Automatic 3D modeling and exploration of indoor structures from panoramic imagery

Giovanni Pintore
giovanni.pintore@crs4.it
CRS4
Cagliari, Italy
National Research Center in HPC, Big Data and Quantum Computing
Italy

Marco Agus
magus@hbku.edu.qa
College of Science and Engineering
Hamad Bin Khalifa University
Doha, Qatar

Enrico Gobbetti
enrico.gobbetti@crs4.it
CRS4
Cagliari, Italy
National Research Center in HPC, Big Data and Quantum Computing
Italy

ABSTRACT

Surround-view panoramic imaging delivers extensive spatial coverage and is widely supported by professional and commodity capture devices. Research on inferring and exploring 3D indoor models from 360° images has recently flourished, resulting in highly effective solutions. Nevertheless, challenges persist due to the complexity and variability of indoor environments and issues with noisy and incomplete data. This course provides an up-to-date integrative view of the field. After introducing a characterization of input sources, we define the structure of output models, the priors exploited to bridge the gap between imperfect input and desired output, and the main characteristics of geometry reasoning and data-driven approaches. We then identify and discuss the main sub-problems in indoor reconstruction from panoramas and review and analyze state-of-the-art solutions for indoor capture, room modeling, integrated model computation, visual representation generation, and immersive exploration. Relevant examples of implemented pipelines are described, focusing on deep-learning solutions. We finally point out relevant research issues and analyze research trends.

CCS CONCEPTS

• **Computing methodologies** → **Computer graphics**; *Shape modeling*; **Computer vision**; *Computer vision problems*; *Shape inference*; *Reconstruction*; • **Applied computing** → *Computer-aided design.*

KEYWORDS

panoramic images, surround-view images, omnidirectional images, indoor reconstruction, structured reconstruction, exploration, extended reality

ACM Reference Format:
Giovanni Pintore, Marco Agus, and Enrico Gobbetti. 2024. Automatic 3D modeling and exploration of indoor structures from panoramic imagery. In *SIGGRAPH Asia 2024 Courses (SA Courses '24), December 03-06, 2024.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3680532.3689580

SA Courses '24, December 03-06, 2024, Tokyo, Japan

ACM ISBN 979-8-4007-1135-0/24/12
https://doi.org/10.1145/3680532.3689580

1 INTRODUCTION

This paper accompanies the course given at SIGGRAPH Asia 2024 on the topic of automatic 3D modeling and exploration of indoor structures from panoramic imagery. In addition to describing the course format, prerequisite, and content, we provide a brief discussion of the covered state-of-the-art with relevant bibliographic references.

2 FORMAT AND PREREQUISITES

Format. Half-Day Course (3 hours and 45 minutes, including one 15-minute break).
Presenters. The course is organized by Enrico Gobbetti (co-author of the material) and delivered in presence by Giovanni Pintore and Marco Agus (speakers and material co-authors). See Appendix A for bio sketches.
Necessary background. The course is at the intermediate level. Basic computer vision and deep learning backgrounds are prerequisites.
Intended audience. The target audience includes graduate students, researchers in 3D modeling and scene understanding, and practitioners in the relevant application fields. Researchers will find a structured overview of the field, which organizes the various problems and existing solutions, classifies the existing literature, and indicates challenging open problems. Domain experts will, in turn, find a presentation of the areas where automated methods are already mature enough to be ported into practice, as well as an analysis of the kind of indoor environments that still pose major challenges.

3 COURSE DESCRIPTION

The automatic 3D reconstruction, modeling, and exploration of indoor scenes has become a prominent and increasingly well-defined research topic in recent years [Pintore et al. 2020b]. Current efforts are particularly focused on developing specialized techniques for common, highly structured multi-room environments, such as residential, office, or public buildings, which have a substantial impact on architecture, civil engineering, digital mapping, urban geography, real estate, and more [Ikehata et al. 2015]. In this context, the emphasis has shifted from creating dense 3D models that assemble every measured geometric and visual detail to abstracting high-level structured models that are optimized for specific application-dependent characteristics and incorporate a degree of semantic information [Hu et al. 2020; Ikehata et al. 2015; Pintore

et al. 2020b]. Central to this research are the tasks of identifying architectural elements (such as rooms, walls, windows, and doors) and indoor objects, and integrating them into a coherent structured 3D representation and visual model.

Many options exist for performing capture, ranging from very low-cost commodity solutions to professional devices and systems. Among the many possible options, 360° imagery is attracting a lot of interest [Zou et al. 2021], since it provides the widest cost-effective coverage with just a few shots [Yang et al. 2020].

Furthermore, omnidirectional imagery is increasingly recognized as a critical element for creating immersive content from real-world scenes. A single-shot 360° image, which captures the entire surrounding environment, inherently supports a more dynamic form of exploration compared to traditional 2D imagery. When viewed through a Head-Mounted Display (HMD), it encourages viewers to explore the content by making natural head movements, thereby facilitating an intuitive virtual reality (VR) interface [Xu et al. 2020]. For this reason, 360° image viewing has emerged as a primary mode for exploring real-world scenes in VR [Matzen et al. 2017] and is extensively used in applications such as indoor navigation. However, to provide essential depth cues—such as stereopsis or motion parallax—images alone are insufficient, and scene modeling or view synthesis is required.

Even with the extensive context provided by 360° images, recovering accurate indoor models from visual input remains a highly challenging task due to the intrinsic characteristics of indoor environments, such as confined spaces, windows, textureless surfaces, non-cooperative materials, and abundant clutter. In response to these challenges, various indoor reconstruction techniques that leverage wide contextual information and specific geometric and holistic priors have been proposed in recent years [Pintore et al. 2020b]. Notably, the growing availability of large-scale synthetic and reality-based data collections has facilitated the rise of data-driven and deep-learning approaches capable of relaxing the priors imposed by pure geometric reasoning by learning hidden relations from examples.

In this course, we provide an up-to-date integrative view of the field. After introducing a characterization of input sources, we define desired output structures, the priors exploited to bridge the gap between imperfect sources and the desired output, and the main characteristics of geometry reasoning and data-driven approaches. We then identify and discuss the main sub-problems in structured reconstruction, reviewing state-of-the-art solutions for 3D room and floor-plan modeling and for interactive visual editing and exploration in standard and immersive settings. Examples of data-driven pipelines for depth and layout recovery, 3D floorplan recovery, and integration within interactive Extended Reality (XR) applications will be illustrated. The course closes with a review of relevant research issues and an analysis of research trends.

4 OUTLINE AND SCHEDULE

The course is organized in two sessions, with a 15' break and a final 25' Wrap-up and Q&A open discussion. The schedule is the following.

