Supplementary Material for Domain Adaptive Semantic Segmentation with Self-Supervised Depth Estimation

Qin Wang[1] Dengxin Dai[1,2*] Lukas Hoyer[1] Luc Van Gool[1,3] Olga Fink[1]
[1]ETH Zurich, Switzerland [2]MPI for Informatics, Germany [3]KU Lueven, Belgium
{qwang,lhoyer,ofink}@ethz.ch {dai,vangool}@vision.ee.ethz.ch

In this supplementary, we provide additional analysis and implementation details for the proposed CorDA method.

1. Attention Visualization

To better understand how the learned correlation module plays a role in improving the segmentation performance, we visualize the attention map from depth to semantics in Figure A1. The areas shown in brighter colors indicate that more guidance from depth is used for the final semantic prediction. Certain objects including cars, sidewalks, and poles in general attract stronger attention, which corresponds well with the improved classes in Table 3 of the main paper. In Figure A1, we provide four examples, where the areas with locally increased attention (illustrated in the white squares) improve the model performance, compared to DACS which has no correlation learning module. An interesting observation is that *sidewalk* in general attracts more attention than its easily confusable counterpart *road*. This could be the source of our improvement on both classes.

2. Additional Visual Comparison

We provide further prediction examples in Figure A2 to qualitatively compare our method CorDA with the state-of-the-art DACS [10] as well as FDA [14]. We additionally show the corresponding stereo depth estimates for each image. Note that the pseudo depth estimates are only needed for training and are not used for inference. Performance on objects with strong geometric constraints such as sidewalks are improved, compared to both DACS and FDA.

3. Code Base

To ensure a fair comparison with DACS [10], we adopted the same code base as DACS [10] in all our experiments and added our task feature correlation module and the pseudo-label refinement. The task feature correlation module is implemented based on the PadNet [13] implementation from [11]. For the GTA-to-Cityscapes task, we apply the initial semantic decoder after the ResNet features, and use the initial semantic prediction as pseudo labels for the first 10% training iterations. We found that this helps the model to learn in early stages.

4. Reproducibility with Multiple Runs

We trained our CorDA model for five times on the GTA-to-Cityscapes task, the average performance and standard deviation is $56.9 \pm 0.5\%$ mIoU. This is based on the performance at the end of training (250k iterations) without early stopping. In all five runs, the model performance is significantly better than the state-of-the-art DACS's performance (52.1%). The reported number (56.6%) in the main paper is from the run with median performance. We enclose our code together with this supplementary material.

5. Details on Depth Estimation

We provide details on the generation process of the depth estimates, which are used as pseudo depth ground truth for Cityscapes and GTA5 in the proposed CorDA method.

5.1. Monocular Estimation

For self-supervised monocular depth estimation from image sequences [15], the so-called source image x_i is differentiably warped into the target image x_j based on camera motion and depth, both estimated by a neural network. The loss is calculated from the photometric error of the warped source image $x_{i\to j}$ and the real target image x_j and is back-propagated into the neural network for the weight update. In this work, we follow the implementation of Godard *et al*. [3]. In particular, we use a ResNet50 [4] backbone with a U-Net [7] decoder, which is trained for 200k iterations with a batch size of 4 and an initial learning rate of 1×10^{-5} for the encoder and 1×10^{-4} for the decoders. After 150k iterations, the learning rates are decreased by a factor of 10. In contrast to [3], we deploy an additional ASPP module [2] with dilation rates 3, 6, and 9 between

*The corresponding author

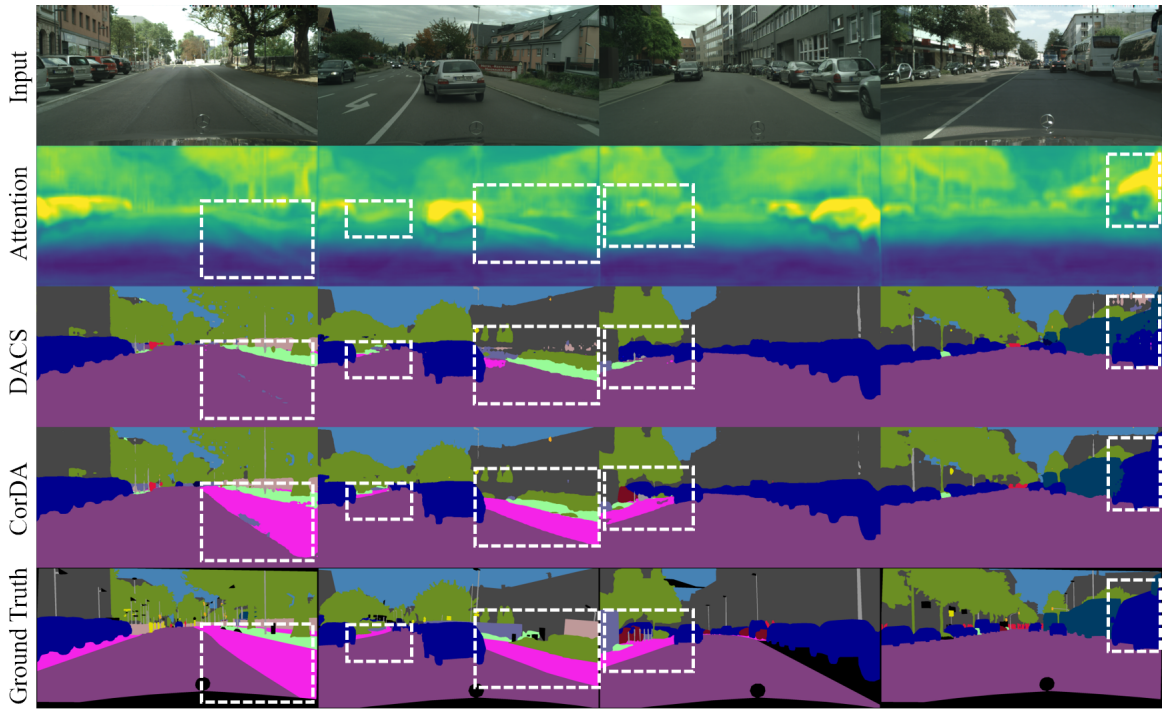

Figure A1. Visualization of our learned attention map. Examples are taken from the GTA-to-Cityscapes task.

