AWM), leveraging gravity alignment—common in most panoramic datasets and using floor/ceiling projections. This approach better accommodates slanted or non-orthogonal walls. Graph-based solutions such as *Deep3DLayout* [1] go further by predicting full 3D watertight meshes of room layouts, freeing the model from predefined priors. Though expressive, these representations complicate the extraction of structured layouts (walls, floor, ceiling) directly from the mesh. Semi-supervised approaches like *SSLayout360* [85] demonstrate that strong performance can be achieved with fewer labels, addressing challenges of annotated panoramic datasets.

Transformers with geometry awareness Transformer-based architectures have recently advanced panoramic layout estimation by capturing long-range dependencies and global structure. *LGT-Net* [86] uses a geometry-aware transformer operating on a 1D horizontal sequence, effectively modeling extended layouts under Manhattan assumptions. *DMH-Net* [87] converts panoramas to cubemaps and reasons in a learnable Hough space to assemble semantic lines into a 3D layout. *DOP-Net* [64]

disentangles features into independent plane streams (floor/ceiling/walls) with triple attention, while *PanelNet* [88] leverages the panorama’s continuity by modeling it as horizontal panels. To address ambiguous layouts, *Bi-Layout* [89] predicts both enclosed and extended layouts, evaluating them with a disambiguation metric. More recently, *Seg2Reg* [90] aligns dense 2D segmentation with 1D layout regression using an occlusion-aware flattened volume rendering module, reducing the mismatch between training and evaluation. *HUSH* [91] instead uses spherical harmonics as basis functions for geometry-aware attention, supporting multi-task outputs for depth, normals, and layout.

Limitations and scalability Despite significant progress, most panoramic layout estimation methods rely on assumptions suited to single-room, low-clutter environments. Manhattan-/IWM-based constraints, reliance on strong geometric features, or explicit 1D/2D layout encodings restrict their ability to model: - irregular architectures, - open or multi-room spaces, - complex connectivity and occlusions, and - partial structural observations.

These limitations motivate the transition from single-room layout methods to multi-room reconstruction frameworks, where depth, layout, and multi-view structural reasoning jointly support scalable indoor modeling.

2.4 Multi-Room Modeling and Floorplans

While single-room depth and layout estimation provide essential building blocks, many practical applications require structured models that span entire apartments, office floors, or complex multi-room environments. Multi-room modeling goes beyond recovering the geometry of individual spaces and instead aims to assemble a coherent global floorplan, along with the connectivity structure that describes how rooms relate to each other and how users can navigate across them.

Traditional multi-view reconstruction pipelines, such as Structure-from-Motion (SfM) or RGB-D SLAM, typically rely on dense image collections or depth scans with significant viewpoint overlap to establish reliable correspondences, recover camera poses, and fuse observations into complete 3D models. These assumptions are fundamentally different from the sparse panoramic scenario, where only one panorama per room is commonly available and visual overlap across rooms is minimal or entirely absent. In such conditions, conventional SfM pipelines break down [92], since feature-based matching cannot be applied. Consequently, multi-room reconstruction from panoramas must rely on *single-view-centric* strategies, reconstructing each panorama

independently and enforcing global consistency at a higher structural level.

A common solution is therefore to estimate the layout of each room, detect structural openings (e.g., doors, windows), and integrate this information into a topological graph that links adjacent panoramas. This graph captures the navigable connectivity of the environment and allows multiple single-room reconstructions to be assembled into a coherent global floorplan. Such representations support not only visualization and analysis but also immersive exploration, enabling VR navigation through the reconstructed multi-room scene.

The fundamental challenge is that a single panorama is strictly limited by the visibility from its capture location. Inferring the existence and arrangement of multiple adjacent rooms requires going beyond local geometry, combining cues such as openings, vanishing lines, or occlusion boundaries with structural priors and learned patterns of indoor organization. Elements like doors or corridors can serve as weak adjacency cues, but they are often cluttered, occluded, or only partially visible. Robust multi-room modeling must therefore integrate local geometric evidence with global reasoning strategies capable of hypothesizing plausible inter-room connections under incomplete or noisy observations.

Early work framed multi-room reconstruction as an explicit geometric reasoning or energy-minimization problem [8, 93, 27, 94, 75], but such pipelines were brittle and sensitive to clutter, noise, and missing data—commonplace in real-world panoramic captures. Recent research organizes the problem into two complementary tasks: *positioning single rooms* and *connecting single rooms*. The former focuses on estimating the relative placement of panoramas based on coarse geometric priors, while the latter requires reconstructing structural elements (layout, openings, room boundaries) to ensure that the individual rooms assemble into a globally coherent floorplan.

Positioning single rooms Spatial registration across panoramas is an essential step for multi-room modeling under sparse-view conditions. Shabani et al. [92] estimate plausible relative room displacements by departing from classical SfM and instead leveraging geometric cues such as predicted single-room layouts (see [section 2.3](#)), openings, and architectural priors. Candidate inter-room connections are then validated and merged into a global optimization framework that refines both the positions and adjacencies of rooms. More recent methods incorporate learned verification and generative refinement, including diffusion-based strategies [95]. SALVe [96] uses a semantic alignment verifier that exploits windows, doors, and openings as robust cues for extreme wide-baseline alignment, while BADGR [97]

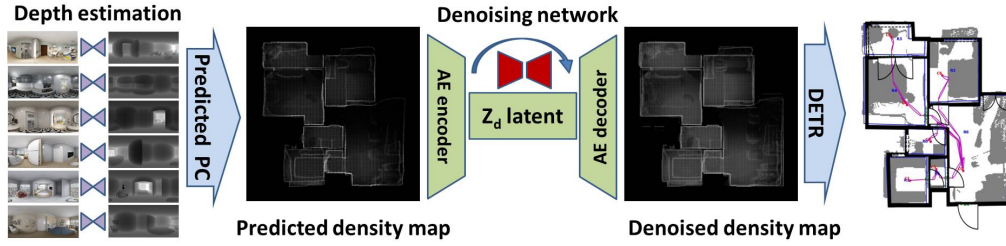


Figure 2.4: **Floorplan layout and connectivity graph recovery example** [10]. Using relative panorama positions and per-view depth, the system reconstructs a global floorplan enriched with semantic elements (layouts, openings, clutter) and forms a connectivity graph describing navigable structure.

proposes a diffusion-based bundle adjustment layer that jointly refines camera poses and floorplan geometry using structural priors such as collinearity and wall adjacency.

These works demonstrate a shift from pure geometry toward hybrid pipelines where local cues, semantics, and generative models collectively support robust positioning under sparse panoramic sampling.

Connecting single rooms Connecting multiple panoramas requires more than just estimating their relative positions; it demands some form of floorplan reconstruction to provide the structural scaffold linking individual rooms. When dense 3D information is available, such as point clouds or RGB-D scans, several works detect low-level primitives (corners, edges, room regions) and assemble them into structured floorplans through optimization. Examples include Floor-SP [98], MonteFloor [99], and FloorNet [100]. Transformer-based architectures further extend this direction: HEAT [101], RoomFormer [17], SLIBO-Net [102], and PolyDiffuse [22] show that polygonal floorplans can be robustly reconstructed from dense or semi-dense data using learned structural priors.

In the context of sparse single-view-per-room panoramas, however, dense point clouds are unavailable and conventional density-map estimation from 3D projections becomes highly unreliable [103]. This has motivated approaches that extend floorplan estimation to the extreme sparse regime. **NadirFloorNet** [103], presented in chapter 5, adapts the transformer-based *RoomFormer* architecture to the single-image-per-room setting by first predicting a clutter-free nadir depth projection and then fusing multi-room cues via deformable attention informed by panorama positions [92, 96, 97]. By jointly estimating ceiling heights and 2D polygons, it provides a metrically coherent multi-room reconstruction from extremely sparse inputs.

PanoFloor [10], described in chapter 6, takes a complementary approach: start-

ing from relative panoramic positions and per-room predicted depths (see [section 2.2](#)), it reconstructs a global floorplan enriched with structural semantics—layouts, doors, windows, and clutter—and builds a connectivity graph describing the navigable topology of the environment (see [Figure 2.4](#)). This representation supports immersive multi-room exploration and provides a VR-ready model that integrates geometry, semantics, and adjacency.

Overall, existing work on multi-room reconstruction illustrates a progression from rigid geometric heuristics to hybrid data-driven and generative pipelines. The contributions presented in this thesis—*NadirFloorNet* for transformer-based multi-room layout reconstruction and *PanoFloor* for connectivity-aware immersive modeling—represent key steps toward scalable, structured modeling of large indoor environments from sparse panoramic imagery.

2.5 Immersive Visualization and VR for Indoor Scenes

Immersive visualization is a central goal of modern indoor reconstruction pipelines, as it enables users to explore real environments in VR with convincing depth cues, motion parallax, and stereoscopic perception. While a single panoramic image naturally captures the full field of view, it lacks the binocular disparity and translational parallax needed for realistic immersion [104]. For this reason, recent research focuses on augmenting panoramas with geometric structure—typically depth, layout, or full 3D proxies—to generate view-consistent renderings that can be deployed in VR, WebXR, and interactive navigation systems.

A foundational component of immersive visualization is *novel-view synthesis* from a single panorama. Early methods used depth-based reprojection [105, 30, 106], often followed by inpainting [107, 108] to fill disocclusions. More advanced techniques design end-to-end networks to generate shifted panoramas at runtime [72, 19] (see [Figure 2.5](#)), exploiting domain-specific priors such as room layout and indoor structure. These solutions enable limited viewpoint translation for headset users, supporting natural head motion while remaining compatible with real-time VR rendering constraints.

To extend immersion beyond small parallax, several works introduce *precomputed 3D representations* derived from a single panorama. Layered structures such as MPIs [109] and their panoramic variants MSI/MCI [110, 104] offer efficient rendering at the cost of blurring in disoccluded regions. Neural representations such as NeRF [111] and 3D Gaussian Splatting [112] provide higher fidelity by optimizing implicit 3D models from synthesized RGB-D views [113, 114, 115], though typically

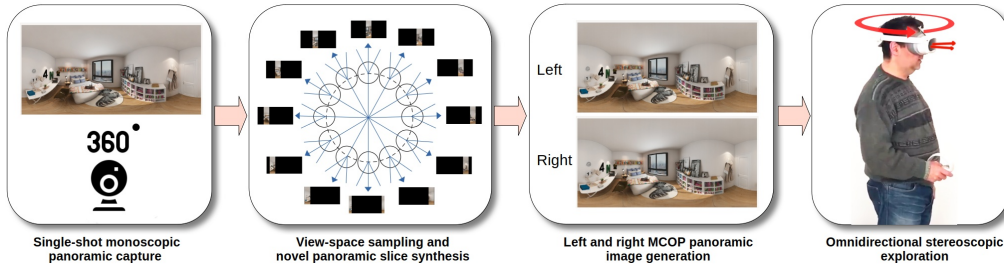


Figure 2.5: **Synthesis of omnidirectional stereo pairs from a single panorama** Slices generated by a view-synthesis network are composed in a single multiple-center-of-projection equirectangular image per eye [11].

with higher memory or computation requirements.

Stereo visualization is particularly important indoors due to close-range objects. Omnidirectional stereo projections [116, 117] capture depth cues in two equirectangular images, while deep approaches generate multi-center-of-projection panoramas directly from a single input [11]. Complementary strategies precompute sets of synthetic views along circular eye paths [118], enabling smooth stereo playback on VR headsets with minimal runtime cost.

Beyond single-room immersion, interactive systems must support *multi-room navigation*. Classical methods interpolate between panoramas via warping or depth-aware blending [105, 119], but struggle under wide baselines and heavy occlusions typical of indoor spaces. Modern pipelines instead integrate geometric priors—e.g., room layouts, door locations, and obstacle awareness—to compute collision-free paths [120, 121] or to synthesize transitions with improved structural coherence [122]. Aesthetic metrics can be used to guide trajectory selection and viewpoint recommendation [106, 123, 124], enhancing perceived quality during exploration.

Overall, immersive visualization from sparse panoramic input requires structured, view-consistent reconstructions. Depth, layout, and global floorplan reasoning provide the geometric scaffolding needed to support stereo, parallax, and multi-room transitions, while neural synthesis and compact 3D representations enable high-quality rendering compatible with VR and WebXR deployment. These elements jointly define modern pipelines for realistic, interactive exploration of indoor environments.

2.6 Summary and Open Challenges

The material reviewed in this chapter highlights how panoramic indoor understanding has evolved from purely geometric pipelines to deep learning methods that exploit architectural priors, gravity alignment, and 360°-specific representations. We have discussed how equirectangular projections and indoor priors constrain the recon-

struction problem, how monocular depth estimation and room layout prediction form the core single-room tasks, how these ingredients can be lifted to multi-room modeling and floorplan estimation under sparse capture, and how immersive visualization pipelines transform structured reconstructions into VR-ready, view-consistent experiences for single- and multi-room environments.

Despite significant progress, several limitations remain. At the single-room level, current depth estimators still struggle to deliver robust, metrically consistent predictions in the presence of clutter, severe occlusions, and complex geometries, and often under-utilize structure-aware constraints tied to indoor organization. Layout estimation, while accurate under Manhattan or Indoor-World assumptions and moderate clutter, degrades in highly irregular or heavily furnished spaces, and most methods are designed for isolated rooms rather than extended apartments or building-scale environments. Multi-room reconstruction methods, in turn, typically assume dense point clouds, RGB-D scans, or richly overlapping multi-view imagery, making them difficult to apply in the sparse panoramic regime that characterizes many practical acquisition scenarios. Finally, immersive visualization and VR pipelines still depend on reconstructions that are both structurally coherent and view-consistent, a requirement that becomes particularly challenging when only one or a few panoramas per room are available.

These gaps motivate the contributions developed in this thesis. First, we address the challenge of estimating depth and geometry in a *structure-aware* manner, explicitly exploiting indoor regularities and density-based constraints to improve robustness and metric reliability from single panoramic views (Chapter 4). Second, we tackle the problem of reconstructing complex multi-room environments from sparsely sampled panoramas, designing transformer-based and layout-centric models that assemble coherent floorplans and connectivity structures even in extreme sparse-view conditions (Chapter 5). Third, we focus on building immersive, explorable models for multi-room indoor scenes, coupling geometric and semantic reconstructions with navigation graphs and VR-oriented representations to support realistic, interactive exploration from sparse panoramic inputs (Chapter 6). Together, these contributions aim to bridge the gap between single-view panoramic analysis and scalable, immersive modeling of large, real-world indoor environments.

3 Methodological Framework

This chapter outlines the methodological framework that underpins the research conducted in this thesis. Our approach is grounded in modern *data-driven* paradigms and follows primarily a *supervised learning* methodology. This choice is motivated by the current landscape of panoramic indoor understanding: supervised models consistently deliver the highest levels of accuracy in depth estimation, layout prediction, and multi-room reconstruction, especially in the presence of clutter, complex geometry, or challenging viewpoints.

While alternative frameworks such as self-supervised or unsupervised learning offer several advantages—most notably reduced reliance on ground-truth annotations—they remain comparatively less mature for 360° indoor scene understanding. Their performance is currently limited by difficulties in enforcing spherical consistency, managing large occlusions, and maintaining stable learning dynamics under extreme distortions. For these reasons, the core contributions of this thesis (Chapters 4, 5, and 6) are primarily developed within a supervised framework. Nonetheless, the methodology presented here remains compatible with future extensions toward weak or self-supervised regimes.

3.1 Data-Driven Learning for Panoramic Indoor Understanding

Given the centrality of training data in supervised paradigms, our methodological choices are shaped by two considerations:

1. **The need for large-scale indoor datasets** containing panoramas, depth maps, layout annotations, and multi-room structural information.
2. **The importance of training/validation protocols** that ensure comparability with existing methods and allow controlled evaluation of generalization capabilities across scenes of varying complexity.

Large-scale datasets play a critical role in the training and benchmarking of panoramic learning models. They enable supervised training on diverse indoor layouts, viewpoint configurations, lighting conditions, and clutter levels, providing models with the variability necessary for robust generalization. Furthermore, the widespread use of common datasets in the literature allows direct and rigorous comparison with state-of-the-art methods.

In this thesis, we leverage several publicly available indoor datasets that include equirectangular panoramas, depth annotations, layout information, and floorplan-level structure. These datasets support the three main tasks addressed in the thesis—dense depth estimation, single-room layout prediction, and multi-room floorplan reconstruction. A detailed discussion of the datasets used in this work is provided later in this chapter.

3.2 Supervised Learning Paradigm

All the proposed models adopt a supervised training scheme, learning a mapping from panoramic images (and additional structured inputs, when available) to target signals such as depth maps, room layouts, or floorplan polygons. Supervision enables:

- precise, metric-level learning of geometric quantities, essential for reliable multi-room modeling;
- controlled optimization through established loss functions (e.g., L_1 , BerHu, ordinal regression, polygonal consistency losses);
- direct benchmarking against state-of-the-art baselines under identical training conditions;
- seamless integration of architectural and geometric priors within the learning objectives.

Across the methods introduced in this thesis, we combine pixel-wise objectives (depth, normals, masks), structural objectives (layouts, polygons), and consistency-driven objectives such as density-map coherence. The detailed formulation of the losses used in each contribution is presented in the corresponding chapters—see Chapter 4 for depth estimation, Chapter 5 for layout and multi-room reasoning, and Chapter 6 for immersive modeling and VR transitions.

3.3 Training Strategy and Evaluation Protocols

All experiments follow the standard dataset splits defined by the respective benchmarks, ensuring comparability with the current state of the art. Depending on the task, evaluation employs widely adopted quantitative metrics: RMSE, δ -accuracy thresholds, and depth error histograms for depth estimation; IoU and 3D IoU for single-room layout prediction; polygon accuracy, topological consistency, and metric floorplan alignment for multi-room reconstruction.

In addition to these quantitative measures, qualitative assessment remains crucial. Evaluations include visual inspection of geometric coherence, robustness to clutter and occlusions, stability of multi-room structural reasoning, and the perceptual fidelity of rendered VR representations. The detailed evaluation procedures, together with task-specific metrics and ablation studies, are provided in the technical chapters—Chapter 4, Chapter 5, and Chapter 6.

3.4 Datasets Related to This Thesis

The development of supervised, data-driven methods for panoramic indoor reconstruction relies fundamentally on the availability of large-scale, high-quality datasets containing 360° panoramas and their associated geometric and semantic annotations. Public datasets enable fair benchmarking, controlled experimentation, and direct comparison with state-of-the-art approaches. Moreover, they provide the diverse training samples necessary for models to learn robust structural and geometric cues under varying room types, clutter levels, and capture conditions.

In this thesis, we make extensive use of publicly released datasets that include equirectangular panoramas, depth maps, layout annotations, and multi-room structural metadata. These datasets reflect different acquisition modalities—ranging from real camera captures to photorealistic synthetic environments—each contributing complementary strengths. Real datasets provide visual realism and naturally occurring clutter, whereas synthetic datasets offer complete and noise-free ground truth with perfect floorplans, pixel-accurate geometry, and consistent semantics.

Below, we summarize the key datasets relevant to the tasks tackled in this thesis.

Matterport3D [14] A large-scale RGB-D dataset containing over 10k panoramas from real residential and commercial buildings. It provides textured meshes, camera poses, region-level semantic labels, and object instances. Depth annotations are commonly derived from reconstructed meshes. It supports tasks such as depth

estimation, semantic segmentation, and sparse multi-view reasoning.

MatterportLayout [125, 38] An extension of Matterport3D including Manhattan-aligned room layout annotations for 2,295 panoramas. It is widely used for benchmarking learning-based layout estimation methods.

Stanford2D-3D-S [126] A large RGB-D dataset of public indoor spaces (e.g., offices, hallways, classrooms) with 552 equirectangular images, depth, normals, semantics, and camera poses. Includes cuboid layouts and aligned 3D meshes.

GibsonV2 [127] A very large mesh-based indoor environment collection with 45k+ panoramas across diverse room types. Depth maps can be generated via ray casting. Useful for training geometric models and evaluating cross-scene generalization.

Pano3D [128, 129] This synthetic dataset targets depth and normal estimation and cross-dataset evaluation. It re-renders 21,203 scenes from *Matterport3D* point clouds [14] and *GibsonV2* meshes [127, 130] 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.

360MonoDepth/M3D [131, 33] A high-resolution extension of Matterport3D with pixel-accurate depth aligned via spherical structure-from-motion. Contains nearly 10k panoramas at up to 1024×2048 resolution. Crucial for supervised panoramic depth estimation.

360MonoDepth/Replica [132] A synthetic high-resolution dataset of 130 photo-realistic panoramas from the Replica scenes, co-registered with ground-truth depth up to 4K resolution. Used mainly for capacity testing and cross-domain generalization.

PanoContext [133] Contains 700 panoramas from the SUN360 collection annotated with room-level cuboid layouts and simplified object geometry. Often used as an additional layout prediction benchmark.

Zillow Indoor Dataset (ZInD) [134] A large-scale real-estate dataset with 67k panoramas captured in unfurnished homes. It includes layout annotations,

door/window positions, and adjacency metadata. Ideal for sparsely captured multi-room reasoning.

Structured3D [135] A massive synthetic dataset with 21k+ panoramas, depth, normals, semantics, and detailed layout information across thousands of architecturally diverse homes in the Atlanta world configuration. A common training resource for depth, layout, and geometric reasoning.

PNVS [136] A synthetic dataset designed for view synthesis with paired panoramas rendered at small and large camera displacements. A key resource for exploring parallax, occlusions, and novel-view synthesis from a single panorama.

SKI360 [13] A Structured3D-based dataset containing paired furnished/unfurnished rooms with layout, depth, and furniture masks. Used to separate architectural structure from movable clutter.

Across these collections, several aspects are especially relevant for methodological design:

Gravity alignment Nearly all datasets provide upright panoramas or metadata to recover alignment, enabling gravity-aware models.

Resolution constraints High-resolution 360° panoramas remain rare, and when available, only synthetic datasets typically provide ground-truth depth at native resolution.

Limited scale of curated datasets Although millions of real estate panoramas exist, annotated research datasets remain relatively small, motivating the need for mixed-dataset training, synthetic augmentation, and semi-supervised strategies.

These datasets collectively support the supervised, data-driven framework adopted in this thesis, enabling rigorous training for depth estimation, layout prediction, multi-room reconstruction, and immersive visualization across a wide variety of real and synthetic environments.

4 Deep Depth Estimation from Panoramic Images

4.1 Problem Definition and Motivation

The automatic 3D modeling of indoor scenes has gained significant research attention in recent years, emerging as a well-defined 3D subfield [9]. Specialized techniques for common, highly structured environments such as residential, office, and public buildings are one of the main targets. Such buildings constitute the majority of the built environment, and 3D reality-based models are required for many purposes [75, 137]. Moreover, the standardized construction methods offer degrees of similarity that can be exploited for improving reconstruction from noisy and partial input [9, 138].

Fast depth estimation from images is a fundamental sub-problem in this context since associating metric information with visual data is necessary for 3D reconstruction, and rapid solutions open the door to many applications, including mobile extended reality, indoor mapping, and autonomous navigation. Although traditional methods have utilized the correlation among multiple views captured simultaneously (e.g., stereo) or sequentially over time (e.g., video), the interest in monocular 360° depth estimation is growing [138].

A 360° image, quickly and easily acquired with affordable cameras, captures the full scene from a single viewpoint, providing rich context for depth inference and scene understanding [139]. Monocular depth estimation remains, however, very challenging in furnished indoor environments. Even though structure priors characterize the architectural shape that bounds the scene, it is often hard to recognize them, since walls are often composed of large untextured regions, and objects can be cluttered and arranged arbitrarily in the near field, masking large portions of a room’s walls and floors.

These challenges have led to the introduction of indoor-specific 360° solutions that have reached impressive results, especially in conjunction with supervised deep-learning approaches that learn hidden relations from large sets of examples [11].

Although such state-of-the-art approaches can predict high-detail depth maps with good accuracy at the pixel level, the salient features of an indoor environment, such as wall planarity and edge sharpness [1], as well as the regularity and consistency of the architectural man-made structures [99, 15] are less well preserved. Such consistency becomes of critical importance when depth is used for room layout reconstruction, exploited for immersive exploration [140, 19, 11], or to merge and analyze multiple single-view reconstructions for structural segmentation of complex multi-room environments [101, 17]. Moreover, it is not uncommon, see [section 4.5](#), to require in the order of hundreds of millions of tunable parameters and over hundreds of GFLOPs to infer depth from a 512×1024 image. Such high memory and computational costs make it difficult to use them for large images or low-latency depth generation [19].

In our work we propose a lightweight end-to-end deep learning approach, dubbed *DDD++*, for depth estimation from a single 360° image in an equirectangular format.

We exploit the characteristics of indoor environments to reduce computational costs and both structural consistency and depth accuracy. To design our network, see [section 4.3](#) and [Figure 4.1](#), we start from the assumption that, due to gravity, world-space vertical and horizontal features have different characteristics in most, if not all, man-made environments. These characteristics are preserved in gravity-aligned or gravity-rectified images [15, 60]. To this end, we perform a contractive encoding to reduce the input equirectangular tensor only along the vertical direction to obtain a compact and flattened sequence of slices made of a set of Gravity-Aligned Features (GAFs). To preserve global information, we perform slicing over different resolution levels, concatenating the result at the end. In addition to optimizing the flow of information contained in features, as done in previous works [15, 60], this representation allows in our design subsequent processing directly through a vision transformer, which takes into account the spherical nature of the input and recovers long- and short-term spatial relationships among features.

To optimize the depth map in terms of its consistency when interpreted as a sampling of a 3D architectural environment, the network is trained through a novel indoor-specific metric and loss function (see [section 4.4](#)). We do that by transforming depths into density maps computed using planar and cylindrical projections and comparing predicted with ground truth ones. These maps, which accumulate the occurrence of 3D points derived from depths and projected onto the floorplan and on two planes orthogonal to it, are known to provide good summaries of the characteristics of indoor environments characterized by vertical walls [101, 17].

The overall approach, combines a network design that enables, through feature

flattening, the direct use of a vision transformer, without the need to sequence the input map by arbitrary patches and positional encoding [62], with the idea to supervise training by minimizing the error on density maps, leading to a better identification and preservation of permanent indoor structures. As a result, the method’s lightweight architecture has a low computational impact and provides greater structural consistency than other current approaches (see section 4.5). Such a lean network can be integrated as a component in multi-stage pipelines, for instance, for multi-room reconstruction (e.g., [17]) or view translation and synthesis for immersive applications [19, 11]. Its fast inference time also makes it ideal for real-time usage.

Furthermore, we devised an improved loss function, which now includes a novel cylindrical, full-view projection to better account for the geometric information available from the spherical view respective to the original proposal of Manhattan-aligned cube maps (see section 4.3); we discuss results of new experiments with commonly used synthetic large-scale datasets and benchmarks where ground truth is fully available [16] and compare them with results obtained by state-of-the-art for indoor panoramic depth estimation [64, 65, 20].

4.2 Task-related works

Depth estimation from monocular input and 3D reconstruction of indoor environments are fundamental computer vision problems, which have recently attracted renewed interest with the emergence of deep learning techniques. A full review is beyond the scope of this chapter, and we refer the reader to established surveys for wider coverage [9, 141, 138]. Here, we focus on the solutions most closely related to our work.

Depth from perspective images. Data-driven monocular depth estimation was introduced over a decade ago (e.g., Make3D [41]). The emergence of deep learning and the availability of large-scale 3D datasets have led to significant performance improvements. After the introduction of CNNs for regressing dense depth maps from a single image [42, 40], Laina et al. [43] introduced the now standard *FCRN* encoder-decoder architecture, combining *ResNet* [44] for the encoding and an up-projection module for decoding and the reverse Huber loss [45] to improve depth estimation. Following these trends, many solutions have been further introduced, including predicting depth from several cropped images combined in the Fourier domain [46], using an ordinal regression loss to preserve the spatial relation among neighboring classes [47], exploiting Conditional random fields (CRF) to refine predictions [48, 49,

[50, 51], and many more follow-ups. However, directly applying perspective methods to 360° images, does not permit the full exploitation of their characteristics, and in particular, their global context, leading to sub-optimal results [52, 53]. As a result, much of the research on reconstruction of indoors from sparse imagery is now focused on 360°-specific solutions.

