Helping Robots see Better: Depth Maps and Segmentation

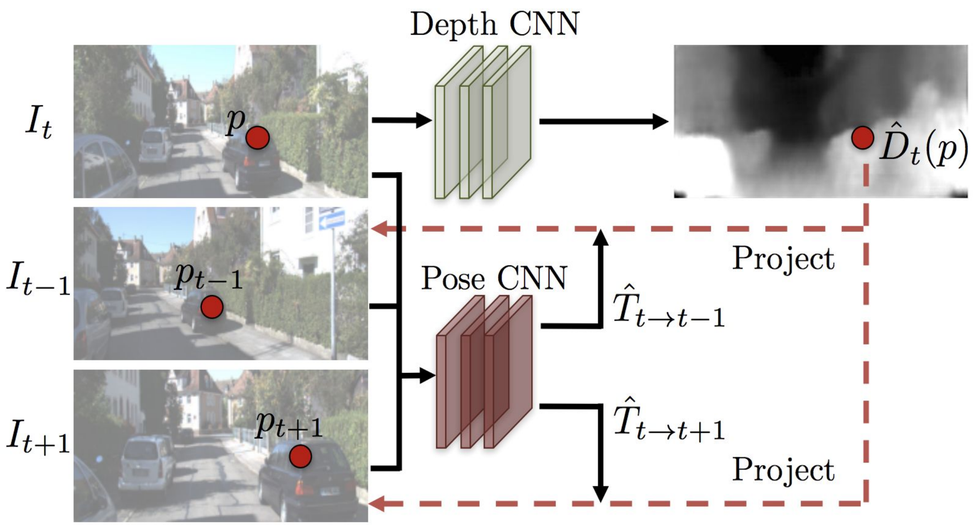
Aarash Heydari, Abhishek Mangla, Sheng-yu Wang
Team Name: group0

Introduction

The baseline model jointly trains these tasks:

1. (Multi-view) Egomotion
2. Single-view Depth Estimation

The two task networks can be used independently at test time.



Goal: Improve the sharpness of the current model's Depth Estimation network by promoting depth gradient at boundaries of objects in the image.

Final Model

1. Unsupervised training of Depth/Pose nets using View Synthesis.
2. Use Mask R-CNN to segment image into different objects
3. Add additional loss to strengthen depth map around object boundaries

View Synthesis Loss

$$\mathcal{L}_{vs} = \sum_{<I_1,\dots,I_N>\in\mathcal{S}} \sum_p \hat{E}_s(p)|I_t(p) - \hat{I}_s(p)|$$

