Visual Computing Institute
Computer Vision

RWTH AACHEN UNIVERSITY

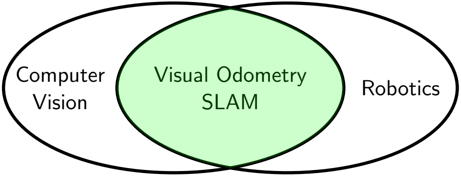
Unsupervised Learning of Depth and Ego-Motion from Monocular Video Using 3D Geometric Constraints

Nikolay Paleshnikov
Advisor: Aljoša Ošep

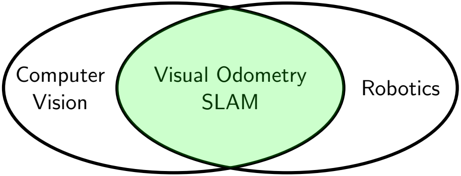
RWTH Aachen
August 02, 2018

Roadmap

Monocular Visual Odometry and SLAM



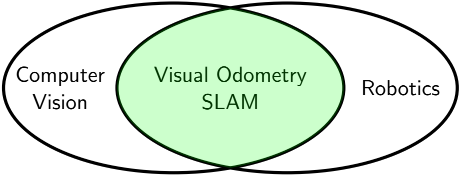
Monocular Visual Odometry and SLAM

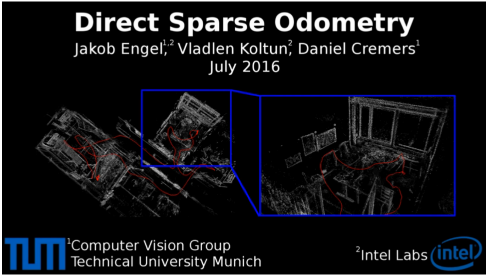


Main applications

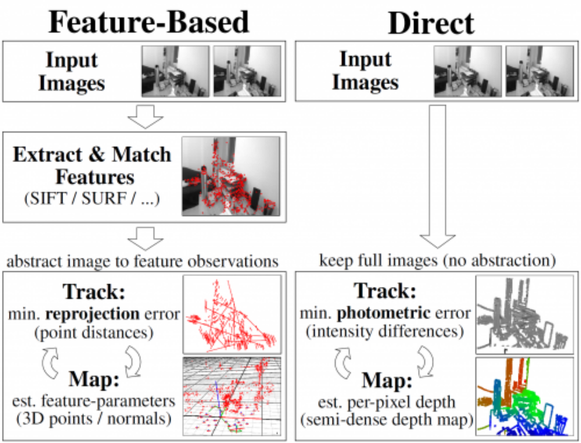
- Virtual and augmented reality
- Unknown surface exploration
- Autonomous navigation

Monocular Visual Odometry and SLAM





Analytic Visual Frameworks



[Engel et al., ECCV 2014]

Roadmap

Euclidean Transformations

Rotation $R \in \mathbb{R}^{3\times 3}$, followed by a translation $t \in \mathbb{R}^3$

$$T = \begin{pmatrix} & R & & t \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Euclidean Transformations

Rotation $R \in \mathbb{R}^{3\times3}$, followed by a translation $t \in \mathbb{R}^3$

$$T = \begin{pmatrix} & R & & t \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$p = \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} \in \mathbb{R}^4$$

$$M_T : \mathbb{R}^4 \to \mathbb{R}^4 \ , \ p \mapsto T \ p$$

Euclidean Transformations

Rotation $R \in \mathbb{R}^{3\times 3}$, followed by a translation $t \in \mathbb{R}^3$

$$T = \begin{pmatrix} & R & & t \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$p = \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} \in \mathbb{R}^4$$

$$M_T : \mathbb{R}^4 \to \mathbb{R}^4 \ , \ p \mapsto T \ p$$

$$T_i^j = T_j^{-1} \ T_i$$

Point Reprojection

$$\Pi : \mathbb{R}^4 \to \mathbb{R}^2 \,,\ \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} \mapsto \begin{pmatrix} x/z \\ y/z \end{pmatrix}$$

$$\Pi^{-1} : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}^4 \,,\ (\begin{pmatrix} u \\ v \end{pmatrix}, d_p) \mapsto \begin{pmatrix} u/d_p \\ v/d_p \\ 1/d_p \\ 1 \end{pmatrix}$$

