STAG: Smart Tools and Applications in Graphics (2024)
A. Caputo, V. Garro, U. Castellani, and A. Giachetti (Editors)

Evaluating AI-based static stereoscopic rendering of indoor panoramic scenes

Sara Jashari[1], Muhammad Tukur[1], Yehia Boraey[2], Mahmood Alzubaidi[1], Giovanni Pintore[3,4], Enrico Gobbetti[3,4], Alberto Jaspe Villanueva[5], Jens Schneider[1], Noora Fetais[2], Marco Agus[1]

[1]College of Science and Engineering, Hamad Bin Khalifa University, Qatar;
[2]College of Engineering, Qatar University, Qatar;
[3]National Research Center in HPC, Big Data, and QC, Italy;
[4]CRS4, Italy;
[5]KAUST, Saudi Arabia.

Abstract

Panoramic imaging has recently become an extensively used technology for the representation and exploration of indoor environments. Panoramic cameras generate omnidirectional images that provide a comprehensive 360-degree view, making them a valuable tool for applications such as virtual tours in real estate, architecture, and cultural heritage. However, constructing truly immersive experiences from panoramic images presents challenges, particularly in generating panoramic stereo pairs that offer consistent depth cues and visual comfort across all viewing directions. Traditional stereo-imaging techniques do not directly apply to spherical panoramic images, requiring complex processing to avoid artifacts that can disrupt immersion. To address these challenges, various imaging and processing technologies have been developed, including multi-camera systems and computational methods that generate stereo images from a single panoramic input. Although effective, these solutions often involve complicated hardware and processing pipelines. Recently, deep learning approaches have emerged, enabling novel view generation from single panoramic images. While these methods show promise, they have not yet been thoroughly evaluated in practical scenarios. This paper presents a series of evaluation experiments aimed at assessing different technologies for creating static stereoscopic environments from omnidirectional imagery, with a focus on 3DOF immersive exploration. A user study was conducted using a WebXR prototype and a Meta Quest 3 headset to quantitatively and qualitatively compare traditional image composition techniques with AI-based methods. Our results indicate that while traditional methods provide a satisfactory level of immersion, AI-based generation is nearing a quality level suitable for deployment in web-based environments.

CCS Concepts

• ***Computing methodologies*** → ***Computer vision; Virtual reality; Neural networks;***

1. Introduction

The advent of panoramic imaging has revolutionized the way we represent and explore indoor environments. Panoramic images provide a comprehensive 360-degree view of a scene, making them highly effective for virtual tours and other immersive experiences [PGGS16, dJ23]. This capability allows users to explore spaces interactively, making it a valuable tool for real estate, architecture, and cultural heritage sectors, among others. However, while panoramic imaging captures a wide field of view, constructing realistic immersive experiences from these images presents significant challenges. One of the main challenges in creating truly immersive experiences from panoramic images is the generation of effective panoramic stereo pairs that are valid for all viewing directions. Traditional stereo imaging techniques, which involve capturing two images from slightly different viewpoints, do not directly apply to the spherical nature of panoramic images. The construction of panoramic stereo pairs that maintain consistent depth cues and visual comfort across all viewing directions requires sophisticated processing to avoid artifacts such as vertical disparities and inconsistent depth perception, which can lead to discomfort or break the immersion [PBEP01]. Various imaging and processing technologies have been developed to address these challenges, ranging from multi-camera rigs that simultaneously capture panoramic stereo pairs to computational methods that generate stereo images from monocular panoramic input [MCE*17]. These methods, while capable of producing high-quality stereoscopic panoramas, often involve complex hardware setups and extensive processing pipelines. The need for precise alignment, calibration, and stitching in multi-camera systems adds to the complexity, making these solutions difficult to implement and apply in practical scenarios. In response to these limitations, recent advancements in deep learning have opened new avenues for generating novel views from single panoramic images [WGD*22, PBAG23]. These deep-learning-based methods leverage neural networks to infer depth and generate stereoscopic views, potentially offering a more flexible and scalable solution compared to traditional approaches. However, despite their promising capabilities, these methods have not yet been thoroughly evaluated in real-world scenarios to determine their effectiveness in supporting immersive exploration and their suitability for different types of content.

In this paper, we describe a series of evaluation experiments designed to assess various technologies for creating static stereoscopic environments from omnidirectional imagery, specifically targeting 3DOF (Degrees of Freedom) immersive exploration scenarios. Our objective is to provide a comprehensive comparison of these methods, ranging from traditional image composition techniques to advanced AI-based synthetic generation methods. To achieve this, we have conducted a user study involving both quantitative and qualitative assessments of the stereoscopic experiences generated by these different methods. For the evaluation, we implemented a WebXR prototype and conducted user tests using a Meta Quest 3 headset. This setup allowed us to analyze the performance of each method in a realistic, interactive environment, reflecting typical use cases for immersive applications. Our preliminary findings suggest that while traditional methods for static image composition can achieve a satisfactory level of immersion, AI-based synthetic generation techniques are rapidly approaching a level of quality that is suitable for deployment in web-based environments, such as those envisioned for the Metaverse. This research aims to provide valuable insights and guidelines for developers and researchers working on immersive content creation. By highlighting the strengths and limitations of different stereoscopic generation techniques, we hope to contribute to the development of more accessible and effective tools for panoramic imaging, ultimately enhancing the user experience in virtual environments.

2. Related work

Our work deals with the generation and assessment of immersive panoramic stereoscopic environments. In the following, we discuss the literature most closely related to our study, while we refer readers to the recent surveys about visual computing for omnidirectional imagery [dSJ23], their application in the extended reality domain [ZZZZ23], and the methodologies for assessing immersiveness in VR environments [BMB24, MCMB24].

