3D Pano Inpainting: Building a VR Environment from a Single Input Panorama

Shivam Asija* Edward Du† Nam Nguyen‡
California Polytechnic State University
San Luis Obispo, CA USA

Stefanie Zollmann§
University of Otago
Dunedin, New Zealand

Jonathan Ventura¶
California Polytechnic State University
San Luis Obispo, CA USA

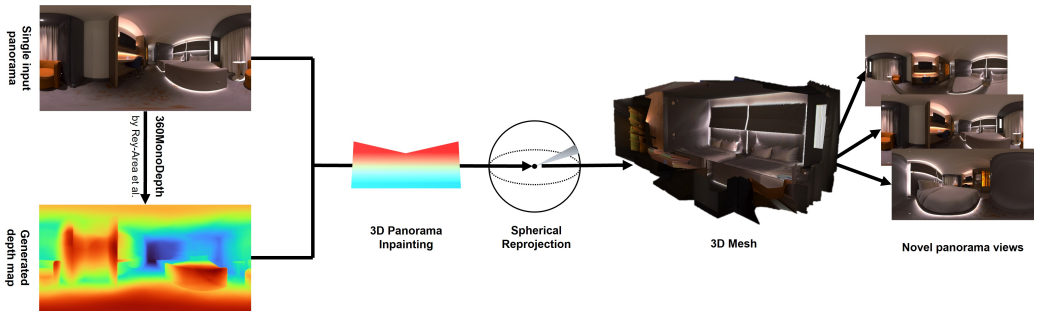

Figure 1: Our novel system takes as input a single panoramic image, generates a depth map [8], and applies an inpainting method [10] adapted for 360° content to produce a 3D textured mesh for real-time 6DOF view synthesis in a VR headset.

Abstract

Creating 360-degree 3D content is challenging because it requires either a multi-camera rig or a collection of many images taken from different perspectives. Our approach aims to generate a 360° VR scene from a single panoramic image using a learning-based inpainting method adapted for panoramic content. We introduce a pipeline capable of transforming an equirectangular panoramic RGB image into a complete 360° 3D virtual reality scene represented as a textured mesh, which is easily rendered on a VR headset using standard graphics rendering pipelines. We qualitatively evaluate our results on a synthetic dataset consisting of 360 panoramas in indoor scenes.

Index Terms: Computing methodologies—Computer graphics—Graphics systems and interfaces—Virtual reality Computing methodologies—Artificial intelligence—Computer vision

1 Introduction

The rise of consumer virtual reality (VR) headsets has led to increasing demand for easy methods for casual users to create immersive 360° 3D content 360° cameras provide the capability for conveniently capturing 360° images and videos by casual as well as professional content creators. While the 360° format allows users to explore the photograph in a VR headset by fully filling the users' field of view, they lack support for motion parallax during translational head movements, resulting in an unnatural experience that may disrupt immersion and induce discomfort or even nausea in some users.

To support a full 6 Degree-of-Freedom (6DoF) experience, earlier work used depth estimation to create 360° RGBD images [1, 3, 8]. However, simply using textured meshes built on 360° RGBD images falls short in delivering a truly immersive user experience, since in places where the depth map exhibits significant discontinuities, the mesh displays a stretching effect between the foreground and background. While one option is to disconnect the mesh in such areas, this approach introduces gaps that can become apparent when the user shifts their perspective.

An alternative is to predict a multi-cylinder image (MCI) representation using a convolutional neural network to support motion parallax [11]. The MCI representation uses soft blending between many cylindrical layers and thus enables inpainting behind occluded objects. However, this approach exhibits reduced effectiveness in perceptual metrics due to blurriness.

Neural Radiance Fields (NeRFs) achieve high-quality novel view synthesis from a collection of posed input images [5]. These methods usually require many input views to generate novel views [6], and the volumetric scene representation leads to slow rendering time and is not suitable for rendering in current VR headsets. In contrast, our method operates on a single panoramic input, which is easily obtained with even a smartphone, and produces a textured mesh representation that is compact enough to be rendered in a VR headset.

Using panoramas for view synthesis is popular because they allow for surround-view view synthesis, enabling an immersive experience in a VR headset. Lin et al. propose a method to perform view synthesis by leveraging Multi-Depth panoramas [4]. They require a multi-camera rig whereas we only limit our input to a single image. Serrano et al. [9] present a system that is similar in spirit to ours, but uses many hand-crafted and traditional image processing methods, whereas we use a more modern learning-based approach.

We introduce a method called 3D Pano Inpainting which achieves sharp and high-quality view synthesis by converting an input panorama to a textured mesh. We address occlusion artifacts by modifying a 3D inpainting method designed for perspective images [10] so that it can process 360 panoramic images. The resulting inpainted textured mesh can be easily rendered on a VR headset using standard graphics pipelines. A complete overview of our novel system is shown in Figure 1.

2 Methods

Depth Estimation The first step in our pipeline is to produce a depth map for the input panorama if it is not already available (e.g. from a

*e-mail: asija@calpoly.edu
†e-mail: eddu@calpoly.edu
‡e-mail: nnguy158@calpoly.edu
§e-mail: stefanie.zollmann@otago.ac.nz
¶e-mail: jventu09@calpoly.edu

(a) Before inpainting

(b) After inpainting

Figure 2: Example inpainting result from our system. (a) The black regions are gaps in the mesh caused by depth discontinuities. (b) The gaps have been filled with mesh geometry and inpainted.