Point Reprojection

$$\Pi : \mathbb{R}^4 \rightarrow \mathbb{R}^2 \; , \; \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} \mapsto \begin{pmatrix} x/z \\ y/z \end{pmatrix}$$

$$\Pi^{-1} : \mathbb{R}^2 \times \mathbb{R} \rightarrow \mathbb{R}^4 \; , \; \left(\begin{pmatrix} u \\ v \end{pmatrix}, d_p\right) \mapsto \begin{pmatrix} u/d_p \\ v/d_p \\ 1/d_p \\ 1 \end{pmatrix}$$

$$K = \begin{pmatrix} f_x & 0 & c_x & 0 \\ 0 & f_y & c_y & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Point Reprojection

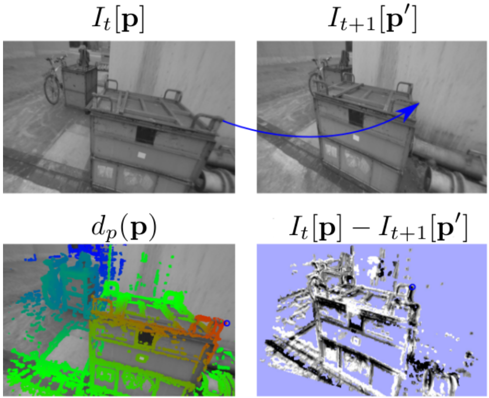
$$\Pi : \mathbb{R}^4 \to \mathbb{R}^2 \,, \quad \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} \mapsto \begin{pmatrix} x/z \\ y/z \end{pmatrix}$$

$$\Pi^{-1} : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}^4 \,, \; (\begin{pmatrix} u \\ v \end{pmatrix}, d_p) \mapsto \begin{pmatrix} u/d_p \\ v/d_p \\ 1/d_p \\ 1 \end{pmatrix}$$

$$K = \begin{pmatrix} f_x & 0 & c_x & 0 \\ 0 & f_y & c_y & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\mathbf{p}' = \Pi\left(K \; T_i^j \; K^{-1} \; \Pi^{-1}\left(\mathbf{p}, d_p\right)\right)$$

Point Reprojection



$$\mathbf{p}' = \Pi\left(K\ T_i^j\ K^{-1}\ \Pi^{-1}\left(\mathbf{p}, d_p\right)\right)$$

Bundle Adjustment

$$E_{total} = \sum_{i \in F} \sum_{\mathbf{p}^* \in sp(i)} \|\mathbf{p}^* - \mathbf{p}\|_2^2$$

Bundle Adjustment

$$E_{total} = \sum_{i \in F} \sum_{\mathbf{p}^* \in sp(i)} \|\mathbf{p}^* - \mathbf{p}\|_2^2$$

$$\underset{P, \{T_i | i \in F\}, K}{\operatorname{argmin}} \quad E_{total}$$

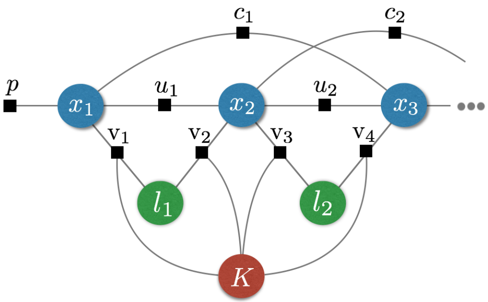
Bundle Adjustment

$$E_{total} = \sum_{i \in F} \sum_{\mathbf{p}^* \in sp(i)} \|\mathbf{p}^* - \mathbf{p}\|_2^2$$

$$\underset{P, \{T_i | i \in F\}, K}{\operatorname{argmin}} \quad E_{total}$$

- Maximum a Posteriori (MAP) estimation
- Non-linear non-convex least-squares optimization problem
- Good initialization required

SLAM Represented as a Factor Graph



robot poses landmarks
camera matrix ■ factors
[Cadena et al., IEEE T-RO 2016]

Probabilistic Interpretation of SLAM

$$\underset{X}{\operatorname{argmax}}\, p(X|Z) = \underset{X}{\operatorname{argmax}}\, p(Z|X)p(X)$$

Probabilistic Interpretation of SLAM

$$\operatorname*{argmax}_X p(X|Z) = \operatorname*{argmax}_X p(Z|X)p(X)$$

$$\operatorname*{argmax}_X p_0\ p(Z|X) = \operatorname*{argmax}_X p_0\ \prod_{i=1}^{n} p(z_i|X) = \operatorname*{argmax}_X p_0\ \prod_{i=1}^{n} p(z_i|X_i)$$

