A Downstream Task Informed Domain Adaptation GAN for Semantic Street Scene Segmentation

Annika Mütze
Stochastics Group, IZMD
University of Wuppertal
Wuppertal, Germany
muetze@uni-wuppertal.de

Matthias Rottmann
Stochastics Group, IZMD
University of Wuppertal
Wuppertal, Germany
rottmann@uni-wuppertal.de

Hanno Gottschalk
Stochastics Group, IZMD
University of Wuppertal
Wuppertal, Germany
hanno.gottschalk@uni-wuppertal.de

***Index Terms*—Generative Adversarial Networks, Domain adaptation, semantic segmentation, semi-supervised learning, image-to-image translation**

For automatically understanding complex visual scenes from RGB images, semantic segmentation is a common but challenging method. To classify each pixel in an image, the actual state-of-the-art results are achieved by deep neural networks (DNNs). These models need plenty of labeled images to generalize well on unseen scenes. But a manual label process is time and cost consuming and usually error-prone. Additionally, in the recent years computer simulations of urban scenes like CARLA [1] were developed and improved. Controlling simulations enables us to generate data with labels at no (or little) extra cost. The possibility of generating arbitrary many labeled data samples enables us theoretically to train an expert network on the simulated domain via supervised learning. This leads to a nearly perfect model on that domain. Our idea is to use this network to infer on more difficult domains or domains where we have less annotated data. But it has been shown that DNNs trained on one domain can perform arbitrary poor when switching to another domain, i.e. changing the data generating distribution [2]. This phenomenon is often called domain gap or domain shift. Techniques to overcome this gap are called domain transfer techniques or domain adaptation techniques and have become an active research area.

We present an approach to mitigate the domain gap via style transfer and guidance towards the down stream task on the domain where labeled data is rare. Unsupervised style transfer has shown good performance for transferring images from one domain to another [3], but this translation is task agnostic and therefore potentially misses important features when transferring the style from one domain to another. Based on the work of [3] we extend the GAN based domain adaptation approach with a semi-supervised learning component and guide the generator with the help of a small amount of labeled data to the down stream task. To guide the generator, the cross entropy between the prediction of style transferred data and the ground truth is used as a regularization term. Furthermore, our method is independent of the generator network's architecture as only the loss is adapted.

The method can be split into three steps:

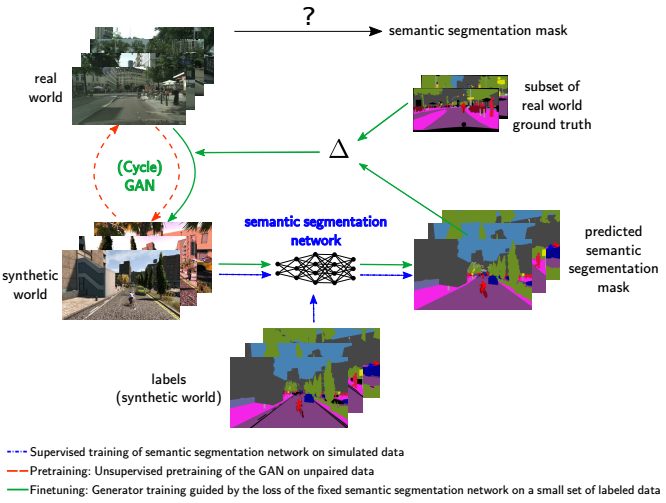


Fig. 1: Methodology: First we train independently a semantic segmentation network on the simulated data (blue) and a CycleGAN [3] based on unpaired data to transfer real data into the synthetic domain (vermilion). Finally, we freeze the semantic segmentation network and fine tune the generator of the CycleGAN with the help of a few labeled data points by guiding it based on the loss of the segmentation network (green).

- *Semantic Segmentation:* Training a semantic segmentation network on the synthetic domain (expert network)
- *Pre-training:* Transfer the real images into the synthetic domain via Image-to-Image translation
- *Fine-tuning:* Guide the generator to the downstream task with the help of a couple of ground truth masks and the semantic segmentation loss.

The concept of our method is shown in Figure 1.

With this approach we mitigate the domain gap between the fully controllable domain (simulation) and the domain where the downstream task original should be solved (e.g. real world) Our main contributions are:

- Feasibility analysis of exploiting a fully controllable domain
- A semi-supervised method to guide the generator to a downstream task without retraining the downstream task network

