RAUM-VO

Subjects: Robotics

Contributor: Jose Luis Sanchez-Lopez , Claudio Cimarelli , Hriday Bavle , Holger Voos

Unsupervised learning for monocular camera motion and 3D scene understanding has gained popularity over traditional methods, which rely on epipolar geometry or non-linear optimization. Notably, deep learning can overcome many issues of monocular vision, such as perceptual aliasing, low-textured areas, scale drift, and degenerate motions. In addition, concerning supervised learning, researchers can fully leverage video stream data without the need for depth or motion labels. RAUM-VO is presented here, an approach based on a model-free epipolar constraint for frame-to-frame motion estimation (F2F) to adjust the rotation during training and online inference.

visual odometry depth estimation unsupervised learning deep learning robotics

1. Introduction

One of the key elements for robot applications is autonomously navigating and planning a trajectory according to surrounding space obstacles. In the context of navigation systems, self-localization and mapping are pivotal components, and a wide range of sensors—from exteroceptive ones, such as the Global Positioning System (GPS), to proprioceptive ones, such as inertial measurement units (IMUs), as well as light detection and ranging (LiDAR) 3D scanners, and cameras—have been employed in the search for a solution to this task. As humans experience the rich amount of information coming from vision daily, exploring solutions that rely on a pure imaging system is particularly intriguing. Besides, relying only on visual clues is desirable as these are easy to interpret, and cameras are the most common sensor mounted on robots of every kind.

Visual simultaneous localization and mapping (V-SLAM) methods aim to optimize the tasks of motion estimation, that is, the 6 degrees of freedom (6DoF) transform that relates one camera frame to the subsequent one in 3D space, and 3D scene geometry (i.e., the depth and structure of the environment), in parallel. Notably, due to the interdependent nature of the two tasks, an improvement on the solution for one influences the other. On the one hand, the mapping objective is to maintain global consistency of the locations of the landmarks, that is, selected points of the 3D world that SLAM tracks. In turn, revisiting a previously mapped place may trigger a loop-closure [1], which activates a global optimization step for reducing the pose residual and smoothing all the past trajectory errors [2]. On the other hand, visual odometry (VO) [3] intends to carry out a progressive estimation of the ego-motion without the aspiration of obtaining a globally optimal path. As such, researchers can define VO as a sub-component of V-SLAM without the global map optimization routine required to minimize drift [4]. However, even VO methods construct small local maps composed by the tracked 2D features, to which a depth measurement is

associated either through triangulation [5] or probabilistic belief propagation [6][7]. In turn, these 3D points are needed to estimate the motion between future frames.

Unsupervised methods have gained popularity for camera motion estimation and 3D geometry understanding in recent years [8]. Especially regarding monocular VO, approaches such as TwoStreamNet [9] have shown equally good or even superior performances compared to traditional methods, such as VISO2 [10] or ORB-SLAM [11]. The unsupervised training protocol [12] bears some similarities with the so-called direct methods [13]. Both approaches synthesize a time-adjacent frame by projecting pixel intensities using the current depth and pose estimations and minimizing a photometric loss function. However, the learned strategy differs from the traditional one because the network incrementally incorporates the knowledge of the 3D structure and the possible range of motions into its weights, giving better hypotheses during later training iterations. Moreover, through learning, researchers can overcome the typical issues of traditional monocular visual odometry. For example, the support of a large amount of example data during training can help solve degenerate motions (e.g., pure rotational motion), scale ambiguity and scale drift, initialization and model selection, low or homogeneously textured areas, and perceptual aliasing [4]. However, being aware of the solid theory behind the traditional methods [14] and their more general applicability, researchers leverage geometrical image alignment to improve the pose estimation.