2.1. Generation of immersive omnidirectional environments

When displaying a single panoramic image on a VR headset, the conventional approach involves projecting the image onto a spherical dome positioned around the user's head. In this method, the eye position is factored in to produce the correct perspective for each eye. However, because all points in the scene are projected onto the dome at the same distance (determined by the dome's radius), the resulting parallax effects are minimal. To achieve a more realistic depth perception, it is necessary to incorporate the scene's geometry into the view synthesis. Small changes in eye position can alter the visibility of scene elements, making it essential to not only estimate the geometry but also manage occlusions and disocclusions effectively. To enable 6DOF VR exploration, various methods try to recover 3D geometry information through proxy representations: for example, OmniPhotos [BYLR20] obtain motion parallax by considering a single sweep with a consumer 360° video camera as input, and by treating vertical distortion with a novel deformable proxy geometry, that is fit to a sparse 3D reconstruction of captured scenes. By considering a few seconds of casually captured 360 video, EgoNerf [CKK23] builds a neural radiance field representation enabling high-quality rendering from novel viewpoints, and by accelerating NeRF using feature grids adopting spherical coordinate instead of conventional Cartesian coordinate. Very recently, the approach was extended to Omnidirectional Local Radiance Fields (OmniLocalRF) [CJK24] to deal with the problem of synthesizing novel views in the presence of dynamic objects including the photographer. Another alternative approach for fast novel viewpoint synthesis involves using layered depth representations, where each pixel is assigned multiple depth values. These layered representations allow for view synthesis by extrapolating and in-painting to fill in missing areas [HK18]. This technique has been successfully adapted for use with single panoramic images [SKC*19]. Additionally, different layered approaches have been explored to improve accuracy: Broxton et al. [BFO*20] introduced light field videos based on layered mesh representations, while Lin et al. [LXM*20] proposed a multi-depth panorama method. Another variation of layered depth representations uses multiple flat planes at fixed depths to create multi-plane images (MPI), which can be processed using convolutional neural networks [ZTF*18, TS20]. However, MPIs are generally limited to viewpoints near the original position, and their quality diminishes as the viewpoint moves further away. To overcome this limitation, adaptive sampling strategies have been proposed [LSR*20]. The concept of capturing scenes at multiple fixed depths has also been extended to panoramic imaging using alternative capture proxies, such as multi-spherical images (MSI) [ALG*20] and multi-cylinder images (MCI) [WGD*22]. Very recently, Pintore et al. [PBAG23, PJVH*23, PJVH*24] proposed various methods for enabling immersive exploration of indoor omnidirectional scenes, based on lightweight deep learning architecture for depth estimation and inpainting of dis occluded areas in a way to enable real-time novel view synthesis. In this work, we further exploit the latter framework, by integrating and comparing schemes for the generation of static 3DOF stereoscopic environments [GD13, HZZ*24], and we assess it on real-world scenarios by comparing it to scenes created through acquired imagery.

2.2. Assessment of panoramic stereoscopic environments

Given the nature of stereoscopic omnidirectional images (SOI), a variety of quality issues may arise in the creation, transmission, and display processes that can negatively affect the correct perception and immersiveness. In recent years, various methods have been proposed to develop accurate and easy-to-use omnidirectional image quality assessment (OIQA) methodologies [ZW24]. These can be subdivided into two main categories: subjective and objective methods. For the subjective evaluation, several quality databases have been compiled in recent years, where each panoramic environment is associated with a perceptual visual quality, depth quality, and general QoE [XLZ*19, QJY*20]. These databases are then used to create data-driven solutions that try to perform automatically quality assessment by matching the subjective assessment. For example, Chai et al. [CSJ*21] use deformable convolutions to ensure the invariant receptive fields of convolutional kernels on Equi-Rectangular Projection (ERP), in a three-channel network involving left-view, right-view, and binocular-difference. Chen et al. [CXLZ20] propose a stereoscopic omnidirectional image quality evaluator (SOIQE) involving a predictive coding theory-based binocular rivalry module and a multi-view fusion module. In the binocular rivalry module, predictive coding theory is considered to simulate the competition between high-level patterns and calculate the similarity and rivalry dominance to obtain the quality scores of viewport images. Differently from previous methods, the quantitative assessment techniques try to predict perceptual quality automatically, in a way that may be integrated into working VR processing systems for optimal performance. Zhou et al. [ZW24] propose a depth quality index (DQI) for efficient no-reference (NR) depth quality assessment of stereoscopic omnidirectional images, built upon multi-color-channel, adaptive viewport selection, and interocular discrepancy features. You et al. [YJJ*23] introduce a visual perception-oriented quality metric for stereo omnidirectional videos, based on features that reflect distortions to be extracted from viewports and equiangular cubemap (EAC) projections, together with features extracted from wavelet decomposition for temporal domain bandpass filtering on consecutive frames, to feed a random forest model to predict the quality score. Another strategy for assessing immersive environments consists of user studies involving questionnaires for assessing quality and immersiveness. Recently, assessment questionnaires specific for VR have been designed [FKTK20], extending the established NASA-TLX questionnaire for usability of computer systems. In our experiments, we consider a slightly updated version of Virtual Reality Questionnaire Toolkit (VRQT) [FKTK20], and similarly to Kuntzer et al. [KSSR24] we integrate it inside the immersive experience, in a way to reduce the bias due to the loss of immersiveness and eventual memory effects during the experiments [SHKP21].

3. Methods

3.1. Overview

Our system prototype follows a typical end-to-end processing workflow for Spherical Omnidirectional Images (SOIs), which consists of three main steps [QJY*20]. The first step involves creating SOIs by composing together several images captured with an imaging device, composed by panoramic cameras or fish-eye lenses, allowing for a full field of view (FoV) of 360° × 180°. In the second step, because of constraints related to storage and transmission, the SOIs are converted from a spherical representation to an Equirectangular Projection (ERP) format. This conversion allows the images to be encoded using standard 2D image or video encoding techniques. The final step occurs when the encoded ERP images are sent to the client for viewing. In general, an inverse projection transforms the ERP format back to a spherical surface, and viewport rendering is applied to reconstruct the scene as seen by the user. These adaptation processes, which are unique to immersive environments, introduce additional complexity to the perceptual characteristics of SOIs compared to typical 2D or 3D images. This complexity can be observed in three key elements: viewport rendering, user interaction, and stereoscopic perception.

- **Viewport Rendering:** During the viewport rendering process, geometric transformations of compression artifacts in ERP images occur, especially in regions near the poles.
- **User Interaction:** While viewing, users can move their heads to change the position of the viewport, selecting the areas they are interested in exploring. Within each viewport, some regions may attract more attention due to eye movements, particularly those with prominent objects.
- **Stereoscopic Perception:** When the quality of the left and right views of SOIs is consistent, binocular fusion happens. However, if there are significant disparities between the two views, binocular rivalry can occur, potentially causing visual discomfort. This is more pronounced in Head-Mounted Displays (HMDs), which are closer to the eyes and provide a wider FoV than traditional 3D displays.

3.2. Stereoscopic Environment Creation

The processing pipeline for creating stereoscopic omnidirectional scenes consists of the following steps:

- **Image acquisition:** in our work we use omnidirectional cameras for creating the input image for the AI-based image synthesis component as well as ground truth for evaluating the latter;
- **Image synthesis:** in this work we exploit a recent novel pose generation scheme [PJVH*23,PJVH*24] for creating poses able to form effective panoramic stereo couples;
- **Image composition:** the various poses need to be stitched and blended in a way to generate stereo couples as artifact-free as possible and able to provide correct stereo cues in all directions. We designed a composing scheme taking into account the viewing direction as well as the depth estimation of the various images.
- **Environment deployment:** we developed a WebXR-based client-server application for stereoscopic viewport rendering.

In the following, we detail the various steps.

