Supplementary

A. Network Training Details

Both DepthNet and PoseNet are implemented with PyTorch [55] and trained on a single Titan X Pascal GPU. We resize the images to 512×256 for both KITTI [25] and EuRoC MAV [5]. We use ResNet-18 [31] as the encoder of DepthNet and it is initialized with ImageNet [60] pre-trained weights. Note that since EuRoC MAV provides grayscale images only, we duplicate the images to form 3-channel inputs. The decoder of DepthNet and the entire PoseNet are initialized randomly. We use a batch size of 8 and the Adam optimizer [38] with the number of epochs 20 and 40 for KITTI and EuRoC MAV, respectively. The learning rate is set to 10^{-4} initially and decreased to 10^{-5} for the last 5 epochs.

The predicted brightness transformation parameters are the same for the 3 channels of the input images. We mask out the over-exposure pixels when applying affine brightness transformation, since we found they negatively affect the estimation of the brightness parameters. Engel et al. also find similar issues in [18].

For the total loss function

$$L_{total} = \frac{1}{s}\sum_{s}(L^{s}_{self} + \lambda^{s} L^{s}_{reg}), \tag{1}$$

we use $s = 4$ output scales with and $\lambda^s = 10^{-3} \times \frac{1}{2^{s-1}}$. For the regularization

$$L_{reg} = L_{smooth} + \beta L_{ab} \tag{2}$$

with

$$L_{smooth} = \sum_{\mathbf{p}\in V} |\nabla_x D_t|\, e^{-|\nabla_x I_t|} + |\nabla_y D_t|\, e^{-|\nabla_y I_t|} \tag{3}$$

and

$$L_{ab} = \sum_{t'}(a_{t'} - 1)^2 + b^2_{t'}, \tag{4}$$

we set $\beta = 10^{-2}$.

B. Network Architectures

DepthNet. We adopt ResNet-18 [31] as the encoder of DepthNet with the implementation from the *torchvision* package in PyTorch [55]. The decoder architecture is built upon the implementation in [26] with skip connections from the encoder, while the difference is that our final outputs contain 3 channels including D_t, D^s_t and Σ_t. Table 1 shows the detailed architecture of DepthNet decoder.

PoseNet. The architecture of PoseNet is similar to [86] without the explainability mask decoder. PoseNet takes 2 channel-wise concatenated images as the input and outputs the relative pose and the relative brightness parameters a and b. The predicted pose is parameterized with translation vector and Euler angles.

DepthNet Decoder				
layer	**chns**	**scale**	**input**	**activation**
upconv5	256	32	econv5	ELU [7]
iconv5	256	16	↑upconv5, econv4	ELU
upconv4	128	16	iconv5	ELU
iconv4	128	8	↑upconv4, econv3	ELU
disp_uncer4	**3**	1	iconv4	Sigmoid
upconv3	64	8	iconv4	ELU
iconv3	64	4	↑upconv3, econv2	ELU
disp_uncer3	**3**	1	iconv3	Sigmoid
upconv2	32	4	iconv3	ELU
iconv2	32	2	↑upconv2, econv1	ELU
disp_uncer2	**3**	1	iconv2	Sigmoid
upconv1	16	3	iconv2	ELU
iconv1	16	1	↑upconv1	ELU
disp_uncer1	**3**	1	iconv1	Sigmoid

Table 1: Network architecture of DepthNet decoder. All layers are convolutional layers with kernel size 3 and stride 1, and ↑ is 2×2 nearest-neighbor upsampling. Here **chns** is the number of output channels, **scale** is the downscaling factor relative to the input image. Note that the disp_uncer layers have 3-channel outputs that contain D_t, D^s_t and Σ_t.

PoseNet						
layer	**k**	**s**	**chns**	**scale**	**input**	**activation**
conv1	3	2	16	2	$I_{t\pm1}$,I_t	ReLU
conv2	3	2	32	4	conv1	ReLU
conv3	3	2	64	8	conv2	ReLU
conv4	3	2	128	16	conv3	ReLU
conv5	3	2	256	32	conv4	ReLU
conv6	3	2	512	64	conv5	ReLU
conv7	3	2	1024	128	conv6	ReLU
avg_pool	-	-	1024	-	conv7	-
pose	1	1	6	-	avg_pool	-
a	1	1	1	-	avg_pool	Softplus
b	1	1	1	-	avg_pool	TanH

Table 2: Network architecture of PoseNet. Except for the global average pooling layer (avg_pool), all layers are convolutional layers with **k** the kernel size, **s** the stride, **chns** the channels and **scale** the downscaling factor relative to the input image.

