OmniSyn: Synthesizing 360 Videos with Wide-baseline Panoramas

David Li[†,‡,*] Yinda Zhang[†] Christian Häne[†] Danhang Tang[†] Amitabh Varshney[‡] Ruofei Du[†,*]
[†] Google Research [‡] University of Maryland, College Park

Abstract

Immersive maps such as Google Street View and Bing Streetside provide true-to-life views with a massive collection of panoramas. However, these panoramas are only available at sparse intervals along the path they are taken, resulting in visual discontinuities during navigation. Prior art in view synthesis is usually built upon a set of perspective images, a pair of stereoscopic images, or a monocular image, but barely examines wide-baseline panoramas, which are widely adopted in commercial platforms to optimize bandwidth and storage usage. In this paper, we leverage the unique characteristics of wide-baseline panoramas and present OmniSyn, a novel pipeline for 360° view synthesis between wide-baseline panoramas. OmniSyn predicts omnidirectional depth maps using a spherical cost volume and a monocular skip connection, renders meshes to 360° images, and synthesizes intermediate views with a fusion network. We envision our work may inspire future research for this unheeded real-world task and eventually produce a smoother experience for navigating immersive maps.

Index Terms: Computing methodologies—Computer graphics—Image manipulation—Image-based rendering

1 Introduction

Recent advances in 360° cameras and virtual reality headsets have promoted the interests of tourists, renters, and photographers to capture or explore 360 images on commercial platforms such as Google Street View [1], Bing Streetside[1], and Matterport[2]. These platforms allow users to virtually walk through a city or preview a floorplan by interpolating between panoramas. However, the existing solutions lack the visual continuity from one view to the next and suffer from ghosting artifacts caused by warping with inaccurate geometry. While prior art reports successful view synthesis experiments in a set of perspective images [3, 4, 8–11], a single image [14, 21], and ODS video [2], limited prior work addresses how we could synthesize an omnidirectional video with large movements, *i.e.*, using a *wide-baseline* pair of panoramas. Since wide-baseline panoramas are broadly adopted for capturing and streaming on commercial platforms, we envision view synthesis on this data can reduce the additional effort of converting to perspective images and leverage the full field-of-view for better alignment between the two panoramas.

Our goal is to synthesize 360° videos between wide-baseline panoramas and stream to consumer devices for an interactive and seamless experience (Fig. 1). Unlike past research which only synthesizes novel views within a limited volume [3, 14] or along a trajectory in rectilinear projection [8], our generated 360° video allows users to move forward/backward, stop at any point, and look around from any perspective. This unlocks a wide range of virtual reality applications such as cinematography [19], teleconferencing [20], and virtual tourism [5].

*Corresponding authors: dli7319@umd.edu and me@duruofei.com

[1]Bing Streetside: https://microsoft.com/en-us/maps/streetside
[2]Matterport: https://matterport.com

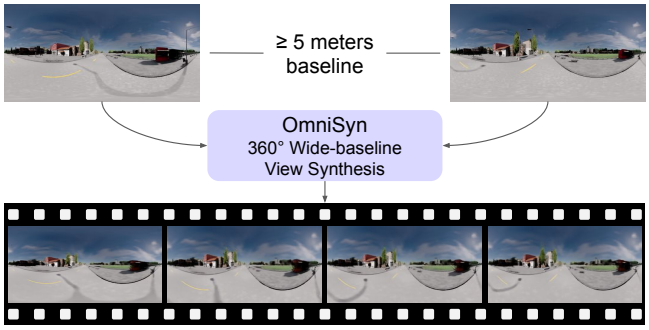

Figure 1: Given two wide-baseline 360° images and poses, our goal is to synthesize a video sequence of intermediate frames with plausible movement and alignment between the input images.

Classical methods for view synthesis [9, 11] often rely on structure-from-motion or multi-view stereo pipelines to perform a sparse 3D reconstruction and develop algorithms to densify the reconstruction. Modern view synthesis methods either rely on additional inputs such as 3D models [17] or denser images sets in the case of NeRF [16]. On one hand, most existing works target perspective images which encounter *visual discontinuities* when objects move outside their field of view. On the other hand, applying monocular methods to multi-view scenarios leads to alignment issues between images as intermediate images are not fused from multiple views. Further, real-world street view images do not have a sufficiently dense layout to apply multiview stereo methods. So our research questions are: How can we achieve novel view synthesis between a pair of wide-baseline 360° images? How can we leverage the full field of view to align the pair of panoramas and inpaint occluded regions?