Therefore, in this work, researchers present RAUM-VO. Researchers' approach, shown in **Figure 1**, combines unsupervised pose networks with two-view geometrical motion estimation based on a model-free epipolar constraint to correct the rotations. Unlike recent works [15][16] that train optical flow and use complex or computationally demanding strategies for selecting the best motion model, researchers' approach is more general and efficient. First, researchers extract 2D keypoints using Superpoint [17] from each input frame and match the detected features from pairs of consecutive frames with Superglue [18]. Subsequently, researchers estimate the frame-to-frame motion using the solver proposed by Kneip et al. [19], which researchers name F2F, and use the rotation to guide the training with an additional self-supervised loss. Finally, RAUM-VO efficiently adjusts the rotation predictions with F2F during online inference, while retaining the scaled translation vectors from the pose network.

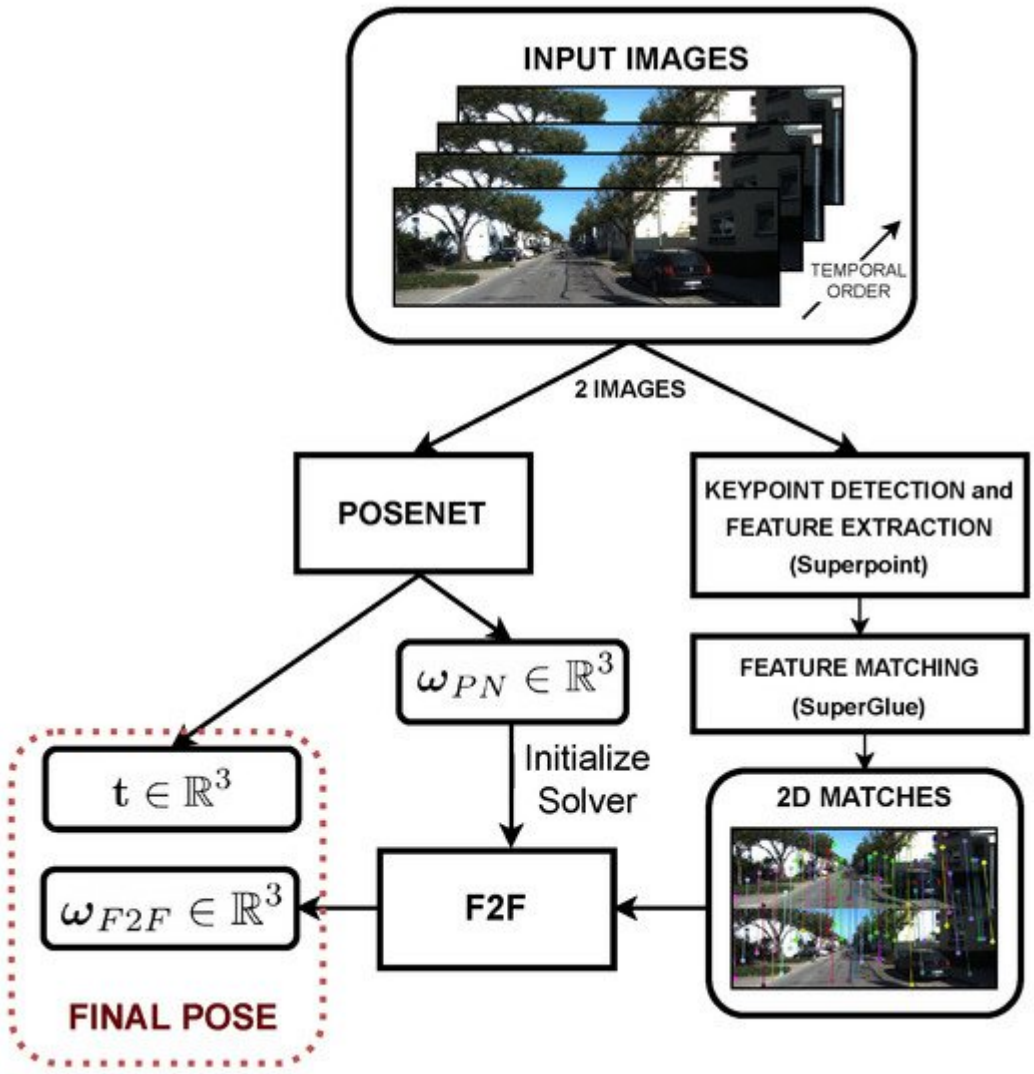

Figure 1. RAUM-VO block diagram. The figure shows the flow of information inside RAUM-VO from the input image sequence to the final estimated pose between each pair of consecutive image frames.

