Towards Scale-Aware, Robust, and Generalizable Unsupervised Monocular Depth Estimation by Integrating IMU Dynamics

Sen Zhang, Jing Zhang, Dacheng Tao
The University of Sydney, Australia

Motivation:

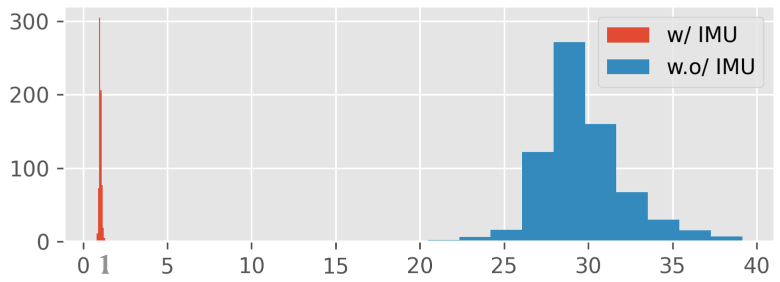
Intrinsic problems of of current deep learning-based unsupervised monocular depth estimation methods:

- (a) Scale ambiguity: The backwarping process is equivalent up to an arbitrary scaling factor w.r.t. depth and translation;
- (b) Robustness: The photometric error is sensitive to illumination change and moving objects;
- (c) Generalizability: The learned neural network may overfit to the training dataset.

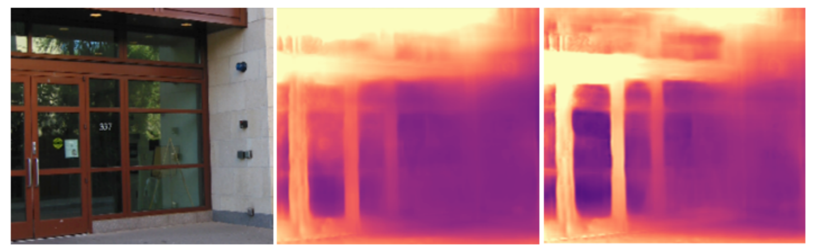
IMU as a remedy. IMU presents a commonly-deployed sensor in modern sensor suites on vehicles that is advantageous in that:

- (a) The absolute scale metric can be recovered by inquiring the IMU motion dynamics;
- (b) It is robust to the scenarios when vision fails such as in illumination-changing and textureless regions;
- (c) It does not suffer from the visual domain gap, leading to a better generalization ability across datasets.

(b) Histograms of the Scaling Ratios



(c) Generalization Results on Make3D



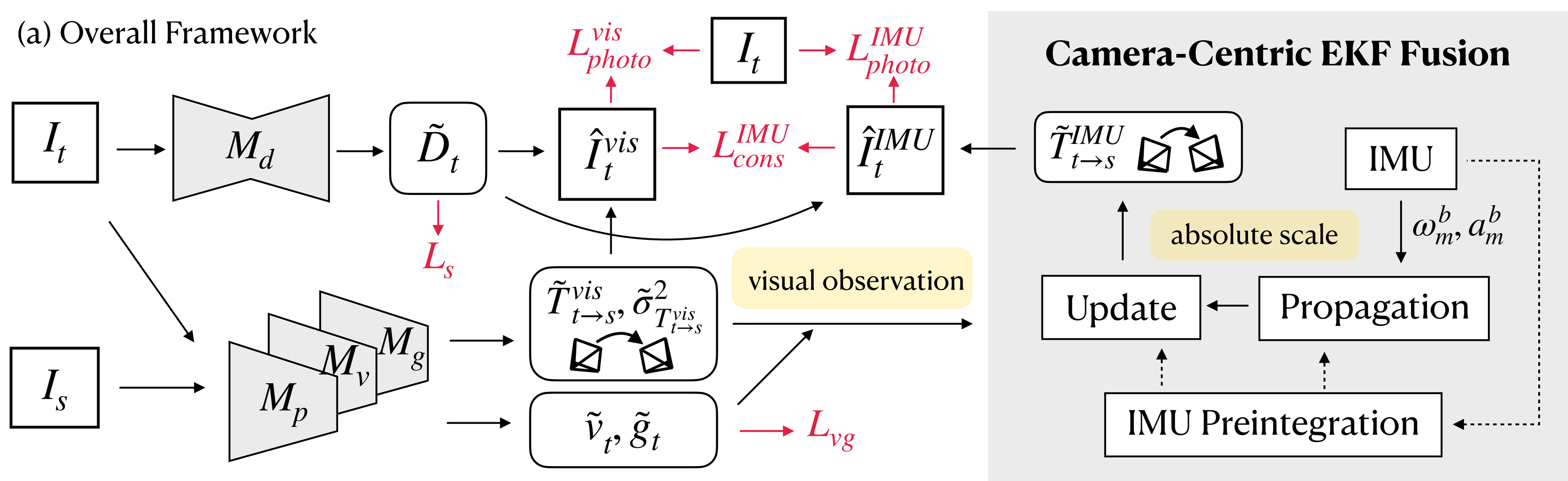
Input Image w.o/ IMU w/ IMU

Methodology

Overall Framework of DynaDepth

- (a) A scale-aware IMU photometric loss and a cross-sensor photometric consistency loss are proposed to introduce IMU motion dynamics into the system;
- (b) The IMU preintegration technique is adopted to avoid redundant computation to accelerate the training process;
- (c) A differentiable camera-centric EKF framework is derived to facilitate the fusion of vision and IMU information.

