Automatic 3D modeling and editing of immersive indoor environments from a single omnidirectional image

Giovanni Pintore

Doctor of Philosophy *(Ph.D.)* Thesis 2024

Supervisors Dr. Alberto Jaspe Villanueva
Research Scientist, KAUST

Dr. Julián Dorado de la Calle
Professor, UDC

Tutor Dr. Julián Dorado de la Calle

UNIVERSIDADE DA CORUÑA

PhD Programme in Information and Communications Technology

Automatic 3D modeling and editing of immersive indoor environments from a single omnidirectional image

Doctor of Philosophy *(Ph.D.)* Thesis 2024

PhD Programme in Information and Communications Technology

Author: Giovanni Pintore

Supervisors: Dr. Alberto Jaspe Villanueva (University of A Coruña)
Dr. Julián Dorado de la Calle (VCC, KAUST)

Tutor: Dr. Julián Dorado de la Calle (University of A Coruña)

University of A Coruña

Department of Computer Science

Faculty of Computer Science

Campus de Elviña, S//N

15071 – A Courña, Spain

Giovanni Pintore Dr. Alberto Jaspe Villanueva Dr. Julián Dorado de la Calle

Cagliari (Italy), Thuwal (Saudi Arabia), and A Coruña (Spain)April 2024.

Dr. D. Alberto Jaspe Villanueva, Científico Investigador, King Abdullah University of Science and Technology (KAUST), Arabia Saudi.

Dr. D. Julián Dorado de la Calle, Catedrático del Departamento de Tecnologías de la Información y las Comunicaciones, Universidade da Coruña (UDC), España.

Dr. Mr. Alberto Jaspe Villanueva, Research Scientist, King Abdullah University of Science and Technology (KAUST), Saudi Arabia.

Dr. Mr. Julián Dorado de la Calle, Full Professor of the Information Technologies Department, University of A Coruña (UDC), Spain.

Atestan

Attest

Que la memoria titulada "**Automatic 3D modeling and editing of immersive indoor environments from a single omnidirectional image**" presentada por **Giovanni Pintore**, ha sido realizada bajo nuestra dirección. Considerando que el trabajo constituye tema de Tesis Doctoral, se autoriza su presentación en la Universidade da Coruña.

That the dissertation entitled "Automatic 3D modeling and editing of immersive indoor environments from a single omnidirectional image" presented by Giovanni Pintore, has been developed under our advising. Considering that the work is subject of Doctoral Thesis, we authorize its presentation at the University of A Coruña.

Y para que así conste, se expide el presente certificado en A Coruña (España) y Thuwal (Saudi Arabia), en Abril del 2024.

This certificate is issued in A Coruña (Spain) and Thuwal (Saudi Arabia) in April 2024.

Fdo. Dr. D. Alberto Jaspe Villanueva

Fdo. Dr. D. Julián Dorado de la Calle

Alla mia meravigliosa famiglia

A mi familia maravillosa

To my wonderful family

Acknowledgements

I would like to thank a number of people who supported me, both professionally and personally, on my journey to complete this work. I thank Alberto Jaspe Villanueva and Julian Dorado de la Calle, who made this project possible. I thank my collaborators at CRS4 and beyond, Marco Agus, Eva Almansa, Fabio Bettio, Fabio Marton, Ruggero Pintus, Antonio Zorcolo and Moonisa Ahsan, and especially Enrico Gobbetti, who has actively supported me for more than 20 years, and to whom I owe everything I (hopefully) learned.

However, i would like to thank my family, who supported me in every way and allowed me to achieve this goal, especially my beloved wife Cristina, my wonderful daughters Elisabetta and Isabella, my parents Giuseppina and Giuseppe, and my brother Emanuele.

Graciñas, y gracias, e grazie, e gràtzias and thanks.

Cagliari, April 2024 *Giovanni Pintore*

Resumo

Nos últimos anos, houbo un interese significativo na investigación da reconstrución 3D automática e o modelado de escenas de interior, o que resulta nun campo emerxente ben definido[1]. Neste contexto, a adquisición de panorámicas 360 graos xurdiu como unha solución eficaz para as contornas de interior. Ofrece unha cobertura rápida e completa, mesmo desde un único punto de vista, e é compatible cunha ampla gama de dispositivos de adquisición profesionais e de consumo, o que fai que a captura de interiores sexa eficiente e rendible[2]. As imaxes panorámicas tamén se converteron en parte integrante da creación de contidos inmersivos directamente a partir de escenas do mundo real e en soporte de diversas aplicacións de Realidade Virtual (RV)[3]. En particular, as visitas virtuais baseadas en imaxes esféricas gañaron popularidade no sector inmobiliario, especialmente durante o período de pandemia. Para lograr unha inmersión total, o sistema debe responder tamén á translación do punto de vista. Aínda que se propuxeron moitas solucións para as configuracións de captura multivista (por exemplo, [4, 5]), realizar a síntese da vista a partir de panoramas dunha soa toma é de gran importancia, debido á conveniencia e difusión da captura dispersa a través de cámaras monoculares de 360°[6]. Con todo, a síntese de vistas arbitrarias depende da estimación do modelo xeométrico da contorna da imaxe, de forma explícita ou implícita, para realizar unha reproyección consciente da oclusión e sintetizar o contido non obstruído. Este aspecto é aínda máis crucial se o propósito é tamén derivar outra información non obvia da vista orixinal, como, por exemplo, derivar un modelo da estrutura permanente sen desorde[7]. Para lograr unha visualización inmersiva e unha edición eficaz na reconstrución 3D en interiores, é necesario abordar varias cuestións de investigación fundamentais, que están relacionadas coas tarefas de estimación da profundidade e a disposición e a síntese do renderizado de novos puntos de vista.

Neste proxecto de investigación, propuxémonos ampliar a estado da arte nestas tarefas fundamentais e, en particular, na súa combinación orientada á exploración

e edición inmersiva en interiores, partindo dunha única imaxe de 360 graos. Con este fin, investigamos novos enfoques para explotar as características arquitectónicas previas en interiores, que teñen en conta as características moi específicas da contorna creada polo home, e solucións eficaces baseadas en datos, que aprenden relacións ocultas a partir de exemplos de big data.

As nosas contribucións dan lugar a varias solucións innovadoras de tipo *end-to-end*, como unha nova metodoloxía para a síntese de escenas 3D de interiores de tipo *Atlanta-world* a partir dunha única imaxe omnidireccional, un enfoque novo para a síntese e exploración en profundidade de contornas estereoscópicos omnidireccionales a partir dunha imaxe panorámica monoscópica, así como unha técnica innovadora para o baleirado automático e instantáneo de escenas de interiores panorámicas, que permite limpar de mobles e outras oclusiones a imaxe e mostrar a arquitectura basee das habitacións. Este tese presenta as meteodologías e resultados obtidos durante a devandita investigación.

Palabras chave: Informática Visual, Visión por Computador, Gráficos por Computador, Captura Esférica, Captura Omnidireccional, Captura Panorámica, Proxección Equirectangular, Reconstrución 3D, Contorna Interior, Visión Monocular, Estimación da Profundidade, Estimación do Trazado 3D, Síntese da Vista, Exploración Inmersiva, Realidade Virtual, Realidade Diminuída.

Resumen

En los últimos años, ha habido un interés significativo en la investigación de la reconstrucción 3D automática y el modelado de escenas de interior, lo que resulta en un campo emergente bien definido [1]. En este contexto, la adquisición de panorámicas 360 grados ha surgido como una solución eficaz para los entornos de interior. Ofrece una cobertura rápida y completa, incluso desde un único punto de vista, y es compatible con una amplia gama de dispositivos de adquisición profesionales y de consumo, lo que hace que la captura de interiores sea eficiente y rentable [2]. Las imágenes panorámicas también se han convertido en parte integrante de la creación de contenidos inmersivos directamente a partir de escenas del mundo real y en soporte de diversas aplicaciones de Realidad Virtual (RV) [3]. En particular, las visitas virtuales basadas en imágenes esféricas han ganado popularidad en el sector inmobiliario, especialmente durante el periodo de pandemia. Para lograr una inmersión total, el sistema debe responder también a la traslación del punto de vista. Aunque se han propuesto muchas soluciones para las configuraciones de captura multivista (por ejemplo, [4, 5]), realizar la síntesis de la vista a partir de panoramas de una sola toma es de gran importancia, debido a la conveniencia y difusión de la captura dispersa a través de cámaras monoculares de 360° [6]. Sin embargo, la síntesis de vistas arbitrarias depende de la estimación del modelo geométrico del entorno de la imagen, de forma explícita o implícita, para realizar una reproyección consciente de la oclusión y sintetizar el contenido no obstruído. Este aspecto es aún más crucial si el propósito es también derivar otra información no obvia de la vista original, como, por ejemplo, derivar un modelo de la estructura permanente sin desorden [7]. Para lograr una visualización inmersiva y una edición eficaz en la reconstrucción 3D en interiores, es necesario abordar varias cuestiones de investigación fundamentales, que están relacionadas con las tareas de estimación de la profundidad y la disposición y la síntesis del renderizado de nuevos puntos de vista.

En este proyecto de investigación, nos propusimos ampliar el estado del arte en estas tareas fundamentales y, en particular, en su combinación orientada a la exploración y edición inmersiva en interiores, partiendo de una única imagen de 360 grados. Con este fin, investigamos nuevos enfoques para explotar las características arquitectónicas previas en interiores, que tienen en cuenta las características muy específicas del entorno creado por el hombre, y soluciones eficaces basadas en datos, que aprenden relaciones ocultas a partir de ejemplos de big data.

Nuestras contribuciones dan lugar a varias soluciones innovadoras de tipo *end-to-end*, como una nueva metodología para la síntesis de escenas 3D de interiores de tipo *Atlanta-world* a partir de una única imagen omnidireccional, un enfoque novedoso para la síntesis y exploración en profundidad de entornos estereoscópicos omnidireccionales a partir de una imagen panorámica monoscópica, así como una técnica innovadora para el vaciado automático e instantáneo de escenas de interiores panorámicas, que permite limpiar de muebles y otras oclusiones la imagen y mostrar la arquitectura base de las habitaciones. Este tesis presenta las meteodologías y resultados obtenidos durante dicha investigación.

Palabras clave: Informática Visual, Visión por Computador, Gráficos por Computador, Captura Esférica, Captura Omnidireccional, Captura Panorámica, Proyección Equirectangular, Reconstrucción 3D, Entorno Interior, Visión Monocular, Estimación de la Profundidad, Estimación del Trazado 3D, Síntesis de la Vista, Exploración Inmersiva, Realidad Virtual, Realidad Disminuida.

Abstract

Over the past few years, there has been significant research interest in the automatic 3D reconstruction and modeling of indoor scenes, resulting in a well-defined emerging fieldc [1]. Within this context, 360-degree panoramic acquisition has emerged as an effective solution for indoor environments. It offers rapid and comprehensive coverage, even from a single viewpoint, and is compatible with a wide range of professional and consumer acquisition devices, making indoor data capture efficient and cost-effective [2]. Panoramic images have also become integral to creating immersive content directly from real-world scenes and supporting various Virtual Reality (VR) applications [3]. Notably, virtual tours based on spherical images have gained popularity in the real estate industry, especially during the pandemic period. To fully support immersion, a system must thus also respond to viewpoint translation. While many solutions have been proposed for multiview capture setups (e.g., [4, 5]), performing view synthesis from single-shot panoramas is of primary importance, due to the convenience and diffusion of sparse capturing through monocular 360° cameras [6]. However, view synthesis relies on estimating the geometric model of the imaged environment, explicitly or implicitly, to perform occlusion-aware reprojection and synthesize disoccluded content. This aspect is even more crucial if the purpose is also to derive other non-obvious information from the original view, such as, for example, deriving a model of the permanent structure without clutter [7]. To achieve immersive visualization and effective editing in indoor 3D reconstruction, it is necessary to address several fundamental research questions related to depth and layout estimation and novel view synthesis.

In our research project, we proposed to extend the state of the art in these fundamental tasks and particularly in their combination aimed at indoor immersive exploration and editing, just starting from a single 360-degree image. To this end, we researched novel approaches to exploit indoor architectural priors, that take in account the very specific man-made environment features, and effective data-driven solutions, that

learn hidden relations from big-data examples. Our contributions result in several, innovative, end-to-end solutions, such as a novel methodology for 3D scene synthesis of Atlanta-world interiors from a single omnidirectional image, a novel approach for deep synthesis and exploration of omnidirectional stereoscopic environments from a monoscopic panoramic image, an innovative end-to-end technique for instant automatic emptying of panoramic indoor scenes. This thesis presents the results obtained during such a research.

Keywords: Visual Computing, Computer Vision, Computer Graphics, Spherical Capture, Omnidirectional Capture, Panoramic Capture, Equirectangular Projection, 3D Reconstruction, Indoor Environment, Monocular Vision, Depth Estimation, 3D Layout Estimation, View Synthesis, Immersive Exploration, Virtual Reality, Diminished Reality.

Preface

THIS thesis summarizes the candidate's research activities in the field of 3D reconstruction of indoor environments aimed at immersive experience with modern VR devices. The presented results come from a deeper background in the field of reconstruction and visualization of large complex 3D models, gained by the candidate over the years within the Visual Computing Group of the CRS4 (Center for Advanced Studies, Research and Development in Sardinia, Italy). The candidate, in particular, has been involved in the recent period in a fruitful research activity in indoor reconstruction from panoramic images with deep learning methods, managing in this context several international projects and actively participating in the scientific community with frequent publications, invited talks and courses in the most important international venues.

Thanks to the help and supervision of Dr. Alberto Jaspe Villanueva and Prof. Julián Dorado de la Calle, a significant part of such a research has been finalized in this PhD project and thesis.

Contents

1 Introduction **1**

1.1 Background and motivation 1
1.2 Objectives 3
1.3 Achievements 6
1.4 Organization 8

2 General requirements, work hypotheses, and means of verification **10**

2.1 Basic definitions 10
2.2 Research questions 12
2.3 Hypotheses supporting the prospected solutions 13
2.4 Means of verification 14

3 Recovering a 3D indoor model for novel view-synthesis **16**

3.1 Introduction 16
3.2 Contributions 19
3.3 Related Work 20
3.3.1 Depth estimation from panoramic images 21
3.3.2 Layout estimation from panoramic images 21
3.3.3 Novel view synthesis 22
3.4 Methods 23
3.4.1 Depth and 3D layout prediction 25
3.4.2 Novel view synthesis 29
3.4.3 Training and losses 32
3.5 Results 34
3.5.1 Training and testing datasets 34
3.5.2 Setup and computational performance 35
3.5.3 Run-time performance 37
3.5.4 Performance vs. ground truth and competitors 37
3.5.5 Discussion and ablation study 41
3.6 Client server design 42
3.7 Neural Network architecture details 45
3.7.1 Depth estimation block architecture 45
3.7.2 Room contour extraction and metric scaling 47
3.7.3 View synthesis network details 47

3.8 Additional experiments . . . 48
3.8.1 Additional prediction examples . . . 48
3.8.2 Examples with predicted work area . . . 48
3.8.3 Additional comparisons with PanoSynthVR . . . 49
3.8.4 Failure cases . . . 49
3.9 Conclusions . . . 50
3.10 Bibliographic notes . . . 50

4 Immersive exploration of indoor stereoscopic environments 51
4.1 Introduction . . . 51
4.2 Contributions . . . 54
4.3 Related work . . . 55
4.3.1 Depth estimation from a single panorama . . . 55
4.3.2 Novel view synthesis . . . 56
4.3.3 View interpolation . . . 56
4.3.4 Omnidirectional stereo display . . . 57
4.4 Method overview . . . 58
4.5 Single panorama depth estimation . . . 59
4.6 Synthesis of novel views . . . 63
4.7 Omnidirectional stereo generation and rendering . . . 65
4.8 Results . . . 68
4.8.1 Dataset and training . . . 68
4.8.2 Computational performance . . . 70
4.8.3 View-synthesis performance . . . 70
4.8.4 Stereoscopic exploration on HMD . . . 72
4.9 Conclusions . . . 75
4.10 Bibliographic notes . . . 76

5 Automatic-assisted editing of immersive indoor models 77
5.1 Introduction . . . 78
5.2 Contributions . . . 80
5.3 Related Work . . . 81
5.3.1 Diminished reality for indoor spaces . . . 82
5.3.2 Data-driven inpainting . . . 83
5.3.3 Image-to-image translation . . . 83
5.3.4 Uncluttered depth estimation . . . 84
5.4 Methods . . . 85
5.4.1 Clutter mask prediction . . . 85
5.4.2 Empty scene synthesis . . . 87
5.4.3 Training and losses . . . 89
5.5 Results . . . 91
5.5.1 Training and testing datasets . . . 92

5.5.2 Setup and computational performance 93
5.5.3 Performance vs. ground truth and competitors 95
5.5.4 Performance in-the-wild 97
5.5.5 Discussion and ablation study 98
5.6 Conclusions 100
5.7 Bibliographic notes 101

6 Conclusion **102**
6.1 Overview of achievements 103
6.2 Discussion and future directions 105
6.3 Publications 108

Bibliography **112**

Appendix A Sinopsis (thesis summary in Spanish) **124**
A.1 Contexto, motivación e hipótesis 125
A.2 Objetivos 129
A.3 Reconstrucción de un modelo 3D de interiores para la síntesis de nuevos puntos de vista 132
A.4 Exploración inmersiva de entornos interiores estereoscópicos 135
A.5 Edición semi-automática de modelos inmersivos de interiores 137
A.6 Logros y conclusiones 140

Appendix B Curriculum Vitae **146**

List of Figures

2.1 **Architectural priors.** A list of architectural priors used in 3D reconstruction, in order of complexity (Pintore et al., CVPR 2023 [43]). 11

3.1 Given a single 360° panorama of an indoor scene, we compute an enriched geometric and structural representation, from which novel panoramas from other close-by viewpoints can be synthesized at interactive rates in response to user motion. 17

3.2 **Approach overview.** At loading time, we process the input equirectangular image to recover depth D_s, Atlanta structure A_s, occupancy map O_m and latent scene representation L_s (ADM section 3.4.1). When moving from the source position, the generation of the new translated views is done by a *soft* z-buffer and a gated neural network, dubbed *Gated View Synth network* - GVS (section 3.4.2). Free viewpoint images can then be generated by extracting perspective views from the translated panoramas taking into account rotations. Supervised training of the GVS network combines visual and perceptual losses with novel indoor-specific losses (section 3.4.3). 20

3.3 **Atlanta depth module (ADM).** ADM is an end-to-end network that returns scene depth D_s, latent representation L_s, Atlanta-world 3D room shape, and floor occupancy map. Here we illustrate the two main cascading blocks: the depth estimation block (DEB) (a) and the layout estimation block (LEB) (b). DEB recovers from the input image the depth and its latent representation, while LEB recovers the layout from the predicted depth. 25

3.4 We present qualitative performance and comparison vs. ground truth and PanoSynthVR [6] on the Structured3D dataset [52]. The average movement for each scene is about $50cm$ distributed on x, y, z axis. . . . 35

3.5 We present our qualitative performance on scenes with structural occlusions. The average movement for each new pose is about $60\ cm$ distributed on x, y, z axis. 38

3.6 We present our qualitative performance on scene acquired by non-professional users. Input resolution here is 6720×3360. The average movement for each scene is about $40\ cm$. 38
3.7 **Comparison to NeRF.** we show a comparison to OmniNeRF [18], using data released by the authors. As the details clearly show, our solution provides better accuracy in many parts of the scene. 41
3.8 **Client-server system architecture.** A thin WebXR client performs image display and head tracking, offloading translated panoramic image computation to a rendering server. 43
3.9 **Forward pipeline.** At loading time, we process the input equirectangular image to recover depth D_s, Atlanta structure A_s, occupancy map O_m, and latent scene representation L_s. When moving from the source position, the generation of the new views is done by a *soft* z-buffer and a gated neural network, dubbed *Gated View Synth network* - GVS. The GVS network is trained in a supervised way combining visual and perceptual losses with novel indoor-specific losses. 44
3.10 **Structure and synthesis.** We show some examples with the predicted floor occupancy map and novel poses generated inside it 45
3.11 **Scenes with significant occlusions from architectural structure.** We present additional qualitative performance on scenes with structural occlusions. 46
3.12 **Visual synthesis results.** Additional visual comparison with PanoSynthVR [6]. 49
3.13 **Failure case.** View synthesis performance strongly depends on depth accuracy and visible feature projection. 49

4.1 **Overview**. Taking as input a single panoramic image, a data-driven architecture synthesizes a comprehensive coverage of the scene's portion visible to both eyes during head rotation, encoding the views in the form of panoramic slices. The slices are then combined into an omni-directional stereo representation composed of two multiple-center-of-projection (MCOP) images, tuned for the left and right eye. A lightweight WebXR viewer presents the suitable portions of these images on an HMD, responding to rotational head motions and delivering both stereo and motion parallax. 52

4.2 **Viewing geometry.** (a): we consider that the two eyes are positioned along a circle whose center is at the center of rotation of the head and their central gaze direction is aligned with the head rotation; their position is uniquely determined by the head radius r, the inter-pupillary distance IPD, and by the angular position of the head θ. (b): during head rotation, at a given position, there is an angular slice of size ω that contains both the right and the left gaze direction; the slice's angular size can be solely determined by r and the IPD. (c): to cover all potential gaze positions and directions, we compute all angular slices placed at closely spaced positions on the circle. 58

4.3 **Panoramic slice** One of the slices produced by the network, placed at its position within the equirectangular image. The red line shows the area sampled by the right eye, while the green line shows the area sampled by the left eye (see Figure 4.2b). In the background, we see the original panorama from the central viewpoint (note the large shift due to parallax effects). 62

4.4 **Novel view synthesis.** Given the source image I_s and its predicted depth D_s, a new, sliced viewport $\widetilde{I_{t,\theta}}$ is rendered from a new viewpoint translated by $T_{\theta,r}$, according to a given direction θ and an offset r. The field-of-view ω, designed to cover both eyes' viewport (section 4.7), is assumed constant. To generate an $I_{t,\theta}$ slice, we exploit the gated architecture already exploited for depth but adapted to have greater accuracy on the 3 RGB channels of the spherical section S. Legend: *in,out* channels; *k* convolution kernel; *s* stride; *u* upsample; *d* dilation. 62

4.5 **Reprojected vs. synthesized image** Two examples from different scenes. On the left, we see a detail of a reprojected image slice, where disoccluded areas are apparent. On the right, we see the output of the view synthesis network. 63

4.6 **Depth estimation.** Two examples of depth prediction on PNVS [17] dataset scenes and an example of depth prediction from a user-acquired panoramic scene taken with a Ricoh Theta 360° camera. 68

4.7 **Ominidirectional stereo panoramas.** Three representative scenes from the PNVS [17] dataset (testing split). Top row: source panorama. Middle row: automatically generated MCOP panorama for the left eye; Bottom row: automatically generated MCOP panorama for the right eye. The vertical alignment clearly shows the parallax effects. 69

4.8 **The WebXR viewer.** The user on the left wears a Pico4 HMD. The images to the right present the left and right images, as rendered by our WebXR viewer running on the PicoBrowser. The source image is a single-shot monoscopic 360° capture of a real environment, transformed to omnidirectional stereo by our framework. 73
4.9 **Ominidirectional stereo panoramas.** Example from real-world capture. Central: the source panorama captured with a Ricoh Theta. Left/Right: automatically generated MCOP panorama for the left and right eyes. . 73
4.10 **Comparison of omnidirectional stereo approximation with ground truth.** The top portion of the two images shows the perspective generated using the multiple-center-of-projection image for the left and the right eye, while the bottom portion shows the ground truth image generated with a center of projection placed at the eye position. As we can see, the perspective is indistinguishable at the center but slowly degrades in the periphery. 74

5.1 Given a 360 panoramic photo of a cluttered indoor scene, our end-to-end approach automatically returns a photorealistic view and depth of same scene emptied of furniture and clutter. Both visual appearance and depth, estimated at interactive speed, are highly suitable for compelling and immersive XR applications, such as (re-)furnishing or planning of interior spaces. 78
5.2 **Model architecture.** We process the input equirectangular image to identify the cluttered area in the scene, exploiting a light-weight network (purple blocks - section 5.4.1). The clutter mask and the input image are passed to the *empty scene synthesis network* (section 5.4.2), including a gated encoder (red blocks), a dilation bottleneck (yellow blocks) and a gated decoder (blue blocks), whose last layer is split in 2 layers: one for the photorealistic equirectangular representation of the emptied scene and one for its depth. The scene synthesis network is trained end-to-end through the methods and losses described in section 5.4.3. 81
5.3 Examples of inference of color and depth of the empty room from a single-shot panorama 92
5.4 Qualitative performance and comparison vs. ground truth and other approaches on the Structured3D dataset [52]. We compare to PanoDR [32], which has the best panoramic performances among the available methods. We additionally show our output depth paired with our visual results (Figure 5.4d). 93

5.5 We present qualitative performance on data for which no ground truth or training set was available. Here, we show cases from the large scale real-world dataset Matterport3D [41] and from typical user-acquired scenes, where captured images are not perfectly aligned and the photographer is visible. 94
5.6 **Predicted depth and its point cloud.** Example of 3D point cloud generated from the predicted depth. 98
5.7 **Limiting cases.** Due to the particular lighting condition our network returns a blurred model. 100

List of Tables

3.1 **Depth-layout estimation computational performance.** We show our computational performance compared to other specific state-of-the-art works for a 512×1024 image. 36

3.2 **View-synthesis computational performance.** We show our computational performance compared to other deep-learning approaches for view synthesis for a 512×1024 image. 36

3.3 **Depth and layout performance.** We show our quantitative performance compared to other representative state-of-the-art works. 39

3.4 **GVS quantitative performance.** We show GVS quantitative performance compared to other state-of-the-art works. 39

3.5 **Ablation stats.** We show the effect of several key choices of our approach. In bold is the adopted configuration. GAF: gravity aligned features-based depth estimation; MW: standard Manhattan World layout estimation; ATL: Atlanta transform structure estimation; GPS: geometric perceptual and style loss. 42

4.1 **Computational performance.** We show the computational performance and latency time of our gated architecture for different tasks. In bold modes are the current architecture choices. 70

4.2 **Quantitative performance comparison on depth reconstruction.** Our results are compared to other state-of-the-art works. 71

4.3 **View synthesis performance.** We show the quantitative performance of view synthesis compared to other state-of-the-art methods, all of which operate at a minimum resolution of 1024×512 image size (i.e., 512×1024 tensor size). The last line shows our sliced solution compared with our baseline trained to reconstruct the whole image. 72

5.1 **Computational performance.** We show our computational performance compared to other state-of-the-art works on a single NVIDIA RTX 2080Ti GPU. 94

5.2 **Computational scalability.** We show our computational performance and latency time for different input resolution. Our results demonstrate how we diminish images with a very low latency even when resolution increase. 95
5.3 **Quantitative performance.** We show our quantitative performance compared to other state-of-the-art works. 96
5.4 **Depth prediction performance.** We show our quantitative performance compared to other state-of-the-art works. 97
5.5 **Ablation facts.** We show the effect of several key choices of our approach. In bold the adopted configuration. PWG: pixel-wise geometry loss; HOG: high-order geometric loss; UI: user intervention; GAN: adversarial-loss; LPIPS,SSIM and δ_1 metrics described in section 5.5.3. 99

1 Introduction

In recent years, significant research attention has been directed toward the automatic 3D reconstruction and modeling of indoor scenes, establishing it as an emerging and well-defined sub-field within 3D reconstruction [1]. The primary goal is to convert input data from real-world interior environments into compact structured models that encapsulate geometric, structural, and visual abstractions [8]. Our focus in this study lies in extracting information from panoramic images. These images offer rapid and comprehensive coverage from a single shot and are compatible with a wide range of professional and consumer capture devices, ensuring efficient and cost-effective data acquisition. Furthermore, panoramic images have become a key component for creating immersive content directly from real scenes and for supporting a range of Virtual Reality (VR) applications [3].

In our work, we explored such modern trends, focusing in several assumed challenging for research: deriving 3D models from a single spherical image of an indoor scene and exploring such models in a photorealistic and immersive way. This chapter outlines the scientific motivation behind this work, provides a brief summary of research achievements, and presents the overall organization of this thesis.

1.1 Background and motivation

The automated reconstruction of 3D models from acquired data, such as images or geometric measurements, has been a central focus in computer graphics and computer vision for several decades. The growth of this field can be attributed to the convergence of scientific, technological, and market advancements. These developments align with the increasing accessibility and affordability of high-quality visual and 3D sensors, which are now widely available. Within this context, automatic reconstruction of indoor environments has garnered significant attention [1]. Data input can be sourced from various sensors. Visual input, such as photographic images, has

garnered significant interest due to its widespread availability, ease of capture, and affordability [9]. However, a single perspective image offers a limited view, and capturing multiple images introduces complexities related to registration. Consequently, 360-degree capture has become an attractive solution in recent years. It provides rapid and comprehensive coverage from a single image and is well-supported by a diverse range of professional and consumer capture devices, ensuring efficient and cost-effective data acquisition. Spherical cameras, also known as $360°$, *panoramic*, or *omnidirectional*, or *surround-view* cameras, provide cost-effective and efficient solutions for rapidly capturing in a single shot the full context around the viewer of an entire environment [10]. A single panoramic image encompasses the complete scene visible from a specific viewpoint within a $360°$ field of view at a given instant. When experienced through a Head-Mounted Display (HMD), users dynamically explore this image by directing their attention to the desired content through head movements, leading to Virtual/Augmented/Extended Reality (VR/AR/XR) experiences with a natural interface and good degree of immersion [11]. For these reasons, omnidirectional imagery is increasingly recognized as a foundational element for generating immersive content from real-world scenes and for supporting a variety of VR/AR/XR applications [3]. Notably, $360°$ virtual tours have gained widespread popularity in the real estate sector [12]. Furthermore, omnidirectional images are easily shareable across various devices and platforms, making them highly versatile and accessible. Since they can be seamlessly integrated into websites, VR/AR/XR applications, or mobile platforms, they enable a broad audience to engage with indoor environments irrespective of their location or their available equipment [10]. Serving as representations of the user's surroundings, panoramic images also promise to be one of the essential building blocks for the construction of the shared physical and digital realities envisioned by the Metaverse concept [13]. 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 location from which the panorama was taken leads to the loss of 3D cues, which are very important to provide a sense of presence [6]. The fact that panoramas appear flat is a particularly strong limitation in indoor environments, given the relatively short distance from the viewer to the architectural surfaces and the objects.

Indeed, to fully support immersion, a system must be able to generate, in real time, images that respond not only to changes of orientations, but also to changes of the viewpoint position. While many solutions have been proposed for multi-view capture setups, performing view synthesis from single-shot panoramas is of primary importance, due to the convenience and diffusion of sparse capturing through

monocular 360 cameras [6]. To achieve view synthesis, it is necessary to either explicitly or implicitly estimate the geometric model of the imaged environment. This, at least, enables occlusion-aware reprojection and the synthesis of disoccluded content, as well as avoid artifact and improve the sense of presence.

To this end, automatically modeling a 3D immersive indoor environment from real-world capture involves several challenging research tasks, where depth and 3D layout estimation from a single image are among the most important. Depth information enhances the input visual representation with per-pixel data representing the distance of each visible pixel from the viewer, however not only the geometric representation of the scene as observed from the point of view is sufficient to model an indoor scene, but a more abstract model of the permanent architectural structure, also known as 3D layout, is required, usually cleared of furniture and other objects [14]. Recovering such information is, however, complex because of the inherent characteristics of interior rooms, where furniture and other interior elements mask large areas of the structures of interest, and concave room shapes generate a large amount of self-occlusion. Therefore, reconstruction requires information from a very large context and must exploit very specific geometric priors for structure reconstruction [1].

In recent years, solutions based on deep-learning have emerged as a very efficient way to deal with these problems [15]. Because of the ability of these techniques to discover hidden relationships from large collections of data, many priors usually imposed by analytical and heuristic approaches based on geometric reasoning can be relaxed. Building on the aspects mentioned above, our research aimed to advance the state of the art in both automatic indoor modeling and novel view-synthesis from a single panoramic image as input. To overcome the inevitable problems and ambiguities due to such challenging input, we sought new approaches that exploit priors typical of man-made structures and the potential of the latest data-driven techniques.

1.2 Objectives

In our research project we studied the state-of-the-art, relative to topics mentioned above, and we identified specific objectives:

1. **Recovering a 3D indoor model for novel view-synthesis.**
   Current 360° cameras offering viable low-cost and energy-efficient solutions

for full-context single-shot capture are increasingly popular in many application fields [10]. Since the captured 360° content, also known as *panoramic*, *spherical*, or *omnidirectional* imagery, covers the entire sphere around the viewer, even a single shot cannot be statically experienced at once, making it fundamentally different, more immersive and more dynamic, than traditional 2D imagery [16]. However, the reduction in degrees of freedom to just the rotation around the center of the panorama, leads to constraints and artifacts [6], especially since only one or two shots per room are available in a typical virtual tour [17]. Moreover, binocular stereo and motion parallax, which are important aspects of immersion in VR, are totally missing. To fully support immersion, a system must thus also respond to viewpoint translation. While many solutions have been proposed for multiview capture setups (e.g., [4, 5]), performing view synthesis from single-shot panoramas is of primary importance, due to the convenience and diffusion of sparse capturing through monocular 360° cameras [6]. View synthesis requires the explicit or implicit estimation of the geometric shape of the imaged environment (i.e., *3D indoor model*), in order to perform occlusion-aware reprojection and to synthesize the disoccluded content. Current state-of-the-art approaches (e.g., [18, 6]) focus on extending to single-shot panoramas the general data-driven view synthesis approaches designed for perspective views of objects and environments, such as Multiplanar images (MPI) [19] or Neural Radiance Fields (NeRF) [20]. The mixing of large untextured surfaces, clutter, and non-cooperative materials in interior environments poses, however, important challenges to generic solutions [1]. In this context, it has been demonstrated that the knowledge of additional information, such as the position location of the room corners and edges, significantly improves the realism of the synthesis [17]. However, recovering the indoor layout directly from the input image is extremely challenging. Even the latest dedicated methods [21, 22, 23] still heavily rely on approximations and expensive heuristic post-processing [24], which significantly limit overall performance. As a result, their use for VR applications necessitating interactive-rate image generation is inhibited.

2. **Immersive exploration of indoor stereoscopic environments.**

   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 location from which the panorama was taken leads to the loss of binocular stereo, which is very important to provide a sense of presence [6]. The fact that panoramas appear flat is a particularly

strong limitation in indoor environments, given the relatively short distance from the viewer to the architectural surfaces and the objects. In order to provide stereo cues for full 360-degree rotations, views from a continuous set of shifted viewpoints must be available to the renderer (*stereoscopic 3D model*). Omnidirectional stereo techniques [25, 26] are employed for that purpose but require the creation of stereo panoramas using cameras moving on a circular path [27, 26, 25] or multiple synchronized 360 cameras [3]. These acquisition approaches, however, reduce the possibility of quickly capturing, experiencing, and sharing a 360° scene using consumer hardware. In particular, while a number of low-cost cameras are widely available for monocular 360° capture (e.g., GoPro, Ricoh Theta, LadyBug, or Insta360), also due to the booming "action-camera" market, stereo 360° solutions (e.g., Vuze+) are more costly and limited, and also typically offer only a low number (i.e., six to eight) of different point of views, leading to stereo and stitching artifacts. Moreover, while rotating camera solutions provide more viewpoints, they do not share the same simplicity and flexibility of single-shot instantaneous capture. For this reason, research has concentrated on view synthesis methods that generate stereo contents from a single 360° panorama. However, current methods either require complicated representations or are too heavy to run directly on HMDs and interactive rates.

3. **Automatic-assisted editing of immersive indoor models.**
   A pure exploration of existing environments through the original spherical photos, is, however very limiting. Prominent examples of additional needs include the emptying of rooms before their presentation to virtual visitors (if only for privacy reasons), or the refurnishing or redecorating of interior spaces [28]. In this context, fast and effective Diminished Reality (DR) techniques, which conceal real-life parts from the view field, are paramount to remove the furniture and other clutter that masks the architectural structure. In particular, DR features are essential to allow users to immediately compare the furnished and unfurnished scene, and to support Augmented Reality (AR) applications in placing objects in the empty scene [29, 30]. Making these features available on novel environments with minimum latency, ideally in real-time, would, in addition, enable their usage in remote collaboration contexts, without the need for prior modeling [31]. While a variety of object erasing and image inpainting solutions have been presented in the literature (see section 5.3), DR for interior environments must generate images of empty indoor spaces that not only have a realistic appearance, but respect the context in stricter ways, in particular

by inferring a plausible organization of the permanent architectural structure that bounds the room's interior [32]. Data-driven solutions, that learn hidden relations from examples, are emerging as viable approaches for this class of problems. However, state-of-the-art methods for image inpainting are mostly focused on photorealism [33, 34], and additional information about the scene is exploited only from the semantic point-of-view [35, 36, 32]. Current pipelines make limited use of the structure of the observed scene, and reconstruction accuracy is achieved at the price of high computational complexity or increased user intervention, using, for example, recursive networks [37], multi-branch architectures [34], and manual definition of specific parts of the original image to be removed [35].

1.3 Achievements

Starting from the targeted objectives, we researched novel techniques, network structure, loss functions, and training methods. Our research produced a series of scientific results (listed at section 6.3), which advanced the state of the art in many aspect. We select among them the main achievements (see section 1.2):

- **A novel methodology for 3D scene synthesis of Atlanta-world interiors from a single omnidirectional image** (section 3). A new data-driven approach, targeting the objective 1, for extracting geometric and structural information from a single spherical panorama of an interior scene, and for using this information to render the scene from novel points of view, enhancing 3D immersion in VR applications. The approach copes with the inherent ambiguities of single-image geometry estimation and novel view synthesis by focusing on the very common case of *Atlanta-world* interiors, bounded by horizontal floors and ceilings and vertical walls. Based on this prior, we introduce a novel end-to-end deep learning approach to jointly estimate the depth and the underlying room structure of the scene. The prior guides the design of the network and of novel domain-specific loss functions, shifting the major computational load on a training phase that exploits available large-scale synthetic panoramic imagery. An extremely lightweight network uses geometric and structural information to infer novel panoramic views from translated positions at interactive rates, from which perspective views matching head rotations are produced and up-sampled to the display size. As a result, our method automatically produces new poses around the original camera at interactive rates, within a working

area suitable for producing depth cues for VR applications, especially when using head-mounted displays connected to graphics servers. The extracted floor plan and 3D wall structure can also be used to support room exploration. The experimental results demonstrate that our method provides low-latency performance and improves over current state-of-the-art solutions in prediction accuracy on available commonly used indoor panoramic benchmarks. This work has been published as a TVCG journal paper [38] and presented at the IEEE ISMAR conference 2023. The candidate was responsible for all the aspects of the work, from conceptualization, methodology, software, to validation, writing and review of the related papers.

- **A novel approach for deep synthesis and exploration of omnidirectional stereoscopic environments from a monoscopic panoramic image** (section 4). An innovative approach, targeting the objective 2, to automatically generate and explore immersive stereoscopic indoor environments derived from a single monoscopic panoramic image in an equirectangular format. Once per 360° shot, we estimate the per-pixel depth using a gated deep network architecture. Subsequently, we synthesize a collection of panoramic slices through reprojection and view-synthesis employing deep learning. These slices are distributed around the central viewpoint, with each slice's projection center placed on the circular path covered by the eyes during a head rotation. Furthermore, each slice encompasses an angular extent sufficient to accommodate the potential gaze directions of both the left and right eye and to provide context for reconstruction. For fast display, a stereoscopic multiple-center-of-projection stereo pair in equirectangular format is composed by suitably blending the precomputed slices. At run-time, the pair is loaded in a lightweight WebXR viewer that responds to head rotations, offering both motion and stereo cues. The approach combines and extends state-of-the-art data-driven techniques, incorporating several innovations. Notably, a gated architecture is introduced for panoramic monocular depth estimation. Leveraging the predicted depth, the same gated architecture is then applied to the re-projection of visible pixels, facilitating the inpainting of occluded and disoccluded regions by incorporating a mixed Generative Adversarial Network (GAN). The resulting system works on a variety of available VR headsets and can serve as a base component for a variety of immersive applications. We demonstrate our technology on several indoor scenes from publicly available data. This work has been accepted for publishing as a Computers & Graphics journal paper [39] and awarded with honorable mention at the ACM WEB3D conference 2023. The candidate The

candidate was responsible for all the aspects of the work, from conceptualization, methodology, software, to validation, writing and review of the related papers.

- **An innovative end-to-end technique for instant automatic emptying of panoramic indoor scenes** (section 5). A new data-driven approach, targeting the objective 3, that from an input 360° image of a furnished indoor space, automatically returns an omnidirectional photorealistic view and architecturally plausible depth of the same scene emptied of all clutter. Contrary to previous data-driven inpainting methods that remove single user-defined objects based on their semantics, our approach is holistically applied to the entire scene, and is capable to separate the clutter from the architectural structure in a single step. By exploiting peculiar geometric features of the indoor environment, we shift the major computational load on the training phase and having an extremely lightweight network at prediction time. This work has been published as a TVCG journal paper [40] and presented at the IEEE ISMAR conference 2022. The candidate was responsible for all the aspects of the work, from conceptualization, methodology, software, to validation, writing and review of the related papers.

1.4 Organization

We organize this thesis as following:

- **Chapter 1** (this chapter) introduces the topic and motivation for this dissertation, with a summary of objectives and achievements.

- **Chapter 2** presents general requirements, work hypotheses and means of verification.

- **Chapter 3** describes a novel deep learning approach that extracts geometric and structural information from a single panorama in order to quickly synthesize plausible panoramic images from close-by viewpoints within a workspace suitable for VR applications.

- **Chapter 4** describes an innovative approach to automatically generate and explore immersive stereoscopic indoor environments derived from a single monoscopic panoramic image in an equirectangular format.

- **Chapter 5** describes novel techniques to exploit a photorealistic and geometric consistent indoor model recovered from a single panoramic image for immersive virtual staging applications.

- **Chapter 6** provides a conclusion and short summary of the achievements, a critical discussion of the results obtained and of how they advance the state-of-the-art, as well as discussion on future lines of work.

---

2 General requirements, work hypotheses, and means of verification

In this chapter we provide a summary of general definitions related to the research works, the hypothesis leading the thesis and the means adopted for the verification.

2.1 Basic definitions

The following is a list of important definitions related to the main concepts that appear in the thesis:

- **Omnidirectional images**. An image with a $360°$ horizontal viewing angle and $180°$ vertical viewing angle. Also called in many works (with a little abuse of the term) *panoramic image*, can be obtained by many means, including stitching of a sequence of photos captured with a mobile phone or by a professional device [41], or, in one-shot, by modern commodity spherical cameras, which have become very widespread and increasingly popular in many application fields [10]. Since the sphere is not isomorphic to a plane, representing the capture as an image typically involves a mapping transformation. While some cameras provide access to the original unstitched images, that provide the highest resolution capture, the most common approach, that has become a de-facto standard in indoor capture and processing, is to extract from the device an equirectangular projection sampled into a regular rectangular 2D grid [42], obtaining what is often called a *full panoramic image*.
- **Structured indoor model**. It is an abstract structure defined by a graph of rooms bounded by walls, floor, and ceiling, as well as connected by doors/passages and containing objects, such as furniture and wall-mounted items. The structured model thus combines a topological part (the connection graph), a geometric part (the shape of the various components) and a visual part (the

appearance model of the different nodes). In our work, since, we are interested in modeling the environment surrounding an observer, the structured indoor model is related to one hypothetical environment or room, the topological and geometrical part of which is the indoor layout (described below).