2. Background on SLAM

The difference between SLAM and VO is the absence of a mapping module that performs relocalization and global optimization of the past poses. Aside from this aspect, researchers can consider contributions in monocular SLAM works seamlessly with those in the VO literature. A primary type of approach to SLAM is filter-based, either using extended Kalman filters (EKFs) (as in MonoSLAM [20]) or particle filters (as in FastSLAM [21]), and keyframe-based [5], referred in robotics to as smoothing [22]. This name entails the main difference between keyframe-based and filtering. While the first optimizes the poses and the landmarks associated with keyframes (a sparse subset of the complete history of frames) using batch non-linear least squares or bundle adjustment (BA) [23], the latter marginalizes past poses' states to estimate the last at the cost of accumulating linearization errors [24]. In favor of bundle adjustment, Strasdat et al. [25] show that the accuracy of the pose increases when the SLAM system tracks more features and that the computational cost for filtering is cubic in the number of features' observations,

compared to linear for BA. Thus, using BA with an accurate selection of keyframes allows more efficient and robust implementations of SLAM. Unsupervised methods are more similar to the keyframe-based SLAM. The motion is not the result of a probabilistic model propagation and a single-step update but of an iterative optimization to align a batch of image measurements.

Motion estimation approaches fall into either direct or indirect categories based on the information or measurements included in the optimized error function. The direct method [13][26] includes intensity values in a non-linear energy function representing the photometric difference between pixels' or patches' correspondences. These are found by projecting points from one frame to another using the current motion and depth estimation, which is optimized either through the Gauss–Newton or Levenberg–Marquardt method. Instead, indirect methods [5][11] leverage epipolar geometry theory [14] to estimate motion from at least five matched 2D point correspondences, in the case of calibrated cameras [27], or eight, in the case of uncalibrated cameras [28]. After initializing a local map from triangulated points, perspective-n-point (PnP) [29] can be used with a random sample consensus (RANSAC) robust iterative fitting scheme [30] to obtain a more precise relative pose estimation. Subsequently, local BA refines the motion and the geometrical 3D structure by optimizing the reprojection error of the tracked features.

3. Related Work

Unsupervised Learning of Monocular VO

The pioneering work of Garg et al. [31] represents a fundamental advancement, because they approached the problem of depth prediction from a single frame in an unsupervised manner for the first time. Their procedure consists of synthesizing a camera's depths in a rectified stereo pair by warping the other using the calibrated baseline and focal lengths. Godard et al. [32] use the stereo pair to enforce a consistency term between left and right synthesized disparities, while adopting the structural similarity (SSIM) metric [33] as a more informative visual similarity function than the L1

loss. SfM-Learner [12] relies entirely on monocular video sequences and proposes the use of a bilinear differentiable sampler from ST-Nets [34] to generate the synthesized views.

Because the absolute metric scale is not directly observable from a single camera (without any prior knowledge about object dimensions), stereo image pairs are also helpful to recover a correct metric scale during training while maintaining the fundamental nature of a monocular method [35][36][37]. Mahjourian et al. [38] impose the scale consistency between adjacent frames as a requirement for the depth estimates by aligning the 3D point clouds using iterative closest point (ICP) and approximating the gradients of the predicted 6DoF transform. Instead, Bian et al. [39], arguing that the previous approach ignores second-order effects, show that it is possible to train a globally consistent scale with a simple constraint over consecutive depth maps, allowing one to reduce drift over long video sequences. In [40], a structure-from-motion (SfM) model is created before training and used to infer a global scale, using the image space distance between projected coordinates and optical flow displacements. More recently, several approaches [15][16][41] have leveraged learned optical flow dense pixel correspondences to recover

up-to-scale two-view motion based on epipolar geometry. Therefore, they resolve the scale factor by aligning a sparse set of points with the estimated depths.