(a) Overall Framework



Methodology:

Camera-Centric IMU Preintegration

- We derive the IMU preintegration formula at the camera frame to ease training since the losses are defined on image appearance.
- We predict the velocity and the gravity vectors directly from images using neural networks to avoid the complicated initialization step that is commonly used in classical methods.

$$\boldsymbol{R}_{c_k \check{c}_{k+1}} = \boldsymbol{R}_{cb}\mathcal{F}^{-1}(\boldsymbol{q}_{b_k b_{k+1}})\boldsymbol{R}_{bc},$$
$$\boldsymbol{p}_{c_k \check{c}_{k+1}} = \boldsymbol{R}_{cb}\boldsymbol{\alpha}_{b_k b_{k+1}} + \boldsymbol{R}_{c_k \check{c}_{k+1}}\boldsymbol{R}_{cb}\boldsymbol{p}_{bc} - \boldsymbol{R}_{cb}\boldsymbol{p}_{bc} + \boldsymbol{v}^{\tilde{c}_k}\Delta t_k - \frac{1}{2}\boldsymbol{g}^{\tilde{c}_k}\Delta t_k^2,$$

IMU-Related Losses

- **IMU Photometric Loss**: This loss is proposed to provide dense supervisory signals for both the depth and the ego-motion networks

$$L^{IMU}_{photo} = \frac{1}{N}\sum_{i=1}^{N}\min_{\delta\in\{-1,1\}} \mathcal{L}(\boldsymbol{I}(\boldsymbol{y}_i), \boldsymbol{I}_\delta(\psi(\boldsymbol{K}\hat{\boldsymbol{R}}_\delta \boldsymbol{K}^{-1}\boldsymbol{y}_i + \frac{\boldsymbol{K}\hat{\boldsymbol{p}}_\delta}{\tilde{z}_i}))),$$

$$\mathcal{L}(\boldsymbol{I}, \boldsymbol{I}_\delta) = \alpha\frac{1 - SSIM(\boldsymbol{I}, \boldsymbol{I}_\delta)}{2} + (1-\alpha)||\boldsymbol{I} - \boldsymbol{I}_\delta||_1,$$

- **Cross-Sensor Photometric Consistency Loss**: This loss is proposed to align the ego-motions predicted from IMU and vision.

$$L^{cons}_{photo} = \frac{1}{N}\sum_{i=1}^{N}\min_{\delta\in\{-1,1\}} \mathcal{L}(\boldsymbol{I}_\delta(\psi(\boldsymbol{K}\tilde{\boldsymbol{R}}_\delta \boldsymbol{K}^{-1}\boldsymbol{y}_i + \frac{\boldsymbol{K}\tilde{\boldsymbol{p}}_\delta}{\tilde{z}_i})), \boldsymbol{I}_\delta(\psi(\boldsymbol{K}\hat{\boldsymbol{R}}_\delta \boldsymbol{K}^{-1}\boldsymbol{y}_i + \frac{\boldsymbol{K}\hat{\boldsymbol{p}}_\delta}{\tilde{z}_i}))),$$

The Camera-Centric EKF Framework

We derive the EKF framework at the camera frame to facilitate the learning process since the training process takes images as input.

- **EKF Propagation**: We separate the states into the nominal states and the error states, where the continuous-time propagation model for the error states is derived as: $\delta\dot{\boldsymbol{x}}_{\boldsymbol{b_t}} = \boldsymbol{F}\delta\boldsymbol{x}_{\boldsymbol{b_t}} + \boldsymbol{Gn}$

$$\boldsymbol{F} = \begin{bmatrix} -[\bar{\boldsymbol{w}}^{b_t}]^\wedge & 0 & 0 & 0 & -\boldsymbol{I}_3 & 0 \\ 0 & 0 & \boldsymbol{I}_3 & 0 & 0 & 0 \\ -\bar{\boldsymbol{R}}_{c_k b_t}[\bar{\boldsymbol{R}}^T_{c_k b_t}\bar{\boldsymbol{g}}^{c_k} + \bar{\boldsymbol{a}}^{b_t}]^\wedge & 0 & 0 & -\boldsymbol{I}_3 & 0 & -\bar{\boldsymbol{R}}_{c_k b_t} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \boldsymbol{G} = \begin{bmatrix} -\boldsymbol{I}_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -\bar{\boldsymbol{R}}_{c_k b_t} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & \boldsymbol{I}_3 & 0 & 0 \\ 0 & 0 & 0 & \boldsymbol{I}_3 \end{bmatrix}$$