Depth from a single omnidirectional image. Several solutions adapted perspective established methods to 360° depth prediction by using projections into a cube map [54] or by replacing regular convolutions with spherical convolutions to cope with distortions [55, 56, 57, 53, 58]. Wang et al. [59] combined the approaches through a two-branch network, respectively for the equirectangular and the cube map projection, based on a distortion-aware encoder [53] and the FCRN decoder [43]. Several recent methods leverage perspective views sampled on panoramic images [63, 33] before combining depth maps using patch-based vision transformers [64, 65, 20]. Building on this trend, Wang et al. [70] propose Depth Anywhere, a framework that enhances 360° monocular depth estimation by distilling knowledge from a pretrained perspective depth estimator into a panoramic model. Complementary to this, Cao et al. [142] address the challenge of high-resolution 360° depth estimation without high-res depth ground truth by introducing a weakly-supervised framework that transfers structural knowledge through a Scene Structural Knowledge Transfer (SSKT) module. Another breed of solutions for panoramic depth estimation in indoor spaces [60, 15] proposes, instead, to work directly on the equirectangular images produced by spherical cameras and to leverage the concept of gravity-aligned features to reduce network size while supporting the exploitation of short- and long-range relations. In this work, we show how to directly use gravity-aligned features [15] to feed a self-attention vision transformer, without the need to arbitrarily partition the image into patches [64, 65, 20]. Moreover, we incorporate the concept of density maps, as used in reconstruction and segmentation tasks [101, 17], into 360° depth prediction to define a structural loss that enhances the accuracy and consistency of depth predictions with architectural structures in indoor models. We exploit here both a cylindrical and a planar projection to recover structural information. As a result, we achieve state-of-the-art performance at a lower inference cost than previous solutions.

4.3 Network architecture

Our network takes as input a 360° gravity-aligned image in equirectangular format and produces as output its per-pixel depth. Assuming gravity-alignment allows us to

design a particularly efficient solution, while not limiting the domain of application of the method. Gravity-aligned capture is very common, and nearly all public 3D indoor datasets commonly used for training and testing reconstruction solutions exhibit minimal orientation deviations. [15, 60]. This is because maintaining the upright position for capturing, besides being natural for free-form single-shot images, is usually enforced by exploiting data from the IMUs present in most modern capture devices or by mechanical setups such as tripods. Moreover, even in the few cases where these assumptions are not verified at capture time, many orthogonal and fast solutions can be applied to gravity-rectify images in a preprocessing step to connect the direct output from the capture device to our depth estimation network (e.g., [35, 36, 34]).

Our lightweight network architecture for depth estimation in indoor environments combines gravity-aligned features obtained by asymmetric convolution of the input with multi-head self-attention. The structure of our network is depicted in Figure 4.1.

From the input image, a cascade of five residual layers [44] returns four feature maps having different depths and spatial sizes. Given the spherical nature of the image, we also adopt circular padding along the horizon for convolutions, to overcome the longitudinal boundary discontinuity, and reflection padding to alleviate the singularities at the poles [73].

To support an efficient gathering of information from the extracted features, we perform a specifically indoor-designed feature compression exploiting our knowledge of preferential directions, based on the fact that gravity-aligned images preserve the fact that world-space vertical and horizontal features have different characteristics in most, if not all, man-made environments [60, 80, 15, 1]. For instance, it is fairly natural, if only for physical reasons, to have horizontal planes both in architectural (e.g., floors) and impermanent (e.g., tabletops) structures, as well as vertical ones (e.g., walls and supporting parts of furniture). Exploiting this assumption, we perform an *anisotropic contractive encoding* that reduces the vertical direction while keeping the horizontal direction unchanged, so that separated vertical features can be better preserved. Specifically, we reduce the vertical dimension by a factor of 8 through an asymmetric convolution module with stride (2, 1), applied three times, that contains a 2D convolution and an ELU module. We apply such compression for each encoded feature map (i.e., four maps), obtaining a set of latent features $L_s = (l_1 \dots l_4)$. Compressed features L_s are reshaped to the same size and joined in a flattened latent feature, as a single sequence of s feature vectors of dimension l (i.e., s horizontal size of the less deep feature map - $s = 1024$ and $l = 256$ for a 512×1024 input). Such a compressed representation contains a wealth of information about

the scene’s local and global geometry, which can be exploited to recover depth and layout and provide a latent representation of the scene.

Note that our flattening of gravity-aligned features constructs a structured linear sequence that can be directly used as input to a self-attention-based vision transformer. This design bypasses the need for arbitrary image patching or complex positional encodings, as commonly required in transformer-based architectures [64, 65]. By leveraging the inherent alignment with the gravity direction, our representation preserves the spatial coherence of the scene, particularly the vertical semantic structure typical of indoor environments. To exploit long-range dependencies in this structured representation, we integrate a single-layer multi-head self-attention (MHSA) module [143]. This module effectively captures contextual relationships across distant regions of the image—an especially valuable property for omnidirectional imagery, where objects and structural cues may span wide angular fields.

Crucially, our approach avoids the need for recurrent modules, which are commonly used in prior GAF-based methods such as SliceNet [15] to model sequential dependencies. While effective, recurrent neural networks (RNNs) introduce significant computational overhead and latency due to their inherently sequential nature.

In contrast, our MHSA implementation processes all tokens in parallel and maintains a much lower computational footprint, enabling faster training and inference without sacrificing the ability to reason over global context. This design choice not only improves scalability but also enhances the model’s capacity to integrate complementary information from spatially distant but semantically related regions, making it particularly well-suited for the rich geometric structure captured in panoramic indoor scenes.

Our self-attention module takes the latent features $L \in \mathbb{R}^{s \times l}$ as input, and outputs a self-attention weight matrix $A \in \mathbb{R}^{s \times s}$:

$$A = \text{softmax} \left(\frac{(LW_q)(LW_k)^T}{\sqrt{l}} \right) \quad (4.1)$$

where $W_q, W_k \in \mathbb{R}^{l \times l}$ are learnable weights. The MHSA module has a particularly lightweight design with four heads and only one inner layer. We have verified experimentally that increasing the number of layers and heads heavily increases the number of parameters and computational load without significantly improving reconstruction accuracy. Once passed to the MHSA module, the decoding of the latent feature ($1 \times 1 \times s$) is very fast, through convolutions, upsampling modules, and ELU activations, until we reach the target output resolution ($1 \times h \times w$).

4.4 Indoor-specific loss function and training strategy

To train our network, we designed a loss function that is a combination of a conventional equirectangular loss term ($\mathcal{L}_{eq}$) with a novel, structure-driven component ($\mathcal{L}_{ds}$), i.e., $\mathcal{L} = \mathcal{L}_{eq} + \mathcal{L}_{ds}$.

The equirectangular loss term $\mathcal{L}_{eq}$ penalizes per-pixel deviations of the inferred depth from the ground truth value. As common for depth estimation frameworks, we build it on top of the robust *Adaptive Reverse Huber Loss* (*BerHu*) [45]:

$$H(e) = \begin{cases} |e| & |e| \leq c \\ \frac{e^2 + c^2}{2c} & |e| > c \end{cases} \quad (4.2)$$

where e is the error term and the parameter c determines where to switch from L1 to L2. To set the c value adaptively, we follow the approach originally introduced by Laina et al. [43], so that c is set, in every gradient step, to 20% of the maximal error of the current batch. When applied to the depth maps, we have $e = D_{ij} - D_{ij}^*$ at each pixel (i, j) , where D and D^* are, respectively, the predicted and the ground-truth depth maps, and, thus:

$$\mathcal{L}_{eq}(D, D^*) = \sum_{ij} H(D_{ij} - D_{ij}^*) \quad (4.3)$$

Using only this term, however, that measures, per-pixel, distances from training data, would not take into account the peculiar features of indoor environments, and especially of the architectural structures, that we expect made of large fairly regular surfaces with preferential orientations. For instance, we expect to find mostly horizontal floors and mostly vertical walls, rather than curved/wobbly surfaces, that can, instead, more commonly be found on objects.

To drive the solutions toward plausible depth reconstructions, we introduce in this work a structural term $\mathcal{L}_{ds}$, rather than using a regularization term. Using such an approach allows us to learn these regularities from data, rather than imposing them upfront through specific penalty functions.

In our original conference contribution [12], we have shown how density maps computed from point clouds can provide an adequate structural summary to be exploited for loss computation. The basic idea was to extract important features through projections on horizontal and vertical directions. This was done, however, by imposing an arbitrarily oriented Manhattan world prior (i.e., Figure 4.2c and Fig-

ure 4.2d), introducing unwanted discontinuities and wide variations in the summary characteristics, for the same environment, depending on the relative alignments of the arbitrary Manhattan axes with the main equirectangular view direction. Here, we significantly improve $\mathcal{L}_{ds}$ by introducing a loss based on a single density map computed by vertically projecting the horizontal planes and a single density map computed by horizontally projecting on a cylinder (i.e., Figure 4.2f), thereby encompassing the entire visible horizon and seamlessly capturing better and more stable structural features.

To compute the $\mathcal{L}_{ds}$ structural term, we transform D and D^* into the equivalent point clouds P_D and P_D^* in Cartesian coordinates using the spherical transformation associated to the equirectangular projection. We then scale 3D points to the same absolute scale, by setting a maximum distance from the observer (20 meters in the examples presented in this work). Assuming the gravity-vertical direction as the Z axis of our reference system, we produce predicted and ground truth density maps from P_D^* and P_D with respect to vertical and horizontal direction. Specifically, we render two density maps, Oz (i.e., vertical projection looking to the floor) and Oc (i.e., horizontal projection, looking to the walls), respectively from the from depth prediction and from the ground truth depth-point cloud. Since O represents a map of the occurrences of 3D points falling on the same pixel, the structural parts of the scenes become more evident. For instance, the vertical projection along Oz highlights the floor plan (e.g., Figure 4.2e), since the many vertically aligned points on walls in ground truth data identify room boundary locations. For this reason, such a projection is often used to automatically derive the floor plan of one or more rooms from a point cloud [98], but, to the best of our knowledge, has not been used to define indoor-specific cost functions for depth recovery. At the same time, the horizontal projection Oc (e.g., Figure 4.2f) emphasizes the shapes of horizontal planes in the scene, as well as geometric patterns distributed along the horizon.

Given a 3D point cloud $P_D = \{\mathbf{p}_i\}_{i=1}^N$, where each point $\mathbf{p}_i = (x_i, y_i, z_i)$ represents the Cartesian coordinates of the i -th point, the vertically projected density map Oz , along Z axis, is easily obtained just removing z_i and counting the occurrences of two points (x_i, y_i) on the floorplan, after normalization and rescaling to the density map size (512×512 in our experiments). For computing the horizontally projected density map, we transform 3D points into cylindrical coordinates.

Specifically, assuming the cylindrical occupancy map $O_c \in \mathbb{R}^{h \times w}$ (256×1024 in our experiments), for each point $\mathbf{p}_i = (x_i, y_i, z_i)$, we compute the azimuthal angle θ_i in the xy -plane as:

$$\theta_i = \text{atan2}(y_i, x_i).$$

The angle θ_i is normalized to the range $[0, 2\pi)$ using:

$$\theta_i = \begin{cases} \theta_i + 2\pi & \text{if } \theta_i < 0, \\ \theta_i & \text{otherwise.} \end{cases}$$

Given the minimum and maximum z -values in the point cloud as:

$$z_{\min} = \min_i z_i, \quad z_{\max} = \max_i z_i.$$

The z -coordinate is normalized to the range $[0, 1]$ using:

$$z'_i = \frac{z_i - z_{\min}}{z_{\max} - z_{\min}}.$$

The normalized cylindrical coordinates (θ_i, z'_i) are then discretized into indices (j, k) for the occupancy map:

$$j = \left\lfloor \frac{\theta_i}{2\pi} \cdot w \right\rfloor, \quad k = \lfloor z'_i \cdot h \rfloor.$$

The indices are clamped to ensure they lie within the valid range:

$$j = \text{clamp}(j, 0, w - 1), \quad k = \text{clamp}(k, 0, h - 1).$$

The occupancy map O_c is initialized as a zero matrix of size $h \times w$. For each point $\mathbf{p}_i$, the corresponding bin in O_c is incremented:

$$O_c(k, j) \leftarrow O_c(k, j) + 1.$$

The resulting cylindrical occupancy map O_c is a 2D grid where each cell (j, k) represents the number of points projected onto the corresponding cylindrical bin. The z -axis is oriented with the origin at the bottom.

Starting from the predicted and ground-truth density maps, respectively (O_z, O_c) and (O_z^*, O_c^*) , we calculate the structural loss term $\mathcal{L}_{ds}$ as the sum of the adaptive Reverse Huber loss of the individual predicted density map value relative to ground truth for each pixel (k, l) in the projections:

$$\begin{aligned} \mathcal{L}_{ds}(O_z, O_c, O_z^*, O_c^*) = & \sum_{kl} \mathcal{H}(O_{z_{kl}} - O_{z_{kl}}^*) + \\ & \sum_{kl} \mathcal{H}(O_{c_{kl}} - O_{c_{kl}}^*) \end{aligned} \quad (4.4)$$

The same parameters used for tuning Equation 4.2 for depth values are used for the density maps. In section 4.5, we show how we achieve good performance using only these data terms even without adding other regularization terms.

It should be noted that our approach does not require strict alignment of the panorama and layout to the Manhattan World axes but only needs the more common gravity alignment (see section 4.2). This allows us to limit geometric data augmentation to flips and random rotations around the Z axis during training. Furthermore, our geometric augmentation accounts for the fact that our density maps are mutually orthogonal and gravity-aligned, but arbitrarily rotated around the gravity vector. The augmentation through random rotations helps uncover hidden relationships that are independent of the view’s alignment with world-space axes.

4.5 Results

Our approach was implemented using *PyTorch* and has been tested on several kinds of indoor scenes. In the following, we discuss the datasets used in this work (subsection 4.5.1). We then briefly illustrate the training setup and the computational performance, also comparing inference times and costs to other state-of-the-art solutions (subsection 4.5.2). Finally, we discuss the quantitative and qualitative results on depth reconstruction (subsection 4.5.3), compared to the state-of-the-art.

4.5.1 Datasets

Following the latest state-of-the-art works [144, 20], we adopt for training and testing Structured3D [16] (see section 3.4). This benchmark provides full spherical coverage and provides very realistic, but artifact-free ground-truth color and depth information. This allows for a device-independent, reliable, and consistent evaluation of equirectangular depth estimation methods, free from the stitching or occlusion-related inconsistencies present in other datasets. For these reasons, Structured3D [16]) is the main benchmark used in this work.

In addition, to illustrate performance on real-world capture data, we include results obtained on Matterport3D [14]. Although Matterport3D [14] is widely adopted as a benchmark for equirectangular depth estimation, it is important to note that the dataset was originally captured in perspective format using multiple RGB-D sensors. As such, transforming the data into a consistent equirectangular representation requires a non-trivial preprocessing pipeline involving stitching and blending of multiple perspective views. This preprocessing step introduces variability in the resulting images and depth maps, which in turn affects the comparability of

results reported across different studies. Consequently, significant discrepancies can be observed among Matterport3D results in the literature, depending on the specific processing pipeline adopted. In this work, to ensure a fair and consistent comparison, we adhere to the preprocessing procedure introduced by Elite360D [20]. We also report their published results as a reference for all comparisons on this dataset.

Furthermore, to provide direct comparisons to the baseline [12] and to support the ablation study, we also discuss results obtained with the publicly available Shanghaitech-Kujiale Indoor 360° (SKI360) dataset [13]. The dataset contains 1,775 panoramic RGB images of scenes of furnished rooms accompanied by ground truth depth maps. The images are synthesized from 3D models with a photorealistic renderer based on path tracing to achieve realistic rendering [21]. As in our previous work [12], this benchmark is also used to compare our performance relative to other solutions exploiting geometric cues as priors and regularizers [21].

4.5.2 Training setup and computational performance

We trained our DDD++ network using a single NVIDIA RTX 4090 GPU equipped with 24GB of VRAM. The training process employed the Adam optimizer with default momentum parameters, specifically $\beta_1 = 0.9$ and $\beta_2 = 0.999$, which are well-suited for stabilizing convergence in dense prediction tasks. We initialized the learning rate at 1×10^{-4} and employed an adaptive learning rate schedule that decayed based on validation loss plateaus, promoting efficient convergence without manual tuning. Training was performed with a batch size of 16, chosen to maximize GPU utilization while maintaining sufficient stability in gradient updates.

All experiments were conducted at a native resolution of 512×1024 , preserving the full equirectangular spatial structure of the input panoramas. This resolution is the native one for Structured3D [16]). Under these conditions, the average training time per image was approximately 32 milliseconds, enabling efficient iteration over large datasets. At inference time, our model achieves a latency of just 7 milliseconds per image on the same RTX 4090, making it well-suited for real-time or near-real-time applications.

During training, invalid depth values—such as points corresponding to views through windows or regions missing in the ground truth—are masked out and excluded from the loss computation. Additionally, these regions are set to zero in the output to prevent the network from hallucinating unreliable depth information.

Table 4.1 presents the computational performance of inference with our network. We compare it to major state-of-the-art depth estimation solutions for 360° indoor imagery, for which performance is reported in the original publication or the code is

Method	Parameters↓	FLOPs↓	Inf. time↓
Bifuse [59]	253 M	682 G	144 ms
SliceNet [15]	79 M	101 G	21 ms
Panoformer [62]	20 M	78 G	17 ms
EGFormer [144]	15 M	74 G	16 ms
Elite360D [20]	25 M	65 G	14 ms
DDD++ (our)	23 M	38 G	7 ms

Table 4.1: **Computational performance of depth inference.** We show our computational performance compared to other state-of-the-art works for a 512×1024 image. We also show an example of inference time on a NVIDIA RTX 4090.

available for testing. As we can see, our approach has, by far, the lowest computational complexity (FLOPs) of the compared methods (see subsection 4.5.3). We also show, for a more intuitive comparison, the average inference time for all methods on an NVIDIA RTX 4090 (24GB VRAM). Our computational cost is, in particular, less than half of the currently fastest method (Elite360D [20]).

The number of parameters is also in the ballpark of recent solutions based on vision transforms (Panoformer [62] and EGFormer [144]) and much less than prior solutions (SliceNet [15] and Bifuse [59]). Our method’s reduced cost and footprint make it possible to scale our solution to larger image sizes than competitors when suitable higher-resolution training data will be available.

4.5.3 Evaluation and comparison with the state-of-the-art

DDD++ is a simple, lightweight architecture that quickly produces depth estimation and is trained by a simple loss function exploiting density maps. Since the density-map-based loss function is one of our main contributions, we strive to demonstrate that good results can be achieved at a low computational cost and without introducing other loss terms, based, e.g., on smoothness, planarity, verticality of walls, or other specific geometrical and architectural priors.

Table 4.2 presents a comparative evaluation of depth estimation performance across state-of-the-art methods on both real-world (Matterport3D [14]) and synthetic (Structured3D [16]) datasets. To reduce problems stemming from variations in architecture complexity, hyperparameters, and validation procedures used across different methods, we refer to the latest state-of-the-art work in this context, Elite360D [20], which performed retraining and evaluation under the same conditions as Jiang et al. [145]. The error metrics used, where lower values indicate better performance (↓), include **Absolute relative error (Abs Rel)**, **Squared relative error (Sq Rel)**, and **Root mean squared error (RMSE)**. Additionally, the evaluation utilizes

Table 4.2: **Depth estimation performance compared to SoA works.** We present results and comparisons on both real-world and synthetic datasets. For comparison, we follow the same setup of the recent publication Elite360D [20], which provides exhaustive results about the latest state-of-the-art approaches. To support further evaluations, we train, under the same conditions [20], *SliceNet*, *DDD*, and *DDD++*.

Datasets	Backbone	Method	Abs Rel ↓	Sq Rel ↓	RMSE ↓	$\delta_1(\%)$ ↑
M3D [14]	Transformer	EGFormer [144]	0.1473	0.1517	0.6025	81.58
		PanoFormer [62]	0.1051	0.0966	0.4929	89.08
	ResNet-18	UniFuse [145]	0.1191	0.1030	0.5158	86.04
		Elite360D [20]	0.1272	0.1070	0.5270	85.28
	ResNet-34	BiFuse [59]	0.1126	0.0992	0.5027	88.00
		UniFuse [145]	0.1144	0.0936	0.4835	87.85
		Elite360D [20]	0.1115	0.0914	0.4875	88.15
	ResNet-50	UniFuse [145]	0.1185	0.0984	0.5024	86.66
		Elite360D [20]	0.1112	0.0980	0.4870	86.70
	ResNet-18+MHSA	DDD [12]	0.1457	0.1522	0.6550	81.48
		DDD++(Ours)	0.1130	0.1026	0.5054	84.28
S3D [16]	Transformer	EGFormer [144]	0.2205	0.4509	0.6841	79.79
		PanoFormer [62]	0.2549	0.4949	0.7937	74.70
	ResNet-34	BiFuse [59]	0.1573	0.2455	0.5213	85.91
		UniFuse [145]	0.1506	0.2319	0.5016	85.42
		Elite360D [20]	0.1480	0.2215	0.4961	87.41
	ResNet-50+RNN	SliceNet [15]	0.1225	0.2214	0.5024	90.82
	ResNet-18+MHSA	DDD [12]	0.0607	0.1128	0.1594	96.14
		DDD++(Ours)	0.0504	0.1092	0.1483	97.18

three threshold percentages, denoted as $\delta_1(\%)$. These measure the percentage of pixels where the depth prediction error is less than a threshold defined as $\delta < \alpha^t$, where $\alpha = 1.25$ and $t = 1$. For these threshold metrics, a higher value indicates better performance ($\uparrow$).

Our method, DDD++, achieves the best performance on the synthetic Structured3D dataset across all metrics. This result is particularly significant as Structured3D provides complete and clean equirectangular views along with full ground-truth depth maps, ensuring the comprehensive geometric context that makes it possible to fully exploit our proposed occupancy-aware loss.

Conversely, on the real-world Matterport3D dataset, while DDD++ performs comparably with other top-performing methods, it does not achieve the best scores, especially when compared with solutions with a much increased computational cost. In addition to computational considerations (see [Table 4.1](#)), this can be attributed to two key factors. First, the equirectangular views in Matterport3D are incomplete and obtained through stitching, which leads to missing or distorted geometric information. Second, many of the competing approaches incorporate additional supervision signals – such as gradient consistency, surface normals, or structural priors – that, in addition to helping to compensate for the incomplete view, provide a better resistance to noise and increased detail preservation. These auxiliary loss terms are particularly effective in noisy real-world scenarios where geometric context is partially missing. We expect that, by incorporating some of these additional terms in our loss function, an even better solution could be achieved.

It is important to note that in Structured3D, such auxiliary losses play a more marginal role due to the dataset’s completeness and reduced noise. This underscores the robustness and potential of our proposed loss, which directly exploits full geometric cues when available. Moreover, we emphasize that our novel loss formulation is agnostic to the network architecture and can be seamlessly integrated with other models and additional loss terms. We leave this broader integration and ablation analysis for future work.

Another important aspect highlighted in [Table 4.2](#) is the impact of network backbone design on performance. While methods such as SliceNet employ structural priors to achieve strong results (particularly on the Structured3D dataset), these approaches often rely on heavier architectures like ResNet-50 combined with recurrent modules (RNN), resulting in high computational costs (see [Table 4.1](#)). In contrast, our DDD++ model leverages a lighter ResNet-18 backbone augmented with multi-head self-attention (MHSA), striking a better balance between efficiency and performance. Notably, the structural consistency enforced by our proposed loss term on density

maps offers a compelling alternative to architectural complexity. Unlike general depth estimation methods such as Elite360D [20], which primarily optimize per-pixel depth accuracy, our formulation encourages data-driven regularization that captures broader structural patterns. We believe this performance gain stems from the synergy between our network design and the proposed loss, which jointly enable the model to better exploit medium- and large-scale regularities characteristic of indoor environments. This suggests that enhancing depth estimation through density-based geometric reasoning can be an effective and scalable strategy, especially in domains where structural coherence plays a central role.

Figure 4.3 presents qualitative examples of our method’s predictions, showcasing both the estimated depth maps and the corresponding point clouds. These visualizations highlight the effectiveness of our approach in preserving fine architectural details. Notably, our method is capable of maintaining sharp geometric features, such as edges and corners, as well as ensuring the smoothness of continuous surfaces. This is achieved without relying on any explicit regularization terms or post-processing heuristics, underscoring the intrinsic strength of our formulation. Furthermore, Figure 4.4 and Figure 4.5 provide a comparative analysis on the Matterport3D [14] and Structured3D [16] datasets, respectively. In these figures, we compare our predictions with those of SliceNet [15], an established method that, like ours, exploits gravity alignment to streamline the network design, but uses a loss function that measures errors on data and gradient components.

4.5.4 Comparison with structured-guided baselines and ablation

To demonstrate the improvements of *DDD++* compared with other structure-guided methods, including the *DDD* baseline [12], we present results on SKI360 dataset [13], which is the benchmark adopted by *DDD* work. With this benchmark, specific results with another structure-guided method (i.e., Jin et al. [21]) are available. The framework introduced by Jin et al. [21] represents a notable state-of-the-art pipeline for indoor scene understanding, jointly predicting per-pixel depth and room layout. The layout is expressed as a structural representation consisting of corners, boundaries, and planes, and serves as a strong geometric prior. The correlation between depth and layout introduces a robust form of structural consistency, which the method exploits through geometric structure-aware and regularized depth estimation.

To ensure a fair comparison, we used the same settings adopted by Jin et al. [21], to train, test, and validate our proposed models, the original *DDD* [12] model and *DDD++*, as well as SliceNet [15]. For Jin et al. [21], we report their official depth

estimation results on the same dataset. It is important to note that, due to the high computational cost of their model, the results by Jin et al. are reported at a downsampled resolution of 256×512 , while our method runs at the resolution of 512×1024 . This reduced resolution may lead to slightly overestimated performance metrics for Jin et al. [21], as finer structural details are lost and prediction becomes less sensitive to high-frequency errors.

Table 4.3 summarizes this comparison. The first two rows show the results from Jin et al.’s network, where ("with SC") and ("no SC") indicate the presence or absence of structural consistency enforced by layout priors. The last three rows present the results obtained with SliceNet [15], DDD [12], and our proposed DDD++ model.

All three of these latter methods share the same underlying principle of using gravity-aligned features (GAF) to better represent the vertical structure of indoor environments. However, they differ significantly in architectural choices and computational efficiency. SliceNet employs a ResNet-50 backbone combined with a recurrent neural network (RNN), which, while effective, is computationally expensive and less suitable for real-time applications. In contrast, both DDD and DDD++ adopt a more lightweight and efficient architecture, based on ResNet-18 augmented with a transformer-based self-attention mechanism. The primary difference between DDD and DDD++ lies in the introduction of a novel loss function in DDD++, which leverages structural priors in the form of density maps, further enhancing accuracy without additional computational cost. The improved performance in DDD++ demonstrates the advantage of the new cylindrical and planar projections.

Method	MRE↓	RMSE↓	$\delta_1(\%)$ ↑
Jin [21] no SC	0.114	0.721	89.4
Jin [21] with SC	0.103	0.666	91.0
SliceNet [15]	0.102	0.273	90.4
DDD [12]	0.063	0.254	91.9
DDD++	0.050	0.242	92.8

Table 4.3: **Comparison with structure-guided baselines.** The first two rows report the results obtained with the network of Jin et al. [21] without (no SC) and with (with SC) the inclusion of geometric priors and regularizers. The third row reports the results obtained by SliceNet [15], which uses GAFs but no structural consistency terms. The last two rows report the results obtained with the DDD baseline [12] and with DDD++.

Table 4.4 presents an ablation study aimed at evaluating the impact of each component in our proposed architecture. The bottom row corresponds to the full DDD++ configuration, which integrates all key contributions: gravity-aligned features (GAF) for encoding structural priors, the lightweight multi-head self-attention

(MHSA) module for capturing long-range contextual dependencies, and our novel cylindrical density loss term. This full-featured setting achieves the best overall performance across all metrics, confirming the effectiveness of our full design.

Moving upward, the row labeled *DDD* [12] uses the earlier version of the loss term (basic DL), based on planar density projection, while still exploiting GAF and MHSA. Despite being competitive, it underperforms compared to *DDD++*, highlighting the contribution of the improved cylindrical density loss (cyl DL).

The upper three rows further dissect the contribution of the architectural components. Specifically, we progressively disable the density loss term (no DL), the MHSA (direct decoding), and GAF, observing significant degradation in accuracy, particularly in the configuration that lacks both modules (top row), which corresponds to a plain CNN without structural priors or attention mechanisms. This baseline suffers from poor geometric understanding and achieves the lowest accuracy, demonstrating the importance of each proposed design choice.

