Monocular Visual Odometry based on joint unsupervised learning of depth and optical flow with geometric constraints

X Y Meng¹,², B Y Sun¹,²*

¹Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, China/P. R. China
²University of Science and Technology of China, Hefei 230026, China/P. R. China
*Corresponding author’s e-mail: bysun@iim.ac.cn

Abstract. Inferring camera ego-motion from consecutive images is essential in visual odometry (VO). In this work, we present a jointly unsupervised learning system for monocular VO, consisting of single-view depth, two-view optical flow, and camera-motion estimation module. Our work mitigates the scale drift issue which can further result in a degraded performance in the long-sequence scene. We achieve this by incorporating standard epipolar geometry into the framework. Specifically, we extract correspondences over predicted optical flow and then recover ego-motion. Additionally, we obtain pseudo-ground-truth depth via triangulating 2D-2D pixel matches, which makes the depth scale is closely relevant to the pose. Experimentation on the KITTI driving dataset shows competitive performance compared to established methods.

1. Introduction
Simultaneous localization and mapping (SLAM) is one of the most essential techniques for robotics, intelligent transportation, and surveillance, among many others. SLAM comprises simultaneously estimating a robot’s state and building a model representation of its environment [1]. Visual Odometry (VO) is an integral part of SLAM, it, therefore, is likewise a fundamental problem, inferring camera motion from multi-view frames. Obtaining accurate and reliable VO estimation results can minimize the demand for global optimization. Traditional geometry approaches are built on feature matching and multi-view geometry, they can get reliable results only under appropriate conditions, such as suitable illumination, adequate texture. Recently, some researchers have been committed to achieving VO system in an end-to-end manner[2-5]. These pure learning-based methods estimate relative transformation by training deep networks so that the accuracy of the estimated scale is dependent on the learning ability of the CNNs. Although the performance of the CNNs can be improved by using massive training data and design appropriate loss functions, it is inevitable to encounter large predicted errors or even collapse. The fundamental reason is that they do not take full advantage of stable geometric information.

To solve this, several methods [6-10] that jointly learn the depth and ego-motion come into being. Since these subproblems are coupled through internal geometric constraints, solving them jointly is synergistic. Motivated by these methods, we present an unsupervised learning framework in this work, which jointly estimates a single-view depth map, relative pose, and optical flow. Our approach is based on the work in [10] and achieves better results. Different from [10], we directly estimate the camera pose by decomposing an essential matrix determined from optical flow correspondences. Meanwhile, the triangulation operation is employed on the keypoints to recover the 3D point via standard multi-view
geometry. This triangulated 3D cloud is then used as pseudo-ground-truth depth. Besides, we utilize the bidirectional flow consistency to explicitly handle occlusions. We show that our coupled unsupervised approach provides a competitive result compared to conventional methods.

2. Related Work

2.1. Geometry-based VO

VO is an extensively studied problem, which estimates camera ego-motion between image frames. Previous research can be traced back to the 1980s[11]. Geometry-based VO approaches can be divided into two categories, one is feature-based [1, 12, 13] and the other is direct methods[14-16]. The former exploits correspondences to estimate pose via 8-point algorithm[17], PnP[18], or bundle adjustment [19]. The latter directly optimize the photometric error overall pixel pairs. This method, however, encounters dynamic and lighting issues. Approaches combining the above two methods are proposed recently, e.g. [20-23]. Our proposed method addresses the scale inconsistency issue, which is thorny in monocular systems, while the above methods have failed.

2.2. Learning-based VO

With the rapid development of deep neural networks recently, learning-based methods [2-5, 24-26] have been proposed to figure visual odometry puzzles.

Agrawal et al.[26] estimate relative camera pose by learning stable visual representation. DeMoN[25], which solves the VO problem in an end-to-end manner for the first time, estimated dense depth map and poses simultaneously. DeepTAM [24] extends DTAM[15] through two separate subnetworks for the camera-motion and depth estimation simultaneously. The above two methods suggest that optical flow can be treated as an input of pose estimators rather than directly processing raw images.

Although several of the supervised methods above achieve promising results, they require huge quantities of ground-truth annotations. To get rid of it, self-supervised learning methods have been proposed. Zhou et al. [27] proposed to jointly learn camera ego-motion and depth from monocular videos. Following that, several methods[6, 9, 28-31]integrating supplementary constraints, thus, achieve more desirable results. Vid2Depth[9] adopts ICP regularization to enforce consistency depth consistency. SC-SfMLearner [10] alleviated the scale-inconsistent problem by introducing geometry consistency loss across multiple views. Unlike the above methods, we directly solve relative pose from optical flow correspondences and employ multi-view geometry. Our system shows significant improvement in accuracy.