Probabilistic Interpretation of SLAM

$$\underset{X}{\operatorname{argmax}}\, p(X|Z) = \underset{X}{\operatorname{argmax}}\, p(Z|X)p(X)$$

$$\underset{X}{\operatorname{argmax}}\, p_0 \; p(Z|X) = \underset{X}{\operatorname{argmax}}\, p_0 \; \prod_{i=1}^{n} p(z_i|X) = \underset{X}{\operatorname{argmax}}\, p_0 \; \prod_{i=1}^{n} p(z_i|X_i)$$

$$p(z_i|X_i) \propto exp(-\frac{1}{2}\|h_i(X_i) - z_i\|^2_{\Omega_i})$$

Probabilistic Interpretation of SLAM

$$\underset{X}{\operatorname{argmax}}\, p(X|Z) = \underset{X}{\operatorname{argmax}}\, p(Z|X)p(X)$$

$$\underset{X}{\operatorname{argmax}}\, p_0 \; p(Z|X) = \underset{X}{\operatorname{argmax}}\, p_0 \; \prod_{i=1}^{n} p(z_i|X) = \underset{X}{\operatorname{argmax}}\, p_0 \; \prod_{i=1}^{n} p(z_i|X_i)$$

$$p(z_i|X_i) \propto exp(-\frac{1}{2}\|h_i(X_i) - z_i\|^2_{\Omega_i})$$

$$\underset{X}{\operatorname{argmin}} -ln\;(p_0 \; \prod_{i=1}^{n} p(z_i|X_i)) = \underset{X}{\operatorname{argmin}} \sum_{i=1}^{n} \|h_i(X_i) - z_i\|^2_{\Omega_i}$$

Probabilistic Interpretation of SLAM

$$\operatorname*{argmax}_X p(X|Z) = \operatorname*{argmax}_X p(Z|X)p(X)$$

$$\operatorname*{argmax}_X p_0 \; p(Z|X) = \operatorname*{argmax}_X p_0 \; \prod_{i=1}^{n} p(z_i|X) = \operatorname*{argmax}_X p_0 \; \prod_{i=1}^{n} p(z_i|X_i)$$

$$p(z_i|X_i) \propto exp(-\frac{1}{2}\|h_i(X_i) - z_i\|^2_{\Omega_i})$$

$$\operatorname*{argmin}_X -ln \; (p_0 \; \prod_{i=1}^{n} p(z_i|X_i)) = \operatorname*{argmin}_X \sum_{i=1}^{n} \|h_i(X_i) - z_i\|^2_{\Omega_i}$$

Advantages in comparison with bundle adjustment:

- Simultaneous incorporation of various sensors
- Incremental solution possible

Visual SLAM

Visual SLAM rendered possible by means of a simplified scene model:

- Rigid Lambertian world
- Temporal coherence and constant illumination
- Pinhole camera model and epipolar geometry
  $\Rightarrow$ 6 degrees of freedom for motion, 1 for depth

Visual SLAM

Visual SLAM rendered possible by means of a simplified scene model:

- Rigid Lambertian world
- Temporal coherence and constant illumination
- Pinhole camera model and epipolar geometry
  $\Rightarrow$ 6 degrees of freedom for motion, 1 for depth

Open problems

- Life-long operation
- High-level geometry understanding
- Resilience in a variety of environments

Deep Learning for SfM

Paradigm shift from analytic to statistical solutions

Self-supervision: no explicit labels, geometric consistency

Deep Learning for SfM

Paradigm shift from analytic to statistical solutions

Self-supervision: no explicit labels, geometric consistency

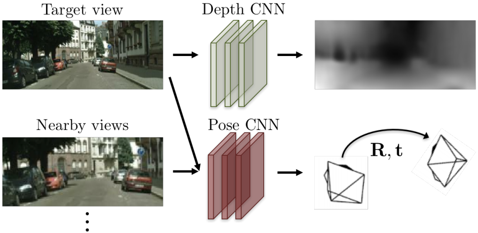
Success stories

- Monocular depth estimation
  (Garg et al., Godard et al., Kuznietsov et al.)
- Joint monocular depth and ego-motion estimation
  (Zhou et al., Vijayanarasimhan et al.)
- Rigid body detection and motion tracking
  (Byravan et al.)