- **Indoor layout**. An abstract representation of a 3D indoor space (usually a room), as a joining of walls, ceilings and floors planes, representing the permanent architectural structure, usually cleared of furniture and other objects [14].

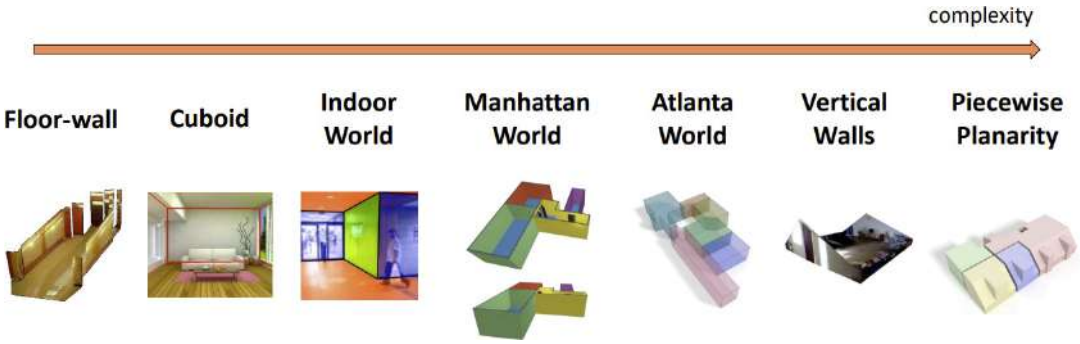


Fig. 2.1.: **Architectural priors.** A list of architectural priors used in 3D reconstruction, in order of complexity (Pintore et al., CVPR 2023 [43]).

- **Indoor priors**. Architectural indoor priors makes it possible to reduce the solution space, making reconstruction more tractable. Figure 2.1 summarizes, in order of complexity, the most commonly geometric priors used in indoor for surface reconstruction. They include *Floor-Wall* (FW) [44], composed by a single flat floor and straight vertical walls; *cuboid* (CB) [45], being a single room of cuboid shape; *Indoor World Model* (IWM) [46], with a single horizontal floor, a single horizontal ceiling, and vertical walls that meet at right angles; *Manhattan World* (MW) [47], an IWM without the restriction of a single floor and celing; *Atlanta World* (AW) [48], similar to MW, without the restriction of walls connecting at right angles; *Vertical Walls*, and Atlanta-World model with possibly sloped ceilings and floors [48], and piecewise planarity, that simply bounds the interior with large planar surfaces [49].
- **Pixel-wise information**. Given a map of pixels, (e.g., an equirectangular image), information is pixel-wise is that assigned to each pixel, in addition to color, such as a depth value. in the case of equirectangular images, each position in the image represents a view direction (i.e., longitude and latitude), as well as the associated depth is the euclidean distance of the point form the camera origin. Other pixel-wise information, if associated to the image, can be normals or semantic segmentation.

- **Novel view-synthesis**. It is a specific task in computer graphics and computer vision. It involves generating images of a specific subject or scene from a particular viewpoint, even when the only available information consists of pictures taken from different viewpoints. Essentially, it allows you to create new views of an object or scene that were not originally captured, based on existing images and their associated camera poses.
- **Structural consistency**. It refers to the coherent and regular arrangement of elements within an indoor environments. In our studies we expect that the generated novel views keep such consistency according to indoor priors. This is crucial to avoid artifact and unpleasant experiences during immersive exploration or interaction.
- **Immersive exploration**. Immersive exploration in the context of virtual reality (VR) refers to the captivating experience of fully engaging with a three-dimensional (3D) digital environment. Immersive exploration involves creating a simulated environment that completely surrounds and engulfs the user's senses and it typically requires specialized hardware, such as a head-mounted display (HMD).
- **Diminished reality**. Diminished Reality (DR) is a set of methodologies that allow users to visually remove, hide, or see through real-life objects in a perceived environment. Unlike augmented reality (AR) and mixed reality (MR), which superimpose virtual objects onto the real world to enhance reality, DR focuses on diminishing the perception of reality. In our work, removing clutter from one indoor environment means generating another realistic indoor environment for, for example, virtual staging applications. In that sense any subsequent augmented or extended reality applications, where new synthetic objects are placed in the environment, is outside the scope of our modeling.

2.2 Research questions

Building on the objectives mentioned in section 1.2, our research aimed to advance the state of the art in both automatic indoor modeling and immersive exploration from a single panoramic image as input. Below we summarize the main specific research questions to address such objectives:

1. *How to associate a pixel-wise depth to the input RGB image?* Generating new views at positions other than the original, necessarily requires knowledge

of the spatial position of the points seen by the observer, in order to apply the proper translation (see section 3.4). In other words, each RGB pixel in the equirectangular image must be associated with a depth, such that the 2D pixel can be transformed, by spherical coordinates to a 3D colored point, and the entire point cloud can be rendered from a new equirectangular view. Estimating the depth an indoor panoramic image is a very trendy research topic (see section 3), and for its nature, an open research problem.

2. *How to recover an occlusion-free 3D layout of the room from a single pose?* The mere translation of the point cloud derived with depth not contain all the information necessary to define a full translated viewport. In fact, some parts of the scene will be occluded or disoccluded. It has been demonstrated that the knowledge of additional information, such as the position location of the room corners and edges (i.e., room layout), especially those hidden in the original point-of-view, significantly improves the realism of the synthesis [17]. However, recovering the indoor layout directly from a single input image is extremely challenging [14]. To this end, specific techniques to recover the 3D layout from a single equirectangular image need to be researched.

3. *How to generate both photorealistic and structural consistent novel views?* In order to generate new views required for immersive experience, scene generation and impainting techniques are employed (see section 4). Such photorealistic inpainting techniques work very well in cases of small disocclusions due to the natural movements of the observer in the scene (see section 4.3), but may break down in the case of large disocclusions. This is a typical case during scene editing if, for example, clutter has to be removed from the scene for virtual staging applications (see section 5.3). In this situation the problem of how to complete the missing parts emerges, not only from the photorealistic point of view, but also in terms of geometric-structural consistency.

2.3 Hypotheses supporting the prospected solutions

The answers to the research questions, which are explained and detailed in the following chapters, are based on the following hypotheses:

1. *How to associate a pixel-wise depth to the input RGB image?* Starting from the fact that gravity plays an important role in the design and construction of man-made

indoor scenes [15], we research novel depth estimation techniques that exploit gravity-aligned features (GAFs) to take into account the fact that world-space vertical and horizontal features have different characteristics in man-made environments (see section 2.1). Such design starts from the assumption that capture of the scene through an equirectangular image is aligned to the gravity vector (i.e., camera is placed on an horizontal-ground plane).

2. *How to recover an occlusion-free 3D layout of the room from a single pose?* Deriving a scene layout has a different complexity than estimating single depth, since it does not simply assign a geometric value to each visible pixel, but must extrapolate large portions of the invisible structure, which may be occluded not only by objects but by the structure itself. In our work, we assumed that an indoor facility can be reasonably represented as an Atlanta World model, that is, we assume that our rooms have flat floors and ceilings. This assumption is less restrictive than the classic Manhattan World, as it admits free wall shapes and corners (see section 2.1).

3. *How to generate both photorealistic and structural consistent novel views?* While some structural information, such as the location of corners and edges disoccluded by translation, can be provided as input to an eventual neural network for the synthesis of a new view, conditioning the generation of the final scene so that it is structurally consistent is not straightforward (see section 2.1). In our work we have confronted this problem by assuming that different indoor priors can be represented as specific losses, in the same way that a perceptual style is imposed [50].

2.4 Means of verification

The hypotheses above have been researched and verified, and have become the basis for the technical contributions of peer-reviewed publications published by the author during the research project (see section 1.3). Furthermore, full pipelines have been implemented of top of the publications, the performance of which exceeds the relative state of the art in many aspects. Operationally, the validation adopts the following approaches:

1. We implement our deep learning algorithms on PyTorch [51], a flexible and powerful framework for creating and deploying deep learning models, widely used by the works related to ours. The use of this framework provides a baseline that facilitates comparison with other work, and enables the use of large-scale

datasets for training, validation and testing by leveraging multi-GPU distributed computing.

2. We adopt commonly used datasets for benchmarking the current state of the art, providing ground truth for depth, layout, and alternative views from different locations of common indoor environments. Specifically, we exploit Structured3D[52], a large-scale synthetic database of indoor scenes comprising 21,000 photorealistic scenes, which provides ground truth depth and layout information for each panoramic image, and PNVS [17], a subset of Structured3D scenes providing, for each source panoramic image, three views translated by 0.2-0.3m along random directions, and three views translated by 1.0-2.0m.

3 Recovering a 3D indoor model for novel view-synthesis

In this chapter, we present a new data-driven approach which addresses the specific problems of item 1. Such an approach aims to extract geometric and structural information from a single spherical panorama of an interior scene, and to use this information to render the scene from novel points of view, enhancing 3D immersion in VR applications. The approach copes with the inherent ambiguities of single-image geometry estimation and novel view synthesis by focusing on the very common case of *Atlanta-world* interiors, bounded by horizontal floors and ceilings and vertical walls. Based on this prior, we introduce a novel end-to-end deep learning approach to jointly estimate the depth and the underlying room structure of the scene. The prior guides the design of the network and of novel domain-specific loss functions, shifting the major computational load on a training phase that exploits available large-scale synthetic panoramic imagery. An extremely lightweight network uses geometric and structural information to infer novel panoramic views from translated positions at interactive rates, from which perspective views matching head rotations are produced and upsampled to the display size. As a result, our method automatically produces new poses around the original camera at interactive rates, within a working area suitable for producing depth cues for VR applications, especially when using head-mounted displays connected to graphics servers. The extracted floor plan and 3D wall structure can also be used to support room exploration. The experimental results demonstrate that our method provides low-latency performance and improves over current state-of-the-art solutions in prediction accuracy on available commonly used indoor panoramic benchmarks.

3.1 Introduction

A single-shot 360° image, containing the entire scene around the viewer, is not consumed at once, but inherently requires a more dynamic exploration with respect

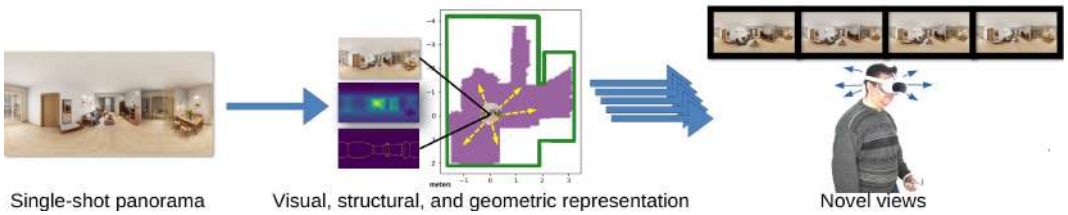


Fig. 3.1.: Given a single 360° panorama of an indoor scene, we compute an enriched geometric and structural representation, from which novel panoramas from other close-by viewpoints can be synthesized at interactive rates in response to user motion.

to traditional 2D imagery. When it is presented through a Head-Mounted Display (HMD), the viewer is encouraged to actively focus on the desired content via natural head movements, leading to an intuitive VR interface [11]. For this reason, 360° image viewing is becoming one of the main exploration modes of real-world scenes in VR [3] and has widespread use in indoor navigation [53].

The reduction in degrees of freedom to just the rotation around the center of the panorama, leads, however, to constraints and artifacts [6], especially since only one or two shots per room are available in a typical virtual tour [17]. Moreover, binocular stereo and motion parallax, which are important aspects of immersion in VR, are totally missing. To fully support immersion, a system must thus also respond to viewpoint translation. While many solutions have been proposed for multiview capture setups (e.g., [4, 5]), performing view synthesis from single-shot panoramas is of primary importance, due to the convenience and diffusion of sparse capturing through monocular 360° cameras [6].

View synthesis requires the explicit or implicit estimation of the geometric shape of the imaged environment, in order to perform occlusion-aware reprojection and to synthesize the disoccluded content. Current state-of-the-art approaches (e.g., [18, 6]) focus on extending to single-shot panoramas the general data-driven view synthesis approaches designed for perspective views of objects and environments, such as Multi-planar images (MPI) [19] or Neural Radiance Fields (NeRF) [20] (section 3.3). The mixing of large untextured surfaces, clutter, and non-cooperative materials in interior environments poses, however, important challenges to generic solutions [1]. In this context, it has been demonstrated that the knowledge of additional information, such as the position location of the room corners and edges, significantly improves the realism of the synthesis [17]. However, recovering the indoor layout directly from the input image is extremely challenging. Even the latest dedicated methods [21, 22, 23] still heavily rely on approximations and expensive

heuristic post-processing [24], which significantly limit overall performance. As a result, their use for VR applications necessitating interactive-rate image generation is inhibited.

In our work, we research a new end-to-end data-driven solution that, from a single 360° indoor panorama, assumed captured with approximate gravity alignment, produces with low latency a newly translated pose from which new perspective images can be extracted that respond to both position and orientation changes. While some HMD solutions strive to fully run on the embedded platform, an alternative design is to compute images on high-performance servers. This approach, extensively employed for high-quality gaming, is made possible by the availability of low-latency tethered or wireless connections with sufficient bandwidth to feed the displays [54]. In our approach, a thin WebXR client directly handles head rotation, while relying on server-computed images to also respond to head translations. Our main novelty is in the indoor-specific deep-learning techniques that synthesize the views. Once per scene, we enrich the original panorama with geometric and structural information, and once per frame, we exploit pre-computed information to quickly perform view synthesis.

The approach copes with the inherent ambiguities of single-image geometry estimation and novel view synthesis in indoor environments by focusing on the very common case of interiors following the *Atlanta world* model (AWM) [1], in which the environment is expected to have horizontal floor and ceiling and vertical walls. Based on this prior, we introduce a novel end-to-end network to jointly estimate the depth and the underlying room structure of the scene, thus efficiently handling occlusions and disocclusions and enabling a plausible prediction even in the case of extensively occluded structures. The prior drives the network structure, which also exploits gravity-aligned features (GAFs) to take into account the fact that world-space vertical and horizontal features have different characteristics in man-made environments. In particular, AWM makes it possible to derive the 3D layout by extruding its 2D floor projection, while GAFs perform vertical compression, exploiting the fact that vertical lines, common in indoor scenes, are not deformed in equirectangular projections. Because of these characteristics, we expect scene GAFs to be inter-related by both short-term and long-term spatial dependencies, improving the quality of depth prediction and layout prediction [15], and, therefore, visual synthesis.

Starting from the enriched panoramic representation, a lightweight network infers at interactive rates novel panoramic images from translated positions in a working

area around the original point of view suitable for VR applications. From these views, perspective images matching head rotations are produced and upsampled to display size. Moreover, the geometric and structural information recovered can also be used to support VR applications, e.g., to define walkable areas.

3.2 Contributions

Our main novel contributions are the following:

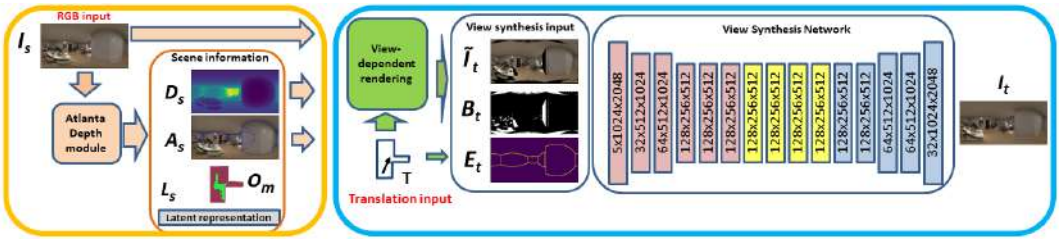
- We present a novel approach, dubbed *Atlanta Depth Module - ADM*, to jointly estimate, starting from a single equirectangular image, the scene depth, a scene latent representation, the 3D room shape and a floor occupancy map (section 3.4.1). ADM achieves state-of-the-art results on both geometric and structural reconstruction (section 3.5), and provides many advantages for VR applications. First, it is much more lightweight than current solutions for depth or layout estimation commonly adopted in this context [55, 24, 15]. Second, the recovered AWM structure is segmented into ceiling, walls, and floor, and represented in metric units, including the prediction of ceiling-floor heights. This, besides improving view synthesis, supports the creation of a *floor occupancy map*, to generate consistent trajectories inside the room without collisions.

- We introduce novel objective functions to take into account the indoor structural consistency (section 3.4.3) of the view synthesis. Such functions, based on GAF encoding [15, 56], support direct (i.e., target predicted depth loss) and latent-space losses. Latent-space losses guide a consistent structural reconstruction during training and are dual to visual losses, called *geometric perceptual* and *geometric style*, Such losses, combined with standard perceptual style transfer and adversarial losses, improve reconstructed scene quality (section 3.5), shifting much of the computational load to the training phase, and making the inference phase much lighter.

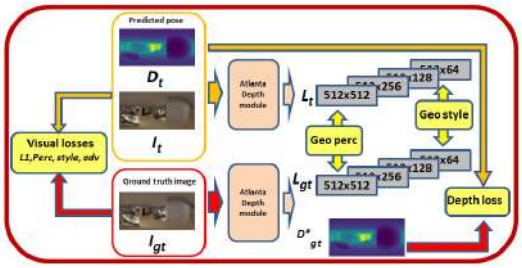
- We introduce a fully data-driven, versatile, and lightweight approach to generate novel panoramic views from a single indoor panorama. Such a deep learning approach does not need dedicated processing for each scene [18, 6], but generalizes over indoor scenes that just follow AWM (section 3.5). Once latent deep features and structural priors are applied at training time, novel

pose synthesis is obtained through a network (GVS) without deep layers or complex pipelines. In fact, GVS consists of a limited number of layers, combining gated and dilated convolutions, focused on maximizing the level of detail (section 3.4.2). As a result, we have a network with a limited and constant number of learnable parameters (section 3.5.2), even as generated image resolution varies.

Our results (section 3.5) improve over state-of-the-art approaches on common benchmarks with measurable ground truth, in terms of accuracy, quality and computational complexity. Moreover, compelling predictions are produced even on images where no ground truth is available for training, as well as on novel user-captured images.



(a) Forward pipeline



(b) Training scheme

Fig. 3.2.: **Approach overview.** At loading time, we process the input equirectangular image to recover depth D_s, Atlanta structure A_s, occupancy map O_m and latent scene representation L_s (ADM section 3.4.1). When moving from the source position, the generation of the new translated views is done by a *soft* z-buffer and a gated neural network, dubbed *Gated View Synth network* - GVS (section 3.4.2). Free viewpoint images can then be generated by extracting perspective views from the translated panoramas taking into account rotations. Supervised training of the GVS network combines visual and perceptual losses with novel indoor-specific losses (section 3.4.3).

3.3 Related Work

Effective view synthesis requires comprehensively understanding the 3D structure of a scene given an image [57]. Full coverage of this topic is outside the scope of

this paper. In the following, we focus on the most closely related approaches, with a particular focus on data-driven solutions for panoramic images.

3.3.1 Depth estimation from panoramic images

Monocular depth estimation is a classic task in computer vision. While early solutions used various combinations of feature detection, matching, and geometric reasoning, recent research is increasingly focusing on data-driven solutions that derive hidden relations from large amounts of examples [15]. Since it has been shown that directly applying perspective methods to 360° depth estimation in indoor environments produces suboptimal results [58], research has started to focus on explicitly exploiting the characteristics and wide geometric context present in omnidirectional images. A first breed of solutions concentrated on handling distortion through spherical convolution [59, 60, 61, 62, 58]. Wang et al. [63] proposed instead a two-branch network, respectively for the equirectangular and the cubemap projection, based on a distortion-aware encoder [58] and the FCRN decoder [64]. Recent solutions for panoramic depth estimation in indoor spaces [65, 15] have proposed to work directly on equirectangular images, as well as to leverage the concept of gravity-aligned features to reduce network size [24, 15]. A recent trend to mitigate panoramic distortion is to leverage perspective views sampled on panoramic images [66, 67] prior to combining depth maps using transformers. In this work, we leverage gravity-aligned features [15] to flatten image features and then process them with a lightweight network designed for interactive applications (section 3.4.1). Compared to previous works, we achieve state-of-the-art performance at a much lower computational cost (section 3.5).

3.3.2 Layout estimation from panoramic images

While depth estimation methods have shown impressive performances, they cannot produce seamless 3D boundary surfaces in case of self-occlusions, since they can only generate a single 3D position per view ray. For this reason, layout-specific approaches are being actively researched. Since man-made interiors often follow very strict rules, early pin-hole methods used geometric reasoning to match image features to simple constrained 3D models [1]. The effectiveness of geometric reasoning methods is, however, heavily dependent on the count and quality of extracted features (e.g., corners, edges, or flat patches). More and more research is thus now focusing on data-driven approaches [68]. Prominent examples are *LayoutNet* [55], which

predicts the corner probability map and boundary map directly from a panorama, and *HorizonNet* [24], which simplifies the layout as three 1D vectors. The 2D layout is then obtained by fitting Manhattan World Model (MWM) segments on the estimated corner positions. 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 [69, 48, 70, 23, 22]. These methods, however, require heavy pre-processing, such as detection of main Manhattan-world directions from vanishing lines analysis [68, 2, 71] and related image warping, or complex layout post-processing, such as Manhattan-world regularization of detected features [55, 24, 69]. LayoutNet [55], for instance, has been used to support view synthesis of individual panoramic images [17] by providing the location of corners in the image, but cannot run at interactive rates. Several methods have, thus, sought to relax the constraints of the Manhattan World model, while decreasing the computational load required by exploiting more general features of man-made structures [1, 56]. These methods, however, target the general reconstruction of the overall room shape but are not usable for the completion of photorealistic views, lacking well-defined parts and edges. In this work, we propose, instead, a new approach for fast estimation of a structured layout, where, unlike the mentioned methods, the estimation is done not from the *RGB* image but from its depth, appropriately transformed (section 3.4.1). Moreover, we apply Atlanta World projection to depth values, and projection is not done on an arbitrary plane as in other transform-based approaches [69, 1].

3.3.3 Novel view synthesis

Our solution exploits recovered depth and layout for novel view synthesis from monocular input. Most view synthesis approaches exploit, instead, multi-view input, such as NeRF [20], the methods based on depth, proxy geometry, and flow [72, 27, 73, 74], or, for 360° views, those using a layered image representation [75, 76], multi-depth panoramas [77], or layered mesh representations [5]. The method of Serrano et al. [78] extended the layered image representation approach to work with a single panoramic input image, but with only a few depth layers and extrapolation and in-painting to fill holes. Layered solutions have been also extended to be used as a target representation in an end-to-end learning pipeline. In multi-plane images (MPI) [79], each layer is a flat plane placed at a fixed depth from the capture point. The regularity of the representation makes it suitable to be the output of a convolutional neural network. Tucker and Snavely [19] introduced a method to infer

an MPI from a single perspective image. *PanoSynthVR* [6] extended this approach by exploring the use of a multi-cylinder representation to approximate a 360 view. However, the method does not exploit information specific to indoors, produces blurry images at disocclusions and severely degrades quality when the viewpoint moves too far from the origin (section 3.5). Xu et al. [17] recently extended to panoramic images the approach of SynSin [57], which uses a neural network to produce both a depth map and a feature map which can then be rendered to new perspective viewpoints. Moreover, similarly to us, they incorporate prior knowledge, in the form of screen corners, demonstrating the importance of using additional information that does not depend on the input viewpoint. Such an approach is capable of generating new, sparse views, but its performance depends on externally computed information and requires a considerable computational cost. An alternative solution is proposed by *OmniNeRF* [18], which proposes a self-supervised approach to generate novel views given a single panoramic image and its depth, with the goal to feed a NeRF [20] pipeline with multiple poses. Although it is one of the first works to adapt the NeRF concept to a panoramic image, the synthesized images feeding the training are a simple interpolated splatting of the original view, so that new views obtained at run time suffer from significant artifacts. In contrast, by introducing and exploiting important indoor priors at inference and training levels (section 3.4.3), we generate new views with greater accuracy than the methods mentioned above and with a particularly lightweight end-to-end network. Our reconstruction is also not only visually but also spatially consistent, unlike other representations [19, 4, 6].

3.4 Methods

In our approach, a thin client explores an HMD synthesized panoramic images that are adapted to position changes through server-side computation.

The extraction, server side, of structural information from a room and the generation of a translated panorama are the most complex operations and are performed through a novel deep learning architecture, whose structure is depicted in Figure 5.2.

Once per scene, we assign a viewer-independent geometry context to input source pixels, to create a room structure that can be used for various purposes and to propagate enriched input information to the new pose. Jointly with depth estimation, we infer a structured model of the underlying architectural structure through a deep network that exploits the Atlanta World prior. The network, dubbed Atlanta

Depth Module (ADM) (orange module in Figure 3.9), feeds the second step of the pipeline, that, every time the viewer position changes, generates the translated panoramic image. While both modules could be trained as a whole, in our design, for performance reasons, we pre-trained ADM separately from view synthesis.

The pre-trained ADM returns the scene depth, a latent scene representation, the room 3D shape, and, additionally, a floor occupancy map (section 3.4.1). Gravity-aligned features (GAFs) encode a panoramic image into a multi-resolution latent scene representation. GAFs are basically used in two ways. On one hand, their decoding produces a pixel-wise depth of the scene (Figure 3.9, orange); on the other hand, the GAF encoding supports specific geometric loss functions in latent space that guide the training of the view synthesis model (Figure 3.2b). Moreover, downstream network layers process the recovered depth to predict a 3D model of the room. As this 3D model is viewer-independent, it is computed once at the time of image loading, albeit with low computational cost, and does not need to be re-executed at each view synthesis. Moreover, the 3D model can be used for a variety of purposes. For instance, it makes it possible to define a walkability map, as well as to limit the legal area for new viewpoints so that there are no walls crossed and we remain inside the room.

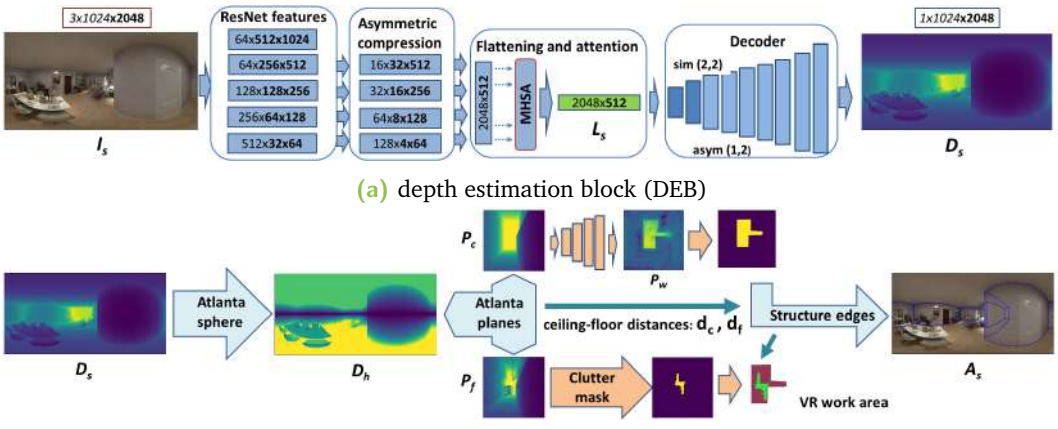
The panoramic view synthesis phase, depicted in the orange block of Figure 3.9, is performed at each viewer position change through a very light network that comprises a *soft* z-buffer block and a gated neural network, dubbed *Gated View Synth* (GVS) (section 3.4.2). The GVS network is trained in a supervised way through a combination of specific losses (section 3.4.3). In particular, in addition to standard metrics in view synthesis, we introduce novel geometric and structural metrics, that also exploit latent space GAFs.

The output of the view synthesis phase is the translated $360°$ view from which free-viewpoint images that respond to translation and rotation can be extracted. In our reference design, the view synthesis machinery is exploited in a client-server application, where a thin WebXR client runs directly on the HMD's embedded platform and communicates with the server that handles the compute-intensive tasks. Both client and server are initialized at each scene change with the initial view, which is used by the client for display and by the server to compute from visual data the augmented panoramic scene representation. The client is designed as a foveated panoramic image viewer, that maintains, in two textures, a low-res and a hi-res representation of portions of the scene's panorama in an equirectangular format centered approximately around the current lookat point. The low-res panorama

typically comprises a 180x180 degree portion (full frontal view), while the high-res panorama covers at double resolution a 90x90 degree area (HMD FOV). A fragment shader combines the two textures at each frame to produce a seamless view. At each head position change, the head transformation is sent to the server. The position is communicated to the view synthesis network for producing the translated panorama. The rotation is used to determine the current look-at point. Since view synthesis, as for all current deep learning solutions, is performed at a resolution that is lower than current HMD capabilities, we perform image upsampling of the viewing region around the look-at point using state-of-the-art deep learning super-resolution methods [80]. The low-res image and hires image, together with the associated parameters are then sent to the client for display. The supplementary material provides detailed information on the structure of the client-server application.

The novel aspects of our work reside in the methods that are employed to generate the translated panorama, which, in addition to supporting view extraction, also generate auxiliary structural and geometric information that can be exploited for other needs. In the following we will describe the main components of this block, first focusing on depth and 3D layout prediction (section 3.4.1), and then on novel view synthesis (section 3.4.2) and training methods (section 3.4.3).

3.4.1 Depth and 3D layout prediction



(a) depth estimation block (DEB)

(b) layout estimation block (LEB)

Fig. 3.3.: **Atlanta depth module (ADM).** ADM is an end-to-end network that returns scene depth D_s, latent representation L_s, Atlanta-world 3D room shape, and floor occupancy map. Here we illustrate the two main cascading blocks: the depth estimation block (DEB) (a) and the layout estimation block (LEB) (b). DEB recovers from the input image the depth and its latent representation, while LEB recovers the layout from the predicted depth.

The Atlanta Depth Module (ADM) can be subdivided into two cascading blocks: depth estimation (Figure 3.3a) and layout estimation (Figure 3.3b). The first network block, dubbed depth estimation block (DEB), receives as input a single panoramic image ${I_s}^{h \times w}$ and returns as output a depth ${D_s}^{h \times w}$ in metric units (i.e., meters), coupled with a latent representation of the scene L_s, which is also used for the specific losses described in section 3.4.3. The second block, dubbed layout estimation block (LEB), receives as input ${D_s}^{h \times w}$, and returns as output an Atlanta World representation A_s (Figure 3.4.1). The rest of this section summarizes the main aspects of DEB and LEB. We refer the reader to the supplementary material for more details.

Depth estimation block. From the input image, a cascade of five residual layers [81] creates four feature maps having different depth and spatial size (Figure 3.3a). We consider, then, both the spherical and indoor nature of the scene to further process this representation. First, we adopt circular padding along the horizon for convolutions, to overcome the longitudinal boundary discontinuity, and reflection padding to alleviate the singularities at the poles [32]. Then, in order 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 (i.e., gravity direction), assuming that world-space vertical and horizontal features (GAF, gravity aligned features) have different characteristics in most, if not all, man-made environments [15, 56]. Compressed latent features $L_s = (l_1 \dots l_4)$ contain a wealth of information on the geometry of the scene, both local and non-local, which can be exploited to recover depth and layout, as well as to provide a latent representation on the scene for further processing (section 3.4.3).

For depth estimation, we aim to leverage complementary features in distant portions of the image rather than only local regions, to maximize the wide contextual information provided by omnidirectional images while keeping the computational cost low. To do that, we adopt a single-layer multi-head self-attention (MHSA) scheme [82] to process the latent feature (see Supplementary material). Once passed to the MHSA module, the decoding of the latent feature is very fast, through convolutions, upsampling modules, and ELU activations, until we reach the target output resolution ($1 \times h \times w$ in Figure 3.3a).

Layout estimation block. According to the Atlanta World model [1], the indoor scene is expected to have a horizontal floor and ceiling and vertical walls, but without the restriction of walls meeting at right angles (supporting, e.g., curved walls). The indoor model can, thus, be fully represented as a 2D polygon around the observer,

that defines the wall footprint, and two scalars d_c and d_f that respectively represent the distance of the observer from the ceiling and floor planes. Without loss of generality, we assume the observer is placed at the origin of a reference frame (Y front, Z up, and X right), and all measures are in metric units. In our network, we represent the layout as a probability map ${P_w}^{wp \times wp}$ tensor, along with the two scalars d_c and d_f. The probability map P represents the probability that a point in the floor plane is inside the room boundary. While the resolution of the depth map $h \times w$ only depends on the input image resolution, the probability map, by design, has a fixed size (i.e., $wp = 512$ in our experiments).

Since layout and depth are inter-related, as the layout must be consistent with the depth assigned to pixels that are part of the architectural structure, we estimate layout jointly with depth, with the added benefit for interactive applications of avoiding the use of separate branches that would increase computational burden (Table 5.1). Moreover, we assume that the prediction of the geometric layout can be extended to non-visible part by exploiting geometric features derived from the depth map, such as flatness, sharpness, and smoothness [56]. Since the recognition of planar and curved structures is not immediate in spherical space, we transform the equirectangular map D_s, representing the Euclidean distances of pixels from the camera center, according to the Atlanta World model. To do this, we apply a planar transformation to D_s, such that the distances are expressed not with respect to a point but with respect to the horizontal plane containing the camera center (Figure 3.3b). This is accomplished by scaling each value in D_s, 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 specific transform imposes that horizontal structures, such as ceilings, floors, or even large tables or furniture, have a constant value so that structures identification is simpler, as shown in the Figure 3.3b example.

This representation is in an equirectangular format that still has its natural spherical distortion. In order to further simplify structure recognition, we then apply the Atlanta World transform, which has been proven to be effective on equirectangular image analysis [69, 1]. Specifically, we project the equirectangular depth map ${D_h}^{h \times w}$ to two projective planes, ${P_c}^{wp \times wp}$ and ${P_f}^{wp \times wp}$ perpendicular to the Z-axis, so that ceiling projection P_c represents depth information belonging to the upper hemisphere of D_h, while the floor projection P_f to the bottom hemisphere. For every

pixel in the perspective image at position (p_x, p_y), we recover a depth value from the corresponding pixel in the equirectangular map by defining a focal length:

$$f = \frac{wp}{2} \cot \frac{fov_p}{2}, \tag{3.1}$$

with $fov_p = 165°$ in our experiments. Then, we sample the equirectangular panorama at the coordinates

$$(u_s, v_s) = \left(\frac{\arctan(s_x/s_z)}{\pi}, \frac{\arcsin(s_y)}{\pi/2} \right), \tag{3.2}$$

where (s_x, s_y, s_z) is the direction vector from the origin to the point (p_x, p_y, f). Since the whole process is differentiable, it can be used in conjunction with back-propagation. Furthermore, both the depth transformation and perspective projection Equation 3.2 are implemented with simple GPU operations (section 3.5) with negligible computational cost.

Since the top view P_c is clearly more clutter-free, as it represents the ceiling, we use that view to derive the shape of the room, while we exploit the floor projection P_f to recover an occupancy map. Furthermore, the respective distances of the ceiling and floor planes from the center of the room are $d_c = max(P_c)$ and $d_f = max(P_f)$.

Simply extrapolating the layout of walls from P_c would return an incomplete reconstruction for eventually occluded parts (Figure 3.3b, first transform). This is a common problem in almost all single view approaches (section 3.3), which is commonly solved by a heuristic post-processing step [24]. However, such a solution, in addition to adding computational cost, works only for very simple cases.

In order to have a more versatile reconstruction we decide to introduce a specific data-driven solution. Starting from the projected depth P_c, we include in our ADM module a further multi-layer perceptron, named layout estimation network (LEN), which estimates a map P_w representing the probability of being inside the room footprint on the floorplan. As the visible shape of the room is already highlighted in the input P_c, the LEN exploits such contextual geometric information to efficiently complete the missing parts, as shown in the example of Figure 3.3b. The LEN, integrated into the same ADM network, is very simple and lightweight (section 3.5.2), as it is realized as a lightweight encoder-decoder network based on the U-Net architecture, using just 256 channels as a bottleneck ($4M$ parameters) and skip-connections [83]. We also tried different configurations for this task, experiencing no performance increment with deeper layers. Once a contour C_{xy} is obtained from P_w and scaled to metric

units (see supplementary material for details), a full 3D layout A_s is obtained (i.e., d_c and d_f, respectively, z-up and z-down components).

Exploiting structural information. The information in the floor projection P_f can be used to recover a floor occupancy map, useful, for example, to define a collision-free VR work area (e.g., Meta Quest room-scale), or to generate a reliable trajectory for new poses, if only to ensure that new views do not cross wall boundaries and remain in the room interior. To efficiently perform those operations, we enrich the representation with a clutter binary mask ${O_m}^{h \times w}$ in an equirectangular format. To obtain that, many lightweight pre-trained networks are available [32, 40]. Using O_m on D_h, the floor projection will be automatically cleaned-up from clutter before transforming (Equation 3.2), so that P_f will return the valid work area of the floor (Figure 3.3b). Some examples of this feature are presented in the supplementary material.

3.4.2 Novel view synthesis

While the previous operations are performed once per image, the *gated view synthesis network (GVS)*, illustrated in Figure 3.9 (cyan block), is activated whenever a movement occurs. Its purpose is to compute a new plausible spherical image from a translated position (i.e., applying a translation T to the camera). Such spherical images can then be sampled with regular means by obtaining all possible rotated views. The GVS network includes two cascading steps: a differential rendering step, which exploits depth D_s and translation T to move pixels information to the new position, and a panoramic view synthesis step, which transforms the reprojected information into a full output image. Such a network takes as input the translated pixels $\widetilde{I}_t$, the disocclusion mask B_t and the layout edges E_t, returning as output the novel view I_t.

Differential rendering. Reprojecting source pixels to their target position requires finding the mapping from the source-view pixels (u_s, v_s) to the target-view pixels (u_t, v_t), which is obtained by converting source pixels to a 3D point cloud PC_s, translating it by T and converting back the resulting point cloud PC_t to image space. As pixel coordinates $(u, v)^{h \times w}$ in an equirectangular image are associated to the view-directions (θ, ϕ), we apply the pixel-wise depth D_s at the same location to recover each point in PC_t as $(m_x, m_y, m_z) = (D_s \cos\theta \cos\phi - T_x, D_s \sin\theta \cos\phi - T_y, D_s \sin\phi - T_z)$.

Then, for each triplet $(m_x, m_y, m_z) \in {PC_t}^{3\times h\times w}$ we obtain coordinates in the source image:

$$(u_s, v_s) = \left(\frac{w \arctan(m_x/m_y)}{2\pi} + \frac{w}{2}, \frac{h \arctan(m_z)}{\sqrt{{m_x}^2 + {m_y}^2}} + \frac{h}{2} \right). \tag{3.3}$$

Since many source points may contribute to the same target image pixel, we want the closer ones to occlude the further ones. In traditional rendering, this can be achieved using a z-buffer, with only the closest point contributing to the rendering of a pixel. However, this process results in a discontinuous and non-differentiable rendering function that is unsuitable for a learning framework [84]. To this end, we adopt a *soft z-buffer* approach to assign a value to each pixel of $\widetilde{I}_t$:

$$\widetilde{I}_t(u_t, v_t) = \frac{\sum_{(u_s,v_s)} I_s(u_s, v_s)(\exp(-D_s(u_s, v_s)/\tau)}{\sum_{(u_s,v_s)} \exp(-D_s(u_s, v_s)/\tau + \varepsilon}. \tag{3.4}$$

The exponential factor, modulated by the temperature τ (i.e., $\tau = 20$ in our experiments), enforces higher precedence for points closer to the camera. A large value of τ results in *softer* z-buffering, whereas a small value yields a rendering process analogous to standard z-buffering [84]. ϵ is a small constant for numerical stability.

After forward splatting through soft z-buffering, as also shown in Figure 5.2, the pixels visible in both the original and translated viewpoint get the expected content, but all disoccluded areas (i.e., the areas visible from the new viewpoint but invisible in the original one) remain empty.

Panoramic view synthesis. The goal of view synthesis is to produce a complete image from the partial information obtained after reprojection, exploiting all the auxiliary information we have generated in previous steps. To this end, several approaches splat source-view features, filling missing holes using feature interpolation approaches [57, 17]. This approach aims to convey more semantic content, but at the same time, it irretrievably loses details of the original image, thus requiring much deeper networks to arrive at the synthesis of the image [85] and much computational effort. To overcome these problems and better adapt novel view synthesis to the VR interactive context, in our approach we address the problem as an image completion and inpainting task that leverages the recovered indoor structure to ensure consistency at various levels.

We start from the rendered image $\widetilde{I}_t^{3\times h\times w}$. As in typical inpainting approaches, we define a binary inpainting mask ${B_t}^{1\times h\times w}$, identifying missing parts in the rendered image. In contrast to pure image-domain approaches [33, 36], we further enrich

the input with the additional information provided by the indoor structure to guide image completion. Since the recovered layout L_s is represented in 3D space, we first project it to the target equirectangular pose. We experienced that the most effective format is through an *edgemap* with occlusions ${E_t}^{1\times h\times w}$, that is, a map storing the visible edges of the room boundary layout (Figure 3.3b). E_t is then concatenated to $\widetilde{I}_t$ and with the mask B_t (i.e., along the batch dimension - 5 layers input). It should be noted that such representation acts as edge guiding[85], but with the main difference that E_t provides information even of parts inside the mask B_t, which are invisible in the source image I_s.

To process such input, we adopt the architecture illustrated in Figure 3.9. The overall encoder-decoder scheme follows a typical design for image inpainting [86], exploiting gated convolutions for encoding/decoding [87] and dilated convolutions as bottleneck [88]. Compared to common inpainting baselines [86, 87], our architecture is thinner, deeper, and with fewer parameters. Moreover, it has only a single branch and it includes several solutions to improve accuracy and reduce computational complexity.

To simplify training and guarantee low latency at inference time, our network uses a modified version of gated convolution called Light Weight Gated Convolutions (LWGC), which reduces the number of parameters and processing time while maintaining the effectiveness [89]. The input is then encoded through a sequence of lightweight gated convolutions having different strides (Figure 5.2, red blocks). Repeated dilations [88] are instead used for the bottleneck (Figure 5.2, yellow blocks). The *dilated convolution operator* is implemented as a modified gated convolution:

$$D_{y,x} = \sigma(b + \sum_{i=-k'_h}^{k'_h} \sum_{i=-k'_w}^{k'_w} W_{k'_h+i,k'_w+j} \cdot I_{y+\eta i,x+\eta j}), \tag{3.5}$$

where η is a dilation factor, $\sigma(\dot{})$ is a component-wise non-linear transfer function and $b \in \mathbb{R}^{C_{out}}$ is the layer bias vector. In our model, we adopt, respectively, $\eta = 2, 4, 8, 16$ for the four bottleneck layers. The network decoder follows a scheme that is symmetrical with respect to the scheme of the encoder until the input resolution is reached.

3.4.3 Training and losses

We train both the ADM (section 3.4.1) and GVS (section 3.4.2) networks using a supervised training approach (Figure 3.2b) on synthetic data (see section 3.5.1). ADM requires a dataset in which ground truth depth is also available, while GVS requires only original and translated views, as it can exploit a pre-trained ADM for geometric and structural information.