3. Method

3.1. Method Overview

Under a unified learning framework, we predicting depth maps, optical flow, and transform matrix together, as depicted in Figure 1. We, in this part, will first elaborate on our framework composition. Then, we solve relative pose transformation from an essential matrix. Finally, we explain how the triangulated 3D point served as a supervisory signal in detail.

3.2. Network Architectures

As depicted in Figure 1, Our system consists of the DepthNet and FlowNet, representing depth prediction, and flow estimation respectively.

For dense optical flow network, we adopt a light-weighted network architecture PWC-Net [32] as the backbone to learn a robust matching. The original PWC-Net presents a compact network design, which combines sophisticated traditional strategies feature pyramid, warping, and cost volume, achieving remarkable performance on KITTI [33] and MPI Sintel [34]. We recommend readers to Sun
et al.[32] for comprehensive reviews of the literature. For single-view depth network, we employ a fully convolutional network with skip-connections to predict depth. We imitate [35] to train DepthNet.

3.3. Training Losses

We adopt the corresponding loss function to effectively train the network.

3.3.1. FlowNet Loss functions. We use the photometric loss for the brightness constancy assumption and smoothness loss for smoothness constraints.

First, we borrow the idea from [6], which uses a combination of $L_1$-norm loss and Structural Similarity (SSIM) loss[36] to alleviate the appearance difference of view synthesis. Our photometric cost can be formulated as:

$$L_{fp} = (1 - \alpha)|I_i - I_{i-1}| + \frac{\alpha}{2} (1 - SSIM(I_i, I_{i-1}))$$ (1)

$$I_{i-1} = f_{warp}(I_i, Re(K, D_i, T_{i-1}^{i-1}))$$ (2)

$\alpha = 0.85$, here, is a balancing parameter, $f_{warp}(\cdot)$ is a warping function and $K$ represents the camera intrinsics. $D_i$ and $T_{i-1}^{i-1}$ the dense depth map and relative transformational matrix, respectively. $Re(\cdot)$ indicate the reprojection from the source image $I_{i-1}$ to consecutive image $I_i$. Meanwhile, we adopt an image gradient-based edge-aware smoothness loss[37] to encourage the predicted flow to be locally smooth. The spatial smoothness loss denotes as below:

$$L_{fs} = \sum_i |\nabla F(i)| \cdot \exp(-|\nabla I_i|)^T$$ (3)

where $\nabla$ is the vector 2D differential operator, $|\cdot|$ expresses the element-wise absolute value.

3.3.2. DepthNet Loss functions. The total loss includes triangulation depth loss $L_{dtri}$, rigid flow loss $L_{rig}$, and smoothness loss $L_{ds}$.

We minimize the gap between the transformed depth $D_{tri}(i)$ and pseudo depth $D_{tri}$.

$$L_{dtri} = \left(\frac{D_{tri} - D_{tri}(i)}{D_{tri}}\right)^2$$ (4)
Where $D_0(i) = S_i D_i$. $D_x(i)$ is explicitly aligned to the triangulated 3D point, whose scale is determined by relative pose scale. Therefore, scale inconsistency is solved.

Meanwhile, given the scale-aligned depth map $D(I)$ of a single image and $K$, we can back project the points on the image plane to 3D world coordinates. Then with the estimated transformation matrix $T_{i-1-i}$, we can formulate view transformation and 2D rigid flow as equation (5) and equation (6).

\[
I_{i-1-i} = K T_{i-1-i} D_0(i) K^{-1} I_i
\]
\[
f_{i-1-i}(l_i) = I_{i-1-i} - I_i
\]

We directly use the explicit optical flow $f_{i-1-i}(l_i)$ generated from FlowNet as supervision. The Rigid flow loss function $L_{rig}$ is formulated as equation (5):

\[
L_{rig} = \| f_{i-1-i}(l_i) - f'_{i-1-i}(l_i) \|_\delta
\]

Where $\| \cdot \|_\delta$ is berHu [38] norm. $L_{ds}$, which formula is similar to equation (3), could preserve the object edges.