Roadmap

SfMLearner Overview

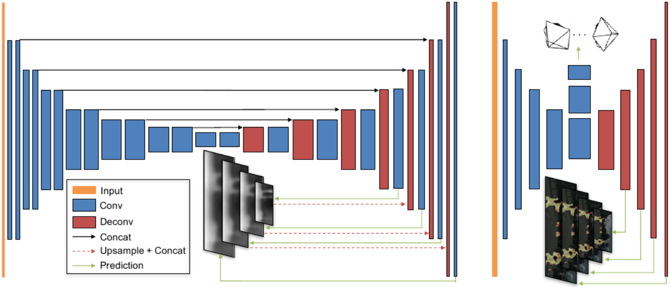
(a) Training: unlabeled video clips.



(b) Testing: single-view depth and multi-view pose estimation.

[Zhou et al., CVPR 2017]

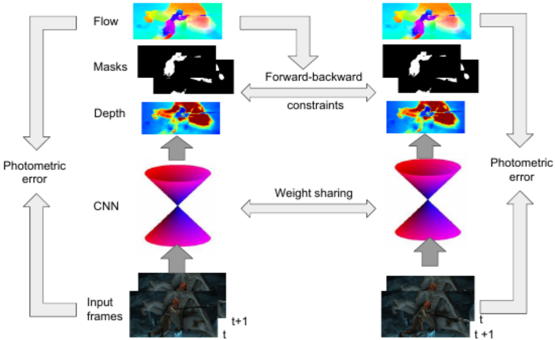
SfMLearner Network



(a) Single-view depth network (b) Pose/explainability network

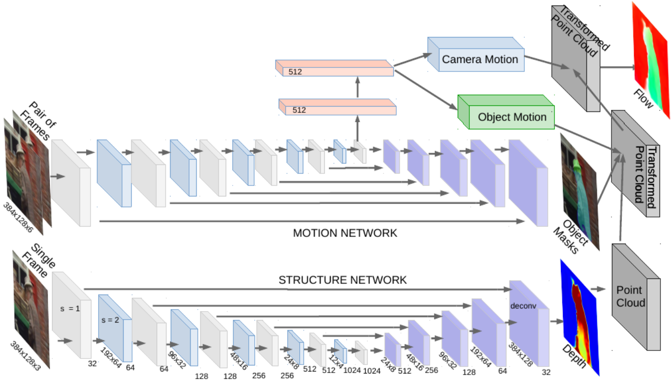
[Zhou et al., CVPR 2017]

SfM-Net Motion Subnetwork



[Vijayanarasimhan et al., ArXiv 2017]

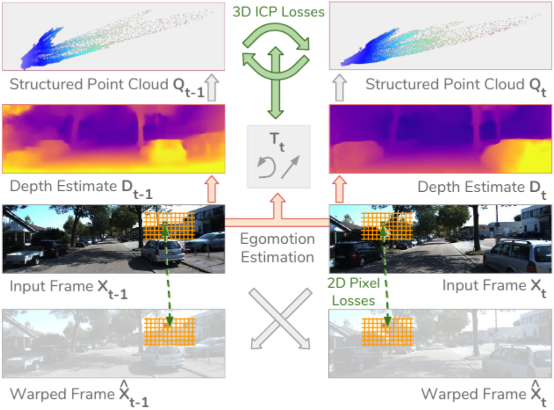
SfM-Net Architecture



[Vijayanarasimhan, arXiv 2017]

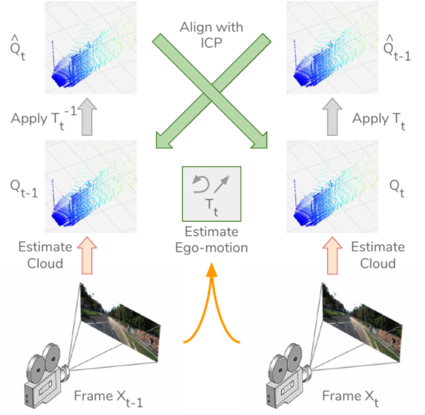
Roadmap

Method Overview



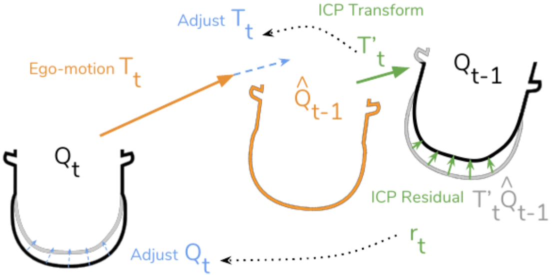
[Mahjourian et al., CVPR 2018]

