

Supplementary Material for Shallow Features Guide Unsupervised Domain Adaptation for Semantic Segmentation at Class Boundaries

Adriano Cardace Pierluigi Zama Ramirez Samuele Salti Luigi Di Stefano
Department of Computer Science and Engineering (DISI)
University of Bologna, Italy
{adriano.cardace2, pierluigi.zama}@unibo.it

Layer	K	In/Out	Input
Edge Module			
Conv1 + BN + ReLU	3	256/128	input
Conv2 + BN + ReLU	3	128/64	Conv1
Conv3	1	64/1	Conv2
Warping Module			
Conv1 + BN + ReLU	3	256/256	input
Conv2 + BN + ReLU	3	256/256	Conv1
Conv3	1	256/2	Conv2
Final Classifier			
Conv1 + BN + ReLU	3	128/128	input
Conv2 + BN + ReLU	3	128/64	Conv1
Conv3	1	64/C	Conv2

Table 1: Detailed structure of the additional components used in the *adaptation* step.

1. Network Architecture

In this section, we provide a more detailed description of our architecture used for the *adaptation* step. We rely on the widely adopted Deeplab-v2 backbone and feed the output of *conv2* to two additional blocks to predict semantic edges and estimate the 2D displacement grid. Moreover, as explained in Sec. 3 of the main paper, a third component that acts as a classifier is introduced to predict the final segmentation mask given the fine-grained feature map. Table 1 reports the architectural details of the three additional components. For each layer, we report kernel size (K) and number of input/output channels.

2. Training details

We provide here some additional details for each experiment.

2.1. Synthetic-to-real adaptation

As explained in Sec. 3.2 of the main paper, we exclude certain classes in the data augmentation pipeline to avoid

scene occlusion in the newly generated images. Indeed, classes such as *road* and *building* typically cover a large portion of the input images and are more unlikely to create plausible scenes when pasted into source images. For this reason, we exclude *road*, *sidewalk*, *building*, *vegetation*, *sky* for both GTA5→Cityscapes and SYNTHIA→Cityscapes. At inference time, we follow common practice [4, 7, 1, 6], and resize the input image to 512×1024 pixels, while the final prediction is bi-linearly up-sampled to 1024×2048 pixels to compute the mIoU score based on the provided annotation.

2.2. Real-to-real adaptation

In the real-to-real adaptation scenario, we train on Cityscapes [3] and test on the four different cities of the NTHU [2] dataset. Since the source domain contains only 2975 training pairs, we crop images at 512×1024 . Following standard practise [7, 1, 5], we merge *pole*, *fence*, *wall* into the class *buildings*, *truck* into *car* and *terrain* into *vegetation*. Akin to the synthetic-to-real setting, at test time we resize the input image to 512×1024 pixels and up-sample the prediction to the same resolution of the ground-truth, i.e. 1024×2048 pixels.

3. Qualitative Results

In this section, we provide additional qualitative results for the three benchmarks in which we tested our method. Fig. 1 shows predictions obtained for the GTA5→Cityscapes benchmark, while Fig. 2 and Fig. 3 deal with SYNTHIA→Cityscapes and Cityscapes→CrossCity, respectively. We observe that in all cases our method achieves excellent performance along class boundaries. In particular, our method can yield very precise segmentation masks for objects such as *pole*, *persons* and *traffic sign*.

4. Displacement Maps

We highlight in Fig. 4 the effect of our warping module, qualitatively proving the intuition the by means of the

displacement map we can obtain sharp segmentation masks in the target domain as well. In Fig. 4 (a), for example, it is possible to appreciate how the legs of the rider (bottom-right) are less noisy compared to those obtained by a model trained on translated images only (bottom-left). Similarly, Fig. 4 (d) shows how the correct shape of the traffic sign is recovered thanks to the displacement map.

5. Data Augmentation

In this section, we show several training pairs generated with our data augmentation pipeline, described in Sec. 3.3 of the main paper. Fig. 5 and Fig. 6 deal with GTA5→Cityscapes and SYNTHIA→Cityscapes, respectively. It is worth noticing that we perform our data augmentation pipeline on translated source images (left-side). Indeed, these images resemble in color the gray-stylish images typical of Cityscapes[3]. In the top row of Fig. 5 for example, we can appreciate how the class fence, which is correctly classified (see predictions on the right-side), is pasted into a translated source image on both sides of the road. Similarly, in the top row of Fig. 6, an instance of the class *bus* is correctly classified and inserted into the source scene.