### 3.4. Estimate pose from flow Correspondence

The pose estimation task can be separated into two stages: First, FlowNet is employed to predict the optical flow between given adjacent images $(I_i, I_{i-1})$. We can convert the estimated flow to a collection of correspondences. Then, we choose $N$ reliable 2D-2D matches with the least bidirectional flow inconsistency in non-occluded regions. Second, those selected matches are then used to solve essential metric $E$ through Eight-point algorithm[39]. Since the optical flow in the VO task is predominantly caused by camera motion[40], it satisfies the epipolar geometric constraint[41]. 3D point $P$, as shown in Figure 2, is in world coordinate, $O_1$ and $O_2$ are camera optical center, projective points $p_1$ and $p_2$ are constrained to be on their respective epipolar lines $l_1$ and $l_2$, $e_1$ and $e_2$ are the epipole points, and the plane where triangle $PO1O2$ lies is the epipolar plane.

![Figure 2. Triangulation from two-view geometry: frame $I_1$ and frame $I_2$](image)

As for our system, we recover the transformation matrix $[R, t]$ from fundamental matrix $F$ or essential matrix $E$ with known $p_1$ and $p_2$. In this article, we adopt the essential matrix $E$ since it is simpler. We defined epipolar constraint and Essential Matrix as equation (8) and (9):

\[
p_1^T K^{-T} t^T R K^{-1} p_1 = 0
\]
\[
E = t^T R
\]

Where $R$, $t$, $K$ denotes rotational matrix, translational vector, and camera intrinsics separately. Combine with chirality condition check, i.e., the depth is positive, while four possible solutions of $[R, t]$ exist, only one is correct.

### 3.5. Triangulated point as pseudo-ground-truth depth

The scale ambiguity issue in the monocular VO system is an inevitable problem since the translation vector $t$ decomposed from essential matrix $E$ is up-to-scale.

The monocular depth scale $t$, commonly, is set to 1 in conventional VO methods. It is not on a real-world scale. Unlike these methods, we attempt to optimize scale. Instinctively, there are two feasible solutions exist, one is adjusting the depth to be consistent with pose transformation, the other is aligning
pose with predicted depth by solving PnP (3D-2D). Unfortunately, it is still not accurate enough to obtain a camera pose from the existing depth estimation methods. More attractively, optical flow estimation is more sophisticated with desirable results. Meanwhile, we adopt the triangulated 3D point as pseudo-ground-truth depth. However, the viewing rays from frame $i-1$ and frame $i$ do not exactly intersect due to pose estimation error in the ordinary case. Thus, we choose the best 3D position as the middle point along the shortest distance between the two viewing rays. It can be formulated as follows:

$$D_{tri} = \arg \min_p [d(r_1, p) + d(r_2, p)]$$

Where $r_1$ and $r_2$ denote the two viewing rays.

4. Experiments
We will explain in this section the evaluation results of our method in detail. We conduct massive experiments on the KITTI dataset[33, 42]. We compare and analyze, then, with other methods.

4.1. Implementation Details

4.1.1. Datasets. During the training stage, we adopt Eigen et al.’s split [43]. We evaluate our pipeline using KITTI odometry sequences, with publicly available ground-truth 6-DoF poses obtained from IMU/GPS. We use sequences 09-10 for qualitative evaluation. Although the dataset provides stereo image pairs, we merely use the left image during testing.

4.1.2. Training strategy. We implement the neural nets using the PyTorch [45] framework. During training, we adopt Adam[42] optimizer with batch size $b = 8$ and learning rate $lr = 10^{-4}$. In the first training stage, we only train the flow network for 20 epochs. Train the depth network, then, with the flow network frozen for another 20 epochs. Finally, both two of them are trained together for the last 10 epochs.

4.2. Comparison with classic methods
We compare our results against two learning methods Depth-VO-Feat[46], SC-SfMLearner[10], two geometry-based methods VISO2[47], and ORB-SLAM2[1] without loop closure. Since ORB-SLAM2 occasionally fails to track or initialize, we repeatedly run it three trials and adopt the best result. The result of VISO2 was referenced from [48]. Following the common evaluation criterion, we report the averaged translational errors $t_{rerr}$ and rotational errors $r_{rerr}$ over 100-800 meters. Absolute trajectory error (ATE), which is a general metric, denotes the root-mean-square error. Relative pose error (RPE) evaluates inter-frame error. We align all predicted trajectories to ground-truth by minimizing ATE.

As shown in Table 1, our experimental results substantially achieve better performance than deep learning methods in all metrics. Besides, it outperformed VISO2, which is a geometry-based method. This is expected since our framework retains and exploits more geometry information. Furthermore, our method surpasses ORB-SLAM2 without loop closing, in which error is reduced by global bundle adjustment. Besides, our method can be accelerated on the GPU while ORB-SLAM2 cannot. Figure 3 shows the estimated 3D trajectories on two testing sequences respectively. As we can see, Depth-VO-Feat, although trained with binocular images, encountered increasing error accumulation.