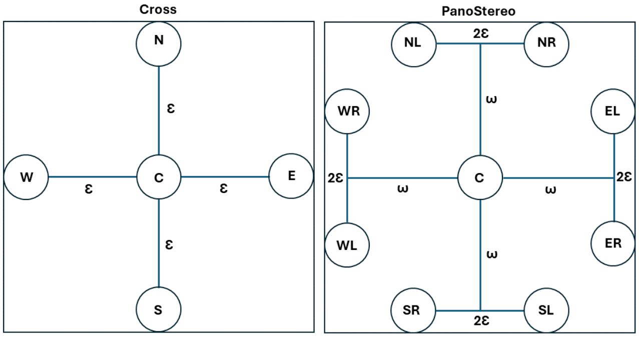

Figure 1: **Image acquisition scheme.** Left: cross pattern. Right: PanoVerse/Vuze+ pattern. The parameters are the Intra-Pupil-Distance ϵ and the axis distance ω. In our experiments, we used $\epsilon = 3cm$, and $\omega = 6, 10cm$.

Image acquisition To create panoramic stereo couples, we acquired panoramic images through the Insta360 X3 omnidirectional camera. The acquisition process involved capturing images at varying distances from a central reference point and according to different patterns, as represented in Figure 1:

- A cross pattern with the camera oriented towards the north, and pictures acquired in the north (N), south (S), west (W), and east (E) positions;
- A pattern named PanoStereo and emulating the scheme employed in PanoVerse [PJVH*23], in which eight images with the camera pointing towards north are acquired and representing stereo viewing for the different quadrants, North-Left(NL), North-Right(NR), East-Left(EL), East-Right(ER), South-Left(SL), South-Right(SR), West-Left(WL) and West-Right(WR).

All images were taken with the same camera: for ensuring the consistency of positions during the acquisition process, we attached a ruler in the table before placing a tripod holding the camera, while the orientation is kept consistent to simplify the composition between the various images.

In the schemes depicted in Figure 1, the parameter ϵ represents the average half intra-pupil distance, and in our experiments was fixed to 3cm. For the PanoStereo setting, the parameter ω is variable, and we acquired scenes with two different distances: 6 cm and 10 cm (corresponding to distances of 6.7 and 10.4 cm from the center), to fit with the design used by the Vuze+ camera (in the first case) and the average distance between eyes and head axis (for the second case). In both configurations, the central image is not used for composing the final stereo couple, but as input for generating the synthetic environments.

Image synthesis Given a single panoramic image, a novel view synthesis model is able to generate another panoramic image from a different camera position. To this end, we used the recent architecture proposed by Pintore et al. [PBAG23, PJVH*24]. The network is composed of two modules: the first one estimates a depth map from a single panoramic input, and the second reprojects the views to the desired position, synthesizing a complete image that fills dis occluded areas with plausible content. The network employs a lightweight gated architecture with a dilated bottleneck, ensuring scalability to larger images or embedded hardware while maintaining high visual detail during view reprojection. The framework is characterized by a unified network architecture with custom training strategies for both depth estimation and view synthesis. The same lightweight network is used for both tasks by adjusting the activation function and training mode. For novel view synthesis, a specific photometric loss is combined with a GAN approach, enabling the generation of photorealistic views at low computational cost. Additionally, we consider a super-resolution GAN architecture to enhance stereo image resolution [WYW*18]. For our view synthesis experiments, we considered the same patterns used for image acquisition, and represented in Fig. 1: the image acquired in the central position (C) is used as input for generating the other images, either for the cross pattern (W, N, E, and S), for the PanoVerse pattern (NL, NR, EL, ER, SL, SR, WL, WR), and for the extended PanoStereo pattern consisting by multiple inferences composing a Multi Center of Projection image (in our experiments we considered 32 images).

Image composition For obtaining panoramic stereo couples, we need to stitch and blend together various portions from the acquired or generated panoramic image. In this work, we extend the angular blending scheme proposed by Pintore et al. [PJVH*23] by exploiting the depth estimation signal provided by the model. Specifically, given two panoramic images I_c and I_n that need to be blended in an angular portion w, and the corresponding estimated depths d_c and d_n, the blended image I is obtained as follows:

$$I = \gamma I_c + (1-\gamma) I_n, \tag{1}$$

where γ is a blending factor depending on the angle between two adjacent views, and the depth difference. Given normalized pixel coordinates (x, y) relative to two adjacent views and a percent window w, the angular blending factor is computed as follows:

$$\tau(x) = \begin{cases} 1 & x \leq (\frac{1}{2} - w) \\ \frac{1}{2}(1+\cos(\pi \frac{x+w-\frac{1}{2}}{2w})) & (\frac{1}{2} - w) < x < (\frac{1}{2} + w) \\ 0 & x \geq (\frac{1}{2} + w), \end{cases} \tag{2}$$

that we further compose with a sigmoid function depending on the depth differences between the corresponding pixels:

$$\gamma(x,y) = \delta\tau(x) + (1-\delta)\sigma\left(\frac{\sin(\pi y)(d_n - d_c) + \eta}{\eta}\right), \tag{3}$$

where δ is a tunable weight, and η is the distance between the camera centers of the two adjacent images ($\sqrt{2}\epsilon$ for the cross pattern, $\sqrt{\omega^2 + \epsilon^2}$ for the PanoVerse pattern). In all our experiments, we considered $\delta = 0.2$. For our experiments, we considered various patterns for generating stereoscopic panoramic images:

- **Standard:** no blending is applied, and the stereo is composed by the two images W and E from the cross pattern. This composition has the drawback of not providing correct stereo cues when observing the scene towards the south direction.
- **Composed:** in this case the same W and E images are used for generating the stereo couple, but blending is performed to switch the portion of the panoramic images when pointing to the south direction (W becomes E, and E becomes W). This composition solves the problem for stereo cues in the south direction, but not for lateral viewing since the two views would result occluded.
- **Cross:** in this case the images of the cross pattern are blended according to quadrants (W and E towards the north direction, N and S towards the east direction, E and W towards the south direction, and S and N towards the west direction).
- **PanoVerse:** the same composition scheme used by Pintore at al. [PJVH*23] and considering the PanoStereo acquisition pattern. Even in this case, the blending is applied according to quadrants, but the views are separated between the eyes. This pattern better represents the head rotation but is prone of blending and stitching artifacts. The same scheme is employed by Vuze+ virtual reality camera.
- **PanoStereo:** the same composition scheme used by Pintore at al. [PJVH*24] and considering the blend of multiple images to form a Multi Center of Projection Image (MCOP) with a radius of 3 cm. It can be considered an extension of the cross pattern, and it significantly reduces blending and stitching artifacts. On

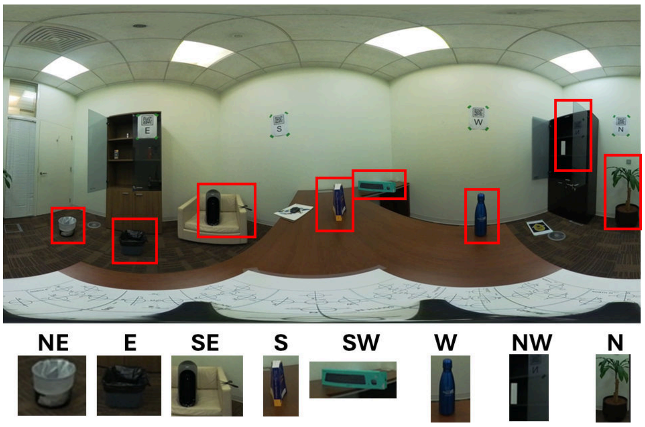

Figure 2: **Experimental setup:** we acquired images in an office with various objects placed at variable distance from the camera and according to specific directions.

the other side, it is very complicated to acquire images according to this pattern without specialized hardware.

