Cityscapes Segmentation

K.H.W. Stolle k.h.w.stolle@student.tue.nl

TABLE I
TARGET LABELS

Category	Labels
Flat	road, sidewalk, parking, rail track
Human	person, rider
Vehicle	car, truck, bus, on rails, motorcycle, bicycle, trailer
Construction	building, wall, fence, guard rail, bridge, tunnel
Object	pole, pole group, traffic sign, traffic light
Nature	vegetation, terrain
Sky	sky
Void	ground, dynamic, static

***Abstract*—In this paper, a possible solution for the Cityscapes Pixel-Level Segmentation task is discussed. The U-Net architecture is taken as a baseline.**

***Index Terms*—Cityscapes, Segmentation, CNN**

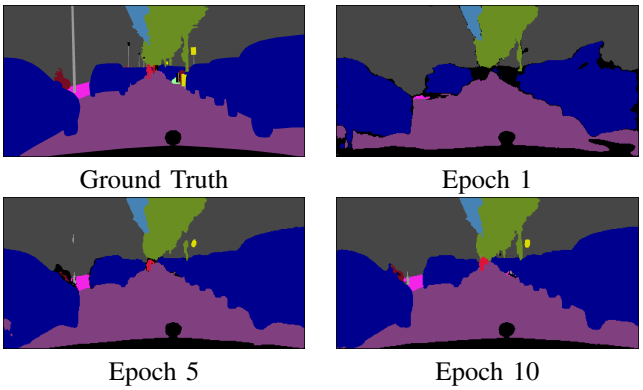

Fig. 1. Ground truth and predicted masks after training for 1, 5 and 10 epochs of the sample `munster_000098_000019_leftImg8bit.png` using the U-Net model with thresholds and data augmentation

I. Introduction

The Cityscapes Dataset involves a large number of pictures taken from the front of a car while driving through various cities in Germany [1].

In this paper, a practical solution to the pixel-level semantic labeling task is discussed. This involves predicting a per-pixel semantic labeling of the image without consiering higher-level object instance or boundary information.

First, a baseline is set using an off-the-shelf architecture for semantic segmentation. This baseline network is then improved by data augmentation and iterative design, after which this solution is tested. The results of testing are finally interpreted and discussed.

II. Method

A. The dataset

As with any machine learning project, first the available data has to be parsed into a practical format.

The dataset consists of a collection of PNG-encoded images (the input) and a corresponding segmentation mask (the ground truth). The segmentation mask is formatted as a JSON-document containing a list of objects that pair a label with a polygon that describes the shape of this object in the image.

When an item in the dataset is requested, the PNG-file is read and converted to an array of matrices, where each channel (red, green and blue) is an element of the array. The polygons file corresponding to this item is then loaded and the polygons are converted to rasters and combined into a single image, with each pixel of this image corresponding to one label (see Table I) or nothing.

B. Setting a baseline

The baseline implementation sets a reference point to improve upon and score our method against.

The architecture U-Net is suitable for semantic segmentation of biological cells under a microscope [2]. In this paper, the same implementation is used as a baseline for the segmentation of the cityscapes.

Our implementation of the UNet is based on [3] with some modifications to work with the Cityscapes Dataset.

C. Data augmentation

A straightforward way to improve the accuracy of the model is to increase the size of the training set by applying a set of transforms. The following transforms were used:

- Zoom & rotate
- Mirror over the vertical axis

D. Measuring performance

In order to measure how well the network is performing, the Intersection-over-Union (IoU) metric is implemented.

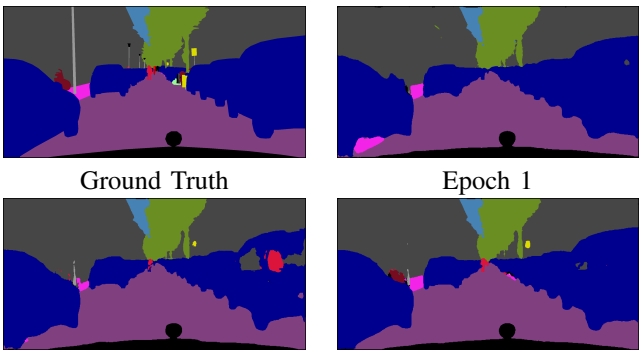

Fig. 2. Ground truth and predicted masks after training for 1, 5 and 10 epochs of the sample `munster_000098_000019_leftImg8bit.png` using the U-Net model with reduced dimensions

E. Network architecture

F. Tweaking hypterparameters

III. Results

A. Baseline

B. Data augmentation

C. Threshold

IV. Discussion

V. Conclusion

References

[1] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: http://arxiv.org/abs/1505.04597

[3] milesial, "U-net: semantic segmentation with pytorch," https://github.com/milesial/Pytorch-UNet, 2018.