References

- [1] Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. 1
- [2] Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. No more discrimination: Cross city adaptation of road scene segmenters. In *ICCV*, 2017. 1
- [3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 2
- [4] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schuster, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. 1
- [5] Haoran Wang, Tong Shen, Wei Zhang, Lingyu Duan, and Tao Mei. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In *The European Conference on Computer Vision (ECCV)*, August 2020. 1
- [6] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-mei Hwu, Thomas S. Huang, and Honghui Shi. Differential treatment for stuff and things: A simple unsupervised domain adaptation method for semantic segmentation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2020. 1
- [7] Yang Zou, Zhiding Yu, Xiaofeng Liu, B. V. K. Vijaya Kumar, and Jinsong Wang. Confidence regularized self-training. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. 1

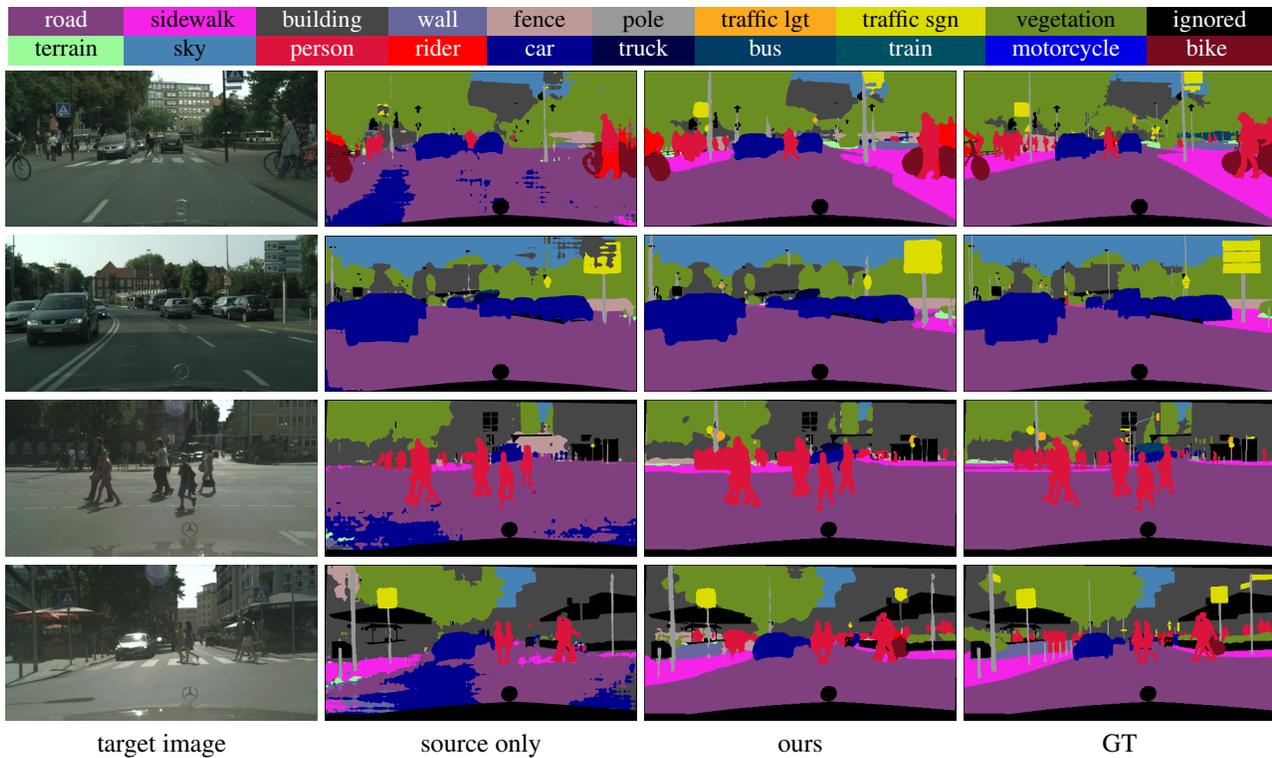


Figure 1: Qualitative results on GTA5→Cityscapes. From left to right: input image, source only, our method, ground-truth.

