Optimizing against Infeasible Inclusions from Data for Semantic Segmentation through Morphology

Supplementary Material

A. Additional Qualitative Comparisons

ADE20K. In Fig. 7 showing additional qualitative results on the ADE20K val set, we observe that in the prediction of the basic OCRNet [8] network on the first example, the building includes sky, which is rectified in the InSeIn-upgraded version of the network. In the second example, we observe that a segment of bed is infeasibly included in a cupboard segment, which is corrected by our InSeIn model. In both of the cases, the infeasible region is marked with *green* bounding boxes.

ACDC. In the first example of Fig. 8, we observe that a segment of sky is completely included in a building marked with *green* bounding boxes, which is a physical anomaly and is solved by InSeIn.

Cityscapes. In Fig. 9, we note that a rider segment corresponding to the rider's arm is infeasibly included in a bus segment in the prediction of Mask2Former [3], whereas InSeIn manages to correctly connect this part of the arm with the rest of the rider's body. In the second example, a segment of the truck is infeasibly included by the building segment. InSeIn rectifies the anomaly. The infeasible regions are marked with *green* bounding boxes.

B. Discussion over choosing SOTA models

We have chosen DeepLab series [1, 2], HRNet [6], OCRNet [8] because they are top-performing CNN-based methods for semantic segmentation. SegFormer [7] is chosen since it is a top-performing CNN and transformer-based method. Mask2Former [3] and OneFormer [5] are chosen as they are standard mask-based methods. SegMAN [4] is the latest transformer-based method outperforming Mask2Former. On the other hand, InSeIn is a light-weight (as it does not contain any learned parameter like these baselines) plug-and-play module, which, when plugged to these baselines and retrained on the 3 datasets, we have observed a consistent increase in the performance across the 3 datasets.

C. Additional discussion on class-wise comparison

From Tab.3, we can observe that out of 95 comparisons, there are 16 comparisons where our method has a lower IOU score than its corresponding baseline.

This is due to the few new False Positive cases arising in the retrained version with InSeIn. This happens since, while re-training with InSeIn, we are initialising the baseline model's weights randomly, which leads the optimisation to converge to different local minima than the original baseline trained without InSeIn in the loss landscape.

And as InSeIn only penalises infeasible pair inclusion, some of these False Positives do not belong to the category of infeasible inclusion, and they remain in the prediction feature map. For example, in Fig. 10, we can observe that in the re-trained version, a small segment of the person class is included by the road class (marked with green bounding box), which is a False Positive according to the ground truth, but it is not an infeasible inclusion with respect to the taxonomy of the Cityscapes dataset.

Table 10. Wall-clock time for one SGD iteration on full input images and GPU memory usage during training for a batch size of 4.

Model	Space (GB)	#Params	Time (sec.)
Mask2Former	21	216M	7.2
Mask2Former w/ InSeIn	41	216M	8.4
SegFormer-B4	8	64.1M	3.5
SegFormer-B4 w/ InSeIn	29	64.1M	4.0

D. Discussion over memory usage and time complexity

Tab. 10 demonstrates the memory usage and time complexity of InSeIn. The memory overhead induced by InSeIn only applies to *training*, while at inference, we do not require any additional memory compared to the baseline networks. The parameters remain the same as the baselines since InseIn is free of any learnable parameters.

References

[1] 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, 2018. 1

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 1

[3] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1280–1289, 2022. 1, 3

[4] Yunxiang Fu, Meng Lou, and Yizhou Yu. Segman: Omni-scale context modeling with state space models and local at-