Baseline	Loss	GAF	MHSA	MRE↓	RMSE↓	δ_1 ↑
DDD cnn	×	×	×	0.382	0.627	75.7
DDD direct	×	✓	×	0.162	0.312	88.9
DDD no DL	×	✓	✓	0.075	0.278	90.7
DDD [12]	basic DL	✓	✓	0.063	0.254	91.9
DDD++	cyl DL	✓	✓	0.050	0.242	92.8

Table 4.4: **DDD++ Ablation study.** The full model (DDD++, bottom row) combines gravity-aligned features (GAF), multi-head self-attention (MHSA), and cylindrical density loss (cyl DL), achieving top performance. Removing cyl DL (DDD row) causes a measurable drop, validating its improvement over planar density loss. Progressively disabling MHSA and GAF (upper rows) further degrades accuracy, with the plain CNN baseline (top row, no GAF/MHSA/DL) performing worst.

As additional experiments, Figure 4.6 provides a qualitative comparison of the reconstructions obtained by our method to both ground truth and the SliceNet approach [15].

In general, these experiments demonstrate how including structural consistency terms strongly benefits depth estimation, since both Jin et al. [21] and DDD improve the most important metrics when incorporating their structural terms, independently from the different paths taken to design the networks. Moreover, our method, when including the structural consistency terms, achieves state-of-the-art performance despite the much lower computational burden compared to the other baselines.

The results in Table 4.3 show also that our method also provides increased performance when compared to a reference method that enforces a stronger architectural layout structure, such as Jin et al. [21], which focuses on polyhedral rooms and was

specifically designed using the dataset employed in this specific benchmark. Although further analysis is required to draw definitive conclusions, we hypothesize that for pure depth estimation, relying solely on density map similarities – rather than using corners, boundaries, and planes as priors and regularizers – makes our approach more robust to variations in the actual layout compared to the imposed prior model. Moreover, our depth inference solution is much leaner, since the complexity of generating and evaluating density maps is relevant only to the training phase.

4.6 Failure cases

Our network produces a pixel-wise depth map, whose accuracy depends on how closely the scene adheres to the indoor structural assumptions it was trained on. Learning this structure from data, as done with our density-map-based approach, makes the network less rigid than approaches that force architectural priors (e.g., Manhattan-based methods). Nonetheless, the proposed loss term, based on structural verticality and regularity, is less effective when these basic assumptions are violated. [Figure 4.7](#) illustrates what happens in an edge case, where the scene is part of a construction that opens to the outside and contains predominantly outdoor elements ([Figure 4.7a](#)). In this scenario, it is evident that the resulting density map, particularly on the horizontal plane ([Figure 4.7c](#)), lacks the typical features visible in typical indoor settings, where, e.g., walls have important prominence. While this does not prevent the network from estimating a plausible depth map, the result is not a very good match ([Figure 4.7b](#)). This failure case highlights both the advantages and limitations of structure-guided approaches, which improve over general-purpose solutions in the restricted cases when their underlying assumptions are met.

4.7 Conclusions

Our work introduces a novel deep neural network designed for fast and structurally consistent monocular 360° depth estimation in indoor environments. This network infers a depth map from a single gravity-aligned or gravity-rectified equirectangular image, ensuring that the predicted depth matches the typical depth distribution and features of cluttered interior spaces. This is achieved by a network architecture that leverages the unique characteristics of vertical and horizontal features present in man-made interior environments through gravity-aligned feature flattening, feeding specialized vision transformers. To improve structural consistency, we introduced a novel purely data-driven loss function that measures the difference between the

density maps constructed by projecting predicted depth values onto horizontal (i.e., full-view cylindrical projection) and vertical planes and those built from training data.

Our experiments show that this approach achieves very good depth estimation results while maintaining a lightweight architecture with the low computational demands required by real-time usage in applications such as extended reality exploration and autonomous navigation. The solution offers greater structural consistency compared to existing methods that focus on optimizing pixel-wise depth estimation accuracy. Moreover, consistency is achieved by learning hidden relations from example sets, rather than implicitly or explicitly forcing the alignment with strict planar layouts. The results presented in this work highlight the benefit of the new density map computation and loss function design relative to the original conference contribution [12], and include a much extended evaluation on real and synthetic data, a comparison with state-of-the-art solutions, a complete ablation study, and a critical discussion.

Since our work focused on evaluating the benefits of our novel density-map-based loss, we did not complement it with other loss components, demonstrating how we can achieve good structural preservation and depth estimation performance even in simplified settings. In the future, to further improve performance, we plan to extend this work by supplementing the loss with other gradient-based terms and normals, working especially at the fine-detail scale, as in other state-of-the-art methods.

Finally, while this work focused on the depth estimation task, we also plan to exploit our method as a building block inside a full processing pipeline. The two use cases that we are targeting are the extraction of multi-room 3D models from very sparse sampling (e.g., one image per room) and the generation of depth to support the synthesis and exploration of stereoscopic environments from a single surround-view panoramic image in extended reality settings. In both cases, the depth estimation task is very important, as is the preservation of structural consistency.

4.8 Bibliographic notes

The contents of this chapter report the article *Giovanni Pintore, Marco Agus, Alberto Signoroni and Enrico Gobbetti: DDD++: Exploiting Density map consistency for Deep Depth estimation in indoor environments* [61], of which the candidate is the lead author.

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 the contribution of the Italian National Research Center in High-Performance Computing, Big Data, and Quantum Computing (Next Generation EU PNRR M4C2 Inv 1.4). The findings herein reflect the work and are solely the responsibility of the authors.

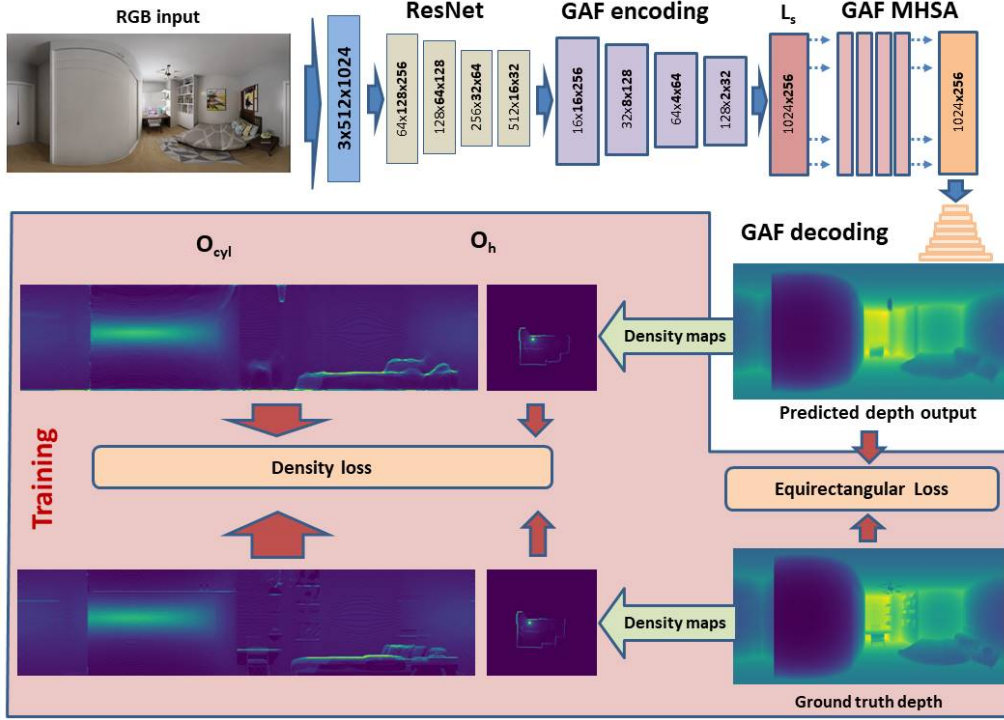


Figure 4.1: **Deep depth estimation overview.** The network maps a gravity-aligned 360° image to its depth. The input image is first transformed by a ResNet block (light green) into feature maps with different depths and spatial sizes. Through a Gravity-Aligned Features (GAF) encoding block (light purple), we perform a gravity-aligned anisotropic contractive encoding to obtain latent features. Once assembled in a single sequence, latent features are processed by a single-layer multi-head self-attention scheme (light pink) to produce the final set of features whose decoding, through convolution and upsampling, produces the desired depth map. During training, we produce density maps respectively from predicted and ground truth depths and exploit them for structural loss computation. It should be noted that density maps are computed only at training time. Thus, the memory and computational costs do not affect network performance at inference time. Note that the depth image and density map colors have been visually enhanced for illustration purposes.

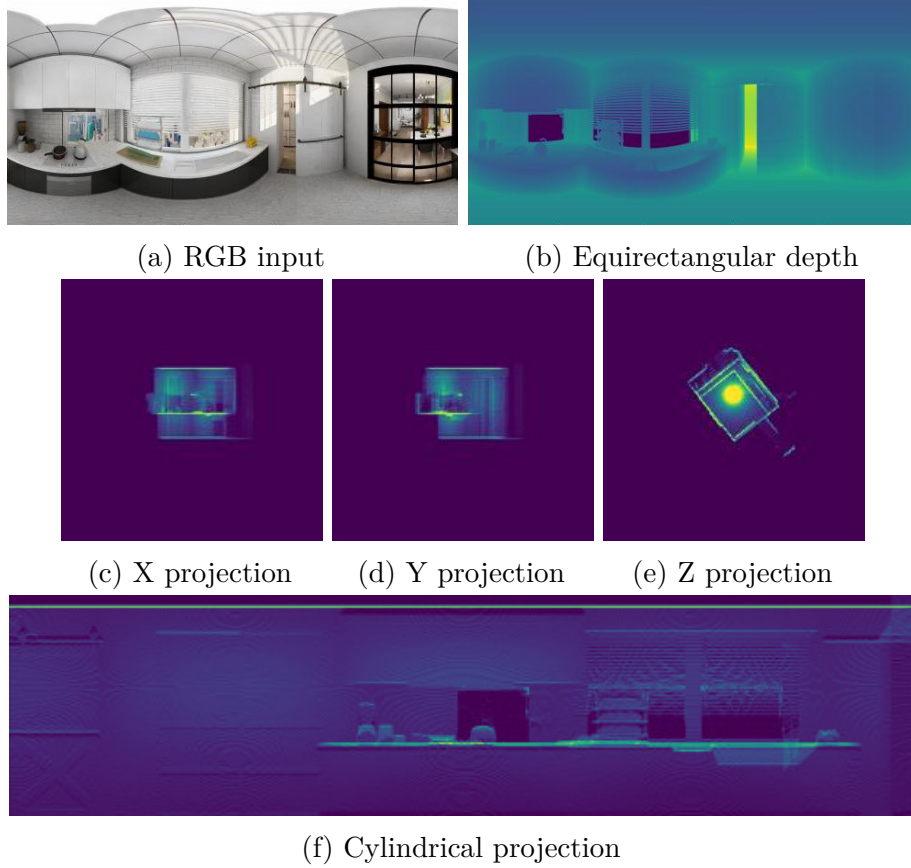


Figure 4.2: **Different density maps examples.** In the first row we show the input image [Figure 4.2a](#) and the equirectangular depth [Figure 4.2b](#). In the second row we show examples of density maps recovered from the equirectangular depth. In this case we illustrate the density maps adopted by the DDD baseline [12]: [Figure 4.2c](#) and [Figure 4.2d](#), respectively projections along the X and Y axes; [Figure 4.2e](#) along Z axis, that is the density map projected on the floorplan (i.e., vertical projection). Here we keep [Figure 4.2e](#), which is highly representative of the room shape, and we replace the arbitrary projections [Figure 4.2c](#) and [Figure 4.2d](#) with a more general and comprehensive cylindrical projection [Figure 4.2f](#), which cover the whole horizon (i.e., horizontal projection). As also evident visually, this projection best captures seamless structural details.

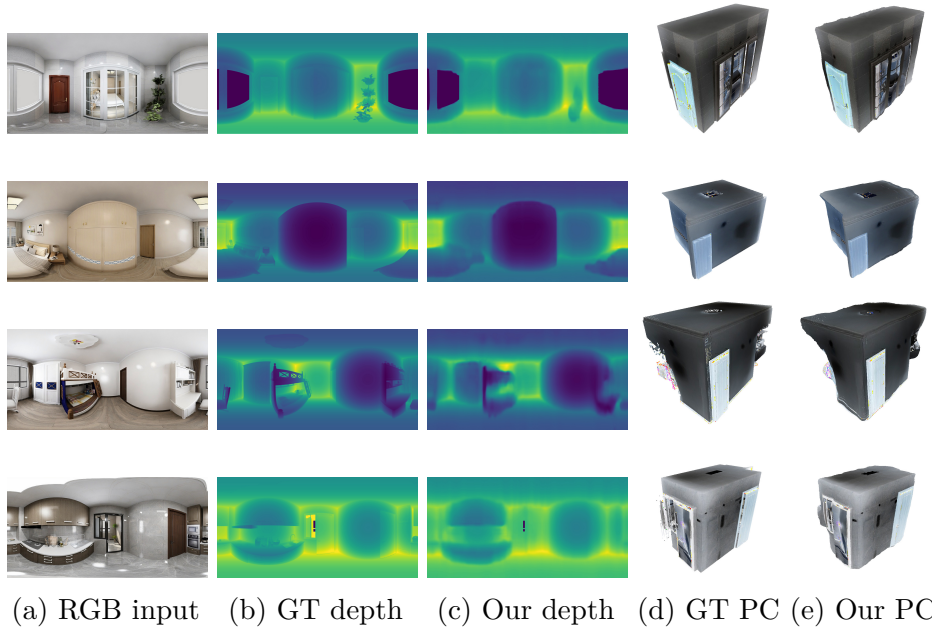


Figure 4.3: **Depth estimation qualitative performance.** We show our reconstructions and ground-truth models as depth maps and as the associated point clouds from the SKI360 dataset [13]. Thanks to structural consistency and without specific post-processing, the method effectively preserves architectural details, such as sharp edges and smooth planes.

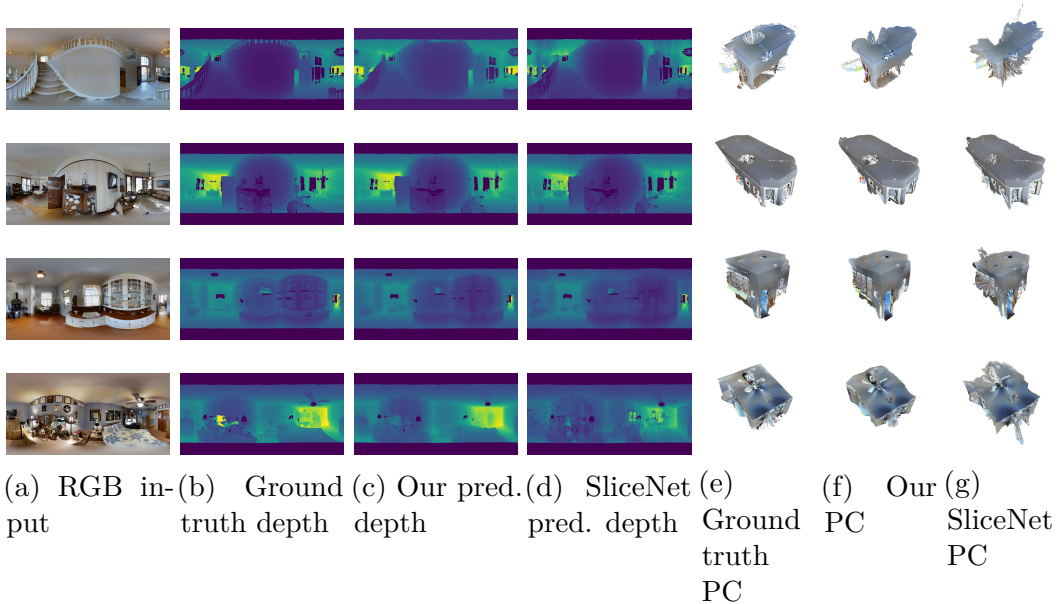


Figure 4.4: **Depth estimation qualitative performance and comparisons on Matterport3D dataset [14].** We illustrate our qualitative performance compared to the competitor [15] on real-world scenes. We show reconstructions and ground-truth models as depth maps and as canonical views of the associated point clouds.

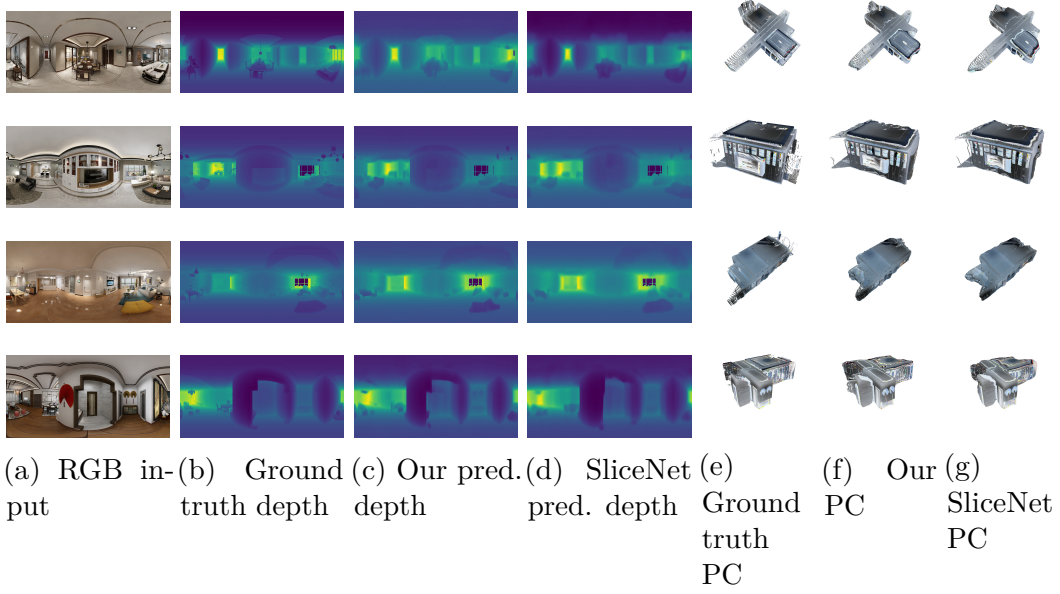


Figure 4.5: **Depth estimation qualitative performance and comparisons on Structured3D dataset [16].** We illustrate our qualitative performance compared to the competitor [15] that reported the best performance in Table 4.2. We show reconstructions and ground-truth models as depth maps and as canonical views of the associated point clouds.

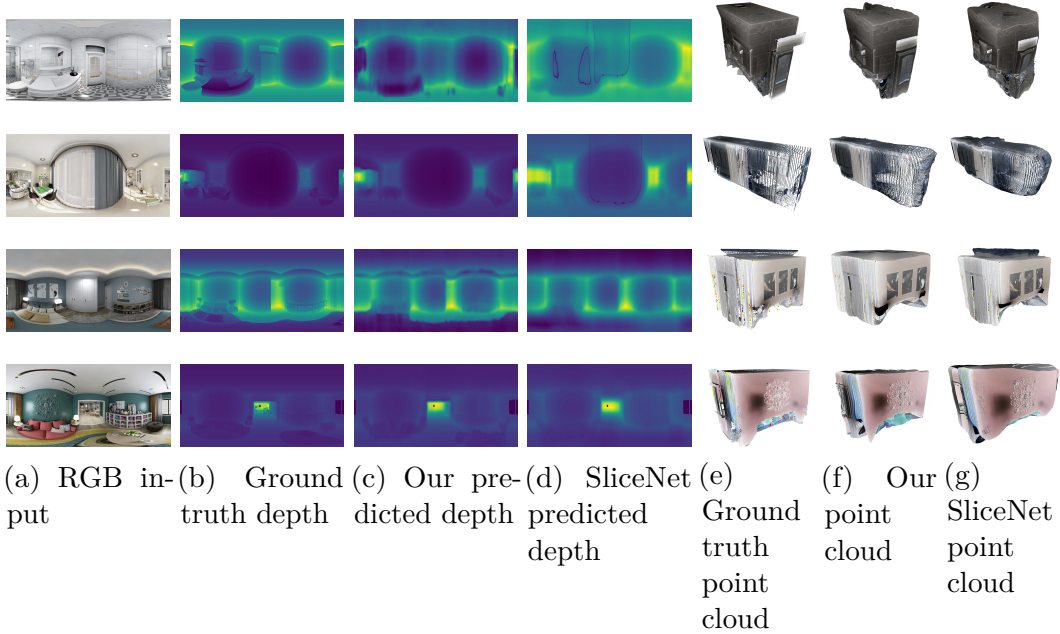


Figure 4.6: **Depth estimation qualitative comparison with structure-guided indoor depth estimation baseline.** We illustrate our qualitative performance on SKI360 dataset [13], compared to a state-of-the-art solution that, as ours, exploits gravity alignment (SliceNet [15]). We show reconstructions and ground-truth models as depth maps and as canonical views of the associated point clouds.

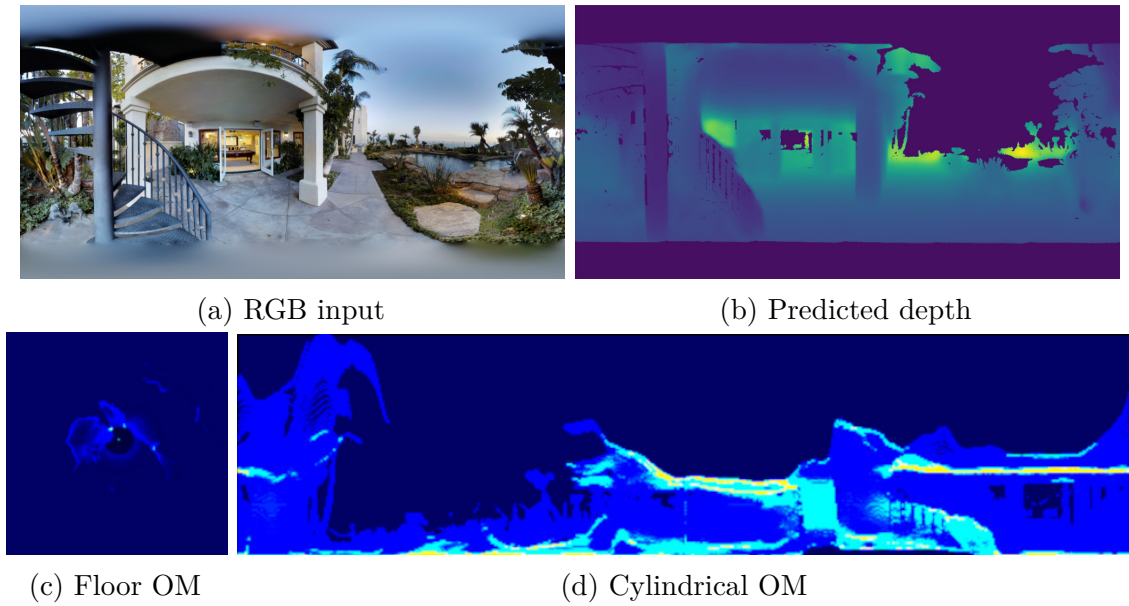


Figure 4.7: **Failure case illustrating the limitations of the proposed loss term.** The input RGB image (Figure 4.7a) depicts a semi-outdoor scene that violates the indoor structural assumptions of the training data. As a result, the density map, particularly on the horizontal plane (Figure 4.7c), lacks meaningful features. Despite this, the network is still able to produce a plausible pixel-wise depth prediction (Figure 4.7b).

5 Multi-Room 3D Floorplan Reconstruction

5.1 Problem Definition and Motivation

The automatic 3D reconstruction of indoor structures from purely visual input is a very active visual computing research topic [9]. The focus is on creating quick and efficient methods to reconstruct the permanent architectural structure of common environments, like homes, offices, and public buildings [75], by exploiting their regularities to aid reconstruction from incomplete or noisy data [9]. In this context, 360° photography is one of the most prevalent capture methods [137, 99, 138].

This work addresses the challenge of reconstructing 3D floorplans from a minimal set of mutually registered gravity-aligned 360° images of cluttered rooms. This scenario is representative of common practices in the real estate industry, where millions of casual users generate and share virtual tours by capturing about a single image per room with consumer-grade cameras, exploiting a small amount of mutually visible space, e.g., through open doors, for registration [146]. Leveraging this sparse data to automatically generate floorplans could enable a range of important applications, including enhanced navigation during virtual tours, improved price estimation, and integration with building information models [18]. However, this minimalistic capture approach poses significant challenges due to the sparsity and ambiguity of available information (section 5.2).

Recent deep-learning approaches have obtained important successes in 3D inference tasks by exploiting available large-scale building datasets of panoramic images coupled with annotated layouts, both synthetic [16] and real [18] (section 5.2). However, while good performance may be achieved for pixel-wise depth inference (e.g., [60, 65]), 3D room layout reconstruction from monocular input is still limited by the need to extrapolate large portions of the invisible structure, which can be occluded not only by objects but by the structure itself [1], forcing the incorporation of heavy priors and heuristic post-processing [80, 38, 147]. Moreover, inference at the

individual room level may produce inconsistent global results, especially in invisible areas, leading to the need for specialized integrated model computation methods [9]. However, state-of-the-art multi-room reconstructors usually require the availability of reliable density maps for their analyses [101, 17]. 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.

We overcome these challenges using a bottom-up approach, in which information on the architectural room structure is inferred for each image before being used, globally, to predict the multi-room shape from multi-view fusion. Leveraging the availability of annotated datasets of room layouts and exploiting the vertical walls assumption, we train a transformer-based neural network to predict, from a single panoramic image of a cluttered room, the room’s geometric structure – excluding furniture and objects – and its projection onto a horizontal plane (the *Nadir plane*). This process yields, for each room, an estimate of ceiling heights and a map of the disoccluded floor area (the *Nadir shape*). The Nadir shapes of all rooms within an environment are then integrated into a consistent, unified floor plan using a deformable attention transformer, which employs mutual information to resolve structural occlusions. Our main contributions are the following:

- We introduce a transformer-based neural network that predicts, from an equirectangular image of a cluttered room, the depth of the emptied scene and its projection onto the Nadir Plane (section 5.4). It combines gated convolution for residual feature extraction with gravity alignment for feature compression. By transforming and projecting the uncluttered Euclidean depth and incorporating indoor-specific transformations and loss functions, we obtain a metrically consistent regularized and self-completed Nadir shape, only assuming vertical walls rather than imposing heavy constraints, such as MWM or AWM (Atlanta World) [81, 84]. Focusing on learning to infer the uncluttered scene proves more effective than analyzing the depth of the cluttered scene for structural predictions (section 5.6.3). While the Nadir shape is not a full room layout, its accuracy is comparable to that of dedicated layout prediction methods, without requiring additional post-processing or heuristic completion.
- We join all the Nadir shapes in the same Nadir floorplan, exploiting a deformable transformer network [148] to process inter-relations, completing self-occluded parts, and refining room shapes (section 5.5). In contrast to previous work, we do not take a dense multi-view occurrences map as input to the fusion step [99, 101, 17], but we start from the self-completed *Nadir shape* information

extracted at the single image level to drive the solution of the final floorplan with ceiling height (section 5.6.4), supporting reconstructions with a single image per room.

- We propose a fully data-driven, end-to-end solution, termed *NadirFloorNet*, that integrates into a single deep network architecture per-image Nadir shape inference with global multi-room floor plan reconstruction. *NadirFloorNet* directly maps a set of panoramic equirectangular images, each with its associated reference frame, to a complete floor plan with room heights, without any intermediate processing steps (section 5.3). Although *NadirFloorNet* functions as a single network during inference, the Nadir shape prediction module is trained first and acts as a pre-trained module during the training of the floor plan module. Given the differentiability of the entire network, we support future fine-tuning, allowing all combined network parameters to be learnable.

We evaluated our method on large-scale synthetic and real benchmarks [16, 18]. Our results (section 5.6) demonstrate that *NadirFloorNet* outperforms current state-of-the-art approaches starting from a minimal set of panoramic images.

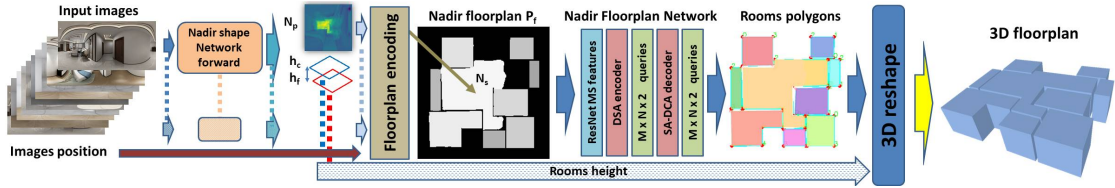


Figure 5.1: **NadirFloor overview.** For each input panoramic image, a Nadir shape network forward pass infers a Nadir shape and the heights of floor-ceiling planes (see section 5.4). These representations are then merged into a fixed-size Nadir floorplan P_f , exploiting the available mutual registration. The Nadir floorplan is processed through the Nadir floorplan Network (see section 5.5) to predict the 2D polygonal multi-room floorplan. Finally, a 3D floorplan is recovered by extruding the 2D plan using the room heights recovered by the Nadir shape network.

