PanoVerse: automatic generation of stereoscopic environments from single indoor panoramic images for Metaverse applications

Giovanni Pintore*
gianni.pintore@crs4.it
CRS4
Cagliari, Italy

Alberto Jaspe-Villanueva*
alberto.jaspe@kaust.edu.sa
VCC, KAUST
Thuwal, Saudi Arabia

Markus Hadwiger
markus.hadwiger@kaust.edu.sa
VCC, KAUST
Thuwal, Saudi Arabia

Enrico Gobbetti
enrico.gobbetti@crs4.it
CRS4
Cagliari, Italy

Jens Schneider
jeschneider@hbku.edu.qa
College of Science and Engineering
Hamad Bin Khalifa University
Doha, Qatar

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

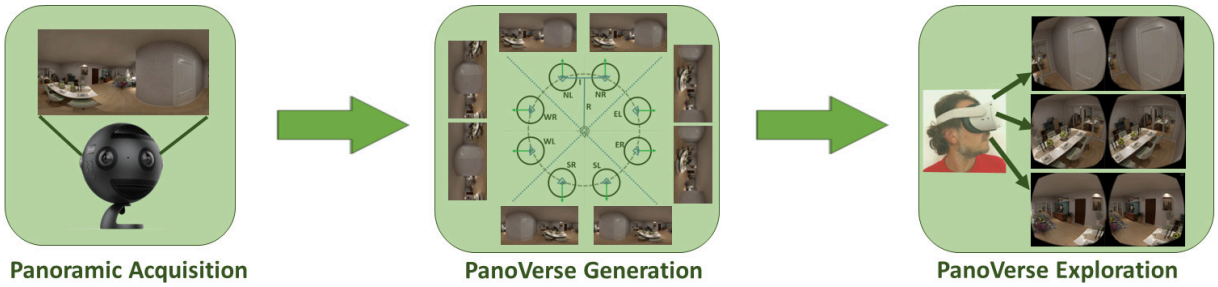

Figure 1: PanoVerse. We present a framework for generation and exploration of immersive scenes representing indoor environments. Starting from single panoramic images (left), we generate through a data-driven architecture stereo couples covering the viewing workspace (middle), that can be explored by VR setups on lightweight WebXR viewers ready for Metaverse applications (right).

ABSTRACT

We present 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. To realize this system, we introduce several novel techniques that combine and extend state-of-the art data-driven techniques. In particular, we present a gated architecture for panoramic monocular depth estimation and, starting from the re-projection of visible pixels based on predicted depth, we exploit the same gated architecture for inpainting the occluded and disoccluded areas, introducing a mixed GAN with self-supervised loss to evaluate the stereoscopic consistency of the generated images. At interactive rates, we interpolate precomputed panoramas to produce photorealistic stereoscopic views in a lightweight WebXR viewer. The system works on a variety of available VR headsets and can serve as a base component for Metaverse applications. We demonstrate our technology on several indoor scenes from publicly available data.

CCS CONCEPTS

• **Human-centered computing**; • **Computing methodologies**; • **Information systems**;

KEYWORDS

Indoor Environments, Omnidirectional Images, Data-driven Methods, Immersive Stereoscopic Exploration, Metaverse Applications, WebXR

*Joint first authors.

Web3D '23, October 9–11, 2023, San Sebastian, Spain

ACM ISBN 979-8-4007-0324-9/23/10.
https://doi.org/10.1145/3611314.3615914

ACM Reference Format:
Giovanni Pintore, Alberto Jaspe-Villanueva, Markus Hadwiger, Enrico Gobbetti, Jens Schneider, and Marco Agus. 2023. PanoVerse: automatic generation of stereoscopic environments from single indoor panoramic images for Metaverse applications. In *The 28th International ACM Conference on 3D Web Technology (Web3D '23), October 9–11, 2023, San Sebastian, Spain.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3611314.3615914

1 INTRODUCTION

In the context of digital content creation and immersive experiences, omnidirectional panoramic images have emerged as a powerful tool for constructing realistic and captivating indoor environments. A panoramic image can be created by stitching together multiple photographs or in single shots through specialized cameras (e.g., Ricoh Theta, LadyBug, or Insta360) and contains the entire scene context visible from a viewpoint within a 360° field of view. When presented through a Head-Mounted Display (HMD), the viewer dynamically explores it by focusing on the desired content via natural head movements, just like in the real world, leading to a natural VR interface with a good degree of immersion [Xu et al. 2020].

In this context, panoramas have become one the main exploration modes of real-world scenes in VR [Matzen et al. 2017]. In particular, single panoramic images can be easily shared and accessed across various devices and platforms, making them highly versatile and accessible. They can be seamlessly integrated into websites, virtual reality applications, or mobile devices, allowing a broad audience to experience and interact with indoor environments regardless of their location or the equipment they possess. As a representation of the environment around the user, they also have the potential to be an important building block for the construction of the shared physical and digital realities popularized by the Metaverse concept [Dong and Lee 2022].

Even though capturing a single shot panorama is a very appealing way to create a virtual clone of a real environment, the limitation to viewer to rotating 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 [Waidhofer et al. 2022]. 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. Capturing those views would require complex setups, such as rotating stereo rigs, that are incompatible with the need of quickly capturing, experiencing and sharing a 360° scene using consumer hardware. For this reason, research has concentrated on view synthesis methods that, however, either require complicated representations or are too heavy to run directly on HMDs and interactive rates (Sec. 2).