One of the main assumptions of the original unsupervised training formulation is that the world is static. Hence, many works investigate informing the learning process about moving objects through optical flow [42][43][44][45][46][47][48][49][50][51][52][53]. The optical flow, which represents dense maps of the pixel coordinates displacement, can be separated into two components. The first, the rigid flow, is caused by the camera's motion. The second, the residual flow, is caused by dynamic objects that move freely in relation to the camera frame. Therefore, these methods train specific networks to explain the pixel shifts inconsistent with the two-view rigid motion. However, these methods focus principally on the depth and optical flow maps quality and give few details about the impact of detecting moving objects on the predicted two-view motion. Notably, they use a single metric to benchmark the relative pose that is barely informative about the global performance and cannot distinguish the improvements clearly.

A recent trend is to translate traditional and successful approaches such as SVO [54], LSD-SLAM [26], ORB-SLAM [11], and DSO [13] into their learned variants, or to take them as inspiration for creating hybrid approaches, where the neural networks usually serve as an initialization point for filtering or pose graph optimization (PGO) [55][56][57][58][59][60][61][62]. However, RAUM-VO focuses on improving the predicted two-view motion of the pose network without introducing excessive computation overhead as required by a PGO backend.

Instead of training expensive optical flow, RAUM-VO leverages a pre-trained Superpoint [17] network for keypoint detection and feature description and Superglue [18] for finding valid correspondences. Unlike optical flow, the learned features do not depend on the training dataset and generalize to a broader set of scenarios. In addition, using Superglue, researchers avoid heuristics for selecting good correspondences among the dense optical flow maps, which researchers claim could be a more robust strategy. However, researchers do not use any information about moving objects to discard keypoints lying inside these dynamic areas. Finally, differently from other hybrid approaches [15][16], researchers do not entirely discard the pose network output, but researchers look for a solution that improves its predictions efficiently and sensibly. Thus, the adoption of the model-free epipolar constraint of Kneip and Lynen [19] allows researchers to find the best rotation that explains the whole set of input matches without resorting to various motion models and RANSAC schemes. To the best of researchers' knowledge, researchers are the first to test such an approach combined with unsupervised monocular visual odometry.

References

1. Gálvez-López, D.; Tardos, J.D. Bags of binary words for fast place recognition in image sequences. IEEE Trans. Robot. 2012, 28, 1188–1197.
2. Dellaert, F.; Kaess, M. Factor Graphs for Robot Perception. Found. Trends Robot. 2017, 6, 1–139.

3. Scaramuzza, D.; Fraundorfer, F. Visual Odometry . IEEE Robot. Autom. Mag. 2011, 18, 80–92.
4. Taketomi, T.; Uchiyama, H.; Ikeda, S. Visual SLAM algorithms: A survey from 2010 to 2016. IPSJ Trans. Comput. Vis. Appl. 2017, 9, 16.
5. Klein, G.; Murray, D.W. Parallel Tracking and Mapping for Small AR Workspaces. In Proceedings of the Sixth IEEE/ACM International Symposium on Mixed and Augmented Reality, ISMAR 2007, Nara, Japan, 13–16 November 2007; IEEE Computer Society: Washington, DC, USA, 2007; pp. 225–234.
6. Vogiatzis, G.; Hernández, C. Video-based, real-time multi-view stereo. Image Vis. Comput. 2011, 29, 434–441.
7. Engel, J.; Sturm, J.; Cremers, D. Semi-dense Visual Odometry for a Monocular Camera. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, 1–8 December 2013; IEEE Computer Society: Washington, DC, USA, 2013; pp. 1449–1456.
8. Ming, Y.; Meng, X.; Fan, C.; Yu, H. Deep learning for monocular depth estimation: A review. Neurocomputing 2021, 438, 14–33.
9. Ambrus, R.; Guizilini, V.; Li, J.; Pillai, S.; Gaidon, A. Two Stream Networks for Self-Supervised Ego-Motion Estimation. In Proceedings of the 3rd Annual Conference on Robot Learning, CoRL 2019, Osaka, Japan, 30 October–1 November 2019; Kaelbling, L.P., Kragic, D., Sugiura, K., Eds.; Proceedings of Machine Learning Research. PMLR: London, UK, 2019; Volume 100, pp. 1052–1061.
10. Geiger, A.; Ziegler, J.; Stiller, C. StereoScan: Dense 3d reconstruction in real-time. In Proceedings of the IEEE Intelligent Vehicles Symposium (IV), Baden-Baden, Germany, 5–9 June 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 963–968.
11. Mur-Artal, R.; Montiel, J.M.M.; Tardós, J.D. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. IEEE Trans. Robot. 2015, 31, 1147–1163.
12. Zhou, T.; Brown, M.; Snavely, N.; Lowe, D.G. Unsupervised Learning of Depth and Ego-Motion from Video. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; IEEE Computer Society: Washington, DC, USA, 2017; pp. 6612–6619.
13. Engel, J.; Koltun, V.; Cremers, D. Direct Sparse Odometry. IEEE Trans. Pattern Anal. Mach. Intell. 2018, 40, 611–625.
14. Harltey, A.; Zisserman, A. Multiple View Geometry in Computer Vision, 2nd ed.; Cambridge University Press: Cambridge, UK, 2006.