5.2 Task-related works

Indoor reconstruction is a broad research topic. We analyze only the closely related approaches, referring the reader to established and recent surveys for wider coverage [9, 138].

Depth and layout estimation from panoramic images. 360° images are increasingly used for data-driven depth estimation in indoor spaces for their ability to capture the full surroundings. Their exploitation includes adapting perspective

methods through spherical convolutions [55, 56, 57, 58, 53], joint processing in mixed equirectangular and cube-map projections spaces [59], leveraging perspective views sampled on panoramic images before combining depth maps using transformers [63, 33, 65], as well as exploiting gravity-aligned features for direct processing in equirectangular space [80, 15]. Our approach emphasizes the structural nature of the desired output [21] by seeking to obtain the depth of the permanent structure deprived of the various objects in the room, combining and extending image inpainting and structural reconstruction concepts [15]. Instead of computing per-pixel depth data, structural extraction from single panoramas can also be obtained by layout estimation methods targeting the production of seamless 3D boundary surfaces in case of self-occlusions. Prominent examples include *LayoutNet* [79], which predicts the corner probability map and boundary map directly from a panorama, and *HorizonNet* [80], which simplifies the layout as three 1D vectors before extracting the 2D layout by fitting MWM segments to the estimated corner positions, as well as a breed of transformer-based approaches that project the equirectangular input image to planar surfaces [81, 84, 82, 147, 87]. All these methods, however, require heavy preprocessing, such as detection of main MWM directions from vanishing lines analysis and related image warping [4, 52, 38], or complex layout post-processing, such as MWM regularization of detected features [79, 80, 81, 64]. In this work, we propose, instead, a fully data-driven approach for estimating the Nadir plane footprint and 3D cues, not from the cluttered RGB image but from its uncluttered depth, appropriately transformed (section 5.4), deferring to multi-view computation the definition of consistent layout details also in invisible areas, without imposing MWM priors or requiring the hallucination of non-visible shapes. Our method only requires overlap if needed for mutual registration, enabling reconstruction with most points seen only once – typically using one image per room registered via a few feature points from open doors. Unlike multi-view approaches that rely on multiple images per room, even if only a few [149, 150, 151, 152], we support a single image per room by estimating monocular uncluttered depth before merging.

Data-driven floorplan reconstruction. Early approaches combined low-level image processing with geometric reasoning and energy minimization solvers to extract room layouts [8, 93, 27, 94, 75]. Many recent solutions adopt a hybrid approach where low-level primitives detected by neural networks are assembled by optimization techniques into the final models. These include methods that detect room corners and then generate wall segments through integer programming [100], methods that detect room segments with Mask R-CNN [153] and reconstruct individual room polygons by sequentially solving shortest path problems [98, 154], approaches that

uses Monte-Carlo Tree-Search to select room proposals [99], and diffusion techniques that generate plausible room arrangements by combining graph neural networks with constrained diffusion [155, 156]. In contrast to the hybrid solutions, several recent approaches employ, like ours, an end-to-end deep-learning architecture. In particular, HEAT [101] integrates the DETR deformable transformer [148], into a bottom-up pipeline that first detects corners and then classifies edge candidates connecting corners. Also based on DETR [148], RoomFormer [17] predicts floorplans from a dense point cloud using a single-stage, end-to-end trainable neural network. Differently from previous data-driven approaches [98], RoomFormer encodes the floorplan as a variable-size set of polygons, which are variable-length sequences of ordered vertices. SLIBO-Net [102] focuses on improving RoomFormer’s semantic and local geometric quality by incorporating additional MWM priors and post-processing steps. More recently, PolyDiffuse [22] refines polygonal reconstructors from point cloud density maps through a conditional generation procedure. We can consider it an orthogonal work that could be used as a post-processor for our baseline, similar to how it was applied to RoomFormer [17]. We also exploit deformable transformers [148] encoding floor plans with a variable sequence of polygons and vertices [102, 17], but start from pure, sparse imagery input, without needing dense coverage, dense point clouds or MWM-based post-processing. Extreme registration approaches [92, 146], like ours, take as input sparse panoramic images, but exploit them for relative pose estimation with minimal (or no) overlaps. They can offer a complement for initial camera registration.

5.3 Overview

fig. 5.1 summarizes our reconstruction approach, focusing on the forward prediction pass for a more clear illustration. We refer the reader to the supplementary material for network and training details. We take as input a small set of spatially registered and gravity-aligned equirectangular images. For each input image, the *Nadir shape* network predicts a regularized probability map of the walkable floor area N_p and the floor-ceiling planes distance from the observer, respectively h_f and h_c (section 5.4). We exploit then the relative pose of the input images to encode all Nadir shapes in a common, fixed-size, floor plan map P_f . The recovered metric scaling is stored separately to reshape the final 3D floorplan. Given this joined representation P_f , named *Nadir floorplan*, an encoder-decoder transformer-based network processes inter-relations, completes self-occluded parts and refines room shapes as 2D polygons (section 5.5). Finally, a 3D floorplan is generated by combining ceiling-floor heights,

metric aspect, and 2D polygons.

5.4 Nadir shape generation

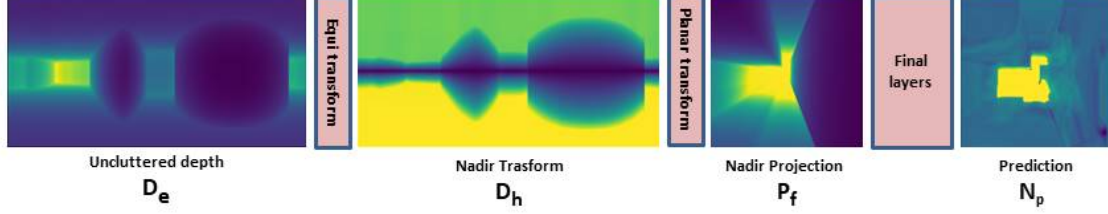


Figure 5.2: **Nadir shape network last layers.** Starting from the intermediate uncluttered depth D_e , the fully differentiable transform and projection recover the Nadir projection P_f . Finally, the last autoencoding layers return the Nadir Shape N_p .

The Nadir shape generation is performed, for each input image, by the Nadir shape network (Figure 5.1). The network takes as input an equirectangular image I_c of a cluttered room, and outputs in the forward pass a regularized probability map of the free floor area N_p with the floor-ceiling planes distances h_f and h_c . The same network also outputs the equirectangular depth of the emptied scene D_s as an intermediate result, which is used in loss computation and weight update during the backward pass of training (Figure 5.2). Internally, the network predicts an attention mask, segmenting the input image into cluttered or uncluttered zones. A *residual-gated* network extracts features from the input image and its mask. Features are then encoded as a fixed-size sequence of vertically-compressed features, leveraging the indoor structural nature of the desired output, and then processed by a transformer to decode an *uncluttered* depth map of the scene. The depth map is then transformed and projected to recover the Nadir shape.

Attention mask and feature extraction.. Starting from the input equirectangular image I_c , we predict an attention mask M_e through a pre-trained, lightweight network [21] (see supplementary material for implementation details), roughly segmenting I_c into cluttered or empty zones to bootstrap the gated-residual encoder. We concatenate the attention mask M_e with the input image I_c and we extract multi-layers features (i.e., four layers with different depth and spatial sizes) through a residual-gated encoder (see supplementary materials for implementation details). In contrast to canonical residual schemes [44], often adopted for generic depth estimation, our approach, for each convolution layer, introduces a dynamic gating approach to process the masked scene efficiently and to complete missing parts, corresponding

to structural bounding surfaces of the room occluded by the clutter.

Gravity-aligned-features transformer.. To make optimum use of the pixel-wise information of the equirectangular image, our network internally predicts an equirectangular clutter-free depth map D_s , by introducing a specific transformer-based approach. In contrast to a standard vision transformer baseline [157], which would require the conversion of feature maps into a sequence of pixel-wise patches, our approach aims to encode the feature maps into a sequence by accounting for the indoor nature of the scene and its spherical representation. Thus, as in other depth estimation works [15, 60], starting from the assumption that gravity plays an important role in the design and construction of interior environments, we assume that world-space vertical and horizontal features have different characteristics. Such an assumption is even stronger in our case since our goal here is to recover only the structure (i.e., walls, ceiling, and floor), without the need to model clutter details, that typically contain the most free-form data. We perform, thus, an anisotropic, contractive encoding that reduces the vertical direction (i.e., gravity direction) while keeping the horizontal direction unchanged, generating gravity-aligned features (GAF). We apply such compression to each encoded feature map. Finally, compressed features are reshaped to the same size and joined in a flattened GAF feature, $L_s = (l_0 \dots l_s)$, as a sequence of s feature vectors of dimension l (i.e., s horizontal size of the less deep feature map - $s = 256$ and $l = 1024$ for a 512×1024 input). Such a sequence L_s contains both local and non-local geometric features, which are exploited to recover missing depth samples through a multi-head self-attention transformer [143]. Once passed to the transformer, sequence decoding is very fast, through convolutions, upsampling modules, and ELU activations, until we reach the target output resolution.

Nadir shape recovery.. To recover the Nadir shape for each camera, we need to efficiently translate the information from the emptied depth map D_e to the Nadir plane. Since recognizing man-made structures is not immediate in spherical space, the network processes 3D data by storing at each pixel of the equirectangular map, rather than the depth D_e , the distance to the horizontal plane passing through the camera location, i.e., the relative height with respect to the eye. This is accomplished by scaling each value in D_e , which corresponds to an azimuth angle θ (along w) and polar angle ϕ (along h), by $\|\sin \phi\|$ to obtain the map D_h . This representation D_h (fig. 5.2, D_h), under the vertical walls assumption better highlights the room structure, although still in equirectangular format. The height value is also exploited to recover additional 3D cues, such as floor and ceiling distances from the camera, to complete the 3D layout of the rooms. To recover the footprint in the Nadir plane,

which corresponds to the predicted floor, exploiting the Vertical Walls prior [9], we project the equirectangular depth map D_h to the Nadir plane, $P_f^{wp \times wp}$ perpendicular to the Z-axis, so that P_f represents information belonging to the bottom hemisphere of D_h (fig. 5.2, P_f). This transformation and projection operation is fully differentiable and integrated into the network. However, because such an intermediate representation P_f directly comes from the predicted emptied depth, much structural information, especially on the edges, may be missing or noisy. Therefore, the final layers of the network consist of an UNet-based (see supplementary material for details) autoencoder, which takes the P_f representation and output a the regularized probability map N_p (see fig. 5.2 and fig. 5.1). Projecting onto the floor, which is assumed to be a single plane, lends itself better to the merging of multiple footprints, also in the case of rooms with different ceiling heights. In contrast, projecting onto the ceiling, as done by traditional layout recovery methods working on cluttered input images [84, 81] would necessitate a consistent horizontal plane across the entire multi-room environment, limiting the solution space.

Nadir shape network training.. We train the Nadir shape network supervising the training with single room layouts from real-world and synthetic annotated floorplans (section 5.6.2). We do not need additional annotations for this training task, but we use layouts from the main floorplan reconstructor sets (i.e., adopting the same train/valid/test split - section 5.5). To directly supervise the Nadir shape prediction and the layout heights, we adopt, respectively, $\mathcal{L}_l$, the binary cross entropy with logits loss for the predicted probability map N_p section 5.4 and $\mathcal{L}_h$, the $L1$ distance error for the predicted average ceiling-floor distances. Furthermore, starting from the assumption that indoor structure and depth are interrelated [21], we enforce the training by exploiting the intermediate clutter-free depth D_e (see section 5.4). To recover ground truth data for D_e , we render such depth from the annotated layout. We adopt then several specific losses for such depth. The robust *Adaptive Reverse Huber Loss (BerHu)* $\mathcal{L}_d$; the Structural Similarity Index Measure (SSIM) $\mathcal{L}_{dss}$, which measures the preservation of highly structured signals with strong neighborhood dependencies. Furthermore, to enforce the indoor nature of the output, we include more specific indoor structural losses, respectively the normal consistency loss L_n and the gradient of normals consistency loss L_g , which supervise the smoothness of walls and sharpness of edges. Assuming vertical walls, we restrict the normal map to the horizon [82], computing the cosine similarity to get the normal loss [147] $\mathcal{L}_n = \frac{1}{W} \sum_{i=1}^W (-n_i \cdot \bar{n}_i)$, where W is the width of the equirectangular image, n_i is the predicted normal, $\bar{n}_i$ is the ground truth normal. To supervise the turning of corners, we compute the loss term as $\mathcal{L}_g = \frac{1}{W} \sum_{i=1}^W (g_i - \bar{g}_i)$,

where as $g_i = \arccos(n_{i-1} \cdot n_{i+1})$ is the angle between consecutive normals. As a result, the Nadir shape network is trained by combining indoor depth and layout losses $\mathcal{L}_{sv} = \lambda_d \mathcal{L}_d - \lambda_{dss} \mathcal{L}_{dss} + \lambda_l \mathcal{L}_l + \lambda_h \mathcal{L}_h + \lambda_n \mathcal{L}_n + \lambda_g \mathcal{L}_g$ (see supplementary for lambda values).

5.5 Nadir floorplan reconstruction

The Nadir shapes recovered for each image are joined and encoded in the Nadir floorplan (section 5.3). We cast the floorplan reconstruction as a prediction problem of a polygon set, where each polygon represents a room modeled as an ordered sequence of vertices. It should be noted that, differently to pure prediction from an unstructured point cloud [101, 99, 17], here we start from a partial room segmentation, so the main goal is to process inter-relations, complete self-occluded parts, and refine room shapes.

To do this, we scale and fit our Nadir floorplan, which is in metrically-scaled dimensions, into a fixed-size map of dimension $F_n^{wf \times wf}$ (fig. 5.1), according to common, available annotations [101, 99]. To join and encode the Nadir shapes in this representation, we transform each probability map N_p into a binary mask N_s using a softmax, displace N_s exploiting its camera position, then encode each N_s mask with a progressive integer value (fig. 5.1, Floorplan encoding).

The floorplan prediction problem is then cast as $Q \in \mathbb{R}^{M \times N \times 2}$ queries [17] and solved using a deformable transformer [148]. We set M to the maximum number of rooms and N to the maximum number of vertices for each room ($M = 20, N = 40$ in our experiments), and associate each vertex to a binary value, indicating whether it is valid or not. The prediction is implemented through a network (called Nadir Floorplan network in fig. 5.1) which includes multi-scale features extraction and flattening, deformable self-attention (DSA) for transformer encoding, and self-attention and deformable cross-attention (SA-DSA) for decoding.

Nadir Floorplan network.. Similarly to the Nadir shape encoder (section 5.4), the first layers extract multi-scale features (four levels) from the Nadir floor map F_n using a plain *ResNet* scheme [44] without gated convolutions, and flatten them to a feature sequence to serve as input to a six layers transformer encoder. Each encoder layer consists of a multi-scale deformable *self-attention* module and a simple feed-forward network [148] (DSA in fig. 5.1). Similarly to encoding, decoding is performed by six stacked layers, but, differently from the encoder, each layer consists of a self-attention module, a multi-scale deformable cross-attention module, and a feed-forward network (SA-DSA in fig. 5.1). In the decoder, we adopt a combination

of content and positional queries [17], so that the decoder performs self-attention on all corner-level queries, regardless of the room they belong to. This design not only allows the interaction between corners of a single room but also enables interaction among corners across different rooms, thus combining information from different polygons to better solve structural occlusions. The decoder output is passed to a shared feed-forward network to predict binary class labels c for each query indicating its validity as a corner. Finally, the decoder outputs a $M \times N \times 2$ polygons map and a $M \times N \times 1$ classification map, from which a set of valid 2D polygons is extracted [99].

Floorplan network training.. The NadirFloor network predicts room logits and corners. Its training is supervised with the same strategy and losses employed by *RoomFormer* [17], with the important difference that we take as input the registered representations inferred for each room by the pre-trained Nadir shape prediction network rather than occurrence maps from point clouds. The transformer decoder outputs M polygons of N vertices (including non-valid ones, mapped to 0) while the ground truth contains an arbitrary number of polygons with an arbitrary number of vertices. As in *RoomFormer* [17], we adapt Deformable-DETR *Hungarian matcher* [148] to find the optimal match between prediction and ground truth. Note that, since the NadirFloor and NadirShape networks are concatenated and form, together, a fully differentiable end-to-end network mapping a set of panoramic images to a floorplan, it is possible to further improve quality through a fine-tuning pass, where the parameters of the NadirShape network are also made learnable. We plan to explore this path in future work.

5.6 Results

Our approach is implemented with *PyTorch* and has been tested on a large variety of synthetic and real-world indoor scenes. In the following we report the costs of inference and training, as well as qualitative and quantitative results on public datasets.

5.6.1 Training and inference cost

The computational complexity of both neural network modules is very low. The Nadir shape prediction network, including the initial attention mask prediction (section 5.4), has 31.81M parameters with 98.97Gflops. The average inference time for a 512×1024 equirectangular input image is 67ms on an NVIDIA RTX2060 GDD6GB laptop (e.g., building a 20-room Nadir floorplan takes less than 1.5

seconds). While the architecture consists of a single differentiable network, for practical reasons we train the two modules separately, also saving intermediate results (i.e., individual Nadir shapes and empty depths) for analysis. We trained the Nadir shape network with the Adam optimizer, with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and an adaptive learning rate from 0.0001, on an NVIDIA RTX A5000 (24GB VRAM) with a batch size of 8, with average training time 613 ms/image. The weights are: $\lambda_d = 1.0$; $\lambda_{ss} = \lambda_l = 0.5$; $\lambda_h = \lambda_n = \lambda_g = 0.1$. The Nadir floorplan reconstruction module was trained for 500 epochs following the same settings of RoomFormer (RF) [17]. The average inference time for a 256×256 floorplan map is 20ms on a NVIDIA RTX2060 GDD6GB laptop.

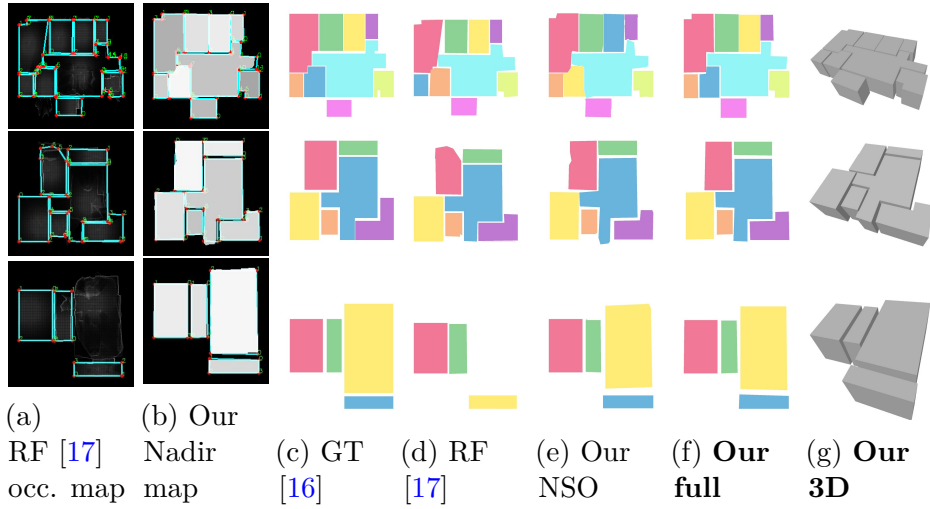


Figure 5.3: **Floorplan qualitative performances on Structured3D (S3D) [16] dataset.** Each row shows: RoomFormer (RF) [17] point cloud occupancy map (RF occ. map), our Nadir floorplan map (Our Nadir map), ground truth floorplan (GT), RoomFormer prediction (RF), our floorplan prediction just composing the Nadir shapes (our NSO - Nadir shapes only), our full floorplan prediction (Our full), our 3D prediction with predicted metric information (Our 3D).

5.6.2 Datasets

To train and test our approach, we used public datasets for which ground truth annotated floor plans and registered panoramic images were available. In addition to Structured3D (S3D) [16] and ZinD [18](see section 3.4). To have a consistent comparison, we converted ZinD to S3D’s format, following the authors’ original instructions [16, 18].

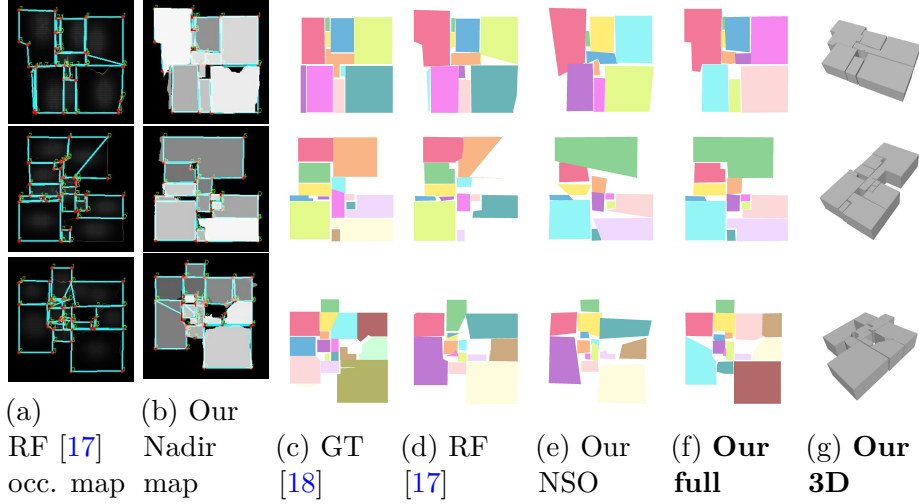


Figure 5.4: **Floorplan qualitative performance on ZInD dataset [18]**. For each row we show: RoomFormer (RF) [17] point cloud occupancy map (RF occ. map), our Nadir floorplan map (Our Nadir map), ground truth floorplan (GT), RoomFormer prediction (RF), our floorplan prediction just composing the Nadir shapes (our NSO - Nadir shapes only), our full floorplan prediction (Our full), our 3D prediction with predicted metric information (Our 3D).

5.6.3 Nadir shape reconstruction performance

The Nadir shape network outputs room shapes and floor-ceiling heights from single equirectangular images (section 5.4) to support the subsequent joint and comprehensive final reconstruction (section 5.5). Although such a representation is not intended to be used as a complete room layout (as opposed to dedicated methods using additional pre- and post-processing steps [38]), we can still evaluate the layout reconstruction performance on the single image, since our network was trained on the same data used by state-of-the-art methods dedicated to this task.

Method	IoU2D	IoU3D
Led2-Net [82]	90.36	88.49
LGT-Net [147]	91.77	89.95
DOPNet [64]	91.94	90.13
Our Nadir Shape	90.14	86.41

Table 5.1: **Nadir shape single room layout performance.** We show our quantitative performance (in bold) compared to other representative state-of-the-art works. We compute IoU2D (pixel-wise intersection over union) and IoU3D (volumetric intersection over union) on ZInD [18] benchmark.

To this end, table 5.1 presents our performance (in bold) compared to dedicated state-of-the-art works for layout estimation [82, 147, 64] on the ZInD dataset benchmark, using the results since officially results on that are already available in their

papers. For this comparison, we used the same train and test splits used for the evaluation of layout tasks (see original papers for details). Our solution achieves performances in line with specialized state-of-the-art works, demonstrating how this specific solution is particularly effective in the specific task of supporting floorplan reconstruction, although performance is obtained without post-processing (see, in particular, iou2D) and MWM assumptions for the ceiling (see, in particular, iou3D).

5.6.4 Floorplan reconstruction performance

table 5.2, fig. 5.3, and fig. 5.4 show quantitative and qualitative results on S3D [16] and ZInD [18] floorplans. As in many floorplan reconstruction works [99, 101, 17], for each ground truth room, we find the best-matching reconstructed room among all predictions in terms of IoU, reporting its Precision, recall, and F1 scores.

Method	Set	Room			Corner			Angle		
		Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
RF	S3D	0.892	0.876	0.884	0.704	0.681	0.692	0.643	0.623	0.633
Our NSO	S3D	0.939	0.930	0.928	0.553	0.542	0.647	0.424	0.522	0.466
Our full	S3D	0.943	0.939	0.940	0.834	0.801	0.817	0.746	0.717	0.731
RF	ZInD	0.811	0.756	0.788	0.712	0.615	0.660	0.528	0.444	0.505
Our NSO	ZInD	0.786	0.733	0.758	0.652	0.538	0.627	0.505	0.365	0.424
Our full	ZInD	0.824	0.767	0.794	0.765	0.626	0.689	0.594	0.488	0.536

Table 5.2: **Floorplan performance.** We show our quantitative performance (Our full - in bold) on Structured3D (S3D) [16] and ZInD [18], compared to other representative state-of-the-art works (RoomFormer (RF) [17]) adapted to the same 360° images input. We also include the performance of our method just spatially registering individual Nadir shapes (NS) - Nadir shapes only).

As discussed in section 5.1 and section 5.2, current deep-learning floorplan reconstruction methods segment a 2D occupancy map recovered from a dense point cloud, typically generated by dense coverage with regular cameras. Only a few particular methods take panoramic images as direct input, but to find their registration under extreme conditions rather than producing full floor plans [92, 146]. Both benchmarks used here present a challenging situation for reconstructing an articulated multi-room environment from just panoramic data, since only about one image/room is available for S3D, and about 1.5 image/room for ZInD.

To fairly evaluate our method against alternatives, we assume that no 3D data is available, and we predict the occupancy maps required by other methods by predicting the depth of the room using monoscopic panoramic depth inference followed by vertical projection and accumulation. We experimentally determined (see supplementary material), that when trained with original scenes rather than

uncluttered ones our Nadir shape network has a depth estimation performance in line with other state-of-the-art solutions. We thus selected our depth estimator, and combined it, to form a representative baseline, with RoomFormer (RF) [17], which is a state-of-art approach for floor plan segmentation from point cloud input. The same solution is adopted as a baseline by recent orthogonal works [102, 22].

We fully retrained RF on both Structured3D (S3D) [16] and ZInD [18] predicted point clouds, obtaining the performance illustrated in table 5.2, fig. 5.3 and fig. 5.4. With both datasets, our approach (table 5.2, *Our full*) achieves the best performance in all metrics. Furthermore, table 5.2 analyses the performance of a version of our method without the floorplan integration network (table 5.2, *Our NSO* (i.e., Nadir Shape Only)). The floorplan in this case (NSO) is the separate polygonization of the contours of each Nadir shape, and the joining of them in the same reference frame thanks to the available mutual registration of each shape, without further optimization. This comparison numerically highlights the interaction between Nadir shapes in the Floorplan network to improve the reconstruction and completion of rooms. This difference is evident qualitatively in fig. 5.3 and fig. 5.4. In both figures, each row shows: a comparison between a predicted point cloud occupancy map (RF [17] occ. map) and our Nadir floorplan map (Our Nadir map section 5.4); the ground truth floorplan (GT), the compared method prediction (RF [17]), our floorplan prediction just composing the Nadir shapes (our NSO - Nadir shapes only), our full-pipeline floorplan prediction (Our full), our 3D prediction with predicted metric information (Our 3D) and metric aspect. The proposed scenes highlight how much of the error in the compared solutions comes from the deterioration of structural information in the occupancy map (RF occ. map), as opposed to the Nadir floorplan, where, instead, planarity, sharpness, and other structural features are already visible from single-view processing (our NSO - Nadir shapes only). These differences become more pronounced in the real-world case (fig. 5.4), where the contribution of disocclusion and regularization from multi-view integration (section 5.5) also becomes more evident (Our full).

5.6.5 Ablation study

We further analyze our design choices in the ablation study in table 5.3). The first two rows (NS and UNC+NS) summarize the performance of our pipeline without the floorplan integration network (section 5.5), highlighting the effect of performing the uncluttering of the scene (section 5.4). In the first case (NS), the Nadir shapes are obtained from cluttered scenes (i.e., intermediate depth D_e includes clutter), while the second case (UNC+NS), adopting the uncluttering layers, corresponds to the

Method	Room			Corner			Angle		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
NS	0.902	0.897	0.822	0.540	0.526	0.602	0.412	0.516	0.424
UNC+NS (NSO)	0.939	0.930	0.928	0.553	0.542	0.647	0.424	0.522	0.466
NS+NF	0.940	0.936	0.932	0.788	0.768	0.712	0.546	0.688	0.698
UNC+NS+NF	0.943	0.939	0.940	0.834	0.801	0.817	0.746	0.717	0.731

Table 5.3: **Ablation study.** NS: spatially registering of individual Nadir shapes from cluttered depth maps; UNC+NS: NS from uncluttered depth maps; NS+NF: our full pipeline but without depth uncluttering; UNC+NS+NF: our full pipeline with uncluttered depth-shapes.