To overcome these limitations, we propose in this paper a novel framework, dubbed **PanoVerse** for fast and effective automatic generation and exploration of stereoscopic immersive scenes from single panoramic images. In our approach, we split the problem in two phases. First, once per shot, we infer, through a novel data-driven solution, a fixed set of panoramic stereo pairs distributed around the current central view-point. Then, once per frame, directly on the HMD, we rapidly fuse the closest sampled panoramas based on the current view direction to seamlessly cover the exploration workspace. Our main contributions are the following:

- we introduce a novel architecture that generates shifted views of an indoor panoramic image; the end-to-end architecture first estimates a depth map from a single panoramic input, and then generates views by reprojection and inpainting. Unlike other state-of-the-art approaches in the literature [Pintore et al. 2021; Sun et al. 2021], the network is based on a lightweight gated architecture, to ensure scalability and possibility of deployment on commodity hardware, and a dilated bottleneck, to maintain 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 for novel view synthesis of a specific photometric loss, combined with a GAN approach. As a result, photorealistic novel views for both right and left eye are generated with low latency. We moreover use super-resolution GAN-based architectures to further increase the resolution between the stereo images [Wang et al. 2018].
- we achieve real-time stereoscopic exploration by transferring, once per scene, a set of precomputed stereo views distributed around the central viewpoint, and producing, once per frame, seamless stereo couples that respond to head motion with low-latency and high frequency. The system is integrated in WebXR and achieves real-time performance on current HMD displays.

We evaluate the components of the architecture for generating stereo couples, and we show how depth inference and inpainting networks achieve state-of-the-art performance. Moreover, we demonstrate our technology on a variety of indoor high-resolution scenes, and we test it on different head-mounted displays, ranging from Meta Quest to HTC Vive and Google Cardboard. The proposed framework is easy to integrate in current panoramic viewers, just replacing the current monoscopic renderers, and is intended to work as a practical building block for deliver engaging and realistic experiences that captivate audiences and enable them to virtually explore and interact with indoor spaces in current and future Metaverse applications.

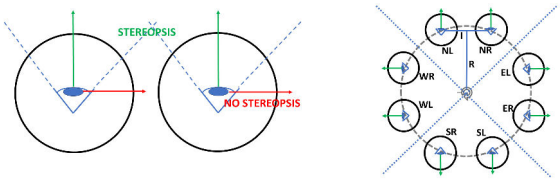

Figure 2: Left: panoramic stereo couples work correctly for the directions orthogonal to the baseline, while don't provide correct stereopsis for directions parallel to the baseline. Right: PanoVerse generates seamless stereo couples covering the four quadrants of the approximated viewing workspace.

2 RELATED WORK

Our work deals with the generation of immersive content from a single panoramic image of an interior environment. In the following, we briefly review the most closely related work, referring the reader to recent surveys on indoor reconstruction [Pintore et al. 2020], panoramic imaging and applications for scene understanding [Gao et al. 2022], and 3D geometry extraction from 360° imagery [da Silveira et al. 2022] for a wider coverage.

Given a single panoramic image, the classical solution for presenting it on a VR headset consists of projecting the image on a spherical dome centered around the user's head, eventually taking into account the eye position to generate each eye's perspective. Since all scene points are at the same position, given by the dome radius, parallax effects are limited. For this reason, view synthesis must take into account the geometry of the scene. Since visibility of scene elements may change even for small shifts of the eye position, this requires not only geometry estimation, but also the handling of occlusions and disocclusions.

Depth estimation from monocular input is increasingly focusing on data-driven solutions that derive hidden relations from large amounts of examples, while applying priors that fit specific use cases, in particular interior environments [Pintore et al. 2020]. Since it has been shown that directly applying perspective methods to 360° depth estimation in indoor environments produces suboptimal results [Zioulis et al. 2018], current research explicitly exploit the wide geometric context present in omnidirectional images, while also handling wrap-arounds and distortions present in equirectangular projections [Coors et al. 2018; Martin et al. 2020; Pintore et al. 2021; Rey-Area et al. 2022; Su and Grauman 2017; Tateno et al. 2018; Zioulis et al. 2018]. In this work, we follow this trend, proposing a simple light-weight pipeline that shares the same architecture as the view synthesis network.

The recovered depth map can then be exploited for view synthesis in various ways, that range from simply performing interpolation through the rendering and infilling of point clouds [Huang et al. 2017] or the generation and rendering of view-independent meshes from depth maps [Tukur et al. 2022] to the integration and blending of depth maps or generated meshes with multiple images or signals [Bertel et al. 2020; Luo et al. 2018]. Recently, end-to-end view synthesis networks have been proposed to generate shifted panoramic views at run time [Pintore et al. 2023; Xu et al. 2021]. These networks have proven to be able to infer compelling views in a small neighborhood around the viewer (e.g., 50cm), but are too demanding to be run directly on the embedded platforms. For this reason, HMDs are supported only through remote rendering [Pintore et al. 2023]. In this work, we also employ an end-to-end reprojection and synthesis network, but only to generate a small set of stereo pairs once per panoramic view, without stringent frequency requirements. Real-time rendering is then performed on the HMD starting from these inferred views through a simple interpolation method. Since we limit per-frame generation to stereo pairs, our networks are much simpler than general prior solutions for free-viewpoint synthesis [Pintore et al. 2023; Xu et al. 2021].

