# Pixel-by-Pixel Cross-Domain Alignment for Few-Shot Semantic Segmentation Supplementary Material

Antonio Tavera[1], Fabio Cermelli[1], Carlo Masone[2], and Barbara Caputo[1]
[1]Polytechnic University of Turin, Turin, Italy
[2]CINI - Consorzio Interuniversitario Nazionale per l'Informatica, Rome, Italy
[1]{antonio.tavera, fabio.cermelli, barbara.caputo}@polito.it

## 1. Robustness to hyperparameters

**Sample Selection threshold.** As stated in the main paper, forcing the alignment to the source samples far away and detached from the target distribution results in negative transfer and performance degradation, thus reducing the generalization capability of the overall framework. In this section we demonstrate this assumption by training the 1-shot GTA→Cityscapes scenario with source data sampled with different thresholds ($\delta$). The results in the Fig. 1 confirm that iteratively subsampling the source dataset and selecting source samples that are better aligned with the target semantic distribution leads to a boost in terms of accuracy. The best result is achieved for $\delta = 0.4$, doubled at each epoch. Note that using values that range from 0.1 to 1 changes the result by little more than 1.6% while always achieving state-of-the-art results, demonstrating a good robustness.

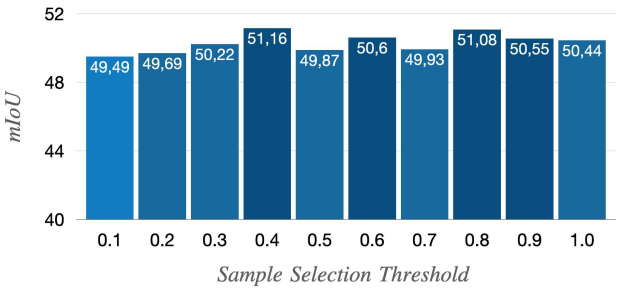


Figure 1. Ablation study about the choice of the sample selection threshold $\delta$, performed on the GTA→Cityscapes 1-shot scenario.

**Lambda.** We also tune the hyper-parameter $\lambda$ which controls the PixAdv loss, comparing four different values: 1, 0.1, 0.01 and 0.001. As shown in Tab. 1, we discovered that in the 1-shot setting, in the case of GTA→Cityscapes the best results are achieved for $\lambda = 0.1$ while in the case of SYNTHIA→Cityscapes, the best results are achieved for $\lambda = 1.0$. Nevertheless, the difference in the performance between $\lambda = 1$ and $\lambda = 0.1$ is 1.9% and 1.4% on GTA and SYNTHIA, respectively. Thus, even by selecting a sub-optimal hyperparameter PixDA still outperforms all the baselines by a good margin.

| $\lambda$ | mIoU | |
|---|---|---|
| | GTA5→City | SYNTHIA→City |
| 1 | 49.23 | **55.92** |
| 0.1 | **51.16** | 54.51 |
| 0.01 | 50.59 | 53.38 |
| 0.001 | 49.05 | 53.74 |

Table 1. Ablation study about the effect of $\lambda$ on the PixAdv loss, performed on the GTA→Cityscapes and SYNTHIA→Cityscapes 1-shot scenario.

## 2. Qualitative Results

Figures 2 and 3 present segmentation results for all the 1 to 5 shot settings and provide a qualitative comparison between our method and the baselines tested in the companion paper. These images intuitively demonstrate the challenge in handling the domain shift in the few-shot setting with imbalanced semantic content and they show that PixDA is able to better align the underrepresented semantic classes.

**GTA5.** At a quick glance, Fig. 2 confirms that the visual domain shift between GTA5 [4] and Cityscapes [2] is lower in some major classes, *e.g.*, "road", "sky" and "car", since the source-only model succeed in correctly classifying them without domain adaptation.

The same cannot be said for classes that are underrepresented in the target realm (either because not present or small in size) such as "bus", "train", "rider", "truck", "traffic sign" and "traffic light". Notably, even the "sidewalk" category, albeit frequent, is completely mishandled by the source-only model due to a higher domain shift. This fact is clearly visible in the 1, 4 and 5 shot settings, where we can see that both the source-only model and the joint training baseline struggle to classify the sidewalk.

Looking at the other baselines, *i.e.*, unsupervised domain adaptation with FDA [7] and transfer learning with NAAE [6], we can see that despite the improvements w.r.t. the source-only model they still struggle with the finer details, such as bicycles, motorcycles, traffic signs. Even fine-tuning with few images suffers from similar problems and in some instances it produces completely wrong predictions (*e.g.*, the bus in the 4-shot setting).