(a) Ground Truth

(b) Mesh Without Inpainting

(c) Our Method

Figure 3: View synthesis comparison between (a) ground truth, (b) textured mesh without inpainting, and (c) our method.

stereo panorama camera). We apply the 360MonoDepth method by Rey-Area et al. [8] which projects the 360° input onto perspective tangent images, predicts depth maps for each tangent image using the MiDaS network [7], corrects misalignments within the spherical domain, and merges them back together to form a high-resolution spherical depth map.

Initial Mesh Formation We form the initial mesh by creating a sphere geometry with one vertex per pixel in the equirectangular panorama. We determine each vertex location using spherical coordinates: $X = d\sin(\theta)\cos(\phi), Y = d\sin(\phi), Z = d\cos(\theta)\cos(\phi)$ where θ, ϕ are the horizontal and vertical angles, respectively, of the pixel, and d is depth of the pixel according to the estimated depth map. We assign the vertex color from the corresponding pixel in the input panorama. We form triangular faces by connecting neighboring vertices, with horizontal wrapping to connect the pixels on the left and right sides of the panorama.

Mesh Inpainting As mentioned earlier, depth discontinuities cause noticeable artifacts in the mesh, and since the depth map only provides one depth value per pixel, the initial mesh cannot represent occluded surfaces. The 3D Photo Inpainting method by Shih et al. [10] addresses these artifacts by "tearing" the mesh at depth edges, synthesizing occluded geometry at the depth edges, and generating image content for the synthesized geometry. They provide pre-trained neural network models for these various inpainting steps. However, their method is designed for perspective images and so does not support the spherical coordinate system used in an equirectangular panorama, and does not respect horizontal wrapping between vertices on the left and right edges of the panorama.

We modified the 3D Photo Inpainting codebase to addresses these issues and support panoramic inputs. In particular, we: modified the geometry creation steps to use spherical coordinates; took care to connect mesh vertices between the left and right panorama edges; and disabled border extrapolation, since in a panorama there is no border to the image. An example result is shown in Figure 2.

View Synthesis After initial mesh formation and inpainting, the textured mesh is stored in a PLY file for compatibility with many rendering software packages. We wrote a custom web viewer in three.js with support for viewing in a VR headset.

3 Results and Conclusions

We used the 360 Replica Dataset Generator [2] to produce synthetic 2048×1024 panoramic images of indoor scenes. For each input panorama we rendered three nearby panoramas for ground truth comparison. We qualitatively evaluated the results of our method in comparison to the initial textured mesh (without inpainting).

An example view synthesis result is shown in Figure 3. The mesh without inpainting shows noticeable stretching artifacts, whereas the results for our method are smooth and have a plausible appearance in occluded regions.

The video example provided in supplementary material shows a complete 360° spin around an indoor scene processed using our method, with the virtual camera moving in and out as it spins to highlight the extent of parallax achieved with our method.

Our framework builds an immersive 360° VR environment from only a single equirectangular panorama as input. By adapting the 3D Photo Inpainting method to panoramas, we generate plausible geometry and texture for occluded regions, and support straightforward compatibility with standard rendering software through a textured mesh representation. Future work lies in scaling the approach to higher-resolution panoramas to ensure a comfortable and immersive experience in a VR headset.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 2144822 and the New Zealand Marsden Council through Grant UOO1724.

References

[1] G. Albanis, N. Zioulis, P. Drakoulis, V. Gkitsas, V. Sterzentsenko, F. Alvarez, D. Zarpalas, and P. Daras. Pano3d: A holistic benchmark and a solid baseline for 360deg depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3727–3737, 2021.

[2] B. Attal, S. Ling, A. Gokaslan, C. Richardt, and J. Tompkin. MatryODShka: Real-time 6DoF video view synthesis using multi-sphere images. In *European Conference on Computer Vision (ECCV)*, Aug. 2020.

[3] L. He, B. Jian, Y. Wen, H. Zhu, K. Liu, W. Feng, and S. Liu. Rethinking supervised depth estimation for 360 panoramic imagery. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 5169–5177. IEEE, 2022.

[4] K.-E. Lin, Z. Xu, B. Mildenhall, P. P. Srinivasan, H.-G. Yannick, S. DiVerdi, Q. Sun, K. Sunkavalli, and R. Ramamoorthi. Deep multi depth panoramas for view synthesis. In *ECCV*, 2020.

[5] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[6] M. Niemeyer, J. T. Barron, B. Mildenhall, M. S. Sajjadi, A. Geiger, and N. Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5480–5490, 2022.

[7] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020.

[8] M. Rey-Area, M. Yuan, and C. Richardt. 360MonoDepth: High-resolution 360deg monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3762–3772, June 2022.

[9] A. Serrano, I. Kim, Z. Chen, S. DiVerdi, D. Gutierrez, A. Hertzmann, and B. Masia. Motion parallax for 360 RGBD video. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):1817–1827, 2019.

[10] M.-L. Shih, S.-Y. Su, J. Kopf, and J.-B. Huang. 3D photography using context-aware layered depth inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[11] J. Waidhofer, R. Gadgil, A. Dickson, S. Zollmann, and J. Ventura. PanoSynthVR: Toward light-weight 360-degree view synthesis from a single panoramic input. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 584–592, 2022. doi: 10.1109/ISMAR55827.2022.00075