Project Update 2 - Motion Segmentation

Dieker, Jonas and Hiebl, Christian

Informatics - Technische Universität München

Introduction

As discussed in last week's meeting we started off with a supervised network and an already existing dataset *KittiMoSeg Masks* with 1200 images. This dataset has a lot of stationary vehicles and almost no dynamic objects. Afterwards we used the *Extended KittiMoSeg Masks* as described in FuseModNet [1] which extends the MoSeg dataset to over 12000 images. Again, a lot of data was from sequences where the test car was driving through a neighborhood of only parked cars. With visual inspection we used only 12 out of about 30 sequences.

Neural Network Architecture

We used a U-Net structure [2] using skip connections, feature maps of sizes 64, 128, 256 and 512 as well as the BatchNorm which was not originally in the U-Net implementation but very common in autoencoder structure now. The network takes two successive concatenated RGB frames as its input. The 6 channel image was then fed into a U-Net architecture.

Focal loss vs BCE Loss We have first a binary cross entropy with a sigmoid activation in the last layer to predict the binary pixels. The network would overfit to our training data. However, the network would only predict a black mask for all images. Additionally, these all black predictions would lead to a 100% accuracy on a single validation sample since this sample was coincidentally all stationary pixels. This lead to us to switch to the extended dataset as well as use the focal loss, which would weigh the hard negative samples more, such that the network is forced to train on the dynamic pixels instead of predicting all black pixels.

Tensorboard We augmented our train script to use Tensorboard for easy training and validation loss curves, as well as ground truth masks and prediction masks from the validation step.

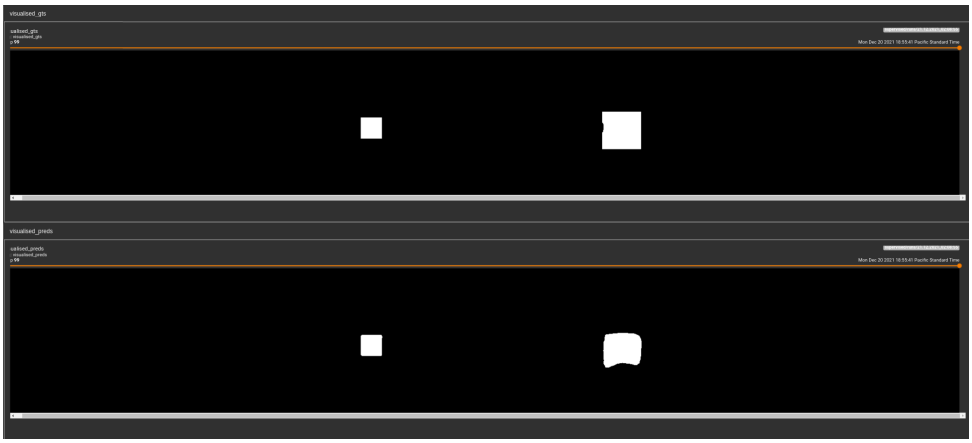

Figure 1: Ground truth masks (top) and predicted masks after 100 epochs (bottom)

The above image shows 100 epochs of training with a relatively high learning rate of 5e-3 using the annotated Kitti sequence 05 which has about 150 images, of which only 80% were used for training.

Dataset The datasets between the regular and extended Kitti Masks were different in folder structure so two dataset classes were implemented which could be used with the same dataloader in the train file. It was ensured that the pair of images would be sequential and stem from the same Kitti sequence.

More Data - CARLA

In order to have more control over the data and also collect more, the next step would be to collect our own data using Carla. Carla 0.9.13's instance segmentation would be useful to distinguish between different instances but we propose the following way to obtain our ground truth motion segmentation masks.

Use two consecutive optical flow images. Use semantic segmentation to use keypoint matching of stationary pixels (from buildings streets etc). This will give the static scene flow of the ego vehicle. The potentially dynamic objects, classified as vehicle or pedestrian in Carla can be classified as dynamic or static by checking descriptors in the semantic mask against the static scene flow. Here instance segmentation could be beneficial if vehicles' segmentation masks overlap. After checking the instance's descriptors, the whole instance mask can be marked as static/dynamic.

References

[1] Hazem Rashed, Mohamed Ramzy, Victor Vaquero, Ahmad El Sallab, Ganesh Sistu, and Senthil Kumar Yogamani. "FuseMODNet: Real-Time Camera and LiDAR Based Moving Object Detection for Robust Low-Light Autonomous Driving". In: *2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*. IEEE, 2019, pp. 2393–2402.

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *CoRR* abs/1505.04597 (2015). arXiv: 1505.04597.