ADM loss functions. We train the ADM network by combining and extending the standard depth and layout losses:

$$\mathcal{L}_{adm} = \lambda_d \mathcal{L}_d - \lambda_{ss} \mathcal{L}_{ss} + \lambda_l \mathcal{L}_l + \lambda_h \mathcal{L}_h. \tag{3.6}$$

$\mathcal{L}_d$ is the robust *Adaptive Reverse Huber Loss (BerHu)* [90] for the predicted depth;$\mathcal{L}_{ss}$ is the Structural Similarity Index Measure (SSIM), which measures the preservation of highly structured signals with strong neighborhood dependencies; $\mathcal{L}_l$ is binary cross entropy wuth logits loss for the predicted probability map P_w section 3.4.1; $\mathcal{L}_h$ is the $L1$ distance error for the predicted ceiling-floor distances D_c and D_f. The λ weights in our experiments are $\lambda_d = 1.0, \lambda_{ss} = 0.5, \lambda_l = 0.5, \lambda_h = 0.1$.

GVS loss functions. The novel view synthesis network is trained by combining visual terms and indoor-domain geometric terms: $\mathcal{L}_{vsn} = \mathcal{L}_{vis} + \mathcal{L}_{geo}$.

Visual terms include losses that measure the photorealistic quality of the output:

$$\mathcal{L}_{vis} = \lambda_{px} \mathcal{L}_{px} + \lambda_{perc} \mathcal{L}_{perc} + \lambda_{style} \mathcal{L}_{style} + \lambda_{adv} \mathcal{L}_{adv}. \tag{3.7}$$

Here the first term is a pixel-based $L1$ loss between the predicted RGB image I_t and the ground truth target image I_{gt}, $\mathcal{L}_{perc}$ and $\mathcal{L}_{style}$ are the data-driven perceptual and style losses [50], enforcing I_{out} and I_{gt} to have a similar representation in the feature space as computed by a pre-trained $VGG-19$ [91], while $\mathcal{L}_{adv}$ is a discriminator-based loss (i.e., PatchGAN [92]). λ weights are $\lambda_{px} = 1.0,\ \lambda_{style} = 100.0,\ \lambda_{perc} = 1.0,\ \lambda_{adv} = 0.2$. Such components are a common and effective solution for many single pose inpainting problems [87].

However, in our problem the scene to be reconstructed is from a different pose, thus standard inpainting techniques return many artifacts, especially for disoccluded structures, such as wall edges or hidden corners (Figure 3.4). To this end, to better

exploit the guiding of the structural information, we introduce in our method specific geometric and indoor-domain loss terms, exploiting the capabilities of ADM network to return indoor scene latent representations.

As described in section 3.4.1, our latent, compressed scene representation is given by 4 layers $L = (l_1 \dots l_4)$, which shapes (i.e., $l \times s$), for a 1024×2048 input, are: 512×512, 512×256, 512×128, 512×64.

Since the network is pre-trained to recover pixel-wise depth and the Atlanta World model of the room, including the parts occluded in the source view, we assume that L features contain important characterizing features of the indoor scene we want to reconstruct.

Analogous to the fundamental concepts of style-transfer[50], we expect that the $L1$ distance between the latent representation of I_t and I_{gt}, respectively the predicted and ground truth target images, preserves the *high-level* content of the scene, and thus global similarity:

$$\mathcal{L}_{geocont} = \sum_{n}^{4} \|L_n(I_t) - L_n(I_{gt})\|_1. \tag{3.8}$$

According to the same concepts, we also define an objective function giving more importance to local similarity, acting as a kind of *geometric style* loss, based on the *Gram matrix* function of the same 4 layers:

$$\mathcal{L}_{geostyle} = \sum_{n}^{4} \left\| K_n(L_n(I_t)^T L_n(I_t)) - L_n(I_{gt})^T L_n(I_{gt}) \right\|_1, \tag{3.9}$$

where K_n is the Gram matrix normalization factor $1/s * l$ for the nth selected layer.

In addition, a direct depth loss term $\mathcal{L}_{tdepth}$ is included (i.e., (BerHu) [90]) to enforce geometric consistency. It should be noted that, since the available datasets[17] do not provide a ground truth depth for the target pose, the loss is calculated by assuming as ground truth the depth predicted by the ADM network, respectively on the target ground truth image I_{gt} (dubbed $D^*{}_{gt}$) and on the predicted image I_t (dubbed D_t) (Figure 3.2b). For completeness, we also tried a self-supervised loss in order to estimate the target depth without ground truth [93], but with significantly less accurate results.

As a result, the full geometric term is:

$$\mathcal{L}_{geom} = \lambda_{gcont}\mathcal{L}_{gcont} + \lambda_{gstyle}\mathcal{L}_{gstyle} + \lambda_{tdepth}\mathcal{L}_{tdepth}. \tag{3.10}$$

Here λ weights are $\lambda_{gcont} = 1.0$, $\lambda_{gstyle} = 100.0$, $\lambda_{tdepth} = 0.01$.

3.5 Results

Our approach was implemented using *PyTorch* [51] and has been tested on a large variety of indoor scenes. The accompanying video shows its usage for the exploration of free viewpoint exploration of panoramic images with HMDs. In this section, we focus on analyzing the performance of our approach for depth and layout estimation and view synthesis.

3.5.1 Training and testing datasets

For training our solutions, we harness the availability of public panoramic scene datasets where ground truth is available. In particular, for training and testing ADM, we exploit Structured3D [52]), a large-scale synthetic database of indoor scenes comprising 21,000 photorealistic scenes, which provides ground truth depth and layout information for each panoramic image. To train and test GVS, instead, we exploit PNVS[17], a subset of Structured3D scenes providing, for each source panoramic image, three views translated by 0.2-0.3m along random directions, and three views translated by 1.0-2.0m. In contrast to the original PNVS setup, we included the zero-motion case for a fraction of the samples (i.e., 15%) to better adapt the dataset to a common VR use case where users may remain still for a portion of the time. It should be noted that in PNVS, the ground truth depth and layout are provided only for the source pose, while only the visual rendering is provided for the target views. The pre-trained ADM network is therefore used for providing the additional geometric and structural information, which is regarded as ground truth for GVS training. All these datasets provide data at a resolution of 1024x512 that we have used for all training.

In this paper, we also use Structured3D [52] and PNVS [17] as test datasets, using official splits that do not replicate data between training and testing sets, so as to make it possible to have a comparison with other solutions. Furthermore, to demonstrate

transfer learning capabilities and versatility, we present results on real-world scenes captured by non-professional users.

We also considered other commonly used datasets, but none of them were fully suitable for our task. As an example, Matterport3D [41], provides incomplete depth maps not reliable for accurate rendering of points on novel views, while MatterportLayout [14] (annotated layouts for the Matterport3D dataset), only provides layout for a limited number of rectified scenes, with manual annotation not always coincident with the underlying image[14].

In our tests, we handle novel poses in a range of $50cm$, which is well above what is required for stereo (6-7cm) and assumed consistent with natural head movements to avoid full hallucination of image content [94, 57]. The accompanying video shows typical allowed motion.

(a) RGB input **(b)** ground truth **(c)** other [6] **(d)** our

Fig. 3.4.: We present qualitative performance and comparison vs. ground truth and PanoSynthVR [6] on the Structured3D dataset [52]. The average movement for each scene is about $50cm$ distributed on x, y, z axis.

3.5.2 Setup and computational performance

We trained both the ADM (section 3.4.1) and the GVS (section 3.4.2) networks with the Adam optimizer [95], 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 for ADM and 2 for GVS. When using the Structured3D [52] 512×1024 native resolution for both training tasks [52, 17], the average training time for the ADM model is 76 ms/image and 428 ms/image for the GVS model. Inference time on the same NVIDIA RTX A5000 is 18 ms/image for ADM and 29 ms/image for GVS.

Method	Params↓	GFLOPS↓	Output type
Bifuse [63]	253 M	682	only depth
SliceNet [15]	79 M	101	only depth
AtlantaNet [48]	100 M	273	only layout
ADM (our)	**29 M**	**79**	**depth+layout**

Tab. 3.1.: **Depth-layout estimation computational performance.** We show our computational performance compared to other specific state-of-the-art works for a 512×1024 image.

Table 5.1 presents the computational performance of our ADM compared to state-of-the-art depth and layout estimation solutions.

For depth estimation, we compare with SliceNet [15] and Bifuse [63], which are state-of-the-art methods commonly used as benchmarks in the latest panoramic works [67, 66]. Both works [15, 63] adopt as backbone a ResNet, which is often employed for depth estimation in stand-alone view synth pipelines [17], but do not use patch projection or transformers, that would add additional load to the pipeline making it less suitable for VR applications.

For layout estimation, we compare our method with AtlantaNet [48], a fast state-of-the-art solution that also handles the same scene types as ours. In particular, AtlantaNet does not use pre- and post-processing (i.e., usually done in CPU with considerable computational load), as done by pipelines based on LayoutNet [55, 17], or HorizonNet [24]. Our ADM approach is clearly the most lightweight and has lower computational complexity (GFLOPS) than the compared methods, even though it jointly performs both tasks.

Method	Params↓	GFLOPS↓	Output type
PNVS [17]	13.9 M	359	rgb
DeepFillv2 [33]	13.8 M	163	rgb
GVS (our)	**1.7 M**	**83**	**rgb-d**

Tab. 3.2.: **View-synthesis computational performance.** We show our computational performance compared to other deep-learning approaches for view synthesis for a 512×1024 image.

Table 3.2 presents the computational performance for the view synthesis network (GVS). Here we compare pipelines that are, like ours, end-to-end deep learning networks. Thus, we compare to the PNVS [17] view synthesis branch, as well as, for completeness, with a state-of-the-art network for generic image inpainting (*DeepFill* [33]). Since the PNVS [17] source code is not available, we evaluate its computational cost from the information provided by the authors in the original paper, since the view synthesis network is an adaption of *EdgeConnect* [85] network.

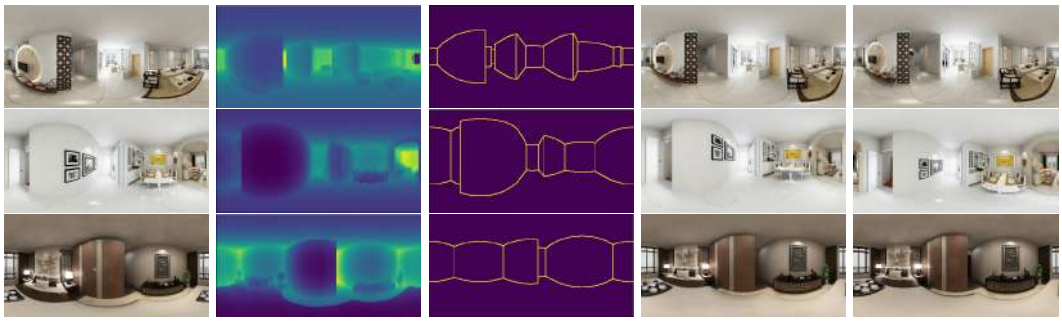
Other types of approaches, not directly comparable in these terms, are also included for completeness. PanoSynthVR [6] exploits a pre-trained MPI network [19] to build, for each input panoramic scene, an MCI (multi-cylinder image) structure is about in $383ms$ (declared by the authors on an NVIDIA V100 GPU). Similar considerations apply to NeRF adaptations to equirectangular images. In this case, the training time is about $8h42m$ an NVIDIA RTX A5000 (24GB VRAM), with subsequent $14s$ inference time for each individual new scene view generated at 512×1024. In this case, much of the computational load is from generating new views around the main view.

3.5.3 Run-time performance

We ran the server connected to the display on a desktop machine equipped with an NVIDIA RTX 2080Ti. Predicting, once per scene, the enriched representation with ADM takes 32 ms, while performing per-frame view synthesis with GVS takes 39 ms. Cropping the image to 90x90° and upsampling it (2x) using *Real-ESRGAN* (with model *realesr-animevideov3*) [80] takes 39 ms. Transfering to CPU and image encoding, which in our prototype is done using TurboJPG takes an additional 5 ms/image. Server-side, thus, image computation can be performed at about 12fps, and is reduced to about 11fps including encoding and transmission. It should be noted that the machine is over 30% slower than the A5000 used for training. Moreover, encoding time could be reduced by integrating hardware JPEG encoding [96] or using alternative codecs, in particular for the ETC2 texturing format, which is widely supported on mobile platforms [97]. Client side, the WebXR application running on A PICO 4 VR headseat refreshes the display, in response to head rotations, at 72fps and updates the current panorama at the server speed. As a result, the proof-of-concept implementation supports about 70Hz refresh rate while updating panoramas at 11fps with a latency of $\approx 0.1s$, for a working volume of $\approx \pm 30cm$ around the original viewpoint. Latency was measured on the client by computing the difference from the time at which the position was sent to the server and the time the corresponding updated panorama is first displayed.

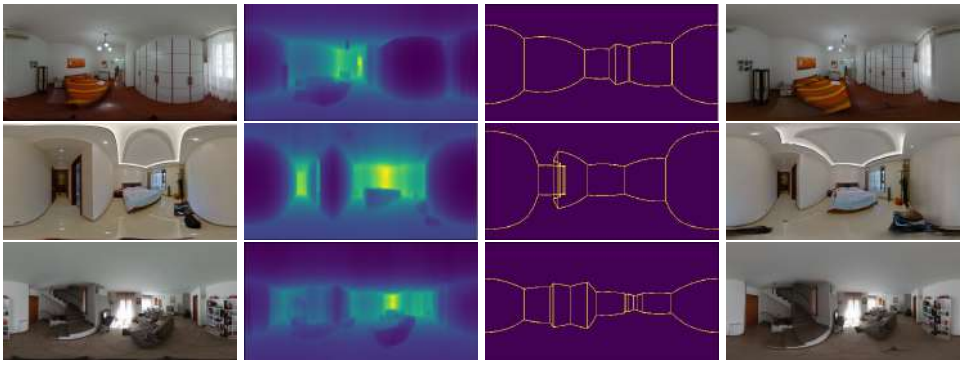
3.5.4 Performance vs. ground truth and competitors

As discussed in section 3.4, the quality of the novel pose generation results strongly depends on the geometric information available, since the input image to be completed depends on the accuracy of reprojection, guided by the estimated depth, and the complementary information to guide inpanting depends on the indoor structure



(a) RGB input (b) P. source depth (c) P. target layout (d) Predicted view (e) Ground truth

Fig. 3.5.: We present our qualitative performance on scenes with structural occlusions. The average movement for each new pose is about $60\ cm$ distributed on x, y, z axis.



(a) RGB input (b) Pred source depth (c) Pred target layout (d) Predicted view

Fig. 3.6.: We present our qualitative performance on scene acquired by non-professional users. Input resolution here is 6720×3360. The average movement for each scene is about $40\ cm$.

inferred (section 3.4.2). For this reason, we present specific results for depth and layout estimation, followed by the results on the quality of the synthesized scene.

For depth estimation, in Table 5.4 we summarize our performance compared to SliceNet [15], and to the work of Jin et al. [98], which is a representative pipeline that jointly predicts depth and layout, where layout is predicted through LayoutNet [55, 17]. While the source code is available for SliceNet, for Jin et al. [98] we compare with their official results, where only depth estimation performance is available. For a fair comparison, we adopt the Structured3D [52] splitting of Jin et al. [98], adapting both SliceNet and our code to it. We adopt common metrics, i.e., mean squared error (MSE) and root mean square error of linear measures (RMSE) and relative accuracy δ_1, defined as the fraction of pixels where the relative error is within a threshold of 1.25. For layout estimation, we compare, instead, with AtlantaNet [48], an end-to-end solution that, like ours, does not require Manhattan World pre and post-processing to work [14]. Here, we adopt the common metrics IoU3D (volumetric intersection over union) and IoU2D (pixel-wise intersection over union). The results demonstrate

Method	mse↓	rmse↑	δ_1 ↑	iou3d↑	iou2d↑
Jin et al. [98]	0.103	0.666	0.91	-	-
SliceNet [15]	0.044	0.174	0.93	-	-
AtlantaNet [48]	-	-	-	82.45	85.78
AVN (Our)	**0.008**	**0.043**	**0.96**	**84.56**	**88.86**

Tab. 3.3.: Depth and layout performance. We show our quantitative performance compared to other representative state-of-the-art works.

how our method achieves state-of-the-art performance in both tasks, despite the lower computational burden compared to those baselines. To evaluate the gated

Method	PSNR↑	SSIM↑	LPIPS↓
MPI 32 [19, 6]	17.59	0.768	0.263
MPI 64 [19]	17.93	0.783	0.258
MPI 128 [19]	18.22	0.789	0.252
GVS (Our)	**22.97**	**0.817**	**0.178**

Tab. 3.4.: GVS quantitative performance. We show GVS quantitative performance compared to other state-of-the-art works.

view synthesis network (GVS), we compared our performance to the one achieved by state-of-the-art methods [19, 18, 6], which are representative and suitable for VR applications, as discussed in section 3.3, and for which source code was available.

In Table 3.4, we present our quantitative results compared to the solutions already exploited for panoramic VR applications [6] that can be trained end-to-end and for which a comparison with respect to ground truth was possible [19, 6]. Specifically, MPI [19] is adopted by PanoSynthVR [6] to generate multi-cylinder-images (in their case with 32 layers) from a single panoramic view, as well as by MatryODska [4] to generate low-resolution views form stereoscopic panoramic input. We also considered Synsin [57], but no official panoramic implementation was available, and only low-resolution results have been presented [17].

Since such multi-layer approaches have a limited working range, we adopt the PNVS benchmark called *easy set* [17], which limits motion to 0.1-0.2cm. Differently from the experiments proposed in the PNVS paper [17], we choose to test on Structured3D full resolution (i.e., 512×1024), since the benchmark proposed in the PNVS paper [17] was run at a resolution of 256×512, way too low for our applications.

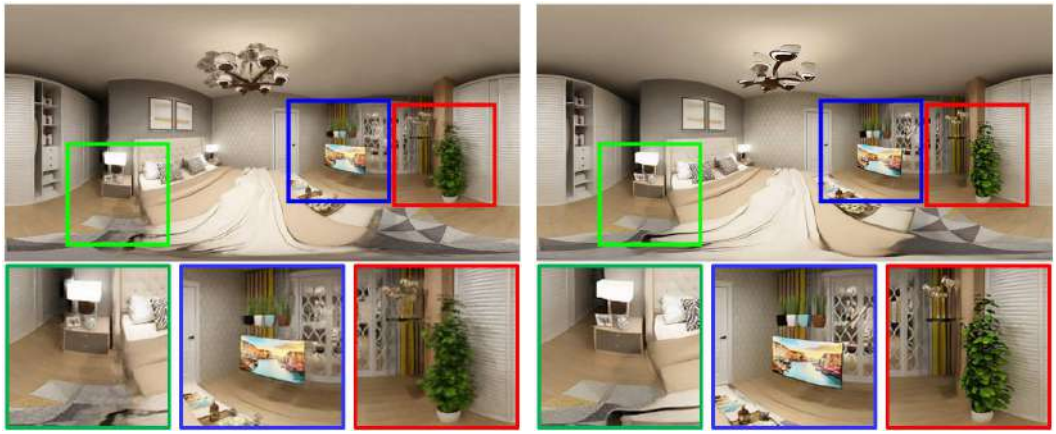
Table 3.4 summarizes the results with standard metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [99], and the Learned Perceptual Image Patch Similarity (LPIPS) [100]. These results show how our method outperforms

other solutions in terms of accuracy in all metrics. It should be noted that the current approaches are adaptations of methods based on planar or semi-planar projections, while our method fully exploits depth and layout data.

Qualitative results on a variety of indoor scenes are presented in Figure 3.4, Figure 3.10, Figure 3.6, and Figure 3.7. Figure 3.4 shows our performance compared with a recent approach for VR applications based on MPI [6], using the same target position and converting the cylindrical output of that system to an equirectangular map. As in Table 3.4, we adopt data for which ground truth is available [17]. In this regard, the same images provided by PanoSynthVR are not usable for direct comparison, since these are cylindrical crops and not full equirectangular scenes. The comparison shows how our method is able to predict occluded and disoccluded parts even in the presence of significant structural occlusions, such as corridors to particularly bulky furniture. Furthermore, besides our superior performance in terms of accuracy, it should be noted that the compared solution, although returning a perceptual plausible view, does not reconstruct a spatially reliable scene, probably due to approximation with a limited number of planes/cylinders.

In Figure 3.10, we present instead qualitative examples of our performance, illustrating different tasks. Alongside the input source image, we show the predicted source depth, the predicted layout translated at the target position, our prediction at the target position, and the ground truth target. In this case, it is noticeable the correlation of depth and layout quality with the generated output view.

In addition to the results on synthetic scenes, we present in Figure 3.6 qualitative performances on real-world, user-captured scenes. Here we exploit the training with Structured3D [52] to predict depth, layout, and novel views, from an input 6720×3360 images captured by a Ricoh Theta S. Also in this case our method returns visually realistic reconstructions. Finally, we present a qualitative comparison with a NeRF approach. As recent omnidirectional image-based methods attempt to build a NeRF structure from a single image (section 3.3), in Figure 3.7 we show a comparison to a NeRF-based approach for equirectangular images, OmniNeRF [18]. In this case, as no ground truth is available, we present a qualitative assessment of the data made available by the authors, considering their same spatial range (i.e., $30cm$). As clearly highlighted in Figure 3.7 details, our solution provides better accuracy in many parts of the scene. In this context, we expect that our work could be used to generate input views for further NeRF processing.



(a) OmniNeRF [18] (b) Our

Fig. 3.7.: **Comparison to NeRF.** we show a comparison to OmniNeRF [18], using data released by the authors. As the details clearly show, our solution provides better accuracy in many parts of the scene.

3.5.5 Discussion and ablation study

Table 5.5 illustrates the results of an ablation study made to analyze the major technical choices of our method. The first test (first and second row in Table 5.5) shows the importance of effective depth estimation. The accuracy of depth is critical to synthesize the new pose, as the correct displacement of elements visible from the new view depends on it [57]. In our ablation study, the first row presents results where depth is estimated with a domain-independent approach with a baseline that exploits only multi-resolution aggregation [17] but without compression according to a preferred direction. Using gravity-aligned features here results in a great increment in performance.

Row 3 and 4 in Table 5.5 show the contribution of layout knowledge to handle occlusion and disocclusion. In this case, we detail the performance difference between using a standard layout estimation with Manhattan World post-processing [24], vs. our approach (i.e., ATL - row 4). Finally, in row 5 we show our full-configuration performance, even using our novel geometric perceptual and style losses (see section 3.4.3). It should be noted how this contribution mainly affects SSIM and LPIPS.

Our model proves versatile on different types of indoor scenes, even as the type of real or synthetic input data varies. However, there are cases, mainly in real-world scenes very different from training data, where our method did not produce plausible images. In such bad cases (i.e., illustrated and discussed in the supplementary

GAF	MW	ATL	GPS	PSNR↑	SSIM↑	LPIPS↓
-	-	-	-	16.87	0.693	0.325
✓	-	-	-	18.28	0.751	0.206
✓	✓	-	-	19.12	0.776	0.186
✓	-	✓	-	22.01	0.798	0.181
✓	-	✓	✓	**22.97**	**0.817**	**0.178**

Tab. 3.5.: **Ablation stats.** We show the effect of several key choices of our approach. In bold is the adopted configuration. GAF: gravity aligned features-based depth estimation; MW: standard Manhattan World layout estimation; ATL: Atlanta transform structure estimation; GPS: geometric perceptual and style loss.

material), the lack of performance depends on the quality of the depth associated with the initial view. This depth, in fact, depends on the projection of the visible pixels or any associated features, which is, in fact, the real input to the actual view synthesis network. In this sense, even the commonly used soft z-buffer method [84] may be subject to error and may be the subject of future work.

We also noted that, even if the structure and depth of the scene are predicted once and remain the same for all frames, visible instabilities may occur in the form of flickering in the application in which a new image is generated per frame. Examples are visible in the accompanying video. This is due, mainly, to small changes in the reprojection that trigger large changes in predicted images. We plan to mitigate this problem through the inclusion of regularization terms, as well as by augmenting the training sets.

Another important discussion point is the resolution of the generated images. Currently, the biggest limitation is the resolution of the available training datasets, which is still below the capabilities of modern VR viewers. Although new datasets will soon be available at higher resolutions, one practical solution is now the use of fast super-resolution methods [80, 101].

3.6 Client server design

The diagram in Figure 3.8 illustrates the operation of a client-server application for immersive rendering of synthesized indoor environments from new viewpoints using a pre-trained neural network. A user wears a head-mounted device (HMD) where the client runs on a browser enabled for immersive WebXR applications. The server consists of a multi-threaded C++ program that continuously receives the head position and control state from the client via a WebSocket and emits to the client the

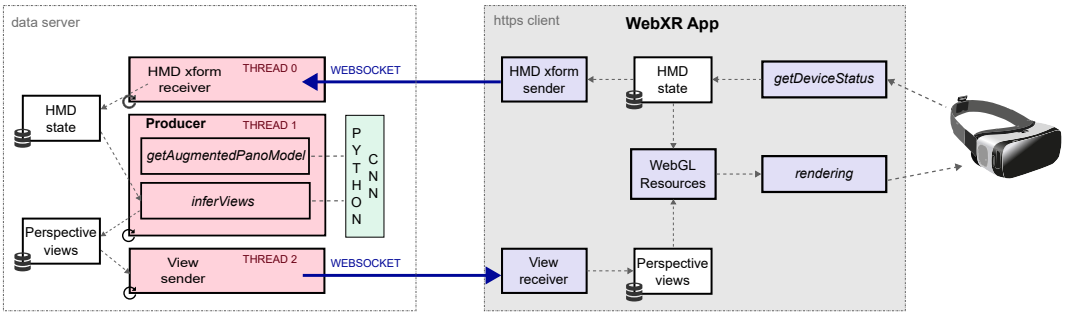


Fig. 3.8.: Client-server system architecture. A thin WebXR client performs image display and head tracking, offloading translated panoramic image computation to a rendering server.

images synthesized by the neural network and corresponding to what is visible from the new viewpoint. Upon receiving new images, the JavaScript client updates the WebGL resources from the server and handles the interactive rendering of the scenes by merging the received information with the HMD device state data.

Specifically, the server is implemented in C++ on Linux using the QT framework, which enables multithreading. The main thread (HMD xform receiver) is the main QT application, and its sole purpose is to establish a WebSocket connection with the client, wait for HMD position messages, and keep the HMD state updated using appropriate synchronization primitives. A second thread (Producer) retrieves the current HMD state and infers the views corresponding to the new state through the CNN network using the CPython interface. During initialization, the Producer requests the first network to produce the augmented representation (getAugmentedPanoModel) and configures the synthesis network to be ready to produce translated views using this augmented representation.

When a new HMD state (new viewpoint) is available, the Producer synthesizes two cropped equirectangular views that cover the field of view at different resolutions. By default, for maximum coverage, the low-resolution image spans a large area (fov_lrt to 180x180 degrees) encompassing the full frontal view. This allows the user to perceive a continuous immersive space, even during rapid head rotations. The high-resolution image, instead, is limited to cover only the field of view of the headset itself (fov_hr = 90x90 degrees), to provide greater visual detail of the projected content. When the user is about static, the full image is in high resolution, while when the user rotates the head, low-resolution areas start to appear only in peripheral regions.

The production process consists in estimating the full low-res panoramic image with the view synthesis network, providing the current translation, and using the scene representation previously calculated. The resulting image tensor is then cropped to the field of view and encoded in JPEG format using TurboJPEG. The high-resolution image is, instead, produced by first cropping the region of interest of the tensor, and then upsampling it using a super-resolution neural network (currently *Real-ESRGAN* with model *realesr-animevideov3* [80]) prior to JPEG encoding.

The two cropped images are calculated based on the viewing direction and position of the headset in space. Computed images and their associated auxiliary information (look-at vector, fov_lr, fov_hr, image size, etc.) are stored in a synchronized data structure (Perspective views). Another thread retrieves the most up-to-date information from the views data structure, serializes it, and sends it to the JavaScript client via a WebSocket.

The three highly specialized threads run their cycles in parallel, maximizing the system's capabilities and ensuring efficient handling of all data acquisition and production processes.

The JavaScript client uses the ThreeJS library and its dedicated components for immersive systems (WebXR). The HMD consists of two screens, one for each eye. The client renders the cropped equirectangular images received from the server using a custom WebGL shader developed for this purpose. Upon receiving an update, the images are decoded and stored into two 2D textures. A fragment shader calculates the foveated rendering color by obtaining the camera ray, converting it to equirectangular UV coordinates, and blending the colors from the high-resolution and low-resolution textures based on the alpha value of the high-resolution texture.

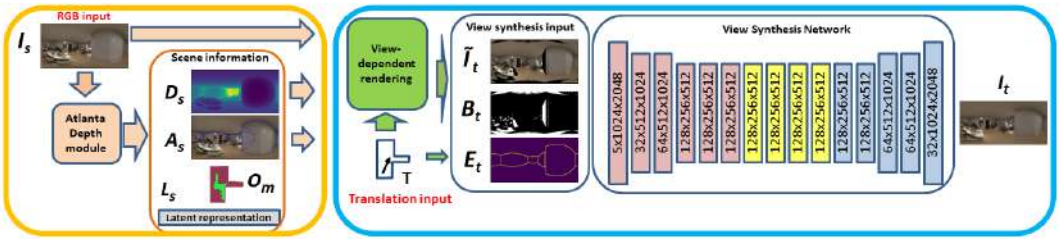


Fig. 3.9.: **Forward pipeline.** At loading time, we process the input equirectangular image to recover depth D_s, Atlanta structure A_s, occupancy map O_m, and latent scene representation L_s. When moving from the source position, the generation of the new views is done by a *soft* z-buffer and a gated neural network, dubbed *Gated View Synth network* - GVS. The GVS network is trained in a supervised way combining visual and perceptual losses with novel indoor-specific losses.

(a) RGB input (b) Occupancy map (c) Predicted frame

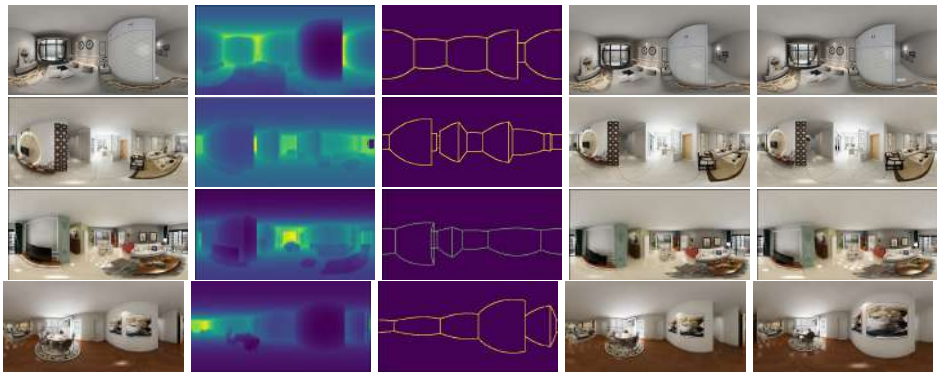
Fig. 3.10.: **Structure and synthesis.** We show some examples with the predicted floor occupancy map and novel poses generated inside it

3.7 Neural Network architecture details

As a complement of the overall description provided in the main paper, we illustrate here the detailed architecture of the depth estimation block (section 3.7.1), the details on the computation of floor plan and metric scaling (section 3.7.2), and the architectural details of the view synthesis network (section 3.7.3).

3.7.1 Depth estimation block architecture

From the input image, a cascade of 5 residual layers [81] returns 4 feature maps having different depth and spatial size (Figure 3.9). 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 [32]. In order 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, so we assume that world-space vertical and horizontal features have different characteristics in most, if not all, man-made environments [65, 24, 15, 56]. According to 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 3 times, that contains a 2D convolution and an ELU module. We apply such compression



(a) RGB input (b) Pred source depth (c) Pred target layout (d) Predicted view (e) Ground truth

Fig. 3.11.: Scenes with significant occlusions from architectural structure. We present additional qualitative performance on scenes with structural occlusions.

for each encoded feature map (i.e., 4 maps), obtaining a set of latent features $L_s = (l_1 \dots l_4)$. Such representation will be also exploited at training time to support specific loss functions in latent space. To support depth and layout prediction, instead, 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 - $s = 2048$ and $l = 512$ for a 1024×2048 input).

Such a compressed representation contains a variety of information about the geometry of the scene, both local and non-local, which can be exploited to recover depth, layout and also to provide a latent representation of the scene.

For the depth estimation, we aim to leverage complementary features in distant portions of the image rather than only local regions, to maximize the wide contextual information provided by omnidirectional images while keeping the computational cost low. To do that, we adopt a single-layer multi-head self-attention (MHSA) scheme [82]. Our self-attention module takes the latent features $L \in \mathbb{R}^{s \times l}$ as input, and outputs a self-attention weight matrix $A \in \mathbb{R}^{s \times s}$:

$$A = softmax\left(\frac{(LW_q)(LW_k)^T}{\sqrt{l}}\right) \tag{3.11}$$

where $W_q, W_k \in \mathbb{R}^{l \times l}$ are learnable weights. The MHSA module has a particularly lightweight design with 4 heads and only 1 inner layer. We have verified experimentally that increasing the number of layers and heads does not affect performance. 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 ($1 \times h \times w$).

3.7.2 Room contour extraction and metric scaling

Our ADM module includes a multi-layer perceptron, named layout estimation network (LEN) (Figure 3.9), that estimates a map P_w representing the probability of the room footprint on the floorplan. In order to estimate the 2D shape of the room, a set of corners C_{ij} in image space (i.e., here $wp \times wp$) is recovered from the contour of the probability map P_w (i.e., softmax and polygonal approximation of contour). By exploiting the metric distance of ceiling and floor, we recover the 2D corners C_{xy} positions in model space:

$$C_{xy} = \frac{d_c}{\tan(\pi - fov_p) * wp} * (C_{ij} - wp/2) \tag{3.12}$$

The full 3D layout A_s is then obtained, according to Atlanta model, by combining d_c and d_f, respectively, z-up and z-down components, with C_{xy}.

3.7.3 View synthesis network details

To simplify training and guarantee low latency at inference time, our network uses a modified version of gated convolution called Light Weight Gated Convolutions (LWGC), which reduces the number of parameters and processing time while maintaining the effectiveness [89].

The input is encoded through a sequence of light-weight gated convolutions having different strides (the 6 layers in red in Figure 3.9) so that the original size is reduced by a factor of four in each direction. Each encoding convolution is followed by instance normalization [102] and ReLU activation.

Repeated dilations [88] are instead used for the bottleneck (Figure 3.9, yellow blocks), thus increasing the area that each layer can use as input. It should be noted that this is done without increasing the number of learnable weights, but obtained by spreading the convolution kernel across the input map. The *dilated convolution operator* is implemented as a modified gated convolution:

$$D_{y,x} = \sigma(b + \sum_{i=-k'_h}^{k'_h} \sum_{i=-k'_w}^{k'_w} W_{k'_h+i,k'_w+j} \cdot I_{y+\eta i,x+\eta j}) \tag{3.13}$$

where η is a dilation factor, $\sigma(\dot{})$ is a component-wise non-linear transfer function and $b \in \mathbb{R}^{C_{out}}$ is the layer bias vector. With $\eta = 1$, the equation becomes the standard

convolution operation. In our model, we adopt, respectively, $\eta = 2, 4, 8, 16$ for the four bottleneck layers.

Using this strategy, we aggregate multi-scale contextual information without losing resolution, thus capturing the global context efficiently by expanding the receptive field, avoiding additional parameters, and preventing information loss. This is important for the image completion task, as capturing sufficient context is critical for realism. By using dilated convolutions at lower resolutions, the model can effectively cover a larger area of the input image when computing each output pixel than with standard convolutional layers [86].

The network decoder (7 blue layers in Figure 3.9) follows a scheme that is symmetrical with respect to the scheme of the encoder. Five layers, based on gated convolutions, restore the resolution of the output to the original input resolution, and a final layer is dedicated to the synthesis of *RGB image* by a $tanh$ activation function.

3.8 Additional experiments

To complement the results discussed in the main paper, we provide here additional examples of predictions (section 3.8.1), also illustrating the ability to compute the floor occupancy map (section 3.8.2), an expansion of visual comparison with PanoSynthVR [6] (section 3.8.3), and discussing a number of failure cases (section 3.8.4).

3.8.1 Additional prediction examples

In Figure 3.11 we present additional examples of our qualitative performance on scenes with structural occlusions. The average movement for each new pose is about $60\ cm$ distributed on the x, y, z axis. We show predicted depth, layout, and novel pose.

3.8.2 Examples with predicted work area

In Figure 3.10 we present additional results, showing, besides novel poses, the predicted floor occupancy map.

(a) Input pose (b) Our novel pose (c) Our detail (d) PanoSynthVR novel pose (e) PanoSynthVR detail

(f) Input pose (g) Our novel pose (h) Our detail (i) PanoSynthVR novel pose (j) PanoSynthVR detail

(k) Input pose (l) Our novel pose (m) Our detail (n) PanoSynthVR novel pose (o) PanoSynthVR detail

Fig. 3.12.: **Visual synthesis results.** Additional visual comparison with PanoSynthVR [6].

3.8.3 Additional comparisons with PanoSynthVR

In Figure 3.12 we present additional qualitative performance and comparison vs. ground truth and PanoSynthVR [6] on the Structured3D dataset [52]. The average movement for each scene is about $50cm$ distributed on the x, y, z axis.

(a) Input (b) Depth (c) Projection (d) Prediction

Fig. 3.13.: **Failure case.** View synthesis performance strongly depends on depth accuracy and visible feature projection.

3.8.4 Failure cases

As shown in the bad case of Figure 5.7, the performance of our approach, as well as of most current methods, depends on the quality of the depth associated with the initial view (i.e., Figure 3.13b). This depth in fact depends on the projection of the visible pixels or any associated features (i.e., Figure 3.13c), which is, in fact, the real input to the actual view synthesis network. If, as in the case shown, the

image projected at the new position is incorrect, it compromises any synthesis of the disoccluded parts as well (i.e., Figure 3.13d).

3.9 Conclusions

In this chapter We have presented a novel deep learning approach that extracts geometric and structural information from a single panorama in order to quickly synthesize plausible panoramic images from close-by viewpoints within a workspace suitable for VR applications. This end-to-end approach is particularly compact and lightweight, and introduces several innovations. In particular, our novel integrated network for estimating an environment's depth and permanent structure produces elements that are crucial requirements for ensuring reliable view synthesis. By incorporating novel domain-specific loss functions, we shift the major computational load on the training phase, and obtain an extremely lightweight network at prediction time. As a result, our method automatically produces compelling new poses ready for interactive use. Moreover, the extracted floor plan and 3D wall structure can also be used to support room exploration. A possible future work will concentrate on further improving the performance, especially on larger-size images, as well as the stability for its use in real-time exploration. We also plan to integrate this work with other solutions for the dynamic exploration of panoramic images, such as automatic room emptying and editing [40, 103].

3.10 Bibliographic notes

The contents of this chapter reports the journal article *Giovanni Pintore, Fabio Bettio, Marco Agus, and Enrico Gobbetti: Deep scene synthesis of Atlanta-world interiors from a single omnidirectional image* [104], of which the candidate is the lead author.

This publication was made possible by NPRP-S grant NPRP14S-0403-210132 by Qatar National Research Fund (a member of Qatar Foundation) and by Sardinian Regional Authorities through project XDATA. The findings herein reflect the work, and are solely the responsibility of the authors.

4 Immersive exploration of indoor stereoscopic environments

This chapter presents an innovative approach, targeting the objective 2, to automatically generate and explore immersive stereoscopic indoor environments derived from a single monoscopic panoramic image in an equirectangular format. Once per 360° shot, we estimate the per-pixel depth using a gated deep network architecture. Subsequently, we synthesize a collection of panoramic slices through reprojection and view-synthesis employing deep learning. These slices are distributed around the central viewpoint, with each slice's projection center placed on the circular path covered by the eyes during a head rotation. Furthermore, each slice encompasses an angular extent sufficient to accommodate the potential gaze directions of both the left and right eye and to provide context for reconstruction. For fast display, a stereoscopic multiple-center-of-projection stereo pair in equirectangular format is composed by suitably blending the precomputed slices. At run-time, the pair is loaded in a lightweight WebXR viewer that responds to head rotations, offering both motion and stereo cues. The approach combines and extends state-of-the-art data-driven techniques, incorporating several innovations. Notably, a gated architecture is introduced for panoramic monocular depth estimation. Leveraging the predicted depth, the same gated architecture is then applied to the re-projection of visible pixels, facilitating the inpainting of occluded and disoccluded regions by incorporating a mixed Generative Adversarial Network (GAN). The resulting system works on a variety of available VR headsets and can serve as a base component for a variety of immersive applications. We demonstrate our technology on several indoor scenes from publicly available data.

4.1 Introduction

Spherical cameras, also known as 360°, *panoramic*, or *omnidirectional*, or *surround-view* cameras, provide cost-effective and efficient solutions for rapidly capturing

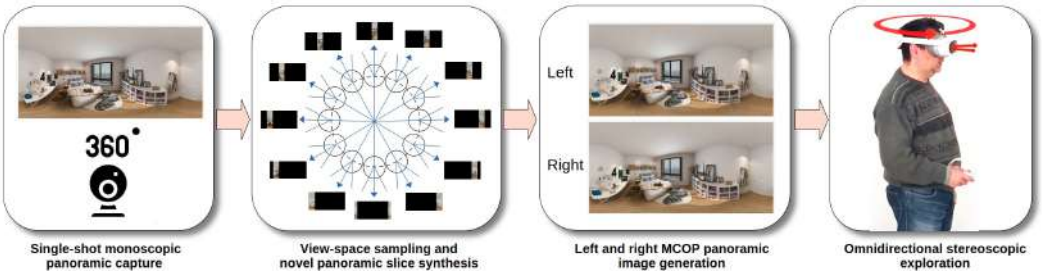


Fig. 4.1.: **Overview**. Taking as input a single panoramic image, a data-driven architecture synthesizes a comprehensive coverage of the scene's portion visible to both eyes during head rotation, encoding the views in the form of panoramic slices. The slices are then combined into an omnidirectional stereo representation composed of two multiple-center-of-projection (MCOP) images, tuned for the left and right eye. A lightweight WebXR viewer presents the suitable portions of these images on an HMD, responding to rotational head motions and delivering both stereo and motion parallax.

in a single shot the full context around the viewer of an entire environment [10]. A single panoramic image encompasses the complete scene visible from a specific viewpoint within a 360° field of view at a given instant. When experienced through a Head-Mounted Display (HMD), users dynamically explore this image by directing their attention to the desired content through head movements, leading to Virtual/Augmented/Extended Reality (VR/AR/XR) experiences with a natural interface and good degree of immersion [11].

For these reasons, omnidirectional imagery is increasingly recognized as a foundational element for generating immersive content from real-world scenes and for supporting a variety of VR/AR/XR applications [3]. Notably, 360° virtual tours have gained widespread popularity in the real estate sector [12]. Furthermore, omnidirectional images are easily shareable across various devices and platforms, making them highly versatile and accessible. Since they can be seamlessly integrated into websites, VR/AR/XR applications, or mobile platforms, they enable a broad audience to engage with indoor environments irrespective of their location or their available equipment [10]. Serving as representations of the user's surroundings, panoramic images also promise to be one of the essential building blocks for the construction of the shared physical and digital realities envisioned by the Metaverse concept [13].

