Consistent Video Depth Estimation
影像深度估計

Group Number: EECS003 Member: 陳映寰 、楊侑霖

Abstract

We present a novel method for estimating depth and camera pose in dynamic scenes featuring moving objects, leveraging videos captured by moving cameras. Building upon the monocular depth estimator MiDAS, our approach enhances accuracy by fine-tuning MiDAS[1] with geometric and temporal constraints to mitigate inconsistencies in frame-by-frame predictions. Our method demonstrates superior depth estimation capabilities, comparable to state-of-the-art results.

Our Method

Overview

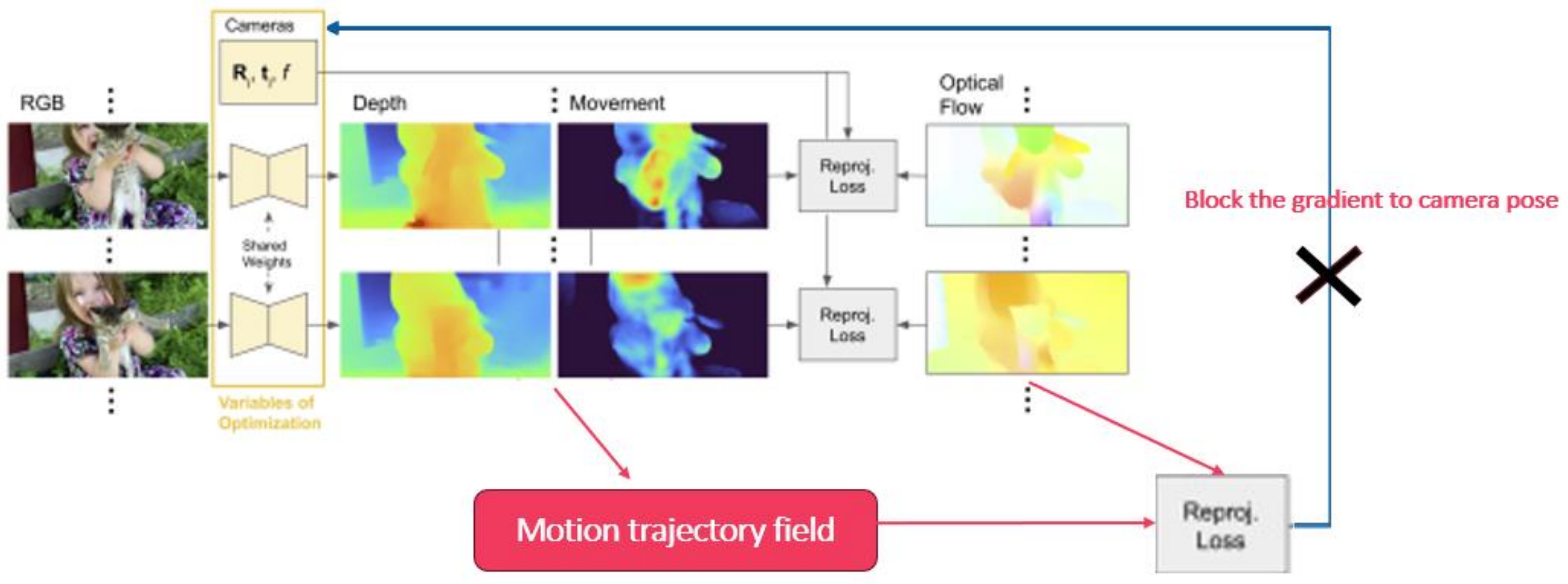

Figure 1: Overview of our model

In Figure 1, we present an overview of our model. We employ a test-time training approach to train on specific video segments, fine-tuning the camera pose, depth network, and movement estimator with Reprojection Loss. Additionally, we introduce an auxiliary motion trajectory field for 3D point scene flow prediction. This fine-tuning process enables our depth network to produce consistent video depth.