Environment deployment We developed a WebXr viewer for stereoscopic panoramic images enabling users wearing Head Mounted Displays to inspect the environments with 3 degrees of freedom. The two panoramic images generated through the various composition schemes are displayed to the viewer using the same method as regular panoramas, where the left image is shown to the left eye and the right image to the right eye. The high-resolution stereo panoramas act as textures applied to two spheres—one surrounding the left eye and the other surrounding the right eye. This setup creates a distinct environmental map for each eye, and the system retrieves head position data from the headset's sensors during each frame of the animation loop. This information is used to calculate the correct perspective projections for each eye. Both panoramas share the same viewing transformation, ensuring alignment with head orientation. As the viewer looks in a specific direction, the correct perspectives for each eye are rendered in the headset, with accurate horizontal parallax in the central view and slight degradation towards the edges. The stereoscopic effect is achieved because human depth perception is concentrated in the central field of vision, with minimal effect in the peripheral areas. The website panostereo.onrender.com contains all scenes generated in this study and used for the assessment. It can be accessed through WebXR enabled browsers, and the scenes can be explored with most HMD headsets: we tested with Meta Quest 2 and 3, and with Google cardboard on an Android smartphone.

4. Experimental setup

We designed the assessment experiments in a way to evaluate the quality of stereo perception, the immersiveness, and the artifacts due to image stitching and generation. To this end, we chose an indoor environment represented by an office with a variety of objects at different distances from the viewpoint and with specific directional reference points, in a way to be able to ask specific questions related to the observation towards specific directions. We placed the camera at the center of the room, and we used the patterns represented in Fig. 1 to acquire ground truth images and to compose them according to the composition schemes described above. The scene is represented in Fig. 2 and it contains specific focus objects along the eight wind rose directions. The images acquired with the Insta360 camera were blended to form the following ground truth stereo couples (represented in Fig. 3): standard, composed, cross, and PanoVerse pattern. On the other side, we considered the image in the center C of the acquisition pattern (see Fig. 1) to run the novel view synthesis model in a way to construct various generated stereo couples (represented in Fig. 4): stereo, compose, cross, and PanoStereo. We also considered some generated synthetic stereo couples obtained by processing scenes from the dataset Structured3D [ZZL*20]: Fig. 5 shows an example of generated stereo couples with the cross pattern and the stereo pattern. We rescaled all equirectangular composed scenes to the same resolution 3072X1536: for the ground truth ones we reduced them through standard bicubic filtering, while for the generated ones we used a deep-learning based super-resolution method [WYW*18]. With all these stereo couples, we performed a series of assessment tests involving quantitative methods and qualitative user evaluation.

Table 1: **Qualitative expert assessment**: experts evaluated the office scenes and provided their indications with respect to the various directions (S for indicating the presence of stereo cue, and D for indicating the presence of distortions and ghosts). The ∗ in place of S indicates that experts do not have consensus over correct stereo perception, while instead of D indicates all of them did not perceive artifacts.

	GT-B	GN-B	GT-S	GN-S	GT-C	GN-C	GT-P	GN-P
N	S*	S*	S*	S*	S*	S*	S*	S*
NE	S*	S*	S*	S*	S*	S*	SD	S*
E	S*	SD	SD	SD	S*	S*	SD	S*
SE	*D	*D	**	**	SD	SD	*D	S*
S	*D	*D	SD	SD	S*	SD	SD	SD
SW	*D	*D	*D	*D	SD	SD	*D	SD
W	**	**	S*	S*	S*	SD	S*	SD
NW	**	*D	S*	S*	S*	S*	S*	S*

4.1. User study procedure

For what concerns the user assessment, the main goal was to compare the immersiveness of the various scenes obtained through the various composition schemes applied to the ground truth and the generated images. To this end, we considered two kinds of sessions: one accurate qualitative session involving subjects with VR experience, and one general user study involving subjects without VR experience. All experiments consisted of letting the users sit on an office rolling chair to conform to the same height as the camera, and observe the scenes with a Meta Quest 3 HMD (see Fig. 7). In the following we denote the scenes created by composing acquired images as ground truth (GT), and the scenes crated by composing AI-generated images as generated (GN).

Expert assessment We performed a preliminary expert assessment, in which five subjects were requested to observe the scene

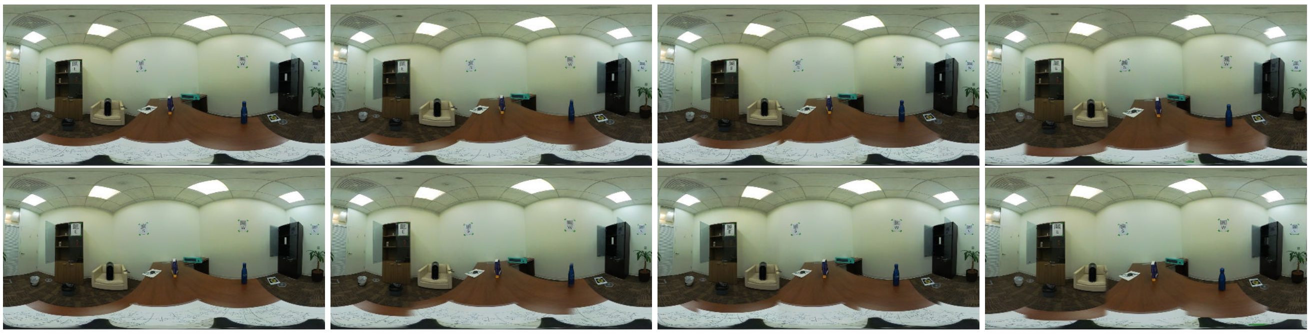

Figure 3: **Ground truth stereo couples:** on the top the left image, and on the bottom the right image. From left to right: standard stereo, composed stereo, cross stereo, and PanoVerse pattern [PJVH*23]