Finally, we observe that PixDA performs consistently well in all settings, providing good predictions across all classes and being overall the closest to the ground truth. Inspecting the levels of the PixDA loss reveals that the loss provides more power (dark blue) to all those areas that are proven to be challenging for the other methods, *e.g.*, the semantic classes with a higher domain shift ("sidewalk") or with fewer pixels in the target domain ("bicycle", "motorcycle" or "signs"). What is most remarkable is that PixDA manages to achieve very good results even in the 1-shot setting, where all other methods struggle with the under-represented classes ("bicycle", "motorcycle", "traffic sign", "traffic light") and some of them even on the predominant ones.

**SYNTHIA.** Similar remarks can be made for the results obtained with the SYNTHIA→Cityscapes scenario in Fig.3. The major difference w.r.t. the GTA experiment is seen for the "road" and "sidewalk" classes. All baselines, including FDA and NAAE, display poorer results in classifying these semantic categories due to a higher domain shift between SYNTHIA [5] and Cityscapes. The increased difficulty in predicting the "road" and "sidewalk" is corroborated by the PixAdv loss levels. We notice that the PixAdv loss puts a higher emphasis on parts of the road and sidewalk w.r.t. to the GTA experiment. As a result, the prediction of these classes with PixDA is the best across all settings. Once again, we remark that PixDA also shows very consistent results, providing good predictions even in the 1-shot setting where all the other methods clearly struggle.

## 3. Assessing the PixAdv loss in UDA

**Introduction.** The research presented in this manuscript focuses on the few-shot DA setting, which is promising for the autonomous driving application. In the main paper we discuss how having access to only few data samples from the target distribution leads to an extreme imbalance in the number of pixels exposed for each class in the target domain. Yet, this problem is still present in the unsupervised DA setting, although to a lesser extent. Even though the UDA setting is outside the scope of this paper, we can demonstrate that our new PixAdv is effective even in the context of Unsupervised Domain Adaptation (UDA). The UDA setting considers having access at training time to a source dataset $S$ with segmentation labels $Y_S$ and a large target dataset $T$ with no labels. The goal is to train a semantic segmentation network on both $S$ and $T$ and get it to perform as near as possible to the model trained on $T$ but with ground truth labels $Y_T$.

**Experiments description.** Because our PixAdv loss is dependent on target ground truth information and the UDA setup does not supply target annotations, we first train the model unsupervised with a regular binary cross-entropy (BCE) functioning as adversarial loss (we denote this model as $M(0)_{BCE}$). We use the model $M(0)$ predictions on the target data to generate high confidence pseudo labels following the "max probability threshold" approach mentioned in [3]. Once we have them, we can move forward with the adversarial training by computing the segmentation loss on the target as well. Two parallel experiments are conducted for comparison, the first $M(1)_{BCE}$ continuing the prior adversarial training with the $BCE$ as the adversarial loss and the second $M(1)_{PixAdv}$ replacing the BCE with our novel PixAdv loss. As the model's accuracy improves, this phase is repeated (we denote this new rounds as $M(2)_{BCE,PixAdv}$).

**Implementation details.** The experiments use the same segmentation model and hyperparameters as the main paper. The DeepLab V2 [1] model ($M$) is trained using batch size $4$ and SGD with initial learning rate $2.5 \cdot 10^{-4}$, adjusted at each iteration with a "poly" learning rate decay with a power of $0.9$, momentum $0.9$ and weight decay to $0.0005$. The discriminator is trained using Adam optimizer, with learning rate $10^{-5}$ and the same decay schedule of the segmentation model. The momentum for Adam is set to $\{0.9, 0.99\}$. The model $M$ is trained for $30k$ iterations for every of the three phases (0,1,2). At each $30k^{th}$ iteration, the model is used to generates pseudo labels. The pseudo labels generation threshold is set to 0.9. We want to emphasize that neither sample selection nor knowledge distillation and fine-tuning are employed in this set of experiments. They are only intended to demonstrate the effectiveness of our novel loss in a different setting with respect to the one in the companion paper.

**Experiments results.** Results for the GTA5→Cityscapes and SYNTHIA→Cityscapes scenarios are shown in Tab. 2 and in Tab. 3 respectively. Tab. 2 reveals that our PixAdv loss improves the performance of the baseline method $M(0)$ by $+4.39\%$ and $+5.11\%$ after the first and second rounds of pseudo labelling, respectively. The use of the BCE, on the other hand, yields fewer results, with a $+3.41\%$ and $+4.13\%$. This tendency is also corroborated in Tab. 3, where our loss outperforms the baseline by $+4.58\%$ and $+5.22\%$, respectively, compared to the $+3.7\%$ and $+3.15\%$ produced by the conventional BCE loss. It should be noted that the BCE is much less successful in the second round when compared to the results acquired in the first round.