Duration	Lecturer	Topic	Sub-topics
10'	Pintore	Opening and introduction	Course motivation and outline; Presenters introduction; course overview
25'	Agus	Indoor capture, modeling, and exploration basics	Definitions & Applications; Tasks and model, Data capture; Panoramic cameras; Artifacts; Reconstruction priors; Open research data
45'	Pintore	Room modeling	Bounding surfaces; Exploiting priors; Deep learning solutions; Examples of data-driven pipelines for depth and layout recovery
15'		**BREAK**	
45'	Pintore	Integrated model computation	Multi-rooms; Multi-view; Segmentation and localization; Examples of data-driven pipelines for 3D floorplan recovery
60'	Agus	Visual representation generation and exploration	Appearance; Immersive panoramic exploration; Example of integration within interactive XR applications
25'	Agus, Pintore	Wrap-up and discussion + Q&A	Summary of techniques and assessment of capabilities; Open problems; Open discussion

5 COURSE CONTENT OVERVIEW

The content of each session is summarized in the following sections.

5.1 Opening and introduction

The introductory section introduces the organizer and speakers (see Appendix A) and provides a global overview of the course motivation and organization.

The tutorial's content is based on the authors' relevant experience, which has produced surveys, tutorials, and publications that have advanced the state-of-the-art in the various sub-fields targeted by this course. A comprehensive review and analysis of the 3D indoor reconstruction field has been published in Computer Graphics Forum [Pintore et al. 2020a] and presented in a talk at Eurographics and a half-day tutorial at SIGGRAPH 2020. A course focused on omnidirectional images has also been presented at CVPR2023. This course significantly expands the sub-topic centered on panoramic images, updating the survey of recent techniques, and expanding the section on exploration in standard and XR environments. These prior surveys and courses provide significant background material. We also direct the reader to complementary surveys on scene understanding from panoramic imaging [Gao et al. 2022], as well as extraction of 3D geometry from 360° imagery [da Silveira et al. 2022] for expanding the coverage of the subject matter.

On those topics, the authors have introduced innovations that will be discussed in this course, together with major publications (see references in Pintore et al. [Pintore et al. 2020a], plus relevant subsequent ones (e.g.,[Nauata et al. 2021; Zou et al. 2021]). These include deep-learning solutions (e.g., scene synthesis [Pintore et al. 2023], depth estimation and completion [Pintore et al. 2021a, 2024a], single-shot automatic emptying [Pintore et al. 2022], inference of Atlanta-world layouts using a slice-based representation [Pintore et al. 2021a] or of general 3D layouts with graph-convolutional networks [Pintore et al. 2021b]), geometry-reasoning or mixed techniques (e.g., reconstruction of multiroom environments from overlapping images [Pintore et al. 2019, 2018] or concurrent extraction of geometric, material and semantic signals [Shah et al. 2024]), as well as solutions for real-time exploration in XR settings (e.g., [Pintore et al. 2023, 2024b]), for photorealistic style transfer between indoor environments [Tukur et al. 2023b], and many others (see Appendix A).

5.2 Indoor capture, modeling, and exploration basics

The goal of structured 3D indoor reconstruction is to transform an input source containing a sampling of a real-world interior environment into a compact structured model containing both geometric and visual abstractions. A characterization of the typical structured indoor models and the main problems to be solved to create such models from the given input data was provided by Ikehata et al. [Ikehata et al. 2015]. In this tutorial, we focus on the main problems of individual room modeling (subsection 5.3), integrated model computation (subsection 5.4), and visual representation generation and exploration (subsection 5.5).

Regardless of the specific sub-problem addressed, each input source typically provides only partial coverage and imperfect sampling, complicating the reconstruction process and introducing ambiguities. Berger et al. [Berger et al. 2017], focusing on point clouds, have classified the most common artifacts into *uneven sampling density*, *noise*, *outliers*, *misalignment*, and *missing data*. Such artifacts, also prevalent in single- and multi-view 360° inputs, take specific forms when combining indoor environments and panoramic imagery. First of all, While 360° images capture the full context surrounding the viewer, they often suffer from uneven angular coverage and distortions due to acquisition settings (i.e., camera design) and data interchange formats (i.e., projections such as the equirectangular one). Moreover, indoor scenes further complicate scene analysis and understanding, as they are typically characterized by narrow spaces bounded by architectural elements such as walls, floors, and ceilings, and filled with various objects, including furniture. Thus, the depth distribution in indoor environments is uneven, ranging from furniture close-ups to distant features like ceilings, complicating the accurate prediction of metric depths and information extraction. Although the scene is often contained within architectural boundaries, structure recognition can be difficult due to cluttered and arbitrarily arranged objects that obscure large portions of walls and floors. Additionally, extensive untextured regions, such as bare walls, make associating geometric properties with specific points challenging. The presence of non-cooperative materials, such as mirrors or semi-transparent surfaces, further complicates the challenges.

Thus, without prior assumptions, the reconstruction problem for indoor environments is ill-posed, since an infinite number of solutions may exist that fit under-sampled, partially missing, or ambiguous data. For this reason, indoor reconstruction has focused its efforts on formally or implicitly restricting the target output model by introducing geometric priors for structural recovery such as *floor-wall* [Delage et al. 2006], *cuboid* [Hedau et al. 2009], *Manhattan world* [Coughlan and Yuille 1999], *Atlanta world* (a.k.a. *Augmented Manhattan World*) [Schindler and Dellaert 2004], *Indoor World Model* [Lee et al. 2009], *Vertical Walls* [Pintore et al. 2018], and *Piece-wise planarity* [Furukawa et al. 2009].

While early solutions incorporated these priors within algorithms that combined feature detection, matching, and geometric reasoning, recent years have seen the emergence of data-driven methods designed to address this traditionally ill-posed problem under less restrictive constraints, particularly through deep-learning approaches. The development and benchmarking of data-driven solutions are facilitated by the availability of synthetic and reality-based datasets, including Structured3D [Zheng et al. 2020], PNVS [Xu et al. 2021], Matterport3D [Chang et al. 2017], Stanford-2D-3D-S [Stanford University 2017], 360MonoDepth [Rey-Area et al. 2022], Habitat [Savva et al. 2019], Replica [Straub et al. 2019].

5.3 Room modeling

Room modeling focuses on reconstructing individual spaces. We differentiate between the generation of pixel-wise information – such as per-pixel depth, surface normals, or semantic labels – from the layout reconstruction problem, which involves parsing room spaces into the structural elements that bound their geometry (e.g. floor, ceilings, walls). The main focus, here is on monocular solutions, which are employed either independently or as foundational components in multi-view or multi-room pipelines (subsection 5.4).

