Learning Attention Maps for Unsupervised Depth Estimation in Monocular Videos

Mendoza, Julio
j153646@dac.unicamp.br

Pedrini, Helio
helio@ic.unicamp.br

Institute of Computing, University of Campinas, Campinas-SP, Brazil

Abstract

In this work, we propose an approach to learn attention maps based on the features learned by a convolutional network. This network is trained to estimate dense depth maps using a view synthesis approach, in which sequences of frames are used as a supervised signal. The attention maps are used to increase the error contribution of the meaningful regions on the error computation between the synthesized and the real views. The attention maps are learned simultaneously with the depth network in an end-to-end manner. We evaluate our method on the depth estimation and odometry tasks of the KITTI benchmark [2]. The inclusion of the attention maps shows an improvement in both tasks when compared to a competitive baseline method.

1 Introduction

Depth estimation [3] is an important task in machine perception. It allows to recognize the 3D geometric structure of an environment and can be used in several applications such as autonomous driving, robot collaboration, and localization and navigation systems. Obtaining depth maps from videos is interesting because cameras are cheap acquisition devices when compared to LIDAR.

Recently, deep learning approaches [1, 6, 4, 5] have been proposed to address the deep estimation problem in various scenarios. One of them is the unsupervised scenario, which is motivated by the lack of large datasets with depth ground truth. Several unsupervised approaches [6, 4, 5] estimate depth dense maps and camera motion transformation between subsequent frames following a view synthesis approach.

Based on the intuition that not all regions of the image have the same importance for depth and motion estimation tasks we hypothesize that the error contribution of different regions should not be equal. Therefore, we propose to quantify the importance of the regions through an attention map.

2 Problem

The input to our method is a sequence of 3-frames $\{I_{t-1}, I_t, I_{t+1}\}$ and the camera intrinsic parameters K. Depth and camera motion are learned with neural networks. Given an image I_i its depth map D_i is obtained with a convolutional network. Given two adjacent frames $\{I_i, I_j\}$, where $j \in \{i-1, i+1\}$, the camera pose $P_{i\rightarrow j}$ is estimated through a convolutional network. Using I_i, its depth map D_i and the camera motion between the frames i and j, $P_{i\rightarrow j}$, we can synthesize the frame I_i by projecting the frame I_j. We can use a reconstruction loss between I_i and the synthesized frame $\hat{I}_i$ to train the depth and motion networks. At the test time, both networks can be used to infer depth and camera motion independently.

3 Method

Our method quantifies the error contribution of each pixel through an attention map which is learned from the features on the encoder-part of the depth network for each scale of the attention. The attention map is used to weight error pixel-wise on the reconstruction loss. Moreover, we add a constraint to force the attention map to have a sum proportional to the resolution of the image. This constraint is important since it avoids a trivial solution in which the attention map, filled with zeros, excludes all pixels from the error computation. As well as other competitive methods, we include a depth smoothness loss.

4 Results

Table 1 shows that attention maps improve the performance of the baseline method on all accuracy and error metrics.

Method	↓ Lower is better				↑ Higher is better		
	Abs Rel	Sq Rel	RMSE	Log RMSE	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Baseline	0.1963	2.9921	6.7384	0.2692	0.7559	0.9126	0.9607
Ours	0.1642	1.4190	6.1079	0.2464	0.7713	0.9206	0.9672

Table 1: We compare the results of depth estimation between a baseline method and the same method including the attention map improvement on the KITTI benchmark.

Figure 1 shows the translation and rotation errors by path length. We observe that the translation error increases with the path length due to error accumulation on the baseline and our approach. However, translation error accumulation is approximately one scale factor smaller with our approach than with the baseline. Moreover, our approach reduces the rotation error consistently on all sampled path lengths.

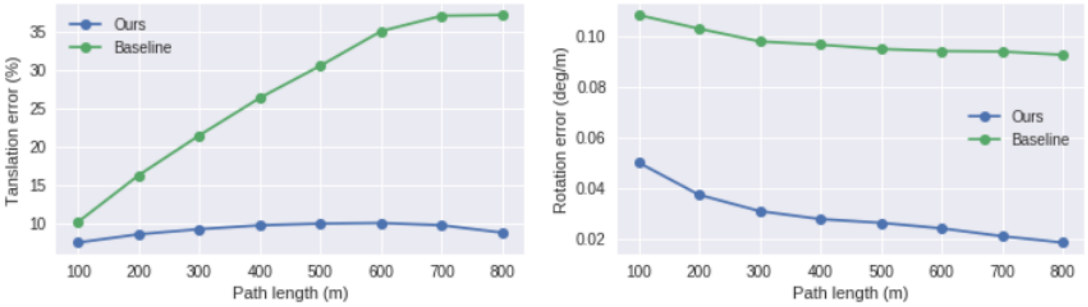

Figure 1: Translation and rotation errors by path length on the KITTI benchmark.

References

[1] David Eigen, Christian Puhrsch, and Rob 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.

[2] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision Meets Robotics: The KITTI Dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[3] J. Mendoza and H. Pedrini. Self-Supervised Depth Estimation Based on Feature Sharing and Consistency Constraints. In *15th International Conference on Computer Vision Theory and Applications (VISAPP)*, pages 134–141, Valletta, Malta, February 2020.

[4] Anurag Ranjan, Varun Jampani, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Adversarial Collaboration: Joint Unsupervised Learning of Depth, Camera Motion, Optical Flow and Motion Segmentation. *arXiv preprint arXiv:1805.09806*, 2018.

[5] Zhichao Yin and Jianping Shi. GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[6] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised Learning of Depth and Ego-Motion From Video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.