Interpolating images from different viewpoints to generate novel views has been widely researched, and effective solutions have been proposed, even without the support of a prior depth estimation step [Reda et al. 2022; Trinidad et al. 2019]. End-to-end networks that perform this task have, however, the same computational constraints of depth estimation and view synthesis networks, and cannot be executed at interactive rates on HMDs.

A modern trend for fast novel viewpoint synthesis from scenes consists of considering layered depth representations, in which each pixel is associated to multiple depth values. These layered representations are used for view synthesis through extrapolation and in-painting in a way to fill holes [Hedman and Kopf 2018]. This approach has been successfully extended to work with single panoramic images [Serrano et al. 2019]. Moreover, different layered representations have been considered in a way to generate more accurate results: Broxton et al. [Broxton et al. 2020] create light field videos through layered mesh representations, while Lin et al. [Lin et al. 2020] propose multi-depth panorama. An alternative version of layered depth representations consists of using multiple flat planes at fixed depth in a way to capture a multi-plane image (MPI), that can be used together with convolutional neural networks [Tucker and Snavely 2020a; Zhou et al. 2018]. 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 [Li and Khademi Kalantari 2020]. The concept of capturing the scene at multiple fixed depths has been extended for panoramic imaging by considering different capturing proxies like multi-spherical images (MSI) [Attal et al. 2020] or multi-cylinder images (MCI) [Waidhofer et al. 2022]. Our proposed framework uses, instead, a discrete set of stereo panoramic couples covering the exploration workspace spanned during standard head movements, that we can seamlessly blend during immersive exploration with VR setups. Compared to the current state of the art, our system has the advantage of being lightweight both for inference of novel omnidirectional stereo couples and for immersive exploration through WebXR viewers.

3 METHODS

The proposed PanoVerse system is composed by two main components (see a schematic depiction in Fig. 1):

- a data-driven component for automatic generation of multiple discrete stereo couples seamlessly covering the visible scene around the observer (Sec. 3.1). In order to provide a compact and lightweight architecture, we exploit a gated network architecture in which a baseline is shared between depth estimation and image synthesis. In Sec. 3.2 and Sec. 3.3 we explain such architecture and the different specializations respectively, for depth and view synthesis tasks;
- a WebXR rendering component presenting the immersive scene during real-time exploration, selecting the closest stereo couples according tu the current gaze direction and blending them to generate per-eye images (Sec. 3.4).

3.1 Stereo couples generation

Even if an omnidirectional camera can capture an entire 360 field-of-view of the scene, it is well known that, for two images taken by an omnidirectional camera, only some parts of the images can be used for the stereoscopic pair [Vanijja et al. 2006]. Specifically, two panoramic images can be viewed as a stereo pair in a perpendicular direction to the line connecting the two viewpoints, while they will fail to give stereo perception when viewed in the direction of the line connecting the two viewpoints since images from the two cameras are behind each other (Fig. 2 left). To overcome this issue, multiple stereo couples need to be generated to cover the viewing workspace. In our system, we consider that during the standard head motion for exploration of 360 panoramic images, we can approximate the trajectory of two eyes as a circle centered at the center of the head. These circular trajectories can be sampled to

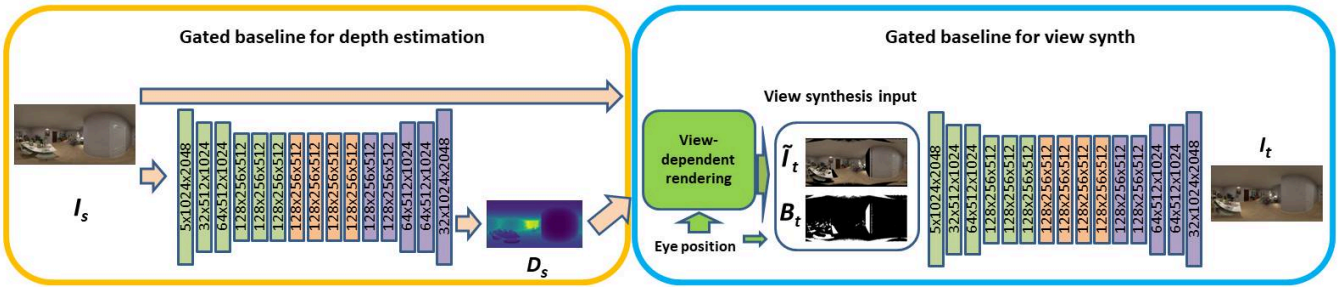

Figure 3: PanoVerse generation. A gated architecture is used to predict depth (left), as well as to synthetize a novel rgb view (right). The encoder-decoder scheme follows a design commonly adopted for image inpainting, exploiting dilated convolutions as bottleneck, and gated convolutions for encoding and decoding.

individuate discrete feasible points of view that can be used during the real-time users head rotation and provide correct stereopsis for a limited portion of the immersive workspace (Fig. 2 right).

In the proposed PanoVerse system, to demonstrate the concept in the most challenging situation, we consider four stereo couples mapping to the four quadrants (North, East, West, South), each of them with a field of view of 90 degrees. Additional stereo couples can be added for increasing granularity or, eventually, also for taking into account limited 6-DOF head motion (see Sec. 5).

For the generation of the stereo poses, we consider the following parameters: the intra-ocular distance I and the distance R between the head center and the eye axis (Fig. 2 right). A panoptic representation for blending between adjacent couples is used on the fly during rendering for taking into account the transition areas between the quadrants and computing convincing stereo images to be displayed on VR headsets.