Monocular depth reconstruction is taken as the main example of pixel-wise inference. While early methods mostly combined feature detection with geometric reasoning, recent research focuses on data-driven solutions information extracted from large training datasets [Pintore et al. 2021a]. In this context, 360° cameras are increasingly used for their ability to capture full surroundings in one image. Their exploitation includes adapting perspective methods through spherical convolutions [Payen de La Garanderie et al. 2018; Su and Grauman 2019; Su and Grauman 2017; Tateno et al. 2018; Zioulis et al. 2018], joint processing in mixed equirectangular and cube-map projections spaces [Wang et al. 2020], leveraging perspective views sampled on panoramic images before combining depth maps using transformers [Ai et al. 2023; Li et al. 2022; Rey-Area et al. 2022], as well as direct processing equirectangular images by exploiting gravity-aligned features to reduce network size [Pintore et al. 2021a; Sun et al. 2019].

3D layout reconstruction is more complex than depth estimation, since, instead of assigning a depth value to each visible pixel, it must also extrapolate substantial portions of the invisible structure, which may be occluded by both objects and the structure itself, resulting in multiple intersections per view ray. Thus, single-view layout computation must be capable of plausibly hallucinating the non-visible geometry. This need is also present in most multi-view cases, as full coverage is impractical in cluttered indoor environments. To address this complexity, several approaches operate within highly restrictive solution spaces. In particular, most methods target variants of the Manhattan World model (MWM: horizontal floors and ceilings, vertical walls meeting at right angles) [Sun et al. 2019; Zou et al. 2021], such as the Indoor World model (IWM: MWM with single horizontal ceiling and floor) [Wang et al. 2021] or the Atlanta World model (AWM: vertical walls with single horizontal ceiling and floor) [Pintore et al. 2020a]. Moreover, the most effective approaches recover the layout by exploiting projections to lower-dimensional spaces before expanding them to 3D. However, combining 1D/2D projections with restrictive priors limits the reconstruction capability to very few regular shapes and makes reconstruction less robust to occlusion. To mitigate spherical distortion and maximize the efficiency of modern deep learning techniques such as transformers, many recent approaches project the equirectangular input image to planar surfaces [Jiang et al. 2022; Pintore et al. 2020a; Wang et al. 2021; Yang et al. 2019; Zhao et al.

2022]. All these methods, however, require heavy preprocessing, such as detection of main MWM directions from vanishing lines analysis and related image warping [Lee et al. 2009; Zhang et al. 2014; Zou et al. 2021], or complex layout post-processing, such as MWM regularization of detected features [Shen et al. 2023; Sun et al. 2019; Yang et al. 2019; Zou et al. 2018]. To expand the solutions space, it has also been proposed to directly infer a watertight 3D mesh representation of the room shape using graph-convolutional networks [Pintore et al. 2021b].

In the course slides (section 6), we briefly summarize the main characteristics of these solutions and discuss the structure of their implementation.

5.4 Integrated model computation

The structured reconstruction of a complex environment requires not only the analysis of isolated structures, permanent or not, but also to ensure their integration into a coherent structured model.

Early approaches to infer vectorized geometries of permanent architectural structures combined low-level image processing with geometric reasoning and energy minimization solvers to extract room layouts [Cabral and Furukawa 2014; Furukawa et al. 2009; Ikehata et al. 2015; Monszpart et al. 2015; Silberman et al. 2012]. Many recent solutions adopt a hybrid approach, where neural networks first detect low-level primitives (e.g., corners, edges, region segments), then optimization techniques assemble them into the final models. Floor-SP [Chen et al. 2019] and Nauata et al. [Nauata and Furukawa 2020], in particular, rely on Mask R-CNN [He et al. 2017] to detect room segments and reconstruct polygons of individual rooms by sequentially solving shortest path problems, while MonteFloor [Stekovic et al. 2021] relies on Monte-Carlo Tree-Search to select room proposals. Alternative bottom-up methods, such as FloorNet [Liu et al. 2018], first detect room corners and then generate wall segments through integer programming. Also in the family of hybrid methods, diffusion approaches generate plausible room arrangements by combining graph neural networks with constrained diffusion [Gueze et al. 2023; Shabani et al. 2023]. Also related is the method of Shabani et al. [Shabani et al. 2021], which takes as input sparse panoramic images to generate plausible room displacements to find camera spatial registration.

In contrast to the hybrid solutions, several recent methods employ an end-to-end deep-learning approach. In particular, HEAT [Chen et al. 2022] proposes an end-to-end model, based on the deformable transformer (DETR) [Zhu et al. 2020], following a bottom-up pipeline: first detect corners, then classify edge candidates connecting corners. Also based on DETR [Zhu et al. 2020], RoomFormer [Yue et al. 2023] predicts floorplans from a dense point cloud using a single-stage, end-to-end trainable neural network. Differently from previous data-driven approaches [Chen et al. 2019], RoomFormer encodes the floorplan as a variable-size set of polygons, which are variable-length sequences of ordered vertices. By incorporating additional MWM priors and post-processing steps, SLIBO-Net [Su et al. 2023] focuses on improving RoomFormer's semantic and local geometric quality. More recently, PolyDiffuse [Chen et al. 2024] refines polygonal reconstructors from point cloud density maps through a conditional generation procedure.

In the course slides (section 6), we briefly summarize the main characteristics of these solutions, expanding on multi-view layout estimation, structured floorplan reconstruction, 3D scene reconstruction, and view localization, also providing an example of a deep-learning 3D floorplan recovery pipeline will be presented.

5.5 Visual representation generation and exploration

The geometric and topological descriptions coming out of the previously described steps may not be enough for the applications that should ultimately visualize the reconstructed model. Thus, the structured representation must often be enriched with information geared towards visual representation.

In this session, we introduce techniques that infer visual information, associate it with the topological and geometric models, and modify them to support editing operations [Tukur et al. 2023b]. We then discuss how these techniques can be exploited to create models suitable for regular or immersive exploration, either statically or by view synthesis. We, in particular, discuss methods for providing stereo cues and motion parallax starting from a single panoramic image, illustrating examples of XR applications exploiting head-mounted displays.

Even though capturing a single shot panorama is a very appealing way to create a virtual clone of a real environment, the limitation of presented content to what was visible around the fixed capture location leads to constraints and artifacts [Waidhofer et al. 2022], due to the reduction in degrees of freedom to just the rotation around the center of the panorama. In particular, binocular stereo and motion parallax, which are important aspects of immersion in VR, are missing. The fact that panoramas appear flat is a strong limitation in indoor environments, given the relatively short distance from the viewer to the architectural surfaces and the objects. Moreover, the large amount of clutter and occlusions encourages users to move their heads not only to see other angular portions of the environment but also to look behind occluding objects or architectural structures [Matzen et al. 2017]. To fully support immersion, a system must thus also respond to viewpoint translation. Even though many solutions have been proposed for multiview capture setups (e.g., [Attal et al. 2020; Broxton et al. 2020]), performing view synthesis from single-shot panoramas is of primary importance, due to the convenience and diffusion of sparse capturing through monocular 360° cameras [Waidhofer et al. 2022].