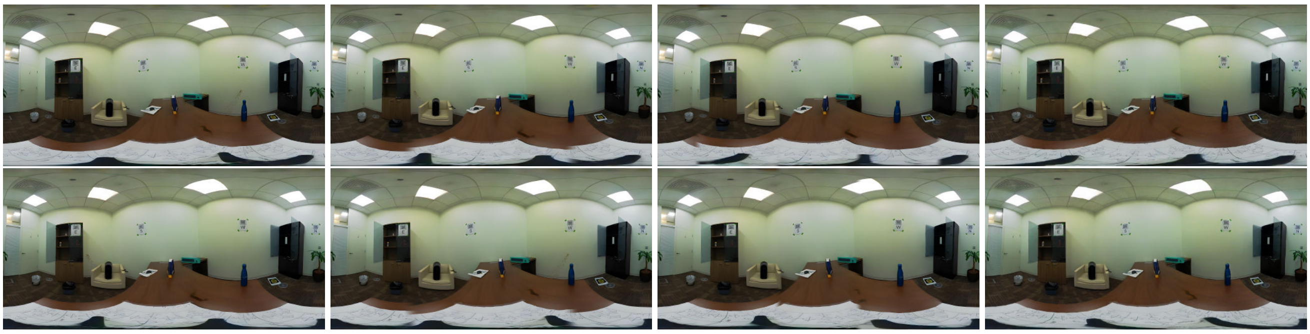

Figure 4: **Generated stereo couples:** on the top the left image, and on the bottom the right image. From left to right: standard stereo, composed stereo, cross stereo, and PanoStereo pattern [PJVH*24]

.

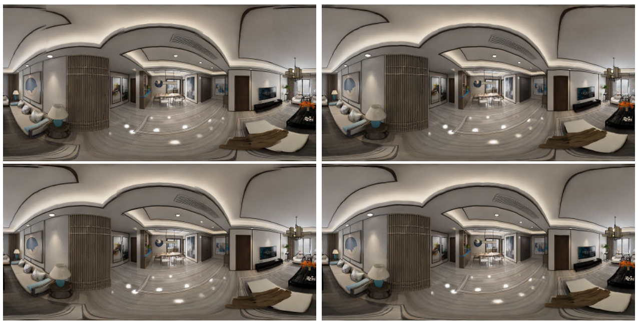

Figure 5: **Synthetic stereo couples:** on the top the left image, and on the bottom the right image. From left to right: cross stereo, and PanoStereo composition [PJVH*24]

.

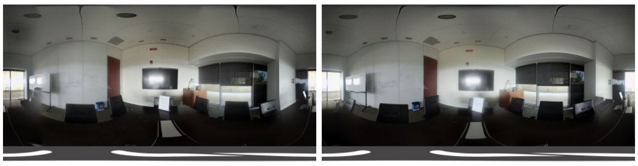

Figure 6: **Vuze+ example:** for comparison, we considered a scene acquired and processed through Vuze+ stereo camera and processing software.

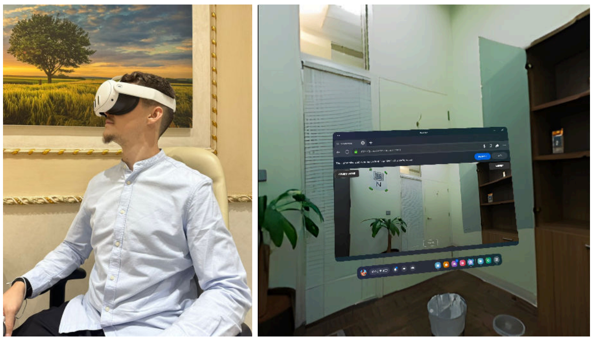

Figure 7: **User study setup:** users can explore the scene through a Meta Quest 3 (left) by sitting on an office wheeled chair and fill directly the VR questionnaire survey with the controller (right).

towards all wind rose directions according to the scheme in Fig. 2, and express their opinion related to the stereo perception, the presence of artifacts due to blending or AI-based generation, or eventual ghosts.

According to this procedure, the experts evaluated the following scenes created from the office setup (see Fig. 2): standard stereo baseline with 2 images (GT-B), composed stereo with 2 images (GT-S), cross with 4 images (GT-C), and PanoVerse with 8 images

Table 2: **VR Questionnaire**

Id	Question	Range
Q1	What would you say about the stereoscopic VR environment's visual clarity?	Very blurry (1) - Very clear (5)
Q2	Please rate the visual artifacts or distortions in the VR environment?	Not noticeable (1) - Extremely evident (5)
Q3	How immersive did you find the VR environment?	Not immersive (1) - Very immersive (5)
Q4	Did you feel a sense of presence in the VR environment?	No presence (1) - Perfect presence (5)
Q5	Did the depth perception enhance your sense of immersion in the VR environment?	No immersion (1) - Perfect immersion (5)
Q6	How satisfied are you with the current state of the VR environment?	Very dissatisfied (1) - Very satisfied (5)

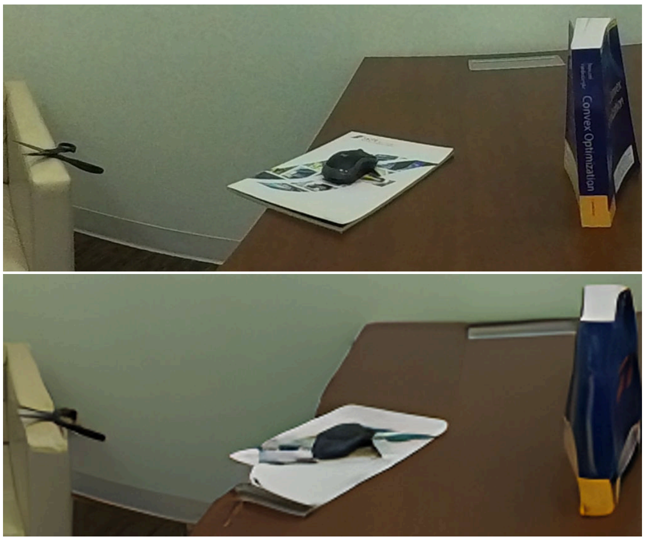

Figure 8: **Detail comparison:** on the top the ground truth acquired image, and on the bottom the AI-generated one.

(GT-P) created from acquired ground truth images, and the standard stereo with 2 images (GN-B), composed stereo with 2 images (GN-S), cross stereo with 4 images (GN-C), and PanoStereo with 32 images (GN-P) created by composing AI-generated images. The aggregated outcomes of this assessment are represented in Tab. 1 where, for each direction and each scene, users marked with S the correct stereo perception along some direction and with D the presence of some artifacts. The $*$ in place of S indicates that experts do not have consensus over correct stereo perception (at least 4 over 5), while instead of D indicates all of them did not perceive artifacts.

