Deep SLAM

Progress report @ Su Lab, 2020/02/12

Presenter: You-Yi Jau, M.S. in UCSD

Outlines

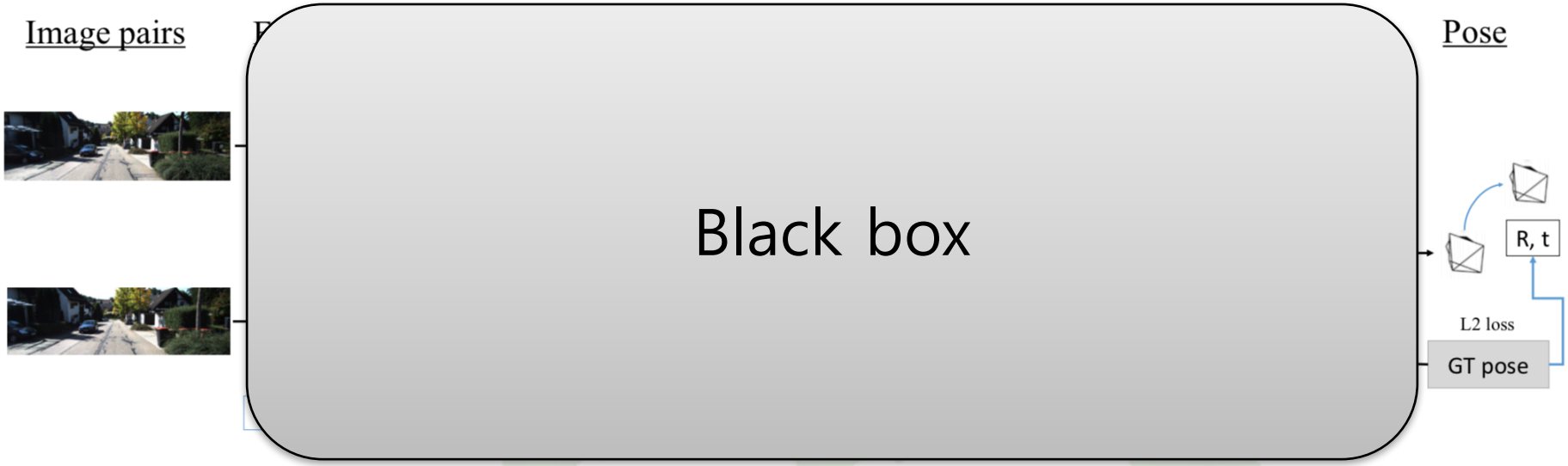
- Deep Keypoint-based Camera Pose Estimation with Geometric Constraints
- Progress: Path to deep SLAM