A first class of approaches targets the problem of providing a restricted motion, e.g., to support just stereo of small head movements. In the first case, the eyes move in a circle around the original capture position, while in the second case, they remain in a small volume (e.g., a ball around 50cm). This knowledge makes it possible to focus on specialized solutions to handle moderately small perspective changes and disocclusions. It also enables the creation of compact and fast-render customized representations valid for the known viewing environment.

A panoramic image with an accompanying depth map can be utilized for view synthesis using diverse approaches, such as directly rendering point clouds [Huang et al. 2017], generating and rendering view-independent meshes from depth maps [Tukur et al. 2023a], or integrating and blending depth maps or generated meshes

with multiple images or signals [Bertel et al. 2020; Luo et al. 2018; Pintore et al. 2016]. Recently, end-to-end view synthesis networks have been proposed to generate shifted panoramic views at run time [Pintore et al. 2023; Xu et al. 2021]. While these networks excel at inferring immersive views within a limited volume around the viewer, their computational demands preclude direct execution on embedded platforms. Consequently, Head-Mounted Displays (HMDs) are exclusively supported using these techniques at runtime via remote rendering [Pintore et al. 2023]. The generation of novel views by interpolating images taken at nearby viewpoints has also been widely researched, with effective solutions being proposed, even in the absence of a prior depth estimation step [Reda et al. 2022; Trinidad et al. 2019]. However, end-to-end networks tackling this task face similar computational constraints as depth estimation, limiting their applicability to interactive-rate frame generation on Head-Mounted Displays (HMDs). For this reason, often these methods are not directly used to generate images in response to head motion, but as building blocks to create precomputed representations that are faster to render.

An emerging approach for rapid novel viewpoint synthesis involves employing layered depth representations, associating each pixel with multiple depth values [Hedman and Kopf 2018]. This methodology has been effectively expanded to operate with single panoramic images [Lin et al. 2020; Serrano et al. 2019], as well as to create light field videos through layered mesh representations [Broxton et al. 2020]. For perspective views, multi-plane panoramas (MPI) have also been proposed as an output representation produced with convolutional neural networks [Tucker and Snavely 2020; Zhou et al. 2018]. However, MPIs are limited to viewpoints close to the origin and degrade when the viewpoint moves further. To address this limitation, adaptive sampling schemes have been proposed [Li and Khademi Kalantari 2020]. The concept of capturing the scene at multiple fixed depths has been extended for panoramic imaging by considering different capturing proxies like multi-spherical images (MSI) [Attal et al. 2020] or multi-cylinder images (MCI) [Waidhofer et al. 2022]. For the particular case of stereo-generation, Pintore et al. [Pintore et al. 2024b] have recently proposed to synthesize a discrete set of panoramic slices that cover the circular trajectory made by both eyes during head rotations and are oriented towards the main view directions. These images are subsequently blended to form an omnidirectional stereo pair comprised of two multiple-center-of-projection (MCOP) equirectangular images. The method builds on the multiperspective technique [Rademacher and Bishop 1998] based on circular projection stereo [Peleg and Ben-Ezra 1999] that aims to combine in a single image all the information required for stereo. For viewing, each vertical column of an equirectangular image has a different center of projection, corresponding to the position of the eye viewing it. By generating an image for the left eye and another one for the right eye, stereo is achieved. However, when viewing such an image in VR, stereo is only correct at the center of the image and degrades for peripheral vision. For this reason, other works in this area have concentrated on generating images that dynamically adapt to the user's gaze through the view-dependent rendering of depth images [Marrinan and Papka 2021].

For larger displacements from the original capture position, the input panorama must be transformed into a complete 3D renderable model. A first set of solutions exploits prior knowledge, learned from large sets of examples, on the semantics of the imaged room. Representative examples are Pano2CAD [Xu et al. 2017], Auto3DIndoor [Yang et al. 2018], DeepPanoContext [Zhang et al. 2021], and PanoContextFormer [Dong et al. 2024], which combine the estimation of single-room geometry or layout (see subsection 5.3) with the recognition of the type and pose of known objects. A full 3D model is then reconstructed. However, in real-world captures, interior environments are filled with objects with undefined/unrecognized semantics, leading to these methods failing to reconstruct complete real-world scenes.

Without requiring semantics, approaches based on Neural Radiance Fields (NeRF) [Mildenhall et al. 2021] or 3D Gaussian Splats (3DGS) [Kerbl et al. 2023] have demonstrated remarkable results in rendering novel views, but in their original formulation require a large number of views to work. Based on their concepts, many single-view novel-view synthesis methods have emerged. The underlying idea for these single-view extensions is to train the NeRF of 3DGS with synthetic novel views inferred by one of the above-described methods from the input panorama, letting the NeRF or 3DGS optimizer perform the merging of possibly inconsistent views. Representative examples of such methods include DietNeRF [Jain et al. 2021], Pix2NeRF [Cai et al. 2022], SinNerF [Xu et al. 2022], NerfDiff [Gu et al. 2023], NerDi [Deng et al. 2023], PERF [Wang et al. 2024], PixelSplat [Charatan et al. 2024], and Pano2Room [Pu et al. 2024]. Single-panorama novel view synthesis is handled, in state-of-the-art solutions [Pu et al. 2024; Wang et al. 2024] by generating images with depths corresponding to novel views and performing collaborative RGB-D inpainting to drive the creation of final NeRF representation. Dense sequential inpainting, as used in PERF [Wang et al. 2024], however, forces the virtual camera used for generating the virtual 3D views to stay on predefined trajectories. Since the views are merged in a sequence and several areas are under-sampled, ghost geometries are generated by underfitted model optimization. In addition, the large differences among views created by individual novel-view inferences lead to heavy blurring in occluded areas. Pano2Room [Pu et al. 2024] improves over this method by first converting the input panorama into a mesh through depth estimation (subsection 5.3) and then iteratively refining the mesh by leveraging a panoramic RGB-D inpainter to generate occluded color and geometry. The new content is gradually incorporated into the inpainted mesh, checking for visibility conflicts at each step. Finally, the inpainted mesh is converted to a 3D Gaussian Splat, training it with collected 3D-consistent pseudo novel views.

In the course slides (section 6), we provide an overview of these solutions and illustrate a reference implementation.