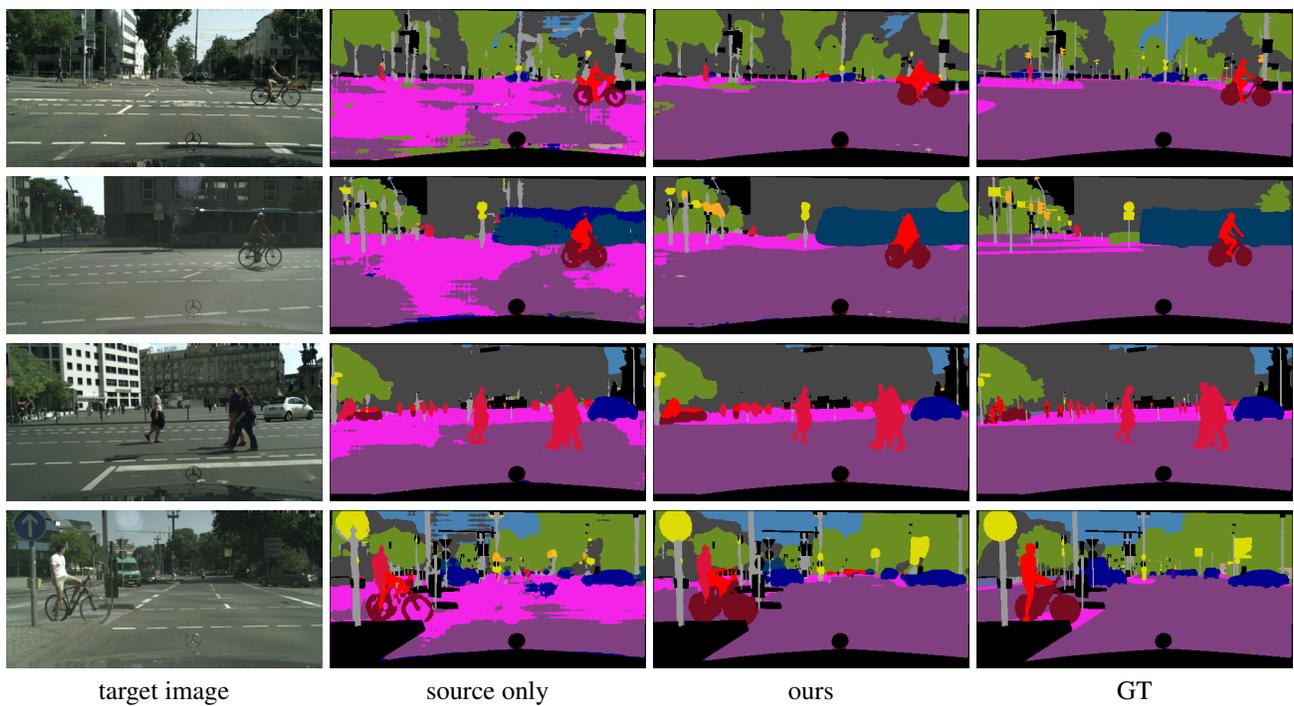


Figure 2: Qualitative results on SYNTHIA→Cityscapes. From left to right: input image, source only, our method.