To answer these questions, we develop a pipeline for 360° view synthesis using *wide-baseline* panoramas. Unlike prior setups, our inputs are a pair of 360° images which are at least 5 meters apart for street view scenes and 2 meters apart for indoor scenes without Lidar or 3D geometry. Our pipeline is comprised of a depth predictor, a 360° mesh renderer, and an image fusion network.

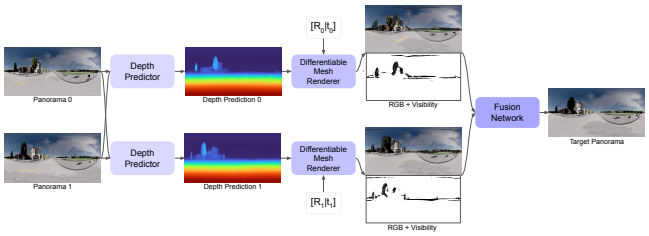

Figure 2: Our 360° view synthesis pipeline consists of a stereo depth predictor, a 360° mesh renderer, and an image fusion network. All the three components are differentiable, while only the depth predictors and image fusion network have learnable parameters.

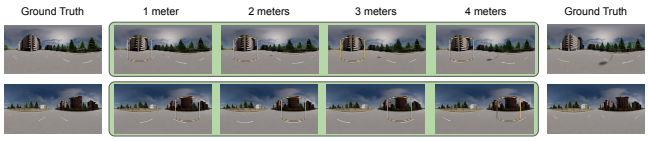

Figure 3: Qualitative results of OmniSyn along two Carla streets.

2 View Synthesis Pipeline

OmniSyn, shown in Fig. 2, consists of three major components to synthesize novel views for wide-baseline panoramas: a stereo depth predictor, a differentiable 360° mesh renderer, and an image fusion network. OmniSyn takes two wide-baseline panoramas in equirectangular projection (ERP) and relative poses as inputs.

Given two 360° ERP panoramas and their relative poses to a target position, we first estimate depth using a spherical sweep cost volume. Then we build a mesh representation for each panorama with discontinuities computed from depth estimates. Each mesh is rendered from the target position into a separate 360° panorama with a corresponding visibility map. Finally, our fusion network joins the two panoramas together resolving ambiguities and inpaints any holes to produce the final 360° panorama.

2.1 Depth Prediction

To perform consistent depth prediction from wide-baseline panoramas, we build a network architecture tailored for stereo 360° depth estimation inspired by StereoNet [13] and 360° depth estimation [7]. Our depth prediction network consists of three components: a 2D feature encoder, a 3D cost volume refinement network, and a 2D depth decoder. Stereo depth estimation allows the network to match features presented in both of the 360° images for aligned depth estimation in contrast to a purely monocular approach.

2.2 Mesh Creation

To render each image from the novel viewpoint, we first create a spherical mesh for each input image similar to the perspective mesh of Worldsheet [12]. For a $W \times H$ image, we instantiate a spherical mesh with $2H$ height segments and $2W$ width segments and offset vertices based on the depth prediction. Large gradients in the depth image correspond to edges of buildings and other structures within the RGB image. For these areas, we discard triangles within the spherical mesh to accurately represent the underlying discontinuity. With the meshes created and discontinuities calculated, we use a modified version of PyTorch3D [18] to render the mesh from the new viewpoint to a 360° RGBD image. Holes due to occlusions in the original images are extracted to a visibility mask as shown in Fig. 2.

2.3 Image Fusion

After rendering each mesh from the new viewpoint, holes appear in each rendering due to the occlusions in the synthesized view. Thus, we use a fusion network to fuse the two mesh renderings and inpaint the holes into a single consistent panorama. We input both RGB mesh renderings and the corresponding binary visibility masks into the fusion network to get the final fused and inpainted RGB image.

3 Conclusion

In this paper, we briefly examine the task of intermediate view synthesis for wide-baseline 360° panoramas, typically ≥ 5 meters apart, and propose OmniSyn which leverages 360° stereo depth estimation, mesh rendering, and 360° fusion to synthesize plausible 360° street view panoramas from static scenes. We refer readers to the full paper on our project page for technical details. In the future, we hope to combine view synthesis with scene understanding [22], foveated video streaming [15], and neural lightfields [6] to deliver on-demand 6-DoF streaming for street view scenes.