Deep Keypoint-based Camera Pose Estimation with Geometric Constraints

You-Yi Jau, Rui Zhu, Hao Su, Manmohan Chandraker

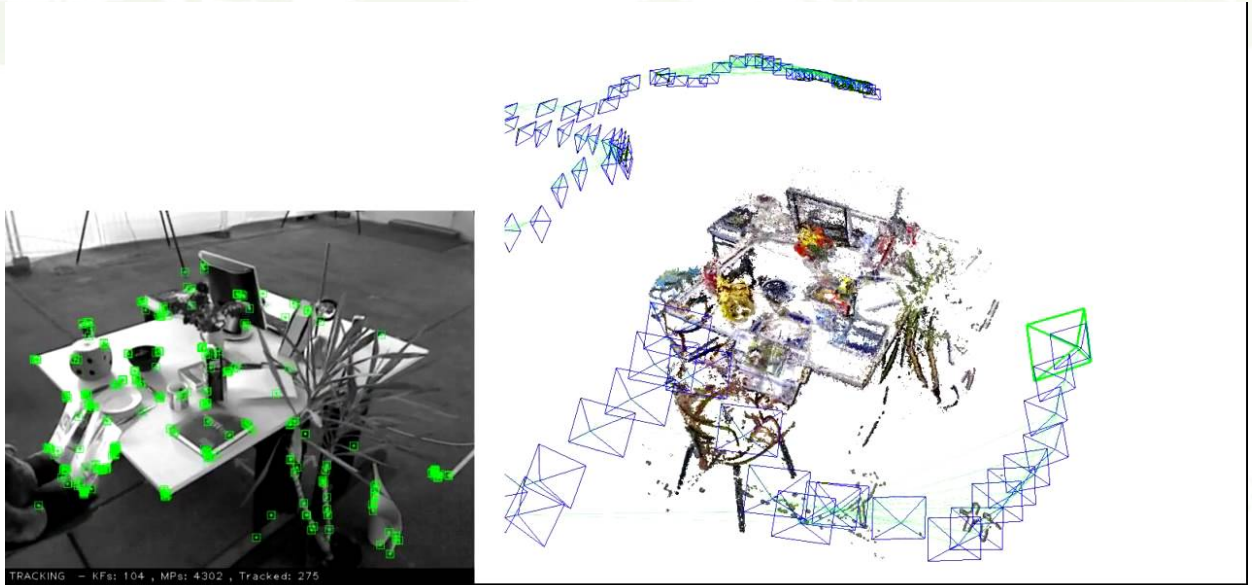
CVPR 2020 Submission

Camera pose estimation



Example: ORB-SLAM

- Image → ORB feature extractor (green points) → correspondences
- Correspondences → Camera pose (R, t) → Trajectory



Motivation and problem description

- Camera pose estimation has been the key to Simultaneous Localization and Mapping (SLAM) systems
- SIFT + RANSAC method has dominated the design of camera pose estimation pipeline for decades.
- Basic challenges for learning-based systems.
  - Not trained and optimized end-to-end for the ultimate purpose of camera poses
  - The over-fitting nature of training-based methods
  - Existing learning-based keypoint detector is weaker than SIFT

Contribution

- A new end-to-end trainable framework for feature extraction, matching, outlier rejection, and relative pose estimation
- The pipeline is tightly connected with the novel *Soft-argmax* bridge, and optimized with geometry-based objective obtained from correspondences
- The thorough study on cross-dataset setting is done to evaluate generalization ability, which is critical but not much discussed in the existing works

Prior works

- Geometry-based visual odometry
  - ORB-SLAM: 8-point algorithm, PnP, RANSAC, bundle adjustment
  - LSD-SLAM
- Learning-based visual odometry
  - CNN-SLAM: depth prediction, PoseNet
  - DeepVO: RNN
- Learning-based feature extraction and matching
  - SuperPoint, … (many new papers)
- Learning-based camera pose estimation
  - Sfm-learner
  - Deep fundemental matrix estimation (DeepF)