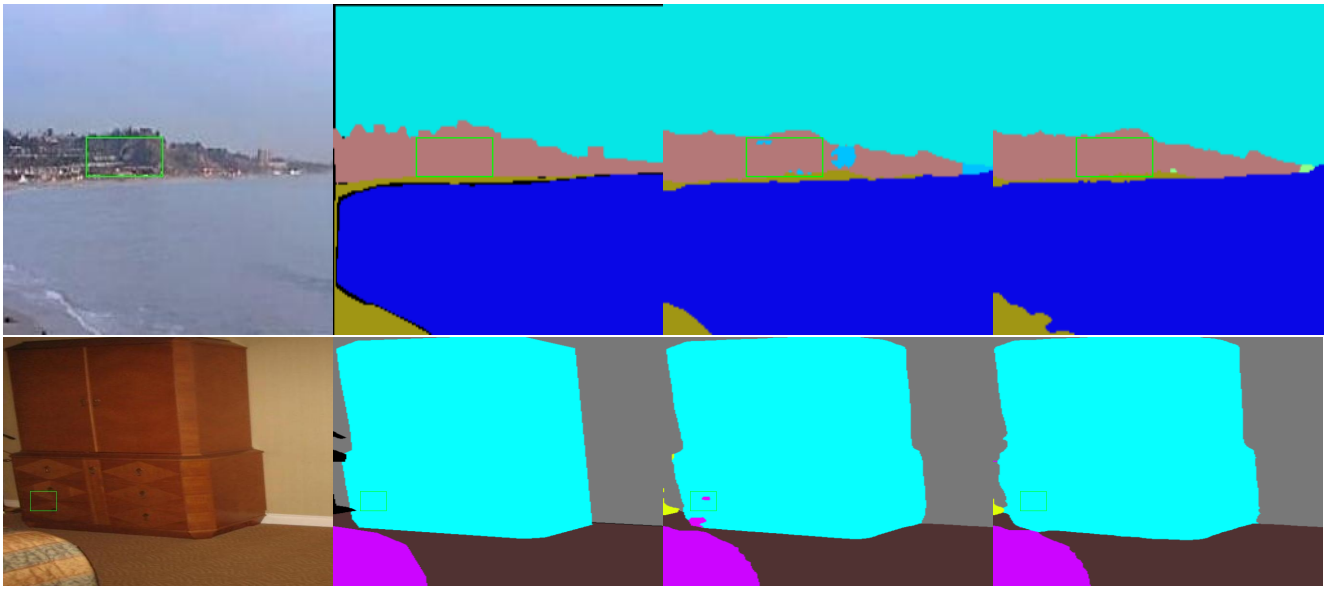

Figure 7. **Additional qualitative comparison on ADE20K.** From left to right: input image, ground-truth semantic labels, and predictions of OCRNet [8] and InSeIn. Best viewed on a screen and zoomed in.

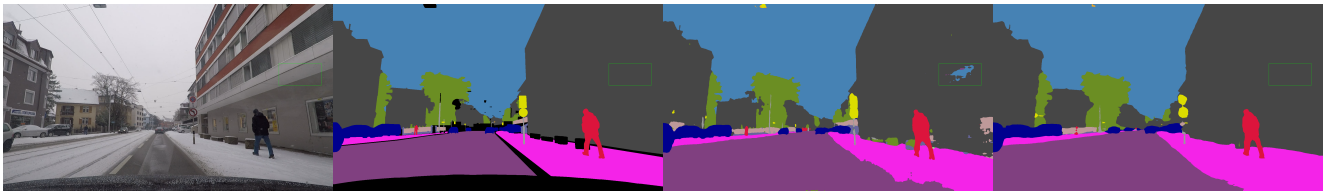

Figure 8. **Additional qualitative comparison on ACDC.** From left to right: input image, ground-truth semantic labels, and predictions of SegFormer-B4 [7] and InSeIn. Best viewed on a screen and zoomed in.

tention for semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 1

[5] Jitesh Jain, Jiachen Li, MangTik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. OneFormer: One Transformer to Rule Universal Image Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[6] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2019. 1

[7] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1, 2

[8] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *ECCV*, 2020. 1, 2, 3

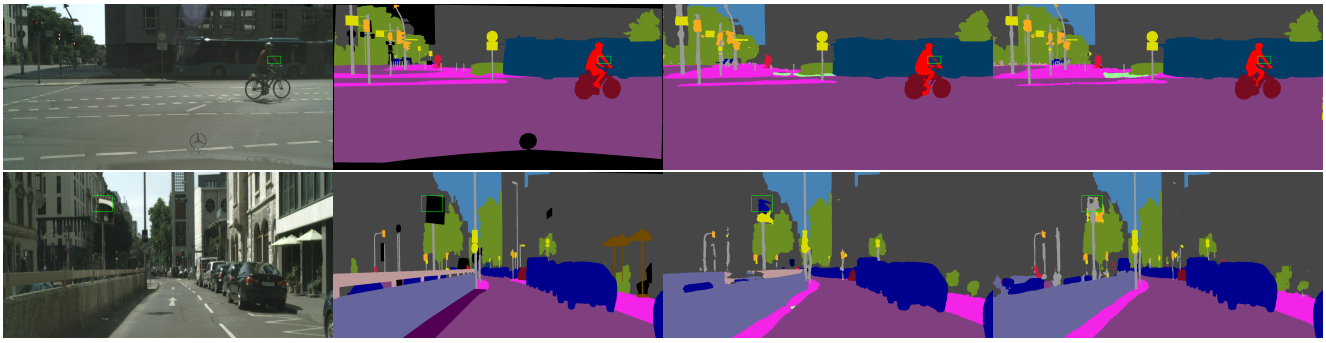

Figure 9. **Additional qualitative comparison on Cityscapes.** From left to right: input image, ground-truth semantic labels, and predictions of Mask2Former [3] and InSeIn. Best viewed on a screen and zoomed in.

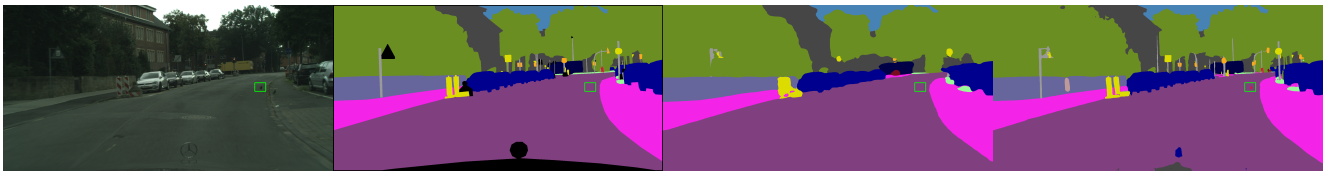

Figure 10. **Failure case in class-wise IOU comparison for Cityscapes.** From left to right:input image,ground truth, baseline-prediction OCRNet [8] and InSeIn. Best viewed on a screen and zoomed in.