5.6 Wrap-up and discussion

Surround-view panoramic imaging provides the quickest and most complete per-image coverage and is supported by a wide variety of professional and consumer capture devices. For this reason, it is the target of much research. This course provides a comprehensive

overview of the rapidly advancing field of 3D indoor model inference from 360-degree images and. more briefly, of the techniques for exploring such models. The course slides (section 6) summarize the main take-home messages.

6 MATERIAL AND RESOURCES

The course website (https://www.crs4.it/vic/sigasia2024-course-pano/) provides the commented slides for all the tutorial sessions. The main discussed works are included in the bibliography of this article. Complementary information can be found in our survey on indoor reconstruction [Pintore et al. 2020a], on the SIGGRAPH 2020 course on the same topic (https://doi.org/10.1145/3388769.3407469) and on our CVPR 2023 tutorial focusing on indoor reconstruction from panoramic imagery (https://www.crs4.it/vic/cvpr2023-tutorial-pano/).

ACKNOWLEDGMENTS

GP and EG acknowledge the contribution of the Italian National Research Center in High-Performance Computing, Big Data and Quantum Computing. MA, GP, and EG received funding from NPRP-Standard (NPRP-S) 14th Cycle grant 0403-210132 AIN2 from the Qatar National Research Fund (a member of Qatar Foundation). The findings herein reflect the work and are solely the responsibility of the authors.

REFERENCES

Hao Ai, Zidong Cao, Yan-Pei Cao, Ying Shan, and Lin Wang. 2023. HRDFuse: Monocular 360° Depth Estimation by Collaboratively Learning Holistic-With-Regional Depth Distributions. In *Proc. CVPR*. 13273–13282.

Benjamin Attal, Selena Ling, Aaron Gokaslan, Christian Richardt, and James Tompkin. 2020. MatryODShka: Real-time 6DoF video view synthesis using multi-sphere images. In *Proc. ECCV*. 441–459.

Matthew Berger, Andrea Tagliasacchi, Lee M. Seversky, Pierre Alliez, Gaël Guennebaud, Joshua A. Levine, Andrei Sharf, and Claudio T. Silva. 2017. A Survey of Surface Reconstruction from Point Clouds. *Computer Graphics Forum* 36, 1 (2017), 301–329.

Tobias Bertel, Mingze Yuan, Reuben Lindroos, and Christian Richardt. 2020. OmniPhotos: Casual 360° VR Photography. *ACM TOG* 39, 6 (2020), 266:1–266:12.

Michael Broxton, John Flynn, Ryan Overbeck, Daniel Erickson, Peter Hedman, Matthew DuVall, Jason Dourgarian, Jay Busch, Matt Whalen, and Paul Debevec. 2020. Immersive Light Field Video with a Layered Mesh Representation. *ACM TOG* 39, 4 (2020), 86:1–86:15.

R. Cabral and Y. Furukawa. 2014. Piecewise Planar and Compact Floorplan Reconstruction from Images. In *Proc. CVPR*. 628–635.

Shengqu Cai, Anton Obukhov, Dengxin Dai, and Luc Van Gool. 2022. Pix2nerf: Unsupervised conditional P-GAN for single image to neural radiance fields translation. In *Proc. CVPR*. 3981–3990.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3D: Learning from RGB-D Data in Indoor Environments. In *Proc. 3DV*. 667–676.

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. 2024. PixelSplat: 3D Gaussian Splats from image pairs for scalable generalizable 3D reconstruction. In *Proc. CVPR*. 19457–19467.

Jiacheng Chen, Ruizhi Deng, and Yasutaka Furukawa. 2024. Polydiffuse: Polygonal shape reconstruction via guided set diffusion models. *NeurIPS* 36 (2024), 1863–1888.

Jiacheng Chen, Chen Liu, Jiaye Wu, and Yasutaka Furukawa. 2019. Floor-sp: Inverse cad for floorplans by sequential room-wise shortest path. In *Proc. CVPR*. 2661–2670.

Jiacheng Chen, Yiming Qian, and Yasutaka Furukawa. 2022. HEAT: Holistic Edge Attention Transformer for Structured Reconstruction. In *Proc. CVPR*. 3866–3875.

James M Coughlan and Alan L Yuille. 1999. Manhattan world: Compass direction from a single image by bayesian inference. In *Proc. ICCV*, Vol. 2. 941–947.

Thiago L. T. da Silveira, Paulo G. L. Pinto, Jeffri Murrugarra-Llerena, and Cláudio R. Jung. 2022. 3D Scene Geometry Estimation from 360° Imagery: A Survey. *ACM Comput. Surv.* 55, 4 (2022), 68:1–68:39.

E. Delage, Honglak Lee, and A. Y. Ng. 2006. A Dynamic Bayesian Network Model for Autonomous 3D Reconstruction from a Single Indoor Image. In *Proc. CVPR*, Vol. 2. 2418–2428.

Congyue Deng, Chiyu Jiang, Charles R Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. 2023. NeRDI: Single-view NeRF synthesis with language-guided diffusion as general image priors. In *Proc. CVPR*. 20637–20647.

Yuan Dong, Chuan Fang, Liefeng Bo, Zilong Dong, and Ping Tan. 2024. PanoContext-Former: Panoramic total scene understanding with a transformer. In *Proc. CVPR*. 28087–28097.

Yasutaka Furukawa, Brian Curless, Steven M Seitz, and Richard Szeliski. 2009. Reconstructing building interiors from images. In *Proc. ICCV*. IEEE, 80–87.

Shaohua Gao, Kailun Yang, Hao Shi, Kaiwei Wang, and Jian Bai. 2022. Review on Panoramic Imaging and Its Applications in Scene Understanding. *IEEE TIM* 71 (2022), 1–34.

Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. 2023. NeRFDiff: Single-image view synthesis with NeRF-guided distillation from 3D-aware diffusion. In *Proc. ICCV*. 11808–11826.

Arnaud Gueze, Matthieu Ospici, Damien Rohmer, and Marie-Paule Cani. 2023. Floor Plan Reconstruction from Sparse Views: Combining Graph Neural Network with Constrained Diffusion. In *Proc. CVPR*. 1583–1592.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. In *Proc. ICCV*. 2961–2969.

V. Hedau, D. Hoiem, and D. Forsyth. 2009. Recovering the spatial layout of cluttered rooms. In *Proc. ICCV*. 1849–1856.

Peter Hedman and Johannes Kopf. 2018. Instant 3D Photography. *ACM TOG*. 37, 4 (2018), 101:1–101:12.

Ruizhen Hu, Zeyu Huang, Yuhan Tang, Oliver Van Kaick, Hao Zhang, and Hui Huang. 2020. Graph2Plan: Learning floorplan generation from layout graphs. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 118–1.

Jingwei Huang, Zhili Chen, Duygu Ceylan, and Hailin Jin. 2017. 6-DOF VR videos with a single 360-camera. In *Proc. IEEE VR*. 37–44.

