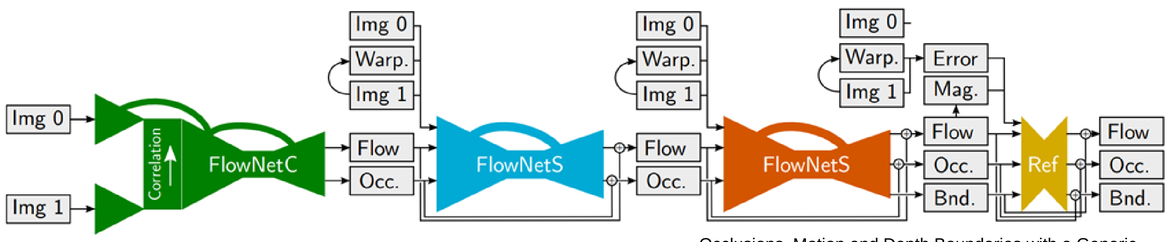
Modalities for semantic segmentation

Lukas Boschert, Simon Ging, Hanna Siemund

Summary

Partitioning an image into meaningful sections and assigning the right class label is a common task in computer vision, needed in many different application fields.
This poster analyses the influence of modalities given as additional input to a semantic segmentation task.
In particular, we focus on optical flow, motion boundaries and occlusions as well as combinations of them. For the main part of the prediction, we use the architecture of PSPNet[1], whereas a pretrained FlowNet3[3] network estimates the required modalities.
Qualitative and quantitative results are provided which illustrates the effects of different inputs on the network's performance. Additional we append an evaluation on class balancing and different batch sizes.

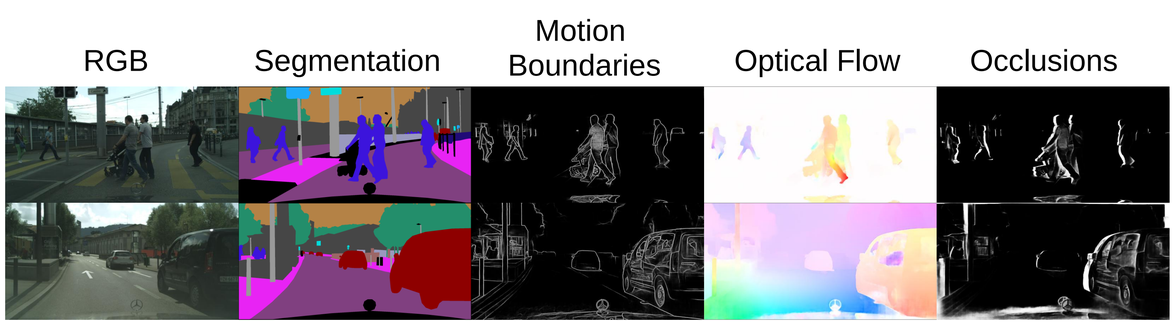
FlowNet3: Optical Flow



Occlusions, Motion and Depth Boundaries with a Generic Network for Disparity, Optical Flow or Scene Flow Estimation

FlowNet3[2] does not only predict Optical Flow but also Occlusions and Motion Boundaries. The flow is computed stepwise. After each step the image is warped with the current flow prediction to create a smaller displacement for the next iteration.