3.2 Single panorama depth estimation

The fundamental task of enabling new view estimation and 3D navigation of an immersive scene from a single image is depth estimation. From the operational point of view, depth is essential to establish the 3D position of visible points in space, and their novel position in case of viewpoint change, which is just verified in the case of stereo image generation. Several methods exist in the literature for estimating depth from a single panoramic image, within particular deep-learning approaches achieving impressive levels of accuracy [Pintore et al. 2021; Rey-Area et al. 2022]. However, most methods pay little attention to the computational complexity and, in particular, the latency times required to achieve the prediction. In the case of web and MetaVerse applications, these aspects become important, in particular to allow for the execution of the code on low-end machines and to reduce the time required from capture to exploration start. To this end, in this work we designed a network for depth prediction that was an efficient compromise between accuracy and cost. To predict depth, as well as to synthesize a novel RGB view, as described in Sec. 3.3, we design the gated architecture illustrated in Fig. 3. The encoder-decoder scheme follows a design commonly adopted for image inpainting [Iizuka et al. 2017], thus exploiting dilated convolutions as bottleneck [Yu and Koltun 2016], and gated convolutions for encoding and decoding [Yu et al. 2018]. Compared to the baseline [Iizuka et al. 2017; Yu et al. 2018], our design has fewer parameters, with a lighter 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 [Gkitsas et al. 2021].

The input of the depth estimation network consists of a single equirectangular image, which is encoded through a sequence of light-weight gated convolutions having different strides (the six layers in green in Fig. 3), so that the original size is reduced by a factor of four in each direction. Each encoding convolution is followed by instance normalization [Ulyanov et al. 2016] and ReLU activation.

It should be noted that, 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 (Sec. 3.3).

We adopt a gated convolution (GC) approach [Yu et al. 2019], expressed as:

$$\begin{aligned} G &= conv(W_g, I) \\ F &= conv(W_f, I) \\ O &= \sigma(G) \odot \psi(F) \end{aligned} \tag{1}$$

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 [Yi et al. 2020]. Specifically, we decompose G from Equation 1 into a depth-wise convolution [Yi et al. 2020] (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 [Yu and Koltun 2016] are used for the bottleneck (Fig. 3, orange 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 1), but with some differences.

It is expressed as:

$$D_{y,x} = \sigma(b + \sum_{i=-k'_h}^{k'_h} \sum_{i=-k'_w}^{k'_w} W_{k'_h+i,k'_w+j} \cdot I_{y+\eta i,x+\eta j}) \quad (2)$$

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, 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 [Iizuka et al. 2017]. The network decoder (5 blue layers in Fig. 3) 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. It should be noted that such baseline is the same for both the depth and RGB task, and differs, besides the input, only in the final activation function, which is respectively *ELU* for a depth output and *tanh* for the RGB output (i.e, following original gated convolution approach[Yu et al. 2019]). Indeed, the use of such a versatile baseline depends on its training. In our approach, we adopt as the loss function for the depth prediction task the robust *Adaptive Reverse Huber Loss (BerHu)* [Lambert-Lacroix and Zwald 2016], combined with a Structural Similarity Index Measure (SSIM), which measures the preservation of highly structured signals with strong neighborhood dependencies. On the other hand, we illustrate in Sec. 3.3 the specific training techniques for generating photorealistic images with parallax. As a result, such a panoramic depth prediction approach returns accurate depth maps for the input pose, as demonstrated by our results at Sec. 4.

3.3 Novel view-synthesis

The goal of this task is to compute a novel photorealistic spherical image from a translated position. In our case, we assume as the origin of the coordinates system the position of the input image, and generate novel views for the left and right eye (Fig. 2). Without loss of generality, in this paper we generate 4 couples of right-left images will be generated, respectively for a North, East, South, and West oriented panoramic view, heading in 4 directions spaced each other $90degrees$. The resulting 8 spherical images, represent a sampling of the whole 360 horizon around the ideal head's center, and will be combined in a unique pan-optic immersive environment as described in Sec. 3.4. No particular change has to be made to the architecture or rendering code to support a finer angular sampling.

In this section, we will focus on the details of generating a single pair of new stereo panoramic images from a single one as input. In order to generate other views, the same procedure is generated by simply re-executing the code with horizontally rotated input images (e.g., by 90 degrees steps for the example discussed in this paper). Assuming the interpupillar distance I (58mm in our experiments) and the eye-axis to head center distance R (100mm in our case), two images (i.e., right and left) are generated for each input panorama, at the positions $(\mp I/2, R)$ (Fig. 2 right).

Considering one of the two possible translations T, a novel image is obtained by applying a translation T to the camera. View-synthesis includes then two cascading steps: a rendering step, which exploits the predicted 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, inpainting disoccluded parts. Such a network takes as input the translated pixels $\widetilde{I_t}$ and the disocclusion mask B_t, returning as output the novel view I_t.