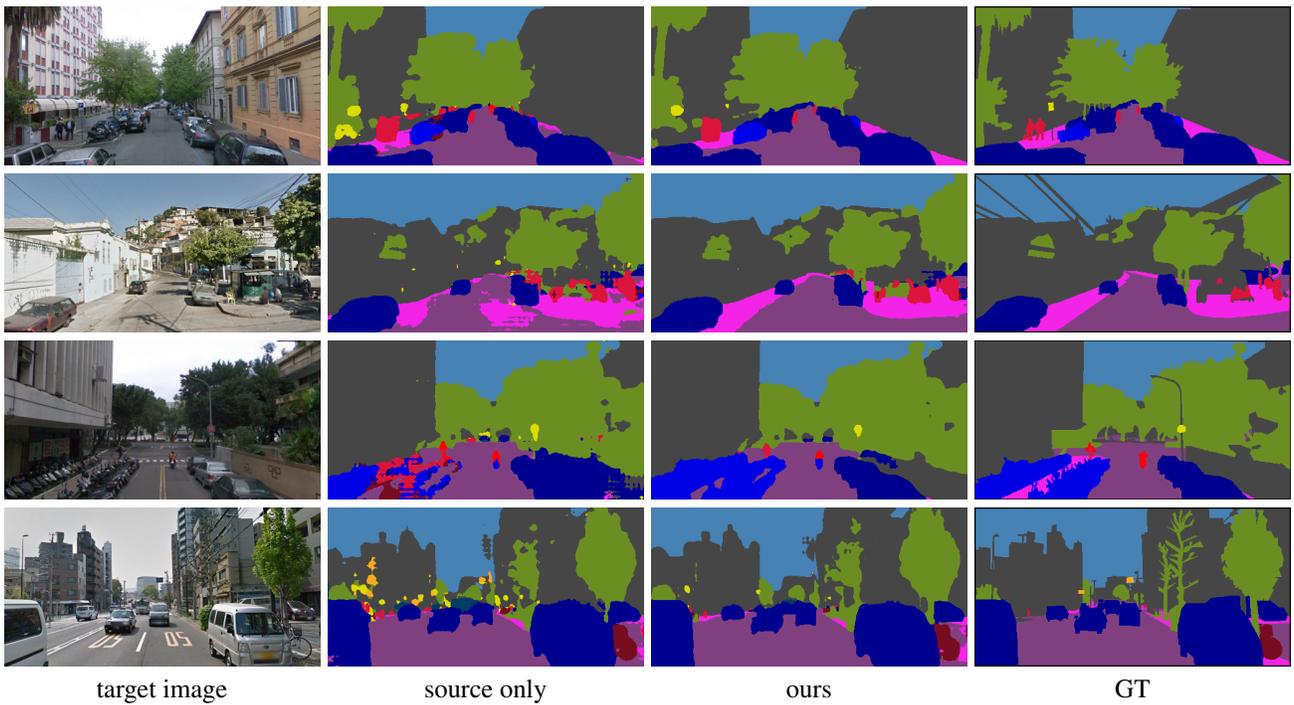


Figure 3: Qualitative results on Cityscapes→CrossCity. From left to right: input image, source only, our method. From top to bottom: Rome, Rio, Taipei, Tokyo.