Reprojection Loss

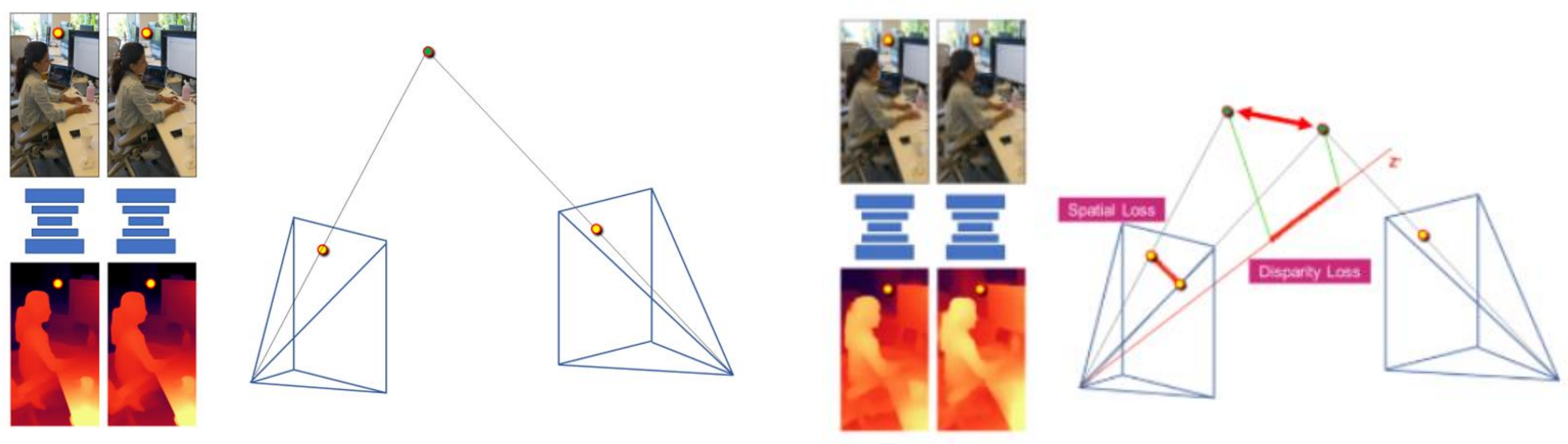

Figure 2: Reprojection Loss, mentioned in [2]

Figure 2 displays two frames from a video. If the two red dots in the upper image point to the same location in space, after being projected to 3D space by estimated depth, the two points should align in 3D space, as shown in the left image. However, due to inconsistencies in depth estimation across frames, projecting the same point into 3D space may lead to misalignment, as illustrated in the right image. This error can be decomposed into two losses (spatial loss and disparity loss in right image), which we aim to optimize.

Dynamic modeling

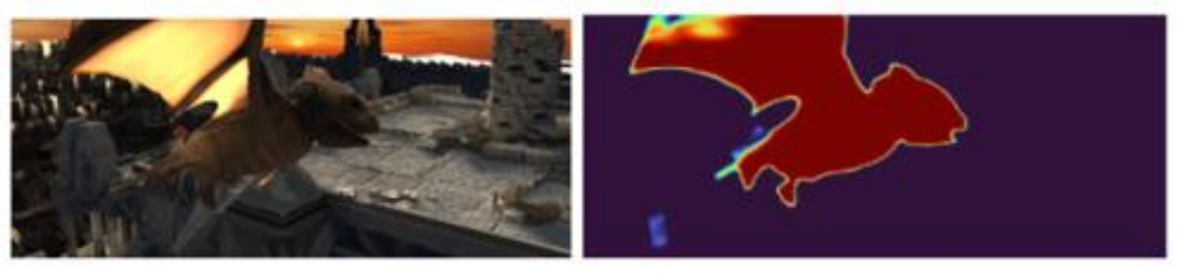

Figure 3: Movement estimation

We will train a CNN-based movement estimator to estimate moving objects, as illustrated in Figure 3. For dynamic regions, we utilize a motion trajectory field, as mentioned in [3], to estimate the scene flow, which models the displacement of 3D points. then supervise depth network and motion trajectory field with a modified reprojection loss, as depicted in Figure 4.

