Using Synthetic Data Generation to Probe Multi-View Stereo Networks

Pranav Acharya ◆ Daniel Lohn ◆ Vivian Ross ◆ Maya Ha ◆ Alexander Rich ◆ Ehsan Sayyad ◆ Tobias Höllerer

UC SANTA BARBARA

Motivation

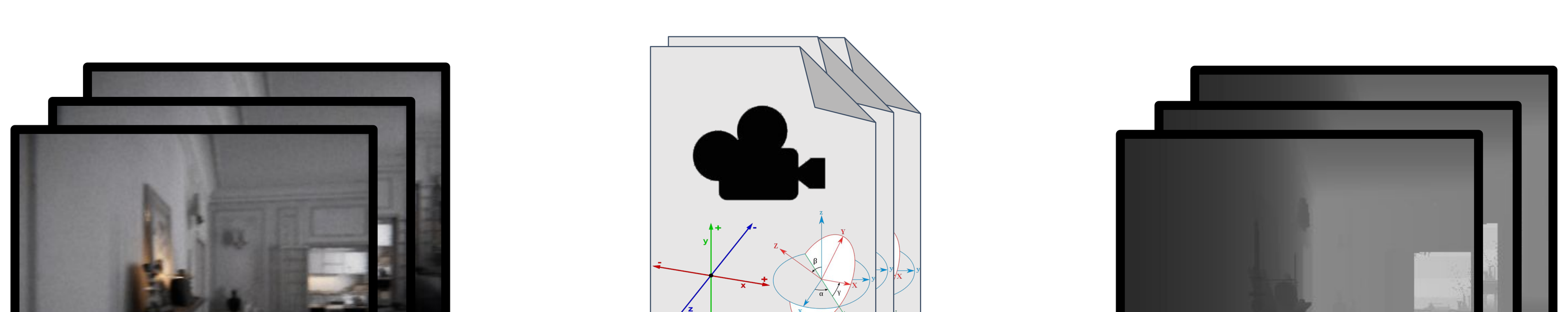
- 3D reconstruction has various applications ranging from autonomous driving to augmented reality. We investigate Multi-View Stereo (MVS), a subtask of 3D reconstruction.
- It is unknown how well pre-trained MVS algorithms are able to generalize to scenarios not resembling the training dataset.
- Our goal is to generate customizable synthetic training data which will allow us to evaluate various existing MVS networks [3, 5, 6, 8] as well as the properties of the data itself.

Our Contribution

- Created a tool in Unity 3D Game Engine to generate 2D datasets from existing 3D datasets given adjustable parameters.
- We test network error across three datasets (Matterport3D [2], ArchViz [1], SUNCG [7]), four parameters (Camera Height, Camera Pitch, Camera Yaw, Sample Distance), and five MVS networks.

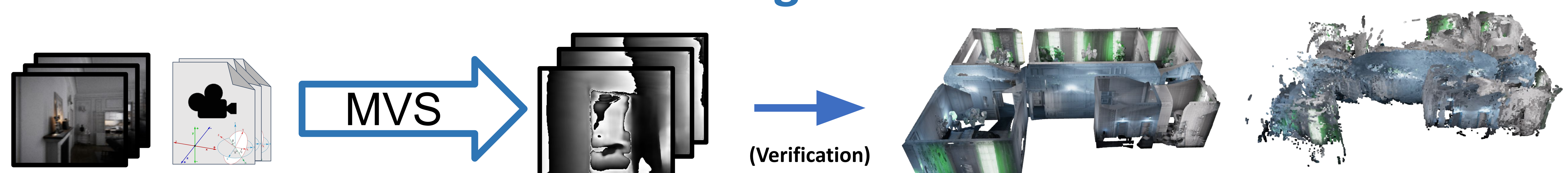
Input: 3D Dataset + Parameters

Output: RGB Images, Extrinsic Matrices, Depth Maps



Extrinsic Matrices contain the camera position and rotation of the camera for every picture taken

Testing Process



RGB Images + Extrinsic Matrices

***Predicted* Depth Maps**

3D Scene reconstructed from ground truth depth maps

3D Scene reconstructed from MVS predicted depth maps

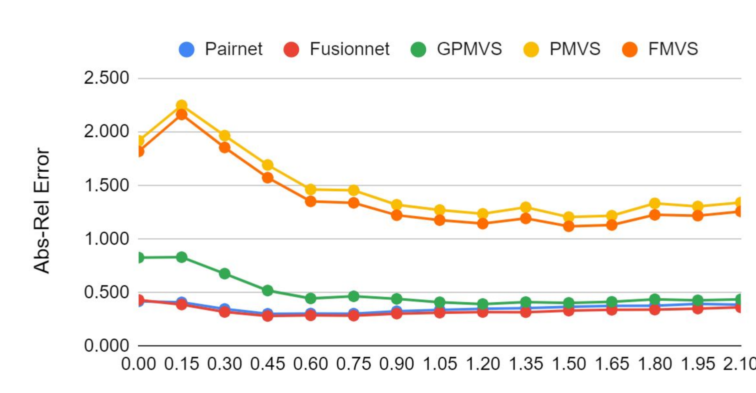
We compare MVS predicted depth map with ground truth (generated) depth map using *abs rel* error, and use these results to inform selection of future parameter settings. MVS networks used: Pairnet [5], Fusionnet [5], PMVS [3], GPMVS [6], FMVS [8].

