Project 4: SfM Learner

Abhinav Modi
Masters of Engineering in Robotics
University of Maryland, College Park
Email: abhi1625@umd.edu

Prateek Arora
Masters of Engineering in Robotics
University of Maryland, College Park
Email: pratique@terpmail.umd.edu

Kartik Madhira
Masters of Engineering in Robotics
University of Maryland, College Park
Email: kmadhira@terpmail.umd.edu

I. Introduction

Deep approaches to predict monocular depth and ego-motion have grown in recent years due to their ability to produce dense depth from monocular images. The main idea behind them is to optimize the photometric consistency over image sequences by warping one view into another, similar to direct visual odometry methods. Zhou *et al.* [1] proposed an unsupervised approach to learn depth and ego-motion from video. We propose minor modification in architecture and loss function to improve the accuracy of the network.

II. Our approach

A. Architectural change

The current network uses an auto encoder-decoder based VGG architecture. This is one of the most used architecture in any vision based neural network methods. We have tried an alternate approach using a Resnet architecture which also incorporates an Autoencoder-decoder to calculate the **photometric loss**.

B. Loss function

The current loss function used by the authors of the paper is a photometric loss which helps in training the depth and pose net in an unsupervised manner. We have changed the loss function to include a SSIM (Structural Similarity Index). It is another well known and robust metric for measuring perceptual differences between two images. The photometric loss assumes *Brightness Constancy* which need not be the case everytime.

SSIM considers three factors, namely luminance, contrast and structure which provide a more robust measure for image similarity. Since, SSIM needs to be maximized, we use the following loss function:

$$L_{SSIM} = \sum_{s} \frac{1 - SSIM(I_t, I_{Swarped})}{2} \tag{1}$$

Thus the loss function for the photometric loss becomes:

$$Loss = \alpha \frac{1 - SSIM(I_t, I_{Swarped})}{2} + (1-\alpha)||I_t - I_S||_1 \tag{2}$$

where α is a constant which was taken as 0.85 according to [2]. Thus the total loss becomes

$$Total_{loss} = photometric_{loss} + L_{ssim} + smooth_{loss} + exp_{loss} \tag{3}$$

C. Other minor changes

We saw that our training was unusually slow than the original SFMLearner training. Thus there are two minor changes that we introduced.

1) adaptive learning rate: We use exponential decay for faster convergence. Learning rate initially is higher and as the training progresses the learning rate exponentially decays.
2) batch norm: This is another standard technique to get faster convergence. Batch normalization reduces the amount by what the hidden unit values shift around (covariance shift).

III. Results

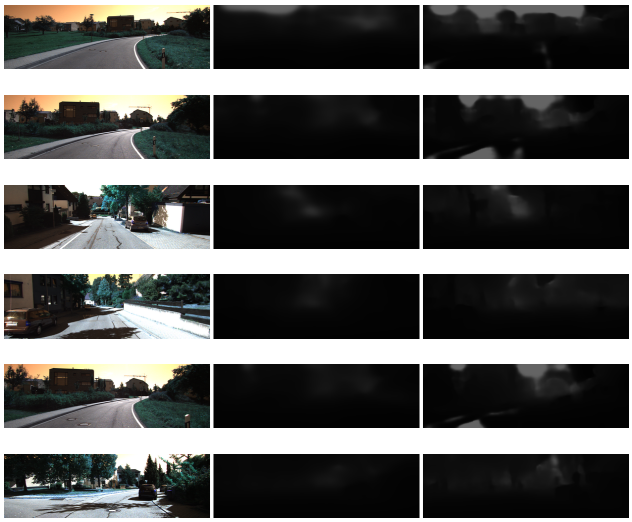

Figure 1: Input image, SFM Learner, Our output
Comparison of depth maps and corresponding rgb image.

IV. Discussion and Conclusion

The changes we have made to the existing pipeline of SFMLearner are sure promising as can be seen in the Table(1). The outputs for the depth map comparison are shown in the fig(1). The output predictions can still be improved by training the network for more number of epochs, data augmentation and including epipolar geometric constraints. The current framework works with the assumption that the scene is static, i.e., there are no dynamic objects. The explainability mask accounts for these challenges but only to some extent.

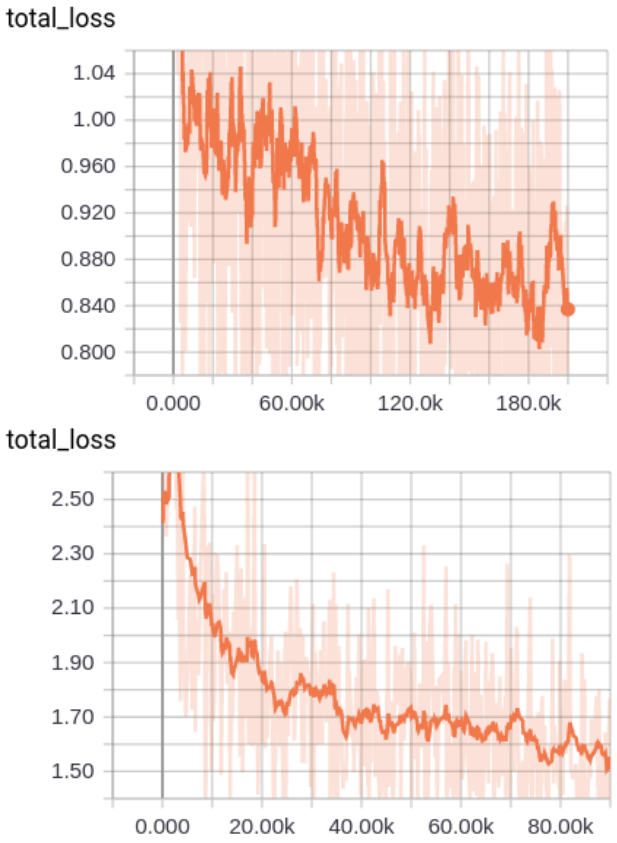

Figure 2: The first image shows total loss as computed by the SFM learner, and the second one is the total loss per iteration calculated using our pipeline

Method	Error Metric			
	Abs Rel	Sq Rel	RMSE	RMSE log
SFMLearner	0.208	1.768	6.856	0.283
Ours-Resnet	0.431	2.455	8.014	0.331
Ours-SSIM loss	0.212	1.699	6.596	0.271
Ours-SSIM+Resnet	0.251	1.823	7.351	0.318

Table I: Single-view depth results on the KITTI dataset generated from the script provided

REFERENCES

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

[2] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1983–1992, 2018.