Satoshi Ikehata, Hang Yang, and Yasutaka Furukawa. 2015. Structured Indoor Modeling. In *Proc. ICCV*. 1323–1331.

Ajay Jain, Matthew Tancik, and Pieter Abbeel. 2021. Putting NeRF on a diet: Semantically consistent few-shot view synthesis. In *Proc. ICCV*. 5885–5894.

Zhigang Jiang, Zhongzheng Xiang, Jinhua Xu, and Ming Zhao. 2022. Lgt-net: Indoor panoramic room layout estimation with geometry-aware transformer network. In *Proc. CVPR*. 1654–1663.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM TOG* 42, 4 (2023), 139:1–139:14.

David C Lee, Martial Hebert, and Takeo Kanade. 2009. Geometric reasoning for single image structure recovery. In *Proc. CVPR*. 2136–2143.

Qinbo Li and Nima Khademi Kalantari. 2020. Synthesizing Light Field From a Single Image with Variable MPI and Two Network Fusion. *ACM TOG* 39, 6 (2020), 229:1–229:10.

Yuyan Li, Yuliang Guo, Zhixin Yan, Xinyu Huang, Ye Duan, and Liu Ren. 2022. Omnifusion: 360 monocular depth estimation via geometry-aware fusion. In *Proc. CVPR*. 2801–2810.

Kai-En Lin, Zexiang Xu, Ben Mildenhall, Pratul P. Srinivasan, Yannick Hold-Geoffroy, Stephen DiVerdi, Qi Sun, Kalyan Sunkavalli, and Ravi Ramamoorthi. 2020. Deep Multi Depth Panoramas for View Synthesis. In *Proc. ECCV*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). 328–344.

Chen Liu, Jiaye Wu, and Yasutaka Furukawa. 2018. FloorNet: A unified framework for floorplan reconstruction from 3D scans. In *Proc. ECCV*. 201–217.

Bicheng Luo, Feng Xu, Christian Richardt, and Jun-Hai Yong. 2018. Parallax360: Stereoscopic 360° Scene Representation for Head-Motion Parallax. *IEEE TVCG* 24, 4 (2018), 1545–1553.

Thomas Marrinan and Michael E Papka. 2021. Real-time omnidirectional stereo rendering: generating 360° surround-view panoramic images for comfortable immersive viewing. *IEEE TVCG* 27, 5 (2021), 2587–2596.

Kevin Matzen, Michael F. Cohen, Bryce Evans, Johannes Kopf, and Richard Szeliski. 2017. Low-cost 360 Stereo Photography and Video Capture. *ACM TOG* 36, 4 (2017), 148:1–148:12.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2021. NeRF: representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (dec 2021), 99—-106. https://doi.org/10.1145/3503250

Aron Monszpart, Nicolas Mellado, Gabriel J Brostow, and Niloy J Mitra. 2015. RAPter: rebuilding man-made scenes with regular arrangements of planes. *ACM TOG* 34, 4 (2015), 103–1.

Nelson Nauata and Yasutaka Furukawa. 2020. Vectorizing world buildings: Planar graph reconstruction by primitive detection and relationship inference. In *Proc. ECCV*. Springer, 711–726.

Nelson Nauata, Sepidehsadat Hosseini, Kai-Hung Chang, Hang Chu, Chin-Yi Cheng, and Yasutaka Furukawa. 2021. House-GAN++: Generative adversarial layout refinement network towards intelligent computational agent for professional architects. In *Proc. CVPR*. 13632–13641.

Gregoire Payen de La Garanderie, Amir Atapour Abarghouei, and Toby P. Breckon. 2018. Eliminating the Blind Spot: Adapting 3D Object Detection and Monocular Depth Estimation to 360 Panoramic Imagery. In *Proc. ECCV*. 812–830.

Shmuel Peleg and Moshe Ben-Ezra. 1999. Stereo panorama with a single camera. In *Proc. CVPR*. 395–401.

Giovanni Pintore, Marco Agus, Eva Almansa, and Enrico Gobbetti. 2022. Instant Automatic Emptying of Panoramic Indoor Scenes. *IEEE Transactions on Visualization and Computer Graphics* 28, 11 (November 2022), 3629–3639.

Giovanni Pintore, Marco Agus, Eva Almansa, Jens Schneider, and Enrico Gobbetti. 2021a. SliceNet: deep dense depth estimation from a single indoor panorama using a slice-based representation. In *Proc. CVPR*. 11536–11545.

Giovanni Pintore, Marco Agus, and Enrico Gobbetti. 2020a. AtlantaNet: Inferring the 3D Indoor Layout from a Single 360 Image Beyond the Manhattan World Assumption. In *Proc. ECCV*. 432–448.

Giovanni Pintore, Eva Almansa, Marco Agus, and Enrico Gobbetti. 2021b. Deep3DLayout: 3D Reconstruction of an Indoor Layout from a Spherical Panoramic Image. *ACM TOG* 40, 6 (2021), 250:1–250:12.

Giovanni Pintore, Eva Almansa, Armando Sanchez, Giorgio Vassena, and Enrico Gobbetti. 2024a. Deep Panoramic Depth Prediction and Completion for Indoor Scenes. *Computational Visual Media* 10 (February 2024), 1–20.

Giovanni Pintore, Fabio Bettio, Marco Agus, and Enrico Gobbetti. 2023. Deep scene synthesis of Atlanta-world interiors from a single omnidirectional image. *IEEE TVCG* 29, 11 (2023), 4708–4718.

Giovanni Pintore, Fabio Ganovelli, Enrico Gobbetti, and Roberto Scopigno. 2016. Mobile Mapping and Visualization of Indoor Structures to Simplify Scene Understanding and Location Awareness. In *Proc. ECCV Workshops*. 130–145.

Giovanni Pintore, Fabio Ganovelli, Alberto Jaspe Villanueva, and Enrico Gobbetti. 2019. Automatic modeling of cluttered floorplans from panoramic images. *Computer Graphics Forum* 38, 7 (2019), 347–358.

Giovanni Pintore, Fabio Ganovelli, Ruggero Pintus, Roberto Scopigno, and Enrico Gobbetti. 2018. 3D floor plan recovery from overlapping spherical images. *Computational Visual Media* 4, 4 (2018), 367–383.

Giovanni Pintore, Alberto Jaspe-Villanueva, Markus Hadwiger, Jens Schneider, Marco Agus, Fabio Marton, Fabio Bettio, and Enrico Gobbetti. 2024b. Deep synthesis and exploration of omnidirectional stereoscopic environments from a single surround-view panoramic image. *Computers & Graphics* 119 (March 2024), 103907.