15. Zhao, W.; Liu, S.; Shu, Y.; Liu, Y. Towards Better Generalization: Joint Depth-Pose Learning without PoseNet. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020; Computer Vision Foundation/IEEE: New York, NY, USA, 2020; pp. 9148–9158.
16. Zhan, H.; Weerasekera, C.S.; Bian, J.; Garg, R.; Reid, I.D. DF-VO: What Should Be Learnt for Visual Odometry? arXiv 2021, arXiv:2103.00933.
17. DeTone, D.; Malisiewicz, T.; Rabinovich, A. Superpoint: Self-supervised interest point detection and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–23 June 2018; pp. 224–236.
18. Sarlin, P.E.; DeTone, D.; Malisiewicz, T.; Rabinovich, A. Superglue: Learning feature matching with graph neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 4938–4947.
19. Kneip, L.; Lynen, S. Direct Optimization of Frame-to-Frame Rotation. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, 1–8 December 2013; IEEE Computer Society: Washington, DC, USA, 2013; pp. 2352–2359.
20. Davison, A.J.; Reid, I.D.; Molton, N.; Stasse, O. MonoSLAM: Real-Time Single Camera SLAM. IEEE Trans. Pattern Anal. Mach. Intell. 2007, 29, 1052–1067.
21. Montemerlo, M.; Thrun, S.; Koller, D.; Wegbreit, B. FastSLAM: A Factored Solution to the Simultaneous Localization and Mapping Problem. In Proceedings of the Eighteenth National Conference on Artificial Intelligence and Fourteenth Conference on Innovative Applications of Artificial Intelligence, Edmonton, AB, Canada, 28 July–1 August 2002; Dechter, R., Kearns, M.J., Sutton, R.S., Eds.; AAAI Press/The MIT Press: Cambridge, MA, USA, 2002; pp. 593–598.
22. Dellaert, F.; Kaess, M. Square Root SAM: Simultaneous Localization and Mapping via Square Root Information Smoothing. Int. J. Robot. Res. 2006, 25, 1181–1203.
23. Triggs, B.; McLauchlan, P.F.; Hartley, R.I.; Fitzgibbon, A.W. Bundle Adjustment—A Modern Synthesis. In Proceedings of the Vision Algorithms: Theory and Practice, International Workshop on Vision Algorithms, held during ICCV ’99, Corfu, Greece, 21–22 September 1999; Triggs, B., Zisserman, A., Szeliski, R., Eds.; Lecture Notes in Computer Science. Springer: Berlin/Heidelberg, Germany, 1999; Volume 1883, pp. 298–372.
24. Scaramuzza, D.; Zhang, Z. Visual-Inertial Odometry of Aerial Robots. arXiv 2019, arXiv:1906.03289.
25. Strasdat, H.; Montiel, J.M.M.; Davison, A.J. Visual SLAM: Why filter? Image Vis. Comput. 2012, 30, 65–77.
26. Engel, J.; Schöps, T.; Cremers, D. LSD-SLAM: Large-Scale Direct Monocular SLAM. In Proceedings of the Computer Vision—ECCV 2014—13th European Conference, Zurich,

Switzerland, 6–12 September 2014; Fleet, D.J., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Part II; Lecture Notes in Computer Science. Springer: Berlin/Heidelberg, Germany, 2014; Volume 8690, pp. 834–849.