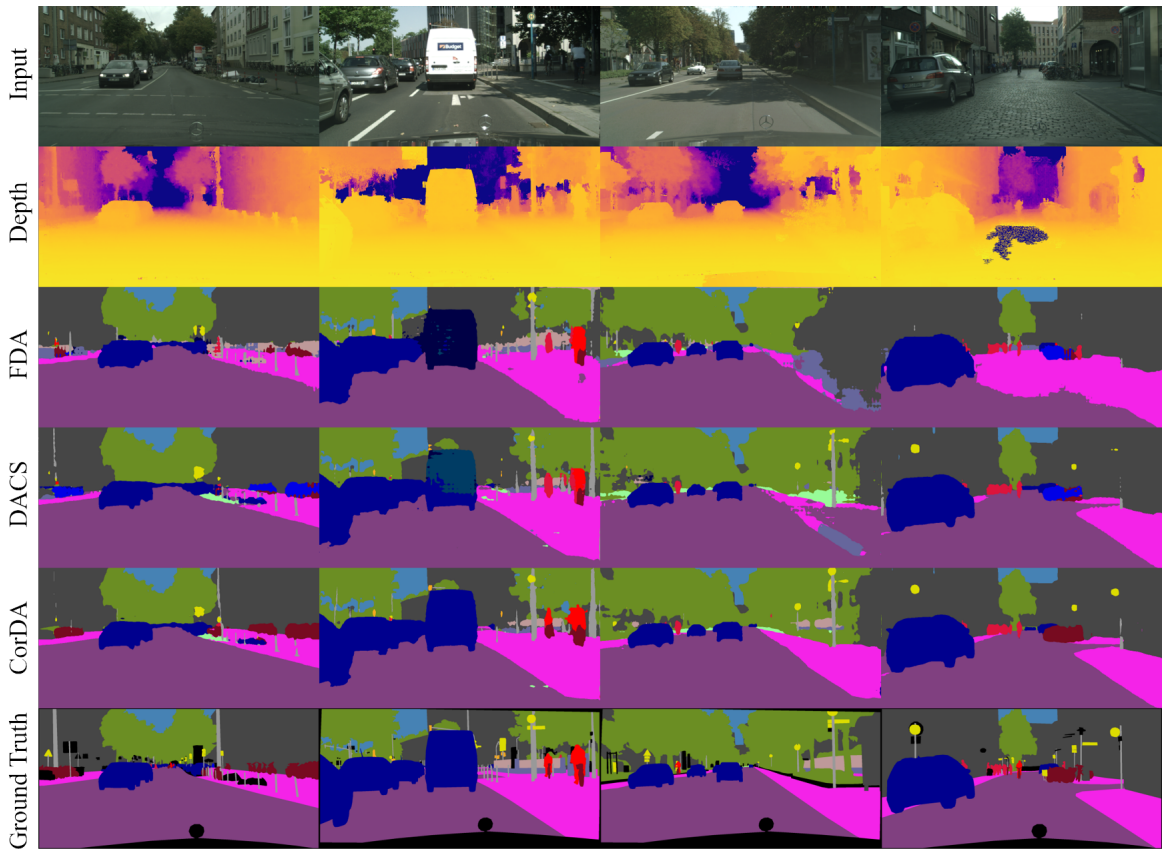

Figure A2. Semantic segmentation results on GTA-to-Cityscapes. We compare our method with DACS [10] and FDA [14].

encoder and decoder for multi-scale context feature aggregation, use BatchNorm [6] in the decoder for faster convergence, and apply random cropping of size 512×512 for data augmentation. This Monodepth2 model is trained on the image sequences.

5.2. Stereo Estimation

The depth estimation can also be generated from stereo pairs. In this work, we use the publicly-available stereo estimates generated by [9, 8]. Disparity maps are first estimated from stereo pairs using the Semi-Global Matching [5]. Camera focal length and the baseline are then used to convert disparity to initial depth. This initial depth with many missing values is then filled by the stereoscopic inpainting [12] approach. The unsupervised superpixels generated by SLIC [1] are used to guide the depth filling process. For more details of the generation process, we refer the readers to [9]. We use it because the generated depth is directly available for Cityscapes. However, our method is directly compatible with other depth estimation approaches.

References

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012. 3

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017. 1

[3] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1

[5] Heiko Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 807–814. IEEE, 2005. 3

[6] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015. 3

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1

[8] Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 687–704, 2018. 3

[9] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, Sep 2018. 3

[10] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1379–1389, 2020. 1, 2

[11] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Mti-net: Multi-scale task interaction networks for multi-task learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 527–543. Springer, 2020. 1

[12] Liang Wang, Hailin Jin, Ruigang Yang, and Minglun Gong. Stereoscopic inpainting: Joint color and depth completion from stereo images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008. 3

[13] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 675–684, 2018. 1

[14] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4085–4095, 2020. 1, 2

[15] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1851–1858, 2017. 1