Three-Dimensional Point Cloud Alignment



[Mahjourian et al., CVPR 2018]

Three-Dimensional Point Cloud Alignment



[Mahjourian et al., CVPR 2018]

Three-Dimensional Geometry Loss Term

Point cloud of frame i at time t:

$$Q_t^i = \{M_i[\mathbf{p}_t^i](K^{-1}\Pi^{-1}(\mathbf{p}_t^i, d_i(\mathbf{p}_t^i)))|\mathbf{p}_t^i \in P\}$$

Warped towards the next local coordinate frame $i+1$:

$$Q_{t-1}^{i+1} = T_t Q_t^i$$

Three-Dimensional Geometry Loss Term

Point cloud of frame i at time t:

$$Q_t^i = \{M_i[\mathbf{p}_t^i](K^{-1}\Pi^{-1}(\mathbf{p}_t^i, d_i(\mathbf{p}_t^i)))|\mathbf{p}_t^i \in P\}$$

Warped towards the next local coordinate frame $i+1$:

$$Q_{t-1}^{i+1} = T_t Q_t^i$$

Objective function of the Iterative Closest Point (ICP) algorithm used for point cloud alignment:

$$\underset{T'}{\operatorname{argmin}} \frac{1}{2} \| T' Q_{t-1}^{i+1} - Q_t^{i+1} \|_2^2$$

Three-Dimensional Geometry Loss Term

Point cloud of frame i at time t:

$$Q_t^i = \{M_i[\mathbf{p}_t^i](K^{-1}\Pi^{-1}(\mathbf{p}_t^i, d_i(\mathbf{p}_t^i)))|\mathbf{p}_t^i \in P\}$$

Warped towards the next local coordinate frame $i+1$:

$$Q_{t-1}^{i+1} = T_t Q_t^i$$

Objective function of the Iterative Closest Point (ICP) algorithm used for point cloud alignment:

$$\operatorname*{argmin}_{T'} \frac{1}{2} \| T' Q_{t-1}^{i+1} - Q_t^{i+1} \|_2^2$$

Loss term:

$$L_{3D} = |T_t' - I| + |r_t|$$

Differentiable Loss Function

Photometric consistency term:

$$L_{ph} = \sum_{\mathbf{p}_t^i \in P} \| M_i[\mathbf{p}_t^i](I_t[\mathbf{p}_t^i] - I_{t+1}[\mathbf{p}_{t+1}^i]) \|_2$$

Differentiable Loss Function

Photometric consistency term:

$$L_{ph} = \sum_{\mathbf{p}_t^i \in P} \| M_i[\mathbf{p}_t^i](I_t[\mathbf{p}_t^i] - I_{t+1}[\mathbf{p}_{t+1}^i]) \|_2$$

Structured similarity term:

$$SSIM(x, y) = \frac{(2\mu_x \mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x + \sigma_y + c_2)}$$

$$L_{SSIM} = \sum_{\mathbf{p}_t^i \in P} M_i[\mathbf{p}_t^i](1 - SSIM(I_t[\mathbf{p}_t^i], I_{t+1}[\mathbf{p}_{t+1}^i]))$$

Differentiable Loss Function

Photometric consistency term:

$$L_{ph} = \sum_{\mathbf{p}_t^i \in P} \| M_i[\mathbf{p}_t^i](I_t[\mathbf{p}_t^i] - I_{t+1}[\mathbf{p}_{t+1}^i]) \|_2$$

Structured similarity term:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x + \sigma_y + c_2)}$$

$$L_{SSIM} = \sum_{\mathbf{p}_t^i \in P} M_i[\mathbf{p}_t^i](1 - SSIM(I_t[\mathbf{p}_t^i], I_{t+1}[\mathbf{p}_{t+1}^i]))$$

Depth gradient smoothness term:

$$L_{sm} = \sum_{\mathbf{p}_t^i \in P} \|\partial_x d_i(\mathbf{p}_t^i)\| e^{-\|\partial_x I_t[\mathbf{p}_t^i]\|} + \|\partial_y d_i(\mathbf{p}_t^i)\| e^{-\|\partial_y I_t[\mathbf{p}_t^i]\|}$$

