Proposal of Final Project
Artificial Intelligence: Principles and Techniques
Prof. Chongjie Zhang

Feng Gao
2015011208
gaof15@mails.tsinghua.edu.cn

Introduction

Understanding the 3D structure of a scene is a fundamental problem in many machine perception tasks, and many computer vision applications need to be based on it, most notably for autonomous vehicles and robotics. Many applications now use radar and laser sensors to acquire 3D structures, while these sensors are expensive and not easy to obtain. Cameras can be a better choice, as cameras are the cheapest, least restrictive and most ubiquitous sensor for robotics. Therefore, monocular depth estimation has become a hot topic. A related problem of inferring ego-motion from a sequence of images is likewise a fundamental problem in autonomous vehicles and robotics, known as visual odometry estimation.

To solve these two problems, there has been lots of works that pose the tasks of monocular depth estimation and visual odometry as supervised learning problems[2, 3, 7, 8]. These methods attempt to directly predict the depth or odometry using models that have been trained offline on large collections of ground truth depth data. While these methods have enjoyed great success, it cannot be ignored that these annotations are expensive to obtain, e.g. expensive laser or depth camera to collect depths, and may introduce their own sensor noise. Recently, Garg et al.[4] proposed to use photometric warp error as a self-supervised signal to train a convolutional neural network for the single view depth estimation and visual odometry estimation. Following [4], some methods, such as [1, 5, 6, 9, 11, 12, 13], also use the photometric error based supervision to learn depth and pose estimators and get comparable results to that of fully supervised methods. Specifically, [4, 5, 12] use the photometric warp error between left-right images in a stereo pair to learn depth and pose, while [1, 6, 9, 11, 13] use monocular videos directly to train the estimators.

Zhou et al.[13] firstly proposed to use monocular sequences to jointly train two neural networks for depth and odometry estimation. However [13] and most of the methods that use monocular sequences in training, like [6, 9, 11], suffer from the per frame scale-ambiguity problem, i.e. the actual scale of depth and camera translations is missing and only direction is learned. To fix this issue, they multiply the predicted depth maps and camera transformation matrices by a scale factor for evaluation, and the scale factor is calculated using the groundtruth. These methods would be invalid in real environment, since we cannot get the true values. The use of stereo sequences can avoid this problem, which can constrains the scene depth and camera motion to be in a common, real-world scale. However, stereo is not nearly as widely available as monocular video, which will limit the methods applicability. A latest research[1] proposed to introduce geometric structure in the learning process by modeling the scene and the individual objects, and removed the scale-ambiguity issue. Their results outperform all of monocular methods on both depth and ego-motion estimation and are competitive to that of models trained with stereo. Moreover, Wang et al.[11] incorporated a Direct Visual Odometry (DVO)[10] as the pose predictor instead of a Pose-CNN. DVO can take the advantage of the geometric relation between camera pose and depth and yield slight improvements.

Inspired by [1] and [10], we propose a novel approach that learns depth directly from monocular videos with DDVO as pose predictor and models 3D motions of moving objects, together with camera ego-motion. Differential DVO (DDVO) is proposed by [10], which can be directly used in end-to-end learning. On the other hand, motion modeling allows our methods effectively learn from highly dynamic scenes in a monocular setting and learn the actual scale of depths and camera's ego-motions.

Reference

[1] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova. Unsupervised learning of depth and ego-motion: A structured approach. In *AAAI*, 2019.

[2] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse. Monoslam: Real-time single camera slam. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29:2007, 2007.

[3] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2366–2374. Curran Associates, Inc., 2014.

[4] R. Garg, B. V. Kumar, G. Carneiro, and I. Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision*, pages 740–756. Springer, 2016.

[5] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017.

[6] Y. Kuznietsov, J. Stueckler, and B. Leibe. Semi-supervised deep learning for monocular depth map prediction. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[7] L. Ladicky, J. Shi, and M. Pollefeys. Pulling things out of perspective. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 89–96, 2014.

[8] F. Liu, C. Shen, G. Lin, and I. Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 38(10):2024–2039, Oct. 2016.

[9] R. Mahjourian, M. Wicke, and A. Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[10] F. Steinbrcker, J. Sturm, and D. Cremers. Real-time visual odometry from dense rgb-d images. In *ICCV Workshops*, pages 719–722. IEEE, 2011.

[11] C. Wang, J. M. Buenaposada, R. Zhu, and S. Lucey. Learning depth from monocular videos using direct methods. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[12] H. Zhan, R. Garg, C. Saroj Weerasekera, K. Li, H. Agarwal, and I. Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

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