C. Factor Graph of Front-end Tracking

In Figure 1, we show the visualization of the factor graphs created for the front-end tracking in D3VO. The non-keyframes are tracked with respect to the reference frame, which is the latest keyframe in the optimization window with direct image alignment. With the predicted relative poses from PoseNet, we also add a prior factor between the consecutive frames. When the new non-keyframe comes, the oldest non-keyframe in the factor graph is marginalized. The figure shows the status of the factor graph for the first (I_t), second (I_{t+1}) and third non-keyframe (I_{t+2}) comes.

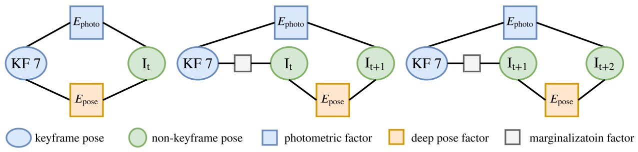

Figure 1: Visualization of the factor graph created for the front-end tracking in D3VO. From left to right are the factor graph when the first (I_t), second (I_{t+1}) and third (I_{t+2}) frame comes after the newest keyframe, which is the reference frame for the front-end tracking, is added to the optimization window. The predicted relative poses from the proposed PoseNet is used as the prior between the consecutive frames.

	avg photometric error
w/o *ab*	0.10
w/ *ab*	**0.03**
w/ *ab (LS)*	*0.07*

Table 3: Average photometric errors on *V2_03_difficult*. We project the visible 3D points with ground-truth depth of the left images onto the corresponding right images fo the stereo pairs, and then calculate the absolute photometric errors. Note that the intensity values are normalized to $[0, 1]$. The results show that by transforming the left images with the predicted *ab*, the average photometric error is largely decreased.

D. Additional Experiments on Brightness Parameters

In our main paper, we have shown that the predictive brightness parameters effectively improve the depth estimation accuracy, especially on EuRoC MAV where the illumination change is quite strong. To further validate the correctness of the predicted brightness parameters, we measure the photometric errors when projecting the pixels from the source images to the next consecutive images using the ground-truth depth and poses in *V2_03_difficult*. An example of the ground-truth depth is shown in Figure 2 for which we use the code from the authors of [28]. We first calculate the photometric errors using the original image pairs and then calculate the absolute photometric errors by transforming the left images with the predicted parameters from PoseNet. We also implemented a simple baseline method to estimate the affine brightness parameters by solving linear least squares (LS). We formulated the normal equation with the dense optical flow method [20] implemented in OpenCV [4]. As shown in Table 3, the average photometric error is decreased by a large margin when the affine brightness transformation is performed and the predicted parameters from PoseNet are better than the ones estimated from LS. We show more examples of the affine brightness transformation in Figure 4.

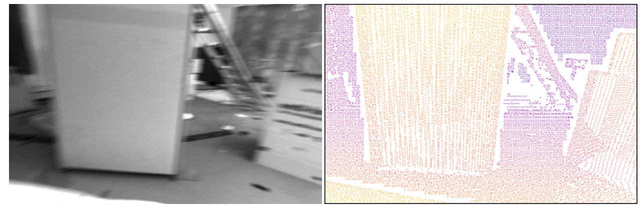

Figure 2: An example of the ground-truth depth map of *V2_03_difficult* in EuRoC MAV.

	01	02	06	08	09	10	mean
ORB2 [53]	**21.4**	*15.0*	3.52	*11.1*	*6.34*	5.25	*10.4*
S. DSO [74]	*26.5*	16.4	*3.11*	**11.0**	9.39	*3.11*	11.6
D3VO	26.9	**10.4**	**2.92**	12.7	**5.30**	**2.44**	**10.1**
ORB2 [53]	9.95	9.55	2.45	3.75	*3.07*	0.99	4.96
S. DSO [74]	*5.08*	*7.82*	*1.93*	**3.02**	4.31	*0.84*	*3.83*
D3VO	**1.73**	**5.43**	**1.69**	*3.53*	**2.68**	**0.87**	**2.65**

Table 4: Absolute translational error (ATE) as RMSE on KITTI. The upper part and the lower part show the results w/o and w/ SE(3) alignment, respectively. Note that ATE is very sensitive to the error occurs at one specific time [84].