In general, we achieve a significant performance improvement since our method essentially mitigates the scale-inconsistency problem. Our method, accordingly, proves lower RPE meaning that our model is more robust between inter-frame tracking.
5. Conclusion

we have presented a jointly unsupervised learning VO model, and demonstrate the advantages of leveraging traditional multi-view geometry. Specifically, we solve the essential matrix from 2D-2D optical flow correspondence as an intermediate for camera ego-motion estimation. Meanwhile, we triangulate these correspondences to reconstruct sparse clouds, the triangulated 3D points are then used as pseudo-ground-truth depth. Explicit align depth estimated from DepthNet with pose via triangulation alleviates the scale inconsistency issue common in monocular VO task. Our key insight is that integrating multi-view geometry can gain the best from both low-level tasks. Experiments demonstrate that our coupled unsupervised framework obtained a competitive result compared to baselines.

Acknowledgments

The work was financially supported by the National Natural Science Foundation of China (61773360).

References

[1] Mur-Artal, R. and J.D. Tardós, ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. IEEE Transactions on Robotics, 2017.
[2] Wang, S., et al., End-to-end, sequence-to-sequence probabilistic visual odometry through deep neural networks. International Journal of Robotics Research, 2018. 37(4-5): p. 513-542.
[3] Clark, R., et al., V1Net: Visual-Inertial Odometry as a Sequence-to-Sequence Learning Problem. 2017.
[4] Wang, S., et al., DeepVO: Towards End-to-End Visual Odometry with Deep Recurrent Convolutional Neural Networks. 2017.

[5] Xue, F., et al., Beyond Tracking: Selecting Memory and Refining Poses for Deep Visual Odometry. 2019.

[6] Yin, Z., J. Shi, and Ieee, GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose, in 2018 Ieee/Cvf Conference on Computer Vision and Pattern Recognition. 2018. p. 1983-1992.

[7] Yang, D., et al., Unsupervised framework for depth estimation and camera motion prediction from video. Neurocomputing, 2020. 385: p. 169-185.

[8] Mun, J.H., M. Jeon, and B.G. Lee, Unsupervised Learning for Depth, Ego-Motion, and Optical Flow Estimation Using Coupled Consistency Conditions. Sensors, 2019. 19(11): p. 20.

[9] Mahjourian, R., 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. 2018.

[10] Bian, J.-W., et al., Unsupervised Scale-consistent Depth and Ego-motion Learning from Monocular Video, in Advances in Neural Information Processing Systems 32, H. Wallach, et al., Editors. 2019.

[11] Cai, M., et al., Camera Relocalization by Exploiting Multi-View Constraints for Scene Coordinates Regression, in 2019 Ieee/Cvf International Conference on Computer Vision Workshops. 2019. p. 3769-3777.

[12] Klein, G. and D. Murray. Parallel Tracking and Mapping for Small AR Workspaces. in IEEE & Acm International Symposium on Mixed & Augmented Reality. 2008.

[13] Geiger, A., J. Ziegler, and C. Stiller. StereoScan: Dense 3d reconstruction in real-time. in Intelligent Vehicles Symposium (IV), 2011 IEEE. 2011.

[14] Engel, J., V. Koltun, and D. Cremers, Direct Sparse Odometry. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016: p. 1-1.

[15] Newcombe, R.A., S.J. Lovegrove, and A.J. Davison. DTAM: Dense tracking and mapping in real-time. in 2011 International Conference on Computer Vision. 2012.

[16] Engel, J., T. Sch, and D. Cremers. LSD-SLAM: large-scale direct monocular SLAM. in ECCV. 2014.

[17] Hartley, R.I., In defense of the eight-point algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997. 19(6): p. 580-593.

[18] Lepeit, V., F. Moreno-Noguer, and P. Fua, EPNP: An Accurate O(n) Solution to the PnP Problem. International Journal of Computer Vision, 2009. 81(2): p. 155-166.

[19] TRIGGS, B., Bundle Adjustment -- A Modern Synthesis. 1999.

[20] Engel, J., T. Schps, and D. Cremers. LSD-SLAM: Large-scale direct monocular SLAM. in European Conference on Computer Vision. 2014.