Overview: Pipeline

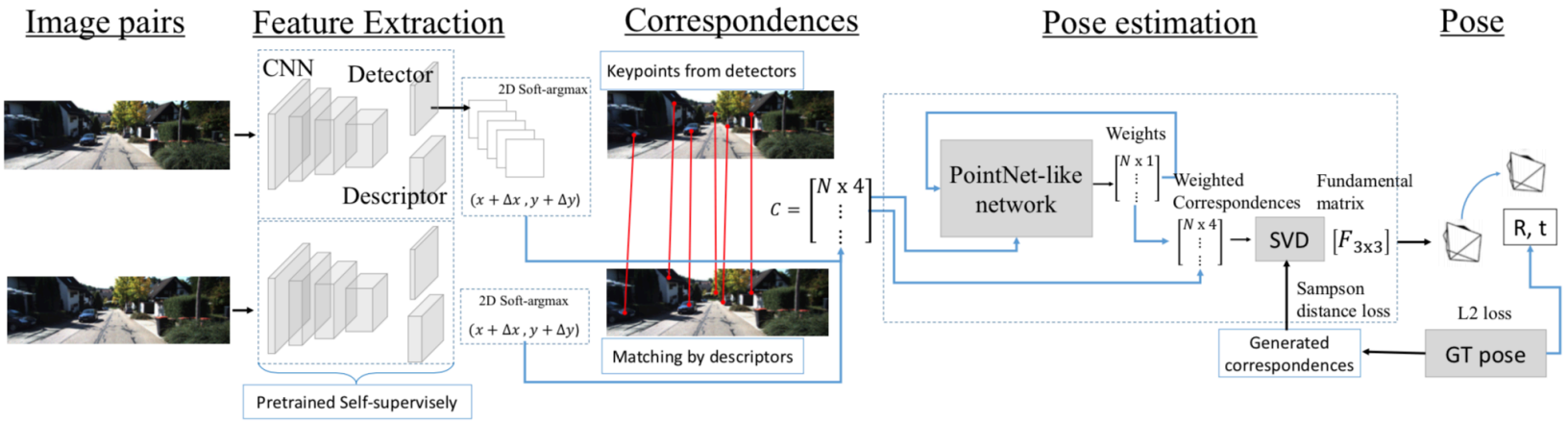
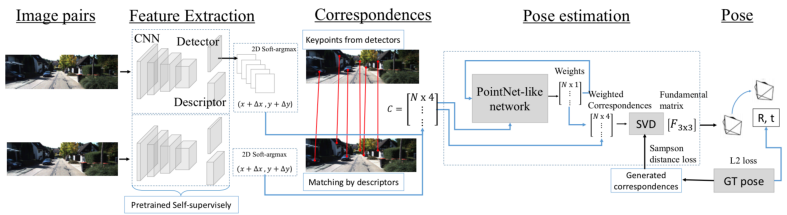


Figure 1: **Overview of the system.** A pair of images are fed into the pipeline to predict the relative pose. Feature extraction predicts detection heatmap and descriptors for finding sparse correspondences. Local 2D `Softargmax` is used to get subpixel prediction with gradients. Matrix $\mathbf{C}$ as size $N \times 4$ is formed from correspondences. $\mathbf{C}$ matrix is the input for pose estimation, where the PointNet-like network predicts weights for all correspondences. Weighted correspondences are passed through SVD to find fundamental matrix F, and further decomposes into pose. Ground truth poses (GT poses) are used to compute L2 loss between rotation and translation. Correspondences generated from GT poses are used to compute fundamental matrix loss (F-loss). See more details in Sec. 3.

Method details and analysis



- Soft-argmax detector head
  - 5 x 5 patch for each keypoint
  - Weighted over u, v direction
  - Differentiable

$$\delta u = \frac{\sum_j \sum_i e^{f(u_i, v_j)} * i}{\sum_j \sum_i e^{f(u_i, v_j)}}, \delta v = \frac{\sum_j \sum_i e^{f(u_i, v_j)} * j}{\sum_j \sum_i e^{f(u_i, v_j)}}. \quad (1)$$

- Geometry-based loss
  - Correspondences → Fundamental matrix
  - Fundamental matrix → solve R, t
  - Optimize over the best R, t (min. error)

$$\mathcal{L}_{pose} = min(\mathcal{L}_{rot}(\mathbf{R}_{est}, \mathbf{R}_{gt}), c_r) + \lambda_{rt} * min(\mathcal{L}_{trans}(\mathbf{t}_{est}, \mathbf{t}_{gt}), c_t), \quad (4)$$

$$\mathcal{L}_{rot} = min(\|q(\mathbf{R}_{est\_i}) - q(\mathbf{R}_{gt})\|_2), i = [1, 2], \quad (5)$$

$$\mathcal{L}_{trans} = min(\|\mathbf{t}_{est\_i} - \mathbf{t}_{gt}\|_2), i = [1, 2], \quad (6)$$

Experiment settings

- Baselines
  - SIFT + RANSAC (Si-base)
  - SuperPoint + RANSAC (Sp-base)
  - SIFT + DeepF[34] (Si-models)
  - Our method – no end-to-end training (Sp-models)
  - Our method - with end-to-end training (DeepFEPE)
- Datasets
  - KITTI
  - ApolloScape

[34] Rene´ Ranftl and Vladlen Koltun. Deep Fundamental Ma- trix Estimation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, Computer Vision – ECCV 2018, volume 11205, pages 292–309. Springer Inter- national Publishing, Cham, 2018.