For the rendering step, many view-synthesis methods adopt a differential-rendering approach, such as a soft z-buffer [Tulsiani et al. 2018]. However, in our case, pixel rendering is not part of the learnable layers by design, so we can directly project visible points according to D_s and T. Direct projection is more accurate than soft z-buffer, and, if no splatting is performed, produces sharp images while leaving several pixels in the new view to be inpainted. This direct solution is better suited to our case with respect to more elaborate spatting methods, since for stereo rendering the distance of a few centimeters with respect to the center generates munch 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 linear mapping between image domain pixels and spherical domain angular coordinates allows for direct transitions between image and spherical-based operations [Zioulis et al. 2019]. Omitting the straightforward relationship between Cartesian and spherical coordinates, the following equation relates spatial (i.e, T) with angular displacements (i.e., $\widetilde{I_t}$ pixels):

$$\begin{bmatrix} \partial d \\ \partial \phi \\ \partial \theta \end{bmatrix} = \begin{bmatrix} \sin(\phi)\sin(\theta) & \cos(\theta) & \cos(\phi)\sin(\theta) \\ \frac{\cos(\phi)}{d\sin(\theta)} & 0 & \frac{-\sin(\phi)}{d\sin(\theta)} \\ \frac{\sin(\phi)\cos(\theta)}{d} & \frac{-\sin(\theta)}{d} & \frac{\cos(\phi)\cos(\theta)}{d} \end{bmatrix} \begin{bmatrix} \partial x \\ \partial y \\ \partial z \end{bmatrix} \quad (3)$$

where d is the depth of the given pixel. According to this mapping rgb values from the source image I_s are scattered to the target image $\widetilde{I_t}$ (Fig. 3).

We assume then $\widetilde{I_t}^{3\times h\times w}$ as the input to the view synthesis network. 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. B_t is then concatenated to $\widetilde{I_t}$ (i.e., along the batch dimension - 4 layers input). To process such input, we adopt the same architecture of Fig. 3, but with a different final activation function (i.e., $tanh$) and training strategy. Then, we train inpainting network by combining visual terms and geometric-consistency terms: $\mathcal{L}_{vs} = \mathcal{L}_{vis} + \mathcal{L}_{gc}$.

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

$$\mathcal{L}_{vis} = \lambda_{px}\mathcal{L}_{px} + \lambda_{perc}\mathcal{L}_{perc} + \lambda_{style}\mathcal{L}_{style} + \lambda_{adv}\mathcal{L}_{adv}. \quad (4)$$

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 [Gatys et al.

2016], enforcing I_{out} and I_{gt} to have a similar representation in the feature space as computed by a pre-trained $VGG-19$ [Simonyan and Zisserman 2014], while $\mathcal{L}_{adv}$ is a discriminator-based loss (i.e., PatchGAN [Isola et al. 2017]). λ weights are common for many single pose inpainting problems [Yu et al. 2018]: $\lambda_{px} = 1.0$, $\lambda_{style} = 100.0$, $\lambda_{perc} = 1.0$, $\lambda_{adv} = 0.2$.

$\mathcal{L}_{gc}$, instead, is a photometric loss, which combines the $L1$ penalty function with structural dissimilarity L_d [Godard et al. 2017], under a relative weighting factor η (i.e., $\eta = 0.85$ in our experiments). The superscript B denotes multiplication with the binary mask B_t:

$$\mathcal{L}_{gc} = \eta \mathcal{L}_{\lceil}(I^B{}_s, I^B{}_{ts}) + (1-\eta)\left|I^B{}_s - I^B{}_{ts}\right| \tag{5}$$

where I_s is the image at the source position, I_{ts} is the novel image at the target position projected back to I_s position using D_s, with both images masked by B_t.

For increasing the resolution of the immersive scene without the need to train and execute the full view synthesis network on large images, we use a pretrained GAN-based super-resolution network [Wang et al. 2018] at 4x magnification in a way that original stereo couples are converted to $4096x2048$ resolution, that is adequate for limiting artifacts on VR headsets.

3.4 Stereoscopic scene rendering

Figure 4: The panoptic representation is obtained by blending the stereo couples computed for the various workspace quadrants.

The immersive rendering component consists of a viewer on top of a panoptic representation built by integrating the stereo panoramic couples representing the quadrants of the viewing workspace.

During the exploration, the view position is extracted by the VR headset and used for sampling the panoptic environment through blending operation (Fig. 4). Specifically, we use the rotation angle α along vertical direction for computing an interpolation factor between two adjacent stereo couples according to the following equation:

$$I_{CE} = \tau I_{SE} + (1-\tau) I_{DE}, \tag{6}$$

where τ is a blending factor depending on the angle between two adjacent views, C is the computed image, S, D are the source and destination directions (one between North, East, South and West), and E is the eye for which the image is computed (Left or Right). For what concerns the blending factor, we designed a smooth function depending on a transition window that is able to provide a trade-off

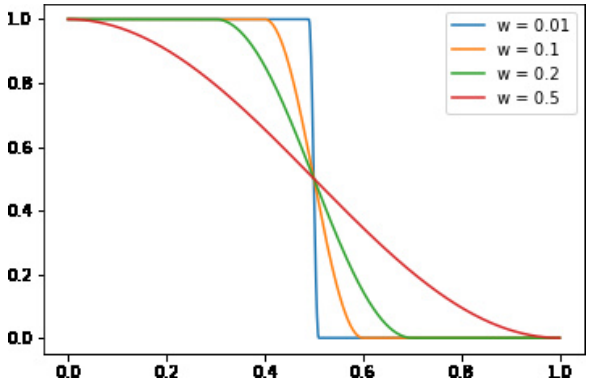

Figure 5: Blending factor. For computing the panoptic blended images we consider a blending function for smoothing the transition between adjacent views, depending on a window parameter w.