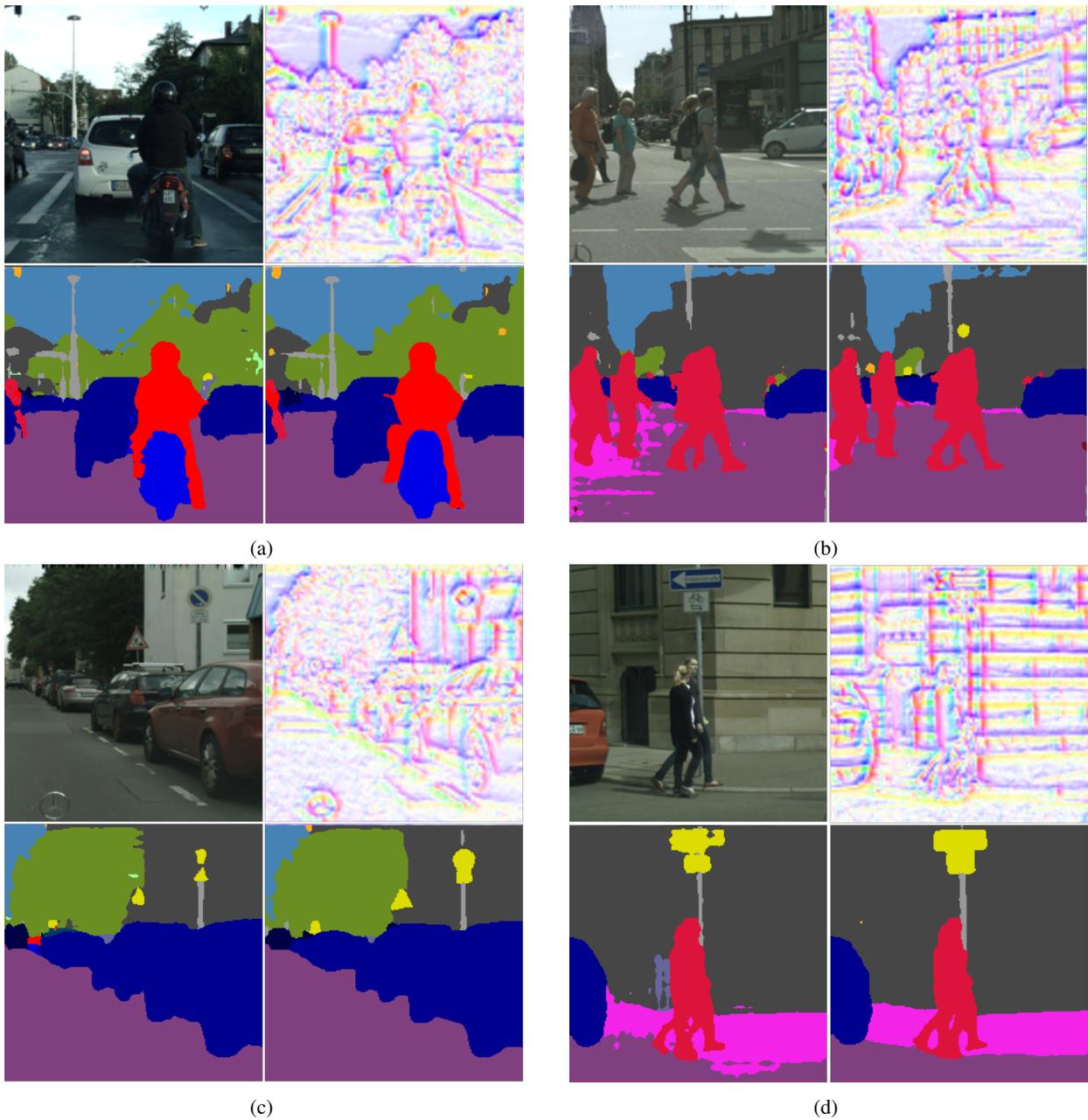


Figure 4: Displacement field visualization in GTA5→Cityscapes. Top left: input image. Top right: estimated displacement field. Bottom left: semantic map from a model trained on translated images. Bottom Right: Our results, providing more accurate segmentation along class boundaries thanks to feature warping based on the estimated displacement field.