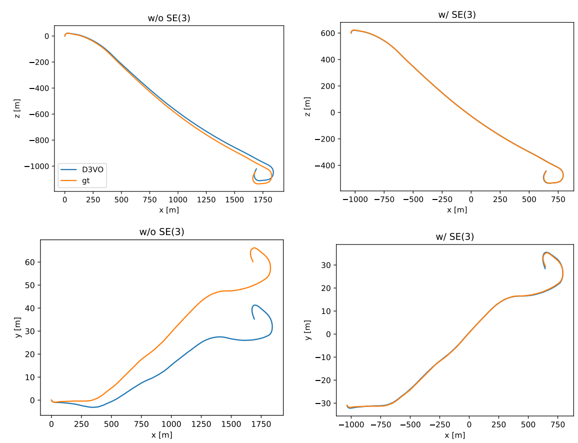

Figure 3: Trajectories on KITTI 01 to compare between w/o and w/ SE(3) alignment for the ATE evaluation. The upper part of the figure shows the trajectories on the x-z plane and the lower part shows the trajectories on the x-y plane. We can see that less accurate pose estimations for the initial frames may result in a large overall ATE, if no SE(3) alignment is performed.

E. Absolute Translational Error on KITTI

The evaluation metrics proposed with the KITTI benchmark [25] measures the relative pose accuracy. It is important to measure the global consistency of the pose estimations. Therefore, we also show the absolute translational error (ATE) as RMSE in Table 4 where the upper part shows the evaluation results without the SE(3) alignment and the lower part shows the results with the SE(3) alignment. For some sequences, e.g., KITTI 01, the ATE without SE(3) alignment is very large, while the ATE with SE(3) alignment dramatically decreases. The trajectories on KITTI 01

Figure 4: Examples of affine brightness transformation in *V2_03_difficult* from EuRoC MAV.

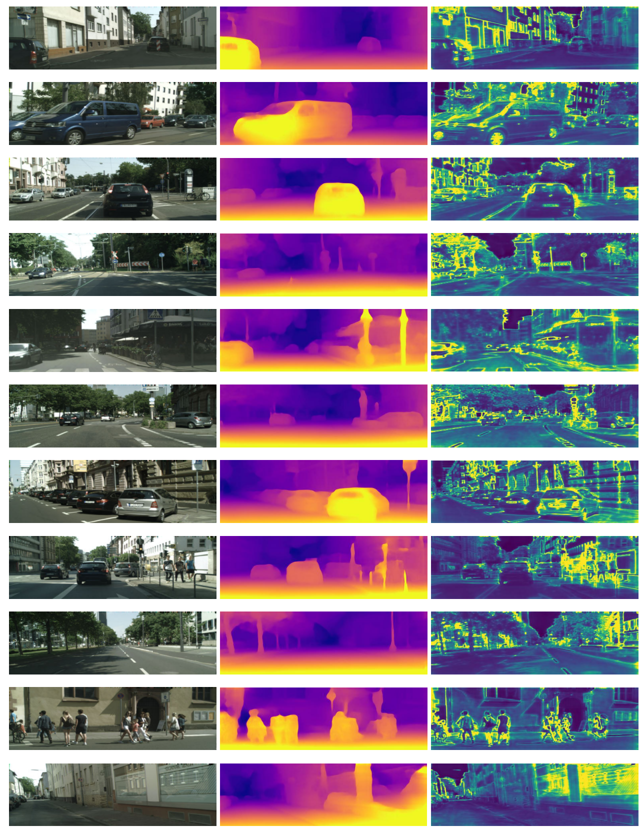

Figure 5: Results on Cityscapes with the model trained on KITTI.

are shown in Figure. 3 where we can see that the less accurate pose estimations for the initial frames may result in a large overall ATE.

F. Cityscapes

Figure 5 shows the results on the Cityscapes dataset [8] with our model trained on KITTI. The results show the generalization capability of our network on both depth and uncertainty prediction. In particular, the network can generalize to predict high uncertainties on reflectance, object boundaries, high-frequency areas, and moving objects.

References

[1] Jia-Wang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth and ego-motion learning from monocular video. In *Thirty-third Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[2] Michael Bloesch, Jan Czarnowski, Ronald Clark, Stefan Leutenegger, and Andrew J Davison. CodeSLAM-learning a compact, optimisable representation for dense visual SLAM. *arXiv preprint arXiv:1804.00874*, 2018.

[3] Michael Bloesch, Sammy Omari, Marco Hutter, and Roland Siegwart. Robust visual inertial odometry using a direct EKF-based approach. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 298–304. IEEE, 2015.

[4] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 2

[5] Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, Markus W Achtelik, and Roland Siegwart. The EuRoC micro aerial vehicle datasets. *The International Journal of Robotics Research*, 2016. 1

[6] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8001–8008, 2019.