27. Nistér, D. An Efficient Solution to the Five-Point Relative Pose Problem. IEEE Trans. Pattern Anal. Mach. Intell. 2004, 26, 756–777.
28. Longuet-Higgins, H.C. A computer algorithm for reconstructing a scene from two projections. Nature 1981, 293, 133–135.
29. Lepetit, V.; Moreno-Noguer, F.; Fua, P. EPNP: Accurate O(n) Solut. PnP Probl. Int. J. Comput. Vis. 2009, 81, 155–166.
30. Cantzler, H. Random Sample Consensus (Ransac); Institute for Perception, Action and Behaviour, Division of Informatics, University of Edinburgh: Edinburgh, UK, 1981.
31. Garg, R.; Kumar, B.G.V.; Carneiro, G.; Reid, I.D. Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue. In Proceedings of the Computer Vision—ECCV 2016—14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Part VIII; Lecture Notes in Computer Science. Springer: Berlin/Heidelberg, Germany, 2016; Volume 9912, pp. 740–756.
32. Godard, C.; Aodha, O.M.; Brostow, G.J. Unsupervised Monocular Depth Estimation with Left-Right Consistency. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; IEEE Computer Society: Washington, DC, USA, 2017; pp. 6602–6611.
33. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. IEEE Trans. Image Process. 2004, 13, 600–612.
34. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial Transformer Networks. In Advances in Neural Information Processing Systems 28: Annual Proceedings of the Neural Information Processing Systems 2015, Montreal, QC, Canada, 7–12 December 2015; Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R., Eds.; NeurIPS: San Diego, CA, USA, 2015; pp. 2017–2025.
35. Li, R.; Wang, S.; Long, Z.; Gu, D. UnDeepVO: Monocular Visual Odometry Through Unsupervised Deep Learning. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation, ICRA 2018, Brisbane, Australia, 21–25 May 2018; pp. 7286–7291.
36. Zhan, H.; Garg, R.; Weerasekera, C.S.; Li, K.; Agarwal, H.; Reid, I.D. Unsupervised Learning of Monocular Depth Estimation and Visual Odometry With Deep Feature Reconstruction. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; Computer Vision Foundation/IEEE Computer Society: Washington, DC, USA, 2018; pp. 340–349.

37. Godard, C.; Aodha, O.M.; Firman, M.; Brostow, G.J. Digging Into Self-Supervised Monocular Depth Estimation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea, 27 October–2 November 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 3827–3837.
38. Mahjourian, R.; Wicke, M.; Angelova, A. Unsupervised Learning of Depth and Ego-Motion From Monocular Video Using 3D Geometric Constraints. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; Computer Vision Foundation/IEEE Computer Society: Washington, DC, USA, 2018; pp. 5667–5675.
39. Bian, J.; Li, Z.; Wang, N.; Zhan, H.; Shen, C.; Cheng, M.; Reid, I.D. Unsupervised Scale-consistent Depth and Ego-motion Learning from Monocular Video. In Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, BC, Canada, 8–14 December 2019; Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R., Eds.; Neural Information Processing Systems: San Diego, CA, USA, 2019; pp. 35–45.
40. Luo, X.; Huang, J.; Szeliski, R.; Matzen, K.; Kopf, J. Consistent video depth estimation. ACM Trans. Graph. 2020, 39, 71.
41. Li, S.; Wu, X.; Cao, Y.; Zha, H. Generalizing to the Open World: Deep Visual Odometry with Online Adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, 19–25 June 2021; Computer Vision Foundation/IEEE: New York, NY, USA, 2021; pp. 13184–13193.
42. Casser, V.; Pirk, S.; Mahjourian, R.; Angelova, A. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8001–8008.
43. Vijayanarasimhan, S.; Ricco, S.; Schmid, C.; Sukthankar, R.; Fragkiadaki, K. Sfm-net: Learning of structure and motion from video. arXiv 2017, arXiv:1704.07804.
44. Yin, Z.; Shi, J. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1983–1992.
45. Zou, Y.; Luo, Z.; Huang, J.B. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 36–53.
46. Zhao, C.; Sun, L.; Purkait, P.; Duckett, T.; Stolkin, R. Learning monocular visual odometry with dense 3D mapping from dense 3D flow. In Proceedings of the 2018 IEEE/RSJ International

Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 6864–6871.

