Project 4: Learning Structure from Motion — An Unsupervised Approach

Nitin Suresh
School of Electrical and
Computer Engineering
University of Maryland - College Park
UID: 113638855

Jo Shoemaker
Department of Computer Science
University of Maryland - College Park
UID: 115506787

I. Introduction

In this project, we estimate the depth and pose information from a sequence of monocular images using an unsupervised approach. The project builds off of the SfMLearner approach by David Lowe's team at Google. The task of estimating depth information from a monocular sequence of images is pretty challenging, and requires using view reconstruction as supervisory signals. Unsupervised depth estimation usually works by estimating depth maps and projecting onto adjacent views, and then training the model by minimizing image reconstruction error. In the next section we present our contributions to improving the technique

II. Proposed changes for improved monocular depth and pose estimation

We have implemented changes in the loss function, data augmentation, and network architecture. For the loss function, the first change we made is in the pixel loss calculation. In SfMLearner, the reprojection error over all the source images is averaged as the total pixel error. This causes issues with pixels that are occluded/disoccluded, between the source and target images. Instead of averaging the error over all source images, we use the minimum photometric error over all source images, as in the equation below [1].

$$L_p = min_{t'} pe(I_t, I_{t' \to t})$$

The second change implemented in the loss function is the usage of an auto-masking method that removes the effect of stationary pixels in the sequence, which lets the network overcome effect of objects moving at camera velocity and frames where the camera is stationary [1]. The mask is binary with values in $\{0, 1\}$. A logical way to think about this is to ignore loss of pixels where warping the source images causes an increase in reprojection error. Such pixels between the source and target image indicate the conditions specified above, and the pixels are thus masked out using the equation below. Data augmentations are carried out by random changes in brightness, contrast, hue and saturation jitter.

$$\mu = [min_{t'} pe(I_t, I_{t'}) < min_{t'} pe(I_t, I_{t' \to t})]$$

We changed the architecture of the monocular depth network to a ResNet [2], using the same network structure as [3], which improved on SfMLearner's results. The encoding portion of the depth network—which in the baseline is comprised of 14 convolutional layers that halve in size every two layers—is replaced by five sets of three layers with "shortcut" connections, each half the size of the previous one. Since this is a deep network, it is expected that adding residual connections will improve learning.

III. Results

Table I displays our results on the evaluation data for pose and depth estimation.

Method	Error Metric			
	Abs Rel	Sq Rel	RMSE	RMSE log
Baseline	N	N	N	N
+ResNet	N	N	N	N
+Res+Loss	N	N	N	N
Baseline	N	N	N	N
+ResNet	N	N	N	N
+Res+Loss	N	N	N	N

TABLE I
Top: Depth Estimation Results. Bottom: Pose Estimation Results

References

[1] C. Godard, O. Mac Aodha, and G. Brostow, "Digging into self-supervised monocular depth estimation," *arXiv preprint arXiv:1806.01260*, 2018.
[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
[3] Y. Zou, Z. Luo, and J.-B. Huang, "Df-net: Unsupervised joint learning of depth and flow using cross-task consistency," in *European Conference on Computer Vision*, 2018.