Our NSO setup of table 5.2. Similarly, in the third and fourth rows (NS+NF and UNC+NS+NF), we show our full pipeline and the effect of uncluttering on the final result, where the fourth configuration (UNC+NS+NF) is our full method (i.e., Our full in table 5.2).

5.7 Conclusion and future works

Our novel deep-learning pipeline proves capable of reconstructing multi-room environments from a minimum number of registered images (i.e., about one per room). Our approach combines indoor priors with single panorama analysis to obtain a clutter-free geometric representation of single views and transformer-based resolution of major structural occlusions while fusing multiple views. Although we achieve state-of-the-art results for reconstruction with minimal data, some limitations remain, and several aspects can be extended in the future. In the supplementary material, we show, for example, failure cases that arise when some of our priors are not met, for instance when structural parts, such as stairs, become dominant in the scene, generating ambiguity in both the uncluttering process and the geometric reconstruction, or when the image includes outdoor parts and the assumption of indoor environment fully bounded by vertical walls lapses. Another point that we expect to develop more in the future is ceiling modeling since our method can already return a heightmap of the whole floorplan, segmented into rooms. However, the current annotated datasets available for floorplans do not provide an unambiguous representation of the ceiling shapes, forcing other works to impose the AWM [99, 101, 17]. In the future, we expect to exploit and extend the already available annotations (e.g., ZinD [18]) to model 3D floorplans of more complex shapes.

5.8 Bibliographic notes

The contents of this chapter report the paper *Giovanni Pintore, Uzair Shah, Marco Agus and Enrico Gobbetti: NadirFloorNet: reconstructing multi-room floorplans from a small set of registered panoramic images* [103], of which the candidate is the lead author.

This publication was made possible 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 the contribution of the Italian National Research Center in High-Performance Computing, Big Data, and Quantum Computing. The findings herein reflect the work and are solely the responsibility, of the authors.

6 Immersive Exploration of Multi-Room Indoor Spaces

6.1 Problem Definition and Motivation

Photorealistic virtual visits of multi-room indoor spaces like homes and offices provide interactive access to real environments, improving accessibility and enabling diverse applications [138]. In real estate, they let users explore properties remotely [158], while in emergency response or facility management, they help users familiarize themselves with spaces for effective task execution [159]. Across domains, essential actions include establishing presence through visual exploration, recognizing room location within the building, and navigating between rooms [160]. Image-based techniques are the most practical way to generate realistic virtual visits, as manual 3D modeling is complex, costly to create and maintain, and often does not match the as-built or as-inhabited situation. Spherical imaging has become standard, with consumer cameras enabling quick, full-scene capture [28]. With just a few spherical images, ideally one per room, a good coverage of the habitable and walkable space of typical interior environments can be obtained. This scenario is representative of common practices, in particular in the real estate industry, where millions of casual users generate and share virtual tours by capturing about a single image per room with consumer-grade spherical cameras, exploiting a small amount of mutually visible space, e.g., through open doors, for mutual registration [18, 146]. Moreover, viewing 360° images in a Head-Mounted Display (HMD) enables intuitive exploration via head movements, leading to easy-to-use Virtual/Augmented/Extended Reality (VR/AR/XR) applications [161, 162].

While creating virtual visits by connecting a few spherical images placed in a multi-room space is very appealing, current approaches face limitations that reduce the sense of presence and increase creation efforts (section 6.2). First of all, the strict locking of the users to the original capture location and the lack of binocular stereo reduces the sense of presence in indoor environments characterized by short to

medium distances from the viewer to the architectural surfaces and the objects [104]. Moreover, and most importantly, immersive multi-room experiences require not only inferring depth to support parallax effects [138], but also detecting walls, doors, and clutter-free areas to map the environment and enable room-to-room paths preserving presence and spatial coherence [160].

To tackle these challenges, we introduce *PanoFloor*, a deep learning approach that creates floor plans and virtual visits from a minimal set of co-registered, gravity-aligned 360° panoramas. Leveraging the broad contextual information available in a panorama and patterns learned from large annotated datasets, we reconstruct a visual, geometric, and semantic model using a single image per room. The extracted model takes the form of a graph, in which nodes represent multiple center-of-projection (MCOP) viewpoints at capture locations, and arcs define spherical videos following optimal paths through doors and around clutter that minimize disocclusions to enhance visual quality.

Our fully automatic pipeline unfolds in sequential steps, progressively incorporating broader global information (section 6.4). We start from the key assumption that much of the geometric and semantic information of built indoors can be extracted by analyzing the vertical distribution of geometry relative to the ground [99]. From each input 360° image, we predict depth and form a global density map by projecting it onto a common ground plane using the available relative poses. Since, differently from methods relying on dense and precise measurements coming from 3D devices or multi-view capture [101, 17], monocular visual inference at the individual room level may produce inconsistent global results, we refine the density map using a latent diffusion model to enhance geometry and semantics. We then segment the floor plan to extract room layouts, openings, and clutter maps. As a result, we obtain a consistent, plausible floor plan approximation from which to extract a full visual representation. First, we create nodes in the resulting graph using the panoramic depth to produce a stereoscopic MCOP image. Nodes are connected by arcs within or across rooms via doors. For each arc, we compute an optimal path that avoids clutter and minimizes visual discontinuities, generating a photorealistic 360° transition video using a view-synthesis network. The resulting precomputed representation in the form of a structured scene graph of stereoscopic views linked by panoramic videos enables fast, immersive, and coherent exploration across VR devices, including HMDs with minimal graphics capabilities. The extracted floor plan, in addition to being of great value for the documentation of the captured building, can also support VR/AR/XR exploration when displayed as a map of where the user is currently located inside the building. The approach introduces the

following important contributions.

- We introduce a latent diffusion method to refine noisy, incomplete density maps by enhancing their semantic content. Standard generative diffusion models [163], designed and trained for common imagery, are suboptimal for this task due to the sampling and detail requirements of occurrence maps. Our lightweight approach models noise as the difference between the diffusion input and ground truth density maps (section 6.4.2).
- We introduce a lightweight autoencoder tailored for density maps to support latent diffusion processing (section 6.4.2). Standard latent variational autoencoders [163] have higher computational costs and, even after fine-tuning, yield suboptimal results (table 6.3). Our gated and dilated convolutions efficiently capture geometric and semantic indoor details (section 6.7).
- We leverage geometric information to enhance view synthesis and navigation (section 6.5), computing optimal paths that balance geometric constraints (e.g., passing through doors and avoiding clutter) with visual quality (e.g., maximizing reconstruction quality by minimizing disocclusions). Additionally, we extend recent depth estimation techniques [19] to improve reconstruction and reduce visual artifacts in view synthesis.
- We propose an end-to-end pipeline to reconstruct an indoor multi-room model, including walls, openings and clutter map just starting from a set of registered images (section 6.4), unlike competing techniques that require dense point clouds acquired with measurement instruments [99, 101, 17, 22].

Our experiments and performance analyses (section 6.7) show that our methods improve over the state-of-the-art in different tasks, in terms of accuracy, quality, and computational complexity, producing image-based models ready to be interactively explored in HMD devices. We also show that our model, trained on synthetic data, can produce compelling predictions on user-captured images.

6.2 Task-related works

Transforming sparse imagery of multi-room environments into image-based representations for VR/AR/XR exploration involves many research areas. In the following, we focus on the most relevant data-driven approaches for panoramic images.

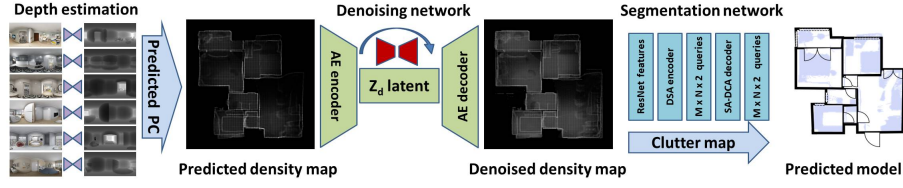


Figure 6.1: **Geometric modeling overview.** Depth maps predicted by a depth network guide both novel view synthesis and structural reconstruction. A denoising network and gated auto-encoder refine a unified density map, which is segmented to extract room layouts, openings, and clutter. The result is scaled to world dimensions for accurate floorplan modeling.

Capture constraints. Our method takes as input a small set of gravity-aligned, registered 360° panoramas. Most, if not all, indoor datasets preserve gravity alignment, as it is guaranteed by tripods or easily achievable in freehand capture by performing upright adjustment using IMU data, image-based vertical direction detection, or a combination of both [15, 60, 18]. Co-registration of images taken in different rooms can be done with printed markers [164], often used also for precise scale detection [18], but can also be achieved from raw capture data. Methods include exploiting the small overlaps that occur through open doors with photogrammetry [165, 166] or RGB-based learning [146], exploiting semantic cues like room types [92], or aligning the detected openings (i.e., doors) [96]. These problems are orthogonal to the ones treated in this paper, and our pipeline does not depend on the methods used for gravity alignment and co-registration.

Multi-room reconstruction. Reconstructing a 3D floor plan from purely visual input in terms of rooms bounded by walls, ceilings, and floors, and connected through passages, provides the 3D geometry and the connections required to implement an effective 3D exploration. Since reconstruction from sparse visual input alone is an ill-formed problem, the regularities present in common environments, like homes, offices, and public buildings, are typically used to aid reconstruction from incomplete or noisy data [138, 75]. Early methods used image processing, architectural priors, and optimization to extract structure from images or depth data [8, 93, 27, 94, 75], but struggled in complex scenes due to reliance on feature detection (e.g., corners and edges) and strong assumptions (e.g., Manhattan World) to support geometric reasoning and optimization. More recent methods combine deep learning for extracting hidden relations with optimization for room reconstruction. Techniques include corner detection with neural networks, followed by integer programming for wall segment reconstruction [100], instance segmentation to identify room regions combined with path solvers for structured polygon generation [98, 154]. Some methods

further refine room shapes using Monte Carlo Tree Search [99], while others use constrained diffusion and graph neural networks for multi-room layout synthesis [155, 156]. In contrast to hybrid approaches, end-to-end deep learning models aim to predict room structures directly from input density maps without complex post-processing. HEAT [101] introduces a bottom-up pipeline that integrates deformable transformers [167] to detect corners and classify edge connections, enabling a flexible representation of room boundaries. Similarly, RoomFormer [17] extends DETR-based architectures to predict structured floorplans from dense density maps, encoding indoor structures as variable-length polygon sequences. SLIBO-Net [102] builds on RoomFormer by incorporating additional geometric and semantic priors, improving accuracy through multi-view consistency and learned post-processing refinements. More recently, PolyDiffuse [22] has introduced a diffusion-based refinement technique that enhances polygonal reconstructions. By treating floorplan reconstruction as a conditional generative process, PolyDiffuse serves as a post-processing step that can refine outputs from various baseline methods, including RoomFormer. Our approach introduces key improvements to the current end-to-end learning solutions. First, we process sparse image inputs directly, predicting dense depth maps and deriving a single density map without the need for 3D point clouds from multi-view capture or measuring instruments. Furthermore, we introduce a novel latent denoising approach to enhance geometric and semantic features in the predicted density map. The proposed method is not constrained to Manhattan-World constraints, and can recover multi-room structures with slanted walls. Enforcing right angles, if needed for some particular application, can be eventually achieved by performing adaptive Manhattan-World corrections as a post-process (i.e., rectifying walls close to meeting at right angles [168]). The generated visual, geometric, and structural data are then used to feed the subsystem generating the visual graph for immersive exploration.

Immersive indoor exploration. Structured representations encode geometry and topology for multi-room scenes and must be enhanced for immersive visualization, enabling stereo cues, motion parallax, and navigation. Single-shot panoramas offer a convenient way to replicate real-world environments but are limited to head rotation with flat appearances, leading to artifacts [104, 162]. To address these constraints, methods targeting restricted viewer motion have been developed, supporting stereo or small head movements by generating compact representations for specific environments. Panoramic images combined with depth maps have enabled diverse view synthesis approaches, such as point cloud rendering [30], depth map-based meshes [169], and blended RGB-D data [170]. Despite advances in immersive deep learning-based

solutions, computational challenges often limit their direct application to embedded devices, imposing remote rendering for Head-Mounted Displays (HMDs) [72, 19]. This has led to the emergence of specialized precomputed representations, such as Layered depth images [171] and multi-plane panoramas (MPI) [172, 173]. For stereo generation, techniques like omnidirectional stereo rendering [11] produce equirectangular images tailored to VR, although peripheral stereo accuracy remains a challenge. Gaze-adaptive rendering offers an alternative by dynamically adjusting depth images [174]. Larger displacements from the capture position require transforming panoramas into 3D models, often leveraging semantics-driven frameworks [77, 175]. However, such methods struggle with real-world environments filled with unrecognized objects. In that case, Neural Radiance Fields (NeRF) [111] and 3D Gaussian Splats (3DGS) [112] provide novel view synthesis, with recent single-view adaptations [176, 95] addressing panorama-based challenges through RGB-D inpainting and mesh refinement. Notably, iterative mesh optimization has shown promise for handling occlusions and underfitted geometries [95]. While methods like PERF [114] focus on generating 3D-consistent novel views and representations, trajectory constraints and blurring in occluded areas persist as limitations. More recent solutions iteratively refine meshes by leveraging inpainting techniques, incorporating occluded color and geometry to improve mesh quality [95]. Generation, however, is still costly and does not provide the quality and resolution required for real-time exploration on an HMD. In our case, we fully precompute a 3D representation made of stereo panoramas connected by transitions on optimized paths. For stereo exploration, we refine a deep-learning solution to generate omnidirectional stereo from a single 360° image [11], exploiting the geometric data available from multi-room reconstruction. For navigation, we introduce an approach to synthesize omnidirectional videos along paths that minimize view-synthesis artifacts in a reconstructed environment. The generation of video transitions from panoramas is a well-researched problem. The classic solution is to match feature points in the start and end panoramas and use them for constraining warping and cross-blending [105]. Since this class of methods suffers from ghosting and deformation artifacts caused by warping the non-matching geometry, later work has exploited depth estimation and view synthesis to achieve a more natural interpolation between wide-baseline panoramas (e.g., [177, 122]). We use the same machinery, as explained above, to synthesize views along transition paths, exploiting the same view synthesis network used for stereo. Indoors, however, straight line interpolation of viewpoints is not sufficient and must be complemented by curved path computation. Classic methods (e.g., [120]) generated a connecting path through collision avoidance and smoothness optimization, eventually taking

into account the maximization of viewpoint selection quality (e.g., [121, 178]). These methods, however, rely on detailed geometry, typically from 3D models or dense multiview capture. Starting from purely visual data, several authors proposed optimizing neural aesthetic metrics [106] on synthesized views to select appealing viewpoints [123, 124] and paths [179], though they focus on single-room navigation. We adopt a similar strategy, but generate paths across rooms by linking door-connected segments, avoiding clutter and minimizing disocclusions.

6.3 Overview

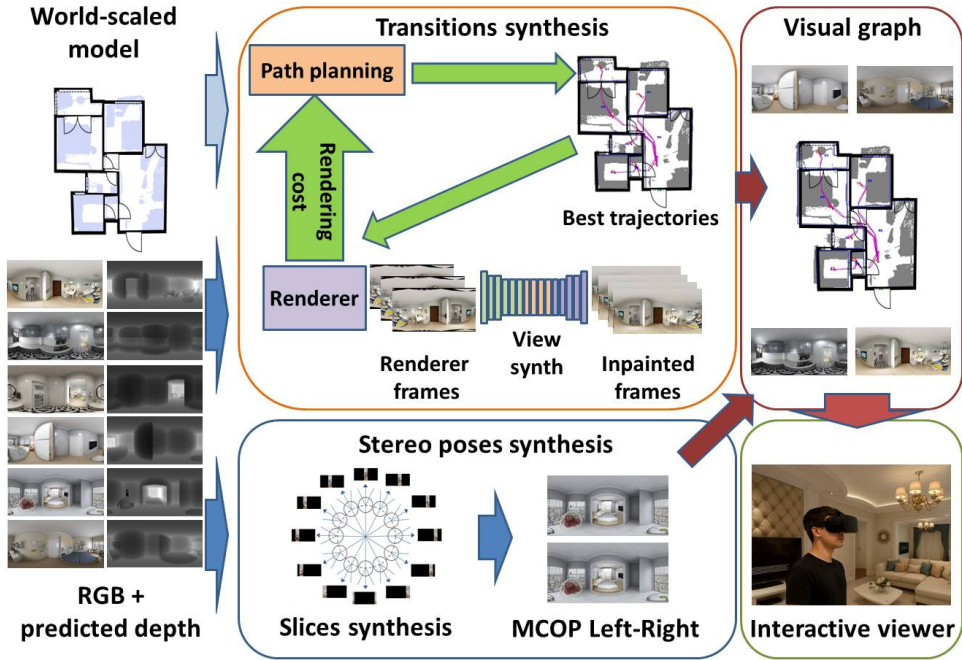


Figure 6.2: **Visual modeling overview.** Visual modeling aligns RGB-D poses to the floorplan, forming a navigable graph with immersive node views and smooth visual transitions. A view-dependent renderer and a view-synthesis network generate stereoscopic images for nodes and 360° videos for arcs, enabling real-time WebXR visualization.

Our method builds an explorable model by combining neural networks in an end-to-end pipeline for geometric (fig. 6.1) and visual (fig. 6.2) modeling. At runtime, it requires only omnidirectional images (at least one per room) and their relative poses. Training uses paired RGB+depth panoramas with pixel-accurate depth that includes clutter as ground truth, as provided by standard indoor datasets like Structured3D [16].

The geometric step (section 6.4) starts with depth prediction by a *depth estimation network* (section 6.4.1), used for both view synthesis and structural reconstruction

via a unified density map (section 6.4.2). A *denoising network* with a gated auto-encoder refines this map, which a *floorplan segmentation network* then converts into room layouts, openings, and clutter (section 6.4.3). The model is scaled to world dimensions using data gathered in density map computation (see *world-scaled floorplan* in fig. 6.2).

In the visual modeling step, original images and predicted depth maps compose RGB-D poses that are aligned to the world-scaled floorplan to create a navigable graph (fig. 6.2), with nodes as stereoscopic omnidirectional views and arcs as transitions. Imagery is generated using a *view-dependent renderer* and a *view-synthesis network*. For each node, we generate an MCOP omnidirectional stereoscopic image (section 6.5.2), and for each arc we exploit the inferred floorplan to determine optimal trajectories, considering walls, openings, clutter, and a renderer-estimated visual disocclusion cost (section 6.5.1). Selected trajectories are used to synthesize interactive 360° videos. The final navigable graph is WebXR-ready for real-time immersive viewing on all devices using fully precomputed data, with the floorplan also serving for a variety of purposes, including the display of a navigable map for VR/AR/XR that also shows where the user is currently located inside the building.

6.4 Geometric modeling

Geometric modeling of the multi-room environment is performed through a sequential approach that progressively incorporates broader global information. First, we determine the shape of the visible portion of each room by associating a depth with each pixel of the captured 360° images (section 6.4.1). Then, exploiting the original co-registration, we project the depths onto a common floor plan to create a density map of the entire multi-room environment. This comprehensive, but partial and noisy, model is then denoised and enhanced in terms of accuracy and semantic richness through a novel latent diffusion-based approach (section 6.4.2). Finally, the global density map is segmented into a floor plan made of rooms bounded by walls, floor, and ceiling, and connected through openings, and a clutter map defining the walkable space (section 6.4.3).

6.4.1 Depth maps prediction

Monocular depth reconstruction from each 360° image is the first step in our pipeline and serves multiple purposes. First, the fusion of the inferred scene depth around the observer provides a first geometric approximation of the environment, which is used as a starting point for 3D floorplan generation. Second, depth is exploited for

occlusion-aware reprojection and disocclusion handling in the construction of the visual model, both for supporting stereo and viewpoint translation and for creating transition paths. Many methods exist for estimating depth from equirectangular indoor images. We build upon the depth estimation branch of a state-of-the-art architecture, named ADM [19], originally designed for depth and layout estimation in view synthesis. The depth estimation network is designed around gravity-aligned features (GAFs), built with asymmetric convolutions, to take into account the fact that world-space vertical and horizontal features have different characteristics in man-made environments (see original work for details [19]). We further build on this concept by including an additional loss term that compares predicted and ground truth density maps obtained by counting the occurrences of the visible 3D points on the floor plan. The loss function $\mathcal{L}_d$ of the predicted depth compared to ground truth (see section 6.7.1 for details) thus becomes:

$$\mathcal{L}_d = \mathcal{L}_e - 0.5\mathcal{L}_{ss} + \mathcal{L}_{om}. \quad (6.1)$$

where $\mathcal{L}_e$ and $\mathcal{L}_{om}$ are respectively, the *Adaptive Reverse Huber Loss* [45] for the predicted depth and for the density map and $\mathcal{L}_{ss}$ is the Structural Similarity Index Measure (SSIM). Explicitly including the density map term allows us to drive the network to produce depths that generate good quality projections on the floor, which is important for our floor plan estimation steps. As a result, the adopted approach achieves both quantitative and computational state-of-the-art performances, as reported in section 6.7.1 and section 6.7.2.

6.4.2 Multi-room density map reconstruction

To fully take advantage of deep learning methods in floorplan reconstruction, it is a common practice to convert 3D point clouds into an intermediate 2D representation, often named density map [98, 101]. Such representation is typically a fixed-size map (e.g., 256×256 in common implementations and benchmarks), where the original point cloud is normalized and scaled to fit the image size. In our work, we first convert each depth map into a local point cloud through a spherical to Cartesian transformation, then we project all of them into a floorplan density map [101] using the co-registration information. However, such a predicted density map is noisy and incomplete, and must be processed to obtain a full, plausible representation of the entire multi-room environment. Our proposed solution is to adopt a novel latent diffusion-based approach to enhance them.

Working in latent space requires the definition of an appropriate auto-encoder [163].

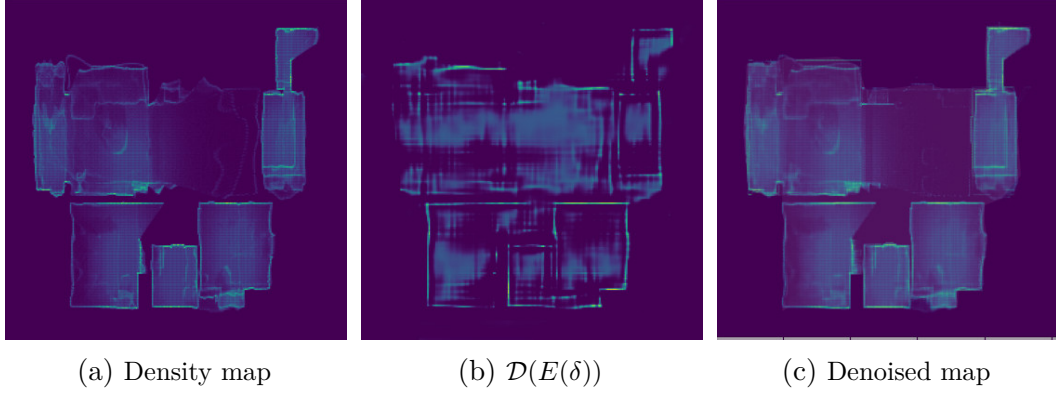


Figure 6.3: **Density denoising.** Comparison between predicted (fig. 6.3c) and denoised (fig. 6.3c) density map (images enhanced for illustration purposes). Although the two maps may look the same, the denoising process recovers some of the missing structural information in the latent space as noise $E(\delta)$. To better show, such latent difference has been decoded with the autoencoder in fig. 6.3b.

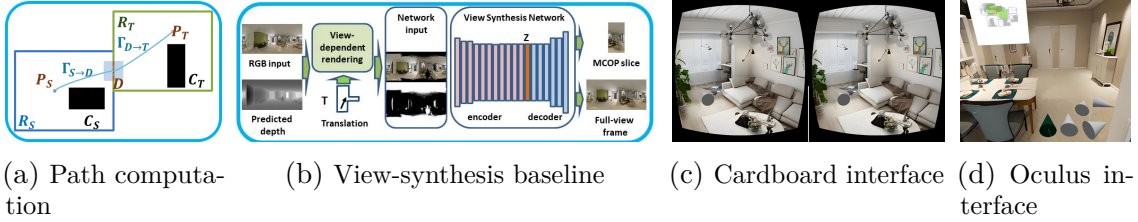


Figure 6.4: **Visual modeling main tasks.** Visual modeling finalizes the geometric model by constructing a navigable graph with immersive visual content for nodes and transitions. fig. 6.4a shows the estimation of optimal navigation trajectories between panoramic viewpoints. This process accounts for structural constraints, including openings and obstacles, while minimizing visual discontinuities and occlusions during transitions. fig. 6.4b shows the baseline [19] adopted for view-synthesis, which, in our use, can output spherical slices to compose MCOP or full-view spherical images for transitions. fig. 6.4c and fig. 6.4d show HMD user interface.

While very effective foundational variational autoencoders exist [180], they are designed and trained on standard imagery for generative tasks. Density maps, which count occurrences of points at every pixel, have, however, different characteristics. Thus, we designed a lightweight auto-encoder targeting only this data type. The selected lightweight architecture, dubbed *GAE*, is based on gated [107] and dilated [181] convolutions. While similar schemes have proven effective for image-to-image view synthesis, we introduce some differentiating features, see fig. 6.4b.

In our network, gating functions as a *self-attention weight mask*, unlike in inpainting, where the mask is provided to specify missing pixels. Here, the network is designed to encode a latent representation z and decode it into a refined density map.

The encoder $\mathcal{E}$ includes 6 LWGC (light-weight gated-convolution [107]) layers, followed by 6 repeated dilations (as view-synth dilations in section 6.5.2), thus increasing the area that each layer can use as input. As a decoder $\mathcal{D}$, 4 LWGC layers restore the original density map resolution, followed by a *Sigmoid* activation function, instead of *tanh* adopted for view-synthesis:

$$z_{\text{gt}} = \mathcal{E}(d_{\text{gt}}) \quad \hat{d}_{\text{gt}} = \mathcal{D}(z_{\text{gt}}) \quad (6.2)$$

Defining a ground truth density map as $d_{\text{gt}} \in \mathbb{R}^{1 \times 256 \times 256}$, ideally, $\hat{d}_{\text{gt}} \approx d_{\text{gt}}$. In our experiments, z_{gt} is $1 \times 32 \times 64 \times 64$, starting with 32 latent channels in the first convolutional layer (section 6.5.2). The GAE autoencoder is trained on the ground truth density maps provided by Structured3D[16] with the following loss function:

$$\mathcal{L}_{\text{GAE}} = \mathcal{L}_d(\mathcal{D}(\mathcal{E}(d_{\text{gt}})), d_{\text{gt}}) + \lambda \cdot \mathcal{L}_{\text{ss}}(\mathcal{D}(\mathcal{E}(d_{\text{gt}})), d_{\text{gt}}) \quad (6.3)$$

where $\lambda = 0.05$ and, similarly to eq. (6.1), $\mathcal{L}_d$ is the *Adaptive Reverse Huber Loss* [45] and $\mathcal{L}_{\text{ss}}$ is the Structural Similarity Index Measure (SSIM). As a result, the GAE network has much less computational and training complexity than foundation autoencoders [163], as shown in table 6.3.

Once the latent space is defined, we model the noise in the input as the difference $\delta = d_{\text{in}} - d_{\text{gt}}$, where d_{in} is the input, noisy density map, which implies that the noisy latent representation can be approximated as:

$$z_{\text{in}} = z_{\text{gt}} + E(\delta) \quad (6.4)$$

In other words, predicting the latent $z_d = E(\delta)$ means predicting the missing information in z_{in} to be more similar to z_{gt} . According to this, the forward diffusion process corrupts z_{gt} by adding Gaussian noise:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, I) \quad (6.5)$$

where t is the diffusion timestep and $\bar{\alpha}_t$ are predefined schedule coefficients. The latent diffusion model, parameterized by a UNet $\epsilon_\theta(z_t, t)$, is trained to predict such a noise component. Inference, given z_{in} , applies denoising by inverting eq. (6.5):

$$\hat{z}_d = \frac{z_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(z_t, t)}{\sqrt{\bar{\alpha}_t}} \quad (6.6)$$

to estimate $\hat{z}_d$ from z_{in} as input. Finally, we reconstruct the denoised density map as the decoding $\hat{d}_d = \mathcal{D}(z_{\text{in}} - \hat{z}_d)$.