The outcomes confirmed the expectations about the perceived stereoscopic cues:

- both the standard schemes (GT-B and GN-B) can provide correct stereo only in the North portion of the scene;
- both the composed schemes (GT-S and GN-S) have issues in the East and West direction (lack of stereo cues, ghost and blending artifacts);
- the cross for both cases (GT-C and GN-C) and the PanoStereo scheme for generated images (GN-P) provide correct stereo cues in all directions;
- the cross scheme exhibits noticeable blending artifacts in some intermediate positions (especially scissors and couch in SE direction);
- the PanoVerse scheme (GT-P) exhibits various annoying blending artifacts in different directions;
- the PanoStereo scheme (GN-P) does not exhibit blending artifacts in any direction;
- the generated scenes exhibit generally fewer blending artifacts, but more visible reconstruction artifacts (reported ones include incomplete scissors, and writing details in the bottle and the book) (see Fig. 8 for a detailed comparison between one ground truth image and the corresponding generated one).

For what concerns the PanoVerse scheme, we also performed a qualitative test with a scene created through the Vuze+ camera (see Fig. 6), and we noticed similar artifacts due to the blending of the various fish-eye images. After this preliminary assessment, we decided to exclude the standard scheme and the PanoVerse scheme from the user study for the quantity and quality of artifacts that could bias novice subjects. For fairness of comparison, we also excluded the PanoStereo generated scene, since we were not able to create a corresponding ground truth scene with a similar number of images.

Novice assessment For the user study involving naive subjects, we considered the following five scenes: a control scene without stereo (named mono M), two scenes with composed stereo scheme (named ground truth stereo GT-S and generated stereo GN-S), and two scenes with the cross scheme (named ground truth cross GT-C and generated cross GN-C). We involved 20 subjects with normal vision and no experience in VR, and to avoid any bias we let them explore three scenes randomly selected from the five, for a total of 60 exploration sessions (12 for each scene). Users could explore freely the scenes for a few minutes, and then answer a questionnaire directly in the immersive session through the controllers (see Fig. 7). We designed the assessment questionnaire according to the standards used for VR assessment [FKTK20], inherited from NASA-TLX usability assessment forms. The questions are represented in Tab. 2 and users were asked to rate the various characteristics of the explored scene in a 5-point Likert scale. Additionally, we asked subjects to report about any eventual issues with comfort, like dizziness, eye strain, headache, nausea, vertigo, blurred vision. The outcomes of the study are discussed in the next section.

5. Results

We report on a quantitative comparison between ground truth images and AI-generated ones, and on the outcomes of the novice user study.

Table 3: **Quantitative Comparison**: metrics for comparison between ground truth and generated images for different distances: 3*cm*, 6.7*cm*, and 10.4.

Distance (in cm)	PSNR ↑	SSIM ↑	RMSE ↓
3.0	17.87	0.710	32.63
6.7	15.78	0.567	41.47
10.4	14.63	0.551	47.32

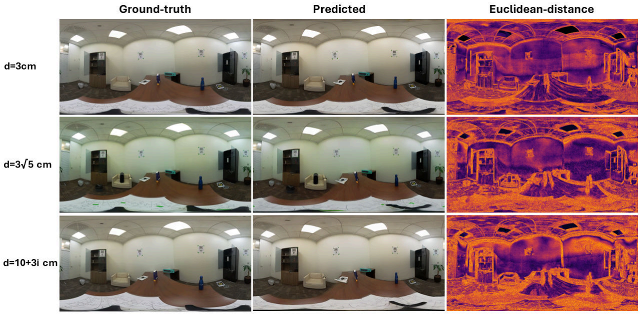

Figure 9: **Qualitative comparisons** Left: ground truth. Middle: predicted image. Right:color-mapped (inferno) L_2 distance between the ground truth and the generated image.

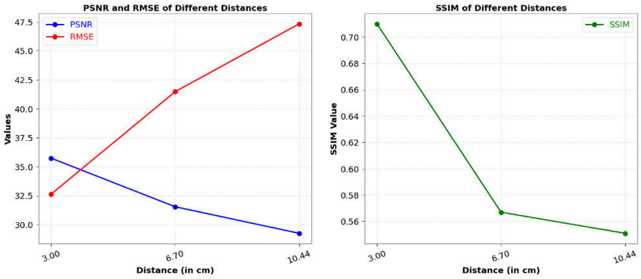

Figure 10: **Performance Metrics:** Graphs showing the average RMSE and PSNR(left), and average SSIM(right) for varying distances.

Quantitative analysis The quantitative evaluation was performed by comparing the performance metrics according to three distinct acquisition distances: 3 cm, 6.7 cm, and 10.4 cm. The evaluation metrics included Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Root Mean Square Error (RMSE). Table 3 presents a comparison of these distances, while Figures 9 and 10 visually highlight the performance variations across the captured panoramic scenes. Figure 9 provides qualitative comparisons of the generated panoramic scenes against the ground truth images, highlighting the variations in PSNR, SSIM, and RMSE across different configurations. Overall, the quantitative results highlight the advantages of using shorter distances in the generation process for enhancing the visual fidelity and structural coherence of panoramic images.

User study outcomes We performed a preliminary statistical analysis of the outcomes of the VR questionnaires submitted by the 20 naive subjects. Fig. 4 shows a table with average and standard deviations related to the Likert scores of the VR questionnaire, specifically visual clarity, distortions, immersiveness, sense of presence, depth perception, and overall satisfaction. Fig. 11 shows the full boxplots of the answers related to the same VR questionnaire. From those values, it appears a slight preference towards the cross scenes GT-C, GN-C with respect to the composed scenes GT-S, GN-S, and a slight preference towards the ground truth images GT-C, GT-S with respect to the generated ones GN-C, GN-S. We also performed a preliminary two-way ANOVA to evaluate the effects of the various composition schemes (composed versus cross) and the AI generation (ground truth versus generated). We found significant effects only for Q6 about overall satisfaction ($p = 0.02$ with $F = 5.96$) for the comparison between cross ground truth GT-C and cross generated GN-C, mostly due to the generation artifacts in small details. For the rest, immersiveness, depth perception, and sense of presence did not reveal any significant effect.

Discussion and limitations From this preliminary evaluation study, we could get the following outcomes:

- the composition schemes exhibit more artifacts and create perceptual issues according to the distance between the consecutive center of views. A scheme like cross and PanoStereo with reduced radius provide better stereo perception with respect to the original PanoVerse/Vuze+ scheme;
- the novel view synthesis process gets deteriorated with the increase of the distance, hence a reduced radius helps in limiting the number of distortions and detail artifacts;
- even the blending process, as already pointed out in [PJVH*24], is depending on the number and the distance between consecutive views;
- the generative method is able to reconstruct scenes perceptually not too distant from ground truth ones. On the other side, we will need more experiments to assess the PanoStereo scheme with respect to the number of images, that currently we excluded from our analysis for lack of time and subjects.