Results

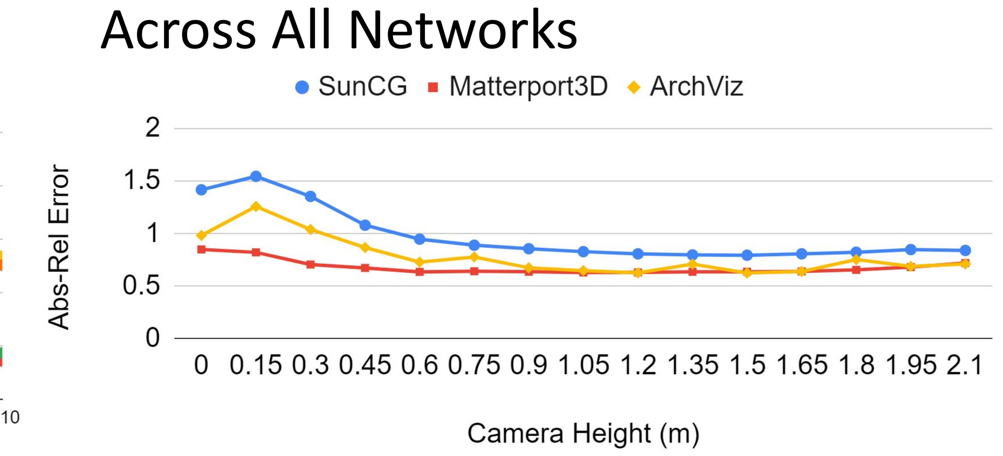
$$\textit{Absolute Relative Error} = \frac{1}{n}\sum \frac{|d - d^*|}{d^*}$$

n = number of pixels in depth map, d = predicted depth map, d* = ground truth depth map

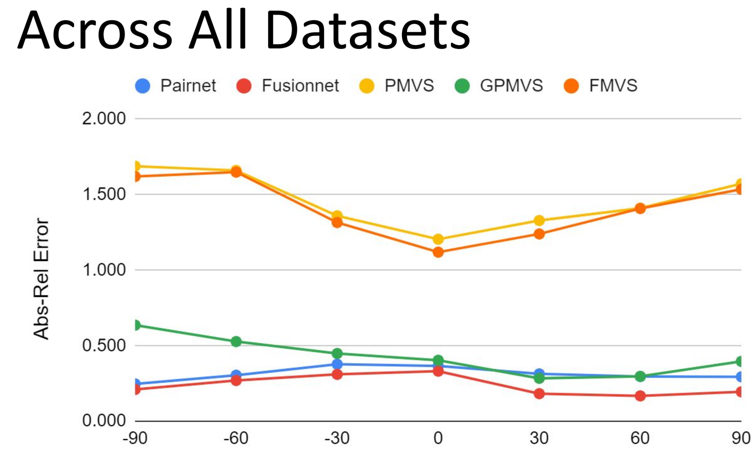
Average Camera Height Error Across All Datasets



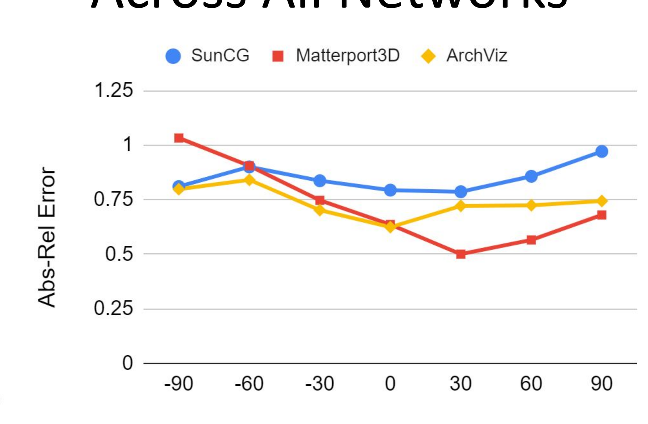
Average Camera Height Error Across All Networks



Average Camera Pitch Error Across All Datasets



Average Camera Pitch Error Across All Networks



Future Work

- Add more parameters to tool to give more control over generated data
- Train machine learning networks on our data to determine how well our data transfers to real-world applications

Discussion

- The different parameter settings offer insights on how network architecture affects performance. Differences in performance between GPMVS [6], Pairnet [5], and Fusionnet [5] are likely caused by the absence of deep features in the cost volume construction of GPMVS.
- Matterport3D [2] and ArchViz's [1] similar textures likely cause network predictions on image sequences derived from these datasets to be more similar to each other than to predictions on sequences from SUNCG [7].
- Variations of camera height and pitch produce the two largest average maximum abs-rel errors. We hypothesize that all of the networks used are most sensitive to varying vertical camera views.
- We found the best choices of the values for each camera parameter vary for each network. The networks trained on ScanNet [4] have the least error, likely due to the similarity between our training data and ScanNet.

Acknowledgements

We would like to thank Noah Stier, Diba Mirza, Aarti Jivrajani, Will Eiers, and Feng Yang.

[1] ArchVizPRO: The best way to learn real-time archviz visualization in unity. https://oneirosvr.com/portfolio/archvizpro/.
[2] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, AndyZeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. International Conference on3D Vision (3DV), 2017.
[3] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In The IEEE International Conference on Computer Vision (ICCV), 2019.
[4] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes, 2017.
[5] Arda Duzceker, Silvano Galliani, Christoph Vogel, Pablo Speciale, Mihai Dusmanu, and Marc Pollefeys. Deep-video mvs: Multi-view stereo on video with recurrent spatio-temporal fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2021.
[6] Yuxin Hou, Juho Kannala, and Arno Solin. Multi-view stereo by temporal nonparametric fusion. In Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV), October 2019.
[7] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. Proceedings of 30th IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[8] Zehao Yu and Shenghua Gao. Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In CVPR, 2020.