Assuming a frozen GAE autoencoder, we train the latent diffusion with the loss:

$$\mathcal{L}_{\text{LDM}} = |\hat{z}_d - z_d|_1 + 0.1|\mathcal{D}(z_{in} - \hat{z}_d) - \mathcal{D}(z_{gt})|_1 \quad (6.7)$$

As an example, fig. 6.3 shows a comparison between the predicted (fig. 6.3a) and the denoised (fig. 6.3c) density map. While the two maps appear visually similar (with enhancements applied for better illustration), the denoising process restores missing structural details in the latent space as noise $E(\delta)$. To better illustrate this latent difference, fig. 6.3b shows its decoding in the density map space.

6.4.3 Floorplan segmentation

Taking the denoised multi-room density map as input, we formulate floorplan segmentation as the prediction of a polygon set, where each polygon can represent a *room*, defined by an ordered sequence of 2D vertices, or an *opening*, encoded by a degenerate polygon with only two vertices. Rooms and openings are predicted as a set of queries, $Q \in \mathbb{R}^{M \times N \times 2}$, using a deformable transformer [167]. We define M as the maximum number of polygons (rooms and openings) and N as the maximum number of vertices per polygon ($M = 50, N = 40$ in our experiments). Each vertex is associated with a binary value indicating its validity. The *segmentation network* (fig. 6.1) carries out the prediction by integrating multi-scale feature extraction, flattening, deformable self-attention (DSA) for transformer encoding, and a combination of self-attention and deformable cross-attention (SA-DSA) for decoding.

The first layers extract multi-scale features (four levels) from the density map using a *ResNet* scheme [44], flatten them to a feature sequence to serve as input to a six-layer transformer encoder. Each encoder layer consists of a multi-scale deformable *self-attention* module and a simple feed-forward network [167] (DSA in fig. 6.1). Similarly to encoding, decoding is performed by six stacked layers; however, unlike the encoder, each layer consists of a self-attention module, a multi-scale deformable cross-attention module, and a feed-forward network (SA-DSA in fig. 6.1). The decoder output is passed to a shared feed-forward network to predict binary class labels c for each query, indicating its validity as a corner. Finally, the decoder outputs a $M \times N \times 2$ polygons map and a $M \times N \times 1$ classification map, from which a set of valid 2D polygons is extracted [17]. While the ground truth contains an arbitrary number of polygons with an arbitrary number of vertices, the transformer decoder always outputs M polygons of N vertices, setting the invalid ones to 0. We train the network with the Structured3D [16] dataset, which provided fully annotated floorplans including geometries and semantic elements (see section 6.7.1 for details),

using the following recently introduced losses [17, 22]:

$$\mathcal{L}_{floor} = \sum_m^M (2\mathcal{L}_{cls}^m + 5\mathcal{L}_{coord}^m + \mathcal{L}_{ras}^m) \quad (6.8)$$

where $\mathcal{L}_{cls}$ is a standard binary cross-entropy (i.e., class loss), $\mathcal{L}_{coord}$ (i.e., coordinates loss) is the L_1 distance, $\mathcal{L}_{ras}$ (i.e., raster loss - only for rooms) is the Dice loss [182].

Quantitative and qualitative results show how our solution predicts rooms, doors, and windows with state-of-the-art accuracy while starting from purely visual input (section 6.7.2, section 6.7.3), ready to be exploited and integrated in the immersive visual model section 6.5.

For our application, in addition, we complete the model with a clutter map, defining the walkable space in the environment used to generate our transitions. To this end, we calculate a partial density map from the fraction of point clouds below the observer’s horizon, therefore removing ceilings and hangings, and consider as clutter the points on the density maps that have occurrences above a certain threshold (10% in our experiments), to remove from clutter the walkable floor (which counts as at least one occurrence) and eventual projection noise.

6.5 Visual modeling

Visual modeling completes the geometric model by creating a navigable graph (section 6.5.1), including the immersive visual content for nodes and transitions (section 6.5.2). The final output is ready for interactive viewing with a WebXR-compatible device (section 6.6).

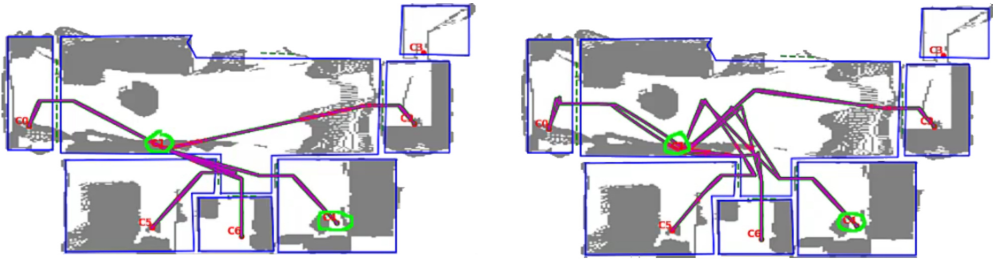


Figure 6.5: **Path optimization.** At initialization, paths connect source to target poses, passing through doors (left). After optimization, paths avoid walls and clutter and minimize disocclusions (right). The green mark identifies a very difficult case, in which the detected door is invisible in one of the views.

6.5.1 Door-based graph topology

From our floorplan segmentation (section 6.4.3), each door is identified by a small two-vertex segment. We first verify that the door connects exactly two distinct rooms. If so, we associate the door with the pair of *nodes* (i.e., cameras) that occupy those rooms. This yields a high-level graph structure: each door triggers a potential connection (arc) between two nodes. This representation is already sufficient to automatically support multi-room navigation through exploration at the original location and teleportation to connected rooms. To improve location and movement awareness, we refine each arc into a continuous path composed of two segments: (1) *Source-to-Door*: a parametric curve from the source viewpoint to the door location; (2) *Door-to-Target*: a parametric curve from the door location to the target viewpoint. Both sub-paths must be stitched into a single smooth trajectory at the door location. To ensure continuous derivatives (i.e., no abrupt orientation changes), we model each sub-path with a *Hermite spline* [183] (see fig. 6.4, left).

Constructing a single multi-segment path from the source camera to the target camera via the door involves optimizing several control points and tangents. We use *dual simulated annealing* (DSA) to minimize a custom cost function [184, 185]. DSA randomly perturbs the free control points and tangent vectors over multiple iterations, gradually “cooling” to a stable solution that locally minimizes our *cost function*. Let the full path Γ be a concatenation of Hermite segments Γ_k , each defined by control points and tangents $(\mathbf{P}_k, \mathbf{P}_{k+1}, \mathbf{T}_k, \mathbf{T}_{k+1})$. The cost function combines four penalty components: *Path Length* C_ℓ favors shorter routes, reducing travel time and video frame count, computed using closed-form Hermite spline integration; *Disocclusion Penalty* C_d accounts for novel-view artifacts by sampling geometry along Γ and counting newly visible pixels in each frame; *Clutter Penalty* C_c discourages paths intersecting high-clutter areas, counting how many sampled path points intersect high-density clutter regions from the precomputed clutter map; *Wall Penalty* C_w ensures that Γ does not cross walls and remains within a safe distance from them, counting how many path points fall within a threshold distance from wall boundary polygons. The complete cost is:

$$C(\Gamma) = \omega_\ell C_\ell(\Gamma) + \omega_d C_d(\Gamma) + \omega_c C_c(\Gamma) + \omega_w C_w(\Gamma), \quad (6.9)$$

with weights $\omega_\ell = 0.01$, $\omega_d = 1.0$, $\omega_c = 100.0$, and $\omega_w = 100.0$. The weights were empirically determined with a few trials and kept fixed for all the experiments.

Once the optimized sub-path $\Gamma_{\text{src} \rightarrow \text{door}}$ is found, we repeat the procedure for $\Gamma_{\text{door} \rightarrow \text{tgt}}$. Merging them produces the complete door-based arc. For each segment, we

sample a discrete set of points in parametric space (e.g., 10–100 points, depending on distance and resolution needs). These sampled points and their corresponding camera orientations form the keyframes used by our view-synthesis network to generate intermediate 360° frames.

By minimizing the *Disocclusion Penalty*, we ensure that the path remains in areas that are well sampled from the original viewpoint, reducing the need to generate novel data, thereby producing paths that give the impression of motion towards the door when leaving a room and away from the door when entering a new one. Good reconstruction in the transition area is only possible when a reasonably large mutually visible area exists, which may not be the case if one of the views does not see the door (e.g., it is around a corner - see fig. 6.5, green highlight). If such a case is detected, we trim the portions of transitions closest to the occluded door to minimize visual artifacts and preserve the viewing experience.

6.5.2 Immersive view synthesis

To generate immersive visual content, it is necessary to synthesize novel panoramic images translated relative to the input poses. In our approach, we leverage a scheme already proven to be effective in VR [19], which integrates a *view-dependent renderer* and a *view synthesis* network. This architecture, illustrated in fig. 6.4b, is very lightweight and suitable for an efficient synthesis of the many images required for visit precomputation. Compared to the inpainting baseline (see [19]), we introduce the minimization of the inpainting area through our path optimization method (section 6.5.1), and incorporate structural constraints directly into the depth (section 6.4.1). The view-dependent renderer exploits the predicted depth to convert source pixels to 3D points and translates and reprojects them to their new location in the target equirectangular image. The view synthesis network inpaints the missing disoccluded pixels.

At each node, we adopt a stereoscopic MCOP model [11] that uses these networks to generate panoramic slices, placing each slice along a circular path matching eye movement during head rotation. Each slice covers enough angular range for both eyes and contextual reconstruction (fig. 6.4b). The final stereoscopic MCOP pair is efficiently composed by blending these precomputed slices in equirectangular format. For a typical human-sized configuration, assuming a head radius of 100mm and an interpupillary distance (IPD) of 65mm, we define a 45° portion of the image as sufficient to cover both eyes. To provide additional context for reconstructing missing areas and supporting eye convergence at finite distances, we expand this region to approximately 56°.

At each arc, instead, we render full-view, monoscopic equirectangular frames, following the optimal trajectories computed in section 6.5.1. In our tests, we have experienced that monoscopic video is sufficient to ensure immersive exploration in motion, while also limiting the sense of discomfort from moving stereo or other artifacts. As a final step, we also perform upsampling [186] of all generated images to match the resolution of the display. This is because, currently, our synthesis is performed at a resolution smaller than the display size (i.e., a vertical slice resolution of 512 pixels vs 2048 for a typical headset). This limitation is not due to our scalable network architectures, but rather, to limitations in available ground-truth training sets (see section 6.7.1 for details).

6.6 Interactive VR exploration

As a proof-of-concept, we develop a virtual tour prototype built with WebXR and a minimal interface using *Three.js*. At runtime, the application imports the previously constructed graph data structure, wherein each *node* is an MCOP-stereo panoramic view of a given room, and each *arc* is a transition path, represented by a pre-rendered 360° video, which corresponds to the optimized trajectory computed in section 6.5.1. Once a user puts on a VR headset (e.g., *Meta Quest* or *Cardboard*), they are placed in a node’s stereoscopic panorama. The interface displays arrow-shaped placeholders indicating all possible door connections to adjacent rooms. Clicking or gazing at one of these placeholders triggers the 360° transition video associated with that arc, smoothly transporting the user to the corresponding target node. This design mimics intuitive room-to-room travel while avoiding abrupt scene jumps, since the underlying arc videos handle the in-between viewpoints and occlusions in a visually coherent manner. To satisfy common constraints on VR headsets, each transition video is encoded at a resolution of 5760×2880 pixels with a frame rate of 30 fps, meeting hardware performance requirements without compromising immersive quality. The system thus achieves fluid scene traversal across multiple rooms, letting users freely inspect each room’s panoramic stereo view, then proceed to any connected room by activating the corresponding door arc. In practice, the entire experience runs within a standard WebXR-capable browser, making it readily deployable to standalone headsets or desktop devices with VR support.

6.7 Results

The presented approach was implemented using *PyTorch* [187] and WebXR and has been tested on a large variety of indoor scenes.

6.7.1 Setup and computational performance

For training and benchmarking our solutions, we mainly exploit Structured3D [16] (see section 3.4). That dataset, in addition to being the larger dataset providing full annotations (including depthmaps and full semantic), provides the most accurate ground truth possible, which is essential, for example, for creating geometric-consistent density maps. We used Structured3D for training and testing the depth estimation network (section 6.4.1), the autoencoder and diffusion networks (section 6.4.2) and the floorplan segmentation network (section 6.4.3). In all cases we follow the official splittings [16]. To train and test the view-synthesis network, instead, we exploit PNVS[72], which is a visual extension of Structured3D scenes providing, for each original panoramic image, three views translated by 0.2-0.3m along random directions, and three views translated by 1.0-2.0m. All models have been trained at the native Structured3D resolution, that is 1024×512 , which have been upsampled to 4096×2048 for the final visual output, using *Real-ESRGAN* (with model *realesr-animevideov3*) [186] It is noteworthy that although the model is trained on synthetic data, it works successfully for reconstructing real-world multi-room scenes, as shown in fig. 6.9. Here we exploit scenes from ZinD [18] (see section 3.4). It should be noted that the proposed real-world scenes are intentionally evaluated in a zero-shot setting — that is, predictions are obtained by training on a synthetic dataset and directly applying the model to entirely different data, potentially in scenarios where ground truth is unavailable. For this experiment, it was not necessary to acquire new ground truth data.

Method	Params↓	FLOPs↓	Inf. time↓	MAE↓	RMSE↓	δ_1 ↑
Panoformer [62]	20 M	78 G	17 ms	0.254	0.793	0.747
EGFormer [144]	15 M	74 G	16 ms	0.220	0.684	0.797
Elite360D [20]	25 M	65 G	14 ms	0.148	0.496	0.874
ADM [19]	29 M	79 G	18 ms	0.080	0.124	0.968
Our depth e.	23 M	38 G	7 ms	0.045	0.135	0.978

Table 6.1: **Performance of depth inference.** We show our performance compared to other state-of-the-art works for a 512×1024 image. All methods were trained and tested using the same conditions. Other results are the same as reported in recent publications [20], while ADM [19] has been retrained by us.

Metric	ADM +PolyDiffuse	ADM +RoomFormer	Our plain density	Our denoised density
Room				
Prec	0.880	0.863	0.882	0.943
Rec	0.851	0.845	0.864	0.918
F1	0.871	0.864	0.873	0.930
Corner				
Prec	0.697	0.679	0.686	0.787
Rec	0.660	0.662	0.677	0.760
F1	0.675	0.652	0.681	0.782
Angle				
Prec	0.662	0.601	0.608	0.705
Rec	0.629	0.589	0.601	0.682
F1	0.612	0.578	0.605	0.694
Openings				
Prec	-	0.556	0.567	0.635
Rec	-	0.588	0.606	0.672
F1	-	0.571	0.582	0.653

Table 6.2: **Floorplan performance and ablation.** We show our quantitative performance on Structured3D (S3D) [16] (Our density denoised). For a fair comparison from predicted images, we adopt an alternative SoA depth prediction method to generate competitors’ density maps (ADM [19]). We compare then with ADM [19]+Polydiffuse [22], which is currently the method that performs best for segmentation of polygonal rooms, and RoomFormer [17], which is the same baseline of PolyDiffuse but provides full floorplan segmentation (ADM [19]+RoomFormer [17]). Finally, as an ablation experiment, we compare to our pipeline but without denoising (Our plain density).

We train all the networks with the Adam optimizer [188], with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and an adaptive learning rate from 0.0001, on a NVIDIA RTX 4090 (24GB VRAM) with a batch size of 8. The average training time on the same machine is 10 *ms/image* for the depth network with a batch size of 8 and 300 epochs, 22 *ms/image* for the view-synthesis network with a batch size of 8 and 300 epochs, 402 *ms/image* for the floorplan network with a batch size of 10 and 500 epochs, 4 *ms/map* for the gated autoencoder with a batch size of 32 and 1000 epochs, 140 *ms/map* for the diffusion network with a batch size of 64, 1000 epochs of 1000 steps. Both the autoencoder and the diffusion model are extremely lighter than common foundation models [163]. Below we collect a comparison based on our experiments in a table 6.3. Here *VQ-VAE* is a standard implementation of a variational autoencoder without hierarchical latent [180] and *Fast latent diffusion* is a lightweight implementation of standard latent diffusion [189]. In table 6.1 we summarize computational complexity for depth inference, compared to other state-of-the-art works, on a NVIDIA RTX 4090 with 24Gb VRAM. Under the same setting, inferring a floorplan from a $1 \times 256 \times 256$ density map takes 4*ms*, and a 100-step denoising takes 142*ms*. Starting from the geometric

	Simple VQ-VAE	GAE (our)	F. latent diffusion	Our denoise
Params↓	20 M	0.11 M	100 M	7.6 M
FLOPs↓	20 G	0.68 G	50 G	31 G

Table 6.3: **AE and diffusion computational performance.** We show our computational performance compared to a standard solution for the same density map size of 256×256 .

model, visual modeling generation time depends on the number of stereoscopic poses and transition frames. The inference time for a full-view equirectangular frame of 1024×512 is $10ms$, with $18ms$ additional time for super-resolution upsampling.

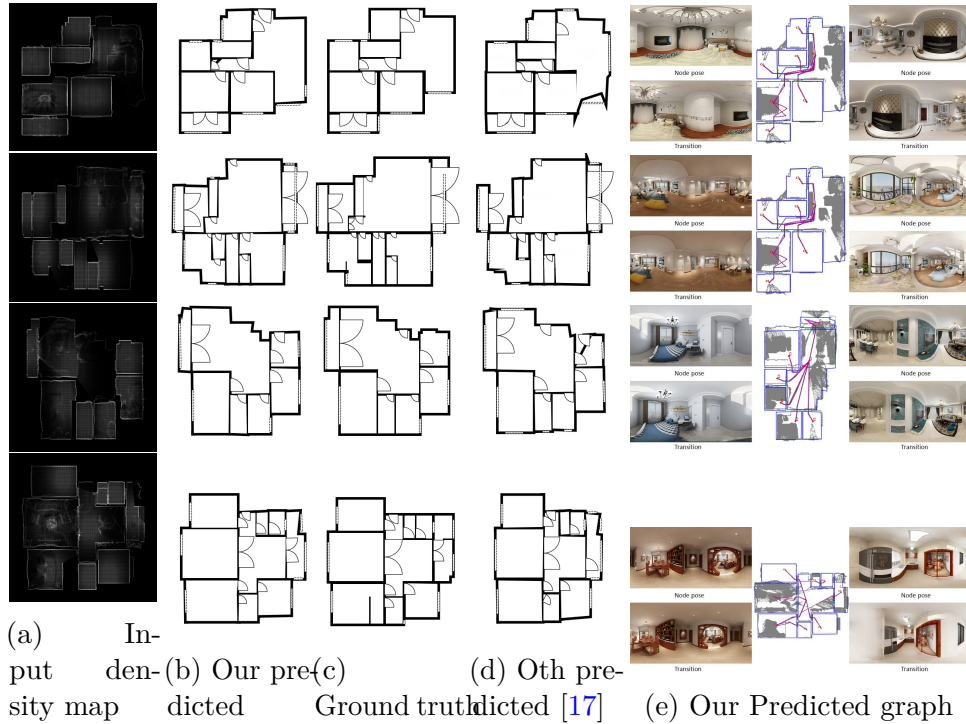


Figure 6.6: **Examples of model reconstructions from Structured3D [16].** For each scene, we show the denoised density map, our predicted model from denoised with openings, the ground truth, the prediction from original noisy density map, and the multi-room scene graph in world coordinates (several rotated for better fitting in the illustration), including clutter maps, computed trajectories and screenshots at primary and transition poses. Notably, our method distinguishes between doors and windows based on relative queries (section 6.4.3).

6.7.2 Quantitative performance analysis

Following the pipeline order, we first show results about depth estimation and view-synthesis, followed by performances specific to our major technical contributions related to the reconstruction of the entire floorplan for navigation. In table 6.1

we show our performance in terms of depth estimation on standard metrics (mean absolute error (MAE), root mean square error of linear measures (RMSE), and relative accuracy δ_1 [20]), compared to the latest state-of-the-art approaches. Although our solution for depth estimation is an enhancement of ADM [19], it outperforms all the competitors in both accuracy and computational performance (table 6.1) for the restricted task focused on structured data and scenes.

In addition, the higher efficiency in depth estimation also translates into better performance in view synthesis. As an example, we show in table 6.4 the performance of our view-synthesis network, compared to the baseline GVS [19] and a baseline for the synthesis of MCOP slices [11], on PNVS [72] benchmark.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
MCOP [11]	18.22	0.789	0.252
GVS [19]	22.97	0.817	0.178
Our	23.02	0.824	0.187

Table 6.4: **View-synthesis performance.** We show our quantitative performance compared to other state-of-the-art works.

table 6.2 summarizes a comprehensive quantitative evaluation of our approach on the Structured3D [16] dataset, focusing on the effectiveness of our denoised density maps in reconstructing indoor environments. To ensure a fair comparison, we generate competitor density maps using an alternative state-of-the-art (SoA) depth prediction method, ADM [19], which serves as a robust baseline for depth-based density estimation (see table 6.1). We then compare our approach against two prominent approaches. The first is ADM combined with PolyDiffuse [22], which currently achieves the best segmentation performance for polygonal room layouts by leveraging diffusion-based polygonal generation. Since PolyDiffuse (as ours) builds upon the RoomFormer [17] architecture, an advanced transformer-based model for full floorplan segmentation (i.e., including room type and openings), we also compare our solution to the combination of ADM and RoomFormer.

Furthermore, to assess the contribution of our denoising step, we conducted an ablation study by comparing our full pipeline to a variant that directly utilizes raw density maps without denoising (Our plain density). This experiment helps isolate the impact of our denoising strategy on downstream tasks such as room segmentation and floorplan reconstruction. For each ground truth room, we find the best-matching reconstructed room among all predictions in terms of Precision, Recall, and F1 scores about Room (i.e., found or not), Corner (i.e., room corners), Angle (i.e., room angles) and openings (i.e, found or not). In all metrics, our proposed pipeline outperforms the other solutions.

6.7.3 Qualitative performance analysis

fig. 6.6 shows several examples of model reconstruction from the Structured3D [16] dataset, for which a comparable ground truth is available. For each scene, we show the input denoised density map, our predicted model including openings, the ground truth model, and the predicted multi-room scene graph in world coordinates, including clutter map and computed trajectories (i.e., last column). Besides the reconstructed scene graph, we show representative equirectangular screenshots at the input poses (i.e., node pose) and synthesized screenshots along the computed trajectory. The third column shows, as comparison, the predicted model using ADM [19]+RoomFormer [17] (see table 6.2), which provides an output comparable with ours. It should be noted that our approach can also distinguish between doors and windows, according to the relative queries (section 6.4.3).

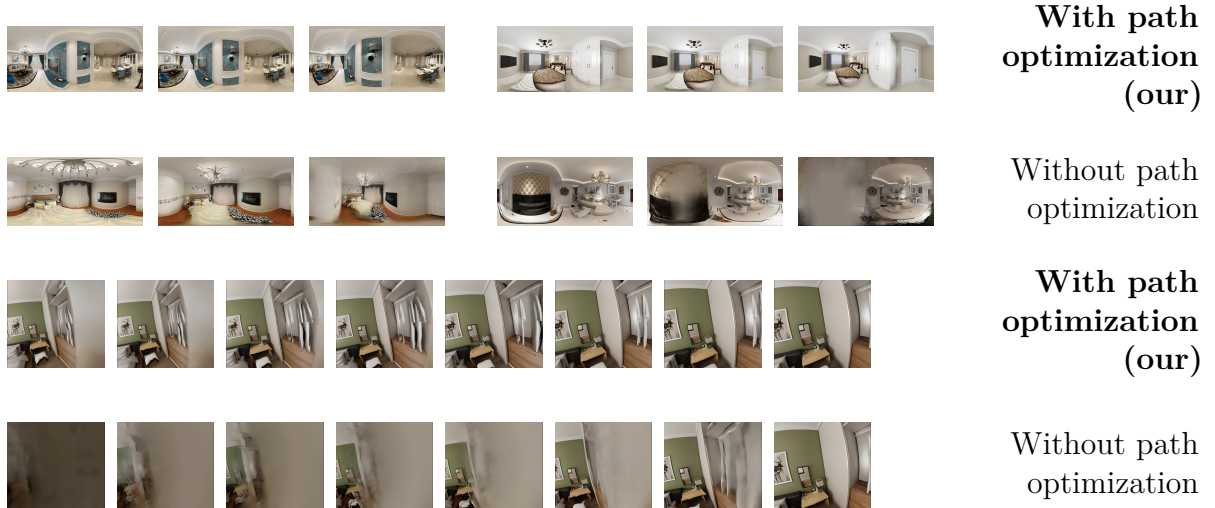


Figure 6.7: **Transition comparison.** Comparison between synthesized equirectangular views (top) and live frames (bottom) with and without our optimized trajectory. Without path optimization, the trajectory passes through walls and clutter to reach the final viewpoint.

Additionally, fig. 6.9 illustrates the capabilities of our approach on real-world captured scenes, leveraging ZinD [18] to obtain multi-room spherical captures. These experiments illustrate the generalization from synthetic to real-world scenarios, demonstrating the capability of domain transfer. Additionally, our results showcase the potential for zero-shot reconstruction, where the model successfully reconstructs previously unseen real-world environments without requiring additional fine-tuning or retraining on real data.

We leverage geometric information to improve view synthesis and navigation (section 6.5), computing optimal paths that account for both geometric constraints—like

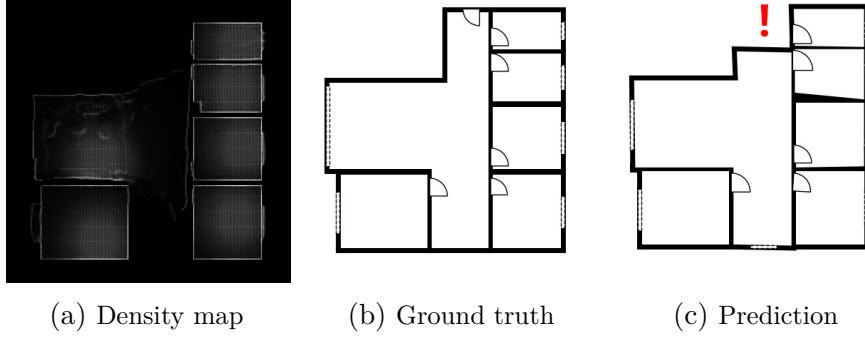


Figure 6.8: **Panoramic floorplan failure example.** We show an example where the accuracy of the reconstruction affects the final result. Although the result appears to be a good reconstruction at first glance, with all doors present, the incorrect positioning of the wall at the top prevents access to the upper right room.

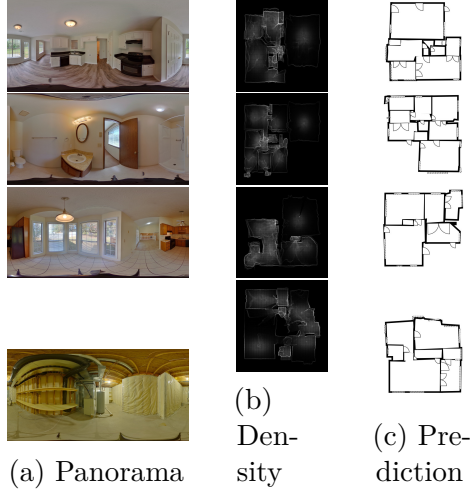


Figure 6.9: **Examples of reconstructions on real-world captures**, using our models trained on synthetic data [16] for prediction. Such experiments demonstrate the capabilities of domain transfer and zero-shot reconstruction.

passing through doorways and avoiding obstacles—and visual quality by minimizing disocclusions. We demonstrate this in fig. 6.7. In the top two rows of the figure, we compare the complete synthesized equirectangular views generated with and without the optimized trajectory designed to avoid clutter and obstacles. The bottom two rows show the same comparison using live frames captured along the respective trajectories. Notably, in the case without our computed path, the trajectory passes through walls and cluttered areas to reach the final position, demonstrating the importance of our path optimization.

6.7.4 Failure cases

The process of reconstructing numerous structural elements, such as walls, doors, and windows, from a predicted density map presents many difficulties and possibilities for error. In many cases, (see fig. 6.6 and fig. 6.9), errors about the shape of rooms or the exact location of an opening do not prevent the construction of an immersive graph, bearing in mind precisely that the purpose of the reconstruction is to support exploration, and not to produce a detailed reconstruction. However, there are cases where the accuracy of reconstruction affects the final result, as in the case we selected in fig. 6.8. Here, although the result may seem a good reconstruction and even all the doors are present, the wrong positioning of the wall at the top makes it impossible to reach the upper right room.

