Depth Anyviewpoint? Evaluating Monocular Depth Over Viewpoint Sequences

Theo Rode*, Rui Zhang*, Ryan Aubrey†, Olivia Li*, Jessica Huang*, Liv Chu*, Healani Dowd*, and Calden Wloka*
*Harvey Mudd College
Email: cwloka@hmc.edu
†Pomona College

***Abstract*—Monocular depth estimation aims to predict depth maps from single RGB images—forgoing the need for stereo camera pairs—and has recently seen impressive leaps in performance due to the introduction of affine-invariant losses and the use of foundation model features. These monocular depth models predict affine-invariant depth, which describes the relative positions of objects in the scene instead of absolute differences. While these models demonstrate strong performance on single image benchmarks, their effectiveness at preserving consistent spatial structure over viewpoint translation has been less thoroughly explored. Many potential real-world applications, such as handling video streams as input, rely on coherent three-dimensional consistency in scenes. We develop a novel evaluation protocol for monocular depth models using a sequence of camera positions to test the consistency of prediction across variation in viewpoint. We demonstrate this methodology in a novel real-world dataset of video sequences where we are able to anchor depth model output to metric depth values using structure from motion, as well as in synthetic scenes where ground-truth metric depth is directly available. Our work demonstrates open challenges in the development of broadly applicable monocular depth foundation models, particularly with their ability to perform consistently under varied viewpoints and camera motion sequences.**

***Index Terms*—Monocular depth estimation, single-image depth prediction, reliability and robustness.**

I. Introduction

Monocular depth estimation (MDE) aims to provide an alternative method for retrieving depth information by enabling depth prediction from a single RGB image. Recent innovations making use of affine-invariant loss and pre-trained foundation models have led to a rapid improvement in the effectiveness of monocular depth models and a proliferation of methods [1]–[7]. While the ability to recover metric depth is limited with these models, they have shown remarkable promise in recovering relative depth within a scene. Where alternative methods for recovering depth in a scene may be impractical due to additional required hardware or the need to capture the scene from many viewpoints [8], MDE requires only a single camera or static image.

Standard metrics for evaluating monocular depth models include absolute relative error (AbsRel) and proportion of pixels exceeding 25% error (δ_1) [1]–[3], [5]. While these metrics together effectively convey performance over a given image, they are typically calculated treating test images as independent and disjoint, potentially missing biases in model effectiveness for less well represented viewpoints, and do not capture potential lapses in spatial coherence across viewpoints. Recent work extending MDE to video streams has considered the need for spatial coherence [3], [6], [7], however, the quantitative metrics proposed focus on consistency between adjacent frames, where critical failures in spatial representation may only become apparent across the entire sequence.

While monocular depth models show significant promise for applications across robotics and other scene understanding tasks, these applications will be dependent on the spatial consistency of these models. Further, while evaluating the relative depth performance of these models gives a baseline understanding of their performance, real-world applications will require the use of metric depth maps scaled from the model output. Hence, for understanding the expected consistency of these models across video sequences, we must explicitly examine the consistency of their metric depth predictions. With this, our contributions are as follows:

- Curation of a monocular depth benchmarking dataset, consisting of multiple classes of camera movements across real and synthetic scenes.
- Development of two novel metrics, *warp* and *drift* error, measuring the temporal consistency of monocular depth models across the entire video sequence.
- Providing a comprehensive benchmark suite demonstrating the spatial coherence of 6 monocular depth models.

II. Methods

Our primary focus of investigation is the spatial consistency of monocular depth models over a range of scenes and camera movements. We examine both real-world and synthetic scenes, building a real-world dataset of 2,245 images across 23 scenes and a synthetic dataset of 2024 images across 7 scenes. In order to examine the effect of camera trajectory, we broadly classify each video into either an arc or a pan camera movement. Further, while our synthetic dataset has access to ground-truth depth information, we use structure from motion—specifically COLMAP [9], [10]—in real-world scenes in order to have metric depth-priors.

In order to scale a given relative depth map—output from the monocular depth model—to metric depth, we compute a least-squares fit for a linear scaling from the relative to the metric depth space. In particular, this scaling aims to map the relative depth to the known ground-truth depth in the scene

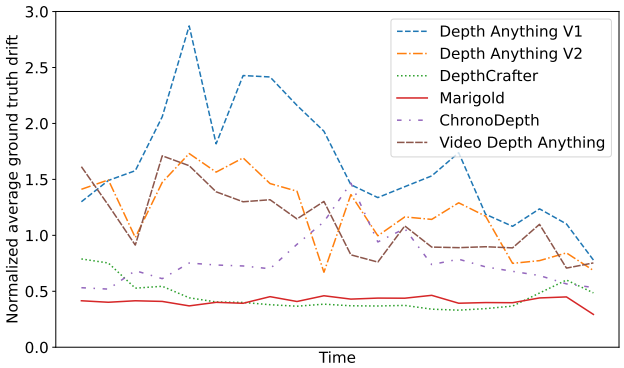

Fig. 1. Warp error reported across all real scenes. Warp error is normalized to average depth of COLMAP points in scene. Outliers (values in the top and bottom first percentile) are removed.