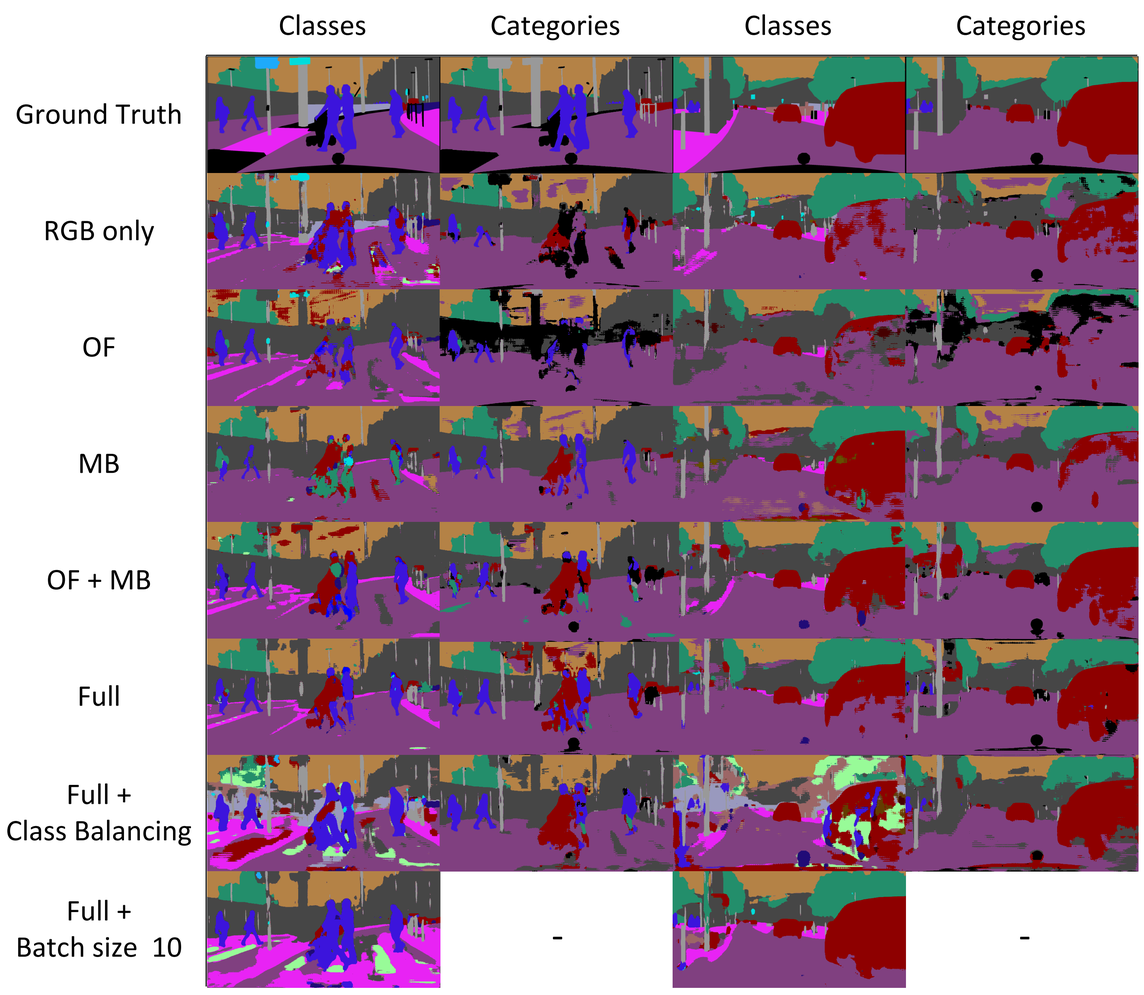
Dataset



The Cityscapes Dataset for Semantic Urban Scene Understanding

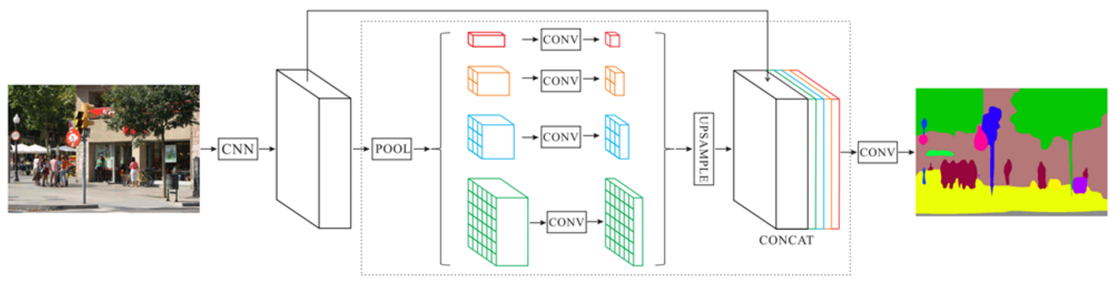
- cityscapes dataset[3]: complex street scenes from 50 different cities
- 30 different classes (20 relevant for prediction), 8 different categories
- 1936 images split into 1672 training and 264 test data points
- optical flow, motion boundaries, occlusions from FlowNet3
- generated by using two consecutive video frames with a temporal distance of 60ms

Qualitative Evaluation



Ground truth and predictions for different models for two data points.