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 location from which the panorama was taken leads to the loss of binocular stereo, which is very important to provide a sense of presence [6]. The fact that panoramas appear flat is a particularly strong limitation

in indoor environments, given the relatively short distance from the viewer to the architectural surfaces and the objects. To provide stereo cues for full 360-degree rotations, views from a continuous set of shifted viewpoints must be available to the renderer. Omnidirectional stereo techniques [25, 26] are employed for that purpose but require the creation of stereo panoramas using cameras moving on a circular path [27, 26, 25] or multiple synchronized 360 cameras [3]. These acquisition approaches, however, reduce the possibility of quickly capturing, experiencing, and sharing a 360° scene using consumer hardware. In particular, while several low-cost cameras are widely available for monocular 360° capture (e.g., GoPro, Ricoh Theta, LadyBug, or Insta360), also due to the booming "action-camera" market, stereo 360° solutions (e.g., Vuze+) are more costly and limited, and also typically offer only a low number (i.e., six to eight) of different point of views, leading to stereo and stitching artifacts. Moreover, while rotating camera solutions provide more viewpoints, they do not share the same simplicity and flexibility of single-shot instantaneous capture. For this reason, research has concentrated on view synthesis methods that generate stereo contents from a single 360° panorama. However, current methods either require complicated representations or are too heavy to run directly on HMDs and interactive rates (section 4.3).

To overcome these limitations, we propose in this paper a novel approach for quickly and automatically generating and experiencing an omnidirectional stereo representation of an indoor environment starting from a single monoscopic panoramic image in an equirectangular format. In our approach, summarized in Figure 4.1 and section 4.4, we start by estimating full-frame per-pixel depth using a gated deep network designed to exploit interior environment constraints and trained on large sets of synthetic examples (section 4.5). Then, we synthesize panoramic slices through reprojection and view-synthesis using a deep network that shares the same design features and training set of the depth estimation one (section 4.6). These slices are placed around the central viewpoint, on the circle formed by the two eyes during head rotations, and cover an angular portion sufficient to accommodate the potential gaze directions of both the left and right eye. A stereoscopic multiple-center-of-projection stereo pair in equirectangular format is then composed by suitably blending the precomputed slices. The resulting pair is loaded into a WebXR viewer for a lightweight, responsive experience with both motion and stereo cues during runtime (section 4.7). In this approach, based on approximating a full stereo experience through an omnidirectional stereo pair (see section 4.3), the run-time costs are minimized, both in terms of storage and bandwidth and in terms

of rendering performance, at the cost of a slight degradation of stereo reconstruction in the peripheral vision (see section 4.7).

4.2 Contributions

Our main contributions are the following:

- we introduce a novel end-to-end deep network architecture that generates shifted views of an indoor panoramic image in equirectangular format; a first network module estimates a depth map from a single panoramic input; then, these views are reprojected to the desired position, and a full image is synthesized through a second network capable to generate plausible content in disoccluded areas. Unlike other state-of-the-art approaches in the literature [65, 15], the network is based on a lightweight gated architecture and a dilated bottleneck; as a result, we ensure scalability to larger image sizes and/or embedded hardware, while maintaining maximum visual detail when re-projecting onto new views;

- we introduce a unified network architecture with custom training strategies for both depth estimation and view synthesis. The same lightweight network is exploited for both tasks, just adapting the final activation function and changing the training mode. To this end, we introduce a specific photometric loss for novel view synthesis, combined with a GAN approach. As a result, photorealistic novel views are generated with a low computational cost. We moreover use super-resolution GAN-based architectures to increase further the resolution between the stereo images.

- we exploit our depth estimation, reprojection, and synthesis approach to generate a set of panoramic slices and use them to compute an omnidirectional stereo image pair that can be directly experienced on WebXR viewers that sample them to generate stereo couples that respond to head motion with low-latency and high frequency. The limitation to panoramic slices greatly simplifies off-line computational costs in comparison with previous solutions [105], and the direct exploitation of standard omnidirectional stereo formats fosters the applicability of the method to a variety of hardware and software platforms.

Our evaluation (section 4.8) illustrates how depth inference and inpainting networks achieve state-of-the-art performance and how they can be exploited to produce seamless omnidirectional stereo images at a high angular sampling rate. Since the proposed framework is easy to integrate into current panoramic viewers, just replacing the current monoscopic renderers, it promises to be a practical building block for delivering engaging and realistic experiences that captivate audiences and enable them to virtually explore and interact with indoor spaces in current and future Metaverse applications.

4.3 Related work

Our research focuses on creating immersive content using as sole input a single monoscopic panoramic image captured within an interior setting. The presentation of an image with stereo-parallax effects requires synthesizing different views for the left and the right eye. These views should respond to user motion by taking into account that the visibility of scene elements may change even for small shifts of the eye position. This requires not only implicit or explicit geometry estimation to take into account depth-dependent stereo parallax but also the handling of occlusions and disocclusions. In the subsequent discussion, we provide a concise overview of only the most closely-related works. We direct the reader to recent surveys on indoor reconstruction [1], scene understanding from panoramic imaging [106], as well as extraction of 3D geometry from 360° imagery [107] for a more comprehensive coverage of the subject matter.

4.3.1 Depth estimation from a single panorama

State-of-the-art monocular depth estimation solutions involve the adoption of data-driven approaches that extrapolate implicit relationships from extensive labeled datasets, incorporating priors tailored to specific applications, particularly within interior environments [1]. Prior studies have demonstrated that the direct application of perspective methods to 360° depth estimation in indoor settings yields suboptimal outcomes [58]. For this reason, ongoing research directly exploits the wide geometric context inherent in omnidirectional images while addressing wraparounds and distortions characteristic of equirectangular projections [59, 60, 108, 58, 109, 15, 110, 111]. Following this trend, our work introduces a streamlined and lightweight

pipeline directly working on an equirectangular image, introducing an architecture that we also exploit for the view synthesis network.

4.3.2 Novel view synthesis

A panoramic image with an accompanying depth map can be utilized for view synthesis using diverse approaches, such as directly rendering point clouds [74], generating and rendering view-independent meshes from depth maps [112, 103], or integrating and blending depth maps or generated meshes with multiple images or signals [27, 73]. Recently, end-to-end view synthesis networks have been proposed to generate shifted panoramic views at run time [17, 104]. While these networks excel at inferring immersive views within a limited volume around the viewer (e.g., 50cm), their computational demands preclude direct execution on embedded platforms. Consequently, Head-Mounted Displays (HMDs) are exclusively supported via remote rendering [104]. For stereo generation, Pintore et al. [105], proposed, instead, to generate a set of stereo pairs off-line and to perform rendering on the HMD starting from these inferred views through a simple interpolation method. Since per-frame generation is confined to stereo pairs, the complexity of view synthesis networks is significantly reduced compared to more general previous solutions for free-viewpoint synthesis [17, 104]. In this work, we further streamline the method by generating off-line a set of panoramic slices optimized for subsequent blending into an omnidirectional stereo panorama. As a result, we further reduce both the off-line computation and the on-line rendering costs.

4.3.3 View interpolation

The generation of novel views by interpolating images taken at nearby viewpoints has been widely researched, with effective solutions being proposed, even in the absence of a prior depth estimation step [113, 114]. 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). An emerging approach for rapid novel viewpoint synthesis involves employing layered depth representations, associating each pixel with multiple depth values [76]. This methodology has been effectively expanded to operate with single panoramic images [78, 115], as well as to create light field videos through layered mesh representations [5]. For perspective views, multi-plane panoramas (MPI) have also been proposed as an output representation produced with convolutional neural

networks [79, 116]. However, MPIs are limited to viewpoints that are close to the origin and degrade when the viewpoint moves further. To address this limitation, adaptive sampling schemes have been proposed [117]. 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) [4] or multi-cylinder images (MCI) [6]. In contrast, our proposed framework synthesizes 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. Compared to the current state-of-the-art, our approach offers the advantage of being lightweight, both in terms of cost of inferring novel views, since they are constrained to small angular portions of the sphere, and for immersive exploration through WebXR viewers, since rendering has about the same cost of monoscopic viewing. Moreover, the solution is compatible with methods employed for conventional stereo panoramas captured with moving rigs [118]. It should be noted that our view synthesis machinery would also be compatible with a run-time presentation of slices sampled by taking into account per-frame precise per-pixel view directions across the entire field of view. For general head displacements, Pintore et al. [104], for instance, proposed to generate novel panoramas on demand on a server in response to head motion changes. The solution, however, can only update panoramas at around 10Hz and with a latency of about 0.1s. Despite our faster networks, the expected speed-up is less than a factor of two, and, in any case, we would still require a high-speed connection to a fast rendering server. Since we only need to respond to rotation, an alternative solution would be to upload the entire set of slices to cover all possible eye directions. Using the same sampling rate employed in this paper would require, however, the uploading of 360 images, increasing bandwidth and storage requirements by over two orders of magnitude. However, using a lower sampling rate would increase ghosting artifacts [105] when employing simple blending or would require more complex precomputations and blending operations [119].

4.3.4 Omnidirectional stereo display

While 360° surround-view panoramas are limited to only the three rotational degrees of freedom, with the location being fixed, stereo presentation on HMDs requires different images for the left and right eyes to provide the stereo depth cue. Omnidirectional stereo projection, used in this work, is a multiperspective technique [120]

based on circular projection stereo [25] that aims to combine in a single representation 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, recent work has concentrated on generating images that dynamically adapt to the user's gaze, in particular through the view-dependent rendering of depth images [121]. Our solution could also be adapted to those methods, given our capability to infer good depth maps. However, using plain omnidirectional stereo-pairs remains an appealing approach for indoor environments viewed on HMDs, since degradation mostly appears at the poles of the equirectangular image, which generally do not contain suitable content due to typical indoor environment shape and capture constraints, and in the peripheral vision, that also incurs in degradation due to foveation [122].

4.4 Method overview

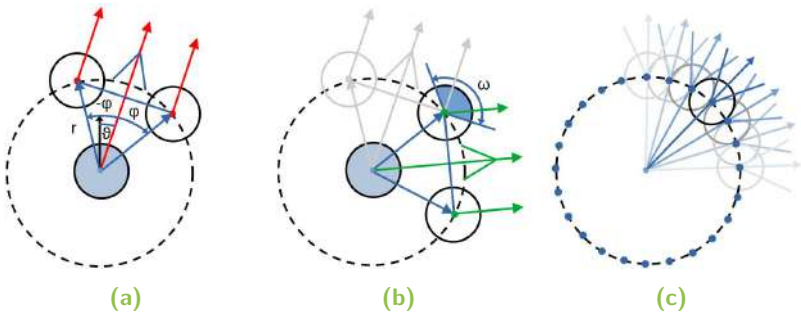


Fig. 4.2.: **Viewing geometry.** (a): we consider that the two eyes are positioned along a circle whose center is at the center of rotation of the head and their central gaze direction is aligned with the head rotation; their position is uniquely determined by the head radius r, the inter-pupillary distance IPD, and by the angular position of the head θ. (b): during head rotation, at a given position, there is an angular slice of size ω that contains both the right and the left gaze direction; the slice's angular size can be solely determined by r and the IPD. (c): to cover all potential gaze positions and directions, we compute all angular slices placed at closely spaced positions on the circle.

Our method automatically and rapidly converts a single monoscopic panoramic image in equirectangular format into an omnidirectional stereo panorama, also in equirectangular format, that can be rapidly explored with stereo and motion parallax on an HMD.

As depicted in Figure 4.2a, we consider that during exploration, the two eyes will be on a circular trajectory centered at the head's rotation center. Thus, their specific positions are defined by the head radius (r), the inter-pupillary distance (IPD), and the angular position of the head (θ). Without loss of generality, the central gaze direction of both eyes is considered in this paper aligned with the head rotation, even though the method can be easily adapted to other gaze directions (see section 4.7).

Given this geometric configuration, during head rotation, any given position on the circle may thus become the center of projection for the left or the right eye. Thus, see Figure 4.2b, from this point of view, there is a constant angular slice of size ω that is guaranteed to contain both the right and the left gaze direction. The angular size of such a slice can be solely determined by r and the IPD (see section 4.7). As shown in Figure 4.2c, it is thus sufficient to compute all angular slices placed at closely spaced positions on the circle. We exploit this geometric configuration to define an efficient approach to synthesize all these views and combine them into an omnidirectional stereo panorama. The first step of our method is to estimate the per-pixel depth of the input panoramic image that we assume is placed at the head center. This depth is computed in a single step by a gated deep network designed to exploit interior environment constraints and trained on large sets of synthetic examples, as detailed in section 4.5. Given this depth and the original color, we synthesize each of the required shifted panoramic slices. For each of these slices, we start by reprojecting the original image into the required slice, defining a bounded vertical section of a panoramic image in an equirectangular format, using as a center of projection the relevant eye position. View-synthesis is performed using a deep network that shares the same design features and training set of the depth estimation one, as detailed in section 4.6. Finally, an omnidirectional stereoscopic image pair in equirectangular format is composed by suitably blending the precomputed slices and used for display in a lightweight WebXR viewer, producing images suitable for HMD consumption. (section 4.7).

In the following, details are provided for each of the individual components.

4.5 Single panorama depth estimation

Augmenting a single image with depth is essential to establish the 3D position of visible points in space to compute their novel position when the viewpoint changes.

Many techniques have been documented in the literature to estimate depth from a single panoramic image (section 4.3). Given the inherent ambiguity in depth estimation from single images, all approaches necessitate leveraging prior information to steer the reconstruction towards plausible architectural forms that align with the input. Notably, there has been a remarkable advancement in data-driven methods within this context, wherein these methods acquire knowledge of such priors from large collections of labeled exemplar data [15, 110]. Following this research trend, we designed a network for depth prediction that was an efficient compromise between accuracy and computational cost, and with an architecture that can be reused for the view synthesis part (section 4.6). Using a lean and scalable network design is also important to support, in the future, larger and larger image sizesand to provide a low latency from the acquisition time to the presentation time, especially when using low-end machines.

To predict depth we designed a gated architecture, which encoder-decoder scheme follows the same design adopted for view-synthesis (see section 4.6), but with several differences to adapt it to the specific task of spherical depth estimation. In particular, here, gating acts as a *self-attention weight mask*, differently from inpainting, where, instead, the mask is given as input to indicate the pixels to be inpainted (section 4.6). Moreover, given the spherical nature of the input, we adopt circular padding along the horizon for convolutions, thus removing longitudinal boundary discontinuity and reflection padding to alleviate the singularities at the poles [32]. Furthermore, considering that our output will be a single channel, we use 32 internal channels instead of the default 64 channels in standard inpainting networks. Finally, since we produce depth, the last layer activation function is an ELU, instead of $tanh$.

The input equirectangular image, which is encoded through a sequence of light-weight gated convolutions having different strides, so that the original size is reduced by a factor of four in each direction. Each encoding convolution is followed by instance normalization [102] and ReLU activation. Generally, compared to view-synthesis baselines [86, 87], our design has fewer parameters, with a lighter single branch, and it includes several solutions, described below, to improve accuracy for the depth estimation task and reduce computational complexity.

The adopted gated convolution (GC) approach [33] is expressed as:

$$\begin{aligned} G &= conv(W_g, I) \\ F &= conv(W_f, I) \\ O &= \sigma(G) \bigodot \psi(F) \end{aligned} \tag{4.1}$$

where σ is the Sigmoid function, which outputs values in the range $[0, 1]$, ψ is an activation function (ReLU in our case), and W_g and W_f are two different sets of convolutional filters, which are used to compute the gates and features respectively. GC enables the network to learn a dynamic feature selection mechanism. In order to simplify training and guarantee low latency at inference time, our network uses a modified version of GC called Light Weight Gated Convolutions (LWGC), which reduces the number of parameters and processing time while maintaining the effectiveness [89]. Specifically, we decompose G from Equation 5.3 into a depth-wise convolution [89] (i.e., 3×3) followed by a 1×1 convolution, having, as a result, the same gating step but with only $k_h \times k_w \times C_{in} + C_{in} \times C_{out}$ parameters. Repeated dilations [88] are used for the bottleneck, thus increasing the area that each layer can use as input. It should be noted that this is done without increasing the number of learnable weights but obtained by spreading the convolution kernel across the input map. The *dilated convolution operator* is then implemented as a gated convolution (i.e., Equation 5.3), but with some differences. It is expressed as:

$$D_{y,x} = \sigma(b + \sum_{i=-k'_h}^{k'_h} \sum_{j=-k'_w}^{k'_w} W_{k'_h+i,k'_w+j} \cdot I_{y+\eta i, x+\eta j}) \tag{4.2}$$

where η is a dilation factor, $\sigma(\dot{})$ is a component-wise non-linear transfer function and $b \in \mathbb{R}^{C_{out}}$ is the layer bias vector. With $\eta = 1$, the equation becomes the standard convolution operation. In our model, we adopt, respectively, $\eta = 2, 4, 8, 16$ for the four bottleneck layers. Using this strategy, we aggregate multi-scale contextual information without losing resolution, thus capturing the global context efficiently by expanding the receptive field, avoiding additional parameters, and preventing information loss. This is important for both depth estimation and the image completion task (section 4.6), as capturing sufficient context is critical for realism. By using dilated convolutions at lower resolutions, the model can effectively cover a larger area of the input image when computing each output pixel than with standard convolutional layers [86]. The network decoder, based on gated convolutions without dilation, restores the resolution of the output to the original input resolution.

The effectiveness of such a versatile baseline also depends on its training. In our approach, we adopt as a loss function for the depth prediction task the robust *Adaptive Reverse Huber Loss (BerHu)* [90], combined with a Structural Similarity Index Measure (SSIM), which measures the preservation of highly structured signals with strong neighborhood dependencies. As a result, such a panoramic depth prediction approach returns accurate depth maps for the input pose, as demonstrated by our results (section 4.8).

Fig. 4.3.: **Panoramic slice** One of the slices produced by the network, placed at its position within the equirectangular image. The red line shows the area sampled by the right eye, while the green line shows the area sampled by the left eye (see Figure 4.2b). In the background, we see the original panorama from the central viewpoint (note the large shift due to parallax effects).

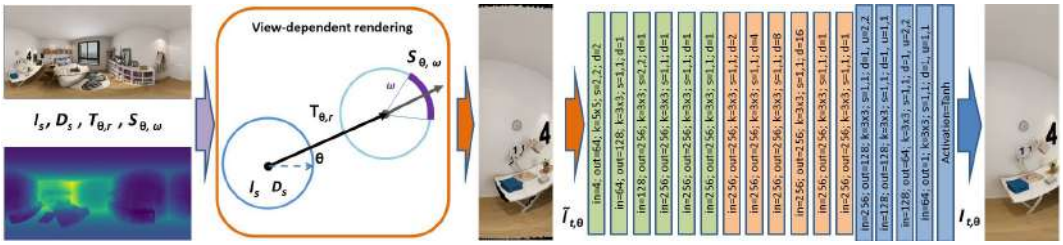


Fig. 4.4.: **Novel view synthesis.** Given the source image I_s and its predicted depth D_s, a new, sliced viewport $\widetilde{I_{t,\theta}}$ is rendered from a new viewpoint translated by $T_{\theta,r}$, according to a given direction θ and an offset r. The field-of-view ω, designed to cover both eyes' viewport (section 4.7), is assumed constant. To generate an $I_{t,\theta}$ slice, we exploit the gated architecture already exploited for depth but adapted to have greater accuracy on the 3 RGB channels of the spherical section S. Legend: *in,out* channels; *k* convolution kernel; *s* stride; *u* upsample; *d* dilation.

4.6 Synthesis of novel views

Taking as input the original panoramic image and the registered panoramic depth map estimated by our deep network, this task aims to synthesize a collection of panoramic slices through re-projection and view synthesis. These slices are distributed around the central viewpoint, with each slice's projection center placed on the circular path covered by each eye during a head rotation. Furthermore, each slice encompasses an angular extent sufficient to accommodate the potential gaze directions of both the left and right eye (section 4.7).

To this end, given a full angular extent of 180 degrees along the vertical direction and a limited viewpoint ω along the horizon, we generate novel spherical images in viewports $S_{\theta,\omega}$ (i.e., *slices*) along the circular path, by translating the original, central view I_s by an offset $T_{\theta,r}$ for each direction θ and distance r from the input view. For a typical human-size configuration, with a head radius of 100mm and an IPD of 65mm, we can safely assume 45 degrees as a portion of the image covering both eyes, which would correspond to 128 pixels for a 1024×512 equirectangular image. We further expand this region by ± 5 degrees to provide context for reconstructing missing areas and to support eye convergence at a finite distance. For a 1024×512 image (corresponding to a 512×1024 tensor), each slice in our experiments is assumed to be 160 pixels wide, while the overall angle ω is about 56 degrees. An example of a slice produced by the network is depicted in Figure 4.3. The View-synthesis pipeline

Fig. 4.5.: **Reprojected vs. synthesized image** Two examples from different scenes. On the left, we see a detail of a reprojected image slice, where disoccluded areas are apparent. On the right, we see the output of the view synthesis network.

is depicted in Figure 4.4 and includes two steps. The first is a view-dependent rendering step, which exploits the predicted depth D_s and translation $T_{\theta,r}$ to move pixel information to the new position. T_θ is given by polar coordinates, which depend on the head radius r (i.e., assuming in our experiment an average radius of $100mm$)

and by the angle θ, while $\widetilde{I_{t,\theta}}$ is the portion of translated pixels viewed from the viewport $S_{\theta,\omega}$. The second step consists in a view-synthesis deep network, which takes as input the translated pixels $\widetilde{I_{t,\theta}}$ and their disocclusion mask B_t (i.e, black pixels in Figure 4.4), returning as output the novel viewport $I_{t,\theta}$. Figure 4.5 shows, for two different scenes, a detail of a reprojected slice, with missing pixels in black, and the corresponding area of the output of the view-synthesis network.

In our case, by design, pixel rendering is not part of the learnable layers, and we assume we can directly project visible points according to D_s depth and T_θ translation using regular z-buffering to obtain the starting view to be optimized. This solution is better suited to our case than more elaborate splatting methods [84], since for stereo rendering, the limited displacement of the eyes from the center generates much narrower disocclusion zones than in the case of free viewpoint motion.

In an equirectangular image, columns correspond to constant longitude/azimuth θ angles, while rows to constant latitude/elevation ϕ angles. Each pixel can be mapped to angular spherical coordinates and vice-versa. This mapping between image domain pixels and spherical domain angular coordinates allows for direct transitions between image-based and spherical-based operations [93]. Omitting the straightforward relationship between Cartesian and spherical coordinates, the following equation relates spatial (i.e, $T_{\theta,r}$) with angular displacements (i.e., $\widetilde{I_{t,\theta}}$ pixels):

$$\begin{bmatrix} \partial d \\ \partial \phi \\ \partial \theta \end{bmatrix} = \begin{bmatrix} \sin(\phi)\sin(\theta) & \cos(\theta) & \cos(\phi)\sin(\theta) \\ \frac{\cos(\phi)}{d\sin(\theta)} & 0 & \frac{-\sin(\phi)}{d\sin(\theta)} \\ \frac{\sin(\phi)\cos(\theta)}{d} & \frac{-\sin(\theta)}{d} & \frac{\cos(\phi)\cos(\theta)}{d} \end{bmatrix} \begin{bmatrix} \partial x \\ \partial y \\ \partial z \end{bmatrix} \tag{4.3}$$

where d is the depth of the given pixel.

For the view-synthesis task, we assume $\widetilde{I_{t,\theta}}^{3\times h\times w}$ as input. As in typical inpainting approaches, we define a binary inpainting mask ${B_t}^{1\times h\times w}$, identifying missing parts in the rendered image. This mask is computed directly in the reprojection step. B_t is then concatenated to $\widetilde{I_t}$ (i.e., along the batch dimension - 4 layers input (Figure 4.4).

To predict the output $I_{t,\theta}$ slice, we adopt the lightweight gated architecture exploited for depth estimation (section 4.5) but adapted for having greater accuracy on the RGB channels of the current spherical viewport $S_{\theta,\omega}$. Here, spherical padding is replaced by replicate padding in all layers. Similarly to other works (e.g., DeepFillV2 [33]), we use $f(x) = \max(0, \tanh(x))$ as activation function for the output layer. Limited to

the [0..1] range, this function behaves similarly to *ReLU* near the lower bound while smoothly saturating at the upper bound.

As shown in Figure 4.4, this network has a higher density at the inner channel level, whose starting value is 64 (first encoder layer in Figure 4.4). This increase in layers, compared to the configuration used for depth, is compensated, from the computational point of view, by the fact that the network processes a smaller portion of the image than the full equirectangular image, leading to a contained computational cost, as demonstrated in section 4.8. This is particularly important since, for each input panorama, the generation of omnidirectional stereo representation requires the generation of hundreds of slices.

We train the inpainting network by including losses that measure the photorealistic quality of the output slice. It should be noted that, in contrast to full image prediction, here the loss is calculated by comparing the predicted slice $I_{t,\theta}$ with the corresponding crop $I_{gt,\theta}$ of the ground truth equirectangular image. Our loss function is expressed as:

$$\mathcal{L}_{vis} = \lambda_{px}\mathcal{L}_{px} + \lambda_{perc}\mathcal{L}_{perc} + \lambda_{style}\mathcal{L}_{style} + \lambda_{adv}\mathcal{L}_{adv} - \lambda_{lpips}\mathcal{L}_{lpips}. \tag{4.4}$$

where the first term is a pixel-based $L1$ loss between the predicted RGB slice $I_{t,\theta}$ and the ground truth target crop $I_{gt,\theta}$, $\mathcal{L}_{perc}$ and $\mathcal{L}_{style}$ are the data-driven perceptual and style losses [50], enforcing I_{out} and I_{gt} to have a similar representation in the feature space as computed by a pre-trained $VGG-19$ [91], while $\mathcal{L}_{adv}$ is a discriminator-based loss (i.e., PatchGAN [92]). Furthermore, in addition to conventional inpainting losses, we introduce a loss based on Learned Perceptual Image Patch Similarity (LPIPS) [100] to enforce similarity due to the restricted field-of-view of the slice. λ weights are common for many single-pose inpainting problems [87]: $\lambda_{px} = 1.0,\ \lambda_{style} = 100.0,\ \lambda_{perc} = 1.0,\ \lambda_{adv} = 0.2,\ \lambda_{lpips} = 1.0$.

4.7 Omnidirectional stereo generation and rendering

Starting from the slices synthesized by the network, we have all the information to provide stereoscopic viewing during head rotations, as these slices contain a plausible scene reconstruction for all the possible points of view of both the left and the right eye. While previous works used a small set of these synthesized images

and combined them at rendering time [105], here we densely sample the position space and construct off line a compact omnidirectional stereoscopic representation by appropriately fusing these slices. The issue of constructing stereoscopic panoramic image pairs has already been addressed in the literature (section 4.3). In this work, we selected to achieve the stereoscopic effect by generating two aligned multiple-center-of-projection (MCOP) images encoded in equirectangular format. As longitude varies in these images, the corresponding pixel column is generated from a different camera position, which corresponds to the position of the eye when it is looking straight in this direction.

The calculation for each eye starts, thus, from the generation of n vertical slices radially uniformly distributed around the head, as shown in Figure 4.2c. Since our networks are computationally efficient, and all the computation is performed offline, we can generate very dense angular samplings in a short time (ideally, even with n equal to the output image width). In practice, we have seen that an angular sampling of one degree (i.e., 360 slices) is sufficient to obtain a very high-quality reconstruction. The single MCOP image for each eye is constructed from the union of vertical slices associated with each θ angle of longitude. The stereo effect is ensured by the fact that, for a given rotation θ of the head, the eyes will point in the same front-facing direction, but each with a slightly different perspective due to the offset caused by the inter-pupillary distance IPD. As seen in Figure 4.2a, when we rotate the head around its vertical axis, for a given θ, the eyes will be positioned on the circle of radius r at angles $\theta - \phi$ (left eye) and $\theta + \phi$ (right eye), where the value of ϕ is given by $\phi = asin(\frac{IPD}{2r})$. When computing the omnidirectional stereo representation for the left eye, we thus loop over all the output columns. For each vertical column at an angle θ, we identify the eye position as $\theta - \phi$ and find the two synthesized panoramic slices with the closest centers of projection (i.e., one to the left and one to the right). The pixels of these slices are then blended with a Gaussian weight based on the angular distance to the output column. The same process is done for the right eye, with the only difference being that the eye position is $\theta + \phi$. Since we are using a very high angular sampling rate in this paper to place slices, i.e., 360 slices per image, the blending area is extremely small – pixel-sized for our typical network outputs, and using such a simple blending does not lead to any noticeable ghosting artifacts, as illustrated in section 4.8.

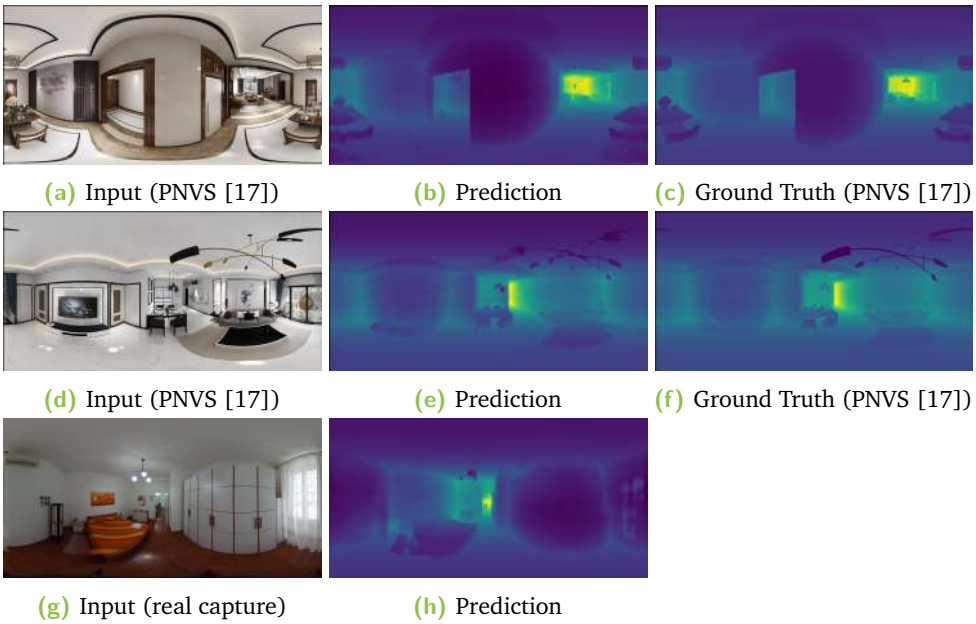
Note that while we have assumed here a view with both eyes looking in the same direction (zero parallax at infinity), we can apply the same approach for calculating MCOP images with eyes that have zero parallax at a finite distance for improved

simulated stereoscopic vision in confined environments. The only variation would be in the extraction of the column, which would not be in the θ direction but towards the focal point.

As a final pre-processing step, we also perform upsampling of the images to match the quality 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 easily scalable network architectures (see section 4.5 and section 4.6), but, rather, to limitations in available ground-truth training sets. In the current work, good quality results are obtained by applying available super-resolution generative adversarial networks capable of zooming images by creating plausible geometric and texture detail. To apply these methods to equirectangular images without boundary effects, we extend, before zooming, the original image to the left by incorporating portions from the right side and vice versa. This extension makes available to the network a sufficient context to define all important areas. After zooming, the image is then cropped only to contain the relevant portion of the equirectangular representation. The two MCOP images resulting from this process can then be presented to the viewer using the same approach used for regular panoramas, presenting the left panorama to the left eye and the right panorama to the right eye, and using the same viewing transformation for both panoramas to adjust for head orientation. When looking in a specific viewing direction, the correct perspective for the left and right eyes will be projected in the headset, with the correct horizontal parallax for the front-facing pixels and a small degradation towards the periphery. The stereoscopic effect of the two images calculated in this way is guaranteed by the fact that human stereoscopic vision is concentrated mostly in the central portion of the field of view and does not exist in the outer peripheral zones.

The rendering on the headset is accomplished using a standard viewer for omnidirectional stereo images through a WebXR API browser. In practice, the high-resolution stereo panoramic images serve as textures for two spheres positioned in the scene — one centered around the left eye and the other around the right eye. The result is a kind of distinct environment map for each eye. As per the WebXR specifications, when XR rendering is enabled, the system retrieves parameters defining the head's position from the headset's sensors during each frame of the animation loop. These parameters are then used to create the correct perspective projections for each eye. For a given longitude θ, the left and right eyes will centrally align with images with

the correct horizontal parallax disparity, thereby providing the effect of stereoscopic perception.



(a) Input (PNVS [17]) (b) Prediction (c) Ground Truth (PNVS [17])

(d) Input (PNVS [17]) (e) Prediction (f) Ground Truth (PNVS [17])

(g) Input (real capture) (h) Prediction

Fig. 4.6.: **Depth estimation.** Two examples of depth prediction on PNVS [17] dataset scenes and an example of depth prediction from a user-acquired panoramic scene taken with a Ricoh Theta 360° camera.

4.8 Results

The processing components to obtain stereo panoramic images for loading onto the headset have been developed in Python, using Pytorch to implement our custom networks, combined with standard image processing and computation libraries (NumPy, Pillow, OpenCV). The generative adversarial network used for zoom operations is *Real-ESRGAN* (with model *realesr-animevideov3*) [80], that has been directly integrated as a post-processing step in our system. The immersive rendering components, instead, have been realized in WebGL and WebXR.

4.8.1 Dataset and training

For training our solutions, we harness the availability of public panoramic scene datasets where ground truth is available. To train and test depth estimation, we exploit Structured3D [52]), a large-scale (21K photorealistic scenes) synthetic database of indoor scenes providing the ground truth depth for each panoramic image.

(a) Central pano (PNVS [17]) (b) Central pano (PNVS [17]) (c) Central pano (PNVS [17])

(d) Left/right MCOP pano (e) Left/right MCOP pano (f) Left/right MCOP pano

Fig. 4.7.: Ominidirectional stereo panoramas. Three representative scenes from the PNVS [17] dataset (testing split). Top row: source panorama. Middle row: automatically generated MCOP panorama for the left eye; Bottom row: automatically generated MCOP panorama for the right eye. The vertical alignment clearly shows the parallax effects.

To train and test view synthesis, instead, we exploit PNVS [17], a subset of Structured3D scenes providing several translated views for each source panoramic image. Since the baseline for stereo view generation is very small, we opt for the PNVS subset known as *easy*, characterized by a maximum range of 300mm. This range comfortably exceeds our default radius of $r = 100$mm. Train and test splits are maintained as in the original papers. The depth estimation network is trained and tested directly on the original images, while the view synthesis network is trained and tested on randomly oriented slices of the provided examples. The generation of randomly oriented slices is implemented as a data augmentation step. Given that the available training and testing data sets are provided at a resolution of 1024×512 pixels, all our processing is done at that size. In the future, we plan to generate synthetic training and testing sets at higher resolution (2K-4K), exploiting the scalability of our networks to directly process images at typical native 360° camera resolution, removing the need for the downscaling and upscaling steps. We trained both networks with the Adam optimizer [95], 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 for depth estimation and 4 for view synthesis. The average training time for the depth estimation network is $150ms/image$, while for view synthesis is $160ms/image$.

4.8.2 Computational performance

Our depth estimation and view synthesis baselines are extremely lightweight. Table 5.2 shows learnable parameters, GFlops, and milliseconds for different tasks and outputs. The benchmarks have been made on the same A5000 machine used for training.

In all presented tasks, we assume 512×1024 as the source image tensor resolution. Indeed, the output resolution is the same for depth estimation, where the network configuration uses 32 internal channels (section 4.5). For the view synthesis task, instead, we compare computational stats to generate a full 360° image (i.e., 512×1024 tensor size) and to generate a slice (i.e., 512×160) that is the final utilization of the network. For this task, we adopt a network configuration with 64 internal channels, as well as other task-designed modifications (section 4.6). The results clearly show the computational advantage in terms of GFlops and inference time. Subsequent results

Tab. 4.1.: **Computational performance.** We show the computational performance and latency time of our gated architecture for different tasks. In bold modes are the current architecture choices.

Mode	Output Res	Params	GFLOPS	ms/frame
Depth	512×1024	6.06 M	164.11	41
Synth	512×1024	6.93 M	326.71	95
Synth sliced	512×160	6.93 M	51.05	58

(section 4.8.3) show that the choice of generating a slice maintains a performance advantage not only in computational terms. With the current approach, a full-quality omnidirectional stereo image, computed with an angular spacing of 1° between slice projection centers, can be generated and experienced in stereo with a latency of only about 21s from the time of the shot.

4.8.3 View-synthesis performance

As depth estimation is a fundamental task to achieve novel view synthesis, Table 4.2 presents the quantitative performance of our gated architecture compared to state-of-the-art panoramic depth solutions. We included in the evaluation the same error metrics used in many prior depth estimation works (e.g., [15, 65]): mean absolute error (MAE), mean relative error (MRE), root mean square error of linear measures (RMSE), and three relative accuracy measures δ_1, δ_2 and δ_3, defined, for an accuracy

δ_n, as the fraction of pixels where the relative error is within a threshold of 1.25^n. The latter measures are useful to illustrate the error distribution.

We compare our performance with SliceNet [15] and HoHoNet [65], which are state-of-the-art methods commonly used as benchmarks in the latest panoramic works [110, 66], and, particularly for the domain of this paper, can provide performance with low latency. In this case, we show the reconstruction performance on the main

Tab. 4.2.: Quantitative performance comparison on depth reconstruction. Our results are compared to other state-of-the-art works.

Method	MAE↓	MSE↓	RMSE↑	δ_1 ↑	δ_2 ↑	δ_3 ↑
HoHoNet [65]	0.081	0.065	0.206	0.958	0.987	0.993
SliceNet [15]	0.082	0.054	0.198	0.961	0.988	0.993
Our	0.061	0.008	0.038	0.962	0.989	0.994

Structured3D [52] test set, for which results from those baselines are available. We show some qualitative results in Figure 4.6 (top two rows) on the PNVS scenes adopted instead for view-synthesis benchmarking. The deformation in the views is present in the original images and is due to the equirectangular projection, which preserves horizontal lines and curves vertical ones. The same deformation is visible in all other images in equirectangular format included in this article. As an illustration of how the same model can be applied to casually captured images, the bottom row of Figure 4.6 shows how our network successfully predicts the depth of a user-acquired scene captured with a hand-held Ricoh Theta 360° camera.

As shown in the qualitative results of Figure 4.6, the overall shape of the room is well preserved, and, similarly to other works [65, 15] the main prediction errors appear on very thin structure (e.g., the lamp in the second row). These thin structures are not very well resolved and, at run time, can cause visual artifacts during exploration. This problem is common to virtually all depth estimators from single images, and we expect to reduce them by increasing the resolution of images in the training set.

Table 4.3 summarizes our performance in terms of view synthesis accuracy, benchmarked on the PNVS [17] test dataset. . Despite presenting much more challenging translations than stereo parallax, this set provides a ground truth on which it is possible to compare with other state-of-the-art methods [19, 57] and among different versions of our architecture.

The results show that our method outperforms other baselines in generating a full equirectangular view (i.e., row 3). Furthermore, we show how reconstructing the

Tab. 4.3.: **View synthesis performance.** We show the quantitative performance of view synthesis compared to other state-of-the-art methods, all of which operate at a minimum resolution of 1024×512 image size (i.e., 512×1024 tensor size). The last line shows our sliced solution compared with our baseline trained to reconstruct the whole image.

Method	Output Res	PSNR↑	SSIM↑	LPIPS↓
SynSin [57]	512×1024	17.28	0.721	0.226
MPI [19]	512×1024	17.59	0.725	0.223
Our full	512×1024	21.55	0.731	0.202
Our full crop	512×160	22.77	0.738	0.196
Our sliced	512×160	23.00	0.744	0.186
Our-sliced-no-LPIPS	512×160	22.38	0.748	0.205

single slice still achieves state-of-the-art performance even though the reconstruction is done by having a smaller context (i.e., 56 degrees vs. 360 degrees). The standard deviation of the error measures on the sliced version amounts to 0.075 for LPIPS, 0.098 for SSIM, and 4.53 for PSNR, and is in very similar ranges for the other versions of the network. We noticed that the main errors found on the set of scenes are due to the imprecise depth reconstruction (see above), especially on thin structures. We thus identify depth estimation as one of the main avenues for improvement.

The last row shows the results obtained with an instantiation of our network trained without the $\mathcal{L}_{lpips}$ loss term. It should be noted how adding this term not only improves the LIPS metric but also has a beneficial effect on the PSNR. Since LPIPS strongly correlates with perceptual quality [100], its addition in the loss improves the final quality of presented images.

To compare the accuracy of the reconstruction, we measured performance by comparing random 512×160 crops on the full generated overview (i.e., training with 512×1024 tensor output), with slices generated with the dedicated network (i.e., training with 512×160). The experiments show that despite the significantly lower computational complexity, performance is on par, if not better, than generating a full equirectangular image for each angle.

4.8.4 Stereoscopic exploration on HMD

We tested the immersive viewer on various devices, including a Meta Quest 2 and an Android mobile device Samsung Galaxy S 22 with a Google Cardboard. Here we report on experiments made on a Pico4, a headset with two 2.56-inch Fast-LCD displays, a global resolution of 4320x2160 pixels (equivalent to 2160x2160 pixels

Fig. 4.8.: **The WebXR viewer.** The user on the left wears a Pico4 HMD. The images to the right present the left and right images, as rendered by our WebXR viewer running on the PicoBrowser. The source image is a single-shot monoscopic 360° capture of a real environment, transformed to omnidirectional stereo by our framework.

(a) Left MCOP panoramas (b) Central panorama (c) Right MCOP panoramas

Fig. 4.9.: **Ominidirectional stereo panoramas.** Example from real-world capture. Central: the source panorama captured with a Ricoh Theta. Left/Right: automatically generated MCOP panorama for the left and right eyes.

per screen), a pixel density (PPI) of 1200, a variable refresh rate ranging between 72 and 90 Hz, and a diagonal Field of View (FOV) of 105 degrees (diagonal).

The web application for rendering is served by a web server that only has to transmit the two panoramic images to the HMD, since the embedded client performs all the rest of the computation and rendering work. The client application is built on the ThreeJS framework, enabling the development of WebGL graphics components, and incorporates mechanisms for interaction with WebXR APIs. On the client side, the application, written in JavaScript ECMAScript 6 following a modular approach, is run on the native PicoBrowser when using Pico4, but the viewer can naturally be run on any headset compatible with WebXR specifications, as demonstrated by our tests on Android Phones with Google Cardboard and on the Meta Quest 2. Figure 4.8 shows an image of the viewer. Other standard viewers supporting the omnidirectional stereo format can also be employed. Using a custom viewer allows us to implement specific operations (i.e., switching among scenes or from mono to stereo, or constraining/freeing the up vector during navigation).

When displayed on the HMD, the images provide immersive stereo cues, as also confirmed by an informal test with ten subjects who were requested to explore the

Fig. 4.10.: Comparison of omnidirectional stereo approximation with ground truth. The top portion of the two images shows the perspective generated using the multiple-center-of-projection image for the left and the right eye, while the bottom portion shows the ground truth image generated with a center of projection placed at the eye position. As we can see, the perspective is indistinguishable at the center but slowly degrades in the periphery.

stereoscopic environment on Oculus Quest and provide their opinion on immersion and stereoscopic perception of the generated scenes. The test was simply carried out by loading either the original 360° version or the omnidirectional stereo one. In all cases, users always differentiated the mono and stereo versions and confirmed that the stereo one was providing a more immersive experience.

Figure 4.7 shows the results of omnidirectional stereo generation for three representative scenes from the PNVS [17] dataset (testing split), while Figure 4.9 shows the results of omnidirectional stereo generation for a real-world captured-scene. In both figures, the top row shows the source panorama, positioned at the center of the head, while the middle and bottom rows show the generated MCOP panoramas for the left and right eye, which incorporate stereo parallax effects.

