Team name Random walker (Team members from grad and undergrad sections)

Project Title Toward interpreting and improving learnt MVS machines on textureless regions

Project summary Multiview stereo (MVS) is a fundamental algorithm to restore the 3D structures (basically equivalent to the depth information) of a scene given a set of calibrated 2D images. There are two rough classes of MVS: feature-based methods and geometry-based methods. The first one can be naturally extended into learning-based methods as most DNNs (e.g. CNNs) [12, 4, 5, 6, 1, 8, 10, 3, 11, 9] are good at extracting features. This will be also the focus of this project. However, the capabilities of extracting deep features might malfunction on textureless regions for no significant features are present. As such, understanding how CNNs perform worse on these regions than texture-rich regions can definitely guide to design a better model capable of extracting global contexts. We also aim to apply some tentative efforts to improve the depth estimation on regions poor in textures.

Approach We will focus on three recent and computationally friendly CNN models [3, 11, 9] and take [11] as the baseline. Firstly, we will train these three models and attempt achieving the results as shown in the original papers. Secondly we will acquire some calibrated images with textureless regions contained to test the performance of the trained models. Thirdly, we will analyze the output of the feature extraction modules (possibly by the visualization techniques) to interpret what kind of features (regions) leads to low/large depth estimation errors. Lastly, we want to replace the feature extraction module with ConvNeXt [7] because it outperforms ViT [2], which demonstrated the ability to capture global contexts. Therefore, we expect ConvNeXt to also extract somewhat *global* semantics on the textureless regions to mitigate the downsides brought by locality of CNNs.

Data set DTU, ETH3D, and manually acquired data for testing.

Team members Yuxin Sun (undergrad), Junze Huang (grad), Huizong Yang (undergrad)

References

[1] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1538–1547, 2019.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[3] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2495–2504, 2020.

[4] Wilfried Hartmann, Silvano Galliani, Michal Havlena, Luc Van Gool, and Konrad Schindler. Learned multi-patch similarity. In *Proceedings of the IEEE international conference on computer vision*, pages 1586–1594, 2017.

[5] Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang. Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2307–2315, 2017.

[6] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. *Advances in neural information processing systems*, 30, 2017.

[7] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *arXiv preprint arXiv:2201.03545*, 2022.

[8] Keyang Luo, Tao Guan, Lili Ju, Haipeng Huang, and Yawei Luo. P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10452–10461, 2019.

[9] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14194–14203, 2021.

[10] Qingshan Xu and Wenbing Tao. Learning inverse depth regression for multi-view stereo with correlation cost volume. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12508–12515, 2020.

[11] Jiayu Yang, Wei Mao, Jose M Alvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4877–4886, 2020.

[12] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018.