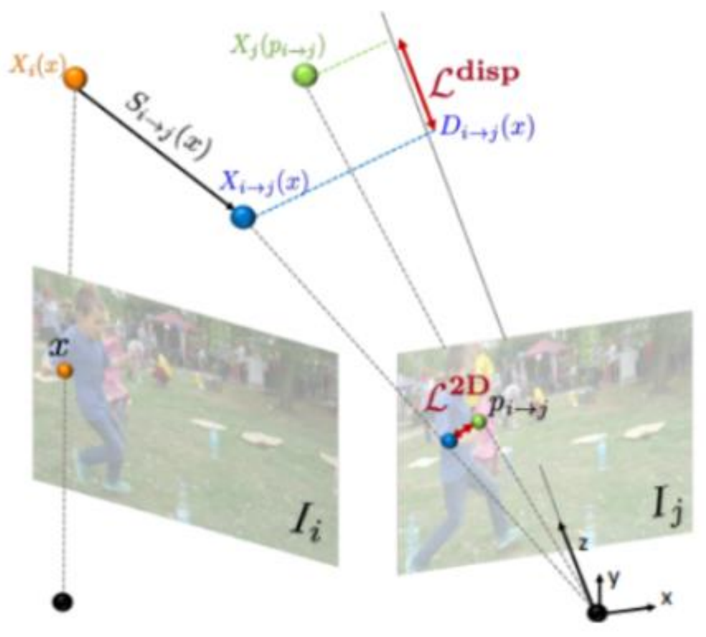

Figure 4: Reprojection Loss in dynamic region, mentioned in [4]

Result and Conclusion

Quantitative result

Our experiments utilize the MPI Sintel dataset [5], which contains 23 animated film segments. This dataset is valuable for evaluating the accuracy of our depth and camera pose estimations, as it provides ground truth depth and camera pose information.

Table 1: Quantitative result of MPI Sintel

Method	Depth Error, Rel, L1↓		
	Avg.	Dynamic	Static
casualSAM	**0.3637**	0.8129	0.3218
ours	0.3742	**0.7352**	0.3327

In Table 1, we compare our method with the state-of-the-art casualSAM [6] in depth estimation, our method exhibits slightly lower overall scores across the 23 videos of the Sintel dataset. However, we significantly outperform casualSAM in dynamic regions, owing to our effective computation of object motion trajectories, which underscores the strength of our motion trajectory field.

Qualitative result

Figures 5 showcase depth prediction results for specific video segments, with five frames selected from each. The displays include the original RGB image, initial depth by MiDAS, results from the previous casualSAM method, our results, and the ground truth. It's evident that MiDAS depth varies greatly across frames. Although casualSAM offers more consistency, its accuracy on moving objects, like the dragon's wings in Figure 3, is less precise. Our method effectively addresses these issues.

