Dynamo-Depth: Fixing Unsupervised Depth Estimation for Dynamical Scenes

Yihong Sun, Bharath Hariharan
Cornell University

Project Page

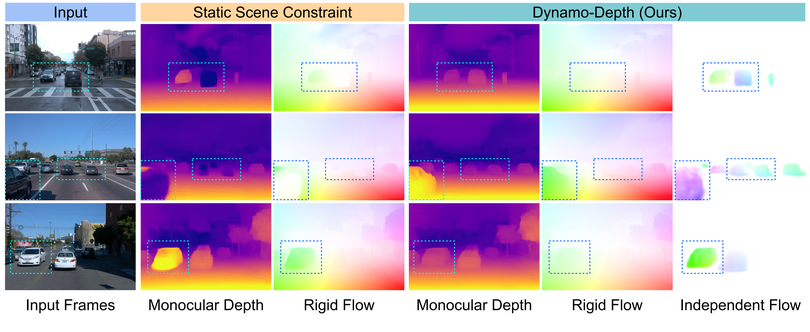
Introduction

Motivation

- To learn scene geometry, methods typically assume a static scene.
- When trained on dynamical scenes, independent motion cannot be adequately modeled which leads to erroneous depth predictions for moving objects.

Insight

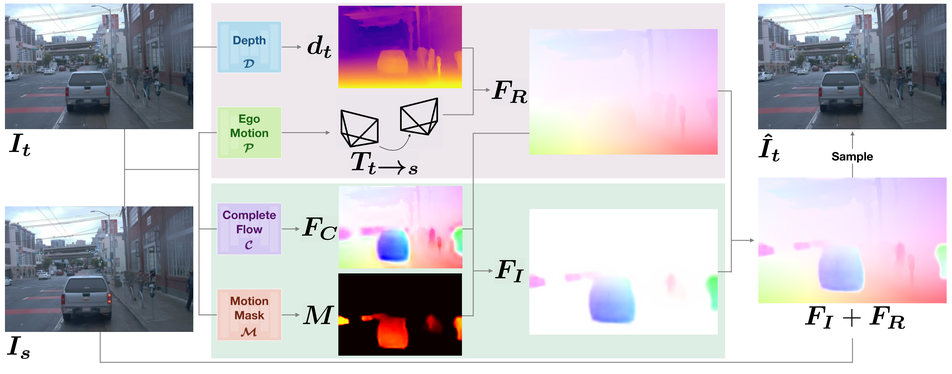
- A good initial estimation of motion segmentation is sufficient for jointly learning depth and independent motion despite the fundamental underlying ambiguity.



Contributions

- We propose **Dynamo-Depth** for learning depth, camera ego-motion, 3D independent flow, and motion segmentation for dynamic scenes solely from **unlabeled videos**.
- We model **independent motion** via the complete scene flow to **facilitate learning and regularization**.
- We learn an early estimation of motion segmentation to explicitly **disentangle independent motion from ego-motion**, which allows E2E training without additional supervision.
- We achieve **state-of-the-art** performance for depth estimation on nuScenes and Waymo Open Dataset with **large improvements** on moving objects.

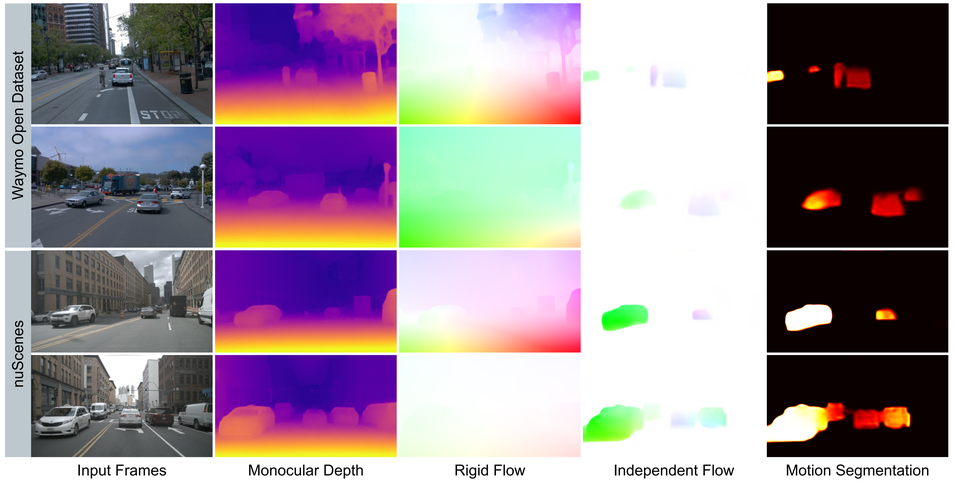
Methods



$$F_I = M \cdot (F_C - F_R) \qquad F_I + F_R = M \cdot F_C + (1 - M) \cdot F_R$$

- **Overview.** We compose the 3D rigid flow F_R and 3D independent flow F_I to reconstruct target frame I_t.

Qualitative Results



Quantitative Results

	D	Abs Rel (↓)				$\delta < 1.25$ (↑)			
		All	*S.B.*	*S.O.*	*M.O.*	*All*	*S.B.*	*S.O.*	*M.O.*
Monodepth2	N	0.425	0.447	0.232	0.418	0.723	0.735	0.704	0.570
Dynamo-Depth (MD2)	N	0.193	0.196	0.196	0.228	0.765	0.761	0.759	0.684
Δ (%)		**-54.6**	**-56.2**	**-15.5**	**-45.5**	**+5.8**	**+3.5**	**+7.8**	**+20.0**
LiteMono	N	0.419	0.431	0.248	0.502	0.720	0.734	0.718	0.517
Dynamo-Depth	N	0.179	0.184	0.182	0.198	0.787	0.781	0.774	0.753
Δ (%)		**-57.3**	**-57.3**	**-26.6**	**-60.6**	**+9.3**	**+6.4**	**+7.8**	**+45.6**
Monodepth2	W	0.173	0.152	0.215	0.749	0.797	0.810	0.683	0.416
Dynamo-Depth (MD2)	W	0.130	0.122	0.175	0.234	0.851	0.862	0.778	0.674
Δ (%)		**-24.9**	**-19.7**	**-18.6**	**-68.8**	**+6.8**	**+6.4**	**+13.9**	**+62.0**
LiteMono	W	0.158	0.140	0.179	0.599	0.816	0.827	0.733	0.506
Dynamo-Depth	W	0.116	0.110	0.155	0.194	0.878	0.891	0.812	0.750
Δ (%)		**-26.6**	**-21.4**	**-13.4**	**-67.6**	**+7.6**	**+7.7**	**+10.8**	**+48.2**

- **Depth performance** with semantic split on the nuScenes (N) and Waymo Open (W) Dataset. *S.B.*, *S.O.* and *M.O.* denotes the pixels that are static background, static moveable object and moving object, respectively.

Conclusions

Compared to Monodepth2 [1] and Litemono [2], from which we adopt our depth network from, our proposed Dynamo-Depth

1) outperforms respective baselines across all metrics on Waymo Open and nuScenes;
2) significantly improves depth performance on the moving objects;
3) and accurately estimates the 3D independent motion field and motion segmentation.

References

[1] Godard, Clément, et al. "Digging into self-supervised monocular depth estimation." in *ICCV* 2019.

[2] Zhang, Ning, et al. "Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation." in *CVPR* 2023.

Acknowledgements

- This research is based upon work supported in part by the National Science Foundation (IIS-2144117 and IIS-2107161).
- Yihong Sun is supported in part by an NSF graduate research fellowship.