Acknowledgments

We would like to thank the anonymous reviewers for the valuable comments on the manuscript. This work has been supported in part by the NSF Grant 18-23321 and the State of Maryland's MPower initiative.

References

[1] D. Anguelov, C. Dulong, D. Filip, C. Frueh, S. Lafon, R. Lyon, A. Ogale, L. Vincent, and J. Weaver. Google street view: Capturing the world at street level. *Computer*, 2010.

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

[3] M. Broxton, J. Flynn, R. Overbeck, D. Erickson, P. Hedman, M. Duvall, J. Dourgarian, J. Busch, M. Whalen, and P. Debevec. Immersive light field video with a layered mesh representation. *ACM Trans. Graph.*, 39(4), July 2020.

[4] G. Chaurasia, S. Duchene, O. Sorkine-Hornung, and G. Drettakis. Depth synthesis and local warps for plausible image-based navigation. *ACM TOG*, 2013.

[5] R. Du, D. Li, and A. Varshney. Geollery: A Mixed Reality Social Media Platform. In *CHI*, 2019.

[6] B. Y. Feng and A. Varshney. Signet: Efficient neural representations for light fields. In *ICCV*, 2021.

[7] B. Y. Feng, W. Yao, Z. Liu, and A. Varshney. Deep depth estimation on 360° images with a double quaternion loss. In *3DV*, 2020.

[8] J. Flynn, I. Neulander, J. Philbin, and N. Snavely. DeepStereo: Learning to predict new views from the world's imagery. In *CVPR*, 2016.

[9] P. Hedman, S. Alsisan, R. Szeliski, and J. Kopf. Casual 3D Photography. *ACM TOG (Proc. SIGGRAPH Asia)*, 2017.

[10] P. Hedman and J. Kopf. Instant 3D Photography. *ACM TOG (Proc. SIGGRAPH)*, 2018.

[11] P. Hedman, J. Philip, T. Price, J.-M. Frahm, G. Drettakis, and G. Brostow. Deep blending for free-viewpoint image-based rendering. *ACM TOG (Proc. SIGGRAPH Asia)*, 2018.

[12] R. Hu, N. Ravi, A. C. Berg, and D. Pathak. Worldsheet: Wrapping the world in a 3d sheet for view synthesis from a single image. In *ICCV*, 2021.

[13] S. Khamis, S. Fanello, C. Rhemann, A. Kowdle, J. Valentin, and S. Izadi. Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. In *ECCV*, 2018.

[14] J. Kopf, K. Matzen, S. Alsisan, O. Quigley, F. Ge, Y. Chong, J. Patterson, J.-M. Frahm, S. Wu, M. Yu, et al. One shot 3D photography. *ACM TOG*, 2020.

[15] D. Li, R. Du, A. Babu, C. D. Brumar, and A. Varshney. A log-rectilinear transformation for foveated 360-degree video streaming. *IEEE TVCG*, 2021.

[16] 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. In *ECCV*, 2020.

[17] J. Park, I.-B. Jeon, S.-E. Yoon, and W. Woo. Instant panoramic texture mapping with semantic object matching for large-scale urban scene reproduction. *IEEE TVCG*, 2021.

[18] N. Ravi, J. Reizenstein, D. Novotny, T. Gordon, W.-Y. Lo, J. Johnson, and G. Gkioxari. Accelerating 3D deep learning with pytorch3d. *arXiv:2007.08501*, 2020.

[19] Y.-C. Su, D. Jayaraman, and K. Grauman. Pano2vid: Automatic cinematography for watching 360 videos. In *ACCV*. Springer, 2016.

[20] T. Teo, L. Lawrence, G. A. Lee, M. Billinghurst, and M. Adcock. Mixed reality remote collaboration combining 360 video and 3D reconstruction. In *CHI*, 2019.

[21] O. Wiles, G. Gkioxari, R. Szeliski, and J. Johnson. SynSin: End-to-end view synthesis from a single image. In *CVPR*, 2020.

[22] C. Zhang, Z. Cui, C. Chen, S. Liu, B. Zeng, H. Bao, and Y. Zhang. Deeppanocontext: Panoramic 3d scene understanding with holistic scene context graph and relation-based optimization. In *ICCV*, 2021.