This study was limited to static 3DOF exploration of panoramic environments, and the system is currently not able to provide proper parallax cues to enable full 6DOF exploration of the scenes. We plan to explore novel view synthesis methods in conjunction with modern Gaussian Splatting technologies [PZL24] to address this limitation. Another limitation is related to the fact that the processing pipeline targets indoor environment and it is not optimized for outdoor scenes. We plan to investigate the generalization to outdoor scenarios in the future.

6. Conclusions

We presented a preliminary assessment of AI-based static stereoscopic rendering techniques for indoor panoramic scenes. Through a series of user studies and quantitative analyses, our findings demonstrate that while traditional stereoscopic methods offer a satisfactory level of immersion, AI-generated techniques are closing the gap in visual quality and performance. The integration of deep learning for novel view synthesis shows potential for improving the accuracy and realism of stereoscopic panoramas, positioning these approaches as viable solutions for web-based and VR applications.

Table 4: **User study outcome:** average and standard deviation of Likert-scores provided by subjects during the exploration of the various scenes.

	Gr. Truth Cross (GT-C)		Gen. Cross (GN-C)		Gr. Truth Compose (GT-S)		Gen. Compose (GN-S)		Mono(M)	
	AVG	SD	AVG	SD	AVG	SD	AVG	SD	AVG	SD
Visual Clarity	4.167	0.718	3.917	0.669	3.750	0.754	3.667	0.651	4.167	1.030
Distortions	2.167	1.115	2.917	1.084	2.083	0.793	2.167	0.718	1.083	0.289
Immersiveness	4.250	0.866	4.083	0.900	3.750	0.754	3.917	0.793	3.417	1.240
Sense of Presence	4.417	0.793	4.167	0.937	4.000	0.739	4.167	0.937	3.583	1.443
Depth Perception	4.500	0.674	4.083	0.669	4.083	0.515	4.083	0.793	3.583	1.443
Satisfaction	4.583	0.669	3.917	0.669	4.167	0.835	4.000	0.603	3.833	0.937

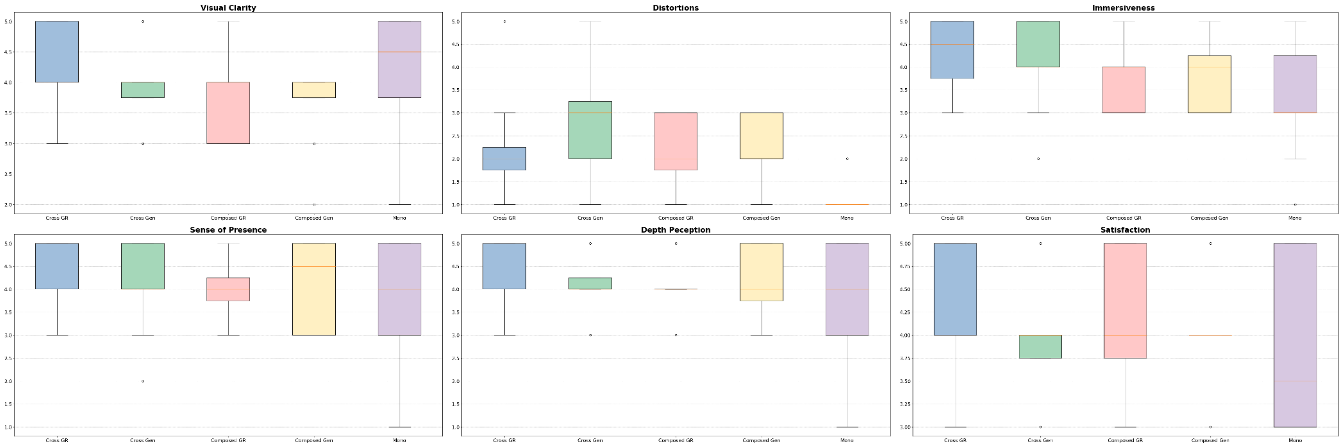

Figure 11: **User study boxplots:** boxplots for the Likert scores for the specific questions related to the exploration of the various environments.

Our preliminary findings validate the potential of AI-based rendering approaches in advancing immersive stereoscopic environments, particularly in applications that demand high-quality visual outputs, such as the Metaverse and other virtual platforms. Future work will focus on addressing the limitations in 6DOF exploration and further enhancing the blending and artifact reduction processes to achieve higher fidelity in immersive environments. Moreover, we plan to integrate these technologies for the development of virtual visits in the cultural and AEC application domain.

Acknowledgments This publication was supported by NPRP-S 14th Cycle grant 0403-210132 AIN2 from the Qatar National Research Fund (a member of Qatar Foundation). GP and EG also acknowledge the contribution of the Italian National Research Center in High-Performance Computing, Big Data, and Quantum Computing (Next Generation EU PNRR M4C2 Inv 1.4). The findings herein reflect the work and are solely the responsibility, of the authors. We thank Fabio Bettio (CRS4) for his work on developing the omnidirectional stereo techniques and tools.

References

[ALG*20] Attal B., Ling S., Gokaslan A., Richardt C., Tompkin J.: Matryodshka: Real-time 6dof video view synthesis using multi-sphere images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I* (2020), Springer, pp. 441–459. 2

[BFO*20] Broxton M., Flynn J., Overbeck R., Erickson D., Hedman P., DuVall M., Dourgarian J., Busch J., Whalen M., Debevec P.: Immersive light field video with a layered mesh representation. 86:1–86:15. 2

[BMB24] Biswas N., Mukherjee A., Bhattacharya S.: "are you feeling sick?" a systematic literature review of cybersickness in virtual reality. *ACM Computing Surveys* (2024). 2

[BYLR20] Bertel T., Yuan M., Lindroos R., Richardt C.: Omniphotos: casual 360 vr photography. *ACM Transactions on Graphics (TOG) 39*, 6 (2020), 1–12. 2

[CJK24] Choi D., Jang H., Kim M. H.: Omnilocalrf: Omnidirectional local radiance fields from dynamic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 6871–6880. 2

[CKK23] Choi C., Kim S. M., Kim Y. M.: Balanced spherical grid for egocentric view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2023), pp. 16590–16599. 2

[CSJ*21] Chai X., Shao F., Jiang Q., Meng X., Ho Y.-S.: Monocular and binocular interactions oriented deformable convolutional networks for blind quality assessment of stereoscopic omnidirectional images. *IEEE Transactions on Circuits and Systems for Video Technology 32*, 6 (2021), 3407–3421. 3