Giovanni Pintore, Claudio Mura, Fabio Ganovelli, Lizeth Fuentes-Perez, Renato Pajarola, and Enrico Gobbetti. 2020b. State-of-the-art in Automatic 3D Reconstruction of Structured Indoor Environments. *Comput. Graph. Forum* 39, 2 (2020), 667–699.

Guo Pu, Yiming Zhao, and Zhouhui Lian. 2024. Pano2Room: Novel View Synthesis from a Single Indoor Panorama. In *Proc. SIGGRAPH Asia Conference Papers*. 11 pages. To appear.

Paul Rademacher and Gary Bishop. 1998. Multiple-center-of-projection images. In *Proc. SIGGRAPH*. 199–206.

Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. 2022. FILM: Frame interpolation for large motion. In *Proc. ECCV*. 250–266.

Manuel Rey-Area, Mingze Yuan, and Christian Richardt. 2022. 360MonoDepth: High-Resolution 360° Monocular Depth Estimation. In *Proc. CVPR*. 3762–3772.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. 2019. Habitat: A Platform for Embodied AI Research. In *Proc. ICCV*. 9339–9347.

G. Schindler and F. Dellaert. 2004. 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. I–I.

Ana Serrano, Incheol Kim, Zhili Chen, Stephen DiVerdi, Diego Gutierrez, Aaron Hertzmann, and Belen Masia. 2019. Motion parallax for 360 RGBD video. *IEEE TVCG* 25, 5 (2019), 1817–1827. Proc. IEEE VR.

Mohammad Amin Shabani, Sepidehsadat Hosseini, and Yasutaka Furukawa. 2023. HouseDiffusion: Vector Floorplan Generation via a Diffusion Model with Discrete and Continuous Denoising. In *Proc. CVPR*. 5466–5475.

Mohammad Amin Shabani, Weilian Song, Makoto Odamaki, Hirochika Fujiki, and Yasutaka Furukawa. 2021. Extreme structure from motion for indoor panoramas without visual overlaps. In *Proc. ICCV*. 5683–5691.

Uzair Shah, Muhammad Tukur, Mahmood Alzubaidi, Giovanni Pintore, Enrico Gobbetti, Mowafa Househ, Jens Schneider, and Marco Agus. 2024. MultiPanoWise: holistic deep architecture for multi-task dense prediction from a single panoramic image. In *Proc. CVPRW - OmniCV*. 1311–1321.

Zhijie Shen, Zishuo Zheng, Chunyu Lin, Lang Nie, Kang Liao, Shuai Zheng, and Yao Zhao. 2023. Disentangling Orthogonal Planes for Indoor Panoramic Room Layout Estimation with Cross-Scale Distortion Awareness. In *Proc. CVPR*. 17337–17345.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. 2012. Indoor segmentation and support inference from rgbd images. In *Proc. ECCV*. Springer, 746–760.

Stanford University. 2017. BuildingParser Dataset. http://buildingparser.stanford.edu/dataset.html. [Accessed: 2019-09-25].

Sinisa Stekovic, Mahdi Rad, Friedrich Fraundorfer, and Vincent Lepetit. 2021. Monte-floor: Extending MCTS for reconstructing accurate large-scale floor plans. In *Proc. CVPR*. 16034–16043.

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. 2019. The Replica Dataset: A Digital Replica of Indoor Spaces. *ArXiv e-print arXiv:1906.05797* (2019), 1–10.

Jheng-Wei Su, Kuei-Yu Tung, Chi-Han Peng, Peter Wonka, and Hung-Kuo Chu. 2023. SLIBO-Net: Floorplan Reconstruction via Slicing Box Representation with Local Geometry Regularization. In *Proc. NeurIPS*. 1–12.

Y. Su and K. Grauman. 2019. Kernel Transformer Networks for Compact Spherical Convolution. In *Proc. CVPR*. 9434–9443.

Yu-Chuan Su and Kristen Grauman. 2017. Learning Spherical Convolution for Fast Features from 360 Imagery. In *Proc. NeurIPS*. 529–539.

Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. 2019. HorizonNet: Learning room layout with 1D representation and pano stretch data augmentation. In *Proc. CVPR*. 1047–1056.

Keisuke Tateno, Nassir Navab, and Federico Tombari. 2018. Distortion-Aware Convolutional Filters for Dense Prediction in Panoramic Images. In *Proc. ECCV*. 732–750.

Marc Comino Trinidad, Ricardo Martin Brualla, Florian Kainz, and Janne Kontkanen. 2019. Multi-view image fusion. In *Proc. ICCV*. 4101–4110.

Richard Tucker and Noah Snavely. 2020. Single-View View Synthesis With Multiplane Images. In *Proc. CVPR*. 548–557.

Muhammad Tukur, Giovanni Pintore, Enrico Gobbetti, Jens Schneider, and Marco Agus. 2023a. SPIDER: A framework for processing, editing and presenting immersive high-resolution spherical indoor scenes. *Graphical Models* 128 (2023), 101182:1–101182:11.

Muhammad Tukur, Atiq Ur Rehman, Giovanni Pintore, Enrico Gobbetti, Jens Schneider, and Marco Agus. 2023b. PanoStyle: Semantic, Geometry-Aware and Shading Independent Photorealistic Style Transfer for Indoor Panoramic Scenes. In *Proc. ICCVW*. 1553–1564.

John Waidhofer, Richa Gadgil, Anthony Dickson, Stefanie Zollmann, and Jonathan Ventura. 2022. PanoSynthVR: Toward Light-weight 360-Degree View Synthesis from a Single Panoramic Input. In *Proc. ISMAR*. 584–592.

Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. 2020. BiFuse: Monocular 360 Depth Estimation via Bi-Projection Fusion. In *Proc. CVPR*. 462–471.

Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. 2021. LED2-Net: Monocular 360 Layout Estimation via Differentiable Depth Rendering. In *Proc. CVPR*. 12956–12965.

Guangcong Wang, Peng Wang, Zhaoxi Chen, Wenping Wang, Chen Change Loy, and Ziwei Liu. 2024. PERF: Panoramic Neural Radiance Field from a Single Panorama. *IEEE TPAMI* (2024), 1–15.

Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. 2022. SinNeRF: Training neural radiance fields on complex scenes from a single image. In *Proc. ECCV*. 736–753.

J. Xu, B. Stenger, T. Kerola, and T. Tung. 2017. Pano2CAD: Room Layout from a Single Panorama Image. In *Proc. WACV*. 354–362.

Jiale Xu, Jia Zheng, Yanyu Xu, Rui Tang, and Shenghua Gao. 2021. Layout-guided novel view synthesis from a single indoor panorama. In *Proc. CVPR*. 16438–16447.