PSPNet: Pyramid Scene Parsing Network



Pyramid Scene Parsing Network

PSPNet[1] introduces the Pyramid Pooling Module to add a global context to the feature vector. The 4-level pyramid uses pooling over the whole, half of, and small portions of the feature map. Using average pooling for the Pyramid Pooling Module empirically outperforms max pooling.

Implementation Details

For the CNN in our PSPNet we use a ResNet50.

Optimizer: Adadelta
- learning rate 1
- decay factor 0.95

Data Augmentation
- random rescaling between 0.5 and 2
- random crop 512 x 512

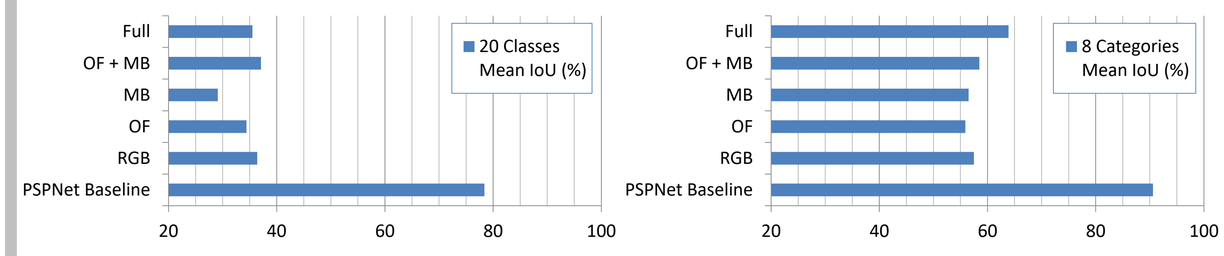
Training
- 80 epochs
- 3000 steps per epoch
- batch size 1

Nvidia GTX 1060 ~ 22min/epoch ~ 29h training

Quantitative Evaluation

	Classes		Categories	
Model	Accuracy (%)	Mean IoU (%)	Accuracy (%)	Mean IoU (%)
PSPNet[1] baseline	-	78.4	-	90.6
RGB only	78.0	36.4	86.0	57.5
Optical Flow (OF)	87.0	34.4	80.0	55.9
Motion Boundaries (MB)	75.6	29.1	84.1	56.5
OF + MB	79.5	**37.1**	83.8	58.5
Full (OF + MB + Occlusion)	77.3	35.5	87.2	**63.9**
Full + Class Balancing	63.0	27.8	84.0	61.1
Full + Batch size 2	80.1	39.9	-	-
Full + Batch size 4	80.5	**43.0**	-	-
Full + Batch size 8	81.9	41.2	-	-
Full + Batch size 10	81.4	40.4	-	-

Results for different experiments compared to the PSPNet[1] baseline.
Modalities: Models where evaluated with different combinations of Optical Flow, Motion Boundaries and Occlusions.
Class Balancing: Classes are weighted reciprocally to their relative occurrence in the training data.
Batch size: With increasing batch size, the steps per epoch are lowered to have all models see the same number of data points.



Visualization of experimental results for different inputs.

Conclusion

The quantitative results show that our models were not able to achieve similar results as the PSPNet[1] baseline. A reason for that can be the use of a ResNet50 trained from scratch in contrast to the pretrained ResNet101 they mention in their paper. We achieved the best performance using a batch size of 4, which needed at least 8 GB of GPU RAM for a Res Net50.
The model using optical flow plus motion boundaries as well as the full model reached highest values for mean IoU. This fits our expectation of modalities helping the network to make better decisions. Surprisingly, the RGB only approach slightly overcame the models which perceived just one modality.
Adding class balancing changed the predictions to the worse which can be clearly seen in the qualitative results. Especially in the case of class labeling, the model predicted the light green class way to frequently, which leads to the assumption of our weights being too extreme.
As conjectured, the use of larger batch sizes improved the results of our model.
Overall, giving multiple modalities to the network slightly improved the results over just using RGB input. Possibly due to noise in our optimization process, this trend is not clearly visible. However, we deem it likely this improvement can be observed in state-of-the art performance models as well.

1. H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, "Pyramid Scene Parsing Network", arXiv:1612.01105
2. E. Ilg, T. Saikia, M. Keuper, T. Brox, "Occlusions, Motion and Depth Boundaries with a Generic Network for Disparity, Optical Flow or Scene Flow Estimation", in European Conference on Computer Vision (ECCV), 2018
3. M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016

code https://github.com/gingsi/segflow