Project Paper Review

Yi Lu

March 2025

1 Paper Basic Information

1.1 Reference

Nekrasov, V., Dharmasiri, T., Spek, A., Drummond, T., Shen, C., & Reid, I. (2019, May). Real-time joint semantic segmentation and depth estimation using asymmetric annotations. In 2019 International Conference on Robotics and Automation (ICRA) (pp. 7101-7107). IEEE.

1.2 Paper Title

Real-Time Joint Semantic Segmentation and Depth Estimation Using Asymmetric Annotations

1.3 Authors

Vladimir Nekrasov, Thanuja Dharmasiri, Andrew Spek, Tom Drummond, Chunhua Shenand and Ian Reid

2 Goal of the paper

The author aimed to adapt a single model to perform multiple tasks in real-time with the limitation of asymmetric datasets in both indoor and outdoor settings, with the main breakthrough as real-time performance achievement and asymmetric and uneven ground truth annotations. The main reason for them to come up with such a research direction is that, few work has been done on how to perform multiple tasks in a single model, and significant obstacle exists in the way as asymmetric dataset annotations.

3 Data

NYUDv2: It is an indoor dataset with 40 semantic labels, containing 1449 RGB images with both segmentation and depth annotations(795 of them are training images, 654 of them are validation images). All of the images were annotated by a teacher network.

KITTI: an outdoor dataset containing 100 images for training and 46 images for testing, pre-processing was conducted and only 6 semantic classes were preserved out of 11. ResNet-38 model was adopted as the teacher network for training image annotation.

4 Algorithm

It deployed backbone network together with joint semantic segmentation and depth estimation. For each dataset, the network was pretrained on the large dataset and fine-tuned on the smaller dataset. In detail, it employed Light-weight RefineNet as the baseline architecture, built upon MobileNet-v2 classification network. The researcher replaced 1×1 convolution in the last CRP(Chained Residual Pooling) block to decrease the influence. They also appended two additional convolutional layers for each task to achieve joint semantic segmentation and depth estimation. Lastly, they relied on an expert model to estimate missing labels for training purpose, which plays a role as the ground truth label. The output of the model is the corresponding label or depth estimation of each pixel.

5 Statistical Results

They mainly used mIoU(mean intersection of union) to evaluate semantic segmentation performance, while using RMSE(root mean squared error) to evaluate depth estimation performance. Besides, they also evaluated the speed of the model by standard of milliseconds.

Ending with 42.02% mIoU and $0.565m$ RMSE on NYUDv2 and 87.02% mIoU and $3.453m$ RMSE, it outperforms most models of previous researchers as a competitive model with a smaller model in real-time. It is also the most efficient method, as it only took 12.8 ms on average, which is much shorter than other models.

6 Interpretation

In general, I believe this paper has made incredible contribution to computer vision problem in autonomous driving in terms of depth estimation and semantic segmentation. The most important contribution of this paper is that it shows it can utilize previous large models and train a single model on multiple datasets in real-time with a parsimonious model size.