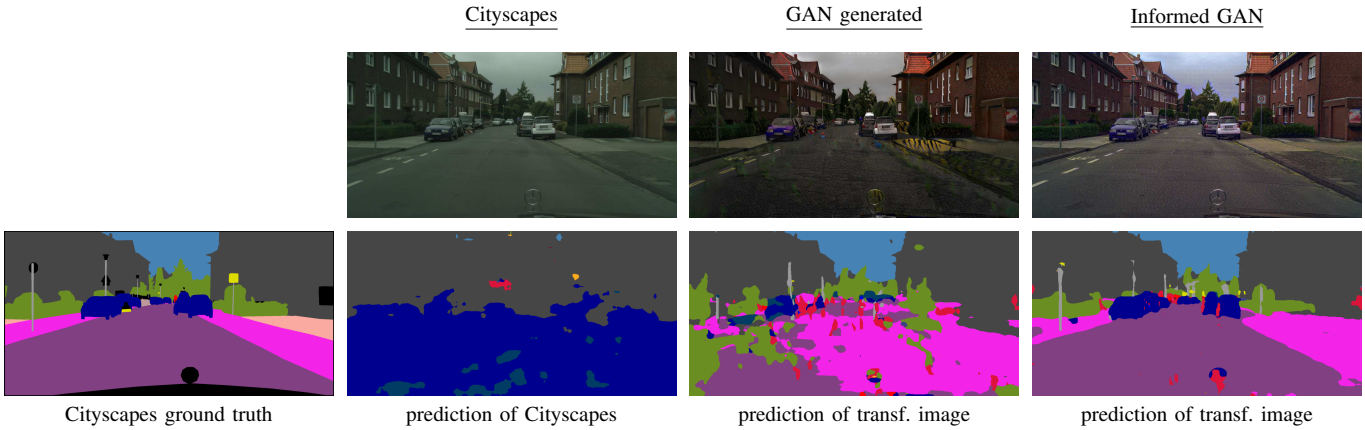


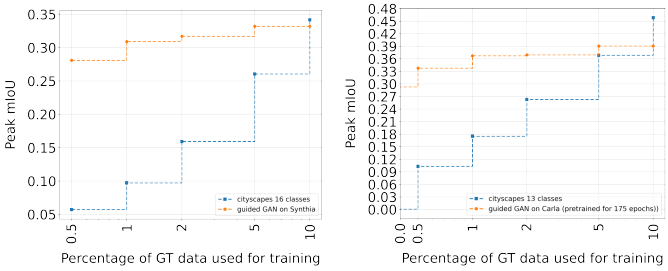
Fig. 2: Comparison of prediction results of original Cityscapes image (left), simple style transfer (mid) and our approach (right)

- A more unbiased domain gap analysis by using a from scratch trained semantic segmentation network.

For our experiments we focus on the domain shift "real to sim(ulation)" as we want to predict semantic segmentation masks on real world examples, where we have few labels and make use of the full controllable simulation to train a performant semantic segmentation network. We evaluate our method on the synthetic data set Synthia [4] and a self generated data set in the CARLA simulator [1]. For the "real" domain we use the Cityscapes data set [5]. To evaluate the domain gap accurately, we train the networks completely from scratch to prevent a bias towards the real world. As a consequence we accept a reduction of the total accuracy. Our experiments show that even though our overall performance lies between $28-39\%$ mean Intersection over Union (mIoU) we can mitigate the domain gap by $18-30$ percent points mIoU depending on the amount of labels. In Figure 2 we show an example image from the validation set of Cityscapes, its transformation with plain CycleGAN and with an informed GAN which was trained with our approach. The corresponding prediction of the network trained on Synthia is shown in the second row. Even though we still find some artifacts, we can see an improvement in the overall scene understanding.

Furthermore, by comparing a Cityscapes training on a small fraction of data with the results of our expert network trained on the synthetic domain, which is fed by images generated by our informed GAN, we have shown that our approach improves segmentation results with only a small fraction of labeled data available. If the amount of labeled data increases, a supervised training approach should be preferred as the information of the labels can be learned directly. The comparison of the strongest performance achieved for a given amount of ground truth (GT) is shown in Figure 3.

Based on our knowledge this is the first time the generator is guided with the help of a semantic segmentation network to focus on the desired downstream task.



(a) Network trained on Synthia (b) Network trained on Carla

Fig. 3: Comparison of the performance given different amount of ground truth data. The blue graph represents the supervised Cityscapes training and the orange graph our approach.

ACKNOWLEDGMENT

This work is funded by the German Federal Ministry for Economic Affairs and Climate Action within the project "KI Delta Learning - Scalable AI for Automated Driving", grant no. 19A19013Q.
The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this project by providing computing time through the John von Neumann Institute for Computing (NIC) on the GCS Supercomputer JUWELS [6] at Jülich Supercomputing Centre (JSC).

REFERENCES

[1] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An Open Urban Driving Simulator," in *Conference on Robot Learning*. PMLR, Oct. 2017, pp. 1–16, iSSN: 2640-3498. [Online]. Available: http://proceedings.mlr.press/v78/dosovitskiy17a.html

[2] G. Csurka, "Domain Adaptation for Visual Applications: A Comprehensive Survey," *arXiv:1702.05374 [cs]*, Mar. 2017, arXiv: 1702.05374. [Online]. Available: http://arxiv.org/abs/1702.05374

[3] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[4] G. Ros, L. Sellart, J. Materzynska, D. Vázquez, and A. M. López, "The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 3234–3243.

[5] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3213–3223. [Online]. Available: https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Cordts_The_Cityscapes_Dataset_CVPR_2016_paper.html

[6] Jülich Supercomputing Centre, "JUWELS: Modular Tier-0/1 Supercomputer at the Jülich Supercomputing Centre," *Journal of large-scale research facilities*, vol. 5, no. A135, 2019. [Online]. Available: http://dx.doi.org/10.17815/jlsrf-5-171