47. Lee, S.; Im, S.; Lin, S.; Kweon, I.S. Learning residual flow as dynamic motion from stereo videos. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 1180–1186.
48. Ranjan, A.; Jampani, V.; Balles, L.; Kim, K.; Sun, D.; Wulff, J.; Black, M.J. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 12240–12249.
49. Luo, C.; Yang, Z.; Wang, P.; Wang, Y.; Xu, W.; Nevatia, R.; Yuille, A.L. Every Pixel Counts ++: Joint Learning of Geometry and Motion with 3D Holistic Understanding. IEEE Trans. Pattern Anal. Mach. Intell. 2020, 42, 2624–2641.
50. Chen, Y.; Schmid, C.; Sminchisescu, C. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 7063–7072.
51. Li, H.; Gordon, A.; Zhao, H.; Casser, V.; Angelova, A. Unsupervised monocular depth learning in dynamic scenes. arXiv 2020, arXiv:2010.16404.
52. Wang, C.; Wang, Y.P.; Manocha, D. MotionHint: Self-Supervised Monocular Visual Odometry with Motion Constraints. arXiv 2021, arXiv:2109.06768.
53. Jiang, H.; Ding, L.; Sun, Z.; Huang, R. Unsupervised monocular depth perception: Focusing on moving objects. IEEE Sens. J. 2021, 21, 27225–27237.
54. Forster, C.; Pizzoli, M.; Scaramuzza, D. SVO: Fast semi-direct monocular visual odometry. In Proceedings of the 2014 IEEE International Conference on Robotics and Automation, ICRA 2014, Hong Kong, China, 31 May–7 June 2014; pp. 15–22.
55. Wang, C.; Buenaposada, J.M.; Zhu, R.; Lucey, S. Learning depth from monocular videos using direct methods. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2022–2030.
56. Yang, N.; Wang, R.; Stuckler, J.; Cremers, D. Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 817–833.
57. Li, Y.; Ushiku, Y.; Harada, T. Pose graph optimization for unsupervised monocular visual odometry. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 5439–5445.

58. Loo, S.Y.; Amiri, A.J.; Mashohor, S.; Tang, S.H.; Zhang, H. CNN-SVO: Improving the mapping in semi-direct visual odometry using single-image depth prediction. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 5218–5223.
59. Tiwari, L.; Ji, P.; Tran, Q.H.; Zhuang, B.; Anand, S.; Chandraker, M. Pseudo rgb-d for self-improving monocular slam and depth prediction. In European Conference on Computer Vision; Springer: Berlin/Heidelberg, Germany, 2020; pp. 437–455.
60. Cheng, R.; Agia, C.; Meger, D.; Dudek, G. Depth Prediction for Monocular Direct Visual Odometry. In Proceedings of the 2020 17th Conference on Computer and Robot Vision (CRV), Ottawa, ON, Canada, 13–15 May 2020; IEEE Computer Society: Washington, DC, USA, 2020; pp. 70–77.
61. Bian, J.W.; Zhan, H.; Wang, N.; Li, Z.; Zhang, L.; Shen, C.; Cheng, M.M.; Reid, I. Unsupervised scale-consistent depth learning from video. Int. J. Comput. Vis. 2021, 129, 2548–2564.
62. Yang, N.; von Stumberg, L.; Wang, R.; Cremers, D. D3VO: Deep Depth, Deep Pose and Deep Uncertainty for Monocular Visual Odometry. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020; Computer Vision Foundation/IEEE: New York, NY, USA, 2020; pp. 1278–1289.