between the need of reducing the misalignment artifacts due to the discretization of the viewing workspace and the need of maximizing the stereo perception. Given a normalized pixel distance x between two adjacent views and a percent window w, the blending factor is computed as follows:

$$\tau(x) = \begin{cases} 1 & x \leq (\frac{1}{2} - w) \\ \frac{1}{2}(1 + \cos(\pi \frac{x + w - \frac{1}{2}}{2w})) & (\frac{1}{2} - w) < x < (\frac{1}{2} + w) \\ 0 & x \geq (\frac{1}{2} + w) \end{cases} \tag{7}$$

Fig. 5 shows various blending functions according to the window parameter w, while Fig. 6 shows different blended panoptic images generated with different parameters w (on the left $w = 0.01$, on the middle $w = 0.1$, and on the right $w = 0.5$). Higher parameter provides a softer transition between adjacent views at the cost of reduction of stereopsis cue and an increase in ghosting artifacts. On the other side, lower w provides abrupt change between adjacent views. The panoptic blended images computed for the left eye and the right eye are then used as textures on a spherical dome representing the immersive environment. By increasing the angular sampling rate from the minimum of 90 degrees as done in this paper, we can increase quality at the cost of data size, but without increase in rendering times, since only the two nearest images are blended together.

4 RESULTS

The PanoVerse system has been developed in Python using PyTorch library for what concerns the automatic generation component, and in WebGL and WebXR using Three.js library for what concerns the immersive rendering component.

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 [Zheng et al. 2020]), a large-scale (21K photorealistic scenes), synthetic database of indoor scenes providing for each panoramic image the ground truth depth. To train and test view synthesis, instead, we exploit PNVS[Xu et al. 2021], a subset of Structured3D scenes providing, for each source panoramic image,

Figure 6: Smoothness trade-off for panoptic blends. Examples of generated panoptic blended images for different window parameters ($w = 0.01$ on the top, $w = 0.1$ on the middle, $w = 0.5$ on the right).

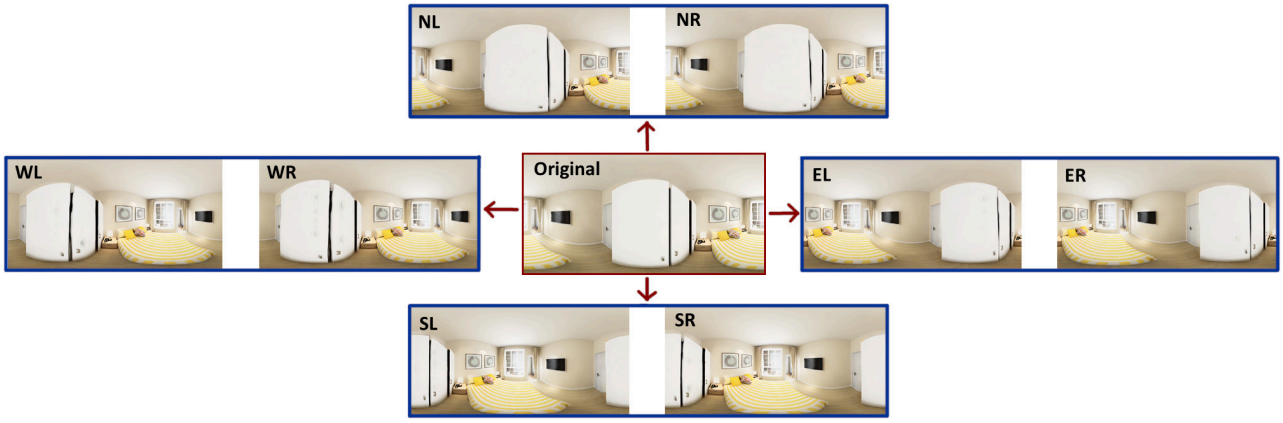

Figure 7: Example of synthesized panoramic stereoscopic views for the main quadrant directions (North, East, South and West).

three translated view. Since the baseline for stereo view is very small, we choose the PNVS subset named *easy* (i.e., maximum 0.2-0.3m range).

Computational performance. Our depth estimation and view synthesis baseline is extremely lightweight. Tab. 1 shows learnable parameters, GFlops and milliseconds to process a single image-frame at different input resolution. Although the expected output scene is static, it should be noted how the time to generate a novel view is very low, reducing the latency from image acquisition to start of exploration. For the low-density sampling used in this paper, a full novel scene requires 8 novel views and 1 depth. Finer resolution can be achieved by just increasing the number of novel views.

Table 1: Computational performance. We show our computational performance and latency time for different input resolutions. Our results demonstrate how we diminish images with very low latency even when resolution increase.

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

Table 2: View synthesis performance: we show view synthesis quantitative performance compared to other state-of-the-art works which works at a minimum resolution of 1024×512.

Method	PSNR↑	SSIM↑	LPIPS↓
SynSin [Wiles et al. 2020]	17.28	0.721	0.226
MPI [Tucker and Snavely 2020b]	17.59	0.725	0.223
Our	**21.57**	**0.731**	**0.205**

Table 3: Depth quantitative performance: we show depth quantitative performance compared to other state-of-the-art works.