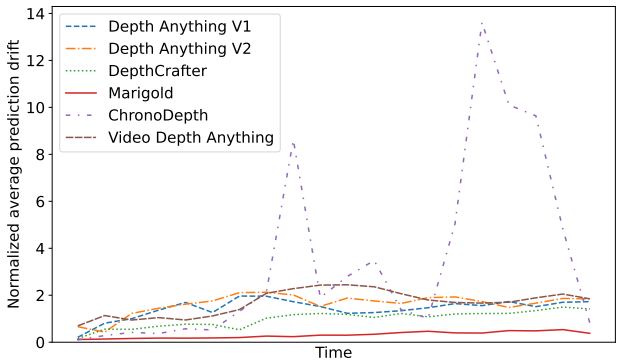

Fig. 2. Drift error reported across all real scenes. Drift error is normalized to average depth of COLMAP points in scene. Outliers (values in the top and bottom first percentile) are removed.

for the synthetic scenes, or to to the metric depth recovered from structure from motion for real scenes (hereafter referred to as ground truth for our purposes). This computation is done per-frame in each video.

We propose two new metrics for evaluating the consistency performance of monocular depth models on video sequences: warp and drift error. For both warp and drift error, we compute error over a set of points with corresponding ground truth information. For each point, we use the predicted metric depth map in order to re-project a 3D world coordinate. Warp error is then defined as the average Euclidean distance between the predicted and ground-truth 3D location of each point in each frame. Instead of using ground-truth locations, drift error computes the Euclidean distance between each predicted location and the *first* predicted 3D location of that point. Namely, we use the known history of which frames each point is visible in from the structure from motion in order to compare the prediction for the first time a point was visible to every subsequent frame it is visible in. In this light, warp error aims to measure the extent to which the predicted depth map "warps" the ground-truth depth map and drift error aims to measure the extent to which a model's prediction drifts throughout the video sequence.

For comparing warp and drift error across scenes, we normalize the errors in each scene by the average ground-truth depth across the entire video. Further, we remove the top and bottom first percentile of error values in order to reduce the impact of major errors and noise. We separately report the average magnitude of these major errors.

III. Results

We show warp and drift error averaged over all real-world scenes in Fig. 1 and Fig. 2, respectively. Each plot shows the trajectory of both error measures—normalized by average scene depth—through 30 temporal buckets to emphasize temporal errors. Notably, we see large drift error for ChronoDepth at the end of scenes, while not seeing deviation in warp. Further, for Depth Anything V1, Depth Anything V2, and Video Depth Anything, we see warp error tends to drop on average through time, even though drift error tends to creep upward.

IV. Conclusion

We explore the temporal consistency of monocular depth models across a novel dataset of real-world and synthetic video. Additionally, we propose two new metrics, warp and drift error, which measure a model's deviation from ground truth and from its original predictions, respectively. We demonstrate how simply measuring temporal consistency against either of these measures fails to completely capture the inconsistencies present in each model. While not outlined here, we additionally demonstrate how these temporal biases extend to different kinds of camera movement.

References

[1] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler, "Repurposing diffusion-based image generators for monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9492–9502, 2024.

[2] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *Advances in Neural Information Processing Systems*, vol. 37, pp. 21875–21911, 2025.

[3] W. Hu, X. Gao, X. Li, S. Zhao, X. Cun, Y. Zhang, L. Quan, and Y. Shan, "Depthcrafter: Generating consistent long depth sequences for open-world videos," *arXiv preprint arXiv:2409.02095*, 2024.

[4] K. Lasinger, R. Ranftl, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *arXiv preprint arXiv:1907.01341*, 2019.

[5] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *CVPR*, 2024.

[6] J. Shao, Y. Yang, H. Zhou, Y. Zhang, Y. Shen, V. Guizilini, Y. Wang, M. Poggi, and Y. Liao, "Learning temporally consistent video depth from video diffusion priors," 2024.

[7] S. Chen, H. Guo, S. Zhu, F. Zhang, Z. Huang, J. Feng, and B. Kang, "Video depth anything: Consistent depth estimation for super-long videos," 2025.

[8] O. Özyeşil, V. Voroninski, R. Basri, and A. Singer, "A survey of structure from motion*.," *Acta Numerica*, vol. 26, pp. 305–364, 2017.

[9] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, "Pixelwise view selection for unstructured multi-view stereo," in *European Conference on Computer Vision (ECCV)*, 2016.

[10] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.