Figure 4.10 shows a comparison between the real-time rendering obtained from the omnidirectional stereo representation and a ground truth image. The top portion of the two images shows the perspective generated using the multiple-center-of-projection image for the left and the right eye, while the bottom portion shows the ground truth view, obtained by placing the center of projection placed at the eye position. As we can see, the two perspectives are indistinguishable at the center but slowly diverge when moving towards the periphery. We noticed that, while this effect is not perceived by the users for most of the scenes, in cases where a strong parallax exists (very nearby objects with details), the users perceive an effect of slight object

motion when rotating the head while still being capable to perceive the parallax. This motion effect is due to the motion of the center of projection across the image. This is, however, an effect perceivable in all multiple-center-of-projection methods and is not introduced by our approach, which aims to present ways for the automated generation of those representations.

4.9 Conclusions

In this chapter we presented a framework for the automatic generation of omnidirectional stereoscopic indoor environments to be used in immersive applications, especially consumed through head-mounted displays. Our method starts from a single panoramic image of an interior environment and uses data-driven architectures for depth estimation and novel view synthesis to quickly generate the images seen by both eyes during head rotation. For this work, these images are combined into an omnidirectional stereo representation, which is consumed on a lightweight WebXR viewer supporting stereoscopic exploration during head rotations. The preliminary results show that the automatic generation components achieve state-of-the-art accuracy, and the visualization component can provide an immersive experience to casual users on a variety of devices. As a result, we can provide a quick method to enhance the exploration of environments acquired with the increasingly ubiquitous and affordable monoscopic panoramic cameras. One of the limitations of the current approach stems from the mismatch between the resolution of the synthesized images and the achievable resolution with nowadays cameras and displays. This mismatch is currently handled by downsampling images before construction and a deep-learning-assisted upsampling before display presentation. The limitation is not due to the lightweight network architecture, which promises to be scalable to much larger image sizes, but instead to the availability of training sets for the depth estimation and view synthesis networks. We plan to tackle this problem by generating higher-resolution training data. In terms of display, we have taken the approach of generating omnidirectional stereo images, which have the major advantage of requiring very limited rendering resources but also introduce a little degradation in the peripheral areas and when the view direction converges towards the poles. Since we have depth available, we can easily improve the method by incorporating state-of-the-art depth-dependent adaptations that have been designed for real captures [121]. In this context, it will be interesting to explore how our deep-learning-based solutions could be further adapted to directly produce the data required for depth-dependent adaptation. We will also evaluate the possibility of exploiting this approach to support a limited

amount of horizontal and vertical head motion in addition to rotation, exploiting the fact that our networks can synthesize arbitrarily displaced images. Finally, we plan to use this panoramic capture and immersive rendering system as a building block for constructing applications that perform actions in shared physical and digital realities. One important direction of work will be to exploit these explorable panoramic environments to serve as interfaces for digital twins of buildings constructed from casually captured real data, that can provide location awareness and be easily annotated in a VR interface.

4.10 Bibliographic notes

This chapter is based on an invited extended version of the ACM Web3D 2023 contribution [105]. In addition to providing a much more thorough exposition, we introduce very significant new material, in particular concerning a full redesign of the view sampling aspects, the exploitation of panoramic slices for the construction of a much more effective view synthesis network, and the off-line computation of omnidirectional stereo panorama in place of the run-time blending of few stereo couples. The contents of this chapter report the journal article *Giovanni Pintore, Alberto Jaspe-Villanueva, Markus Hadwiger, Jens Schneider, Marco Agus, Fabio Marton, Fabio Bettio, Enrico Gobbetti: Deep synthesis and exploration of omnidirectional stereoscopic environments from a single surround-view panoramic image* [39], of which the candidate is the lead author.

GP, FM, EG acknowledge the contribution of the Italian National Research Center in High Performance Computing, Big Data and Quantum Computing. FB acknowledges the contribution of Sardinian regional authorities under the XDATA project. MA 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.

5

Automatic-assisted editing of immersive indoor models

In this chapter we address the specific objective 3, related to immersive editing application from panoramic imagery. Nowadays 360° cameras are increasingly being used in a variety of Extended Reality (XR) applications that require specific Diminished Reality (DR) techniques to conceal selected classes of objects. In this chapter, we present a new data-driven approach that, from an input 360° image of a furnished indoor space automatically returns, with very low latency, an omnidirectional photo-realistic view and architecturally plausible depth of the same scene emptied of all clutter. Such an approach addressed the research problems of objective 3. Contrary to recent data-driven inpainting methods that remove single user-defined objects based on their semantics, our approach is holistically applied to the entire scene, and is capable to separate the clutter from the architectural structure in a single step. By exploiting peculiar geometric features of the indoor environment, we shift the major computational load on the training phase and having an extremely lightweight network at prediction time. This end-to-end approach starts by calculating an attention mask of the clutter in the image based on the geometric difference between full and empty scene. This mask is then propagated through gated convolutions that drive the generation of the output image and its depth. Returning the depth of the resulting structure allows us to exploit, during supervised training, geometric losses of different orders, including robust pixel-wise geometric losses and high-order 3D constraints typical of indoor structures. The experimental results demonstrate that our method provides interactive performance and outperforms current state-of-the-art solutions in prediction accuracy on available commonly used indoor panoramic benchmarks. In addition, our method presents consistent quality results even for scenes captured in the wild and for data for which there is no ground truth to support supervised training.

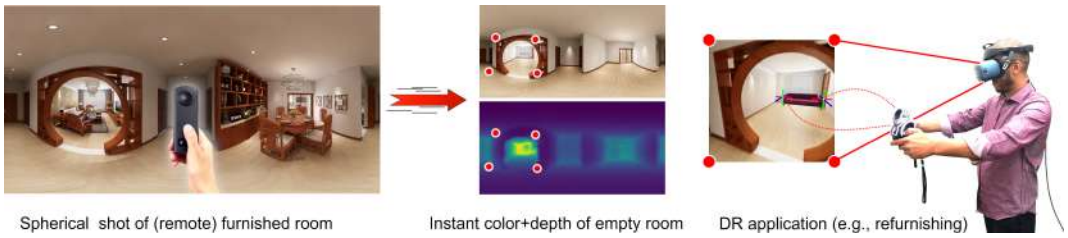


Fig. 5.1.: Given a 360 panoramic photo of a cluttered indoor scene, our end-to-end approach automatically returns a photorealistic view and depth of same scene emptied of furniture and clutter. Both visual appearance and depth, estimated at interactive speed, are highly suitable for compelling and immersive XR applications, such as (re-)furnishing or planning of interior spaces.

5.1 Introduction

Current 360° cameras offering viable low-cost and energy-efficient solutions for full-context single-shot capture are increasingly popular in many application fields [10]. Since the captured 360° content, also known as *panoramic*, *spherical*, or *omnidirectional* imagery, covers the entire sphere around the viewer, even a single shot cannot be statically experienced at once, making it fundamentally different, more immersive and more dynamic, than traditional 2D imagery [16]. In particular, when consumed through Head-Mounted-Displays (HMDs), the user actively focuses on the desired content via natural head movements, just like humans do in real world, achieving a very high degree of immersion [11]. For this reason, omnidirectional imagery is becoming a fundamental component for creating immersive content from real-world scenes, and for supporting a variety of Virtual Reality (VR) applications [3]. Notably, virtual tours based on spherical images are extremely popular in the real estate domain, and have rapidly increased their appeal in the pandemic period [12]. A pure exploration of existing environments through the original spherical photos, is, however very limiting. Prominent examples of additional needs include the emptying of rooms before their presentation to virtual visitors (if only for privacy reasons), or the refurnishing or redecorating of interior spaces [28]. In this context, fast and effective Diminished Reality (DR) techniques, which conceal real-life parts from the view field, are paramount to remove the furniture and other clutter that masks the architectural structure. In particular, DR features are essential to allow users to immediately compare the furnished and unfurnished scene, and to support Augmented Reality (AR) applications in placing objects in the empty scene [29, 30]. Making these features available on novel environments with minimum latency, ideally in real-time, would, in addition, enable their usage in remote collaboration contexts, without the need for prior modeling [31].

While a variety of object erasing and image inpainting solutions have been presented in the literature (section 5.3), DR for interior environments must generate images of empty indoor spaces that not only have a realistic appearance, but respect the context in stricter ways, in particular by inferring a plausible organization of the permanent architectural structure that bounds the room's interior [32]. Data-driven solutions, that learn hidden relations from examples, are emerging as viable approaches for this class of problems. However, state-of-the-art methods for image inpainting are mostly focused on photorealism [33, 34], and additional information about the scene is exploited only from the semantic point-of-view [35, 36, 32]. Current pipelines make limited use of the structure of the observed scene, and reconstruction accuracy is achieved at the price of high computational complexity or increased user intervention, using, for example, recursive networks [37], multi-branch architectures [34], and manual definition of specific parts of the original image to be removed [35].

In this chapter, we present a novel light-weight end-to-end deep network that, from an input 360° image of a furnished indoor space automatically returns, with very low latency, an omnidirectional photorealistic view and architecturally plausible depth of the same scene emptied of all clutter.

By harnessing the availability of large scale, photorealistic synthetic datasets, we train our network on pairs using a set of examples composed of registered equirectangular images of the cluttered environment color, the empty environment color, and their depth. The final end-to-end network is decomposed in two blocks, which are trained separately to reduce training costs. The first block learns an attention mask of the uncluttered parts of the input image, generating training examples from the cluttered input image and the depth pairs. The second block takes as input the attention mask and the cluttered image, and performs the synthesis of the uncluttered scene, using for training indoor-specific losses that embed our knowledge of expected indoor environments. Contrary to other object removal approaches, our approach is holistically applied to the entire scene, removing all clutter in a single step without user intervention. Rapidly emptying the room without manual intervention is the essential building block upon which the other features required for a DR application. For instance, removing a single object (or keeping only a single object) is achieved by compositing the empty room image from our network with the original image, while taking into account the computed object mask (see, e.g., the design of Gkiktas et al. [32]). Moreover, by inferring the room's geometry while removing clutter, we provide support various scene edits, including adding/positioning furniture while resting on floor or attached to a wall (see Figure 5.1).

5.2 Contributions

Our main contributions are summarized as follows:

- We propose a light-weight end-to-end deep-learning technique (section 5.4), which provides, at interactive rate, a panoramic indoor scene emptied automatically without user intervention and suitable for use in XR applications. Our prediction network develops in a linear fashion, with no need to fuse features from parallel branches [34, 32], or to refine the result recursively [37]. In order to alleviate the burden of convolutional gating for generic user-assisted inpainting [33], we adopt instead a depth-separable gated convolution strategy, reducing the number of parameters and processing time while maintaining the effectiveness [89]. Furthermore, both visual and geometric constraints are applied only at training time, where the visual ones follow a strategy of transfer learning [123] and the geometric ones adopt robust and efficient losses that encode our prior knowledge on interior environments (section 5.4.3).

- We predict a geometric representation paired with the output image, that is a dense depth estimation of the empty scene. This geometric representation can be directly used as a basis for further processing in XR application (e.g., to aid object positioning or to compute occlusions). It is obtained jointly with the visual representation and without the need of onerous parallel branches [34, 32]. We also exploit it to define a robust and effective pixel-wise prior together with other 3D priors and losses (section 5.4.3) The generation of a geometric clue as output reduces the need to add additional semantic analysis on the image or to use GAN strategies [92, 124] to disambiguate the results obtained, as demonstrated by our results (section 5.5). By contrast, current inpainting methods are mainly focused on the visual and perceptual output [33, 34], where structure preservation is handled at image-feature level [125] or semantically [35, 36]. Other approaches are based, instead, on manual and simplified annotations of the underlying layout, which does not necessarily represent the true 3D geometry. This information is best interpreted as a 2D semantic prior rather than a geometric one [36, 32].

- We drive our training using a loss function that combines photorealistic and geometric terms. In particular, our geometric terms exploit both pixel-wise information from depth maps and the concept of virtual normals generated by

triples of points at a large distance [126], to efficiently recover the salient characteristics of man-made indoor structures, in terms of flatness and smoothness, without falling into restrictive structures such as Manhattan World, Atlanta World or even vertical walls [1].

Our results show that our method outperforms current state-of-the-art approaches, using common benchmarks with a measurable ground truth, in terms of accuracy, quality and less computational complexity (section 5.5.3). Moreover, our model is also able to produce compelling predictions even on images from common datasets where no ground truth is available for training, as well as on novel images captured by an user (section 5.5.4).

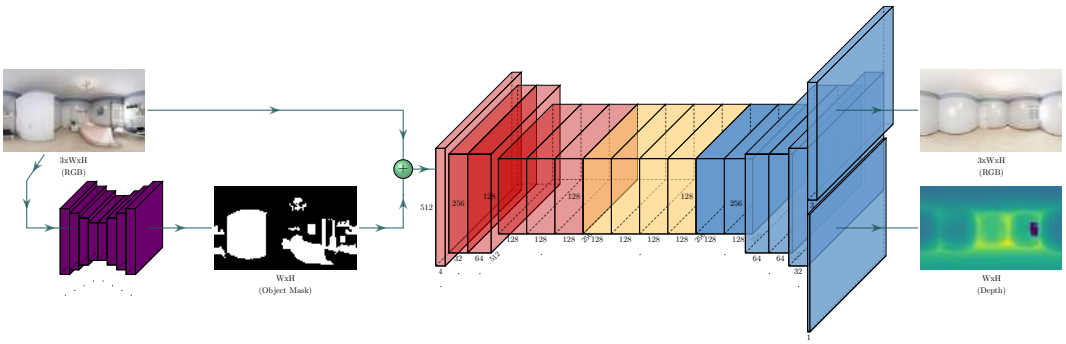


Fig. 5.2.: **Model architecture.** We process the input equirectangular image to identify the cluttered area in the scene, exploiting a light-weight network (purple blocks - section 5.4.1). The clutter mask and the input image are passed to the *empty scene synthesis network* (section 5.4.2), including a gated encoder (red blocks), a dilation bottleneck (yellow blocks) and a gated decoder (blue blocks), whose last layer is split in 2 layers: one for the photorealistic equirectangular representation of the emptied scene and one for its depth. The scene synthesis network is trained end-to-end through the methods and losses described in section 5.4.3.

5.3 Related Work

DR for indoor scenes builds on techniques for data-driven inpainting and image-to-image translation, and must extend them in order to produce a realistic and geometrically consistent environment, eventually estimating the depth of the uncluttered scene. In the following, we focus on the methods that are most closely related to ours.

5.3.1 Diminished reality for indoor spaces

DR applications provide the illusion of concealing, eliminating, and seeing through objects while perceiving an environment. In contrast to AR and MR, which superimpose virtual objects to real-world representations, they require techniques to detect the unwanted objects and replace them with the hidden background in generated images. In most DR applications, the objects to be removed are already determined as targets of interest, and specific techniques are employed for their detection (e.g., pedestrians [127] or buildings [128]). In indoor spaces, the most basic operation is the removal of interior clutter (furniture and other non-permanent objects) [29, 28, 129, 30, 32], which is supported either through interactive mask definitions (e.g., [32]), or through semantic or instance segmentation (e.g., [36]). In this work, instead, we learn, from synthetic examples, a geometric definition of clutter, that includes anything with an appreciable geometric volume that is not part of the permanent architectural structure. A wide variety of approaches have been proposed in the literature for synthesizing the hidden background (see [7] for a comprehensive survey). A number of methods employ reprojections of actual background images, generated through a prior observation of the same scene [130, 131] or a concurrent observation from other points of view, e.g., by employing multiple cameras [132]. Since these approaches require considerable effort and/or specialized setups, much research has focused, instead, on generating plausible background rather than recovering actual ones. Early solutions recovered background textures from the same image, especially analyzing areas nearby removed objects (e.g., [133]). Since these methods are generally limited to small holes and fairly regular scenes, the focus has recently shifted towards data-driven solutions that learn from a large body of prior examples. We follow this trend by generating a plausible background of a novel scene using a single $360°$ observation, exploiting concepts from data-driven inpainting and image-to-image translation, recovering not only the color but also the geometry of the empty scene in the form of a depth map. Shape inference is very important for DR of indoor environments, since it improves texture reprojection [133, 134] and parallax effects [72], and offers a basis for the editing operations [28, 135]. However, prior DR solutions either expected a simplified geometry in the hidden area (e.g., a plane [133, 134]) or required particular capture setup (e.g., multi-view [130, 132, 131] or one or more RGB-D cameras [129, 28, 135]).

5.3.2 Data-driven inpainting

The first data-driven inpainting approaches combined auto-encoders with an adversarial loss [136] or global and local discriminators [86] to produce photo-consistent images. They used regular or dilated convolutions [88] combining valid and masked parts of the image, thus leading to visual artifacts such as color discrepancy and blurriness. To overcome such limitations, Liu et al. [137] introduced *partial convolutions* to handle masking effects. Later, the partial convolution concept was revisited to incorporate structural information (edges) in the reconstructed feature map [125]. Recognizing the importance of edge preservation and generation, *EdgeConnect* [85] introduced an edge generator to hallucinate edges in the missing regions, to use them as structural guidance for the inpainting task. All the above methods assume that the mask is given and concentrate on the infilling part. Yu et al. [33] further extended the idea of partial convolutions by proposing gated convolutions to learn a mask automatically from a large number of examples. Combined with SN-PatchGAN [124, 138], this approach showed the ability of effectively supporting free-form user input as guidance. We also exploit gated convolutions, but learn to separate clutter from architectural structure using examples that exploit the availability of ground truth depths, without resorting to user input in any of our phases. In parallel developments, several authors have shown the importance of feature fusion at different scales, including the pyramid-context encoder approach [139] and the mutual encoder-decoder [140]. Conversely, Li et.al [37] propose a recurrent (i.e., iterative) method to inpaint missing regions from the outer regions of the hole towards the inner ones. Thanks to the data-driven design, the method is superior to previous techniques that assumed that gaps should be filled with similar content to that of the background, and can hallucinate new content for large holes. Zheng et al. [34] further extended the exploitation of global relations by designing a framework to generate multiple plausible results with reasonable content for each masked input, based on a probabilistic approach. To achieve that, they combine generative and variational synthesis approaches. None of the above methods, however, is applied to 360° imagery and exploits and generates geometric data.

5.3.3 Image-to-image translation

DR can be also recast as an image-to-image translation problem [92], as it maps the input cluttered image to the output uncluttered image. Visual content and style preservation is very important in this context [141, 50]. Isola et al. [92]

proposed conditional GANs as a general solution to various translation problems, semantic image synthesis being the most related to ours. The classic approach is to use semantic labeling, and reconstruct images from the semantic maps, preserving boundaries among classes [142]. Conserving semantic information fed to a deep layer built by stacking convolutional, normalization, and non-linear layers is, however, difficult, since normalization layers tend to blur semantic input. For this reason, Park et al. [35] introduced spatially adaptive normalization, in which the input map is exploited for modulating the activation in normalization layers through a spatially-adaptive, learned transformation. The approach has been recently extended by introducing per-region style encoding and allowing the user to select a different style input image for each semantic region [36]. Very recently, *PanoDR* [32] applied the above method to panoramic images of indoor scenes. In their approach, a pixel-wise semantic prior maps each pixel to the ceiling, wall, or floor class, while inpainting is performed exploiting a SEAN module [36]. As for classic inpainting methods, all these techniques are focused on the perceptual aspect and, in order to improve the realism of prediction, they exploit clues of image structure consistency (boundaries, edges), additional semantic information [35, 36] or user input [33]. Using such additional information mainly involves feature fusion from parallel branches [34, 32] or refining the result recursively [37], increasing the computational cost of the methods. In our approach, instead, we propose a linear pipeline avoiding feature fusion and recursion, leveraging the fact that the scene to be reconstructed has a specific geometry, and exploiting such an information only at training time. Furthermore, such geometric constraints reduce the need to use adversarial losses [92, 124] to disambiguate the results obtained, simplifying model structure and training approach. This leads to novel contributions in terms of network structure and loss functions.

5.3.4 Uncluttered depth estimation

In order to integrate models into the empty scenes, or to perform geometric measures, DR requires shape information in addition to color. We do that by also inferring the depth map of the uncluttered image. Learning-based monocular depth estimation was introduced over a decade ago (e.g., Make3D [143]), and the emergence of deep learning, as well as the availability of large-scale 3D datasets, has contributed to significant performance improvements. Since the restricted field of view (FOV) of conventional perspective images inevitably results in a limited geometric context [2], much of the research on reconstruction of indoors from sparse imagery is now focused on inferring depth from a single omnidirectional interior image [63, 65, 15]. While

these methods have been shown to cope with large amounts of clutter, they target the generation of the visible depth of the fully cluttered viewed room, rather than the estimation of the depth of the uncluttered one. For this reason, specific approaches for inferring the architectural layout are being actively researched. As noted by Zou et al. [14], most current data-driven layout reconstruction methods basically follow a pipeline that, based on specific indoor assumptions (e.g., Manhattan World), predicts layout elements in image space, followed by a post-processing for fitting a regularized 3D model to the predicted 2D elements. Recent solutions fully working in 3D [56] produce an approximate result in the form of a low-poly 3D mesh. By contrast, our approach, differently from all prior approaches strives to produce a per-pixel *uncluttered depth*, within a single-branch light-weight network that also produces the uncluttered color.

5.4 Methods

The overall architecture is illustrated in Figure 5.2 and explained in the following sections. We first process the input image (i.e, equirectangular format) to identify the cluttered area in the scene, exploiting a separately-trained light-weight network (purple blocks in Figure 5.2), described in section 5.4.1. The returned clutter mask and the original input image are then passed to the *empty scene synthesis network*, described in section 5.4.2 (main network in Figure 5.2), which returns the photorealistic equirectangular representation of the emptied scene and its registered depth. The scene synthesis network is trained end-to-end through the methods and losses described in section 5.4.3.

5.4.1 Clutter mask prediction

The first stage of our method consists of an identification of the area containing the clutter that should be removed from the image ${I_f}^{h \times w}$ to generate the empty room color and depth panoramas. This identification consists of a binary mask $M^{h \times w}$, that contains 1 for pixels containing clutter, and 0 otherwise. Contrary to many current DR approaches, that are oriented to the removal of single objects, we do not expect that users define it interactively [32], eventually supported by object recognition and segmentation systems [36], but learn how to generate this mask directly from I_f through a lean segmentation network. We do that since we want our mask to identifies all the non-permanent structures that need to be removed at

the same time, differentiating them from the architectural layout of the room. For training, we exploit the large body of information provided by recent large-scale photorealistic synthetic datasets [52], which contain the registered representation of empty and non-empty rooms. Even though, in this thesis, we only exploit the differentiation between clutter and layout, for maximum generality, we cast our problem as a classification problem, since many datasets contain, for each pixel also the type of object and/or the type of layout surface (ceiling, floor, wall). Such a classification might be of interest for reconstruction or when wanting to remove only particular kinds of objects. In our context, we only consider a two-class situation (layout=0, clutter=1), that can be generated, in absence of annotations in the source datasets, by simply comparing the ground-truth depths of the empty (${D_e}^{h\times w}$) and non-empty (${D_f}^{h\times w}$) room representations, including in the clutter class the pixels for which $D_f < D_e$ and to the layout class all others.

With this approach, we define as clutter the portions of the environment that have an appreciable geometric volume in the room, but are not part of the bounding architectural structure of the room. Flat objects such as electric outlets or decorations (or mirror images) thus appear in the empty room by design. Such a definition is also commonly adopted for indoor structured reconstruction approaches [**Pintore:2019:AMC**, 144]. This choice avoids the need for semantic annotations, and lets the system learn a stable association between color and geometric shape using a completely automatic method using commonly available datasets. This approach does not exclude a combination with semantic information (e.g., [36]) to also remove flat objects.

We predict our full-empty mask from the image ${I_f}^{h\times w}$ as a dual channel probability map ${D_m}^{2\times h\times w}$ (i.e., full and empty channel), using a very lightweight encoder-decoder network based on the U-Net architecture, using just 256 channels as bottleneck (i.e., $4M$ parameters) and skip-connections [83]. The training of this network, the purple one in Figure 5.2, is performed independently from the image synthesis network, as we experienced that training the clutter mask network simultaneously with the image synthesis network produces little or no advantages, but imposes an additional load on the entire training process (see section 5.5). For each of the two channels, training of the clutter mask is driven by binary cross-entropy loss:

$$-\frac{1}{n}\sum_{p\in {D_m}^c}(\hat{p}\log p+(1-\hat{p})\log(1-p)) \tag{5.1}$$

where ${D_m}^c$ is the slice c of D_m, p is the predicted probability of one pixel of being of class c, and $\hat{p}$ is the ground truth probability. The final predicted binary mask $M^{h\times w}$,

that feeds the second stage of our complete network, is obtained by assigning each pixel to the class with maximum probability and setting the pixel value according to this classification.

5.4.2 Empty scene synthesis

To generate the empty scene image and depth, we adopt the architecture illustrated in Figure 5.2. The overall encoder-decoder scheme follows a common design for image inpainting [86], exploiting dilated convolutions as bottleneck [88], and gated convolutions for encoding decoding [87]. Compared to the baseline [86, 87], our architecture is thinner, deeper, and with fewer parameters. Moreover, it has only a single branch and it includes several solutions (described below) to improve accuracy and reduce computational complexity. Furthermore, given the spherical nature of the image, we adopt circular padding along the horizon for convolutions, thus removing longitudinal boundary discontinuity, and reflection padding to alleviate the singularities at the poles [32].

The input of the network consists of a masked image of the cluttered room, with white in clutter regions, together with the binary mask indicating the hole regions (section 5.4.1). The paired input is encoded through a sequence of light-weight gated convolutions having different strides (the 6 layers in red in Figure 5.2), so that the original size is reduced by a factor four in each direction. Each encoding convolution is followed by instance normalization [102] and ReLU activation. In our network, we adopt a specific form of gated convolution, that integrates a learnable gating technique when selecting features [87], since vanilla convolutions are ill-fitted for image inpainting [87, 86]. Considering a standard convolutional layer and a $C_{in} - channel$ input feature map, each pixel located at (y, x) in the $C_{out} - channel$ output map is computed as:

$$O_{y,x} = \sum_{i=-k'_h}^{k'_h} \sum_{i=-k'_w}^{k'_w} W_{k'_h+i,k'_w+j} \cdot I_{y+i,x+j} \tag{5.2}$$

where x, y represents the location along the x- and y-axis of the output map, k_h and k_w is the kernel size (e.g. 3×3), $k'_h = \frac{k_h-1}{2}$, $k'_w = \frac{k_w-1}{2}$, $W \in \mathbb{R}^{k_h \times k_w \times C_{in} \times C_{out}}$ are convolutional filters, and $I_{y+i,x+j} \in \mathbb{R}^{C_{in}}$ and $O_{y,x} \in \mathbb{R}^{C_{out}}$ are inputs and outputs. The application of the same filters at each spatial location (y, x) is not appropriate. This is because, for inpainting, the input will need to combine valid pixels/features coming from regions outside holes with invalid pixels/features (in

shallow layers) or synthesized pixels/features (in deep layers) coming from masked regions [33]. Although simple partial convolutions [137] can be used to make the convolution dependent only on valid pixels, they are not suitable for our problem, since, essentially, they act as single-channel hard-gating.

Thus, we adopt a gated convolution (GC) approach [33], expressed as:

$$\begin{aligned} G &= conv(W_g, I) \\ F &= conv(W_f, I) \\ O &= \sigma(G) \bigodot \psi(F) \end{aligned} \tag{5.3}$$

where σ is the Sigmoid function, which outputs values in $[0, 1]$, ψ is an activation function (ReLU in our case), and W_g and W_f are two different sets of convolutional filters, which are used to compute the gates and features respectively. GC enables the network to learn a dynamic feature selection mechanism. It should be noted that, according to Equation 5.2, W_g has $k_h \times k_w \times C_{in} \times C_{out}$ parameters, almost doubling the number of parameters and processing time in comparison to vanilla convolution. In order to simplify training and guarantee low latency at inference time, our network uses a modified version of GC called Light Weight Gated Convolutions (LWGC), which reduces the number of parameters and processing time while maintaining the effectiveness [89]. Specifically, we decompose G from Equation 5.3 into a depth-wise convolution [89] (i.e., 3×3) followed by a 1×1 convolution, having, as a result, the same gating step but with only $k_h \times k_w \times C_{in} + C_{in} \times C_{out}$ parameters.

Repeated dilations [88] are used for the bottleneck (Figure 5.2, yellow blocks), thus increasing the area that each layer can use as input. It should be noted that this is done without increasing the number of learnable weights, but obtained by spreading the convolution kernel across the input map. The *dilated convolution operator* is then implemented as a gated convolution (i.e., Equation 5.3), but with some differences. It is expressed as:

$$D_{y,x} = \sigma(b + \sum_{i=-k'_h}^{k'_h} \sum_{i=-k'_w}^{k'_w} W_{k'_h+i,k'_w+j} \cdot I_{y+\eta i,x+\eta j}) \tag{5.4}$$

where, assuming the same notation of Equation 5.2, η is a dilation factor, $\sigma()\dot{}$ is a component-wise non-linear transfer function and $b \in \mathbb{R}^{C_{out}}$ is the layer bias vector. With $\eta = 1$, the equation becomes the standard convolution operation. In our model, we adopt, respectively, $\eta = 2, 4, 8, 16$ for the four bottleneck layers. Using this strategy, we aggregate multi-scale contextual information without losing resolution,

thus capturing the global context efficiently by expanding the receptive field, avoiding additional parameters and preventing information loss. This is important for the image completion task, as capturing sufficient context is critical for realism. By using dilated convolutions at lower resolutions, the model can effectively cover a larger area of the input image when computing each output pixel than with standard convolutional layers [86].

The network decoder (7 blue layers in Figure 5.2) follows a scheme which is symmetrical with respect to the scheme of the encoder. Five layers, based on gated convolutions, restore the resolution of the output to the original input resolution, and a final double layer (two layers in parallel) is dedicated respectively to the synthesized *RGB image* and its depth (Figure 5.2). These last two layers have two different activation functions, respectively $tanh$ for the RGB output and ELU for the depth output.

5.4.3 Training and losses

During the training phase, we compute the parameters of the network (section 5.4.2) using a supervised training approach. To this end, we currently exploit Structured3D [52]), a large-scale, synthetic database of indoor scenes. For each scene, a photorealistic, equirectangular rendering of the cluttered environment is matched with the rendering of the same empty scene and with its depth map. It should be noted that such a pixel-wise accurate matching between full and empty scenes and their depths is practically only possible with synthetic data. However, an important benefit of our method is the ability to perform transfer learning efficiently, so the model trained on the synthetic dataset also performs very well on real images, as demonstrated in our results (section 5.5).

Our loss functions are designed to combine a visual term, that measures the photorealistic quality of the output, with a geometric term, that drives the solution towards a plausible reconstruction of an indoor environments.

The visual term is a combination of different domain losses to ensure the photorealistic quality of the predictions:

$$\mathcal{L}_{vis} = \lambda_{px}\mathcal{L}_{px} + \lambda_{perc}\mathcal{L}_{perc} + \lambda_{style}\mathcal{L}_{style} \tag{5.5}$$

The first term is a pixel-based $L1$ loss between the predicted RGB image I_{out} and the ground truth empty scene image I_{gt}. $\mathcal{L}_{perc}$ and $\mathcal{L}_{style}$ are the data-driven perceptual and style losses [50]. These enforce I_{out} and I_{gt} to have a similar representation in the feature space as computed by a CNN model ψ, which, as in many image synthesis approaches, is a pre-trained $VGG-19$ [91]. The perceptual loss is, thus, given by:

$$\mathcal{L}_{perc} = \sum_{n}^{N-1} \|\psi_n(I_{out}) - \psi_n(I_{gt})\|_1 \tag{5.6}$$

and computes the $L1$ distance between the projection of I_{out} and I_{gt} into high-level features using the pre-trained network ψ, thus preserving *high-level* content of the image. In Equation 5.6, ψ_n is the activation map of the n-th selected layer. In our loss, we use $relu11$, $relu21$, $relu31$, $relu41$ and $relu51$ layers [141].

The style loss, calculated on the same layers of perceptual loss, is given by:

$$\mathcal{L}_{style} = \sum_{n}^{N-1} \left\| K_n(\psi_n(I_{out})^T \psi_n(I_{out})) - \psi_n(I_{gt})^T \psi_n(I_{gt}) \right\|_1 \tag{5.7}$$

which includes the *Gram matrix* function, where the high level feature $\psi(x)_n$ is of shape $(H_n W_n) \times C_n$, resulting in a $C_n \times C_n$ Gram matrix, and K_n is the normalization factor $1/C_n H_n K_n$ for the nth selected layer. Differently from the perceptual loss, this component gives more importance to local similarity (e.g., texture).

The geometric term is a combination of low- and high-order 3D constraints:

$$\mathcal{L}_{geom} = \lambda_d \mathcal{L}_d + \lambda_n \mathcal{L}_n \tag{5.8}$$

The low-order term $\mathcal{L}_d$ is a robust pixel-wise loss between the predicted depth D_{out} and the ground truth depth of the empty scene D_u (section 5.4.1). Similarly to other recent state-of-the-art solutions (e.g., BiFuse [63] and SliceNet [15]), we adopt as objective function the *Adaptive Reverse Huber Loss (BerHu)* [90]. For the high-order term $\mathcal{L}_n$, we consider a geometric constraint from a global perspective to take long-range relations into account. This is achieved by exploiting the concept of *virtual normal* [126], i.e., the normal vector of a virtual plane formed by three randomly sampled non-collinear points in 3D space. By minimizing the direction divergence between a small set of ground-truth and predicted virtual normals, serving as a high-order 3D geometric constraint, we preserve the global shape of the model. Such an approach is very effective for indoor environments, typical composed of the union of a small set of smooth surfaces.

From the given depth map $D^{h \times w}$, a 3D point cloud is reconstructed by spherical projection, so that, for each pixel $p_i(u_i, v_i) \in D$, we obtain the location $P_i(x_i, y_i, z_i)$ in 3D coordinates with respect to the sphere center (i.e., camera point-of-view). N triples of points are randomly sampled from the point cloud. The three points $\{(P_a, P_b, P_c)\}$ in each triple are restricted to be non-collinear as defined by the following condition:

$$C = \{\alpha \geq \angle(\overrightarrow{P_a P_b}, \overrightarrow{P_a P_c} \leq \beta, \alpha \geq \angle(\overrightarrow{P_b P_c}, \overrightarrow{P_b P_a} \leq \beta\} \tag{5.9}$$

where $\alpha = 150°$ and $\beta = 30°$ in our experiments.

The normal vector n_i of the plane formed by the three points is computed by:

$$n_i = \frac{\overrightarrow{P_a P_b} \times \overrightarrow{P_a P_c}}{\left\| \overrightarrow{P_a P_b} \times \overrightarrow{P_a P_c} \right\|} \tag{5.10}$$

The high order loss $\mathcal{L}_n$ is computed by:

$$\mathcal{L}_n = \frac{1}{N} \sum_{i=1}^{N} \left\| {n_i}^{pred} - {n_i}^{gt} \right\| \tag{5.11}$$

A small number of triples is sufficient to produce effective results. As an example, for each predicted or ground truth depth, having 512×1024 size, we randomly sample 3600 triplets, from which we obtain a pair of 3600 virtual normals, i.e., less than 0.7% of the pixels. The contribution of geometric terms is highlighted in the ablation study at section 5.5.5.

The relative importance of each loss term is determined by the values of the λ_x coefficients. In our experiments we use $\lambda_{px} = 4, \lambda_{perc} = 1, \lambda_{style} = 40, \lambda_d = 0.5, \lambda_n = 0.5$.

5.5 Results

Our approach was implemented using *PyTorch* [51] and has been tested on a large variety of indoor scenes. Here we report on results on the benchmarks used by the majority of state-of-the-art works [52, 41]. In addition, we report on the applications to scenes captured by non-professional users.

Fig. 5.3.: Examples of inference of color and depth of the empty room from a single-shot panorama

Two examples of depth and color representations of empty rooms starting from a single-shot panoramic image of cluttered environments are presented in Figure 5.3.

5.5.1 Training and testing datasets

We use Structured3D [52] to train, validate and numerically compare our results to ground truth and other works. Structured3D [52] is a large-scale photo-realistic dataset containing $3.5K$ house designs created by professional designers with a variety of ground truth 3D structure annotations, including $21,000$ photo-realistic full-panoramic (i.e., equirectangular format) indoor scenes. These panoramic scenes are provided with or without furniture and objects. For all configurations, both RGB images and depth maps are provided, allowing us to immediately use them for training, validating, and testing without further configuration. The official splitting [52] is used, with no overlap among training and testing partitions.

It should be noted that, while our method makes no particular assumption on the architectural structure, Structured3D mostly includes Atlanta World structures [1], leading to a better performance on scenes also meeting these constraints, even though this constraint is not necessary for our network structure. As a further minor limitation of this dataset, we noted that the environment map of the outdoors seen through the windows is replicated in scenes that are part of the training and testing partition. We plan to generate higher variations both of room geometry and outdoor textures in our future works.

In addition to testing with Structured3D, we also adopt Matterport3D [41], a large-scale RGB-D dataset containing $10,800$ panoramic views from $194,400$ RGB-D images of 90 real building-scale scenes, to test our method on real-world scenes. We also

(a) RGB input (b) ground truth (c) other [32] (d) our RGB (e) our depth

Fig. 5.4.: Qualitative performance and comparison vs. ground truth and other approaches on the Structured3D dataset [52]. We compare to PanoDR [32], which has the best panoramic performances among the available methods. We additionally show our output depth paired with our visual results (Figure 5.4d).

exploit this dataset to demonstrate our transfer-learning capabilities. Furthermore, to demonstrate the robustness of the method towards real acquisitions of various types, we selected scenes from a variety of real-world datasets used in the field of automatic building reconstruction [1] or manually acquired by non-professional users using the widely available Ricoh Theta spherical cameras.

5.5.2 Setup and computational performance

We trained both the clutter mask prediction (section 5.4.1) and the scene synthesis (section 5.4.1) networks with the Adam optimizer [95], with $\beta_1 = 0.9$, $\beta_2 = 0.999$, on two NVIDIA RTX 2080Ti GPUs (11GB VRAM) with a batch size of 8 and a learning rate of 0.0001. When using the typical 512×1024 resolution, the average training time for the clutter mask prediction model is $32ms/image$, while it is $196ms/image$ for the scene synthesis model. We adopt for training both networks the Structured3D [52] official splitting. $VGG-19$ [91] pre-trained model is the one provided by TorchVision [51].

Table 5.1 presents our computational performance compared to state-of-the-art inpainting methods. Although our method is fully scalable in resolution (see Table 5.2),

(a) input [41] (b) prediction (c) input [41] (d) prediction

(e) user input (f) prediction (g) user input (h) prediction

Fig. 5.5.: We present qualitative performance on data for which no ground truth or training set was available. Here, we show cases from the large scale real-world dataset Matterport3D [41] and from typical user-acquired scenes, where captured images are not perfectly aligned and the photographer is visible.

Table 5.1 shows the performance with a resolution compatible with the other baselines and adopted in previous comparisons [32], avoiding modifications of the other models (i.e., 256×512). Our method is clearly the most lightweight and has a lower computational complexity (e.g., GFLOPS) than the compared inpainting methods. Moreover, as our approach is designed to remove all the objects in the scene at the same time without user intervention, our presented statistics include the cost of both the clutter mask estimation and scene synthesis networks, while other methods, designed for general infilling, report results only for the synthesis part.

Method	**Params↓**	**GFLOPS↓**	**ms/frame↓**
RFR [37]	30.59 M	412.22	157
Deepfillv2 [33]	13.86 M	163.44	41
PanoDR [32]	20.88 M	122.53	270
Our	**6.06 M**	**41.03**	**17**

Tab. 5.1.: **Computational performance.** We show our computational performance compared to other state-of-the-art works on a single NVIDIA RTX 2080Ti GPU.

Resolution	Params	GFLOPS	ms/frame
256×512	6.06 M	41.03	15
512×1024	6.06 M	164.11	41
1024×2048	6.06 M	656.45	174

Tab. 5.2.: **Computational scalability.** We show our computational performance and latency time for different input resolution. Our results demonstrate how we diminish images with a very low latency even when resolution increase.

Table 5.2 shows, instead, how our approach scales to higher resolutions. As demonstrated in the results, we diminish images with a very low latency, even at the higher tested resolution (1024×2048). Applications can, thus, provide a quick feedback following a camera motion and/or environmental changes. While we currently exploit these advantages for interactively taken single-shot images, the achieved performance makes it possible to consider an extension to real-time room emptying during continuous capture. As a term of comparison, approaches such as PanoDR [32] take 1183 GFLOPS at the 512×1024 resolution (i.e., close to an order of magnitude larger than ours), making it hard to perform the inference on a single commodity graphics board.

5.5.3 Performance vs. ground truth and competitors

We compared our performance to the one achieved by several state of the art inpainting methods [37, 34, 33, 32], which are representative of the most related approaches discussed in section 5.3. To provide a quantitative evaluation with respect to ground truth, we train all methods using Structured3D [52] and used the official implementation, minimally adapted to the equirectangular format, for the computational performance evaluation (section 5.5.2). Table 5.3 presents results on the full Structured3D [52] test set, according to its official split.

We adopt standard metrics: Mean Absolute Error (MAE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [99] and the Learned Perceptual Image Patch Similarity (LPIPS) [100]. LPIPS is a metric that has been shown to better assess the perceptual similarity between two images. It measures the distance between the target and generated images using features extracted from a pre-trained VGG-16 model. Since other compared approaches assume that a mask of the parts of the image to be removed is provided by a user (i.e., the whole clutter), we have used as input the ground truth clutter mask (section 5.4.1). Note that this aspects favors the other approaches, since for our method we use, instead, the mask estimated from the color input.

Even with this difference, our method outperforms the other approaches on all considered metrics. This can be explained by the fact that the currently available methods are designed to remove limited portions of the image or single objects, rather than the entire clutter preserving only the architectural layout, while our method is adapted to that situation. This fact clearly shows the advantage of designing a task-specific network.

Method	LPIPS↓	MAE↓	PSNR↑	SSIM↑
RFR [37]	0.418	0.201	10.885	0.745
PicNet [34]	0.472	0.204	10.922	0.733
Deepfillv2 [33]	0.354	0.188	11.235	0.729
PanoDR [32]	0.310	0.172	11.612	0.754
Our	**0.129**	**0.040**	**24.702**	**0.925**

Tab. 5.3.: **Quantitative performance.** We show our quantitative performance compared to other state-of-the-art works.

Figure 5.4 presents some examples of our qualitative performance (Figure 5.4d), compared to ground truth (Figure 5.4b) and to other methods (Figure 5.4c). We choose to compare our approach with PanoDR [32], since it was specifically designed for diminished reality on panoramic images and, in our tests, it is the best performing among the other tested methods. Moreover, the method embeds several other state-of-the-art solutions for image inpainting [33, 36]. Our method performs well under different conditions, such as near and far objects, poorly or highly textured walls, more or less distorted foreground, as well as background occlusions.

Figure 5.4e shows the depth produced by our method (Figure 5.4e), which is not computed by the other inpainting solutions. To provide a term of comparison, we compared our method with state-of-the-art publicly available networks for panoramic depth prediction, i.e., SliceNet [15] and HoHoNet [65], trained on the Structured3D [52] dataset, and with the work of Jin et al. [98], which released the full-empty dataset adopted in this work. Since we target the estimation of the depth of the uncluttered scene, while competing methods do not differentiate clutter from architectural structure, the comparison is performed in the uncluttered areas for SliceNet [15] and HoHoNet [65], i.e., only for pixels not masked with the ground truth masks (section 5.4.1). To compare ourselves with Jin et al. [98], we use instead the official data provided by the authors on the same data used by our method, since their original code is not available.

Table 5.4 provides depth results using the common metrics, i.e., mean absolute error (MAE), mean squared error (MSE), root mean square error of linear measures

(RMSE) and relative accuracy δ_1, defined as the fraction of pixels where the relative error is within a threshold of 1.25. For MAE, MSE, and RMSE, smaller is better (unit is $meter$), while for δ_1 larger is better.

