Self-Supervised Pillar Motion Learning for Autonomous Driving

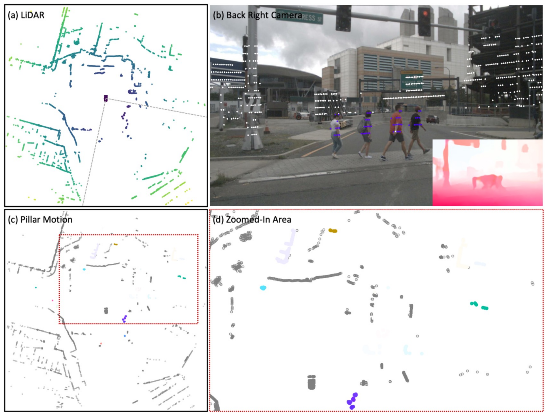
Chenxu Luo[1,2] Xiaodong Yang[1] Alan Yuille[2]
[1]QCraft [2]Johns Hopkins University

CVPR VIRTUAL JUNE 19-25

Introduction

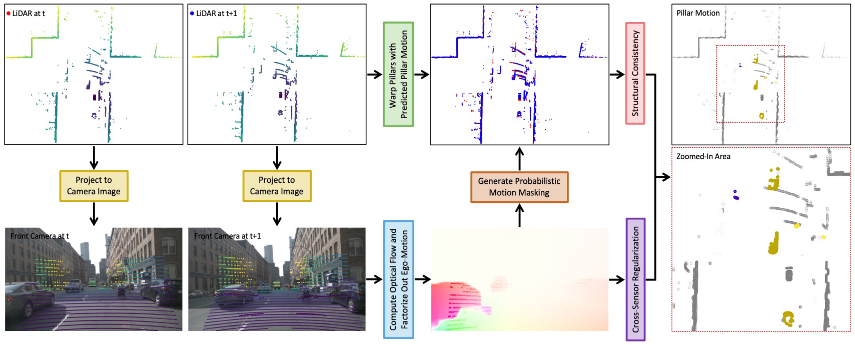
- Motion provides pivotal information to facilitate various self-driving modules.
- Motion should be estimated in an open-set setting irrespective of object classes.
- Self-supervised learning opens the possibility to utilize infinite training data that is continuously collected by the world-wide self-driving fleets.

Overview



(a) illustrates a point cloud in BEV. (b) shows the projected points with color encoding optical flow (ego-motion factorized out) on the back right camera image. Note that the white points are static. Original optical flow is attached for reference. (c) is the predicted pillar motion filed. (d) demonstrates a zoomed-in area of (c).

Method


LiDAR at t
LiDAR at t+1
Warp Pillars with Predicted Pillar Motion
Structural Consistency
Pillar Motion
Project to Camera Image
Project to Camera Image
Generate Probabilistic Motion Masking
Zoomed-In Area
Front Camera at t
Front Camera at t+1
Compute Optical Flow and Factorize Out Ego-Motion
Cross-Sensor Regularization

A schematic overview of the proposed self-supervised learning framework for pillar motion prediction.

- LiDAR based Structural Consistency $\mathcal{L}_{\text{consist}} = \sum_{\tilde{P}_i \in \tilde{\mathcal{P}}} \min_{P_j \in \mathcal{P}} \|\tilde{P}_i - P_j\| + \sum_{P_j \in \mathcal{P}} \min_{\tilde{P}_i \in \tilde{\mathcal{P}}} \|P_j - \tilde{P}_i\|$

- Cross-Sensor Motion Regularization $\mathcal{L}_{\text{regular}} = \sum_i \|\tilde{F}^t(u_i, v_i) - F^t_{\text{obj}}(u_i, v_i)\|_1$

$$F^t(u,v) = F^t_{\text{ego}}(u,v) + F^t_{\text{obj}}(u,v) \quad F^t_{\text{ego}}(u_i, v_i) = KT_{\text{L}\to\text{C}}T_{t\to t+1}P^t_i - KT_{\text{L}\to\text{C}}P^t_i$$

- Probabilistic Motion Masking

$$s^t_i = \exp\{-\alpha \max(\|F^t_{\text{obj}}(u_i, v_i)\| - \tau, 0)\}$$

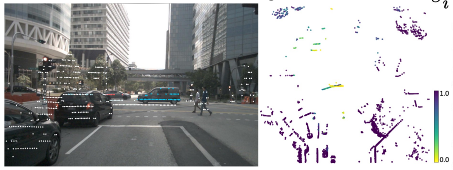
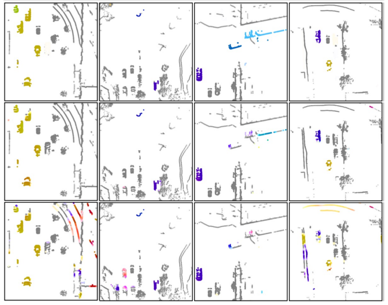
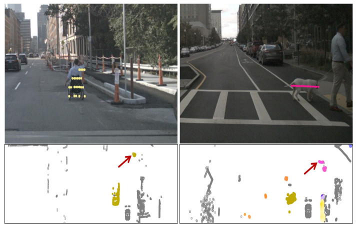


Illustration of the probabilistic motion masking.

Results



Top: ground truth pillar motion. Middle: our full model. Bottom: base model using only structural consistency.



Examples of perceiving rare objects (wheelchair and dog) by pillar motion.

	$\mathcal{L}_{\text{consist}}$	$\mathcal{L}_{\text{regular}}$	Mask	Static		Speed ≤ 5m/s		Speed > 5m/s		Nonempty		Foreground		Moving	
				Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median
(a)	✓			0.3701	0.0063	0.5014	0.1352	1.9405	1.2760	0.3437	0.0081	0.5936	0.1139	0.7516	0.2359
(b)		✓		**0.0285**	**0.0002**	0.3733	**0.0719**	4.2954	3.9788	0.0897	0.0020	0.7914	0.0656	1.1267	0.3948
(c)	✓	✓		0.1688	0.0389	0.4277	0.1694	1.7603	1.2021	0.3133	0.0062	0.5667	0.1017	0.7064	0.1980
(d)	✓		✓	0.0738	0.0038	0.4017	0.1214	1.9384	1.2931	0.1085	0.0007	0.5416	0.0767	0.8064	0.2279
(e)	✓	✓	✓	0.0619	0.0004	**0.3438**	0.1196	**1.7119**	**1.1438**	**0.0846**	**0.0001**	**0.4494**	**0.0507**	**0.5953**	**0.1612**

Comparison of our models using different combinations of the proposed self-supervision components.