[7] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015. 1

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[9] Jeffrey Delmerico and Davide Scaramuzza. A benchmark comparison of monocular visual-inertial odometry algorithms for flying robots. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2502–2509. IEEE, 2018.

[10] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Self-improving visual odometry. *arXiv preprint arXiv:1812.03245*, 2018.

[11] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 224–236, 2018.

[12] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2758–2766, 2015.

[13] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-Net: A trainable CNN for joint detection and description of local features. 2019.

[14] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2658, 2015.

[15] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

[16] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[17] Jakob Engel, Thomas Schöps, and Daniel Cremers. LSD-SLAM: Large-scale direct monocular SLAM. In *European Conference on Computer Vision*, pages 834–849. Springer, 2014.

[18] Jakob Engel, Jörg Stückler, and Daniel Cremers. Large-scale direct SLAM with stereo cameras. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 1935–1942. IEEE, 2015. 1

[19] J. Engel, J. Sturm, and D. Cremers. Semi-dense visual odometry for a monocular camera. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[20] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, pages 363–370. Springer, 2003. 2

[21] Tuo Feng and Dongbing Gu. SGANVO: Unsupervised deep visual odometry and depth estimation with stacked generative adversarial networks. *IEEE Robotics and Automation Letters*, 4(4):4431–4437, 2019.

[22] Christian Forster, Luca Carlone, Frank Dellaert, and Davide Scaramuzza. On-manifold preintegration for real-time visual–inertial odometry. *IEEE Transactions on Robotics*, 33(1):1–21, 2016.

[23] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. SVO: Fast semi-direct monocular visual odometry. In *2014 IEEE international conference on robotics and automation (ICRA)*, pages 15–22. IEEE, 2014.

[24] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, United States, 2018.

[25] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 1, 2

[26] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 1

[27] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. *arXiv preprint arXiv:1609.03677*, 2016.

[28] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2

[29] Ankur Handa, Thomas Whelan, John McDonald, and Andrew J Davison. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In *Robotics and automation (ICRA), 2014 IEEE international conference on*, pages 1524–1531. IEEE, 2014.

[30] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[32] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[33] Hailin Jin, Paolo Favaro, and Stefano Soatto. Real-time feature tracking and outlier rejection with changes in illumination. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 1, pages 684–689. IEEE, 2001.

[34] E. Jung, N. Yang, and D. Cremers. Multi-Frame GAN: Image Enhancement for Stereo Visual Odometry in Low Light. In *Conference on Robot Learning (CoRL)*, 2019.

[35] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017.

[36] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[37] Christian Kerl, Jürgen Sturm, and Daniel Cremers. Dense visual SLAM for RGB-D cameras. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2100–2106. IEEE, 2013.

[38] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[39] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019.

[40] Maria Klodt and Andrea Vedaldi. Supervising the new with the old: learning SFM from SFM. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 698–713, 2018.

[41] Frank R Kschischang, Brendan J Frey, and H-A Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on information theory*, 47(2):498–519, 2001.

[42] Yevhen Kuznietsov, Jörg Stückler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. *arXiv preprint arXiv:1702.02706*, 2017.

[43] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 239–248. IEEE, 2016.

[44] Stefan Leutenegger, Simon Lynen, Michael Bosse, Roland Siegwart, and Paul Furgale. Keyframe-based visual-inertial odometry using nonlinear optimization. *The International Journal of Robotics Research*, 34(3):314–334, 2015.

[45] Bo Li, Chunhua Shen, Yuchao Dai, Anton van den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1119–1127, 2015.

[46] Ruihao Li, Sen Wang, Zhiqiang Long, and Dongbing Gu. UnDeepVO: Monocular visual odometry through unsupervised deep learning. *arXiv preprint arXiv:1709.06841*, 2017.

[47] H-A Loeliger. An introduction to factor graphs. *IEEE Signal Processing Magazine*, 21(1):28–41, 2004.

[48] Shing Yan Loo, Ali Jahani Amiri, Syamsiah Mashohor, Sai Hong Tang, and Hong Zhang. CNN-SVO: Improving the mapping in semi-direct visual odometry using single-image depth prediction. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5218–5223. IEEE, 2019.

[49] R. Mahjourian, M. Wicke, and A. Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5667–5675, June 2018.

[50] Agostino Martinelli. Closed-form solution of visual-inertial structure from motion. *International journal of computer vision*, 106(2):138–152, 2014.

[51] Anastasios I Mourikis and Stergios I Roumeliotis. A multi-state constraint Kalman filter for vision-aided inertial navigation. In *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pages 3565–3572. IEEE, 2007.