[21] Forster, C., et al., SVO: Semidirect Visual Odometry for Monocular and Multicamera Systems. IEEE Transactions on Robotics, 2017. 33(2): p. 249-265.

[22] Engel, J., J. Sturm, and D. Cremers. Semi-dense Visual Odometry for a Monocular Camera. in IEEE International Conference on Computer Vision. 2014.

[23] Forster, C., M. Pizzoli, and D. Scaramuzza*, SVO: Fast semi-direct monocular visual odometry. in IEEE International Conference on Robotics & Automation. 2014.

[24] Zhou, H., B. Ummenhofer, and T. Brox, DeepTAM: Deep Tracking and Mapping. 2018.

[25] Ummenhofer, B., et al. DeMoN: Depth and Motion Network for Learning Monocular Stereo. in IEEE Conference on Computer Vision & Pattern Recognition. 2017.

[26] Agrawal, P., J. Carreira, and J. Malik, Learning to See by Moving. 2015.

[27] Zhou, T., et al., Unsupervised Learning of Depth and Ego-Motion from Video. arXiv, 2017.

[28] Zhan, H., et al. Unsupervised Learning of Monocular Depth Estimation and Visual Odometry with Deep Feature Reconstruction. in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018.
[29] Li, R., et al., UnDeepVO: Monocular Visual Odometry through Unsupervised Deep Learning, in 2018 IEEE International Conference on Robotics and Automation. 2018. p. 7286-7291.
[30] Chen, Y., C. Schmid, and C. Sminchisescu. Self-Supervised Learning With Geometric Constraints in Monocular Video: Connecting Flow, Depth, and Camera. in 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2020.
[31] Casser, V., et al. Depth Prediction without the Sensors: Leveraging Structure for Unsupervised Learning from Monocular Videos. in Thirty-Third AAAI Conference on Artificial Intelligence (AAAI’19). 2019.
[32] Sun, D., et al. PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018.
[33] Geiger, A., P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. in IEEE Conference on Computer Vision & Pattern Recognition. 2012.
[34] Butler, D.J., et al., A Naturalistic Open Source Movie for Optical Flow Evaluation, in Computer Vision - Eccv 2012, Pt Vi, A. Fitzgibbon, et al., Editors. 2012. p. 611-625.
[35] Godard, C., et al., Digging Into Self-Supervised Monocular Depth Estimation, in International Conference on Computer Vision. 2019. p. 3827-3837.
[36] Wang, Z., et al., Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process, 2004. 13(4).
[37] Godard, C., O. Mac Aodha, and G.J. Brostow. Unsupervised Monocular Depth Estimation with Left-Right Consistency. in Computer Vision & Pattern Recognition. 2017.
[38] Laina, I., et al., Deeper Depth Prediction with Fully Convolutional Residual Networks, in Proceedings of 2016 Fourth International Conference on 3d Vision. 2016. p. 239-248.
[39] Hartley, R.I., In defense of the eight-point algorithm. IEEE Pami, 1997.
[40] Wulff, J., L. Sevilla-Lara, and M.J. Black, Optical Flow in Mostly Rigid Scenes. 2017.
[41] Hartley, R. and A. Zisserman. Multiple view geometry in computer vision. 2000.
[42] Kingma, D. and J. Ba, Adam: A Method for Stochastic Optimization. International Conference on Learning Representations, 2014.
[43] Eigen, D., C. Puhrsch, and R. Fergus, Depth Map Prediction from a Single Image using a Multi-Scale Deep Network, in Advances in Neural Information Processing Systems 27, Z. Ghahramani, et al., Editors. 2014.
[44] Ranjan, A., et al., Competitive Collaboration: Joint Unsupervised Learning of Depth, Camera Motion, Optical Flow and Motion Segmentation, in 2019 Ieee/Cvf Conference on Computer Vision and Pattern Recognition. 2019. p. 12232-12241.
[45] Paszke, A., et al., Automatic differentiation in pytorch, in NIPS Autodiff Workshop. 2017.
[46] Zhan, H., et al., Unsupervised Learning of Monocular Depth Estimation and Visual Odometry with Deep Feature Reconstruction. 2018.
[47] Geiger, A., et al., StereoScan: Dense 3d Reconstruction in Real-time, in 2011 Ieee Intelligent Vehicles Symposium. 2011. p. 963-968.
[48] Zhan, H., et al., Visual Odometry Revisited: What Should Be Learnt? 2019.
[49] Geiger, A., et al., Vision meets robotics: The KITTI dataset. International Journal of Robotics Research, 2013. 32(11): p. 1231-1237.