Supplementary Material: Attention Decomposition for Cross-Domain Semantic Segmentation

Liqiang He and Sinisa Todorovic

Oregon State University, Corvallis, OR 97330, USA
{heli,sinisa}@oregonstate.edu

This supplementary material complements our main paper with the following analysis and qualitative results:
1) Comparison to DAFormer.
2) Failure Cases.
3) Societal Impact and Future Work.
4) Additional Qualitative results.

1 Comparison to DAFormer

Fig. 1 compares the results of ADFormer and DAFormer [2] on the SYNTHIA [6]→Cityscapes [1] domain shift. The rightmost two columns also compare the results of the two models in the respective zoomed-in image windows. In the top row, DAFormer predicts the part of the bus as 'vegetation' and 'car', and in the middle row, DAFormer fails to correctly predict the rectangular shape of the 'traffic sign'. In the bottom row, DAFormer oversegments the object in the white dashed window as a mix of 'car', 'pole', 'building', and 'traffic sign'. ADFormer appears to perform better for these examples, and, in particular, in the bottom row, ADFormer predicts the object in the zoomed-in window as 'car'. The three examples demonstrate that ADFormer is successful in segmenting whole objects, whereas DAFormer tends to oversegment them. One reason is that ADFormer uses the query tokens for predicting whole segmentation masks of the classes present in the image, while DAFormer is aimed at pixel-wise prediction.

2 Failure Cases

Fig 2 shows two failure cases of ADFormer on the GTA [5]→Cityscapes domain shift. In the two white dashed windows, ADFormer incorrectly predicts the 'traffic sign' as 'pole'. This suggests that the query tokens of ADFormer were not learned well to reliably distinguish differences between 'traffic sign' and 'pole'. To address this issue, one could try to better fine-tune the hyper-parameter of the query classification loss in the set loss.

3 Societal Impact and Future Work

Societal Impact: Our work makes fundamental contributions to Computer Vision and Deep Learning, and as such may give rise to a wide range of societal

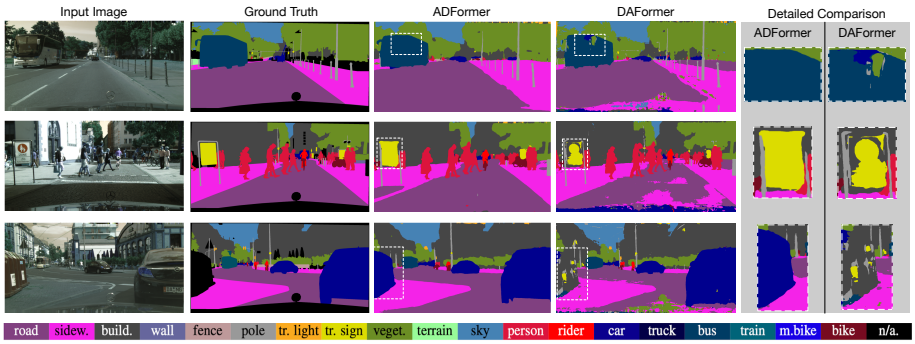

Fig. 1: Comparison of ADFormer and DAFormer [2] on the SYNTHIA→Cityscapes domain shift. The white dashed windows in the semantic segmentations of ADFormer and DAFormer are zoomed-in in the two rightmost columns, for a detailed comparison.

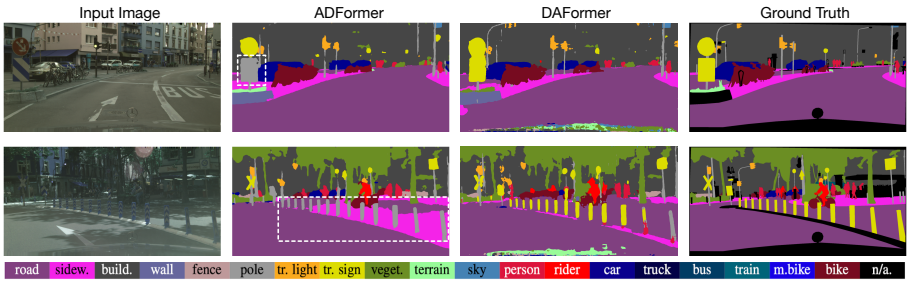

Fig. 2: Failure case analysis on GTA→Cityscapes domain shift.

benefits. As any new model for semantic segmentation, ours could be potentially used for malicious applications of computer vision.