[CXLZ20] Chen Z., Xu J., Lin C., Zhou W.: Stereoscopic omnidirectional image quality assessment based on predictive coding theory. *IEEE Journal of Selected Topics in Signal Processing 14*, 1 (2020), 103–117. 3

[dJ23] da Silveira T. L., Jung C. R.: Omnidirectional visual computing: Foundations, challenges, and applications. *Computers & Graphics 113* (2023), 89–101. 1

[dSJ23] da Silveira T. L., Jung C. R.: Omnidirectional visual computing: Foundations, challenges, and applications. *Computers & Graphics 113* (2023), 89–101. 2

[FKTK20] Feick M., Kleer N., Tang A., Krüger A.: The virtual reality questionnaire toolkit. In *Adjunct proceedings of the 33rd annual ACM symposium on user interface software and technology* (2020), pp. 68–69. 3, 7

[GD13] Gurrieri L. E., Dubois E.: Stereoscopic cameras for the real-time acquisition of panoramic 3d images and videos. In *Stereoscopic Displays and Applications XXIV* (2013), vol. 8648, SPIE, pp. 559–575. 2

[HK18] Hedman P., Kopf J.: Instant 3d photography. *ACM Trans. Graph. 37*, 4 (jul 2018). 2

[HZZ*24] Huang K., Zhang F.-L., Zhao J., Li Y., Dodgson N.: 360° stereo image composition with depth adaption. *IEEE Transactions on Visualization and Computer Graphics 30*, 9 (2024), 6177–6191. 2

[KSSR24] Küntzer L., Schwab S. U., Spaderna H., Rock G.: Rover: A standalone overlay tool for questionnaires in virtual reality. In *Companion Proceedings of the 16th ACM SIGCHI Symposium on Engineering Interactive Computing Systems* (2024), pp. 31–39. 3

[LSR*20] Li Z., Shafiei M., Ramamoorthi R., Sunkavalli K., Chandraker M.: Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020). 2

[LXM*20] Lin K.-E., Xu Z., Mildenhall B., Srinivasan P. P., Hold-Geoffroy Y., DiVerdi S., Sun Q., Sunkavalli K., Ramamoorthi R.: Deep multi depth panoramas for view synthesis. In *Computer Vision – ECCV 2020* (Cham, 2020), Vedaldi A., Bischof H., Brox T., Frahm J.-M., (Eds.), Springer International Publishing, pp. 328–344. 2

[MCE*17] Matzen K., Cohen M. F., Evans B., Kopf J., Szeliski R.: Low-cost 360 stereo photography and video capture. *ACM TOG 36*, 4 (2017), 148:1–148:12. 1

[MCMB24] Magalhães M., Coelho A., Melo M., Bessa M.: Measuring users' emotional responses in multisensory virtual reality: A systematic literature review. *Multimedia Tools and Applications 83*, 14 (2024), 43377–43417. 2

[PBAG23] Pintore G., Bettio F., Agus M., Gobbetti E.: Deep scene synthesis of Atlanta-world interiors from a single omnidirectional image. *IEEE TVCG 29* (November 2023). 2, 4

[PBEP01] Peleg S., Ben-Ezra M., Pritch Y.: Omnistereo: Panoramic stereo imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence 23*, 3 (2001), 279–290. 1

[PGGS16] Pintore G., Ganovelli F., Gobbetti E., Scopigno R.: 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* (October 2016), Springer, pp. 130–145. 1

[PJVH*23] Pintore G., Jaspe-Villanueva A., Hadwiger M., Gobbetti E., Schneider J., Agus M.: Panoverse: automatic generation of stereoscopic environments from single indoor panoramic images for metaverse applications. In *Proceedings of the 28th International ACM Conference on 3D Web Technology* (2023), pp. 1–10. 2, 3, 4, 6

[PJVH*24] Pintore G., Jaspe-Villanueva A., Hadwiger M., Schneider J., Agus M., Marton F., Bettio F., Gobbetti E.: Deep synthesis and exploration of omnidirectional stereoscopic environments from a single surround-view panoramic image. *Computers & Graphics* (2024). 2, 3, 4, 6, 8

[PZL24] Pu G., Zhao Y., Lian Z.: Pano2room: Novel view synthesis from a single indoor panorama. *arXiv e-prints* (2024), arXiv–2408. 8

[QJY*20] Qi Y., Jiang G., Yu M., Zhang Y., Ho Y.-S.: Viewport perception based blind stereoscopic omnidirectional image quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology 31*, 10 (2020), 3926–3941. 2, 3

[SHKP21] Safikhani S., Holly M., Kainz A., Pirker J.: The influence of in-vr questionnaire design on the user experience. In *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology* (2021), pp. 1–8. 3

[SKC*19] Serrano A., Kim I., Chen Z., DiVerdi S., Gutierrez D., Hertzmann A., Masia B.: Motion parallax for 360° rgbd video. *IEEE Transactions on Visualization and Computer Graphics* (2019). 2

[TS20] Tucker R., Snavely N.: Single-view view synthesis with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020). 2

[WGD*22] Waidhofer J., Gadgil R., Dickson A., Zollmann S., Ventura J.: Panosynthvr: Toward light-weight 360-degree view synthesis from a single panoramic input. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (2022), pp. 584–592. 2

[WYW*18] Wang X., Yu K., Wu S., Gu J., Liu Y., Dong C., Qiao Y., Loy C. C.: Esrgan: Enhanced super-resolution generative adversarial networks. In *The European Conference on Computer Vision Workshops (ECCVW)* (September 2018). 4, 5

[XLZ*19] Xu J., Luo Z., Zhou W., Zhang W., Chen Z.: Quality assessment of stereoscopic 360-degree images from multi-viewports. In *2019 Picture Coding Symposium (PCS)* (2019), pp. 1–5. 2

[YJJ*23] You J., Jiang G., Jiang H., Xug H., Jiang Z., Zhu Z., Yu M.: Visual perception-oriented quality assessment for high dynamic range stereoscopic omnidirectional video system. *Displays 80* (2023), 102515. 3

[ZTF*18] Zhou T., Tucker R., Flynn J., Fyffe G., Snavely N.: Stereo magnification: Learning view synthesis using multiplane images. *ACM Trans. Graph. 37*, 4 (jul 2018). 2

[ZW24] Zhou W., Wang Z.: Perceptual depth quality assessment of stereoscopic omnidirectional images. *IEEE Transactions on Circuits and Systems for Video Technology* (2024). 2, 3

[ZZL*20] Zheng J., Zhang J., Li J., Tang R., Gao S., Zhou Z.: Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16* (2020), Springer, pp. 519–535. 5

[ZZZZ23] Zhang F., Zhao J., Zhang Y., Zollmann S.: A survey on 360 images and videos in mixed reality: algorithms and applications. *Journal of Computer Science and Technology 38*, 3 (2023), 473–491. 2