Mai Xu, Chen Li, Shanyi Zhang, and Patrick Le Callet. 2020. State-of-the-art in 360° video/image processing: Perception, assessment and compression. *IEEE Journal of Selected Topics in Signal Processing* 14, 1 (2020), 5–26.

Sheng Yang, Beichen Li, Yan-Pei Cao, Hongbo Fu, Yu-Kun Lai, Leif Kobbelt, and Shi-Min Hu. 2020. Noise-Resilient Reconstruction of Panoramas and 3D Scenes Using Robot-Mounted Unsynchronized Commodity RGB-D Cameras. *ACM TOG* 39, 5 (2020), 152:1–152:15.

Shang-Ta Yang, Fu-En Wang, Chi-Han Peng, Peter Wonka, Min Sun, and Hung-Kuo Chu. 2019. DuLa-Net: A Dual-Projection Network for Estimating Room Layouts from a Single RGB Panorama. In *Proc. CVPR*.

Yang Yang, Shi Jin, Ruiyang Liu, , and Jingyi Yu. 2018. Automatic 3D Indoor Scene Modeling From Single Panorama. In *Proc. CVPR*. 3926–3934.

Yuanwen Yue, Theodora Kontogianni, Konrad Schindler, and Francis Engelmann. 2023. Connecting the Dots: Floorplan Reconstruction Using Two-Level Queries. In *Proc. CVPR*.

Cheng Zhang, Zhaopeng Cui, Cai Chen, Shuaicheng Liu, Bing Zeng, Hujun Bao, and Yinda Zhang. 2021. DeepPanoContext: Panoramic 3D scene understanding with holistic scene context graph and relation-based optimization. In *Proc. ICCV*. 12632–12641.

Yinda Zhang, Shuran Song, Ping Tan, and Jianxiong Xiao. 2014. PanoContext: A Whole-Room 3D Context Model for Panoramic Scene Understanding. In *Proc. ECCV*. 668–686.

Yining Zhao, Chao Wen, Zhou Xue, and Yue Gao. 2022. 3D Room Layout Estimation from a Cubemap of Panorama Image via Deep Manhattan Hough Transform. In *Proc. ECCV*. Springer, 637–654.

Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. 2020. Structured3D: A Large Photo-realistic Dataset for Structured 3D Modeling. In *Proc. ECCV*. 519–535.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. 2018. Stereo Magnification: Learning View Synthesis Using Multiplane Images. *ACM TOG* 37, 4 (2018), 68:1–68:12.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2020. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv preprint arXiv:2010.04159* (2020).

Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, and Petros Daras. 2018. OmniDepth: Dense Depth Estimation for Indoors Spherical Panoramas. In *Proc. ECCV*. 453–471.

Chuhang Zou, Alex Colburn, Qi Shan, and Derek Hoiem. 2018. LayoutNet: Reconstructing the 3D Room Layout from a Single RGB Image. In *Proc. CVPR*. 2051–2059.

Chuhang Zou, Jheng Wei Su, Chi Han Peng, Alex Colburn, Qi Shan, Peter Wonka, Hung Kuo Chu, and Derek Hoiem. 2021. Manhattan Room Layout Reconstruction from a Single 360 Image: A Comparative Study of State-of-the-Art Methods. *International Journal of Computer Vision* 129 (2021), 1410–1431.

A BRIEF BIO SKETCHES OF COURSE AUTHORS

	Giovanni Pintore (co-author and presenter, in person)
Bio	Giovanni Pintore is a senior researcher at the CRS4 research center in Italy and is affiliated with the Italian National Research Center in HPC and Quantum Computing. He has coordinated and managed several international research and industrial projects in various fields, from space exploration to security management in urban environments. His research, widely published in major journals and conferences, spans many computer graphics and computer vision areas, including deep learning architectures, geometry reasoning, panoramic scene understanding, multiresolution representations of large and complex 3D models, 3D multi-view reconstruction, and new generation mobile graphics. He regularly serves the scientific community through participation in conference committees and executive boards. His primary research focus is now on 3D reconstruction and immersive exploration of structured indoor scenes from omnidirectional images, on whose topic he has recently published papers at ISMAR, ECCV, CVPR, and CGF, and given courses at SIGGRAPH, SIGGRAPH Asia, 3DV, and CVPR in recent years.
ORCID	https://orcid.org/0000-0001-8944-1045
More	https://www.crs4.it/vic/cgi-bin/people-page.cgi?name=giovanni.pintore

	Marco Agus (co-author and presenter, in person)
Bio	Marco Agus is an associate professor at Hamad Bin Khalifa University (HBKU) - Qatar Foundation in Doha, Qatar. He was previously a research engineer at King Abdullah University of Science and Technology (KAUST), in Jeddah, Saudi Arabia, and a research scientist at Center of Research, Development and Advanced Studies (CRS4), in Cagliari, Italy. He obtained an M.Sc. and Ph.D. from the University of Cagliari, Italy. His research interests span different domains in visual computing, from haptics and visual rendering for medical applications to real-time exploration of massive models, to machine learning methods for electron microscopy biology data and indoor environments. He published more than 50 peer-reviewed papers on these topics. He taught courses at several important visual computing venues, including CVPR, 3DV, ACM SIGGRAPH, and Eurographics, and he regularly acts as a committee member, reviewer, chair, and associate editor for top journals and conferences in the visual computing domain.
ORCID	https://orcid.org/0000-0003-2752-3525
More	https://www.hbku.edu.qa/en/cse/staff/marco-agus

	Enrico Gobbetti (co-author and organizer)
Bio	Enrico Gobbetti is the director of Visual and Data-intensive Computing (ViDiC) at the Center for Advanced Studies, Research, and Development in Sardinia (CRS4), Italy, and is affiliated with the Italian National Research Center in HPC and Quantum Computing. He holds an Engineering degree (1989) and a Ph.D. degree (1993) in Computer Science from the Swiss Federal Institute of Technology in Lausanne (EPFL), as well as Full Professor Habilitations in Computer Science and Information Processing from the Italian Ministry of University and Research. Before joining CRS4, he held positions at EPFL (Switzerland), UMBC (USA), and NASA/CESDIS (USA). At CRS4, Enrico develops and manages a research program in visual and data-intensive computing supported through institutional, industrial, and government grants, including many national and international collaborative projects. His research spans many visual and data-intensive computing areas and is widely published in major journals and conferences. The primary focus is the creation of innovative solutions for the acquisition, creation, processing, distribution, and exploration of complex and/or massive datasets and real-world objects and environments. He regularly serves the scientific community through participation in editorial boards, conference committees, working groups, and steering boards, as well as through the organization and chairing of conferences. He is a Fellow of the Eurographics Association.
ORCID	https://orcid.org/0000-0003-0831-2458
More	https://www.crs4.it/vic/people/CV-Gobbetti/