Spherical Epipolar Rectification for Deep Two-View Absolute Depth Estimation Supplementary

Pierre-André Brousseau
Université de Montréal
pierre-andre.brousseau@umontreal.ca

Sébastien Roy
Université de Montréal
roys@iro.umontreal.ca

Abstract

Stereo depth estimation relies on triangulation to compute metric depth from disparity. Depth estimations of the same scene are therefore inherently coherent with each other as they encode a 3D world. This supplementary material provides additional experiments with regards to stability. Because animated results cannot be provided in the main paper, they are instead provided as supplementary material. ***This document is accompanied by two*** *(2)* ***videos****.*

1. Full Spherical Rectification Demonstration

The Spherical Crop operator presented in the main paper reduces distortion with respect to other polar rectification methods. The video Rectification-Visualization illustrates Spherical Rectification with Spherical Crop of a full image with almost 100% forward baseline. This forward baseline represents the upper bound on distortion generated by general camera geometries. The video shows that besides directly on the epipole, the Spherical Rectification with Spherical Crop adequately handles distortion and allows for stereo precise matching.

2. Depth Estimation Stability Demonstration

The main paper presents that using more than two frames during depth estimation improves performance. This is because, for a given scene, the estimated depths from triangulation are all naturally coherent. The video Depth-Visualization illustrates the depth estimation stability. As a reminder, the neural network for stereo disparity estimation does not have a context encoder and there is no interframe smoothing. This video demonstrates that monocular stereo matching yields stable depths with triangulation.

3. Comparison with Relative Depth Estimators

As an additional visualization of stability, we present histogram results of depth estimation as compared with a state of the art relative depth estimator Depth Anything V2. The 3D histogram in Fig. 1 (TOP) illustrates that predicted depths are more concentrated around the ground truth in the case of spherical stereo matching with respect to Depth Anything V2. This means that predicted depths are more coherent with the ground truth and other frames in the case of Spherical Stereo. In the case of Depth Anything V2, predicted depths vary from one frame to another, that is there is no coherence with regards to the 3D world.

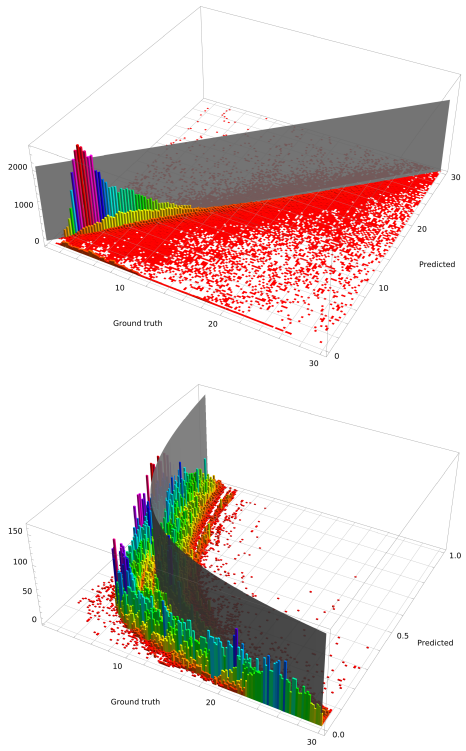

Figure 1. Predicted depth vs ground truth depth for Spherical Stereo (TOP) and Depth Anything V2 (BOT) on the TartanAir carwelding test images. Spherical Stereo has a tight linear relationship with the ground truth. Depth Anything V2 has very few completely erroneous matches but is overall very wide around its rescaling curve.

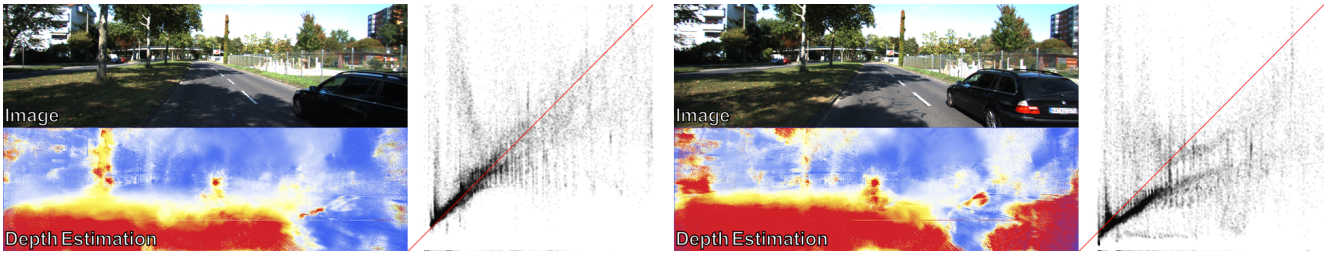

Figure 2. Errors in groundtruth poses. Depth output for images 51 (TOP) and 52 (BOT) from the KITTI Eigen split test set using Spherical Stereo. The depth images (LEFT) look very similar. The histograms (RIGHT) of predicted depth vs. ground truth depth illustrate how an error in the ground truth pose leads to high errors.

4. Analyzing Pose on Kitti using Depth

The presented results give insights on IMU poses and why we suggest there are errors in ground truth poses for KITTI Eigen. Fig. 2 (TOP) shows that Spherical Stereo adequately recovers depths while fig. 2 (BOT) shows that Spherical Stereo underestimates depths. The right histogram illustrates that there is indeed a scaling factor missing between the predicted depth and the actual depth. We attribute this to an incorrect baseline as expressed in Eq. 13 of the main paper. In these two cases, both images should lead to compatible depths from a qualitative standpoint, we postulate that large errors can be the result of small baseline or erroneous poses.

5. Implementation Details.

The feature encoder architecture shown in Fig. 3 of the main paper is 16 ConvNext [3] blocks with a feature size of 64. The rectification, the derectification and the disparity to depth operators are fully differentiable and are implemented as neural network functions. When using spherical crop, for regions with very significant distortion, some pixels cannot be recovered after derectification. These pixels are masked in the training loss. The network is trained with the Adam Optimizer [2] with a learning rate of 10^{-3} and a batch size of 2 on a RTX3090 Ti GPU. The training implementation is made with Mathematica 14.1 [1].

References

[1] Wolfram Research, Inc. Mathematica, Version 14.1, 2024. Champaign, IL, 2024. 2

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[3] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 2