Experiment results

- Models trained on KITTI, tested on KITTI

KITTI Models	KITTI dataset - error(deg.) inlier ratio↑, mean↓, median↓					
	Rotation (deg.)			Translation (deg.)		
	0.1↑	Mean.↓	Med.↓	2.0↑	Mean.↓	Med.↓
Base(Sp-Ran)	0.189	0.641	0.217	0.481	5.798	2.103
Sp-Df-f	0.633	0.100	0.078	0.830	1.476	0.846
Sp-Df-p	0.875	0.130	0.047	0.887	1.719	0.539
Ours(Sp-Df-f-end)	0.915	0.053	0.042	0.905	1.662	**0.489**
Ours(Sp-Df-p-end)	**0.932**	**0.050**	**0.041**	0.905	1.600	0.503
Ours(Sp-Df-fp-end)	0.910	0.054	0.048	**0.917**	**1.062**	0.504

Table 2: **Compare Pose estimation for learning based KITTI models on KITTI dataset.** The set of models are trained in KITTI with learning based feature extraction (FE). It shows significant improvement from RANSAC to DeepF pose estimation, and from separate models to end-to-end training.

KITTI Models	KITTI dataset - error(deg.) inlier ratio↑, mean↓, median↓					
	Rotation (deg.)			Translation (deg.)		
	0.1↑	Mean.↓	Med.↓	2.0↑	Mean.↓	Med.↓
Base(Si-Ran)	0.818	0.391	0.056	0.899	1.895	0.639
Si-Df-f	0.938	**0.051**	0.041	0.914	1.699	**0.484**
Si-Df-p	0.901	0.059	0.044	0.903	1.472	0.513
Si-Df-fp	**0.947**	0.111	**0.038**	0.916	1.741	0.484
Ours(Sp-Df-fp-end)	0.910	0.054	0.048	**0.917**	**1.062**	0.504

Table 3: **Compare Pose estimation for SIFT based KITTI models on KITTI dataset.** The table compare our best `DeepFEPE` model (End-to-end training with pose and F. loss) with SIFT based pose estimation. Our model works better than T-base, and comparable with state-of-the-art Si-models.

Experiment results – cross-dataset setting

- Models trained on KITTI, tested on ApolloScape

KITTI Models	Apollo dataset - error(deg.) inlier ratio↑, mean↓, median↓					
	Rotation (deg.)			Translation (deg.)		
	0.1↑	Mean.↓	Med.↓	2.0↑	Mean.↓	Med.↓
Base(Sp-Ran)	0.407	0.205	0.118	0.583	5.645	1.670
Sp-Df-f	0.725	0.126	0.068	0.754	2.074	1.155
Sp-Df-p	0.730	0.124	0.067	0.827	1.905	0.974
Ours(Sp-Df-f-end)	0.841	0.100	0.051	0.910	**1.122**	**0.589**
Ours(Sp-Df-p-end)	0.686	0.152	0.071	0.747	2.652	1.068
Ours(Sp-Df-fp-end)	**0.864**	**0.092**	**0.051**	**0.924**	1.275	0.659

Table 4: **Compare Pose estimation for our KITTI models on Apollo dataset.** The table compares the learning based approaches in a cross dataset setting. The table shows that our end-to-end DeepFEPE performs the best.