Future Work: To the best of our knowledge, our work represents the first attempt to use a "proposal-based" transformer for cross-domain semantic segmentation. "Proposal-free" transformers constitute the predominant framework of SOTA approaches. For the "proposal-free" framework, the literature presents many advances, such as, e.g., high resolution domain adaptation [3] and masked-image consistency [4]. It seems promising to explore adapting these advances to "proposal-based" transformers in the future.

4 Additional Qualitative results

Fig. 3 and Fig. 4 illustrate additional example results of ADFormer and DAFormer [2] on the GTA→Cityscapes and SYNTHIA→Cityscapes domain shifts. As we

can see, ADFormer has better semantic segmentation than DAFormer on these two domain shift scenarios.

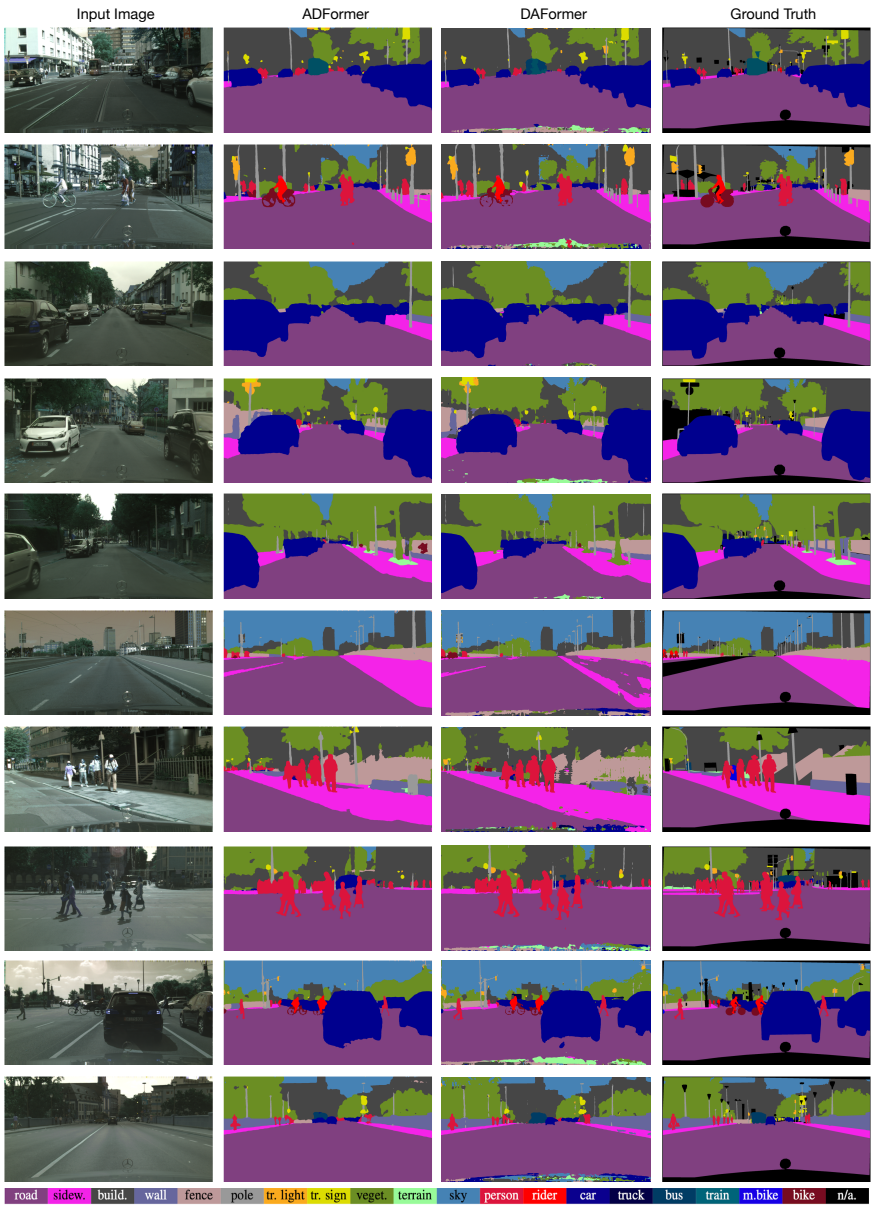

Fig. 3: Inference on example validation images from Cityscapes when ADFormer and DAFormer [2] are trained on GTA→Cityscapes.