Reconstruction Loss

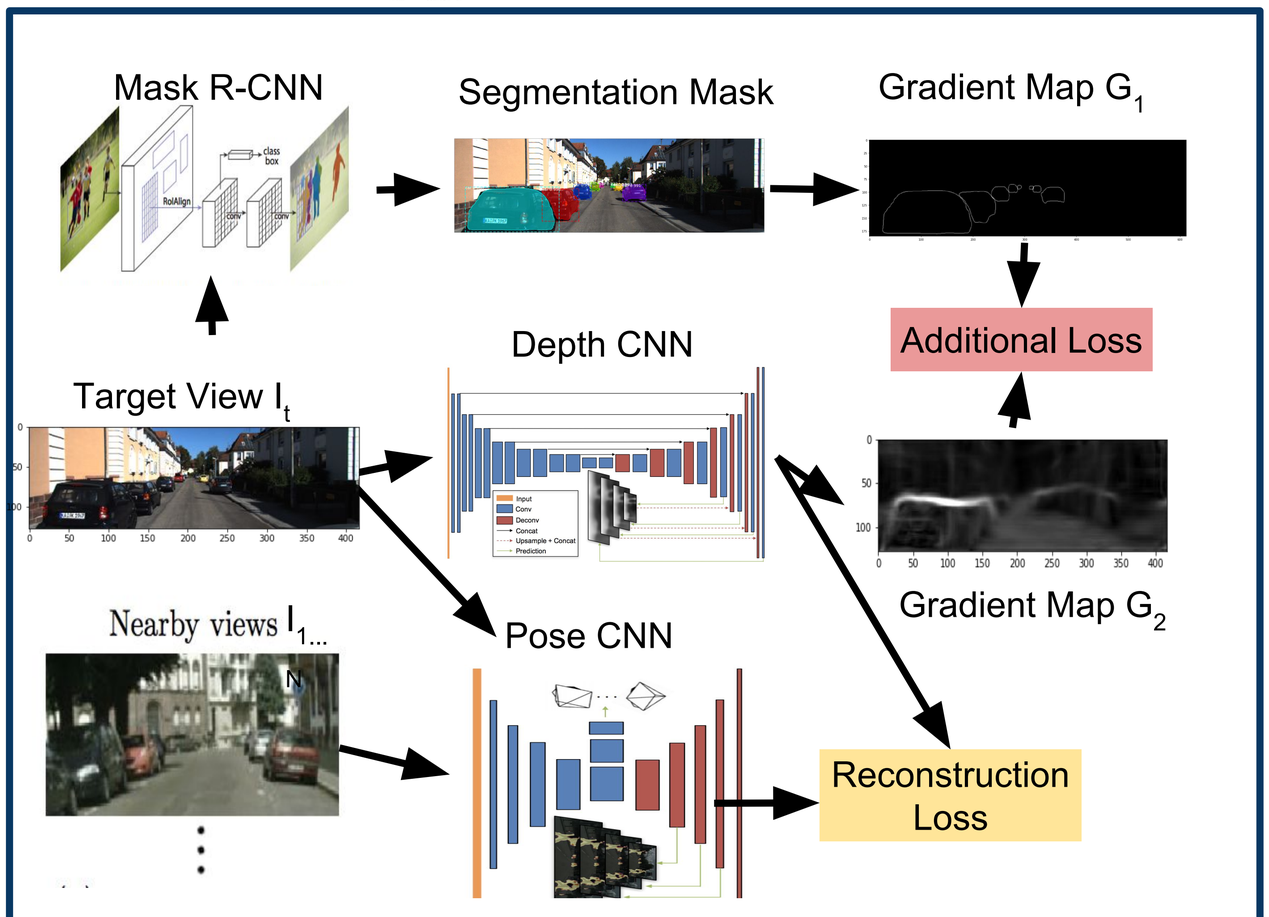
$$\mathcal{L}_{final} = \sum_l \mathcal{L}^l_{vs} + \lambda_s \mathcal{L}^l_{smooth} + \lambda_e \sum_{<I_1,\dots,I_N>\in\mathcal{S}} \mathcal{L}_{reg}(\hat{E}^l_s)$$

(Indexes over different image scales)

- λ_s, λ_e are hyperparameters weighting the smoothness regularization and explainability regularization respectively.
- Smoothness: Penalize the L1 norm of second-order pixel gradients.
- Explainability: View Synthesis assumes a static scene. E_s is a per-pixel mask to discount regions of occlusion and scene dynamics, giving slack to factors not considered by the model.

Additional Loss:

$$\sum_{(x,y)} G_1(x,y) \cdot \max(0, G_1(x,y) - G_2(x,y))$$



Data Processing

Target image: 1226x370

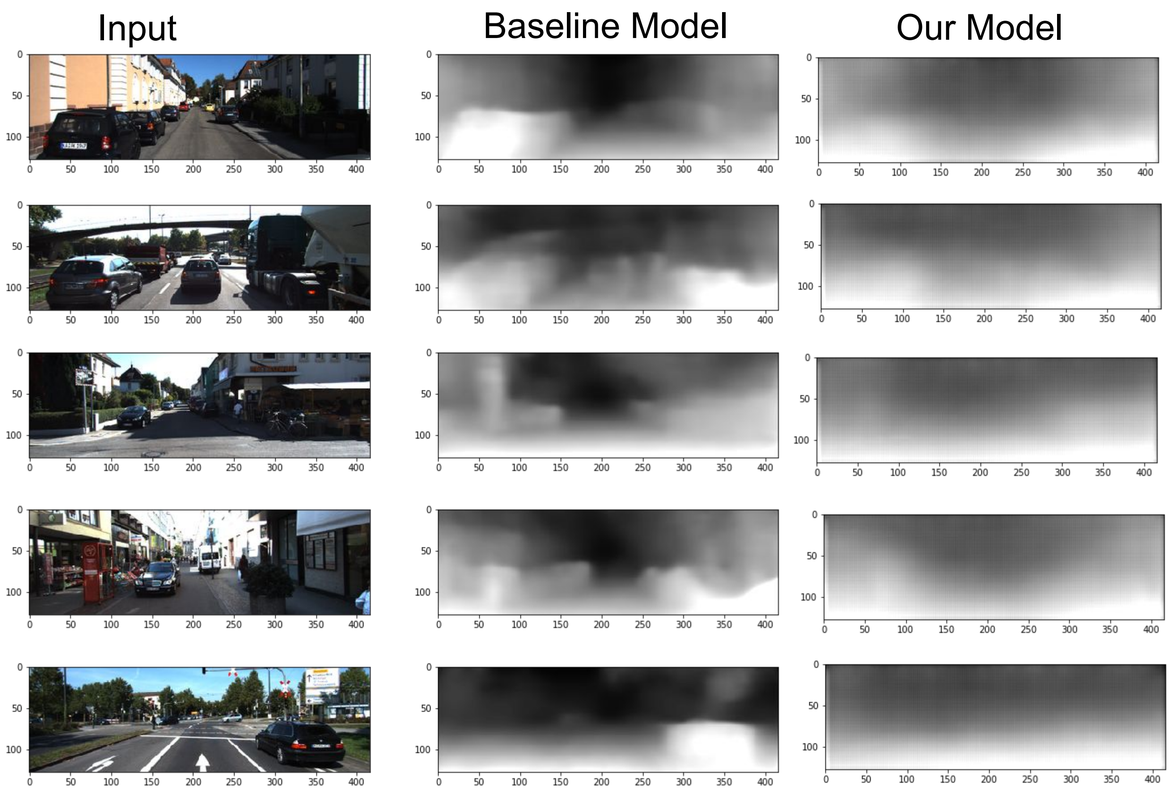
416x128

JPEG compression

Stack the target image, its source images and object boundaries output by Mask R-CNN as a single training sample

For each training sample, include camera intrinsics, including focal lengths and principle points.

Results



- The results output by our model are worse than the baseline model. We believe this is because we calculated the boundary matching loss at multiple scales of the DISP-net output. Although the reconstruction loss is calculated in multiple scales in the original model, it doesn't work that well for our boundary matching loss.
- Moreover, we need to also try to train with more epochs and adjust the hyperparameters.

Challenges

- Gradients of depth maps naturally skyrocket exponentially near the horizon of an image, which made it difficult to match the depth gradient with the object boundary. To resolve this, we instead match gradients with the reciprocal of the depth, the disparity.
- Mask R-CNN does not segment trees, only objects such as pedestrians and cars. We only improve sharpness of depth maps on images that contains objects segmented by mask R-CNN

References

- Zhou, Tinghui, et al. "Unsupervised learning of depth and ego-motion from video." *CVPR*. Vol. 2. No. 6. 2017.
- He, Kaiming, et al. "Mask R-CNN." arXiv:1703.06870, 2017.
- A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite.

UC Berkeley - Computer Science Department - Berkeley, CA

amangla , aheyd, sheng-yu_wang @ berkeley.edu