Differentiable Loss Function

Photometric consistency term:

$$L_{ph} = \sum_{\mathbf{p}_t^i \in P} \| M_i[\mathbf{p}_t^i](I_t[\mathbf{p}_t^i] - I_{t+1}[\mathbf{p}_{t+1}^i]) \|_2$$

Structured similarity term:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x + \sigma_y + c_2)}$$

$$L_{SSIM} = \sum_{\mathbf{p}_t^i \in P} M_i[\mathbf{p}_t^i](1 - SSIM(I_t[\mathbf{p}_t^i], I_{t+1}[\mathbf{p}_{t+1}^i]))$$

Depth gradient smoothness term:

$$L_{sm} = \sum_{\mathbf{p}_t^i \in P} \|\partial_x d_i(\mathbf{p}_t^i)\| e^{-\|\partial_x I_t[\mathbf{p}_t^i]\|} + \|\partial_y d_i(\mathbf{p}_t^i)\| e^{-\|\partial_y I_t[\mathbf{p}_t^i]\|}$$

Weighted sum:

$$L = \sum_s \alpha L_{ph}^s + \beta L_{3D}^s + \gamma L_{sm}^s + \omega L_{SSIM}^s$$

Roadmap

Depth Estimation Evaluation

$$RMSE = \sqrt{\frac{1}{|P|}\sum_{\mathbf{p}_t^i \in P} \|d_i(\mathbf{p}_t^i) - d_i^{gt}(\mathbf{p}_t^i)\|_2^2}$$

Depth Estimation Evaluation

$$RMSE = \sqrt{\frac{1}{|P|}\sum_{\mathbf{p}_t^i \in P} \|d_i(\mathbf{p}_t^i) - d_i^{gt}(\mathbf{p}_t^i)\|_2^2}$$

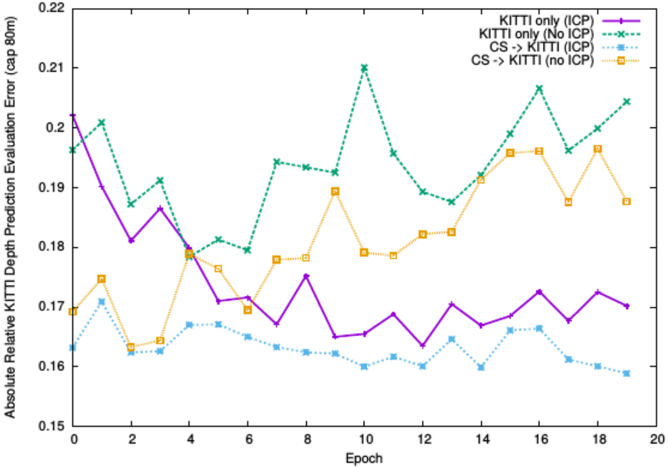
$$\Delta_i = ln\ d_i(\mathbf{p}_t^i) - ln\ d_i^{gt}(\mathbf{p}_t^i)$$

$$RMSE_{scale-invariant}^{log} = \sqrt{\frac{1}{|P|}\sum_{\mathbf{p}_t^i \in P} \Delta_i^2 - \frac{1}{|P|^2}(\sum_{\mathbf{p}_t^i \in P} \Delta_i)^2}$$

Depth Estimation Evaluation

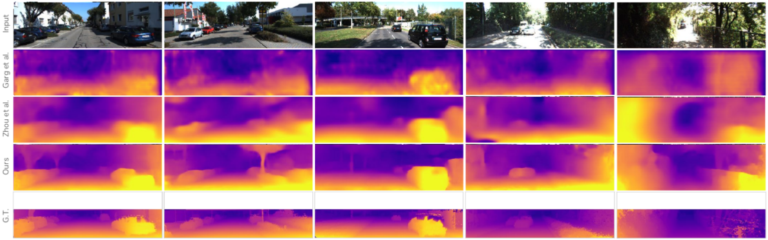
Method	Supervision	Dataset	Depth Cap	*RMSE*	$RMSE^{log}_{scale-invariant}$
All losses	–	Cityscapes + KITTI	0-80m	**5.912**	**0.243**
All losses	–	KITTI	0-80m	6.220	0.250
No ICP loss	–	KITTI	0-80m	6.267	0.252
Zhou et al.	–	Cityscapes + KITTI	0-80m	6.565	0.275
Zhou et al.	–	KITTI	0-80m	6.856	0.283
Eigen et al. Coarse	Depth	KITTI	0-80m	6.563	0.292
Eigen et al. Fine	Depth	KITTI	0-80m	6.307	0.282
All losses	–	Bike dataset	0-80m	7.741	0.309
No ICP loss	–	Bike dataset	0-80m	7.750	0.305
SfM-Net	–	Stereo KITTI 2012	0-80m	N/A	0.45
SfM-Net	–	Stereo KITTI 2015	0-80m	N/A	0.41
All losses	–	Cityscapes + KITTI	1-50m	**4.383**	**0.227**
All losses	–	KITTI	1-50m	4.549	0.231
Garg et al.	Stereo	KITTI	1-50m	5.104	0.273