Method	MAE↓	MSE↓	RMSE↑	δ_1 ↑
HoHoNet [Sun et al. 2021]	0.081	0.065	0.206	0.958
SliceNet [Pintore et al. 2021]	0.082	0.054	0.198	0.961
Our	0.061	0.008	0.038	0.962

Tab. 2 summarizes our performance in terms of view synthesis accuracy, benchmarked on PNVS [Xu et al. 2021]. We also compared our performance to the one achieved by state-of-the-art methods [Tucker and Snavely 2020b; Wiles et al. 2020] which works at a minimum resolution of 1024×512. Tab. 3 right summarizes our performance on depth estimation, using as benchmark Structured3D [Zheng et al. 2020]. In this case, despited the fact that our

Figure 8: Examples of synthesized panoptic stereoscopic views for scenes extracted from Structured3D dataset [Zheng et al. 2020]. On the left the panoptic image for the left eye, on the right the panoptic image for the right eye.

method has a low computational complexity, we achieve performance comparable with other state-of-the-art works.

Visual assessment. We tested the immersive viewer on a Meta Quest 2 attached to a Razer Blade 15 laptop and on an Android mobile device Samsung Galaxy S 22 with a Google Cardboard. The viewer runs on Google Chrome 114 web browser with WebGL 2 and WebXR enabled. Fig. 7 shows an example of the stereoscopic viewpoints generated by our proposed architecture. Fig. 8 shows some examples of the panoptic representations computed by blending the stereo couples representing the four workspace quadrants for both eyes. The scenes are extracted from the public domain dataset Structured3D [Zheng et al. 2020]. From the stereo panoptic

Figure 9: The WebXR viewer enables immersive exploration of the indoor scene around the user on a variety of devices (Android mobile device with Google cardboard in this case).

representations, stereo images are generated in real time and displayed on a variety of devices (Fig. 9), ranging from Oculus Quest to Google Cardboard on mobile devices. We carried out a preliminary qualitative user study to assess the capabilities of the system to provide immersive stereo cues: five subjects were requested to explore the stereoscopic environment on Oculus Quest and provide their opinion on immersion and stereoscopic perception of the generated scenes. All of them were able to perceive stereo cues for the whole 360 viewing workspace, confirming that the system is able to generate an immersive experience. On the other hand, according to the scene and the position, especially in the transition areas, users perceived misalignment artifacts, due to the limited number of views from which the panoptic representation is computed. In the future, we plan to extend the user study in a way to quantify the effects of the discretization and of the blending factors, and to compare with other ways to compute stereoscopic panoramic scenes.

5 CONCLUSIONS AND FUTURE WORK

We presented a framework for automatic generation of stereoscopic indoor environments to be used in immersive Metaverse applications. Our method starts from single panoramic image of an interior environment, and uses data-driven architectures for depth estimation and novel view synthesis to quickly generate a set stereo couples sampling the viewing workspace. At run-time, these precomputed stereo couples are delivered to a lightweight WebXR viewer that supports stereocopic exploration by synthesizing views that respond to head rotations through a composition of the precomputed images. The preliminary results show that the automatic generation components achieve state of the art accuracy and the visualization component is able to provide immersive experience to casual users on a variety of devices, even in the very demanding case of using only four precomputed stereo pairs. As future work, we plan to investigate the trade-off between the number of precomputed images and the quality of experience in a variety of settings. We also plan to further improve immersivity by improving blending also taking into account a new synthesized depth. We will also evaluate the possibility of exploiting this approach to also support a limited amount of horizontal and vertical head motion. Finally, we plan to use our panoramic capture and immersive rendering system as a building block for the construction of applications that perform actions in shared physical and digital realities, as popularized by the Metaverse concept.

ACKNOWLEDGMENTS

This publication was made possible by NPRP-Standard (NPRP-S) 14th Cycle grant 0403-210132 AIN2 from the Qatar National Research Fund (a member of Qatar Foundation). The findings herein reflect the work, and are solely the responsibility, of the authors.

REFERENCES

Benjamin Attal, Selena Ling, Aaron Gokaslan, Christian Richardt, and James Tompkin. 2020. 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.* Springer, 441–459.

Tobias Bertel, Mingze Yuan, Reuben Lindroos, and Christian Richardt. 2020. OmniPhotos: Casual 360° VR Photography. *ACM Transactions on Graphics* 39, 6 (Dec. 2020), 266:1–12. https://doi.org/10.1145/3414685.3417770

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

Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. 2018. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *Proceedings of the European conference on computer vision (ECCV).* 518–533.

Thiago L. T. da Silveira, Paulo G. L. Pinto, Jeffri Murrugarra-Llerena, and Cláudio R. Jung. 2022. 3D Scene Geometry Estimation from 360° Imagery: A Survey. *ACM Comput. Surv.* 55, 4, Article 68 (nov 2022), 39 pages. https://doi.org/10.1145/3519021

Haiwei Dong and Jeannie S. A. Lee. 2022. The Metaverse From a Multimedia Communications Perspective. *IEEE MultiMedia* 29, 4 (2022), 123–127. https://doi.org/10.1109/MMUL.2022.3217627

Shaohua Gao, Kailun Yang, Hao Shi, Kaiwei Wang, and Jian Bai. 2022. Review on Panoramic Imaging and Its Applications in Scene Understanding. *IEEE Transactions on Instrumentation and Measurement* 71 (2022), 1–34. https://doi.org/10.1109/TIM.2022.3216675

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proc. CVPR.* 2414–2423. https://doi.org/10.1109/CVPR.2016.265

Vasileios Gkitsas, Vladimiros Sterzentsenko, Nikolaos Zioulis, Georgios Albanis, and Dimitrios Zarpalas. 2021. PanoDR: Spherical Panorama Diminished Reality for Indoor Scenes. In *Proc. CVPR Workshops.* 3716–3726.

Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. 2017. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 270–279.