Method	MAE↓	MSE↓	RMSE↑	δ_1 ↑
HoHoNet [65]	0.101	0.076	0.206	0.932
Jin et al. [98]	-	0.071	0.642	0.958
SliceNet [15]	0.062	0.054	0.198	0.961
Our	0.091	0.073	0.197	0.954

Tab. 5.4.: **Depth prediction performance.** We show our quantitative performance compared to other state-of-the-art works.

It should be noted that, despite the limited complexity of our network, and the fact that it also targets color estimation, the accuracy of our depth prediction appears comparable to the results of state-of-the-art specific methods for panoramic depth estimation [15, 65]. Our very good results with a much leaner network are due to the fact that, in this particular setting, we target reconstruction only of the fairly regular areas comprising the architectural layout of the room, while methods seeking to reconstruct the full depth [15, 65] must handle much more variable and discontinuous visible shape, due to the high presence of furniture and other objects that have to be measured.

5.5.4 Performance in-the-wild

Figure 5.5 presents qualitative performance on data for which no ground truth or training set was available. This situation is the expected usage of our method.

In the upper part of the figure, we show scenes from the large-scale real-world dataset Matterport3D [41]. In the bottom part of the figure, we show scenes acquired by non-professional users using commodity low-cost devices (i.e., Ricoh Theta V and Ricoh Theta Z), collected or acquired by us. In the case of Matterport data, the blur in the upper and bottom part of the scene is due to the fact that those areas are missing due to hardware limitation of the device, and have been approximated in the input scene with a color diffusion.

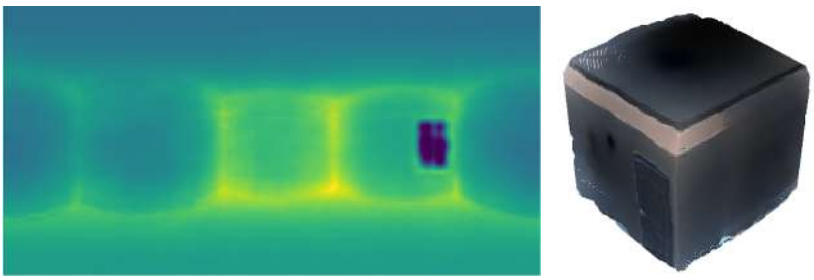
Although the training of our model was done on a synthetic dataset mainly including Atlanta World structures [1], our method makes no special assumption on the indoor scene kind, or about the precise alignment of the camera with respect to the ground [15, 65] (within the limits of rational, even manual, capture). Furthermore, the method automatically removes the photographer who takes a panoramic photo by

holding the camera (i.e., that is considered as clutter). As an example, the last row of Figure 5.5 shows our prediction when capture is not aligned to the ground and when the user is visible in the cluttered scene. In all cases, our method is able to predict compelling empty scenes on real data acquired with different devices, automatically removing various types of clutter and very heterogeneous furniture.

The biggest difference in the results, compared to standard synthetic testing sets, is on the lighting appearance of the resulting scene, which sometimes differs from the setup of the original cluttered scene (Figure 5.5a). One of the evident consequences of this phenomenon is the different color tone of some scenes (Figure 5.5b). This is not surprising, since our model does not, at the moment, make any assumption about lighting. This aspect could be object of future works (section 5.6).

5.5.5 Discussion and ablation study

In this section, we discuss our major technical choices, supported by an ablation study, and several features and limitations of our method. As seen in the previous sections, our approach proves to be light-weight and scalable (see Table 5.2). It should also be noted that the 3D output allows for real-time 3D rendering applications, independent of image-based rendering applications only, for instance the usage of geometric features to help positioning virtual objects on the ground or aligned to walls (Figure 5.1). As shown in Table 5.4, our depth estimation reaches state-of-the-art quality. Figure 5.6 shows an example of the predicted point cloud, which represents a good approximation of the underlying layout of the scene.



(a) Output depth (b) Point cloud

Fig. 5.6.: **Predicted depth and its point cloud.** Example of 3D point cloud generated from the predicted depth.

Throughout Table 5.5, we show the differences in performance, with the same setup as in previous numerical experiments (see section 5.5.2), in exploiting the geometric features of the scene. In particular, PWG indicates the use or not of pixel-wise geometric loss and HOG of high-order geometric loss (section 5.4.3).

In the first case, we tested the model by passing the ground truth masks directly, and found no performance improvement over using masks learned from the model itself. It should be noted that this approach is much more efficient than using adversarial losses, as done in many infilling techniques (see section 5.3). For the GAN option, we have tested a discriminator-based loss that is learned during training (i.e., PatchGAN [92]). We experiment that an adversarial loss gives a boost in performance without geometric hints (second row of Table 5.5), but does not give additional performance when already using geometric losses. Regarding the initial

PWG	HOG	UI	GAN	LPIPS↓	SSIM↑	δ_1 D_{out} ↑
-	-	-	-	0.398	0.698	-
-	-	-	✓	0.302	0.748	-
✓	-	-	-	0.164	0.833	0.905
✓	✓	✓	-	0.136	0.914	0.954
✓	✓	-	✓	0.121	0.918	0.952
✓	✓	-	-	**0.129**	**0.925**	**0.954**

Tab. 5.5.: **Ablation facts.** We show the effect of several key choices of our approach. In bold the adopted configuration. PWG: pixel-wise geometry loss; HOG: high-order geometric loss; UI: user intervention; GAN: adversarial-loss; LPIPS,SSIM and δ_1 metrics described in section 5.5.3.

mask (section 5.4.1), we also experimented that an extreme accuracy of it is not required for our model to work (i.e., 97% IoU), as the gating mask is dynamically propagated and learned, and the purpose of the first network is more related to bootstrapping feature gating. We have also verified that training the clutter mask network in a separate stage is more efficient than training it simultaneously with the scene synthesis model, since it accelerates convergence and reduces the use of memory during training connected to loading all the data at the same time (i.e., both cluttered and uncluttered depth).

Our model proves versatile on different types of indoor scenes, even as the type of real or synthetic input data varies. However, there are cases, mainly in real-world scenes very different from training data, where our method did not produce plausible images. The first row in Figure 5.7 shows one of these cases. The particular conditions of illumination and the presence of many reflections do not lead to a plausible reconstruction. The geometry, in this case, is not sufficient to model the scene, also for the presence of unconventional structures very distant from the domestic training set on which the network has been trained. This drop in performance under these particular lighting conditions is not surprising, as our method does not actually model the lighting of the scene. This aspect will be object of future work.

(a) Input (b) Prediction

Fig. 5.7.: **Limiting cases.** Due to the particular lighting condition our network returns a blurred model.

Moreover, while our method does not explicitly impose the typical restrictive priors of several competitors (e.g., Manhattan World, Atlanta World, vertical walls), and is, therefore, adaptable to more general architectures, the only currently available training dataset [52] is of the Atlanta World type. Thus, irrespective of the generality of our network architecture, performances clearly decay when moving away from this type of scenes. The second row in Figure 5.7 shows a room characterized by a lot of clutter and a very sloping ceiling, where the lower perimeter walls are barely visible. Under these conditions, it is difficult to retrieve contextual and geometric information to reconstruct the missing parts, resulting in several artifacts. Since this limitation is related to lack of training examples, we expect major improvements when datasets containing this kind of architecture will become available.

5.6 Conclusions

In this chapter we have presented a new data-driven approach that, from an input 360° image of a furnished and cluttered indoor space automatically returns, at interactive speed, a 360° photorealistic view and depth of the same scene emptied of all furniture and other clutter. Rather than casting the problem as a simple image infilling problem, we consider the correlation between color and geometry that occurs in indoor spaces. This allows us to exploit, beside perceptual and style objective functions, geometric losses of different orders, including robust geometric pixel-wise and high-order 3D losses targeted for indoor structures, simplifying the prediction model and its computational complexity. The experimental results demonstrate that our method provides interactive performance and outperforms current state-of-the-art

solutions on commonly used indoor panoramic benchmarks and also for indoor scenes captured in the wild and for which there is no ground truth to support supervised training. While the application presented in this chapter focused on emptying of single 360° shots, with subsequent exploration and editing of the produced static environment, the accuracy and speed achieved could make it possible to consider immersive dynamic scenarios with acquisition and modification of the scene, even in motion and in real-time. In the future, such a work can be extended in this sense, also considering the lighting model and the spatial coherence of prediction.

5.7 Bibliographic notes

The contents of this chapter report the journal article *Giovanni Pintore, Marco Agus, Eva Almansa, and Enrico Gobbetti: Instant Automatic Emptying of Panoramic Indoor Scenes* [40], of which the candidate is the lead author.

The article received funding from the European Union's H2020 research and innovation program under grant 813170 (EVOCATION), and from Sardinian Regional Authorities under project VDIC.

6 Conclusion

Panoramic images have become a key component for creating immersive content directly from real scenes and for supporting a range of Virtual Reality (VR) applications. While capturing a single-shot panorama is an effective method to create a virtual replica of a real-world environment, it comes with limitations. The content presented in such panoramas is restricted to what was visible from the fixed location where the panorama was taken. Unfortunately, this limitation results in the loss of crucial 3D cues, which play a significant role in providing a sense of presence. Particularly in indoor environments, where the viewer is in close proximity to architectural surfaces and objects, the flat appearance of panoramas becomes a pronounced drawback. For a truly immersive experience, a system should dynamically generate images that respond not only to changes in orientation but also to shifts in viewpoint position. While multi-view capture setups have been explored, synthesizing views from single-shot panoramas remains essential due to the widespread use of monocular 360 cameras and their convenience. To achieve effective view synthesis, it becomes necessary to estimate the geometric model of the imaged environment, either explicitly or implicitly. This estimation allows for occlusion-aware reprojection, synthesis of disoccluded content, and artifact reduction, ultimately enhancing the sense of presence. Creating a 3D immersive indoor environment from real-world captures involves challenging research tasks, including depth estimation and architectural layout modeling from a single image. Depth information enriches the visual representation by providing per-pixel data on the distance from the viewer. However, beyond the geometric scene representation observed from the viewer's perspective, an abstract model of the permanent architectural structure (referred to as the indoor 3D layout) is also necessary, typically excluding furniture and other objects.

In this context, our thesis followed a precise research path, combining several fundamental research tasks, such as depth and layout estimation, novel-views synthesis and manipulation of them. Along this path, the various tasks were integrated to: recover a 3D indoor model for novel view-synthesis (Chapter 3), provide immersive explo-

ration of indoor stereoscopic environments (Chapter 4) and enable automatic-assisted editing of immersive indoor models (Chapter 5). Along this path we introduced novel techniques that advance the state-of-the-art in 3D reconstruction of indoor environments and immersive interaction. We provide here a brief summary of the achieved results and discussion of the potential directions for future work.

6.1 Overview of achievements

Our contributions result in a number of publications and public talks, enlisted in section 6.3, from which we highlight several innovative, end-to-end solutions, summarized below:

- **A novel methodology for 3D scene synthesis of Atlanta-world interiors from a single omnidirectional image** (section 3). A new data-driven approach for extracting geometric and structural information from a single spherical panorama of an interior scene, and for using this information to render the scene from novel points of view, enhancing 3D immersion in VR applications. The approach copes with the inherent ambiguities of single-image geometry estimation and novel view synthesis by focusing on the very common case of *Atlanta-world* interiors, bounded by horizontal floors and ceilings and vertical walls. Based on this prior, we introduce a novel end-to-end deep learning approach to jointly estimate the depth and the underlying room structure of the scene. The prior guides the design of the network and of novel domain-specific loss functions, shifting the major computational load on a training phase that exploits available large-scale synthetic panoramic imagery. An extremely lightweight network uses geometric and structural information to infer novel panoramic views from translated positions at interactive rates, from which perspective views matching head rotations are produced and upsampled to the display size. As a result, our method automatically produces new poses around the original camera at interactive rates, within a working area suitable for producing depth cues for VR applications, especially when using head-mounted displays connected to graphics servers. The extracted floor plan and 3D wall structure can also be used to support room exploration. The experimental results demonstrate that our method provides low-latency performance and improves over current state-of-the-art solutions in prediction accuracy on available commonly used indoor panoramic benchmarks. This work has been published as a TVCG journal paper [38] and presented at the IEEE

ISMAR conference 2023, held in Sydney, Australia. The candidate followed all the aspects of the work, from conceptualization, methodology, software, to validation, writing and review of the related papers.

- **A novel approach for deep synthesis and exploration of omnidirectional stereoscopic environments from a monoscopic panoramic image** (section 4). An innovative approach to automatically generate and explore immersive stereoscopic indoor environments derived from a single monoscopic panoramic image in an equirectangular format. Once per 360° shot, we estimate the per-pixel depth using a gated deep network architecture. Subsequently, we synthesize a collection of panoramic slices through reprojection and view-synthesis employing deep learning. These slices are distributed around the central viewpoint, with each slice's projection center placed on the circular path covered by the eyes during a head rotation. Furthermore, each slice encompasses an angular extent sufficient to accommodate the potential gaze directions of both the left and right eye and to provide context for reconstruction. For fast display, a stereoscopic multiple-center-of-projection stereo pair in equirectangular format is composed by suitably blending the precomputed slices. At run-time, the pair is loaded in a lightweight WebXR viewer that responds to head rotations, offering both motion and stereo cues. The approach combines and extends state-of-the-art data-driven techniques, incorporating several innovations. Notably, a gated architecture is introduced for panoramic monocular depth estimation. Leveraging the predicted depth, the same gated architecture is then applied to the re-projection of visible pixels, facilitating the inpainting of occluded and disoccluded regions by incorporating a mixed Generative Adversarial Network (GAN). The resulting system works on a variety of available VR headsets and can serve as a base component for a variety of immersive applications. We demonstrate our technology on several indoor scenes from publicly available data. This work has been accepted for publishing as a Computers & Graphics journal paper [39] and its conference form awarded with honorable mention at the ACM WEB3D conference 2023, held in San Sebastian, Spain. The candidate followed all the aspects of the work, from conceptualization, methodology, software, to validation, writing and review of the related papers.

- **An innovative end-to-end technique for instant automatic emptying of panoramic indoor scenes** (section 5). A new data-driven approach that, from an input 360° image of a furnished indoor space automatically returns, with very low latency, an omnidirectional photorealistic view and architecturally

plausible depth of the same scene emptied of all clutter. Contrary to previous data-driven inpainting methods that remove single user-defined objects based on their semantics, our approach is holistically applied to the entire scene, and is capable to separate the clutter from the architectural structure in a single step. By exploiting peculiar geometric features of the indoor environment, we shift the major computational load on the training phase and having an extremely lightweight network at prediction time. This work has been published as a TVCG journal paper [40] and presented at the IEEE ISMAR conference 2022, held in Singapore. The candidate followed all the aspects of the work, from conceptualization, methodology, software, to validation, writing and review of the related papers.

6.2 Discussion and future directions

We provide a short assessment of advantages and limitations of our proposed solutions, and an overview of areas where work is still needed.

As seen in the previous chapters, constructing an immersive indoor model from a single panoramic image requires the ability to extract a structured 3D model from it. This model should include depth information from the camera's perspective, a 3D layout that allows for inferring potentially occluded parts, and the capacity to semantically distinguish between permanent structural elements (i.e., subject to man-made rules) and clutter. Such a model is a crucial requirement for ensuring reliable view synthesis.

All techniques presented here leverage distinct characteristics of the capture setup, particularly gravity alignment, and the spherical environment, specifically world-space alignment with gravity. This alignment enables the exploitation of regularities in vertical features along the horizontal direction. These consistent characteristics have influenced network designs, resulting in the use of asymmetric contractions and various strategies to combine long- and short-range features [38]. The specific networks designed based on these principles offer significant advantages over more generic alternatives. This underscores the value of tailored solutions for interior capture, rather than relying on generic networks designed for outdoor or generic-shape 3D reconstruction. However, creating custom networks also comes with a drawback: they depend on specific environmental characteristics. Consequently, when the imaged environment deviates from expectations, major failures can occur

(e.g., see section 3.8.4). While these solutions are more robust than geometry-reasoning methods, they still exhibit limitations in terms of applicability, as evidenced by the failure case analyses presented in the previous chapters.

Another significant limitation, also shared by the other current approaches, pertains to the input size. Although our presented solutions are generally lightweight and the network design is scalable, the tests have predominantly been conducted on image sizes smaller than what is currently achievable with panoramic cameras. Existing benchmarks typically operate at a resolution of 1024×512, rarely venturing into larger dimensions. In contrast, industrial cameras capture more detailed imagery. An essential path for future research involves evaluating the scalability of these techniques to handle larger datasets and real-world cases. Achieving this will necessitate not only scaling the networks but also generating extensive annotated datasets to serve as ground truth. While depth estimation requires an accurate per-pixel ground-truth, layout reconstruction is a more abstract task which relies on user-made annotation. For practical reasons such annotations are currently mostly limited to Manhattan or Atlanta-world environments. In the future we expect to solve this problem through the increasing availability of devices that capture both high-resolution RGB and depth, and the eventual possibility of integrating annotation systems for semi-automatic strttures of interest, for example by extending uncluttering systems such as the one we presented (i.e., chapter 5).

Modern view-synthesis approaches, including our solutions, are basically based on inpaiting techniques. While in much of the literature such solutions are reinforced with GAN-type training, one of our major contributions in this topic is to also use network architectures and strategies based on geometry information, particularly indoor priors. Such an approach, as demonstrated in previous chapters, greatly improves the realism of the synthesized scene, both in cases of visual occlusion (i.e., a hidden corner that is disoccluded, chapter 3), and in the case of reconstruction of structural parts previously covered by clutter (i.e., chapter 5). This solution is particularly effective in case of an edited scene, such as in the case of uncluttering for diminished reality chapter 5 or for possible virtual staging with addition of 3D objects, i.e., in case the final output is structural or completed with synthetic objects. In the case of pure free-form navigation chapter 3, on the other hand, several artifacts and limitations remain. In this case, indoor structural information cannot help complete any generic objects, which in some cases can generate obvious artifacts as the viewpoint changes. This situation currently limits the observer's range of motion, since in the presence of large movements, layout information and indoor

priors would not be useful in reconstructing large portions of disoccluded objects. In the future this could be addressed by combining multiple images, and increasing the generative capabilities of the synthesis network with diffusion techniques.

In our research work we also presented a framework for the automatic generation of omnidirectional stereoscopic indoor environments to be used in immersive applications (see chapter 5), especially consumed through head-mounted displays. Our method starts from a single panoramic image of an interior environment and uses data-driven architectures for depth estimation and novel view synthesis to quickly generate the images seen by both eyes during head rotation.

One of the limitations of the current approach stems from the mismatch between the resolution of the synthesized images and the achievable resolution with nowadays cameras and displays. This mismatch is currently handled by downsampling images before construction and a deep-learning-assisted upsampling before display presentation. The limitation is not due to the proposed lightweight network architecture, which promises to be scalable to much larger image sizes, but instead to the availability of training sets for the depth estimation and view synthesis networks. We plan to tackle this problem by generating higher-resolution training data, as for the free-form view-synthesis problem discussed above.

Another point regards future directions for immersive automated editing methods. Here we presented a new data-driven approach that, from an input 360° image of a furnished and cluttered indoor space automatically returns, at interactive speed, a 360° photorealistic view and depth of the same scene emptied of all furniture and other clutter. Although the application of diminished reality is self consistent, in that it allows visualizing the structure of the environment free of objects or generic clutter for various purposes, some steps are still necessary, however, to have a complete virtual staging application.First of all, it would be necessary to estimate global illumination, again from single image, in order to have the ability not only to remove objects, but to insert new ones in a photorealistic manner. In this regard several approaches exist for the future to achieve a fully editable model, such as purposed in some recent work [145].

6.3 Publications

The scientific results obtained during this PhD work also appeared in related publications, listed as journal and conference papers.

International journal papers:

- **Deep synthesis and exploration of omnidirectional stereoscopic environments from a single surround-view panoramic image.** Giovanni Pintore, Alberto Jaspe-Villanueva, Markus Hadwiger, Jens Schneider, Marco Agus, Fabio Marton, Fabio Bettio, and Enrico Gobbetti. Computers & Graphics, 2024.
  DOI: cag.2024.103907.
  Here we presented a framework for the automatic generation of omnidirectional stereoscopic indoor environments to be used in immersive applications, especially consumed through head-mounted displays. Our method starts from a single panoramic image of an interior environment and uses data-driven architectures for depth estimation and novel view synthesis to quickly generate the images seen by both eyes during head rotation. For this work, these images are combined into an omnidirectional stereo representation, which is consumed on a lightweight WebXR viewer supporting stereoscopic exploration during head rotations.
  The candidate is the lead author, contributing in all phases, including conceptualization, methodology, implementation, testing, and validation of the method.

- **Deep scene synthesis of Atlanta-world interiors from a single omnidirectional image.** Giovanni Pintore, Fabio Bettio, Marco Agus, and Enrico Gobbetti. IEEE Transactions on Visualization and Computer Graphics, 29, November 2023.
  DOI: TVCG.2023.3320219.
  Here we presented a novel deep learning approach that extracts geometric and structural information from a single panorama in order to quickly synthesize plausible panoramic images from close-by viewpoints within a workspace suitable for VR applications. Such a end-to-end approach is particularly compact and lightweight, and introduces several innovations. In particular, our novel integrated network for estimating an environment's depth and permanent structure produces elements that are crucial requirements for ensuring reliable

view synthesis. By incorporating novel domain-specific loss functions, we shift the major computational load on the training phase, and obtain an extremely lightweight network at prediction time.
The candidate is the lead author, contributing in all phases, including conceptualization, methodology, implementation, testing, and validation of the method.

- **Instant Automatic Emptying of Panoramic Indoor Scenes.** Giovanni Pintore, Marco Agus, Eva Almansa, and Enrico Gobbetti. IEEE Transactions on Visualization and Computer Graphics, 28(11), November 2022.
DOI: TVCG.2022.3202999.
Here we presented a new data-driven approach that, from an input 360° image of a furnished and cluttered indoor space automatically returns, at interactive speed, a 360° photorealistic view and depth of the same scene emptied of all furniture and other clutter. Rather than casting the problem as a simple image infilling problem, we consider the correlation between color and geometry that occurs in indoor spaces. This allows us to exploit, beside perceptual and style objective functions, geometric losses of different orders, including robust geometric pixel-wise and high-order 3D losses targeted for indoor structures, simplifying the prediction model and its computational complexity.
The candidate is the lead author, contributing in all phases, including conceptualization, methodology, implementation, testing, and validation of the method.

- **Deep Panoramic Depth Prediction and Completion for Indoor Scenes.** Giovanni Pintore, Eva Almansa, Armando Sanchez, Giorgio Vassena, and Enrico Gobbetti. Computational Visual Media 2024.
DOI: 10.1007/s41095-023-0358-0.
Here we presented a novel end-to-end deep-learning solution for rapidly estimating a dense spherical depth map of an indoor environment. Our input is a single equirectangular image registered with a sparse depth map, as provided by a variety of common capture setups. Depth is inferred by an efficient and lightweight single-branch network, which employs a dynamic gating system to process together dense visual data and sparse geometric data. We exploit the characteristics of typical man-made environments to efficiently compress multi-resolution features and find short- and long-range relations among scene parts.
The candidate is the first co-author, significantly contributing in all phases,

including conceptualization, methodology, implementation, testing, and validation of the method.

- **SPIDER: A framework for processing, editing and presenting immersive high-resolution spherical indoor scenes.** Muhammad Tukur, Giovanni Pintore, Enrico Gobbetti, Jens Schneider, and Marco Agus. Graphical Models, 128, July 2023.
  DOI: 10.1016/j.gmod.2023.101182.
  Here we presented an interactive-based image processing, editing and rendering system named SPIDER, that takes a spherical 360-degree indoor scene as input. The system is composed of a novel integrated deep learning architecture for extracting geometric and semantic information of full and empty rooms, based on gated and dilated convolutions, followed by a super-resolution module for improving the resolution of the color and depth signals.
  The candidate is the first co-author, significantly contributing in all phases, including conceptualization, methodology, implementation, testing, and validation of the method.

International conference papers:

- **PanoVerse: automatic generation of stereoscopic environments from single indoor panoramic images for Metaverse applications.** Giovanni Pintore, Alberto Jaspe-Villanueva, Markus Hadwiger, Jens Schneider, Marco Agus, and Enrico Gobbetti. . In Proc. Web3D 2023 - 28th International ACM Conference on 3D Web Technology, October 2023. DOI: 10.1145/3611314.3615914. **Honorable mention**.
  DOI: 10.1145/3611314.3615914.
  We presented a novel framework, dubbed PanoVerse, for the automatic creation and presentation of immersive stereoscopic environments from a single indoor panoramic image. Once per 360 shot, a novel data-driven architecture generates a fixed set of panoramic stereo pairs distributed around the current central view-point. Once per frame, directly on the HMD, we rapidly fuse the precomputed views to seamlessly cover the exploration workspace. Conference version of this work, awarded with honorable mention and invited for the journal publication. The method presented here generates a fixed set of panoramic stereo pairs distributed around the current central view-point, differently from the extended approach presented in the journal [39].
  The candidate is the lead author, contributing in all phases, including con-

ceptualization, methodology, implementation, testing, and validation of the method.

- **Deep scene synthesis of Atlanta-world interiors from a single omnidirectional image.** Giovanni Pintore, Fabio Bettio, Marco Agus, and Enrico Gobbetti. Proc. of IEEE ISMAR, November 2023.
  DOI: TVCG.2023.3320219.
  We presented a novel deep learning approach that extracts geometric and structural information from a single panorama in order to quickly synthesize plausible panoramic images from close-by viewpoints within a workspace suitable for VR applications. This is the conference release of the paper.
  The candidate is the presenter and lead author, contributing in all phases, including conceptualization, methodology, implementation, testing, and validation of the method.

- **Instant Automatic Emptying of Panoramic Indoor Scenes.** Giovanni Pintore, Marco Agus, Eva Almansa, and Enrico Gobbetti. Proc. of IEEE ISMAR, November 2022.
  DOI: TVCG.2022.3202999.
  We presented a new data-driven approach that, from an input 360° image of a furnished and cluttered indoor space automatically returns, at interactive speed, a 360° photorealistic view and depth of the same scene emptied of all furniture and other clutter. This is the conference release of the paper.
  The candidate is the presenter and lead author, contributing in all phases, including conceptualization, methodology, implementation, testing, and validation of the method.

Bibliography

[1] Giovanni Pintore, Claudio Mura, Fabio Ganovelli, Lizeth Fuentes-Perez, Renato Pajarola, and Enrico Gobbetti. “State-of-the-art in Automatic 3D Reconstruction of Structured Indoor Environments”. In: *Comput. Graph. Forum* 39.2 (2020), pp. 667–699 (cit. on pp. vi, viii, x, 1, 3, 4, 17, 18, 21, 22, 26, 27, 55, 81, 92, 93, 97, 125, 126, 129, 132, 133, 140).

[2] Yinda Zhang, Shuran Song, Ping Tan, and Jianxiong Xiao. “PanoContext: A Whole-Room 3D Context Model for Panoramic Scene Understanding”. In: *Proc. ECCV*. 2014, pp. 668–686 (cit. on pp. vi, viii, x, 22, 84).

[3] Kevin Matzen, Michael F. Cohen, Bryce Evans, Johannes Kopf, and Richard Szeliski. “Low-cost 360 Stereo Photography and Video Capture”. In: *ACM TOG* 36.4 (2017), 148:1–148:12 (cit. on pp. vi, viii, x, 1, 2, 5, 17, 52, 53, 78, 130, 135).

[4] Benjamin Attal, Selena Ling, Aaron Gokaslan, Christian Richardt, and James Tompkin. “MatryODShka: Real-time 6DoF video view synthesis using multi-sphere images”. In: *Proc. ECCV*. Springer. 2020, pp. 441–459 (cit. on pp. vi, viii, x, 4, 17, 23, 39, 57, 129).

[5] Michael Broxton, John Flynn, Ryan Overbeck, Daniel Erickson, Peter Hedman, Matthew Duvall, Jason Dourgarian, Jay Busch, Matt Whalen, and Paul Debevec. “Immersive light field video with a layered mesh representation”. In: *ACM Transactions on Graphics (TOG)* 39.4 (2020), pp. 86–1 (cit. on pp. vi, viii, x, 4, 17, 22, 56, 129).

[6] John Waidhofer, Richa Gadgil, Anthony Dickson, Stefanie Zollmann, and Jonathan Ventura. “PanoSynthVR: Toward Light-weight 360-Degree View Synthesis from a Single Panoramic Input”. In: *Proc. ISMAR*. IEEE. 2022, pp. 584–592 (cit. on pp. vi, viii, x, 2–4, 17, 19, 23, 35, 37, 39, 40, 48, 49, 52, 57, 125, 126, 129, 130, 132, 134, 135).

[7] Shohei Mori, Sei Ikeda, and Hideo Saito. “A survey of diminished reality: Techniques for visually concealing, eliminating, and seeing through real objects”. In: *IPSJ Transactions on Computer Vision and Applications* 9.1 (2017), pp. 1–14 (cit. on pp. vi, viii, x, 82).

[8] Satoshi Ikehata, Hang Yang, and Yasutaka Furukawa. “Structured Indoor Modeling”. In: *Proc. ICCV*. 2015, pp. 1323–1331 (cit. on p. 1).

[9] Giovanni Pintore, Fabio Ganovelli, Enrico Gobbetti, and Roberto Scopigno. “Mobile Mapping and Visualization of Indoor Structures to Simplify Scene Understanding and Location Awareness”. In: *Computer Vision – ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II*. Springer, 2016, pp. 130–145 (cit. on p. 2).

[10] Tero Jokela, Jarno Ojala, and Kaisa Väänänen. “How people use 360-degree cameras”. In: *Proc. International Conference on Mobile and Ubiquitous Multimedia*. 2019, pp. 1–10 (cit. on pp. 2, 4, 10, 52, 78, 125, 129).

[11] Mai Xu, Chen Li, Shanyi Zhang, and Patrick Le Callet. “State-of-the-art in 360° video/image processing: Perception, assessment and compression”. In: *IEEE Journal of Selected Topics in Signal Processing* 14.1 (2020), pp. 5–26 (cit. on pp. 2, 17, 52, 78).

[12] Mohamad Zaidi Sulaiman, Mohd Nasiruddin Abdul Aziz, Mohd Haidar Abu Bakar, Nur Akma Halili, and Muhammad Asri Azuddin. “Matterport: virtual tour as a new marketing approach in real estate business during pandemic COVID-19”. In: *Proc. IMDES*. 2020, pp. 221–226 (cit. on pp. 2, 52, 78).

[13] Haiwei Dong and Jeannie S. A. Lee. “The Metaverse From a Multimedia Communications Perspective”. In: *IEEE MultiMedia* 29.4 (2022), pp. 123–127. DOI: 10.1109/MMUL.2022.3217627 (cit. on pp. 2, 52, 125).

[14] Chuhang Zou, Jheng Wei Su, Chi Han Peng, Alex Colburn, Qi Shan, Peter Wonka, Hung Kuo Chu, and Derek Hoiem. “Manhattan Room Layout Reconstruction from a Single 360 Image: A Comparative Study of State-of-the-Art Methods”. In: *International Journal of Computer Vision* 129 (2021), pp. 1410–1431 (cit. on pp. 3, 11, 13, 35, 38, 85, 127).

[15] Giovanni Pintore, Marco Agus, Eva Almansa, Jens Schneider, and Enrico Gobbetti. “SliceNet: deep dense depth estimation from a single indoor panorama using a slice-based representation”. In: *Proc. CVPR*. 2021, pp. 11536–11545 (cit. on pp. 3, 14, 18, 19, 21, 26, 36, 38, 39, 45, 54, 55, 60, 70, 71, 84, 90, 96, 97, 126, 128, 133, 134, 136).

[16] Álan LV Guedes, G de A Roberto, Pascal Frossard, Sérgio Colcher, and Simone Diniz Junqueira Barbosa. “Subjective evaluation of 360-degree sensory experiences”. In: *Proc. IEEE MMSP*. 2019, pp. 1–6 (cit. on pp. 4, 78, 129).

[17] Jiale Xu, Jia Zheng, Yanyu Xu, Rui Tang, and Shenghua Gao. “Layout-Guided Novel View Synthesis From a Single Indoor Panorama”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021 (cit. on pp. 4, 13, 15, 17, 22, 23, 30, 33–36, 38–41, 56, 68, 69, 71, 74, 127, 129, 132).

[18] Ching-Yu Hsu, Cheng Sun, and Hwann-Tzong Chen. “Moving in a 360 world: Synthesizing panoramic parallaxes from a single panorama”. In: *arXiv preprint arXiv:2106.10859* (2021) (cit. on pp. 4, 17, 19, 23, 39–41, 129, 132, 134).

[19] Richard Tucker and Noah Snavely. “Single-view view synthesis with multiplane images”. In: *Proc. CVPR*. 2020, pp. 551–560 (cit. on pp. 4, 17, 22, 23, 37, 39, 71, 72, 129, 132).

[20] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. “Nerf: Representing scenes as neural radiance fields for view synthesis”. In: *Communications of the ACM* 65.1 (2021), pp. 99–106 (cit. on pp. 4, 17, 22, 23, 129, 132).

[21] Haijing Jia, Hong Yi, Hirochika Fujiki, Hengzhi Zhang, Wei Wang, and Makoto Odamaki. “3D Room Layout Recovery Generalizing across Manhattan and Non-Manhattan Worlds”. In: *Proc. CVPR Workshops*. 2022, pp. 5188–5197. DOI: 10.1109/CVPRW56347.2022.00567 (cit. on pp. 4, 17, 130, 132).

[22] Yining Zhao, Chao Wen, Zhou Xue, and Yue Gao. “3D Room Layout Estimation from a Cubemap of Panorama Image via Deep Manhattan Hough Transform”. In: *Proc. ECCV*. Springer. 2022, pp. 637–654 (cit. on pp. 4, 17, 22, 130, 132).

[23] Zhigang Jiang, Zhongzheng Xiang, Jinhua Xu, and Ming Zhao. “LGT-Net: Indoor Panoramic Room Layout Estimation with Geometry-Aware Transformer Network”. In: *Proc. CVPR*. 2022, pp. 1654–1663 (cit. on pp. 4, 17, 22, 130, 132).

[24] Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. “HorizonNet: Learning room layout with 1D representation and pano stretch data augmentation”. In: *Proc. CVPR*. 2019, pp. 1047–1056 (cit. on pp. 4, 18, 19, 21, 22, 28, 36, 41, 45, 130, 132, 134).

[25] Shmuel Peleg and Moshe Ben-Ezra. “Stereo panorama with a single camera”. In: *Proc. CVPR*. 1999, pp. 395–401. DOI: 10.1109/CVPR.1999.786969 (cit. on pp. 5, 53, 58, 130, 135).

[26] Christian Richardt, James Tompkin, and Gordon Wetzstein. “Capture, Reconstruction, and Representation of the Visual Real World for Virtual Reality”. In: *Real VR – Immersive Digital Reality: How to Import the Real World into Head-Mounted Immersive Displays*. Ed. by Marcus Magnor and Alexander Sorkine-Hornung. Springer International Publishing, 2020, pp. 3–32. DOI: 10.1007/978-3-030-41816-8_1 (cit. on pp. 5, 53, 130, 135).

[27] Tobias Bertel, Mingze Yuan, Reuben Lindroos, and Christian Richardt. “OmniPhotos: Casual 360° VR Photography”. In: *ACM TOG* 39.6 (Dec. 2020), 266:1–266:12. DOI: 10.1145/3414685.3417770 (cit. on pp. 5, 22, 53, 56, 130, 135).

[28] Edward Zhang, Michael F. Cohen, and Brian Curless. “Emptying, Refurnishing, and Relighting Indoor Spaces”. In: *ACM TOG* 35.6 (2016), 174:1–174:14 (cit. on pp. 5, 78, 82, 131).

[29] Sanni Siltanen, Henrikki Saraspää, and Jari Karvonen. “[DEMO] A complete interior design solution with diminished reality”. In: *Proc. ISMAR*. 2014, pp. 371–372. DOI: 10.1109/ISMAR.2014.6948494 (cit. on pp. 5, 78, 82, 131).

[30] Sanni Siltanen. “Diminished reality for augmented reality interior design”. In: *The Visual Computer* 33.2 (2017), pp. 193–208 (cit. on pp. 5, 78, 82, 131).

[31] Theophilus Teo, Louise Lawrence, Gun A. Lee, Mark Billinghurst, and Matt Adcock. “Mixed Reality Remote Collaboration Combining 360 Video and 3D Reconstruction”. In: *Proc. CHI*. 2019, 201:1—201:14. DOI: 10.1145/3290605.3300431 (cit. on pp. 5, 78, 131).

[32] Vasileios Gkitsas, Vladimiros Sterzentsenko, Nikolaos Zioulis, Georgios Albanis, and Dimitrios Zarpalas. “PanoDR: Spherical Panorama Diminished Reality for Indoor Scenes”. In: *Proc. CVPR*. 2021, pp. 3716–3726 (cit. on pp. 6, 26, 29, 45, 60, 79, 80, 82, 84, 85, 87, 93–96, 131, 137–139).

[33] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. “Free-form image inpainting with gated convolution”. In: *Proc. ICCV*. 2019, pp. 4471–4480 (cit. on pp. 6, 30, 36, 61, 64, 79, 80, 83, 84, 88, 94–96, 131, 138, 139).

[34] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. “Pluralistic image completion”. In: *Proc. CVPR*. 2019, pp. 1438–1447 (cit. on pp. 6, 79, 80, 83, 84, 95, 96, 131, 138, 139).

[35] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. “Semantic image synthesis with spatially-adaptive normalization”. In: *Proc. CVPR*. 2019, pp. 2337–2346 (cit. on pp. 6, 79, 80, 84, 131, 138, 139).

[36] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. “Sean: Image synthesis with semantic region-adaptive normalization”. In: *Proc. CVPR*. 2020, pp. 5104–5113 (cit. on pp. 6, 30, 79, 80, 82, 84–86, 96, 131, 138, 139).

[37] Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao. “Recurrent feature reasoning for image inpainting”. In: *Proc. CVPR*. 2020, pp. 7760–7768 (cit. on pp. 6, 79, 80, 83, 84, 94–96, 131, 138, 139).

[38] Giovanni Pintore, Eva Almansa, Armando Sanchez, Giorgio Vassena, and Enrico Gobbetti. “Deep Panoramic Depth Prediction and Completion for Indoor Scenes”. In: *Computational Visual Media* (2023). DOI: 10.1007/s41095-023-0358-0 (cit. on pp. 7, 103, 105, 141, 143).

[39] Giovanni Pintore, Alberto Jaspe-Villanueva, Markus Hadwiger, Jens Schneider, Marco Agus, Fabio Marton, Fabio Bettio, and Enrico Gobbetti. “Deep synthesis and exploration of omnidirectional stereoscopic environments from a single surround-view panoramic image”. In: *Computers & Graphics* 119 (2024), p. 103907. DOI: https://doi.org/10.1016/j.cag.2024.103907 (cit. on pp. 7, 76, 104, 110, 142).

[40] Giovanni Pintore, Marco Agus, Eva Almansa, and Enrico Gobbetti. “Instant Automatic Emptying of Panoramic Indoor Scenes”. In: *IEEE Transactions on Visualization and Computer Graphics* 28.11 (2022), pp. 3629–3639. DOI: 10.1109/TVCG.2022.3202999 (cit. on pp. 8, 29, 50, 101, 105, 142).

[41] Matterport. *Matterport3D*. https://github.com/niessner/Matterport. [Accessed: 2023-03-03]. 2017 (cit. on pp. 10, 35, 91, 92, 94, 97).

[42] Marc Eder, Mykhailo Shvets, John Lim, and Jan-Michael Frahm. “Tangent Images for Mitigating Spherical Distortion”. In: *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020 (cit. on p. 10).

[43] Giovanni Pintore, Marco Agus, and Enrico Gobbetti. *CVPR2023 Tutorial on Automatic 3D modeling of indoor structures from panoramic imagery*. http://vic.crs4.it/vic/cvpr2023-tutorial-pano. 2023 (cit. on p. 11).

[44] Erick Delage, Honglak Lee, and Andrew Y Ng. “A dynamic bayesian network model for autonomous 3d reconstruction from a single indoor image”. In: *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR’06)*. Vol. 2. IEEE. 2006, pp. 2418–2428 (cit. on p. 11).

[45] V. Hedau, D. Hoiem, and D. Forsyth. “Recovering the spatial layout of cluttered rooms”. In: *Proc. ICCV*. 2009, pp. 1849–1856 (cit. on p. 11).

[46] David C Lee, Martial Hebert, and Takeo Kanade. “Geometric reasoning for single image structure recovery”. In: *2009 IEEE conference on computer vision and pattern recognition*. IEEE. 2009, pp. 2136–2143 (cit. on p. 11).

[47] James M Coughlan and Alan L Yuille. “Manhattan world: Compass direction from a single image by bayesian inference”. In: *Proceedings of the seventh IEEE international conference on computer vision*. Vol. 2. IEEE. 1999, pp. 941–947 (cit. on p. 11).

[48] Giovanni Pintore, Marco Agus, and Enrico Gobbetti. “AtlantaNet: Inferring the 3D Indoor Layout from a Single 360 Image Beyond the Manhattan World Assumption”. In: *Proc. ECCV*. 2020, pp. 432–448 (cit. on pp. 11, 22, 36, 38, 39).

[49] Claudio Mura, Oliver Mattausch, Alberto Jaspe Villanueva, Enrico Gobbetti, and Renato Pajarola. “Automatic Room Detection and Reconstruction in Cluttered Indoor Environments with Complex Room Layouts”. In: *Computers & Graphics* 44 (Nov. 2014), pp. 20–32. DOI: 10.1016/j.cag.2014.07.005 (cit. on p. 11).

[50] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. “Image style transfer using convolutional neural networks”. In: *Proc. CVPR*. 2016, pp. 2414–2423. DOI: 10.1109/CVPR.2016.265 (cit. on pp. 14, 32, 33, 65, 83, 90, 128).

[51] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. “Automatic differentiation in PyTorch”. In: *Proc. NIPS Workshop on Autodiff*. 2017 (cit. on pp. 14, 34, 91, 93).

[52] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. “Structured3D: A Large Photo-realistic Dataset for Structured 3D Modeling”. In: *Proc. ECCV*. 2020, pp. 519–535. DOI: 978-3-030-58545-7_30 (cit. on pp. 15, 34, 35, 38, 40, 49, 68, 71, 86, 89, 91–93, 95, 96, 100).

[53] Mohamed Aly and Jean-Yves Bouguet. “Street view goes indoors: Automatic pose estimation from uncalibrated unordered spherical panoramas”. In: *Proc. WACV*. 2012, pp. 1–8 (cit. on p. 17).

[54] Viktor Kelkkanen, Markus Fiedler, and David Lindero. “Synchronous remote rendering for VR”. In: *Int. Journal of Computer Games Technology* 2021 (2021), pp. 1–16 (cit. on p. 18).

[55] Chuhang Zou, Alex Colburn, Qi Shan, and Derek Hoiem. “LayoutNet: Reconstructing the 3D Room Layout from a Single RGB Image”. In: *Proc. CVPR*. 2018, pp. 2051–2059 (cit. on pp. 19, 21, 22, 36, 38, 134).

[56] Giovanni Pintore, Eva Almansa, Marco Agus, and Enrico Gobbetti. “Deep3DLayout: 3D Reconstruction of an Indoor Layout from a Spherical Panoramic Image”. In: *ACM TOG* 40.6 (2021), 250:1–250:12 (cit. on pp. 19, 22, 26, 27, 45, 85, 134).