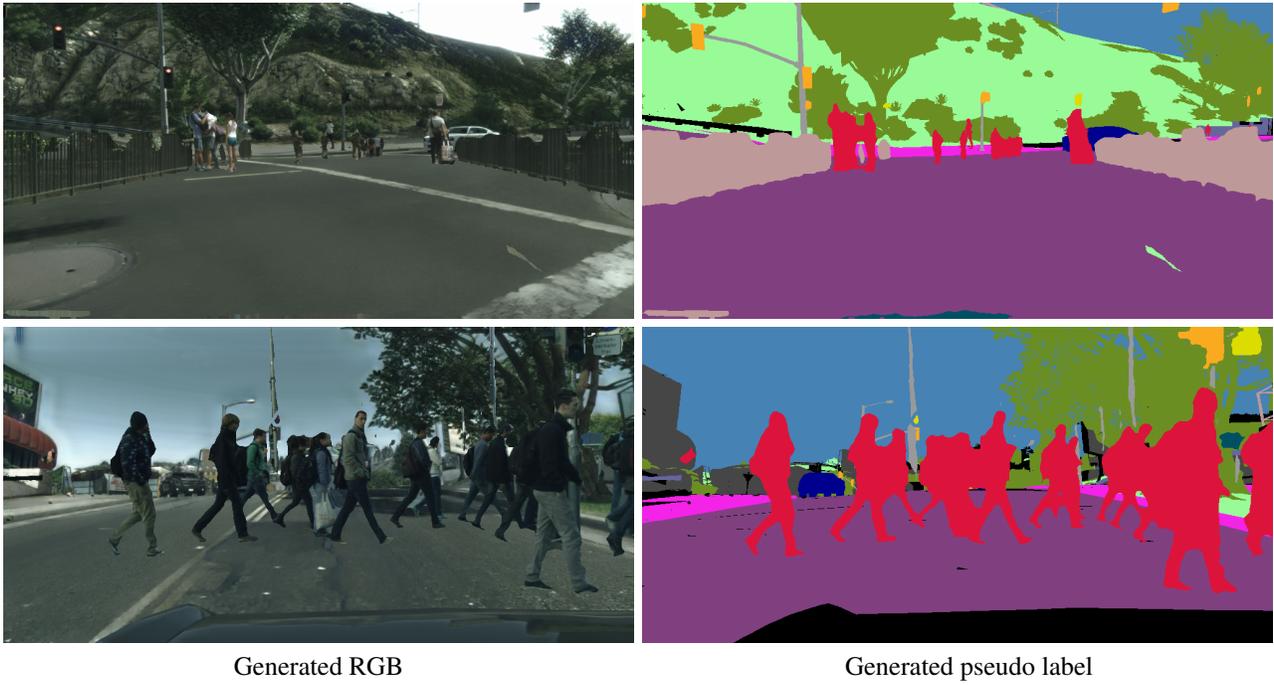


Figure 5: Examples of two newly generated training pairs with our data augmentation for GTA5→Cityscapes. Left images consist of translated source images enhanced with target objects, while right images represent the associated pseudo-labels.

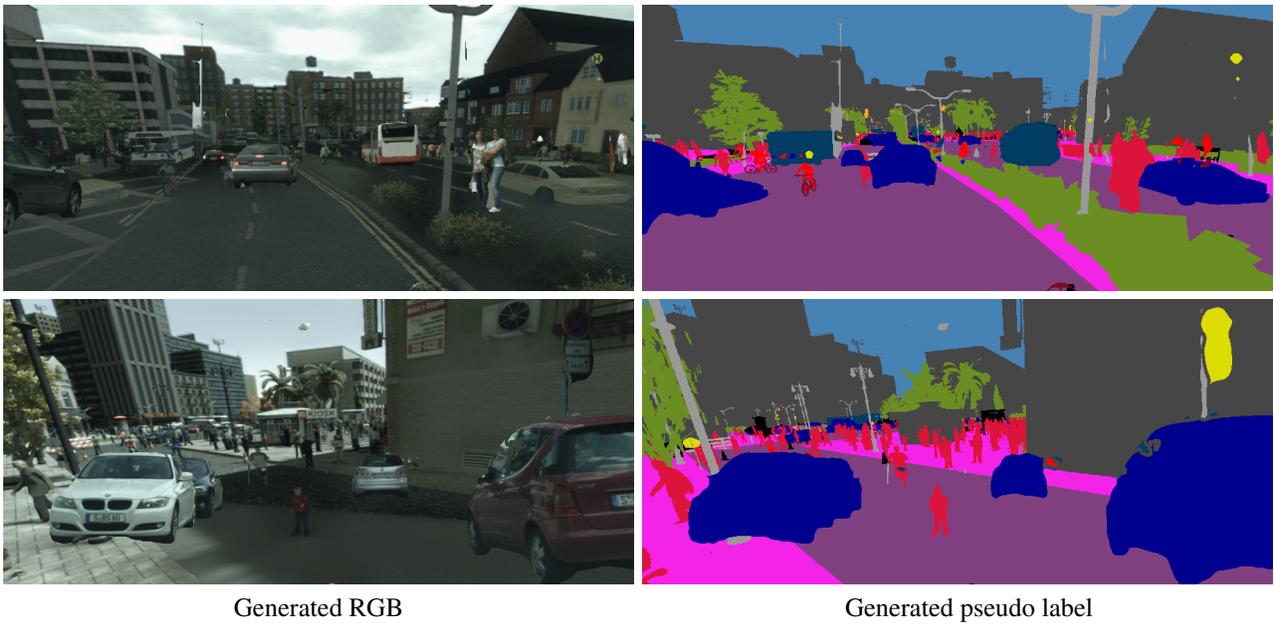


Figure 6: Examples of two newly generated training pairs with our data augmentation for SYNTHIA→Cityscapes. Left images consist of translated source images enhanced with target objects, while right images represent the associated pseudo-labels.