6.8 Conclusions

We presented a method for immersive modeling of complex indoor environments using a minimal set of registered 360° panoramas. Unlike prior single-room or post-constrained approaches, we reconstruct multi-room scenes by refining predicted depth with diffusion denoising and segmenting structure with transformers. By unifying geometry and semantics in a global density map, we recover a floor plan that documents the building and support path planning and immersive navigation in cluttered multi-room spaces. Since our final scene graph contains fully precomputed MCOP panoramic images and panoramic transition videos, scenes can be experienced immersively on commodity HMDs.

The discussed analyses and experiments show that the method achieves state-of-the-art reconstruction from sparse inputs and supports compelling immersive visits. Quantitative measures in multi-room reconstruction, in particular, demonstrate our ability to improve over competing solutions in geometric and structural reconstruction. From the user’s point of view, we currently have only informal feedback on immersion and path planning, and a full user study is an important aspect of future work. In particular, we would like to formally evaluate user satisfaction, improvement in location awareness compared to just teleportation, and quantify any effects of cyber sickness that may arise from reconstruction artifacts.

The primary visual limitations stem from rendering parts that are very distant from the original viewpoint or, even more critically, those that are completely invisible from the original input views. In this regard, one direction for future work is to exploit conditional generative models of latent diffusion instead of deep-learning-based inpainting. We expect that this would allow synthesizing plausible, completely novel

intermediate views and smooth motions through continuous trajectories. This would further close the gap between captured and fully navigable virtual environments.

6.9 Bibliographic notes

The contents of this chapter report the paper *Giovanni Pintore, Sara Jashari, Marco Agus and Enrico Gobbetti: PanoFloor: reconstruction and immersive exploration of large multi-room scenes from a minimal set of registered panoramic images using denoised density maps* [10], of which the candidate is the lead author.

This publication was supported by NPRP-S 14th Cycle grant 0403210132 AIN2 from the Qatar National Research Fund (a member of Qatar Foundation). GP and EG also acknowledge the contribution of the Italian National Research Center in High-Performance Computing, Big Data, and Quantum Computing (Next Generation EU PNRR M4C2 Inv 1.4). The findings herein reflect the work and are solely the responsibility of the authors.

7 Conclusions and Future Work

7.1 Summary of Contributions

This thesis explored how sparse panoramic imagery—often consisting of only one or a few 360° views per room—can support the reconstruction and immersive exploration of large, structurally complex indoor environments. Throughout the thesis, we argued that panoramic images, despite their sparse sampling and inherent distortions, encode rich geometric cues when interpreted under appropriate priors such as gravity alignment, typical indoor structures, and learned shape regularities. Building on these principles, we introduced three major technical contributions, each developed in a dedicated chapter of the thesis.

Deep Depth Estimation from Panoramic Images (Chapter 4). Chapter 4 investigated the problem of depth estimation from a single equirectangular panorama, a task that is both fundamental and challenging due to projection distortions, uneven angular sampling, large structural occlusions, and clutter. The chapter introduced *DDD++*, a structure-aware depth estimation framework that combines gravity-aligned feature encoding with a novel *density-map consistency* loss. This loss models horizontal and vertical projections of the predicted 3D point cloud, extracting structural patterns that correlate with walls, corners, openings, and global room shape.

DDD++ leverages these projections as training-time geometric constraints that guide the network toward metrically coherent and smoother depth predictions, particularly in regions where RGB cues are weak or obstructed. By exploiting indoor priors such as vertical walls, Manhattan/Atlanta tendencies, and the architectural regularities discussed in Chapter 2, the method significantly improves robustness to clutter and occlusions. The resulting depth predictions enable downstream structural reasoning and novel-view synthesis, forming the geometric foundation for the subsequent multi-room modeling stages.

Multi-Room 3D Floorplan Reconstruction (Chapter 5). While single-room understanding is well-studied, multi-room modeling from sparse panoramic imagery remains a major challenge due to minimal viewpoint overlap, strong self-occlusions, and the failure of classical multi-view pipelines. Chapter 5 introduced *NadirFloorNet*, a transformer-based architecture that reconstructs multi-room floorplans by fusing clutter-free nadir projections inferred independently from each panorama.

The chapter presented a two-stage process: (i) a single-view nadir-shape prediction, mapping the panoramic depth into a dense floor-projected representation that captures the structural footprint of each room; and (ii) a multi-view fusion module based on deformable attention, which aligns, aggregates, and reconciles these local representations using the relative positions of panoramas. This formulation directly addresses the sparse-view regime, where the lack of geometric overlap requires the network to reason structurally rather than through correspondence matching.

The resulting method reconstructs coherent multi-room polygons, accounts for room heights, and outputs metrically meaningful 3D floorplans suitable for downstream visualization, analysis, and navigation. This contribution bridges the gap between single-view inference and full-scene modeling under extremely sparse conditions, advancing the field of panoramic indoor reconstruction beyond isolated rooms.

Immersive Exploration of Multi-Room Indoor Spaces (Chapter 6). Chapter 6 extended the reconstruction pipeline by introducing *PanoFloor*, a VR-ready system for immersive, multi-room visualization. The chapter demonstrated how geometric and structural cues derived from sparse panoramas—including depth (from Chapter 4), nadir projections and layouts (from Chapter 5), and density maps—can be combined to produce consistent room boundaries, detect openings, segment clutter, and build navigable connectivity graphs.

The system integrates an autoencoder and diffusion-based denoising strategy for density maps, improving structural detail before polygon extraction. It also employs MCOP-based stereoscopic rendering, door-aware path planning, and view-dependent synthesis to generate smooth transitions between viewpoints. These components enable head-motion parallax, stereo depth cues, and interactive navigation across an entire multi-room environment without requiring dense 3D reconstruction.

Together, these contributions demonstrate an end-to-end pathway from sparse panoramic input to full multi-room reconstruction and immersive VR exploration, showing that panoramic imaging—when paired with appropriate priors and learned representations—is sufficient for producing actionable and perceptually convincing

digital twins.

7.2 Scientific and Practical Impact

Scientifically, this thesis advances the understanding of how panoramic imagery can be used for structural indoor reasoning. The methods developed across Chapters 4–6 illustrate that strong geometric priors—gravity alignment, vertical wall assumptions, characteristic architectural statistics, and nadir-referenced projections—provide a powerful foundation for deep learning models to infer meaningful 3D structure from limited observations.

The thesis shows that, even without traditional multi-view redundancy, spherical imagery encodes patterns related to room shape, openings, occlusion boundaries, and global topology. By exploiting these cues:

- DDD++ demonstrates that structure-aware constraints can significantly enhance depth quality and consistency;
- NadirFloorNet establishes that multi-room reasoning is possible from sparse panoramas through transformer-based fusion of nadir projections;
- PanoFloor demonstrates that these reconstructions can be operationalized for immersive, multi-room VR experiences.

Practically, the contributions reduce acquisition costs and complexity. Instead of requiring LiDAR scans or dense image collections, the entire pipeline relies solely on a small set of consumer-grade panoramas. This makes the approach suitable for real-estate documentation, indoor mapping, cultural heritage, facility management, energy audits, and numerous VR/AR applications.

The VR exploration system developed in Chapter 6 highlights the direct deployability of the reconstructed models in WebXR environments, enabling lightweight, cross-platform visualization that can reach broad audiences.

7.3 Future Directions

The insights gained through this thesis suggest several compelling research directions.

Handling dynamic and time-varying environments. All current models assume static scenes, yet real-world environments contain moving objects, temporary occlusions, and temporal structure. Future work may incorporate video panoramas, temporal consistency constraints, or layered scene representations that disentangle permanent architecture from dynamic elements.

Integrating additional sensing modalities. While this thesis focused exclusively on RGB panoramas, fusing IMU orientation, ultra-lightweight depth sensors, WiFi localization, or low-cost LiDAR could stabilize registration, refine room boundaries, and reduce ambiguity in complex architectures.

Scaling beyond apartment-level environments. Extending NadirFloorNet and PanoFloor to multi-floor buildings or campus-scale environments introduces new challenges: long-range correspondence, hierarchical graph reasoning, cross-floor connectivity, and large-scale optimization for global consistency.

Richer semantic, functional, and physical modeling. Future reconstructions could integrate object-level semantics, affordances, accessibility features, or energy-performance attributes, enabling applications in robotics, building management, and smart-city digital twins.

Enhanced immersive interaction and advanced neural rendering. Adaptive novel-view synthesis, neural radiance fields optimized from sparse 360° inputs, gaze-contingent rendering, or perceptually optimized path planning may further improve the VR experience initiated in Chapter 6.

Self-supervision and foundation models for panoramic reasoning. Large-scale pretraining, multi-modal foundation models, and self-supervised geometric learning represent promising directions for further reducing annotation requirements and improving generalization across diverse indoor environments.

Publications Related to This Thesis

The main scientific contributions of this thesis are documented in the following peer-reviewed publications:

Publications Directly Derived from This Thesis

- G. Pintore, M. Agus, A. Signoroni, and E. Gobbetti, “DDD++: Exploiting density map consistency for deep depth estimation in indoor environments,” *Graphical Models*, vol. 140, 101281, 2025 [61].
- G. Pintore, U. Shah, M. Agus, and E. Gobbetti, “NadirFloorNet: Reconstructing multi-room floorplans from a small set of registered panoramic images,” in *Proc. CVPR Workshop on Urban Scene Modeling*, 2025 [103].
- G. Pintore, S. Jashari, M. Agus, and E. Gobbetti, “PanoFloor: Reconstruction and immersive exploration of large multi-room scenes from a minimal set of registered panoramic images using denoised density maps,” in *Proc. IEEE ISMAR*, October 2025 [10].

These works constitute the core methodological contributions of the thesis, covering: single-view panoramic depth estimation, sparse multi-room reconstruction, and immersive exploration of reconstructed indoor spaces.

Additional Publications Produced During the PhD

Beyond the contributions directly included in this thesis, several other publications have supported, inspired, or contributed foundational concepts to the research presented here. These works span panoramic depth estimation, immersive visualization, stereoscopic rendering, metaverse applications, and large-scale surveys of XR technologies.

International Journals

- G. Pintore, F. Bettio, M. Agus, and E. Gobbetti, “Deep scene synthesis of Atlanta-world interiors from a single omnidirectional image,” *IEEE Transactions on Visualization and Computer Graphics*, October 2023 [19].
- G. Pintore, E. Almansa, A. Sanchez, G. Vassena, and E. Gobbetti, “Deep Panoramic Depth Prediction and Completion for Indoor Scenes,” *Computational Visual Media*, 2024 [190].
- M. Tukur*, G. Pintore*, E. Gobbetti, J. Schneider, and M. Agus, “SPIDER: A framework for processing, editing and presenting immersive high-resolution spherical indoor scenes,” *Graphical Models*, 2023 [169].
- M. Tukur, S. Jashari, M. Alzubaidi, B. A. Salami, Y. Boraey, S. Yong, D. Saleh, G. Pintore, E. Gobbetti, J. Schneider, N. Fetais, and M. Agus, “Panoramic Imaging in Immersive Extended Reality: A Scoping Review of Technologies, Applications, Perceptual Studies, and User Experience Challenges,” *Frontiers in Virtual Reality*, 2025 [191].

International Conference Proceedings

- G. Pintore, A. Jaspe Villanueva, M. Hadwiger, E. Gobbetti, J. Schneider, and M. Agus, “PanoVerse: automatic generation of stereoscopic environments from single indoor panoramic images for Metaverse applications,” in *Proc. ACM Web3D*, 2023 [118]. Honorable Mention.
- G. Pintore, M. Agus, A. Signoroni, and E. Gobbetti, “DDD: Deep indoor panoramic Depth estimation with Density maps consistency,” in *STAG: Smart Tools and Applications in Graphics*, November 2024 [12].
- U. Shah, S. Jashari, M. Tukur, G. Pintore, E. Gobbetti, J. Schneider, and M. Agus, “VISPI: Virtual Staging Pipeline for Single Indoor Panoramic Images,” in *STAG*, November 2024 [192]. Best Paper Award.
- S. Jashari, M. Tukur, Y. Boraey, U. Shah, M. Alzubaidi, G. Pintore, E. Gobbetti, A. Jaspe-Villanueva, J. Schneider, N. Fetais, and M. Agus, “Evaluating AI-based static stereoscopic rendering of indoor panoramic scenes,” in *STAG*, November 2024 [193].

- M. Tukur, Y. Boraey, S. Jashari, A. Jaspe-Villanueva, U. Shah, M. Alzubaidi, G. Pintore, E. Gobbetti, J. Schneider, and M. Agus, “Virtual Staging Technologies for the Metaverse,” in *Proc. IEEE iMeta*, 2024 [194].
- U. Shah, J. Schneider, G. Pintore, E. Gobbetti, M. Alzubaidi, M. Househ, and M. Agus, “EleViT: exploiting element-wise products for designing efficient and lightweight vision transformers,” in *Proc. T4V — CVPR Workshops*, 2024 [195].
- U. Shah, M. Tukur, M. Alzubaidi, G. Pintore, E. Gobbetti, M. Househ, J. Schneider, and M. Agus, “MultiPanoWise: holistic deep architecture for multi-task dense prediction from a single panoramic image,” in *Proc. OmniCV — CVPR Workshops*, 2024 [196].
- S. Jashari, M. Tukur, S. Al Sulaiti, Y. El Sayed, G. Pintore, E. Gobbetti, J. Schneider, and M. Agus, “AI and WebXR Technologies for Interactive Virtual Tours based on 360 Capture: The Doha Design District Case Study,” in *Proc. ACM Web3D*, September 2025 [197].
- M. Tukur, S. Jashari, F. Bettio, G. Pintore, E. Gobbetti, J. Schneider, and M. Agus, “PanoStyleVR: style-based similarity metrics for Web-based immersive panoramic style transfer,” in *Proc. ACM Web3D*, September 2025 [198].
- G. Pintore, M. Agus, and E. Gobbetti, “Automatic 3D modeling and exploration of indoor structures from panoramic imagery,” in *SIGGRAPH Asia 2024 Courses*, December 2024 [138].

Additional related works, surveys, and collaborations produced during the doctoral programme and relevant to the themes of this thesis are also listed in the complete bibliography.

Bibliography

- [1] G. Pintore, E. Almansa, M. Agus, and E. Gobbetti, “Deep3DLayout: 3D reconstruction of an indoor layout from a spherical panoramic image,” *ACM TOG*, vol. 40, no. 6, pp. 250:1–250:12, 2021.
- [2] E. Delage, Honglak Lee, and A. Y. Ng, “A dynamic Bayesian network model for autonomous 3D reconstruction from a single indoor image,” in *Proc. CVPR*, vol. 2, pp. 2418–2428, 2006.
- [3] V. Hedau, D. Hoiem, and D. Forsyth, “Recovering the spatial layout of cluttered rooms,” in *Proc. ICCV*, pp. 1849–1856, 2009.
- [4] D. C. Lee, M. Hebert, and T. Kanade, “Geometric reasoning for single image structure recovery,” in *Proc. CVPR*, pp. 2136–2143, 2009.
- [5] J. M. Coughlan and A. L. Yuille, “Manhattan World: Compass direction from a single image by Bayesian inference,” in *Proc. ICCV*, vol. 2, pp. 941–947, 1999.
- [6] G. Schindler and F. Dellaert, “Atlanta world: an expectation maximization framework for simultaneous low-level edge grouping and camera calibration in complex man-made environments,” in *Proc. CVPR*, vol. 1, pp. I–I, 2004.
- [7] G. Pintore, F. Ganovelli, R. Pintus, R. Scopigno, and E. Gobbetti, “3D floor plan recovery from overlapping spherical images,” *Computational Visual Media*, vol. 4, no. 4, pp. 367–383, 2018.
- [8] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski, “Reconstructing building interiors from images,” in *Proc. ICCV*, pp. 80–87, 2009.
- [9] G. Pintore, C. Mura, F. Ganovelli, L. Fuentes-Perez, R. Pajarola, and E. Gobbetti, “State-of-the-art in automatic 3D reconstruction of structured indoor environments,” *Comput. Graph. Forum*, vol. 39, no. 2, pp. 667–699, 2020.

- [10] G. Pintore, S. Jashari, M. Agus, and E. Gobbetti, “PanoFloor: reconstruction and immersive exploration of large multi-room scenes from a minimal set of registered panoramic images using denoised density maps,” in *Proc. IEEE ISMAR*, October 2025. To appear.
- [11] G. Pintore, A. Jaspe-Villanueva, M. Hadwiger, J. Schneider, M. Agus, F. Marton, F. Bettio, and E. Gobbetti, “Deep synthesis and exploration of omnidirectional stereoscopic environments from a single surround-view panoramic image,” *Computers & Graphics*, vol. 119, p. 103907, March 2024.
- [12] G. Pintore, M. Agus, A. Signoroni, and E. Gobbetti, “Ddd: Deep indoor panoramic depth estimation with density maps consistency,” in *2024 Eurographics Italian Chapter Conference on Smart Tools and Applications in Graphics, STAG 2024*, pp. stag–20241336, Eurographics Association, 2024.
- [13] ShanghaiTech University and Kujiale.com, “Shanghaitech-Kujiale Indoor 360° (SKI360) dataset,” 2020. [Online; accessed 2025-09-16].
- [14] Matterport, “Matterport3D,” 2017. [Accessed: 2025-09-16].
- [15] G. Pintore, M. Agus, E. Almansa, J. Schneider, and E. Gobbetti, “SliceNet: deep dense depth estimation from a single indoor panorama using a slice-based representation,” in *Proc. CVPR*, pp. 11536–11545, 2021.
- [16] J. Zheng, J. Zhang, J. Li, R. Tang, S. Gao, and Z. Zhou, “Structured3D: A large photo-realistic dataset for structured 3D modeling,” in *Proc. ECCV*, pp. 519–535, 2020.
- [17] Y. Yue, T. Kontogianni, K. Schindler, and F. Engelmann, “Connecting the Dots: Floorplan Reconstruction Using Two-Level Queries,” in *Proc. CVPR*, 2023.
- [18] S. Cruz, W. Hutchcroft, Y. Li, N. Khosravan, I. Boyadzhiev, and S. B. Kang, “Zillow indoor dataset: Annotated floor plans with 360° panoramas and 3D room layouts,” in *Proc. CVPR*, pp. 2133–2143, 2021.
- [19] G. Pintore, F. Bettio, M. Agus, and E. Gobbetti, “Deep scene synthesis of Atlanta-world interiors from a single omnidirectional image,” *IEEE TVCG*, vol. 29, no. 11, pp. 4708–4718, 2023.
- [20] H. Ai and L. Wang, “Elite360D: Towards efficient 360 depth estimation via semantic-and distance-aware bi-projection fusion,” in *Proc. CVPR*, pp. 9926–9935, 2024.

- [21] L. Jin, Y. Xu, J. Zheng, J. Zhang, R. Tang, S. Xu, J. Yu, and S. Gao, “Geometric structure based and regularized depth estimation from 360 indoor imagery,” in *Proc. CVPR*, pp. 889–898, 2020.
- [22] J. Chen, R. Deng, and Y. Furukawa, “Polydiffuse: Polygonal shape reconstruction via guided set diffusion models,” *NeurIPS*, vol. 36, pp. 1863–1888, 2024.
- [23] G. Pintore, R. Pintus, F. Ganovelli, R. Scopigno, and E. Gobbetti, “Recovering 3D existing-conditions of indoor structures from spherical images,” *Computers & Graphics*, vol. 77, pp. 16–29, Dec. 2018.
- [24] C. Godard, O. Mac Aodha, and G. J. Brostow, “Unsupervised monocular depth estimation with left-right consistency,” in *Proc. CVPR*, 2017.
- [25] M. Firman, “RGBD datasets: Past, present and future,” in *Proc. CVPR*, pp. 19–31, 2016.
- [26] T. L. T. da Silveira, P. G. L. Pinto, J. Murrugarra-Llerena, and C. R. Jung, “3d scene geometry estimation from 360° imagery: A survey,” *ACM Comput. Surv.*, vol. 55, no. 4, pp. 68:1–68:39, 2022.
- [27] R. Cabral and Y. Furukawa, “Piecewise planar and compact floorplan reconstruction from images,” in *Proc. CVPR*, pp. 628–635, 2014.
- [28] T. Jokela, J. Ojala, and K. Väänänen, “How people use 360-degree cameras,” in *Proc. International Conference on Mobile and Ubiquitous Multimedia*, pp. 1–10, 2019.
- [29] S. Im, H. Ha, F. Rameau, H.-G. Jeon, G. Choe, and I. S. Kweon, “All-around depth from small motion with a spherical panoramic camera,” in *Proc. ECCV*, pp. 156–172, 2016.
- [30] J. Huang, Z. Chen, D. Ceylan, and H. Jin, “6-DOF VR videos with a single 360-camera,” in *Proc. IEEE VR*, pp. 37–44, 2017.
- [31] M. Eder, M. Shvets, J. Lim, and J.-M. Frahm, “Tangent images for mitigating spherical distortion,” in *Proc. CVPR*, 2020.
- [32] M.-L. Champel, R. Doré, and N. Mollet, “Key factors for a high-quality VR experience,” in *Applications of Digital Image Processing XL*, vol. 10396, pp. 183–194, SPIE, 2017.

- [33] M. Rey-Area, M. Yuan, and C. Richardt, “360MonoDepth: High-resolution 360° monocular depth estimation,” in *Proc. CVPR*, pp. 3762–3772, 2022.
- [34] B. Davidson, M. S. Alvi, and J. F. H. Henriques, “360 camera alignment via segmentation,” in *Proc. ECCV*, pp. 579–595, 2020.
- [35] W. Xian, Z. Li, M. Fisher, J. Eisenmann, E. Shechtman, and N. Snavely, “UprightNet: geometry-aware camera orientation estimation from single images,” in *Proc. ICCV*, pp. 9974–9983, 2019.
- [36] R. Jung, A. S. J. Lee, A. Ashtari, and J. Bazin, “Deep360Up: A deep learning-based approach for automatic VR image upright adjustment,” in *Proc. IEEE VR*, pp. 1–8, 2019.
- [37] M. Berger, A. Tagliasacchi, L. M. Seversky, P. Alliez, G. Guennebaud, J. A. Levine, A. Sharf, and C. T. Silva, “A survey of surface reconstruction from point clouds,” *Computer Graphics Forum*, vol. 36, no. 1, pp. 301–329, 2017.
- [38] C. Zou, J. Su, C. Peng, A. Colburn, Q. Shan, P. Wonka, H. Chu, and D. Hoiem, “Manhattan room layout reconstruction from a single 360 image: A comparative study of state-of-the-art methods,” *International Journal of Computer Vision*, vol. 129, pp. 1410–1431, 2021.
- [39] C. Zou, J.-W. Su, C.-H. Peng, A. Colburn, Q. Shan, P. Wonka, H.-K. Chu, and D. Hoiem, “3D Manhattan room layout reconstruction from a single 360 image,” *ArXiv e-print arXiv:1910.04099*, 2019.
- [40] D. Eigen and R. Fergus, “Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture,” in *Proc. ICCV*, pp. 2650–2658, 2015.
- [41] A. Saxena, M. Sun, and A. Y. Ng, “Make3D: Learning 3D scene structure from a single still image,” *IEEE TPAMI*, vol. 31, no. 5, pp. 824–840, 2009.
- [42] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” in *Proc. NeurIPS*, pp. 2366–2374, 2014.
- [43] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, “Deeper depth prediction with fully convolutional residual networks,” in *Proc. 3DV*, pp. 239–248, 2016.

- [44] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. CVPR*, pp. 770–778, 2016.
- [45] S. Lambert-Lacroix and L. Zwald, “The adaptive BerHu penalty in robust regression,” *Journal of Nonparametric Statistics*, vol. 28, pp. 1–28, 06 2016.
- [46] J. Lee, M. Heo, K. Kim, and C. Kim, “Single-image depth estimation based on fourier domain analysis,” in *Proc. CVPR*, pp. 330–339, 2018.
- [47] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, “Deep ordinal regression network for monocular depth estimation,” in *Proc. CVPR*, June 2018.
- [48] F. Liu, Chunhua Shen, and Guosheng Lin, “Deep convolutional neural fields for depth estimation from a single image,” in *Proc. CVPR*, pp. 5162–5170, 2015.
- [49] Peng Wang, Xiaohui Shen, Zhe Lin, S. Cohen, B. Price, and A. Yuille, “Towards unified depth and semantic prediction from a single image,” in *Proc. CVPR*, pp. 2800–2809, 2015.
- [50] Y. Cao, Z. Wu, and C. Shen, “Estimating depth from monocular images as classification using deep fully convolutional residual networks,” *IEEE TCSVT*, vol. 28, no. 11, pp. 3174–3182, 2018.
- [51] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci, “Structured attention guided convolutional neural fields for monocular depth estimation,” in *Proc. CVPR*, pp. 3917–3925, 2018.
- [52] Y. Zhang, S. Song, P. Tan, and J. Xiao, “PanoContext: A whole-room 3D context model for panoramic scene understanding,” in *Proc. ECCV*, pp. 668–686, 2014.
- [53] N. Zioulis, A. Karakottas, D. Zarpalas, and P. Daras, “OmniDepth: Dense depth estimation for indoors spherical panoramas,” in *Proc. ECCV*, pp. 453–471, 2018.
- [54] H. Cheng, C. Chao, J. Dong, H. Wen, T. Liu, and M. Sun, “Cube padding for weakly-supervised saliency prediction in 360 videos,” in *Proc. CVPR*, pp. 1420–1429, 2018.
- [55] Y.-C. Su and K. Grauman, “Learning spherical convolution for fast features from 360 imagery,” in *Proc. NeurIPS*, pp. 529–539, 2017.

- [56] K. Tateno, N. Navab, and F. Tombari, “Distortion-aware convolutional filters for dense prediction in panoramic images,” in *Proc. ECCV*, pp. 732–750, 2018.
- [57] G. Payen de La Garanderie, A. Atapour Abarghouei, and T. P. Breckon, “Eliminating the blind spot: Adapting 3D object detection and monocular depth estimation to 360 panoramic imagery,” in *Proc. ECCV*, pp. 812–830, 2018.
- [58] Y. Su and K. Grauman, “Kernel transformer networks for compact spherical convolution,” in *Proc. CVPR*, pp. 9434–9443, 2019.
- [59] F.-E. Wang, Y.-H. Yeh, M. Sun, W.-C. Chiu, and Y.-H. Tsai, “BiFuse: Monocular 360 depth estimation via bi-projection fusion,” in *Proc. CVPR*, pp. 462–471, 2020.
- [60] C. Sun, M. Sun, and H.-T. Chen, “HoHoNet: 360° indoor holistic understanding with latent horizontal features,” in *Proc. CVPR*, pp. 2573–2582, 2021.
- [61] G. Pintore, M. Agus, A. Signoroni, and E. Gobbetti, “DDD++: Exploiting density map consistency for deep depth estimation in indoor environments,” *Graphical Models*, vol. 140, p. 101281, August 2025.
- [62] Z. Shen, C. Lin, K. Liao, L. Nie, Z. Zheng, and Y. Zhao, “PanoFormer: Panorama transformer for indoor 360 depth estimation,” in *Proc. ECCV*, pp. 195–211, Springer, 2022.
- [63] Y. Li, Y. Guo, Z. Yan, X. Huang, Y. Duan, and L. Ren, “Omnifusion: 360 monocular depth estimation via geometry-aware fusion,” in *Proc. CVPR*, pp. 2801–2810, 2022.
- [64] Z. Shen, Z. Zheng, C. Lin, L. Nie, K. Liao, S. Zheng, and Y. Zhao, “Disentangling orthogonal planes for indoor panoramic room layout estimation with cross-scale distortion awareness,” in *Proc. CVPR*, pp. 17337–17345, 2023.
- [65] H. Ai, Z. Cao, Y.-P. Cao, Y. Shan, and L. Wang, “HRDFuse: Monocular 360° depth estimation by collaboratively learning holistic-with-regional depth distributions,” in *Proc. CVPR*, pp. 13273–13282, 2023.
- [66] Q. Yan, Q. Wang, K. Zhao, J. Chen, B. Li, X. Chu, and F. Deng, “Spherefusion: Efficient panorama depth estimation via gated fusion,” in *Proc. 3DV*, pp. 855–865, 2025.