[57] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. “Synsin: End-to-end view synthesis from a single image”. In: *Proc. CVPR*. 2020, pp. 7467–7477 (cit. on pp. 20, 23, 30, 35, 39, 41, 71, 72).

[58] Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, and Petros Daras. “OmniDepth: Dense Depth Estimation for Indoors Spherical Panoramas”. In: *Proc. ECCV*. 2018, pp. 453–471 (cit. on pp. 21, 55).

[59] Yu-Chuan Su and Kristen Grauman. “Learning Spherical Convolution for Fast Features from 360 Imagery”. In: *Advances in Neural Information Processing Systems 30*. 2017, pp. 529–539 (cit. on pp. 21, 55).

[60] Keisuke Tateno, Nassir Navab, and Federico Tombari. “Distortion-Aware Convolutional Filters for Dense Prediction in Panoramic Images”. In: *Proc. ECCV*. 2018, pp. 732–750 (cit. on pp. 21, 55).

[61] Gregoire Payen de La Garanderie, Amir Atapour Abarghouei, and Toby P. Breckon. "Eliminating the Blind Spot: Adapting 3D Object Detection and Monocular Depth Estimation to 360 Panoramic Imagery". In: *Proc. ECCV*. 2018, pp. 812–830 (cit. on p. 21).

[62] Y. Su and K. Grauman. "Kernel Transformer Networks for Compact Spherical Convolution". In: *Proc. CVPR*. 2019, pp. 9434–9443 (cit. on p. 21).

[63] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. "BiFuse: Monocular 360 Depth Estimation via Bi-Projection Fusion". In: *Proc. CVPR*. 2020, pp. 462–471 (cit. on pp. 21, 36, 84, 90).

[64] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. "Deeper Depth Prediction with Fully Convolutional Residual Networks". In: *Proc. 3DV*. 2016, pp. 239–248 (cit. on p. 21).

[65] Cheng Sun, Min Sun, and Hwann-Tzong Chen. "HoHoNet: 360 Indoor Holistic Understanding with Latent Horizontal Features". In: *Proc. CVPR*. 2021, pp. 2573–2582 (cit. on pp. 21, 45, 54, 70, 71, 84, 96, 97, 136).

[66] Yuyan Li, Yuliang Guo, Zhixin Yan, Xinyu Huang, Ye Duan, and Liu Ren. "Omnifusion: 360 monocular depth estimation via geometry-aware fusion". In: *Proc. CVPR workshop on computer vision for augmented and virtual reality*. 2022, pp. 2801–2810 (cit. on pp. 21, 36, 71).

[67] Manuel Rey-Area, Mingze Yuan, and Christian Richardt. "360MonoDepth: High-Resolution 360deg Monocular Depth Estimation". In: *Proc. CVPR*. 2022, pp. 3762–3772 (cit. on pp. 21, 36).

[68] Chuhang Zou, Jheng-Wei Su, Chi-Han Peng, Alex Colburn, Qi Shan, Peter Wonka, Hung-Kuo Chu, and Derek Hoiem. "3D Manhattan Room Layout Reconstruction from a Single 360 Image". In: *ArXiv e-print arXiv:1910.04099* (2019) (cit. on pp. 21, 22).

[69] Shang-Ta Yang, Fu-En Wang, Chi-Han Peng, Peter Wonka, Min Sun, and Hung-Kuo Chu. "DuLa-Net: A Dual-Projection Network for Estimating Room Layouts from a Single RGB Panorama". In: *Proc. CVPR*. 2019, pp. 3363–3372 (cit. on pp. 22, 27).

[70] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. "LED2-Net: Monocular 360 Layout Estimation via Differentiable Depth Rendering". In: *Proc. CVPR*. 2021, pp. 12956–12965 (cit. on p. 22).

[71] David C Lee, Martial Hebert, and Takeo Kanade. "Geometric reasoning for single image structure recovery". In: *Proc. CVPR*. 2009, pp. 2136–2143 (cit. on p. 22).

[72] Tobias Bertel, Neill DF Campbell, and Christian Richardt. "MegaParallax: Casual 360° panoramas with motion parallax". In: *IEEE TVCG* 25.5 (2019), pp. 1828–1835 (cit. on pp. 22, 82).

[73] Bicheng Luo, Feng Xu, Christian Richardt, and Jun-Hai Yong. "Parallax360: Stereoscopic 360° Scene Representation for Head-Motion Parallax". In: *IEEE TVCG* 24.4 (2018), pp. 1545–1553. DOI: 10.1109/TVCG.2018.2794071 (cit. on pp. 22, 56).

[74] Jingwei Huang, Zhili Chen, Duygu Ceylan, and Hailin Jin. "6-DOF VR videos with a single 360-camera". In: *Proc. IEEE VR*. 2017, pp. 37–44. DOI: 10.1109/VR.2017.7892229 (cit. on pp. 22, 56).

[75] Peter Hedman, Suhib Alsisan, Richard Szeliski, and Johannes Kopf. “Casual 3D photography”. In: *ACM Transactions on Graphics (TOG)* 36.6 (2017), pp. 1–15 (cit. on p. 22).

[76] Peter Hedman and Johannes Kopf. “Instant 3D Photography”. In: *ACM TOG*. 37.4 (2018), 101:1–101:12. DOI: 10.1145/3197517.3201384 (cit. on pp. 22, 56).

[77] Kai-En Lin, Zexiang Xu, Ben Mildenhall, Pratul P Srinivasan, Yannick Hold-Geoffroy, Stephen DiVerdi, Qi Sun, Kalyan Sunkavalli, and Ravi Ramamoorthi. “Deep multi depth panoramas for view synthesis”. In: *Proc. ECCV*. Springer. 2020, pp. 328–344 (cit. on p. 22).

[78] Ana Serrano, Incheol Kim, Zhili Chen, Stephen DiVerdi, Diego Gutierrez, Aaron Hertzmann, and Belen Masia. “Motion parallax for 360° RGBD video”. In: *IEEE TVCG* 25.5 (2019), pp. 1817–1827. DOI: 10.1109/TVCG.2019.2898757 (cit. on pp. 22, 56).

[79] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. “Stereo Magnification: Learning View Synthesis Using Multiplane Images”. In: *ACM TOG* 37.4 (2018), 68:1–68:12. DOI: 10.1145/3197517.3201323 (cit. on pp. 22, 57).

[80] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. “Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data”. In: *Proc. ICCVW*. 2021 (cit. on pp. 25, 37, 42, 44, 68).

[81] K. He, X. Zhang, S. Ren, and J. Sun. “Deep Residual Learning for Image Recognition”. In: *Proc. CVPR*. 2016, pp. 770–778 (cit. on pp. 26, 45).

[82] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017, pp. 6000–6010 (cit. on pp. 26, 46).

[83] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *Proc. MICCAI*. 2015, pp. 234–241 (cit. on pp. 28, 86).

[84] Shubham Tulsiani, Richard Tucker, and Noah Snavely. “Layer-structured 3d scene inference via view synthesis”. In: *Proc. ECCV*. 2018, pp. 302–317 (cit. on pp. 30, 42, 64).

[85] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. “EdgeConnect: Structure Guided Image Inpainting using Edge Prediction”. In: *Proc. ICCVW*. 2019, pp. 3265–3274. DOI: 10.1109/ICCVW.2019.00408 (cit. on pp. 30, 31, 36, 83).

[86] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. “Globally and Locally Consistent Image Completion”. In: *ACM Trans. Graph.* 36.4 (2017). ISSN: 0730-0301. DOI: 10.1145/3072959.3073659 (cit. on pp. 31, 48, 60, 61, 83, 87, 89).

[87] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. “Generative image inpainting with contextual attention”. In: *Proc. CVPR*. 2018, pp. 5505–5514 (cit. on pp. 31, 32, 60, 65, 87).

[88] Fisher Yu and Vladlen Koltun. “Multi-Scale Context Aggregation by Dilated Convolutions”. In: *Proc. ICLR (Poster)*. 2016, 1:1–1:13 (cit. on pp. 31, 47, 61, 83, 87, 88).

[89] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. “Contextual residual aggregation for ultra high-resolution image inpainting”. In: *Proc. CVPR*. 2020, pp. 7508–7517 (cit. on pp. 31, 47, 61, 80, 88, 139).

[90] Sophie Lambert-Lacroix and Laurent Zwald. “The adaptive BerHu penalty in robust regression”. In: *Journal of Nonparametric Statistics* 28 (2016), pp. 1–28 (cit. on pp. 32, 33, 62, 90).

[91] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014) (cit. on pp. 32, 65, 90, 93).

[92] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. “Image-to-image translation with conditional adversarial networks”. In: *Proc. CVPR*. 2017, pp. 1125–1134 (cit. on pp. 32, 65, 80, 83, 84, 99, 139).

[93] Nikolaos Zioulis, Antonis Karakottas, Dimitris Zarpalas, Federic Alvarez, and Petros Daras. “Spherical View Synthesis for Self-Supervised 360° Depth Estimation”. In: *Proc. 3DV*. 2019, pp. 690–699 (cit. on pp. 33, 64).

[94] Po Kong Lai, Shuang Xie, Jochen Lang, and Robert Laganière. “Real-time panoramic depth maps from omni-directional stereo images for 6 dof videos in virtual reality”. In: *Proc. IEEE VR*. IEEE. 2019, pp. 405–412 (cit. on p. 35).

[95] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *ArXiv e-print arXiv:1412.6980* (2014) (cit. on pp. 35, 69, 93).

[96] NVIDIA. *nvJPEG Libraries: GPU-accelerated JPEG decoder, encoder and transcoder*. `https://developer.nvidia.com/nvjpeg`. [Accessed: 2023-06-06]. 2023 (cit. on p. 37).

[97] Jae-Ho Nah. “QuickETC2: Fast ETC2 Texture Compression Using Luma Differences”. In: *ACM Trans. Graph.* 39.6 (2020) (cit. on p. 37).

[98] Lei Jin, Yanyu Xu, Jia Zheng, Junfei Zhang, Rui Tang, Shugong Xu, Jingyi Yu, and Shenghua Gao. “Geometric Structure Based and Regularized Depth Estimation From 360 Indoor Imagery”. In: *Proc. CVPR*. 2020, pp. 889–898 (cit. on pp. 38, 39, 96, 97).

[99] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. “Image quality assessment: from error visibility to structural similarity”. In: *IEEE TIP* 13.4 (2004), pp. 600–612 (cit. on pp. 39, 95).

[100] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. “The unreasonable effectiveness of deep features as a perceptual metric”. In: *Proc. CVPR*. 2018, pp. 586–595 (cit. on pp. 39, 65, 72, 95).

[101] Xin Deng, Hao Wang, Mai Xu, Yichen Guo, Yuhang Song, and Li Yang. “LAU-Net: Latitude Adaptive Upscaling Network for Omnidirectional Image Super-Resolution”. In: *Proc. CVPR*. 2021, pp. 9189–9198 (cit. on p. 42).

[102] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. “Instance normalization: The missing ingredient for fast stylization”. In: *arXiv preprint arXiv:1607.08022* (2016) (cit. on pp. 47, 60, 87).

[103] Muhammad Tukur, Giovanni Pintore, Enrico Gobbetti, Jens Schneider, and Marco Agus. “SPIDER: A framework for processing, editing and presenting immersive high-resolution spherical indoor scenes”. In: *Graphical Models* 128 (2023), 101182:1–101182:11 (cit. on pp. 50, 56).

[104] Giovanni Pintore, Fabio Bettio, Marco Agus, and Enrico Gobbetti. “Deep scene synthesis of Atlanta-world interiors from a single omnidirectional image”. In: *IEEE TVCG* 29 (2023). DOI: 10.1109/TVCG.2023.3320219 (cit. on pp. 50, 56, 57).

[105] Giovanni Pintore, Alberto Jaspe Villanueva, Markus Hadwiget, Enrico Gobbetti, Jens Schneider, and Marco Agus. “PanoVerse: automatic generation of stereoscopic environments from single indoor panoramic images for Metaverse applications”. In: *Proc. ACM Web3D*. 2023, 2:1–2:10. DOI: 10.1145/3611314.3615914 (cit. on pp. 54, 56, 57, 66, 76, 137).

[106] Shaohua Gao, Kailun Yang, Hao Shi, Kaiwei Wang, and Jian Bai. “Review on panoramic imaging and its applications in scene understanding”. In: *IEEE Transactions on Instrumentation and Measurement* 71 (2022), pp. 1–34 (cit. on p. 55).

[107] Thiago L. T. da Silveira, Paulo G. L. Pinto, Jeffri Murrugarra-Llerena, and Cláudio R. Jung. “3D Scene Geometry Estimation from 360° Imagery: A Survey”. In: *ACM Comput. Surv.* 55.4 (2022), 68:1–68:39. DOI: 10.1145/3519021 (cit. on p. 55).

[108] Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. “SphereNet: Learning spherical representations for detection and classification in omnidirectional images”. In: *Proc. ECCV*. 2018, pp. 518–533. DOI: 10.1007/978-3-030-01240-3_32 (cit. on p. 55).

[109] Daniel Martin, Ana Serrano, and Belen Masia. “Panoramic convolutions for 360° single-image saliency prediction”. In: *Proc. CVPR workshop on computer vision for augmented and virtual reality*. 2020, pp. 1–4 (cit. on p. 55).

[110] Manuel Rey-Area, Mingze Yuan, and Christian Richardt. “360MonoDepth: High-Resolution 360° Monocular Depth Estimation”. In: *Proc. CVPR*. 2022, pp. 3752–3762. DOI: 10.1109/CVPR52688.2022.00374 (cit. on pp. 55, 60, 71).

[111] Giovanni Pintore, Eva Almansa, Armando Sanchez, Giorgio Vassena, and Enrico Gobbetti. “Deep Panoramic Depth Prediction and Completion for Indoor Scenes”. In: *Computational Visual Media* (2024). DOI: 10.1007/s41095-023-0358-0 (cit. on p. 55).

[112] M. Tukur, Giovanni Pintore, Enrico Gobbetti, Jens Schneider, and Marco Agus. “SPIDER: Spherical Indoor Depth Renderer”. In: *Proc. Smart Tools and Applications in Graphics (STAG)*. 2022, pp. 131–138. DOI: 10.2312/stag.20221267 (cit. on p. 56).

[113] Marc Comino Trinidad, Ricardo Martin Brualla, Florian Kainz, and Janne Kontkanen. “Multi-view image fusion”. In: *Proc. ICCV*. 2019, pp. 4101–4110. DOI: 10.1109/ICCV.2019.00420 (cit. on p. 56).

[114] Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. “FILM: Frame interpolation for large motion”. In: *Proc. ECCV*. 2022, pp. 250–266. DOI: 10.1007/978-3-031-20071-7_15 (cit. on p. 56).

[115] Kai-En Lin, Zexiang Xu, Ben Mildenhall, Pratul P. Srinivasan, Yannick Hold-Geoffroy, Stephen DiVerdi, Qi Sun, Kalyan Sunkavalli, and Ravi Ramamoorthi. "Deep Multi Depth Panoramas for View Synthesis". In: *Proc. ECCV*. Ed. by Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm. 2020, pp. 328–344. DOI: 10.1007/978-3-030-58601-0_20 (cit. on p. 56).

[116] Richard Tucker and Noah Snavely. "Single-View View Synthesis With Multiplane Images". In: *Proc. CVPR*. 2020, pp. 548–557. DOI: 10.1109/CVPR42600.2020.00063. (cit. on p. 57).

[117] Qinbo Li and Nima Khademi Kalantari. "Synthesizing Light Field From a Single Image with Variable MPI and Two Network Fusion". In: *ACM TOG* 39.6 (2020), 229:1–229:10. DOI: 10.1145/3414685.3417785 (cit. on p. 57).

[118] Paul Bourke. "Capturing omni-directional stereoscopic spherical projections with a single camera". In: *Proc. IEEE VSMM*. 2010, pp. 179–183. DOI: 10.1109/VSMM.2010.5665988 (cit. on p. 57).

[119] Hoang Le and Feng Liu. "Appearance Flow Completion for Novel View Synthesis". In: *Computer Graphics Forum* 38.7 (2019), pp. 555–565. DOI: https://doi.org/10.1111/cgf.13860 (cit. on p. 57).

[120] Paul Rademacher and Gary Bishop. "Multiple-center-of-projection images". In: *Proc. SIGGRAPH*. 1998, pp. 199–206. DOI: 10.1145/280814.280871 (cit. on p. 57).

[121] Thomas Marrinan and Michael E Papka. "Real-time omnidirectional stereo rendering: generating 360° surround-view panoramic images for comfortable immersive viewing". In: *IEEE TVCG* 27.5 (2021), pp. 2587–2596. DOI: 10.1109/TVCG.2021.3067780 (cit. on pp. 58, 75).

[122] Bipul Mohanto, ABM Tariqul Islam, Enrico Gobbetti, and Oliver Staadt. "An integrative view of foveated rendering". In: *Computers & Graphics* 102 (2022), pp. 474–501. DOI: 10.1016/j.cag.2021.10.010 (cit. on p. 58).

[123] Varun Gupta, Rajat Sadana, and Swastikaa Moudgil. "Image style transfer using convolutional neural networks based on transfer learning". In: *International Journal of Computational Systems Engineering* 5.1 (2019), pp. 53–60. DOI: 10.1504/IJCSYSE.2019.098418 (cit. on pp. 80, 139).

[124] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. "Spectral normalization for generative adversarial networks". In: *arXiv preprint arXiv:1802.05957* (2018). DOI: arXiv:1802.05957 (cit. on pp. 80, 83, 84, 139).

[125] Jingyuan Li, Fengxiang He, Lefei Zhang, Bo Du, and Dacheng Tao. "Progressive Reconstruction of Visual Structure for Image Inpainting". In: *Proc. ICCV*. 2019, pp. 5961–5970 (cit. on pp. 80, 83).

[126] W. Yin, Y. Liu, C. Shen, and Y. Yan. "Enforcing Geometric Constraints of Virtual Normal for Depth Prediction". In: *Proc. ICCV*. 2019, pp. 5683–5692 (cit. on pp. 81, 90, 140).

[127] Hanseob Kim, TaeHyung Kim, Myungho Lee, Gerard Jounghyun Kim, and Jae-In Hwang. "Don't Bother Me: How to Handle Content-Irrelevant Objects in Handheld Augmented Reality". In: *Proc. VRST*. 2020. DOI: 10.1145/3385956.3418948 (cit. on p. 82).

[128] Yuichiro Takeuchi and Ken Perlin. “ClayVision: The (Elastic) Image of the City”. In: *Proc. CHI*. 2012, pp. 2411–2420. DOI: 10.1145/2207676.2208404 (cit. on p. 82).

[129] H. Sasanuma, Y. Manabe, and N. Yata. “Diminishing Real Objects and Adding Virtual Objects Using a RGB-D Camera”. In: *Proc. ISMAR-Adjunct*. 2016, pp. 117–120. DOI: 10.1109/ISMAR-Adjunct.2016.0055 (cit. on p. 82).

[130] Shohei Mori, Fumihisa Shibata, Asako Kimura, and Hideyuki Tamura. “Efficient Use of Textured 3D Model for Pre-observation-based Diminished Reality”. In: *Proc. ISMAR Workshops*. 2015, pp. 32–39. DOI: 10.1109/ISMARW.2015.16 (cit. on p. 82).

[131] Glen Queguiner, Matthieu Fradet, and Mohammad Rouhani. “Towards Mobile Diminished Reality”. In: *Proc. ISMAR-Adjunct*. 2018, pp. 226–231. DOI: 10.1109/ISMAR-Adjunct.2018.00073 (cit. on p. 82).

[132] Siim Meerits and Hideo Saito. “Real-time diminished reality for dynamic scenes”. In: *Proc. ISMAR Workshops*. 2015, pp. 53–59 (cit. on p. 82).

[133] Norihiko Kawai, Tomokazu Sato, and Naokazu Yokoya. “Diminished reality based on image inpainting considering background geometry”. In: *IEEE TVCG* 22.3 (2015), pp. 1236–1247 (cit. on p. 82).

[134] Yuka Namboku and Hiroki Takahashi. “Diminished reality in textureless scenes”. In: *International Workshop on Advanced Imaging Technology (IWAIT)*. Vol. 11515. SPIE, 2020, pp. 379 –384. DOI: 10.1117/12.2566248 (cit. on p. 82).

[135] David Lindlbauer and Andy D Wilson. “Remixed reality: manipulating space and time in augmented reality”. In: *Proc. CHI*. 2018, 129:1–128:13 (cit. on p. 82).

[136] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. “Context Encoders: Feature Learning by Inpainting”. In: *Proc. CVPR*. 2016 (cit. on p. 83).

[137] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. “Image Inpainting for Irregular Holes Using Partial Convolutions”. In: *Proc. ECCV*. 2018, pp. 89–105 (cit. on pp. 83, 88).

[138] Chuan Li and Michael Wand. “Precomputed real-time texture synthesis with markovian generative adversarial networks”. In: *European conference on computer vision*. Springer. 2016, pp. 702–716 (cit. on p. 83).

[139] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. “Learning pyramid-context encoder network for high-quality image inpainting”. In: *Proc. CVPR*. 2019, pp. 1486–1494 (cit. on p. 83).

[140] Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. “Rethinking image inpainting via a mutual encoder-decoder with feature equalizations”. In: *Proc. ECCV*. 2020, pp. 725–741 (cit. on p. 83).

[141] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. “Perceptual losses for real-time style transfer and super-resolution”. In: *Proc. ECCV*. 2016, pp. 694–711 (cit. on pp. 83, 90).

[142] Qifeng Chen and Vladlen Koltun. “Photographic image synthesis with cascaded refinement networks”. In: *Proc. ICCV*. 2017, pp. 1511–1520 (cit. on p. 84).

[143] A. Saxena, M. Sun, and A. Y. Ng. “Make3D: Learning 3D Scene Structure from a Single Still Image”. In: *IEEE TPAMI* 31.5 (2009), pp. 824–840 (cit. on p. 84).

[144] Giovanni Pintore, Ruggero Pintus, Fabio Ganovelli, Roberto Scopigno, and Enrico Gobbetti. “Recovering 3D existing-conditions of indoor structures from spherical images”. In: *Computers & Graphics* 77 (Dec. 2018), pp. 16–29. DOI: 10.1016/j.cag.2018.09.013 (cit. on p. 86).

[145] Tiancheng Zhi, Bowei Chen, Ivaylo Boyadzhiev, Sing Bing Kang, Martial Hebert, and Srinivasa G. Narasimhan. “Semantically supervised appearance decomposition for virtual staging from a single panorama”. In: *ACM Trans. Graph.* 41.4 (2022). ISSN: 0730-0301. DOI: 10.1145/3528223.3530148 (cit. on pp. 107, 145).

A

Sinopsis (thesis summary in Spanish)

La reconstrucción y el modelado 3D automáticos de entornos interiores han atraído cada vez más investigación en los últimos años, convirtiéndose en un tema emergente bien definido tanto en gráficos por ordenador como en visión por computador. Gran parte del trabajo en este campo se ha centrado en el desarrollo de técnicas rápidas y eficientes para reconstruir entornos de uso común, como estructuras residenciales, oficinas o edificios públicos, con un sustancial impacto en diversos campos como la ingeniería, arquitectura, planificación urbana y energética y el sector inmobiliario, entre otros. El contenido digital requerido en diversos dominios de aplicación suele ser un modelo estructurado, que no se limita únicamente a la captura geométrica, sino que explota e integra el aspecto visual de la escena y su significado semántico. En este contexto, la adquisición con fotografías panorámicas 360 grados ha surgido como una solución muy eficaz para este tipo de ambientes, ya que proporciona una cobertura rápida y completa, incluso con una sola vista, y es compatible con una amplia gama de dispositivos de adquisición profesionales y de consumo, lo que la coinvierte en una técnica rápida y rentable. Además, las imágenes panorámicas se han convertido en un componente clave para crear contenidos inmersivos directamente a partir de escenas reales y para dar soporte a una serie de aplicaciones de Realidad Virtual (VR). En particular, las visitas virtuales basadas en imágenes esféricas son muy populares en el sector inmobiliario y han aumentado rápidamente su atractivo en el reciente periodo pandémico. Sin embargo, la mera exploración de las habitaciones a través de las fotografías esféricas originales tiene múltiples limitaciones, y en el campo de la exploración de escenas también surge la necesidad de crear modelos tridimensionales vinculados a la imagen. Los ejemplos más obvios de tales necesidades incluyen el vaciado de habitaciones antes de su presentación a los visitantes virtuales (por ejemplo, por razones de privacidad), o el soporte de exploración con 6 grados de libertad (DoF) en el contexto de la VR. En este contexto, en esta tesis se planifcó una ruta de trabajo que combina varias tareas de investigación fundamentales, como la estimación precisa de la profundidad, la disposición, la síntesis de nuevas vistas y la manipulación de las mismas. A lo largo de este camino, las diversas tareas se integraron para, respectivamente, recuperar un modelo 3D de interiores para la síntesis de nuevas vistas (Capítulo 3), la exploración inmersiva de entornos estereoscópicos de interiores (Capítulo 4) y la edición asistida automática de modelos inmersivos de interiores (Capítulo 5). Durate el trabajo, se han introducido técnicas novedosas que suponen un avance en el estado del arte de la reconstrucción 3D de entornos interiores y la interacción inmersiva.

En este anexo se ofrece una sinopsis de la tesis en lengua española, centrándose en las motivaciones, objetivos, y logros conseguidos, junto a las descripciones resumidas de los diversos métodos propuestos.

A.1 Contexto, motivación e hipótesis

La reconstrucción automatizada de modelos 3D a partir de datos adquiridos, como imágenes o mediciones geométricas, ha sido un tema central de los gráficos por ordenador y la visión por ordenador durante varias décadas. El crecimiento de este campo puede atribuirse a la convergencia de avances científicos, tecnológicos y de mercado. Estos avances se alinean con la creciente accesibilidad y abaratamiento de sensores visuales y 3D de alta calidad, que ahora están ampliamente disponibles. En este contexto, la reconstrucción automática de entornos interiores ha recibido una gran atención [1]. La entrada de datos puede proceder de diversos sensores. Los datos visuales, como las imágenes fotográficas, han despertado un gran interés debido a su amplia disponibilidad, facilidad de captura y asequibilidad. Sin embargo, una sola imagen en perspectiva ofrece una visión limitada, y la captura de múltiples imágenes introduce complejidades relacionadas con el registro. En consecuencia, la captura de 360 grados se ha convertido en los últimos años en una solución atractiva. Proporciona una cobertura rápida y completa a partir de una sola imagen y cuenta con el apoyo de una amplia gama de dispositivos de captura profesionales y de consumo, lo que garantiza una adquisición de datos eficaz y rentable. Las cámaras esféricas, también conocidas como cámaras $360°$, *panorámicas*, o *omnidireccionales*, o *surround-view*, proporcionan soluciones rentables y eficientes para capturar rápidamente en una sola toma el contexto completo alrededor del espectador de todo un entorno [10]. Al servir como representaciones del entorno del usuario, las imágenes panorámicas también prometen ser uno de los bloques de construcción esenciales para la construcción de las realidades físicas y digitales compartidas previstas por el concepto de Metaverso [13]. Aunque la captura de una panorámica de una sola toma es una forma muy atractiva de crear un clon virtual de un entorno real, la limitación del contenido presentado a lo que era visible alrededor de la ubicación fija desde la que se tomó la panorámica conduce a la pérdida de señal 3D, que es muy importante para proporcionar una sensación de presencia [6]. El hecho de que las panorámicas aparezcan planas es una limitación especialmente fuerte en entornos interiores, dada la distancia relativamente corta del espectador a las superficies arquitectónicas y los objetos. En efecto, para soportar plenamente la inmersión, un sistema debe ser capaz de generar, en tiempo real,

imágenes que respondan no sólo a los cambios de orientación, sino también a los cambios de posición del punto de vista. Aunque se han propuesto muchas soluciones para configuraciones de captura multivista, la realización de síntesis de vistas a partir de panoramas de una sola toma es de importancia primordial, debido a la conveniencia y difusión de la captura dispersa a través de cámaras monoculares de 360 [6].

Para lograr la síntesis de vistas, es necesario estimar explícita o implícitamente el modelo geométrico del entorno de la imagen. Esto, al menos, permite la reproyección consciente de la oclusión y la síntesis de contenido desocluido, así como evitar artefactos y mejorar la sensación de presencia. La recuperación de dicho modelo es, sin embargo, compleja debido a las características inherentes de las habitaciones interiores, en las que los muebles y otros elementos interiores enmascaran grandes áreas de las estructuras de interés, y las formas cóncavas de las habitaciones generan una gran cantidad de autooclusión. Por lo tanto, la reconstrucción requiere información de un contexto muy amplio y debe explotar presupuestos geométricos muy específicos para la reconstrucción de estructuras [1].

La recuperación de dicho modelo es, sin embargo, compleja debido a las características inherentes de las habitaciones interiores, donde los muebles y otros elementos interiores enmascaran grandes áreas de las estructuras de interés, y las formas cóncavas de las habitaciones generan una gran cantidad de auto-oclusión. Por tanto, la reconstrucción requiere información de un contexto muy amplio y debe explotar presupuestos geométricos muy específicos para la reconstrucción de estructuras [1].

En los últimos años, las soluciones basadas en deep-learning han surgido como una forma muy eficiente de abordar estos problemas [15]. Debido a la capacidad de estas técnicas para descubrir relaciones ocultas a partir de grandes colecciones de datos, se pueden relajar muchos presupuestos que suelen imponer los enfoques analíticos y heurísticos basados en el razonamiento geométrico. Partiendo de los aspectos mencionados, esta investigación se propuso avanzar en el estado del arte del modelado automático de interiores y la síntesis de nuevos puntos de vista a partir de una única imagen panorámica. Para superar los inevitables problemas y ambigüedades debidos a una entrada tan complicada, buscamos nuevos enfoques que explotaran los precedentes típicos de las estructuras artificiales y el potencial de las últimas técnicas basadas en datos.

Partiendo de los objetivos mencionados en section 1.2, nuestra investigación pretendía avanzar en el estado del arte tanto del modelado automático de interiores como de la exploración inmersiva a partir de una única imagen panorámica como entrada. A continuación resumimos las principales preguntas de investigación específicas para abordar tales objetivos:

1. *Cómo asociar una profundidad en píxeles a la imagen RGB de entrada?*. Generar nuevas vistas en posiciones distintas a la original, requiere necesariamente conocer la posición espacial de los puntos vistos por el observador, para poder aplicar la traslación adecuada (ver **??**). En otras palabras, cada píxel RGB de la imagen equirectangular debe asociarse a una profundidad, de forma que el píxel 2D pueda transformarse, mediante coordenadas esféricas, en un punto de color 3D, y toda la nube de puntos pueda renderizarse desde una nueva vista equirectangular. La estimación de la profundidad de una imagen panorámica de interior es un tema de investigación muy en boga (véase section 3), y por su naturaleza, un problema de investigación abierto.
2. *Cómo recuperar un diseño 3D sin oclusión de la habitación a partir de una única pose?*. La mera traslación de la nube de puntos derivada con profundidad no contiene toda la información necesaria para definir un viewport trasladado completo. De hecho, algunas partes de la escena estarán ocluidas o desocluidas. Se ha demostrado que el conocimiento de información adicional, como la posición de las esquinas y los bordes de la habitación (es decir, la distribución de la habitación), especialmente los ocultos en el punto de vista original, mejora significativamente el realismo de la síntesis [17]. Sin embargo, recuperar la distribución interior directamente a partir de una única imagen de entrada es extremadamente difícil [14]. Con este fin, es necesario investigar técnicas específicas para recuperar la distribución 3D a partir de una única imagen equirectangular.
3. *Cómo generar vistas nuevas fotorrealistas y estructuralmente coherentes?*. Para generar las nuevas vistas necesarias para la experiencia inmersiva, se emplean técnicas de generación de escenas e impainting (véase section 4). Estas técnicas de inpainting fotorrealistas funcionan muy bien en casos de pequeñas desoclusiones debidas a los movimientos naturales del observador en la escena (véase section 4.3), pero pueden fallar en el caso de grandes desoclusiones. Este es un caso típico durante la edición de la escena si, por ejemplo, el desorden tiene que ser eliminado de la escena para aplicaciones de puesta en escena virtual (véase section 5.3). En esta situación surge el problema de cómo

completar las partes que faltan, no sólo desde el punto de vista fotorrealista, sino también en términos de coherencia geométrico-estructural.

Las respuestas a estas preguntas de la investigación, que se explican y detallan en los capítulos siguientes, se basan en las siguientes hipótesis:

1. *Cómo asociar una profundidad en píxeles a la imagen RGB de entrada?* Partiendo del hecho de que la gravedad desempeña un papel importante en el diseño y la construcción de escenas de interiores artificiales [15], investigamos nuevas técnicas de estimación de la profundidad que explotan las características alineadas con la gravedad (GAF) para tener en cuenta el hecho de que las características verticales y horizontales del espacio arquitectónico tienen características diferentes en entornos artificiales (véase section 2.1). Este diseño parte del supuesto de que la captura de la escena a través de una imagen equirectangular está alineada con el vector de gravedad (es decir, la cámara está situada en un plano horizontal respecto al suelo).
2. *¿Cómo recuperar una disposición 3D de la habitación sin oclusión a partir de una sola pose?* Derivar una disposición de la escena tiene una complejidad diferente a la estimación de la profundidad simple, ya que no se limita a asignar un valor geométrico a cada píxel visible, sino que debe extrapolar grandes porciones de la estructura invisible, que puede estar ocluida no sólo por objetos, sino por la propia estructura. En nuestro trabajo, asumimos que una instalación interior puede representarse razonablemente como un modelo del de tipo *Atlanta-world*, es decir, suponemos que las habitaciones tienen suelos y techos planos. Este supuesto es menos restrictivo que el clásico *Manhattan-world*, ya que admite formas de paredes y esquinas libres (véase section 2.1).
3. *¿Cómo generar nuevas vistas fotorrealistas y estructuralmente coherentes?* Aunque parte de la información estructural, como la localización de esquinas y bordes desocluidos por la traslación, puede proporcionarse como entrada a una eventual red neuronal para la síntesis de una nueva vista, condicionar la generación de la escena final para que sea estructuralmente consistente no es sencillo (ver section 2.1). En nuestro trabajo hemos afrontado este problema asumiendo que diferentes presupuestos de la geometría de interiores pueden representarse como pérdidas específicas, del mismo modo que se impone un estilo perceptivo [50].

Los objetivos identificados se definen a continuación.

A.2 Objetivos

En nuestro proyecto de investigación estudiamos el estado de la técnica, en relación con los temas antes mencionados, e identificamos objetivos específicos:

- **Reconstrucción de un modelo 3D de interiores para la síntesis de nuevos puntos de vista.**
  Las cámaras actuales de $360°$ que ofrecen soluciones viables de bajo coste y eficiencia energética para la captura de un solo disparo de contexto completo son cada vez más populares en muchos campos de aplicación [10]. Dado que el contenido capturado $360°$, también conocido como imágenes *panorámicas*, *esféricas* o *omnidireccionales*, cubre toda la esfera alrededor del espectador, ni siquiera una sola toma puede experimentarse estáticamente de una vez, lo que la hace fundamentalmente diferente, más inmersiva y más dinámica, que las imágenes 2D tradicionales [16]. Sin embargo, la reducción de los grados de libertad a sólo la rotación alrededor del centro de la panorámica, conduce a limitaciones y artefactos [6], sobre todo porque sólo una o dos tomas por habitación están disponibles en una visita virtual típica [17]. Además, faltan totalmente el estéreo binocular y el paralaje de movimiento, que son aspectos importantes de la inmersión en la VR. Para lograr una inmersión total, el sistema debe responder también a la traslación del punto de vista. Aunque se han propuesto muchas soluciones para configuraciones de captura multivista (por ejemplo, [4, 5]), la realización de síntesis de vistas a partir de panoramas de una sola toma es de importancia primordial, debido a la conveniencia y difusión de la captura dispersa a través de cámaras monoculares de 360° [6]. La síntesis de vistas requiere la estimación explícita o implícita de la forma geométrica del entorno de la imagen (es decir, un modelo 3D de interior), con el fin de realizar una reproyección consciente de la oclusión y sintetizar el contenido ocluido. Los enfoques actuales del estado de la técnica (por ejemplo, [18, 6]) se centran en extender a panoramas de una sola toma los enfoques generales de síntesis de vistas basados en datos diseñados para vistas en perspectiva de objetos y entornos, como las imágenes multiplanares (MPI) [19] o los campos neuronales de irradiación (NeRF) [20]. La mezcla de grandes superficies sin textura, desorden y materiales no cooperativos en entornos interiores plantea, sin embargo, importantes retos a las soluciones genéricas [1]. En este contexto, se ha demostrado que el conocimiento de información adicional, como la localización de la posición de las esquinas y bordes de la habitación, mejora significativamente el realismo de la síntesis [17].

Sin embargo, recuperar la distribución interior directamente a partir de la imagen de entrada es extremadamente difícil. Incluso los últimos métodos específicos [21, 22, 23] siguen dependiendo en gran medida de aproximaciones y costosos postprocesamientos heurísticos [24], que limitan significativamente el rendimiento general. Como resultado, se inhibe su uso para aplicaciones de VR que requieren la generación de imágenes interactivas.

- **Exploración inmersiva de entornos interiores estereoscópicos.**
  Aunque la captura de una panorámica de una sola toma es una forma muy atractiva de crear un clon virtual de un entorno real, la limitación del contenido presentado a lo que era visible alrededor de la ubicación fija desde la que se tomó la panorámica conduce a la pérdida de la estereoscopía binocular, que es muy importante para proporcionar una sensación de presencia [6]. El hecho de que las panorámicas aparezcan planas es una limitación especialmente fuerte en entornos interiores, dada la distancia relativamente corta del espectador a las superficies arquitectónicas y los objetos. Para proporcionar señales estereoscópicas para rotaciones completas de 360 grados, el renderizador debe disponer de vistas desde un conjunto continuo de puntos de vista desplazados (*modelo 3D estereoscópico*). Las técnicas estéreo omnidireccionales [25, 26] se emplean con ese fin, pero requieren la creación de panorámicas estéreo utilizando cámaras que se mueven en una trayectoria circular [27, 26, 25] o múltiples cámaras 360 sincronizadas [3]. Estos enfoques de adquisición, sin embargo, reducen la posibilidad de capturar, experimentar y compartir rápidamente una escena 360° utilizando hardware de consumo. En particular, mientras que una serie de cámaras de bajo coste están ampliamente disponibles para la captura monocular 360° (por ejemplo, GoPro, Ricoh Theta, LadyBug, o Insta360), también debido al auge del mercado de "cámaras de acción", la solución estéreo 360° (por ejemplo, Vuze+) son más costosas y limitadas, y también suelen ofrecer sólo un número bajo (es decir, de seis a ocho) de puntos de vista diferentes, lo que provoca artefactos estereoscópicos y de cosido. Además, aunque las soluciones de cámara giratoria ofrecen más puntos de vista, no comparten la misma sencillez y flexibilidad de la captura instantánea de una sola toma. Por este motivo, la investigación se ha centrado en métodos de síntesis de vistas que generan contenidos estereoscópicos a partir de una única panorámica de 360°. Sin embargo, los métodos actuales requieren representaciones complicadas o son demasiado pesados para ejecutarse directamente en HMD y tipos interactivos.

- **Edición semi-automática de modelos inmersivos de interiores.**
  Una exploración pura de los entornos existentes a través de las fotos esféricas originales es, sin embargo, muy limitante. Ejemplos destacados de necesidades adicionales incluyen el vaciado de habitaciones antes de su presentación a los visitantes virtuales (aunque solo sea por razones de privacidad), o el reamueblamiento o redecoración de espacios interiores [28]. En este contexto, las técnicas de Realidad Disminuida (DR) rápidas y eficaces, que ocultan partes reales del campo de visión, son primordiales para eliminar los muebles y otros elementos desordenados que enmascaran la estructura arquitectónica. En particular, las características de DR son esenciales para permitir a los usuarios comparar inmediatamente la escena amueblada y la no amueblada, y para apoyar las aplicaciones de Realidad Aumentada (RA) en la colocación de objetos en la escena vacía [29, 30]. Hacer que estas características estén disponibles en entornos novedosos con una latencia mínima, idealmente en tiempo real, permitiría, además, su uso en contextos de colaboración remota, sin necesidad de modelado previo [31]. Aunque en la bibliografía se han presentado diversas soluciones de borrado de objetos e inpainting de imágenes (véase section 5.3), la DR para entornos interiores debe generar imágenes de espacios interiores vacíos que no solo tengan un aspecto realista, sino que respeten el contexto de formas más estrictas, en particular infiriendo una organización plausible de la estructura arquitectónica que delimita el interior de la habitación [32]. Las soluciones basadas en datos, que aprenden relaciones ocultas a partir de ejemplos, están surgiendo como enfoques viables para esta clase de problemas. Sin embargo, los métodos más avanzados para el repintado de imágenes se centran principalmente en el fotorrealismo [33, 34], y la información adicional sobre la escena se explota sólo desde el punto de vista semántico [35, 36, 32]. Los *pipelines* actuales hacen un uso limitado de la estructura de la escena observada, y la precisión de la reconstrucción se consigue a costa de una alta complejidad computacional o de una mayor intervención del usuario, utilizando, por ejemplo, redes recursivas [37], arquitecturas multirama [34], y la definición manual de partes específicas de la imagen original que deben eliminarse [35].

En los párrafos siguientes resumimos nuestras contribuciones.

A.3 Reconstrucción de un modelo 3D de interiores para la síntesis de nuevos puntos de vista

La síntesis de vistas requiere la estimación explícita o implícita de la forma geométrica del entorno de la imagen, con el fin de realizar una reproyección consciente de la oclusión y sintetizar el contenido ocluido. Los enfoques actuales (por ejemplo, [18, 6]) se centran en extender a panoramas de una sola toma los enfoques generales de síntesis de vistas basados en datos diseñados para vistas en perspectiva de objetos y entornos, como las imágenes multiplanares (MPI) [19] o los campos neuronales de radiación (NeRF) [20]. (section 3.3). La mezcla de grandes superficies sin textura, desorden y materiales no cooperativos en entornos interiores plantea, sin embargo, importantes retos a las soluciones genéricas [1]. En este contexto, se ha demostrado que el conocimiento de información adicional, como la localización de la posición de las esquinas y bordes de la habitación, mejora significativamente el realismo de la síntesis [17]. Sin embargo, recuperar la distribución interior directamente a partir de la imagen de entrada es extremadamente difícil. Incluso los últimos métodos específicos [21, 22, 23] siguen dependiendo en gran medida de aproximaciones y costosos post-procesamientos heurísticos [24], que limitan significativamente el rendimiento global. Como resultado, se inhibe su uso para aplicaciones de VR que requieren la generación de imágenes interactivas.

En nuestro trabajo, proponemos una nueva solución integral basada en datos que, a partir de una única panorámica de 360° en interiores, supuestamente capturada con una alineación gravitatoria aproximada, produce con baja latencia una nueva pose traducida de la que se pueden extraer nuevas imágenes en perspectiva que respondan tanto a los cambios de posición como de orientación. Mientras que algunas soluciones de HMD se esfuerzan por funcionar totalmente en la plataforma integrada, un diseño alternativo consiste en computar las imágenes en servidores de alto rendimiento. Este enfoque, ampliamente utilizado para los juegos de alta calidad, es posible gracias a la disponibilidad de conexiones inalámbricas de baja latencia con suficiente ancho de banda para alimentar las pantallas. En nuestro enfoque, un cliente WebXR gestiona directamente la rotación de la cabeza, mientras que las imágenes calculadas por el servidor también responden a las traslaciones de la cabeza.