[52] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015.

[53] Raul Mur-Artal and Juan D Tardós. ORB-SLAM2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017. 2

[54] Raúl Mur-Artal and Juan D Tardós. Visual-inertial monocular SLAM with map reuse. *IEEE Robotics and Automation Letters*, 2(2):796–803, 2017.

[55] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*, 2017. 1

[56] Tong Qin, Peiliang Li, and Shaojie Shen. VINS-Mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018.

[57] Tong Qin, Jie Pan, Shaozu Cao, and Shaojie Shen. A general optimization-based framework for local odometry estimation with multiple sensors. *arXiv preprint arXiv:1901.03638*, 2019.

[58] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[59] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[60] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 1

[61] Thomas Schops, Torsten Sattler, and Marc Pollefeys. BAD SLAM: Bundle adjusted direct RGB-D SLAM. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 134–144, 2019.

[62] Hauke Strasdat, JMM Montiel, and Andrew J Davison. Scale drift-aware large scale monocular SLAM. *Robotics: Science and Systems VI*, 2, 2010.

[63] H. Strasdat, J. M. M. Montiel, and A. J. Davison. Real-time monocular SLAM: Why filter? In *2010 IEEE International Conference on Robotics and Automation*, pages 2657–2664, May 2010.

[64] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 573–580. IEEE, 2012.

[65] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8934–8943, 2018.

[66] Richard Szeliski. Image alignment and stitching: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 2(1):1–104, 2006.

[67] Jiexiong Tang, Ludvig Ericson, John Folkesson, and Patric Jensfelt. GCNv2: Efficient correspondence prediction for real-time slam. *IEEE Robotics and Automation Letters*, 4(4):3505–3512, 2019.

[68] Keisuke Tateno, Federico Tombari, Iro Laina, and Nassir Navab. CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction. *arXiv preprint arXiv:1704.03489*, 2017.

[69] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic robotics*. MIT press, 2005.

[70] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. DeMon: Depth and motion network for learning monocular stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5038–5047, 2017.

[71] Vladyslav Usenko, Nikolaus Demmel, David Schubert, Jörg Stückler, and Daniel Cremers. Visual-inertial mapping with non-linear factor recovery. *arXiv preprint arXiv:1904.06504*, 2019.

[72] Lukas Von Stumberg, Vladyslav Usenko, and Daniel Cremers. Direct sparse visual-inertial odometry using dynamic marginalization. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2510–2517. IEEE, 2018.

[73] C. Wang, J. M. Buenaposada, R. Zhu, and S. Lucey. Learning depth from monocular videos using direct methods. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2022–2030, June 2018.

[74] R. Wang, M. Schwörer, and D. Cremers. Stereo DSO: Large-scale direct sparse visual odometry with stereo cameras. In *International Conference on Computer Vision (ICCV)*, Venice, Italy, October 2017. 2

[75] Sen Wang, Ronald Clark, Hongkai Wen, and Niki Trigoni. DeepVO: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 2043–2050. IEEE, 2017.

[76] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[77] N. Yang, R. Wang, X. Gao, and D. Cremers. Challenges in monocular visual odometry: Photometric calibration, motion bias and rolling shutter effect. *IEEE Robotics and Automation Letters (RA-L)*, 3:2878–2885, Oct 2018.

[78] Nan Yang, Rui Wang, Jorg Stuckler, and Daniel Cremers. Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 817–833, 2018.

[79] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. LIFT: Learned invariant feature transform. In *European Conference on Computer Vision*, pages 467–483. Springer, 2016.

[80] Xiaochuan Yin, Xiangwei Wang, Xiaoguo Du, and Qijun Chen. Scale recovery for monocular visual odometry using depth estimated with deep convolutional neural fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5870–5878, 2017.

[81] Zhichao Yin and Jianping Shi. GeoNet: Unsupervised learning of dense depth, optical flow and camera pose. In *CVPR*, 2018.

[82] H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal, and I. M. Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 340–349, June 2018.

[83] Huangying Zhan, Chamara Saroj Weerasekera, Jiawang Bian, and Ian Reid. Visual odometry revisited: What should be learnt? *arXiv preprint arXiv:1909.09803*, 2019.

[84] Zichao Zhang and Davide Scaramuzza. A tutorial on quantitative trajectory evaluation for visual (-inertial) odometry.

In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7244–7251. IEEE, 2018. 2

[85] Huizhong Zhou, Benjamin Ummenhofer, and Thomas Brox. DeepTAM: Deep tracking and mapping. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 822–838, 2018.

[86] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, volume 2, page 7, 2017. 1