KITTI Models	Apollo dataset - error(deg.) inlier ratio↑, mean↓, median↓					
	Rotation (deg.)			Translation (deg.)		
	0.1↑	Mean.↓	Med.↓	2.0↑	Mean.↓	Med.↓
Base(Si-Ran)	**0.922**	0.157	**0.037**	**0.979**	**0.788**	0.388
Si-Df-f	0.845	0.172	0.043	0.895	2.452	0.389
Si-Df-p	0.727	0.333	0.056	0.760	4.918	0.658
Si-Df-fp	0.840	0.148	0.044	0.911	2.103	**0.369**
Ours(Sp-Df-fp-end)	0.864	**0.092**	0.051	0.924	1.275	0.659

Table 5: **Compare Pose estimation for Baseline KITTI models on Apollo dataset.** The table compares our DeepFEPE with other baseline methods in a cross dataset setting. The results show that traditional method (SIFT + RANSAC) maintains the best overall results. Our model performs better than Si-models.

Qualitative results

Ground truth F.

Keypoints

Estimated F.

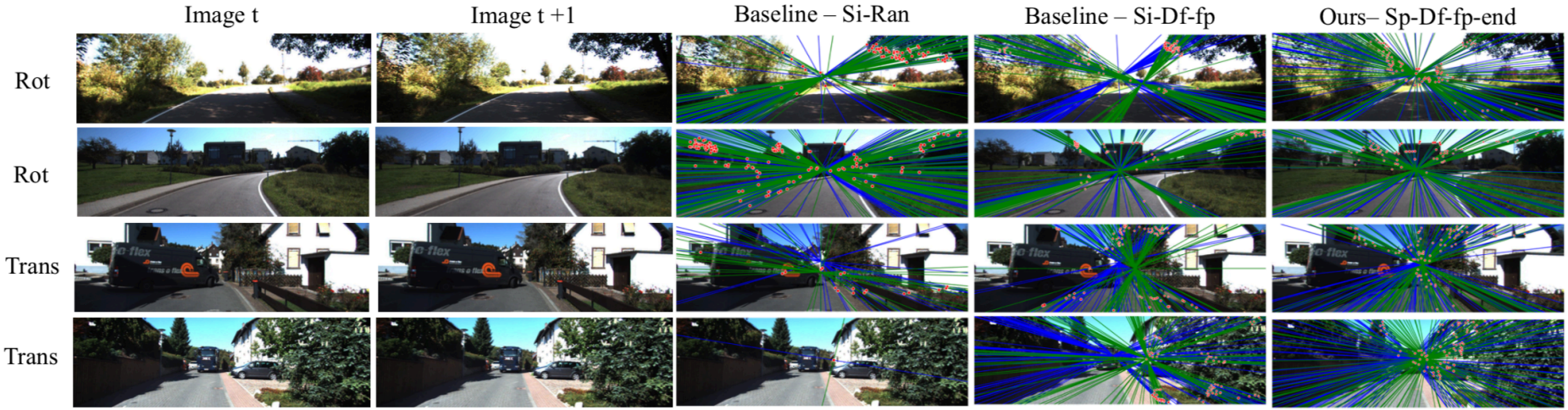


Figure 3: **Pose estimation comparison.** Compare baselines (SIFT + RANSAC) and (SIFT + deepF) with our model. The first two columns show the image pairs. The last three columns compare baseline models and our model. We show 2 examples for rotation and 2 for translation dominated pairs. The distribution of points used for estimating fundamental matrix in our pipeline is more balanced than points in the baselines, which tends to have more accurate pose estimation.