- **EKF Update**: The observation model and the corresponding Jacobian w.r.t. the error states are derived as:

$$h(\check{\boldsymbol{x}}_{k+1}) = \begin{bmatrix} \bar{\boldsymbol{\phi}}_{c_k c_{k+1}} \\ \bar{\boldsymbol{R}}_{c_k b_{k+1}}\boldsymbol{p}_{bc} + \bar{\boldsymbol{p}}_{c_k b_{k+1}} \end{bmatrix}, \boldsymbol{H}_{k+1} = \begin{bmatrix} J_l(-\bar{\boldsymbol{\phi}}_{c_k c_{k+1}})^{-1}\boldsymbol{R}_{cb} & 0 & 0 & 0 & 0 & 0 \\ -\bar{\boldsymbol{R}}_{c_k b_{k+1}}[\boldsymbol{p}_{bc}]^\wedge & \boldsymbol{I}_3 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Experiment:

- Experiment results show that our proposed DynaDepth is scale-aware, robust, and generalizable compared with other methods.
- Ablation studies validate the effectiveness of each components.
- Please refer to our paper for more details on our experiment.

Table 1: Per-image rescaled depth evaluation on KITTI using the Eigen split. The best and the second best results are shown in **bold** and underline. † denotes our reproduced results. Results are rescaled using the median ground-truth from Lidar. The means and standard errors of the scaling ratios are reported in Scale.

Methods	Year	Scale	Error↓				Accuracy↑		
			AbsRel	SqRel	RMSE	$RMSE_{log}$	$\sigma < 1.25$	$\sigma < 1.25^2$	$\sigma < 1.25^3$
Monodepth2 R18 [14]	ICCV 2019	NA	0.112	0.851	4.754	0.190	0.881	0.960	0.981
Monodepth2 R50† [14]	ICCV 2019	29.128±0.084	0.111	0.806	4.642	0.189	0.882	**0.962**	**0.982**
PackNet-SfM [15]	CVPR 2020	NA	0.111	0.785	**4.601**	0.189	0.878	0.960	**0.982**
Johnston R18 [18]	CVPR 2020	NA	0.111	0.941	4.817	0.189	**0.885**	0.961	0.981
R-MSFM6 [49]	ICCV 2021	NA	0.112	0.806	4.704	0.191	0.878	0.960	0.981
G2S R50 [3]	ICRA 2021	1.031±0.073	0.112	0.894	4.852	0.192	0.877	0.958	0.981
ScaleInvariant R18 [38]	ICCV 2021	NA	0.109	0.779	4.641	**0.186**	0.883	**0.962**	**0.982**
DynaDepth R18	2022	1.021±**0.069**	0.111	0.806	4.777	0.190	0.878	0.960	**0.982**
DynaDepth R50	2022	**1.013**±0.071	**0.108**	**0.761**	4.608	0.187	0.883	**0.962**	**0.982**

Table 5: Ablation results of the robustness against vision degradation on the simulated data from KITTI. The best results are shown in **bold**. IC and MO denote the two investigated vision degradation types, i.e., illumination change and moving objects. - means item not available. † denotes our reproduced results.

Methods	EKF	L_{vg}	Type	Scale	Error↓				Accuracy↑		
					AbsRel	SqRel	RMSE	$RMSE_{log}$	$\sigma < 1.25$	$\sigma < 1.25^2$	$\sigma < 1.25^3$
Monodepth2† [14]	-	-	IC	27.701±0.096	0.127	0.976	5.019	0.220	0.855	0.946	0.972
DynaDepth			IC	1.036±0.099	0.124	**0.858**	4.915	0.226	0.852	0.950	0.977
DynaDepth	✓		IC	0.946±0.089	0.123	0.925	**4.866**	**0.196**	**0.863**	**0.957**	**0.981**
DynaDepth	✓	✓	IC	**1.019±0.074**	**0.121**	0.906	4.950	0.217	0.859	0.954	0.978
Monodepth2† [14]	-	-	MO	0.291±0.176	0.257	2.493	8.670	0.398	0.584	0.801	0.897
DynaDepth			MO	0.083±0.225	0.169	1.290	6.030	0.278	0.763	0.915	0.960
DynaDepth	✓		MO	0.087±0.119	0.126	**0.861**	5.312	**0.210**	0.840	0.948	**0.979**
DynaDepth	✓	✓	MO	**0.956±0.084**	**0.125**	0.926	**4.954**	0.214	**0.852**	**0.949**	0.976

Conclusion

- We propose DynaDepth, a scale-aware, robust, and generalizable monocular depth estimation framework by integrating IMU motion dynamics.
- Code link: https://github.com/SenZHANG-GitHub/ekf-imu-depth. For any questions, feel free to contact us via email : szha2609@uni.sydney.edu.au