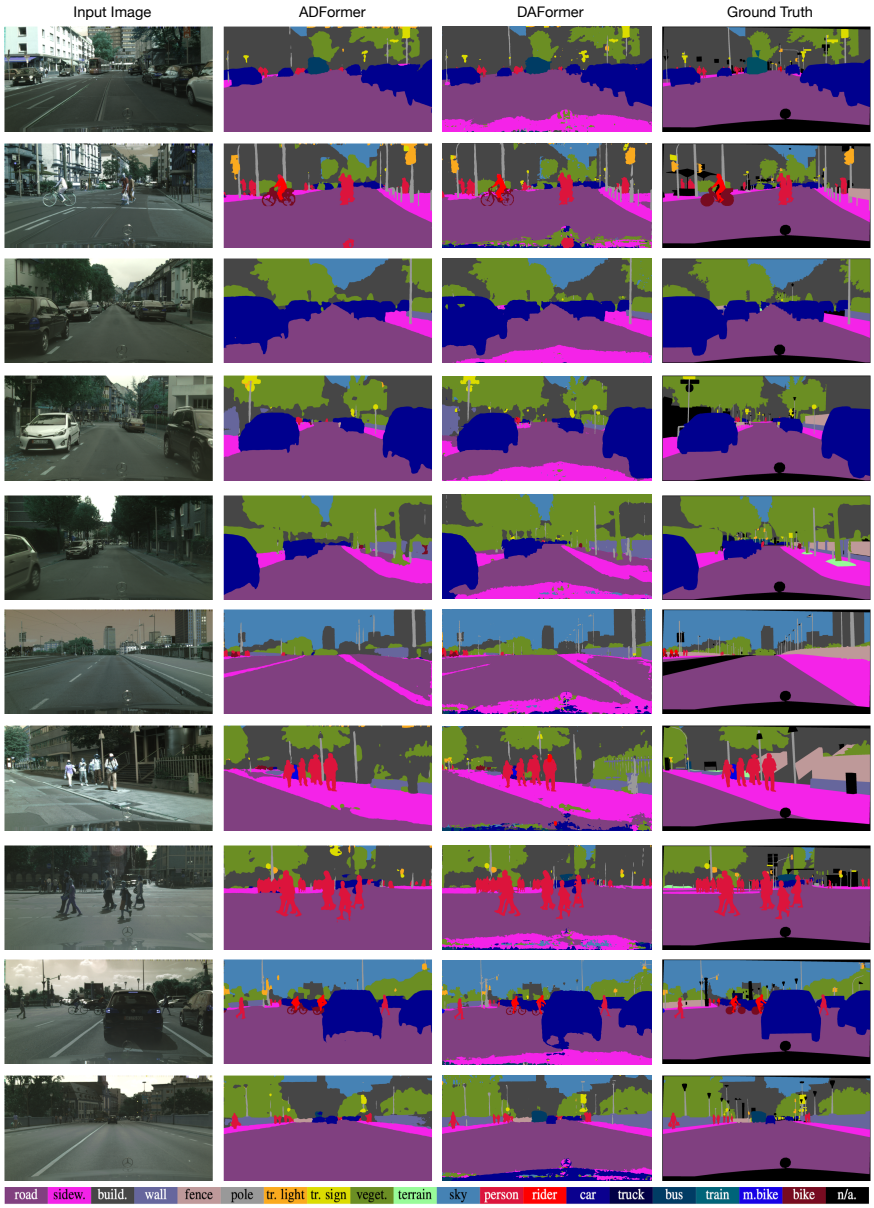

Fig. 4: Inference on example validation images from Cityscapes when ADFormer and DAFormer [2] are trained on SYNTHIA→Cityscapes.

References

1. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016) 1
2. Hoyer, L., Dai, D., Van Gool, L.: DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9924–9935 (2022) 1, 2, 3, 4
3. Hoyer, L., Dai, D., Van Gool, L.: Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In: European Conference on Computer Vision. pp. 372–391. Springer (2022) 2
4. Hoyer, L., Dai, D., Wang, H., Van Gool, L.: MIC: Masked image consistency for context-enhanced domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11721–11732 (2023) 2
5. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 102–118. Springer (2016) 1
6. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3234–3243 (2016) 1