Peter Hedman and Johannes Kopf. 2018. Instant 3D Photography. *ACM Trans. Graph.* 37, 4, Article 101 (jul 2018), 12 pages. https://doi.org/10.1145/3197517.3201384

Jingwei Huang, Zhili Chen, Duygu Ceylan, and Hailin Jin. 2017. 6-DOF VR videos with a single 360-camera. In *2017 IEEE Virtual Reality (VR).* 37–44. https://doi.org/10.1109/VR.2017.7892229

Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2017. Globally and Locally Consistent Image Completion. *ACM Trans. Graph.* 36, 4, Article 107 (jul 2017), 14 pages. https://doi.org/10.1145/3072959.3073659

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proc. CVPR.* 1125–1134.

Sophie Lambert-Lacroix and Laurent Zwald. 2016. The adaptive BerHu penalty in robust regression. *Journal of Nonparametric Statistics* 28 (2016), 1–28.

Qinbo Li and Nima Khademi Kalantari. 2020. Synthesizing Light Field From a Single Image with Variable MPI and Two Network Fusion. *ACM Transactions on Graphics* 39, 6 (12 2020). https://doi.org/10.1145/3414685.3417785

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

Bicheng Luo, Feng Xu, Christian Richardt, and Jun-Hai Yong. 2018. Parallax360: Stereoscopic 360° Scene Representation for Head-Motion Parallax. *IEEE Transactions on Visualization and Computer Graphics* 24, 4 (2018), 1545–1553. https://doi.org/10.1109/TVCG.2018.2794071

Daniel Martin, Ana Serrano, and Belen Masia. 2020. Panoramic convolutions for 360 single-image saliency prediction. In *CVPR workshop on computer vision for augmented and virtual reality*, Vol. 2.

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

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

Giovanni Pintore, Fabio Bettio, Marco Agus, and Enrico Gobbetti. 2023. Deep scene synthesis of Atlanta-world interiors from a single omnidirectional image. *IEEE Transactions on Visualization and Computer Graphics* 29 (November 2023). Proc. ISMAR. To appear.

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

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

Manuel Rey-Area, Mingze Yuan, and Christian Richardt. 2022. 360MonoDepth: High-Resolution 360deg Monocular Depth Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3762–3772.

Ana Serrano, Incheol Kim, Zhili Chen, Stephen DiVerdi, Diego Gutierrez, Aaron Hertzmann, and Belen Masia. 2019. Motion parallax for 360° RGBD video. *IEEE Transactions on Visualization and Computer Graphics* (2019).

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

Yu-Chuan Su and Kristen Grauman. 2017. Learning Spherical Convolution for Fast Features from 360 Imagery. In *Advances in Neural Information Processing Systems 30*. 529–539.

Cheng Sun, Min Sun, and Hwann-Tzong Chen. 2021. HoHoNet: 360 Indoor Holistic Understanding with Latent Horizontal Features. In *Proc. CVPR*. 2573–2582.

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

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

Richard Tucker and Noah Snavely. 2020a. Single-View View Synthesis With Multiplane Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Richard Tucker and Noah Snavely. 2020b. Single-view view synthesis with multiplane images. In *Proc. CVPR*. 551–560.

M. Tukur, Giovanni Pintore, Enrico Gobbetti, Jens Schneider, and Marco Agus. 2022. SPIDER: Spherical Indoor Depth Renderer. In *Proc. Smart Tools and Applications in Graphics (STAG)*. 131–138. https://doi.org/10.2312/stag.20221267

Shubham Tulsiani, Richard Tucker, and Noah Snavely. 2018. Layer-structured 3d scene inference via view synthesis. In *Proc. ECCV*. 302–317.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2016. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022* (2016).

Vajirasak Vanijja, Susumu Horiguchi, et al. 2006. A stereoscopic image-based approach to virtual environment navigation. *The Computer, the Internet and Management* 14, 2 (2006), 68–81.

John Waidhofer, Richa Gadgil, Anthony Dickson, Stefanie Zollmann, and Jonathan Ventura. 2022. PanoSynthVR: Toward Light-weight 360-Degree View Synthesis from a Single Panoramic Input. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 584–592. https://doi.org/10.1109/ISMAR55827.2022.00075

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. 2018. ESRGAN: Enhanced super-resolution generative adversarial networks. In *The European Conference on Computer Vision Workshops (ECCVW)*.

Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. 2020. Synsin: End-to-end view synthesis from a single image. In *Proc. CVPR*. 7467–7477.

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

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

Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. 2020. Contextual residual aggregation for ultra high-resolution image inpainting. In *Proc. CVPR*. 7508–7517.

Fisher Yu and Vladlen Koltun. 2016. Multi-Scale Context Aggregation by Dilated Convolutions. In *Proc. ICLR*, Yoshua Bengio and Yann LeCun (Eds.).

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. 2018. Generative image inpainting with contextual attention. In *Proc. CVPR*. 5505–5514.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. 2019. Free-form image inpainting with gated convolution. In *Proc. ICCV*. 4471–4480.

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

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. 2018. Stereo Magnification: Learning View Synthesis Using Multiplane Images. *ACM Trans. Graph.* 37, 4, Article 65 (jul 2018), 12 pages. https://doi.org/10.1145/3197517.3201323

Nikolaos Zioulis, Antonis Karakottas, Dimitris Zarpalas, Federic Alvarez, and Petros Daras. 2019. Spherical View Synthesis for Self-Supervised 360° Depth Estimation. In *Proc. 3DV*. 690–699.

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