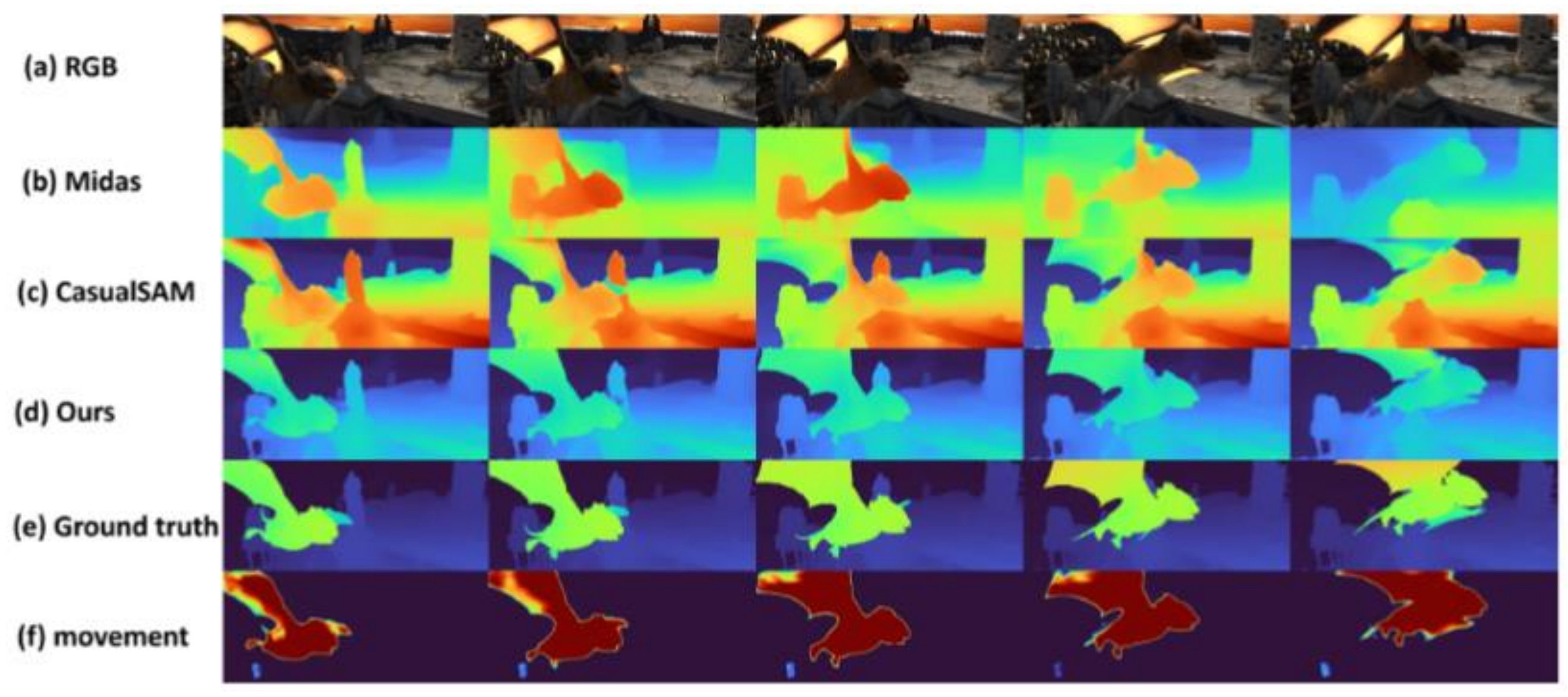

Figure 5: Result of temple_2 sequence in MPI Sintel dataset.

Figures 6 depict the results of camera pose estimation for specific video segments. The red line represents our result, while the black line indicates the ground truth. It's evident that our result highly overlaps with the ground truth.

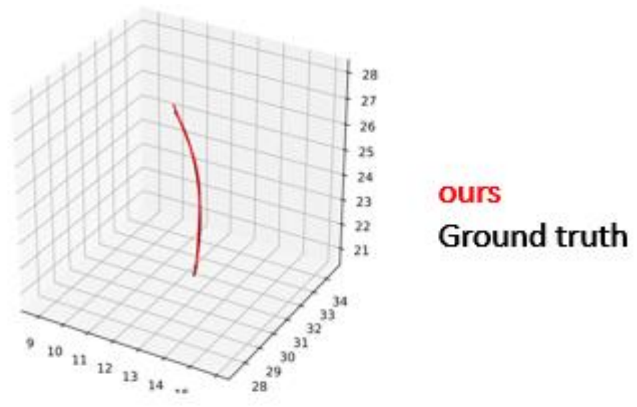

Figure 6: Camera pose of temple_2 sequence in MPI Sintel dataset.

Reference

[1] Ren ́e Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence, 44(3):1623–1637, 2020.
[2] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. ACM Transactions on Graphics (ToG), 39(4):71–1, 2020.
[3] Chaoyang Wang, Ben Eckart, Simon Lucey, and Orazio Gallo. Neural trajectory fields for dynamic novel view synthesis. arXiv preprint arXiv:2105.05994, 2021.
[4] Zhoutong Zhang, Forrester Cole, Richard Tucker, William T Freeman, and Tali Dekel. Consistent depth of moving objects in video. ACM Transactions on Graphics (TOG), 40(4):1–12, 2021.
[5] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12, pages 611–625. Springer, 2012.
[6] Zhoutong Zhang, Forrester Cole, Zhengqi Li, Michael Rubinstein, Noah Snavely, and William T Freeman. Structure and motion from casual videos. In European Conference on Computer Vision, pages 20–37. Springer, 2022.