- [67] F.-E. Wang, H.-N. Hu, H.-T. Cheng, J.-T. Lin, S.-T. Yang, M.-L. Shih, H.-K. Chu, and M. Sun, “Self-supervised learning of depth and camera motion from 360 videos,” in *Proc. ACCV*, pp. 53–68, 2018.
- [68] N. Zioulis, A. Karakottas, D. Zarpalas, F. Alvarez, and P. Daras, “Spherical view synthesis for self-supervised 360° depth estimation,” in *Proc. 3DV*, September 2019.
- [69] F.-E. Wang, Y.-H. Yeh, Y.-H. Tsai, W.-C. Chiu, and M. Sun, “Bifuse++: Self-supervised and efficient bi-projection fusion for 360 depth estimation,” *IEEE TPAMI*, vol. 45, no. 5, pp. 5448–5460, 2022.
- [70] N.-H. A. Wang and Y.-L. Liu, “Depth anywhere: Enhancing 360 monocular depth estimation via perspective distillation and unlabeled data augmentation,” *Proc. NeurIPS*, vol. 37, pp. 127739–127764, 2024.
- [71] Z. Cao, J. Zhu, W. Zhang, H. Ai, H. Bai, H. Zhao, and L. Wang, “PanDA: Towards panoramic depth anything with unlabeled panoramas and mobius spatial augmentation,” in *Proc. CVPR*, pp. 982–992, 2025.
- [72] J. Xu, J. Zheng, Y. Xu, R. Tang, and S. Gao, “Layout-guided novel view synthesis from a single indoor panorama,” in *Proc. CVPR*, pp. 16438–16447, 2021.
- [73] V. Gkitsas, V. Sterzentsenko, N. Zioulis, G. Albanis, and D. Zarpalas, “PanoDR: Spherical panorama diminished reality for indoor scenes,” in *Proc. CVPR Workshops*, pp. 3716–3726, 2021.
- [74] G. Pintore, M. Agus, E. Almansa, and E. Gobbetti, “Instant automatic emptying of panoramic indoor scenes,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, pp. 3629–3639, November 2022. Proc. ISMAR.
- [75] S. Ikehata, H. Yang, and Y. Furukawa, “Structured indoor modeling,” in *Proc. ICCV*, pp. 1323–1331, 2015.
- [76] D. Hoiem, A. A. Efros, and M. Hebert, “Recovering surface layout from an image,” *International Journal of Computer Vision*, vol. 75, pp. 151–172, Oct 2007.
- [77] J. Xu, B. Stenger, T. Kerola, and T. Tung, “Pano2CAD: Room layout from a single panorama image,” in *Proc. WACV*, pp. 354–362, 2017.

- [78] H. Yang and H. Zhang, “Efficient 3D room shape recovery from a single panorama,” in *Proc. CVPR*, pp. 5422–5430, 2016.
- [79] C. Zou, A. Colburn, Q. Shan, and D. Hoiem, “LayoutNet: Reconstructing the 3D room layout from a single RGB image,” in *Proc. CVPR*, pp. 2051–2059, 2018.
- [80] C. Sun, C.-W. Hsiao, M. Sun, and H.-T. Chen, “HorizonNet: Learning room layout with 1D representation and pano stretch data augmentation,” in *Proc. CVPR*, pp. 1047–1056, 2019.
- [81] S.-T. Yang, F.-E. Wang, C.-H. Peng, P. Wonka, M. Sun, and H.-K. Chu, “DuLa-Net: A dual-projection network for estimating room layouts from a single RGB panorama,” in *Proc. CVPR*, 2019.
- [82] F.-E. Wang, Y.-H. Yeh, M. Sun, W.-C. Chiu, and Y.-H. Tsai, “LED2-Net: Monocular 360 layout estimation via differentiable depth rendering,” in *Proc. CVPR*, pp. 12956–12965, 2021.
- [83] W. Zeng, S. Karaoglu, and T. Gevers, “Joint 3D layout and depth prediction from a single indoor panorama image,” in *Proc. ECCV*, pp. 666–682, 2020.
- [84] G. Pintore, M. Agus, and E. Gobbetti, “AtlantaNet: Inferring the 3D indoor layout from a single 360 image beyond the Manhattan World assumption,” in *Proc. ECCV*, pp. 432–448, 2020.
- [85] P. V. Tran, “Sslayout360: Semi-supervised indoor layout estimation from 360deg panorama,” in *Proc. CVPR*, pp. 15353–15362, 2021.
- [86] Z. Jiang, Z. Xiang, J. Xu, and M. Zhao, “LGT-Net: Indoor panoramic room layout estimation with geometry-aware transformer network,” in *Proc. CVPR*, pp. 1654–1663, 2022.
- [87] Y. Zhao, C. Wen, Z. Xue, and Y. Gao, “3D room layout estimation from a cubemap of panorama image via deep Manhattan hough transform,” in *Proc. ECCV*, pp. 637–654, Springer, 2022.
- [88] H. Yu, L. He, B. Jian, W. Feng, and S. Liu, “Panelnet: Understanding 360 indoor environment via panel representation,” in *Proc. CVPR*, pp. 878–887, 2023.

- [89] Y.-J. Tsai, J.-C. Jhang, J. Zheng, W. Wang, A. Y. Chen, M. Sun, C.-H. Kuo, and M.-H. Yang, “No more ambiguity in 360deg room layout via bi-layout estimation,” in *Proc. CVPR*, pp. 28056–28065, 2024.
- [90] C. Sun, W.-E. Tai, Y.-L. Shih, K.-W. Chen, Y.-J. Syu, Y.-C. F. Wang, and H.-T. Chen, “Seg2Reg: Differentiable 2D segmentation to 1D regression rendering for 360° room layout reconstruction,” in *Proc. CVPR*, pp. 10435–10445, 2024.
- [91] J. Lee, H. Park, B.-U. Lee, and K. Joo, “HUSH: Holistic panoramic 3D scene understanding using spherical harmonics,” in *Proc. CVPR*, pp. 16599–16608, June 2025.
- [92] M. A. Shabani, W. Song, M. Odamaki, H. Fujiki, and Y. Furukawa, “Extreme structure from motion for indoor panoramas without visual overlaps,” in *Proc. ICCV*, pp. 5683–5691, 2021.
- [93] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor segmentation and support inference from rgb-d images,” in *Proc. ECCV*, pp. 746–760, Springer, 2012.
- [94] A. Monszpart, N. Mellado, G. J. Brostow, and N. J. Mitra, “Rapter: rebuilding man-made scenes with regular arrangements of planes,” *ACM TOG*, vol. 34, no. 4, pp. 103–1, 2015.
- [95] G. Pu, Y. Zhao, and Z. Lian, “Pano2Room: Novel view synthesis from a single indoor panorama,” in *Proc. SIGGRAPH Asia Conference Papers*, 2024.
- [96] J. Lambert, Y. Li, I. Boyadzhiev, L. Wixson, M. Narayana, W. Hutchcroft, J. Hays, F. Dellaert, and S. B. Kang, “Salve: Semantic alignment verification for floorplan reconstruction from sparse panoramas,” in *Proc. ECCV*, pp. 647–664, Springer, 2022.
- [97] Y. Li, I. Boyadzhiev, Z. Liu, L. Shapiro, and A. Colburn, “Badgr: Bundle adjustment diffusion conditioned by gradients for wide-baseline floor plan reconstruction,” in *Proc. CVPR*, pp. 16785–16795, June 2025.
- [98] J. Chen, C. Liu, J. Wu, and Y. Furukawa, “Floor-SP: Inverse CAD for floorplans by sequential room-wise shortest path,” in *Proc. CVPR*, pp. 2661–2670, 2019.
- [99] S. Stekovic, M. Rad, F. Fraundorfer, and V. Lepetit, “Montefloor: Extending MCTS for reconstructing accurate large-scale floor plans,” in *Proc. CVPR*, pp. 16034–16043, 2021.

- [100] C. Liu, J. Wu, and Y. Furukawa, “FloorNet: A unified framework for floorplan reconstruction from 3D scans,” in *Proc. ECCV*, pp. 201–217, 2018.
- [101] J. Chen, Y. Qian, and Y. Furukawa, “HEAT: Holistic edge attention transformer for structured reconstruction,” in *Proc. CVPR*, pp. 3866–3875, 2022.
- [102] J.-W. Su, K.-Y. Tung, C.-H. Peng, P. Wonka, and H.-K. Chu, “SLIBO-Net: Floorplan reconstruction via slicing box representation with local geometry regularization,” in *Proc. NeurIPS*, pp. 1–12, 2023.
- [103] G. Pintore, U. Shah, M. Agus, and E. Gobbetti, “Nadirfloornet: reconstructing multi-room floorplans from a small set of registered panoramic images,” in *Proc. CVPR*, pp. 1985–1994, 2025.
- [104] J. Waidhofer, R. Gadgil, A. Dickson, S. Zollmann, and J. Ventura, “PanoSynthVR: Toward light-weight 360-degree view synthesis from a single panoramic input,” in *Proc. ISMAR*, pp. 584–592, 2022.
- [105] Q. Zhao, L. Wan, W. Feng, J. Zhang, and T.-T. Wong, “Cube2Video: Navigate between cubic panoramas in real-time,” *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1745–1754, 2013.
- [106] Z. Wei, J. Zhang, X. Shen, Z. Lin, R. Mech, M. Hoai, and D. Samaras, “Good view hunting: Learning photo composition from dense view pairs,” in *Proc. CVPR*, pp. 5437–5446, 2018.
- [107] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, “Free-form image inpainting with gated convolution,” in *Proc. ICCV*, pp. 4471–4480, 2019.
- [108] Y. Zeng, J. Fu, H. Chao, and B. Guo, “Learning pyramid-context encoder network for high-quality image inpainting,” in *Proc. CVPR*, pp. 1486–1494, 2019.
- [109] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely, “Stereo magnification: Learning view synthesis using multiplane images,” *ACM TOG*, vol. 37, no. 4, pp. 68:1–68:12, 2018.
- [110] B. Attal, S. Ling, A. Gokaslan, C. Richardt, and J. Tompkin, “MatryODShka: Real-time 6DoF video view synthesis using multi-sphere images,” in *Proc. ECCV*, pp. 441–459, 2020.

- [111] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, “NeRF: representing scenes as neural radiance fields for view synthesis,” *Commun. ACM*, vol. 65, pp. 99–106, dec 2021.
- [112] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, “3D gaussian splatting for real-time radiance field rendering,” *ACM TOG*, vol. 42, no. 4, pp. 139:1–139:14, 2023.
- [113] S. Kulkarni, P. Yin, and S. Scherer, “360FusionNeRF: Panoramic neural radiance fields with joint guidance,” in *Proc. IROS*, pp. 7202–7209, 2023.
- [114] G. Wang, P. Wang, Z. Chen, W. Wang, C. C. Loy, and Z. Liu, “PERF: Panoramic Neural Radiance Field from a single panorama,” *IEEE TPAMI*, pp. 1–15, 2024.
- [115] M. Rey-Area and C. Richardt, “360° 3D photos from a single 360° input image,” *IEEE TVCG*, vol. 31, no. 5, pp. 2426–2434, 2025.
- [116] S. Peleg and M. Ben-Ezra, “Stereo panorama with a single camera,” in *Proc. CVPR*, pp. 395–401, 1999.
- [117] P. Bourke, “Capturing omni-directional stereoscopic spherical projections with a single camera,” in *Proc. IEEE VSMM*, pp. 179–183, 2010.
- [118] G. Pintore, A. Jaspe Villanueva, M. Hadwiget, E. Gobbetti, J. Schneider, and M. Agus, “PanoVerse: automatic generation of stereoscopic environments from single indoor panoramic images for metaverse applications,” in *Proc. ACM Web3D*, pp. 2:1–2:10, 2023.
- [119] Z. Chen, Y.-P. Cao, Y.-C. Guo, C. Wang, Y. Shan, and S.-H. Zhang, “PanoGRF: generalizable spherical radiance fields for wide-baseline panoramas,” in *Proc. NeurIPS*, 2023.
- [120] M. Di Benedetto, F. Ganovelli, M. Balsa Rodriguez, A. Jaspe Villanueva, R. Scopigno, and E. Gobbetti, “ExploreMaps: Efficient construction and ubiquitous exploration of panoramic view graphs of complex 3D environments,” *Comput. Graph. Forum*, vol. 33, no. 2, pp. 459–468, 2014.
- [121] M. Feixas, M. Sbert, and F. González, “A unified information-theoretic framework for viewpoint selection and mesh saliency,” *ACM Trans. Appl. Percept.*, vol. 6, no. 1, pp. 1:1–1:23, 2009.

- [122] K. Song and L. Zhang, “Novel view synthesis with wide-baseline stereo pairs based on local–global information,” *Computers & Graphics*, vol. 126, p. 104139, 2025.
- [123] H. AlZayer, H. Lin, and K. Bala, “Autophoto: Aesthetic photo capture using reinforcement learning,” in *Proc. IROS*, pp. 944–951, 2021.
- [124] T. Uchida, Y. Kanamori, and Y. Endo, “3D view optimization for improving image aesthetics,” in *Proc. ICASSP*, pp. 1–5, 2025.
- [125] C. Zou, J.-W. Su, C.-H. Peng, A. Colburn, Q. Shan, P. Wonka, H.-K. Chu, and D. Hoiem, “MatterportLayout,” 2017. [Accessed: 2025-09-16].
- [126] Stanford University, “BuildingParser Dataset,” 2017. [Accessed: 2025-09-16].
- [127] F. Xia, A. R. Zamir, Z. He, A. Sax, J. Malik, and S. Savarese, “GibsonV2,” 2018. [Online; accessed 2025-09-17].
- [128] G. Albanis, N. Zioulis, P. Drakoulis, V. Gkitsas, V. Sterzentsenko, F. Alvarez, D. Zarpalas, and P. Daras, “Pano3D,” 2021. [Online; accessed 2025-09-17].
- [129] G. Albanis, N. Zioulis, P. Drakoulis, V. Gkitsas, V. Sterzentsenko, F. Alvarez, D. Zarpalas, and P. Daras, “Pano3D: A holistic benchmark and a solid baseline for 360° depth estimation,” in *Proc. CVPR Workshops*, pp. 3727–3737, 2021.
- [130] F. Xia, A. R. Zamir, Z. He, A. Sax, J. Malik, and S. Savarese, “Gibson Env: real-world perception for embodied agents,” in *Proc. CVPR*, 2018.
- [131] M. Rey-Area, M. Yuan, and C. Richardt, “360MonoDepth Matterport3D,” 2022. [Online; accessed 2025-09-17].
- [132] M. Rey-Area, M. Yuan, and C. Richardt, “360MonoDepth Replica,” 2022. [Online; accessed 2025-09-17].
- [133] Y. Zhang, S. Song, P. Tan, and J. Xiao, “PanoContext dataset,” 2014. [Accessed: 2025-09-16].
- [134] S. Cruz, W. Hutchcroft, Y. Li, N. Khosravan, I. Boyadzhiev, and S. B. Kang, “Zillow Indoor Dataset (ZInD),” 2021. [Accessed: 2025-09-16].
- [135] J. Zheng, J. Zhang, J. Li, R. Tang, S. Gao, and Z. Zhou, “Indoor3Dmapping dataset,” 2020. [Accessed: 2025-09-16].

- [136] J. Xu, J. Zheng, Y. Xu, R. Tang, and S. Gao, “PNVS Dataset,” 2021. [Accessed: 2025-09-16].
- [137] G. Pintore, F. Ganovelli, E. Gobbetti, and R. Scopigno, “Mobile mapping and visualization of indoor structures to simplify scene understanding and location awareness,” in *Proc. ECCV Workshops*, pp. 130–145, 2016.
- [138] E. Gobbetti, G. Pintore, and M. Agus, “Automatic 3D modeling and exploration of indoor structures from panoramic imagery,” in *SIGGRAPH Asia Courses*, pp. 1:1–1:9, 2024.
- [139] Y. Yang, S. Jin, R. Liu, , and J. Yu, “Automatic 3D indoor scene modeling from single panorama,” in *Proc. CVPR*, pp. 3926–3934, 2018.
- [140] O. Wiles, G. Gkioxari, R. Szeliski, and J. Johnson, “Synsin: End-to-end view synthesis from a single image,” in *Proc. CVPR*, pp. 7467–7477, 2020.
- [141] T. L. da Silveira, P. G. Pinto, J. Murrugarra-Llerena, and C. R. Jung, “3D scene geometry estimation from 360° imagery: A survey,” *ACM Computing Surveys*, vol. 55, no. 4, pp. 1–39, 2022.
- [142] Z. Cao, H. Ai, A. V. Vasilakos, and L. Wang, “360 high-resolution depth estimation via uncertainty-aware structural knowledge transfer,” *IEEE Trans. AI*, 2024.
- [143] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Proc. NeurIPS*, vol. 30, 2017.
- [144] I. Yun, C. Shin, H. Lee, H.-J. Lee, and C.-E. Rhee, “EGformer: Equirectangular geometry-biased transformer for 360° depth estimation,” in *Proc. ICCV*, pp. 6078–6089, 2023.
- [145] H. Jiang, Z. Sheng, S. Zhu, Z. Dong, and R. Huang, “Unifuse: Unidirectional fusion for 360 panorama depth estimation,” *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1519–1526, 2021.
- [146] W. Hutchcroft, Y. Li, I. Boyadzhiev, Z. Wan, H. Wang, and S. B. Kang, “CoVisPose: Co-visibility pose transformer for wide-baseline relative pose estimation in 360° indoor panoramas,” in *Proc. ECCV*, pp. 615–633, 2022.

- [147] Z. Jiang, Z. Xiang, J. Xu, and M. Zhao, “LGT-net: Indoor panoramic room layout estimation with geometry-aware transformer network,” in *Proc. CVPR*, pp. 1654–1663, 2022.
- [148] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, “Deformable DETR: Deformable transformers for end-to-end object detection,” *arXiv preprint arXiv:2010.04159*, 2020.
- [149] W. Chang, Y. Zhang, and Z. Xiong, “Depth estimation from indoor panoramas with neural scene representation,” in *Proc. CVPR*, pp. 899–908, 2023.
- [150] H. Wang, W. Hutchcroft, Y. Li, Z. Wan, I. Boyadzhiev, Y. Tian, and S. B. Kang, “PSMNet: Position-aware stereo merging network for room layout estimation,” in *Proc. CVPR*, pp. 8616–8625, 2022.
- [151] B. Solarte, C.-H. Wu, Y.-C. Liu, Y.-H. Tsai, and M. Sun, “360-MLC: Multi-view layout consistency for self-training and hyper-parameter tuning,” *NeurIPS*, vol. 35, pp. 6133–6146, 2022.
- [152] B. Solarte, Y.-C. Liu, C.-H. Wu, Y.-H. Tsai, and M. Sun, “360-dfpe: Leveraging monocular 360-layouts for direct floor plan estimation,” *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6503–6510, 2022.
- [153] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in *Proc. ICCV*, pp. 2961–2969, 2017.
- [154] N. Nauata and Y. Furukawa, “Vectorizing world buildings: Planar graph reconstruction by primitive detection and relationship inference,” in *Proc. ECCV*, pp. 711–726, Springer, 2020.
- [155] M. A. Shabani, S. Hosseini, and Y. Furukawa, “HouseDiffusion: Vector floorplan generation via a diffusion model with discrete and continuous denoising,” *arXiv preprint arXiv:2211.13287*, 2022.
- [156] A. Gueze, M. Ospici, D. Rohmer, and M.-P. Cani, “Floor plan reconstruction from sparse views: Combining graph neural network with constrained diffusion,” in *Proc. CVPR*, pp. 1583–1592, 2023.
- [157] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.

- [158] M. Z. Sulaiman, M. N. A. Aziz, M. H. A. Bakar, N. A. Halili, and M. A. Azuddin, “Matterport: virtual tour as a new marketing approach in real estate business during pandemic COVID-19,” in *Proc. IMDES*, pp. 221–226, 2020.
- [159] P. Boguslawski, S. Zlatanova, D. Gotlib, M. Wyszomirski, M. Gnat, and P. Grzempowski, “3D building interior modelling for navigation in emergency response applications,” *nt. J. Appl. Earth Obs. Geoinf*, vol. 114, p. 103066, 2022.
- [160] J. Zhang, X. Xia, R. Liu, and N. Li, “Enhancing human indoor cognitive map development and wayfinding performance with immersive augmented reality-based navigation systems,” *Advanced Engineering Informatics*, vol. 50, p. 101432, 2021.
- [161] M. Xu, C. Li, S. Zhang, and P. Le Callet, “State-of-the-art in 360° video/image processing: Perception, assessment and compression,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 1, pp. 5–26, 2020.
- [162] K. Matzen, M. F. Cohen, B. Evans, J. Kopf, and R. Szeliski, “Low-cost 360 stereo photography and video capture,” *ACM TOG*, vol. 36, no. 4, pp. 148:1–148:12, 2017.
- [163] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proc. CVPR*, pp. 10684–10695, 2022.
- [164] E. Olson, “AprilTag: A robust and flexible visual fiducial system,” in *Proc. ICRA*, pp. 3400–3407, 2011.
- [165] Z. Yang, J. Z. Pan, L. Luo, X. Zhou, K. Grauman, and Q. Huang, “Extreme relative pose estimation for RGB-D scans via scene completion,” in *Proc. CVPR*, pp. 4531–4540, 2019.
- [166] Z. Yang, S. Yan, and Q. Huang, “Extreme relative pose network under hybrid representations,” in *Proc. CVPR*, pp. 2455–2464, 2020.
- [167] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, “Deformable DETR: Deformable transformers for end-to-end object detection,” in *Proc. ICLR*, 2021.
- [168] H. Jia, H. Yi, H. Fujiki, H. Zhang, W. Wang, and M. Odamaki, “3D room layout recovery generalizing across Manhattan and Non-Manhattan worlds,” in *Proc. CVPR Workshops*, pp. 5188–5197, 2022.

- [169] M. Tukur, G. Pintore, E. Gobbetti, J. Schneider, and M. Agus, “SPIDER: A framework for processing, editing and presenting immersive high-resolution spherical indoor scenes,” *Graphical Models*, vol. 128, pp. 101182:1–101182:11, 2023.
- [170] B. Luo, F. Xu, C. Richardt, and J.-H. Yong, “Parallax360: Stereoscopic 360° scene representation for head-motion parallax,” *IEEE TVCG*, vol. 24, no. 4, pp. 1545–1553, 2018.
- [171] P. Hedman and J. Kopf, “Instant 3D photography,” *ACM TOG*, vol. 37, no. 4, pp. 101:1–101:12, 2018.
- [172] R. Tucker and N. Snavely, “Single-view view synthesis with multiplane images,” in *Proc. CVPR*, pp. 548–557, 2020.
- [173] Q. Li and N. Khademi Kalantari, “Synthesizing light field from a single image with variable MPI and two network fusion,” *ACM TOG*, vol. 39, no. 6, pp. 229:1–229:10, 2020.
- [174] T. Marrinan and M. E. Papka, “Real-time omnidirectional stereo rendering: generating 360° surround-view panoramic images for comfortable immersive viewing,” *IEEE TVCG*, vol. 27, no. 5, pp. 2587–2596, 2021.
- [175] C. Zhang, Z. Cui, C. Chen, S. Liu, B. Zeng, H. Bao, and Y. Zhang, “Deep-PanoContext: Panoramic 3D scene understanding with holistic scene context graph and relation-based optimization,” in *Proc. ICCV*, pp. 12632–12641, 2021.
- [176] A. Jain, M. Tancik, and P. Abbeel, “Putting NeRF on a diet: Semantically consistent few-shot view synthesis,” in *Proc. ICCV*, pp. 5885–5894, 2021.
- [177] D. Li, Y. Zhang, C. Häne, D. Tang, A. Varshney, and R. Du, “OmniSyn: Synthesizing 360 videos with wide-baseline panoramas,” in *Proc. VRW*, pp. 670–671, 2022.
- [178] S. Boorboor, Y. Kim, P. Hu, J. M. Moses, B. A. Colle, and A. E. Kaufman, “Submerse: Visualizing storm surge flooding simulations in immersive display ecologies,” *IEEE TVCG*, vol. 30, no. 9, pp. 6365–6377, 2023.
- [179] D. Xie, P. Hu, X. Sun, S. Pirk, J. Zhang, R. Mech, and A. E. Kaufman, “GAIT: Generating aesthetic indoor tours with deep reinforcement learning,” in *Proc. ICCV*, pp. 7409–7419, 2023.

- [180] A. Razavi, A. Van den Oord, and O. Vinyals, “Generating diverse high-fidelity images with vq-vae-2,” *Advances in neural information processing systems*, vol. 32, 2019.
- [181] F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” in *Proc. ICLR* (Y. Bengio and Y. LeCun, eds.), 2016.
- [182] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” in *Proc. 3DV*, pp. 565–571, 2016.
- [183] P. Wagner, J. Kotzian, J. Kordas, and V. Michna, “Path planning and tracking for robots based on cubic hermite splines in real-time,” in *Proc. IEEE ETFA*, pp. 1–8, 2010.
- [184] B. W. Wah, Y. Chen, and T. Wang, “Simulated annealing with asymptotic convergence for nonlinear constrained optimization,” *Journal of Global Optimization*, vol. 39, pp. 1–37, 2007.
- [185] Y. Xiang, S. Gubian, B. Suomela, and J. Hoeng, “Generalized simulated annealing for global optimization: The gensa package,” *The R Journal*, vol. 5, no. 1, p. 13, 2013.
- [186] X. Wang, L. Xie, C. Dong, and Y. Shan, “Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data,” in *Proc. ICCVW*, 2021.
- [187] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in pytorch,” in *Proc. NIPS Workshop on Autodiff*, 2017.
- [188] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *ArXiv e-print arXiv:1412.6980*, 2014.
- [189] Z. Wu, P. Zhou, K. Kawaguchi, and H. Zhang, “Fast diffusion model,” *arXiv preprint arXiv:2306.06991*, 2023.
- [190] G. Pintore, E. Almansa, A. Sanchez, G. Vassena, and E. Gobbetti, “Deep panoramic depth prediction and completion for indoor scenes,” *Computational Visual Media*, vol. 10, pp. 1–20, February 2024.
- [191] M. Tukur, S. Jashari, M. Alzubaidi, B. Abiodun Salami, Y. Boraey, S. Yong, D. Saleh, G. Pintore, E. Gobbetti, J. Schneider, N. Fetais, and M. Agus,

- “Panoramic imaging in immersive extended reality: A scoping review of technologies, applications, perceptual studies, and user experience challenges,” *Frontiers in Virtual Reality*, vol. 6, 2025.
- [192] U. Shah, S. Jashari, M. Tukur, G. Pintore, E. Gobbetti, J. Schneider, and M. Agus, “VISPI: Virtual staging pipeline for single indoor panoramic images,” in *STAG: Smart Tools and Applications in Graphics*, November 2024.
 - [193] S. Jashari, M. Tukur, Y. Boraey, U. Shah, M. Alzubaidi, G. Pintore, E. Gobbetti, A. Jaspe-Villanueva, J. Schneider, N. Fetais, and M. Agus, “Evaluating AI-based static stereoscopic rendering of indoor panoramic scenes,” in *STAG: Smart Tools and Applications in Graphics*, November 2024.
 - [194] M. Tukur, Y. Boraey, S. Jashari, A. Jaspe-Villanueva, U. Shah, M. Alzubaidi, G. Pintore, E. Gobbetti, J. Schneider, and M. Agus, “Virtual staging technologies for the metaverse,” in *Proc. IEEE iMeta*, pp. 206–213, 2024.
 - [195] U. Shah, J. Schneider, G. Pintore, E. Gobbetti, M. Alzubaidi, M. Househ, and M. Agus, “EleViT: exploiting element-wise products for designing efficient and lightweight vision transformers,” in *Proc. T4V - IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024.
 - [196] U. Shah, M. Tukur, M. Alzubaidi, G. Pintore, E. Gobbetti, M. Househ, J. Schneider, and M. Agus, “MultiPanoWise: holistic deep architecture for multi-task dense prediction from a single panoramic image,” in *Proc. OmniCV - IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1311–1321, 2024.
 - [197] S. Jashari, M. Tukur, S. Al Sulaiti, Y. El Sayed, G. Pintore, E. Gobbetti, J. Schneider, and M. Agus, “AI and WebXR technologies for interactive virtual tours based on 360 capture: The Doha Design District case study,” in *Proc. ACM Web3D*, September 2025.
 - [198] M. Tukur, S. Jashari, F. Bettio, G. Pintore, E. Gobbetti, J. Schneider, and M. Agus, “PanoStyleVR: style-based similarity metrics for web-based immersive panoramic style transfer,” in *Proc. ACM Web3D*, September 2025.