Nuestra principal novedad radica en las técnicas de aprendizaje profundo específicas para interiores que sintetizan las vistas. Una vez por escena, enriquecemos el panorama original con información geométrica y estructural, y una vez por fotograma,

explotamos la información precalculada para realizar rápidamente la síntesis de vistas.

El enfoque hace frente a las ambigüedades inherentes a la estimación de la geometría de una sola imagen y a la síntesis de vistas novedosas en entornos interiores, centrándose en el caso muy común de los interiores que siguen el modelo *Atlanta world* (AWM) [1], en el que se espera que el entorno tenga suelo y techo horizontales y paredes verticales. Basándonos en este modelo previo, introducimos una novedosa red de extremo a extremo para estimar conjuntamente la profundidad y la estructura subyacente de la escena, manejando así eficientemente oclusiones y desoclusiones y permitiendo una predicción plausible incluso en el caso de estructuras extensamente ocluidas. La información previa determina la estructura de la red, que también aprovecha las características alineadas con la gravedad (GAF) para tener en cuenta el hecho de que las características verticales y horizontales del espacio del mundo tienen características diferentes en los entornos artificiales.

En particular, AWM permite derivar el layout 3D extruyendo su proyección 2D de suelo, mientras que los GAFs realizan compresión vertical, explotando el hecho de que las líneas verticales, comunes en escenas de interior, no se deforman en proyecciones equirectangulares. Debido a estas características, esperamos que los GAFs de las escenas estén interrelacionados por dependencias espaciales tanto a corto como a largo plazo, mejorando la calidad de la predicción de la profundidad y la predicción de la disposición [15], y, por lo tanto, la síntesis visual.

A partir de la representación panorámica enriquecida, una red ligera infiere a velocidades interactivas nuevas imágenes panorámicas desde posiciones trasladadas en un área de trabajo alrededor del punto de vista original, adecuadas para aplicaciones de VR. A partir de estas vistas, se producen imágenes en perspectiva que coinciden con las rotaciones de la cabeza y se amplían a tamaño de pantalla. Además, la información geométrica y estructural recuperada también puede utilizarse para apoyar aplicaciones de VR, por ejemplo, para definir zonas transitables.

Nuestras principales aportaciones novedosas son las siguientes:

- Presentamos un nuevo enfoque, denominado "Módulo de Profundidad de Atlanta" (ADM), para estimar conjuntamente, a partir de una única imagen equirectangular, la profundidad de la escena, una representación latente de la escena, la forma 3D de la habitación y un mapa de ocupación del suelo.

ADM consigue los mejores resultados tanto en reconstrucción geométrica como estructural (section 3.5), y proporciona muchas ventajas para las aplicaciones de VR. En primer lugar, es mucho más ligera que las soluciones actuales de estimación de profundidad o disposición comúnmente adoptadas en este contexto [55, 24, 15]. En segundo lugar, la estructura AWM recuperada se segmenta en techo, paredes y suelo, y se representa en unidades métricas, incluyendo la predicción de alturas techo-suelo. Esto, además de mejorar la síntesis de vistas, soporta la creación de un *mapa de ocupación del suelo*, para generar trayectorias consistentes dentro de la habitación sin colisiones.

- Introducimos nuevas funciones objetivo para tener en cuenta la consistencia estructural en interiores (section 3.4.3) de la síntesis de vistas. Dichas funciones, basadas en la codificación GAF [15, 56], admiten pérdidas directas (es decir, pérdida de profundidad predicha por el objetivo) y pérdidas de espacio latente. Las pérdidas del espacio latente guían una reconstrucción estructural coherente durante el entrenamiento y son duales a las pérdidas visuales, denominadas *perceptual geométrica* y *estilo geométrico*, Dichas pérdidas, combinadas con la transferencia de estilo perceptual estándar y las pérdidas adversariales, mejoran la calidad de la escena reconstruida (**??**), desplazando gran parte de la carga computacional a la fase de entrenamiento, y haciendo que la fase de inferencia sea mucho más ligera.

- Presentamos un enfoque totalmente basado en datos, versátil y ligero para generar nuevas vistas panorámicas a partir de una única panorámica de interiores. Este enfoque de aprendizaje profundo no necesita un procesamiento dedicado para cada escena [18, 6], sino que se generaliza sobre escenas de interior que simplemente siguen AWM (**??**). Una vez que las características profundas latentes y los presupuestos estructurales se aplican en tiempo de entrenamiento, se obtiene una síntesis de pose novedosa a través de una red (GVS) sin capas profundas ni pipelines complejos. De hecho, GVS consiste en un número limitado de capas, combinando convoluciones gated y dilatadas, centradas en maximizar el nivel de detalle (section 3.4.2). Como resultado, tenemos una red con un número limitado y constante de parámetros aprendibles (section 3.5.2), incluso cuando varía la resolución de la imagen generada.

Nuestros resultados (section 3.5) mejoran en precisión, calidad y complejidad computacional a los enfoques más avanzados en pruebas de referencia comunes con una verdad sobre el terreno mensurable. Además, se obtienen predicciones convincentes

incluso en imágenes en las que no se dispone de datos reales para el entrenamiento, así como en nuevas imágenes capturadas por el usuario.

A.4 Exploración inmersiva de entornos interiores estereoscópicos

Aunque capturar una panorámica de una sola toma es una forma muy atractiva de crear un clon virtual de un entorno real, la limitación del contenido presentado a lo que era visible alrededor de la ubicación fija desde la que se tomó la panorámica lleva a la pérdida de la estereoscopía binocular, que es muy importante para proporcionar una sensación de presencia [6]. El hecho de que las panorámicas aparezcan planas es una limitación especialmente fuerte en entornos interiores, dada la distancia relativamente corta del espectador a las superficies arquitectónicas y los objetos. Para proporcionar señales estéreo para rotaciones completas de 360 grados, el renderizador debe disponer de vistas desde un conjunto continuo de puntos de vista desplazados. Las técnicas estéreo omnidireccionales [25, 26] se emplean con ese fin, pero requieren la creación de panorámicas estéreo utilizando cámaras que se mueven en una trayectoria circular [27, 26, 25] o múltiples cámaras 360 sincronizadas [3]. Estos enfoques de adquisición, sin embargo, reducen la posibilidad de capturar, experimentar y compartir rápidamente una escena 360° utilizando hardware de consumo. En particular, mientras que varias cámaras de bajo coste están ampliamente disponibles para la captura monocular 360° (por ejemplo, GoPro, Ricoh Theta, LadyBug o Insta360), también debido al auge del mercado de las "cámaras de acción", las soluciones estéreo 360° (por ejemplo, Vuze+) son más costosas y limitadas, y también suelen ofrecer solo un número bajo (es decir, de seis a ocho) de diferentes puntos de vista, lo que provoca artefactos estereoscópicos y de stitching. Además, aunque las soluciones de cámara giratoria proporcionan más puntos de vista, no comparten la misma simplicidad y flexibilidad de la captura instantánea de una sola toma. Por este motivo, la investigación se ha centrado en métodos de síntesis de vistas que generan contenidos estereoscópicos a partir de una única panorámica de 360°. Sin embargo, los métodos actuales requieren representaciones complicadas o son demasiado pesados para ejecutarse directamente en HMD y tasas interactivas (section 4.3). Para superar estas limitaciones, proponemos en este trabajo un enfoque novedoso para generar y experimentar de forma rápida y automática una representación estereoscópica omnidireccional de un entorno interior a partir de una única imagen panorámica monoscópica en formato equirectangular. En nuestro

enfoque, resumido en **??** y section 4.4, comenzamos por estimar la profundidad por píxel de la imagen completa utilizando una red profunda diseñada para explotar las restricciones del entorno interior y entrenada en grandes conjuntos de ejemplos sintéticos (section 4.5). A continuación, sintetizamos cortes panorámicos mediante reproyección y síntesis de vistas utilizando una red profunda que comparte las mismas características de diseño y conjunto de entrenamiento que la de estimación de profundidad (section 4.6). Estos cortes se colocan alrededor del punto de vista central, en el círculo formado por los dos ojos durante las rotaciones de la cabeza, y cubren una porción angular suficiente para acomodar las direcciones potenciales de la mirada tanto del ojo izquierdo como del derecho. A continuación, se compone un par estereoscópico de centro de proyección múltiple en formato equirectangular mezclando adecuadamente los cortes precalculados. El par resultante se carga en un visor WebXR para obtener una experiencia ligera y receptiva con señales de movimiento y estereoscópicas durante el tiempo de ejecución (section 4.7). En este enfoque, basado en la aproximación a una experiencia estereoscópica completa mediante un par estereoscópico omnidireccional (véase section 4.3), se minimizan los costes de ejecución, tanto en términos de almacenamiento y ancho de banda como de rendimiento de renderizado, a costa de una ligera degradación de la reconstrucción estereoscópica en la visión periférica (véase section 4.7).

Nuestras principales aportaciones son las siguientes:

- Introducimos una novedosa arquitectura de red profunda de extremo a extremo que genera vistas desplazadas de una imagen panorámica de interior en formato equirectangular; un primer módulo de red estima un mapa de profundidad a partir de una única entrada panorámica; a continuación, estas vistas se reproyectan a la posición deseada, y se sintetiza una imagen completa a través de una segunda red capaz de generar contenido plausible en áreas desocluidas. A diferencia de otros enfoques de última generación en la literatura [65, 15], la red se basa en una arquitectura de tipo *gated* ligera y de cuello de botella dilatado; como resultado, garantizamos la escalabilidad a tamaños de imagen más grandes y/o hardware embebido, manteniendo el máximo detalle visual al reproyectar en nuevas vistas;

- Introducimos una arquitectura de red unificada con estrategias de entrenamiento personalizadas tanto para la estimación de la profundidad como para la síntesis de vistas. Se explota la misma red ligera para ambas tareas, sólo adaptando la función de activación final y cambiando el modo de entre-

namiento. Para ello, introducimos una pérdida fotométrica específica para la novedosa síntesis de vistas, combinada con un enfoque GAN. Como resultado, se generan nuevas vistas fotorrealistas con un bajo coste computacional. Además, utilizamos arquitecturas basadas en GAN de superresolución para aumentar aún más la resolución entre las imágenes estéreo.

- explotamos nuestro enfoque de estimación de profundidad, reproyección y síntesis para generar un conjunto de cortes panorámicos y utilizarlos para calcular un par de imágenes estéreo omnidireccionales que pueden experimentarse directamente en visores WebXR que los muestrean para generar parejas estéreo que responden al movimiento de la cabeza con baja latencia y alta frecuencia. La limitación a cortes panorámicos simplifica enormemente los costes computacionales off-line en comparación con soluciones anteriores [105], y la explotación directa de formatos estéreo omnidireccionales estándar fomenta la aplicabilidad del método a una variedad de plataformas de hardware y software.

Nuestra evaluación (section 4.8) ilustra cómo las redes de inferencia de profundidad e inpainting alcanzan un rendimiento de vanguardia y cómo pueden explotarse para producir imágenes estereoscópicas omnidireccionales sin fisuras a una alta tasa de muestreo angular. Dado que el marco propuesto es fácil de integrar en los visores panorámicos actuales, sustituyendo simplemente a los actuales renderizadores monoscópicos, promete ser un bloque de construcción práctico para ofrecer experiencias atractivas y realistas que cautiven al público y le permitan explorar e interactuar virtualmente con espacios interiores en aplicaciones Metaverse actuales y futuras.

A.5 Edición semi-automática de modelos inmersivos de interiores

Aunque en la bibliografía se han presentado diversas soluciones de borrado de objetos y repintado de imágenes (section 5.3), la DR para entornos interiores debe generar imágenes de espacios interiores vacíos que no sólo tengan un aspecto realista, sino que respeten el contexto de formas más estrictas, en particular deduciendo una organización plausible de la estructura arquitectónica permanente que delimita el interior de la habitación [32]. Las soluciones basadas en datos, que aprenden relaciones ocultas a partir de ejemplos, están surgiendo como enfoques viables

para esta clase de problemas. Sin embargo, los métodos más avanzados para el repintado de imágenes se centran principalmente en el fotorrealismo [33, 34], y la información adicional sobre la escena se explota sólo desde el punto de vista semántico [35, 36, 32]. Los *pipelines* actuales hacen un uso limitado de la estructura de la escena observada, y la precisión de la reconstrucción se consigue a costa de una alta complejidad computacional o de una mayor intervención del usuario, utilizando, por ejemplo, redes recursivas [37], arquitecturas multirama [34], y la definición manual de partes específicas de la imagen original que deben eliminarse [35].

En este trabajo, presentamos una nueva red profunda ligera de extremo a extremo que, a partir de una imagen de entrada de 360° de un espacio interior amueblado, devuelve automáticamente, con muy baja latencia, una vista fotorrealista omnidireccional y una profundidad arquitectónicamente plausible de la misma escena vaciada de todo desorden.

Aprovechando la disponibilidad de conjuntos de datos sintéticos fotorrealistas a gran escala, entrenamos nuestra red en pares utilizando un conjunto de ejemplos compuestos por imágenes equirectangulares registradas del color del entorno desordenado, el color del entorno vacío y su profundidad. La red final de extremo a extremo se descompone en dos bloques, que se entrenan por separado para reducir los costes de entrenamiento. El primer bloque aprende una máscara de atención de las partes despejadas de la imagen de entrada, generando ejemplos de entrenamiento a partir de la imagen de entrada despejada y los pares de profundidad. El segundo bloque toma como entrada la máscara de atención y la imagen desordenada, y realiza la síntesis de la escena despejada, utilizando para el entrenamiento pérdidas específicas de interiores que incorporan nuestro conocimiento de los entornos interiores esperados. A diferencia de otros métodos de eliminación de objetos, el nuestro se aplica de forma holística a toda la escena, eliminando todo el desorden en un solo paso sin intervención del usuario. El vaciado rápido de la habitación sin intervención manual es el pilar fundamental sobre el que se asientan las demás características necesarias para una aplicación de DR. Por ejemplo, la eliminación de un único objeto (o la conservación de un único objeto) se consigue componiendo la imagen de la habitación vacía de nuestra red con la imagen original, teniendo en cuenta la máscara de objeto calculada (véase, por ejemplo, el diseño de Gkiktas et al. [32]). Además, al inferir la geometría de la habitación y eliminar el desorden, podemos realizar varias ediciones de la escena, como añadir o colocar muebles apoyados en el suelo o pegados a una pared (véase Figure 5.1).

Nuestras principales aportaciones se resumen a continuación:

- Proponemos una técnica ligera de aprendizaje profundo de extremo a extremo (section 5.4), que proporciona, a tasa interactiva, una escena panorámica de interiores vaciada automáticamente sin intervención del usuario y adecuada para su uso en aplicaciones XR. Nuestra red de predicción se desarrolla de forma lineal, sin necesidad de fusionar características de ramas paralelas [34, 32], ni de refinar el resultado recursivamente [37]. Con el fin de aliviar la carga del *gating convolucional* para el *inpainting* genérico asistido por el usuario [33], adoptamos en su lugar una estrategia de convolución tipo *gated* separable en profundidad, reduciendo el número de parámetros y el tiempo de procesamiento mientras se mantiene la efectividad [89]. Además, tanto las restricciones visuales como las geométricas se aplican únicamente en tiempo de entrenamiento, donde las visuales siguen una estrategia de aprendizaje por transferencia [123] y las geométricas adoptan pérdidas robustas y eficientes que codifican nuestro conocimiento previo sobre entornos interiores (section 5.4.3).

- Predecimos una representación geométrica emparejada con la imagen de salida, es decir, una estimación de profundidad densa de la escena vacía. Esta representación geométrica puede utilizarse directamente como base para su posterior procesamiento en la aplicación XR (por ejemplo, para ayudar al posicionamiento de objetos o para calcular oclusiones). Se obtiene obtenida conjuntamente con la representación visual y sin necesidad de onerosas ramas paralelas [34, 32]. También la explotamos para definir una previa robusta y efectiva a nivel de píxel junto con otras previas y pérdidas 3D (section 5.4.3) La generación de una pista geométrica como salida reduce la necesidad de añadir análisis semántico adicional sobre la imagen o de utilizar estrategias GAN [92, 124] para desambiguar los resultados obtenidos, como demuestran nuestros resultados (section 5.5).

  Por el contrario, los métodos actuales de inpainting se centran principalmente en la salida visual y perceptual [33, 34], donde la preservación de la estructura se maneja a nivel de características de la imagen o semánticamente [35, 36]. Otros enfoques se basan, en cambio, en anotaciones manuales y simplificadas del trazado subyacente, que no representa necesariamente la verdadera geometría 3D. Esta información se interpreta mejor como un previo semántico 2D que geométrico [36, 32].

- Dirigimos nuestro entrenamiento utilizando una función de pérdida que combina términos fotorrealistas y geométricos. En particular, nuestros términos geométricos explotan tanto la información de píxeles de los mapas de profundidad como el concepto de normales virtuales generadas por triples de puntos a gran distancia [126], para recuperar eficientemente las características salientes de las estructuras interiores hechas por el hombre, en términos de planitud y suavidad, sin caer en estructuras restrictivas como *Manhattan-world*, *Atlanta-world* o incluso paredes verticales [1].

Nuestros resultados muestran que nuestro método supera a los enfoques más avanzados, utilizando puntos de referencia comunes con una verdad básica medible, en términos de precisión, calidad y menor complejidad computacional (section 5.5.3). Además, nuestro modelo también es capaz de producir predicciones convincentes incluso en imágenes de conjuntos de datos comunes en los que no se dispone de la verdad básica para el entrenamiento, así como en imágenes nuevas capturadas por un usuario (section 5.5.4).

A.6 Logros y conclusiones

Partiendo de los objetivos fijados, investigamos técnicas novedosas, la estructura de las redes, las funciones de pérdida y los métodos de entrenamiento. Nuestra investigación produjo una serie de resultados científicos (enumerados en section 6.3), que hicieron avanzar el estado de la técnica en muchos aspectos. Seleccionamos entre ellos los principales logros (véase section 1.2):

- **Una nueva metodología para la síntesis de escenas 3D de interiores *Atlanta-world* a partir de una única imagen omnidireccional**. (section 3). Un nuevo enfoque basado en datos para extraer información geométrica y estructural de un único panorama esférico de una escena interior, y para utilizar esta información para representar la escena desde nuevos puntos de vista, mejorando la inmersión 3D en aplicaciones de realidad virtual. El enfoque hace frente a las ambigüedades inherentes a la estimación de la geometría de una sola imagen y a la síntesis de puntos de vista novedosos centrándose en el caso muy común de los interiores *Atlanta-world*, delimitados por suelos y techos horizontales y paredes verticales. Basándonos en este modelo, introducimos un nuevo enfoque de aprendizaje profundo de extremo a extremo para estimar conjuntamente la profundidad y la estructura subyacente de la escena. El aprendizaje previo guía

el diseño de la red y de nuevas funciones de pérdida específicas del dominio, desplazando la mayor carga computacional a una fase de entrenamiento que explota las imágenes panorámicas sintéticas a gran escala disponibles. Una red extremadamente ligera utiliza información geométrica y estructural para inferir nuevas vistas panorámicas a partir de posiciones trasladadas a velocidades interactivas, a partir de las cuales se producen vistas en perspectiva que coinciden con las rotaciones de la cabeza y se amplían al tamaño de la pantalla.Como resultado, nuestro método produce automáticamente nuevas poses alrededor de la cámara original a velocidades interactivas, dentro de un área de trabajo adecuada para producir señales de profundidad para aplicaciones de VR, especialmente cuando se utilizan pantallas montadas en la cabeza conectadas a servidores gráficos. El plano extraído y la estructura 3D de las paredes también pueden utilizarse para explorar la sala. Los resultados experimentales demuestran que nuestro método proporciona un rendimiento de baja latencia y mejora la precisión de la predicción de las soluciones más avanzadas en los puntos de referencia panorámicos de interiores más utilizados. Este trabajo ha sido publicado en la revista TVCG [38] y presentado en la conferencia IEEE ISMAR 2023. El candidato fue responsable de todos los aspectos del trabajo, desde la conceptualización, metodología, software, hasta la validación, redacción y revisión de los artículos relacionados.

- **Un novedoso enfoque para la síntesis profunda y la exploración de entornos estereoscópicos omnidireccionales a partir de una imagen panorámica monoscópica**. (section 4). Un enfoque innovador para generar y explorar automáticamente entornos interiores estereoscópicos inmersivos derivados de una única imagen panorámica monoscópica en formato equirectangular. Una vez por cada toma de 360°, estimamos la profundidad por píxel utilizando una arquitectura de red profunda cerrada. Posteriormente, sintetizamos una colección de cortes panorámicos mediante reproyección y síntesis de vistas empleando aprendizaje profundo. Estos cortes se distribuyen alrededor del punto de vista central, con el centro de proyección de cada corte situado en la trayectoria circular recorrida por los ojos durante una rotación de la cabeza. Además, cada corte abarca una extensión angular suficiente para acomodar las posibles direcciones de la mirada de los ojos izquierdo y derecho y proporcionar contexto para la reconstrucción. Para una visualización rápida, se compone un par estereoscópico de centro de proyección múltiple en formato equirectangular mezclando adecuadamente los cortes precalculados. En tiempo de ejecución, el par se carga en un visor WebXR ligero que responde a las rotaciones de la cabeza, ofreciendo pistas tanto de movimiento como estereoscópicas. El

planteamiento combina y amplía las técnicas más avanzadas basadas en datos, incorporando varias innovaciones. En particular, se introduce una arquitectura con compuerta para la estimación de la profundidad panorámica monocular. Aprovechando la profundidad predicha, la misma arquitectura se aplica a la reproyección de píxeles visibles, facilitando el repintado de regiones ocluidas y desocluidas mediante la incorporación de una red generativa adversarial mixta (GAN). El sistema resultante funciona en una gran variedad de cascos de VR disponibles y puede servir como componente base para una gran variedad de aplicaciones inmersivas. Demostramos nuestra tecnología en varias escenas de interiores a partir de datos disponibles públicamente. Este trabajo ha sido aceptado para su publicación como artículo en la revista Computers & Graphics [39] y premiado con mención honorífica en la conferencia ACM WEB3D 2023. El candidato El candidato fue responsable de todos los aspectos del trabajo, desde la conceptualización, metodología, software, hasta la validación, redacción y revisión de los artículos relacionados.

- **Una innovadora técnica de extremo a extremo para el vaciado automático instantáneo de escenas panorámicas de interior** (section 5). Un nuevo enfoque basado en datos que, a partir de una imagen de entrada de 360° de un espacio interior amueblado, devuelve automáticamente, con muy baja latencia, una vista fotorrealista omnidireccional y una profundidad arquitectónicamente plausible de la misma escena vaciada de todo desorden. A diferencia de los anteriores métodos de inpainting basados en datos que eliminan objetos individuales definidos por el usuario en función de su semántica, nuestro enfoque se aplica de forma holística a toda la escena y es capaz de separar el desorden de la estructura arquitectónica en un solo paso. Al explotar las características geométricas peculiares del entorno de interiores, desplazamos la mayor carga computacional a la fase de entrenamiento y disponemos de una red extremadamente ligera en el momento de la predicción. Este trabajo ha sido publicado como artículo de la revista TVCG [40] y presentado en la conferencia IEEE ISMAR 2022. El candidato fue responsable de todos los aspectos del trabajo, desde la conceptualización, metodología, software, hasta la validación, redacción y revisión de los artículos relacionados.

Además, durante el curso de mis estudios de doctorado y siempre en relación con esta tesis, he contribuido también en una serie de publicaciones que no han sido incluidas directamente en este trabajo, y que se listan en la sección dedicada a los resultado bibliográficos (Sec. 6.3).

Conclusiones y futuro. Se presenta en esta sección una breve evaluación de las ventajas y limitaciones de las soluciones, así como una visión de las áreas de trabajo futuras. Como se ha visto en los capítulos anteriores, la construcción de un modelo inmersivo de interiores a partir de una sola imagen panorámica requiere la capacidad de extraer de ella un modelo 3D estructurado. Este modelo debe incluir información de profundidad desde la perspectiva de la cámara, un trazado 3D que permita inferir partes potencialmente ocluidas y la capacidad de distinguir semánticamente entre elementos estructurales permanentes (es decir, sujetos a reglas artificiales) y desorden. Este modelo es un requisito esencial para garantizar la fiabilidad de la síntesis de vistas.

Todas las técnicas que aquí se presentan aprovechan las distintas características de la configuración de captura, en particular la alineación gravitatoria, y del entorno esférico, en concreto la alineación del espacio-mundo con la gravedad. Esta alineación permite explotar las regularidades de las características verticales a lo largo de la dirección horizontal. Estas características consistentes han influido en los diseños de las redes, dando lugar al uso de contracciones asimétricas y diversas estrategias para combinar características de largo y corto alcance [38]. Las redes específicas diseñadas a partir de estos principios ofrecen ventajas significativas frente a alternativas más genéricas. Esto subraya el valor de las soluciones a medida para la captura de interiores, en lugar de confiar en redes genéricas diseñadas para exteriores o para la reconstrucción 3D de formas genéricas. Sin embargo, la creación de redes personalizadas también tiene un inconveniente: dependen de características ambientales específicas. En consecuencia, cuando el entorno de la imagen se desvía de las expectativas, pueden producirse fallos importantes (por ejemplo, véase section 3.8.4). Aunque estas soluciones son más robustas que los métodos de razonamiento geométrico, siguen presentando limitaciones en cuanto a su aplicabilidad, como demuestran los análisis de casos de fallo presentados en los capítulos anteriores.

Otra limitación significativa, también compartida por los demás enfoques actuales, se refiere al tamaño de entrada. Aunque las soluciones que presentamos son en general ligeras y el diseño de la red es escalable, las pruebas se han realizado predominantemente con tamaños de imagen más pequeños de lo que se puede conseguir actualmente con las cámaras panorámicas. Las pruebas comparativas existentes suelen funcionar con una resolución de 1024×512, y rara vez se aventuran en dimensiones mayores. En cambio, las cámaras industriales capturan imágenes más detalladas. Una vía esencial para la investigación futura consiste en evaluar la escala-

bilidad de estas técnicas para manejar conjuntos de datos más grandes y casos del mundo real. Para ello será necesario no sólo escalar las redes, sino también generar amplios conjuntos de datos anotados que sirvan de base. Mientras que la estimación de la profundidad requiere una verdad de base precisa por píxel, la reconstrucción del trazado es una tarea más abstracta que depende de las anotaciones realizadas por el usuario. Por razones prácticas, estas anotaciones se limitan actualmente a los entornos de *Manhattan-world* o *Atlanta-world*. En el futuro esperamos resolver este problema gracias a la creciente disponibilidad de dispositivos que capturen tanto RGB de alta resolución como profundidad, y a la eventual posibilidad de integrar sistemas de anotación semiautomáticos para estructuras de interés, por ejemplo ampliando sistemas de *uncluttering* como el que presentamos (véase chapter 5).

Los enfoques modernos de síntesis de vistas, incluidas nuestras soluciones, se basan básicamente en técnicas de *inpaiting*. Mientras que en gran parte de la literatura tales soluciones se refuerzan con un entrenamiento de tipo GAN, una de nuestras mayores contribuciones en este tema es utilizar también arquitecturas y estrategias de red basadas en información geométrica, en particular presupuestos de los interiores. Tal enfoque, como se ha demostrado en capítulos anteriores, mejora enormemente el realismo de la escena sintetizada, tanto en casos de oclusión visual (es decir, una esquina oculta que se desocluye, chapter 3), como en el caso de reconstrucción de partes estructurales previamente cubiertas por *clutter* (véase chapter 5). Esta solución es especialmente eficaz en el caso de una escena editada, como en el caso de *uncluttering* para disminuir la realidad chapter 5 o para una posible puesta en escena virtual con adición de objetos 3D, es decir, en el caso de que el resultado final sea estructural o se complete con objetos sintéticos. En cambio, en el caso de la navegación de forma libre pura chapter 3, persisten varios artefactos y limitaciones. En este caso, la información estructural de interiores no puede ayudar a completar ningún objeto genérico, lo que en algunos casos puede generar artefactos evidentes al cambiar el punto de vista. En la actualidad, esta situación limita el rango de movimiento del observador, ya que en presencia de grandes movimientos, la información de disposición y las previsiones interiores no serían útiles para reconstruir grandes porciones de objetos desocluidos. En el futuro, esto podría solucionarse combinando varias imágenes y aumentando las capacidades generativas de la red de síntesis con técnicas de difusión.

En nuestro trabajo de investigación también presentamos un marco para la generación automática de entornos interiores estereoscópicos omnidireccionales para su uso en aplicaciones inmersivas (véase chapter 5), especialmente consumidas a través

de pantallas montadas en la cabeza. Nuestro método parte de una única imagen panorámica de un entorno interior y utiliza arquitecturas basadas en datos para la estimación de la profundidad y una novedosa síntesis de vistas para generar rápidamente las imágenes vistas por ambos ojos durante la rotación de la cabeza.

Una de las limitaciones del método actual es el desajuste entre la resolución de las imágenes sintetizadas y la resolución alcanzable con las cámaras y pantallas actuales. Este desajuste se resuelve reduciendo la resolución de las imágenes antes de su construcción y aumentándola con la ayuda del aprendizaje profundo antes de la presentación en pantalla. La limitación no se debe a la arquitectura de red ligera propuesta, que promete ser escalable a tamaños de imagen mucho mayores, sino a la disponibilidad de conjuntos de entrenamiento para las redes de estimación de profundidad y síntesis de vista. Tenemos previsto abordar este problema generando datos de entrenamiento de mayor resolución, como en el caso del problema de síntesis de vistas de forma libre que se ha comentado anteriormente.

Otro punto se refiere a las direcciones futuras de los métodos de edición automática inmersiva. Aquí se presenta un nuevo enfoque basado en datos que, a partir de una imagen de entrada de 360° de un espacio interior amueblado y desordenado, devuelve automáticamente, a velocidad interactiva, una vista fotorrealista de 360° y profundidad de la misma escena vaciada de todos los muebles y demás desorden. Aunque la aplicación de la realidad atenuada es coherente por sí misma, en el sentido de que permite visualizar la estructura del entorno libre de objetos o desorden genérico para diversos fines, siguen siendo necesarios, sin embargo, algunos pasos para disponer de una aplicación de puesta en escena virtual completa.En primer lugar, sería necesario estimar la iluminación global, de nuevo a partir de una sola imagen, para tener la capacidad no sólo de eliminar objetos, sino de insertar otros nuevos de forma fotorrealista. A este respecto, existen varios enfoques de cara al futuro para conseguir un modelo totalmente editable, como el que se propone en algunos trabajos recientes [145].

Curriculum Vitae

B

Giovanni Pintore is a senior engineer and researcher in the Visual Computing (VIC) group at CRS4. He has published more than 60 papers in major international journals and conferences in computer graphics and computer vision. He is present in several technical committees of international conferences and journals, in the board of Eurographics Italy and currently reviewer of IEEE CVPR, ECCV, 3DV, IEEE TVCG, Springer The Visual Computer, Elsevier Computer and Graphics, Computer Graphics Forum and others. He is present in the international scientific community as author and as keynote speaker or program chair. He has been involved as project manager in international and national research and industrial projects and collaborations, particularly in the areas of security, space exploration and smart cities. He is current research interests include data-driven and big-data approaches for automatic 3D reconstruction of architectural structures; emerging technologies for cost-effective acquisition and immersive exploration of the built environment; deep learning and mobile approaches for real-world applications, focused on security and twin digital and green transition; data-driven solutions for image analysis (classification, shape analysis).

Contact Information

Name	Giovanni Pintore
Address	CRS4, Ex-Distilleria Pirri. Via Ampere 2, 09134 - Cagliari, Italy
E-Mail	giannipint@gmail.com
Online	`https://scholar.google.com/citations?user=UDIPaq0AAAAJ&hl=en` `https://orcid.org/0000-0001-8944-1045`

Personal Details

Date of Birth	February 24th, 1975
Place of Birth	Nuoro, Italy
Languages	Italian (native), Sardinian (native), English (fluent), French (basic)

Education

Since 2022	**Ph.D. candidate on Computer Science**. University of A Coruña (UDC), Spain. Doctoral program in Information and Communications Technologies.
January 2003	**Habilitation to the Professional Engineers Association for the Industrial, Information, Civil-Environmental field**. Registration number 5937, Italy
June 2002	**Laurea (M. Sc.) degree (June 2002) in Electronics Engineering**. University of Cagliari, Italy.

Employment History

Since 2019	**Senior Engineer and Researcher** at the Visual Computing Group of the Center for Advanced Studies, Research and Development in Sardinia (CRS4), Italy.
2006-2018	**Expert Engineer and Researcher** at the Visual Computing Group of the Center for Advanced Studies, Research and Development in Sardinia (CRS4), Italy.
2003-2005	**Visual lab engineer** at the Visual Computing Group of the Center for Advanced Studies, Research and Development in Sardinia (CRS4), Italy.

Selected Publications

Journal Articles

- Giovanni Pintore, Alberto Jaspe-Villanueva, Markus Hadwiger, Jens Schneider, Marco Agus, Fabio Marton, Fabio Bettio, and Enrico Gobbetti. Deep synthesis and exploration

of omnidirectional stereoscopic environments from a single surround-view panoramic image. Computers & Graphics, 2024. To appear.

- Giovanni Pintore, Eva Almansa, Armando Sanchez, Giorgio Vassena, and Enrico Gobbetti. Deep Panoramic Depth Prediction and Completion for Indoor Scenes. Computational Visual Media, February 2024.
- Giovanni Pintore, Fabio Bettio, Marco Agus, and Enrico Gobbetti. Deep scene synthesis of Atlanta-world interiors from a single omnidirectional image. IEEE Transactions on Visualization and Computer Graphics, 29, November 2023. Proc. ISMAR..
- Giovanni Pintore, Marco Agus, Eva Almansa, and Enrico Gobbetti. Instant Automatic Emptying of Panoramic Indoor Scenes. IEEE Transactions on Visualization and Computer Graphics, 28(11): 3629-3639, November 2022. Proc. ISMAR.
- Giovanni Pintore, Eva Almansa, Marco Agus, and Enrico Gobbetti. Deep3DLayout: 3D Reconstruction of an Indoor Layout from a Spherical Panoramic Image. ACM Transactions on Graphics, 40(6): 250:1-250:12, December 2021. Proc. SIGGRAPH Asia 2021.
- Giovanni Pintore, Claudio Mura, Fabio Ganovelli, Lizeth Fuentes-Perez, Renato Pajarola, and Enrico Gobbetti. State-of-the-art in Automatic 3D Reconstruction of Structured Indoor Environments. Computer Graphics Forum, 39(2): 667-699, 2020.
- Giovanni Pintore, Fabio Ganovelli, Alberto Jaspe Villanueva, and Enrico Gobbetti. Automatic modeling of cluttered multi-room floor plans from panoramic images. Computers Graphics Forum, 38(7): 347-358, 2019.
- Giovanni Pintore, Fabio Ganovelli, Ruggero Pintus, Roberto Scopigno, and Enrico Gobbetti. 3D floor plan recovery from overlapping spherical images. Computational Visual Media, 4(4): 367-383, December 2018.
- Giovanni Pintore, Ruggero Pintus, Fabio Ganovelli, Roberto Scopigno, and Enrico Gobbetti. Recovering 3D existing-conditions of indoor structures from spherical images. Computers & Graphics, 77: 16-29, December 2018.
- Giovanni Pintore and Enrico Gobbetti. Effective Mobile Mapping of Multi-room Indoor Structures. The Visual Computer, 30(6–8): 707-716, 2014.
- Fabio Marton, Marco Agus, Enrico Gobbetti, Giovanni Pintore, and Marcos Balsa Rodriguez. Natural exploration of 3D massive models on large-scale light field displays using the FOX proximal navigation technique. Computers & Graphics, 36(8): 893-903, December 2012.
- Marco Agus, Fabio Bettio, Andrea Giachetti, Enrico Gobbetti, José Antonio Iglesias Guitián, Fabio Marton, Jonas Nilsson, and Giovanni Pintore. An interactive 3D medical visualization system based on a light field display. The Visual Computer, 25(9): 883-893, 2009.
- Marco Agus, Enrico Gobbetti, José Antonio Iglesias Guitián, Fabio Marton, and Giovanni Pintore. GPU Accelerated Direct Volume Rendering on an Interactive Light Field Display. Computer Graphics Forum, 27(3): 231-240, 2008. Proc. Eurographics 2008.

- Fabio Bettio, Enrico Gobbetti, Fabio Marton, and Giovanni Pintore. Scalable Rendering of Massive Triangle Meshes on Light Field Displays. Computers & Graphics, 32(1): 55-64, February 2008.

Conference Papers

- Giovanni Pintore, Alberto Jaspe Villanueva, Markus Hadwiget, Enrico Gobbetti, Jens Schneider, and Marco Agus. PanoVerse: automatic generation of stereoscopic environments from single indoor panoramic images for Metaverse applications. In Proc. Web3D 2023 - 28th International ACM Conference on 3D Web Technology, October 2023. DOI: 10.1145/3611314.3615914. Honorable mention award in the best paper category at Web3D 2023.
- Giovanni Pintore, Marco Agus, Eva Almansa, Jens Schneider, and Enrico Gobbetti. SliceNet: deep dense depth estimation from a single indoor panorama using a slice-based representation. In Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Pages 11531-11540, 2021. DOI: 10.1109/CVPR46437.2021.01137. Selected as oral presentation.
- Giovanni Pintore, Marco Agus, and Enrico Gobbetti. AtlantaNet: Inferring the 3D Indoor Layout from a Single 360 Image beyond the Manhattan World Assumption. In Proc. ECCV. Pages 432-448, August 2020.
- Giovanni Pintore, Claudio Mura, Fabio Ganovelli, Lizeth Fuentes-Perez, Renato Pajarola, and Enrico Gobbetti. Automatic 3D Reconstruction of Structured Indoor Environments. In SIGGRAPH 2020 Courses. Pages 10:1-10:218, August 2020.
- Giovanni Pintore, Fabio Ganovelli, Ruggero Pintus, Roberto Scopigno, and Enrico Gobbetti. Recovering 3D indoor floor plans by exploiting low-cost spherical photography. In Pacific Graphics 2018 Short Papers. Pages 45-48, October 2018.
- Marco Agus, Enrico Gobbetti, Fabio Marton, Giovanni Pintore, and Pere-Pau Vázquez. Mobile Graphics. In SIGGRAPH Asia 2017 Courses. Pages 12:1-12:259, November 2017.
- Giovanni Pintore, Fabio Ganovelli, Roberto Scopigno, and Enrico Gobbetti. Mobile metric capture and reconstruction in indoor environments. In Proc. SIGGRAPH Asia Symposium on Mobile Graphics and Interactive Applications. Pages 1:1-1:5, November 2017.
- Marco Agus, Enrico Gobbetti, Fabio Marton, Giovanni Pintore, and Pere-Pau Vázquez. Mobile Graphics. In Adrien Bousseau and Diego Gutierrez, editors, Proc. EUROGRAPHICS Tutorials, April 2017.
- Giovanni Pintore, Fabio Ganovelli, Enrico Gobbetti, and Roberto Scopigno. Mobile reconstruction and exploration of indoor structures exploiting omnidirectional images. In Proc. SIGGRAPH Asia Mobile Graphics and Interactive Applications. Pages 1:1-1:4, December 2016.
- Giovanni Pintore, Fabio Ganovelli, Enrico Gobbetti, and Roberto Scopigno. Mobile Mapping and Visualization of Indoor Structures to Simplify Scene Understanding and

Location Awareness. In Computer Vision - ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II. Pages 130-145, October 2016. Springer.

- Giovanni Pintore, Valeria Garro, Fabio Ganovelli, Enrico Gobbetti, and Marco Agus. Omnidirectional image capture on mobile devices for fast automatic generation of 2.5D indoor maps. In Proc. IEEE Winter Conference on Applications of Computer Vision (WACV). Pages 1-9, February 2016.
- Giovanni Pintore, Marco Agus, and Enrico Gobbetti. Interactive mapping of indoor building structures through mobile devices. In Proc. 2nd International Conference on 3D Vision. Volume 2. Pages 103-110, December 2014.
- Giovanni Pintore, Enrico Gobbetti, Fabio Ganovelli, and Paolo Brivio. 3DNSITE: A networked interactive 3D visualization system to simplify location recognition in crisis management. In Proc. ACM Web3D International Symposium. Pages 59-67, August 2012. ACM Press. New York, NY, USA.
- Giovanni Pintore, Enrico Gobbetti, Fabio Marton, Russell Turner, and Roberto Combet. An Application of Multiresolution Massive Surface Representations to the Simulation of Asteroid Missions. In Eurographics Italian Chapter Conference. Pages 9-16. Eurographics Association, November 2010.
- Fabio Bettio, Enrico Gobbetti, Fabio Marton, and Giovanni Pintore. Multiresolution Visualization of Massive Models on a Large Spatial 3D Display. In Proc. Eurographics Symposium on Parallel Graphics and Visualization, May 2007. Eurographics Association. Aire-la-Ville, Switzerland.
- Fabio Bettio, Enrico Gobbetti, Fabio Marton, and Giovanni Pintore. High-quality networked terrain rendering from compressed bitstreams. In Proc. ACM Web3D International Symposium. Pages 37-44, April 2007. ACM Press. New York, NY, USA.
- Marco Agus, Enrico Gobbetti, Giovanni Pintore, Gianluigi Zanetti, and Antonio Zorcolo. Real Time Simulation of Phaco-emulsification for Cataract Surgery Training. In Workshop in Virtual Reality Interactions and Physical Simulations (VRIPHYS 2006), November 2006. Eurographics Association. Conference held in Madrid, Spain, November 6-7.
- Tibor Balogh, Zsuzsa Dobranyi, Tamas Forgacs, Attila Molnar, Laszlo Szloboda, Enrico Gobbetti, Fabio Marton, Fabio Bettio, Giovanni Pintore, Gianluigi Zanetti, Eric Bouvier, and Reinhard Klein. An Interactive Multi-User Holographic Environment. In SIGGRAPH 2006 Emerging Technologies Proceedings. ACM SIGGRAPH. Addison-Wesley, August 2006.
- Fabio Bettio, Francesca Frexia, Enrico Gobbetti, Giovanni Pintore, Gianluigi Zanetti, Tibor Balogh, Tamas Forgacs, Tibor Agocs, and Eric Bouvier. Collaborative immersive visualization without goggles - experiences in developing a holographics display system for medical applications. In Proceedings of the Fifth MIMOS Conference, November 2005. CD ROM Proceedings.

Book chapters

- Giovanni Pintore, Marco Livesu, and Alberto Signoroni, editors. Smart Tools and Apps for Graphics. Eurographics Association, 2018.
- Giovanni Pintore and Filippo Stanco, editors. Smart Tools and Apps for Graphics. Eurographics Association, 2016.
- Fabio Bettio, Francesca Frexia, Andrea Giachetti, Enrico Gobbetti, Giovanni Pintore, Gianluigi Zanetti, Tibor Balogh, Tamas Forgacs, Tibor Agocs, and Eric Bouvier. A Holographic Collaborative Medical Visualization System. In J. D. Westwood, editor, Medicine Meets Virtual Reality 2006, IOS, Amsterdam, The Netherlands, January 2006.
- Fabio Bettio, Francesca Frexia, Andrea Giachetti, Enrico Gobbetti, Giovanni Pintore, and Gianluigi Zanetti. 3D Functional Models of Monkey Brain through Elastic Registration of Histological Sections. In International Conference on Image Analysis and Processing, Springer Verlag, New York, NY, USA, September 2005.