Conclusion and future work

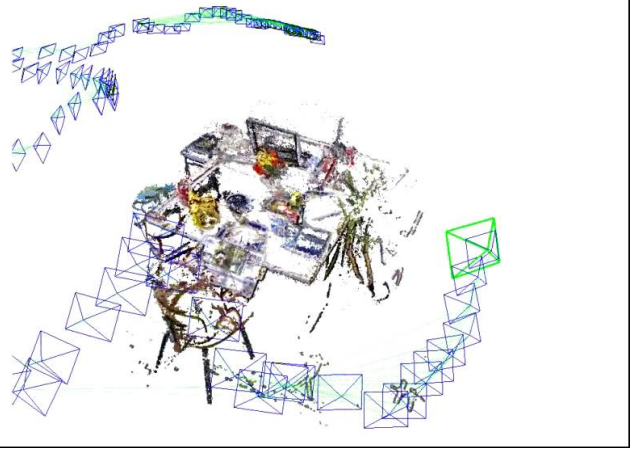
- From modules to end-to-end pipeline
- Camera pose estimation to visual odometry
- Cross-dataset evaluation could be done on indoor datasets

Outlines

- Deep Keypoint-based Camera Pose Estimation with Geometric Constraints
- Progress: Path to deep SLAM

Path to deep SLAM

- A good SLAM system consists of …
  - Initialization, re-localization
  - Long feature tracking
  - Outlier rejection
  - Keyframe-based system
  - Robust bundle adjustment (BA)
- Classic approach: ORB-SLAM



Path to deep SLAM

- A good SLAM system consists of …
  - Initialization, re-localization
  - Long feature tracking
  - Outlier rejection
  - Keyframe-based system
  - Robust bundle adjustment (BA)
- Classic approach: ORB-SLAM

What about deep learning pipeline?

How good are the deep SLAM methods

- Classical methods
  - ORB-SLAM
- Learning-based -- Supervised methods
  - DeepVO [1]
- Learning-based -- Unsupervised methods
  - Depth-VO-Feat [2]

[2] Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction H Zhan, R Garg, C Saroj Weeras ekera, K Li, H Agarwal… - Proceedings of the IEEE …, 2018

Cross-dataset generalizability

me	Link	Affiliation	Year	Platform	Publication	Environment
UZH-FPV Drone Racing	http://rpg.ifi.uzh.ch/uzh-fpv.html	UZH, ETH	2019	UAV	ICRA	Indoor, Outdoor
ADVIO Dataset	https://github.com/AaltoVision/ADVIO	Aalto U	2018	Hand	ECCV	Urban
KITTI	http://www.cvlibs.net/datasets/kitti/index.php	KIT	2013	Veh	IJRR	Urban
TUM-RGBD	https://vision.in.tum.de/data/datasets/rgbd-dataset	TUM	2012	Hand / Mob	IROS	Indoor
EuRoc	http://projects.asl.ethz.ch/datasets/doku.php?id=kmavvisualinertialdatasets	ETH-ASL	2016	UAV	IJRR	Indoor

Discussion and future works

- Cross-dataset evaluation over different methods is in progress.
- Generate thoughts for the deep SLAM pipeline.

References

[1] Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks S Wang, R Clark, H Wen, N Trigoni - 2017 IEEE International Conference on Robotics …, 2017

[2] Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction H Zhan, R Garg, C Saroj Weerasekera, K Li, H Agarwal… - Proceedings of the IEEE …, 2018

Thank you

Backup

Motivation and problem description

- Why visual odometry?
  - Localization
  - Mapping
- Classic methods
  - ORB-SLAM: https://www.youtube.com/watch?v=IuBGKxgaxS0
- Why deep learning?
  - Challenging scenes: textureless, lighting, indoor, outdoor
  - Hand-crafted → Optimize the system end-to-end

Brief summary of recent works

- Pure deep learning visual odometry
  - Beyond tracking: LSTM, relative pose, absolute pose
    - Beyond tracking: Selecting memory and refining poses for deep visual odometry
    - F Xue, X Wang, S Li, Q Wang, J Wang, H Zha - … of the IEEE Conference on Computer …, 2019
  - DSVO: predict left frame from right frame
    - Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry
    - N Yang, R Wang, J Stuckler, D Cremers - … of the European Conference on Computer …, 2018
- Not pure deep learning visual odometry
    - Self-Supervised 3D Keypoint Learning for Ego-motion Estimation
    - Visual Odometry Revisited: What Should Be Learnt?