3D Loss Term Ablation and Pre-Training on Cityscapes



[Mahjourian et al., CVPR 2018]

Qualitative Results



[Mahjourian et al., CVPR 2018]

Ego-Motion Estimation Evaluation

$$ATE = \sqrt{\frac{1}{|T|} \sum_{P_i \in T} \| trans(Q_i^{-1} S P_i) \|_2^2}$$

Ego-Motion Estimation Evaluation

$$ATE = \sqrt{\frac{1}{|T|} \sum_{P_i \in T} \| trans(Q_i^{-1} S P_i) \|_2^2}$$

Method	Sequence 09	Sequence 10
Full ORB-SLAM	0.014 ± 0.008	$\mathbf{0.012 \pm 0.011}$
Zhou et al.	0.021 ± 0.017	0.020 ± 0.015
No ICP loss	0.014 ± 0.010	0.013 ± 0.011
All losses	$\mathbf{0.013 \pm 0.010}$	$\mathbf{0.012 \pm 0.011}$

Roadmap

Conclusion and Future Work

- Main contribution: Novel differentiable three-dimensional geometry loss term
- Attained precision: equal to a full SLAM system
- Robustness: competitive results even after training on a highly irregular custom dataset and evaluation on an unrelated well-calibrated benchmark

Conclusion and Future Work

- Main contribution: Novel differentiable three-dimensional geometry loss term
- Attained precision: equal to a full SLAM system
- Robustness: competitive results even after training on a highly irregular custom dataset and evaluation on an unrelated well-calibrated benchmark

Future research directions:

- Dynamic object detection and tracking
- Optimization over an extended time lapse
- Scene model generalization: non-rigidity, specular reflections
- Learning and evaluation on a richer dataset incorporating all 6 DOF for ego-motion

References I

A. Byravan and D. Fox.
Se3-nets: Learning rigid body motion using deep neural networks.
arXiv, abs/1606.02378, 2016.

C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard.
Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age.
IEEE T-RO, 2016.

D. Eigen, C. Puhrsch, and R. Fergus.
Depth map prediction from a single image using a multi-scale deep network.
NIPS, 2014.

References II

J. Engel, V. Koltun, and D. Cremers.
Direct sparse odometry.
ICRA, 2017.

R. Garg, V. K. B. G, and I. D. Reid.
Unsupervised CNN for single view depth estimation: Geometry to the rescue.
arXiv, abs/1603.04992, 2016.

C. Godard, O. Mac Aodha, and G. J. Brostow.
Unsupervised monocular depth estimation with left-right consistency.
arXiv, abs/1609.03677, 2016.

Y. Kuznietsov, J. Stückler, and B. Leibe.
Semi-supervised deep learning for monocular depth map prediction.
arXiv, abs/1702.02706, 2017.

References III

R. Mahjourian, M. Wicke, and A. Angelova.
Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints.
CVPR, 2018.

N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox.
A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation.
arXiv, abs/1512.02134, 2015.

R. Mur-Artal and J. D. Tardós.
ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras.
arXiv, abs/1610.06475, 2016.

References IV

O. Ozyesil, V. Voroninski, R. Basri, and A. Singer.
A survey on structure from motion.
Acta Numerica, 2017.

S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar, and K. Fragkiadaki.
Sfm-net: Learning of structure and motion from video.
arXiv, abs/1704.07804, 2017.

T. Zhou, M. Brown, N. Snavely, and D. G. Lowe.
Unsupervised learning of depth and ego-motion from video.
CVPR, 2017.