| Experiment | Road | Sidewalk | Building | Wall | Fence | Pole | TLight | TSign | Vegetation | Terrain | Sky | Person | Rider | Car | Truck | Bus | Train | Motorcycle | Bicycle | mIoU$^{19}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $M(0)_{BCE}$ | 87.41 | 40.47 | 74.94 | 20.5 | 25.81 | 30.12 | 29.82 | 20.61 | 56.7 | 27.13 | 35.41 | 44.33 | 8.05 | 59.03 | 16.26 | 5.15 | 10.28 | 20.97 | 1.64 | 32.35 |
| $M(1)_{BCE}$ | 89.99 | 46.05 | 77.48 | **21.31** | 28.14 | 30.7 | 33.63 | 23.67 | 69.71 | 29.25 | 46.19 | 48.35 | 7.63 | 60.91 | 17.9 | **10.49** | 15.4 | 21.43 | 1.28 | 35.76 |
| $M(1)_{PixAdv}$ | **90.91** | **46.35** | **78.2** | 19.56 | **29.02** | **32.78** | **34.37** | **24.77** | **71.07** | **29.96** | **48.75** | **51.09** | **10.58** | **62.65** | **18.46** | 9.34 | **16.25** | **24.48** | **3.34** | **36.94** |
| $M(2)_{BCE}$ | **89.47** | 45.62 | 77.63 | 18.98 | 26.27 | 31.24 | **34.24** | 25.89 | 72.36 | 27.51 | 49.28 | 49.19 | 7.53 | 62.46 | **20.48** | **15.1** | **16.74** | 21.83 | 1.37 | 36.48 |
| $M(2)_{PixAdv}$ | 89.3 | **45.98** | **78.65** | **19.82** | **27.45** | **32.35** | 32.9 | **27.8** | **74.3** | **28.08** | **54.96** | **52.58** | **10.43** | **63.3** | 20.12 | 14.76 | 14.87 | **22.92** | **5.01** | **37.66** |

Table 2. GTA5→Cityscapes Unsupervised Domain Adaptation experiments. In **blue** the best results in term of mIoU after the first round of pseudo labelling. In **green** the best results after the second.

| Experiment | Road | Sidewalk | Building | Wall | Fence | Pole | TLight | TSign | Vegetation | Sky | Person | Rider | Car | Bus | Motorcycle | Bicycle | mIoU$^{16}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $M(0)_{BCE}$ | 59.13 | 25.74 | 75.98 | 4.98 | 0.02 | 18.39 | 2.56 | 9.06 | 59.63 | 60.8 | 20.89 | 6.2 | 54.48 | 24.23 | 2.05 | 11.99 | 25.02 |
| $M(1)_{BCE}$ | 58.35 | 26.02 | **76.44** | 5.66 | 0.03 | 18.77 | 3.58 | **19.71** | 62.52 | 67.77 | 19.36 | **5.23** | 52.36 | **27.82** | 1.66 | 14.26 | 28.72 |
| $M(1)_{PixAdv}$ | **59.24** | **27.35** | 76.28 | **6.9** | **0.06** | **22.17** | **5.56** | 13.24 | **62.61** | **68.68** | **19.87** | 5.14 | **55.19** | 26.44 | **1.82** | **22.97** | **29.6** |
| $M(2)_{BCE}$ | 57.6 | 25.32 | **75.9** | 5.15 | 0.02 | 18.57 | 3.37 | 9.38 | **62.66** | 71.58 | 19.25 | **5.07** | 50.4 | **29.64** | 1.91 | 14.93 | 28.17 |
| $M(2)_{PixAdv}$ | **59.25** | **27.3** | 75.41 | **6.43** | **0.05** | **22.43** | **7.62** | **14.55** | 62.4 | **72.16** | **19.44** | 4.96 | **55.58** | 27.73 | **2.46** | **26.11** | **30.24** |

Table 3. SYNTHIA→Cityscapes Unsupervised Domain Adaptation experiments. In **blue** the best results in term of mIoU after the first round of pseudo labelling. In **green** the best results after the second.

**Conclusion.** Despite the difference in setting with the main paper, all the experiments demonstrate that when our PixAdv loss is applied to the Unsupervised Domain Adaptation scenario, it can improves performance in almost all the semantic classes (both well and underrepresented) and in both phases of pseudo labelling.

## References


[1] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2018.

[2] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[3] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6936–6945, 2019.

[4] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Eur. Conf. Comput. Vis.*, volume 9906 of *LNCS*, pages 102–118. Springer International Publishing, 2016.

[5] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3234–3243, June 2016.

[6] Ruoqi Sun, Xinge Zhu, Chongruo Wu, Chen Huang, Jianping Shi, and Lizhuang Ma. Not all areas are equal: Transfer learning for semantic segmentation via hierarchical region selection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4360–4369, 2019.

[7] Yanchao Yang and Stefano Soatto. FDA: Fourier domain adaptation for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4085–4095, 2020.

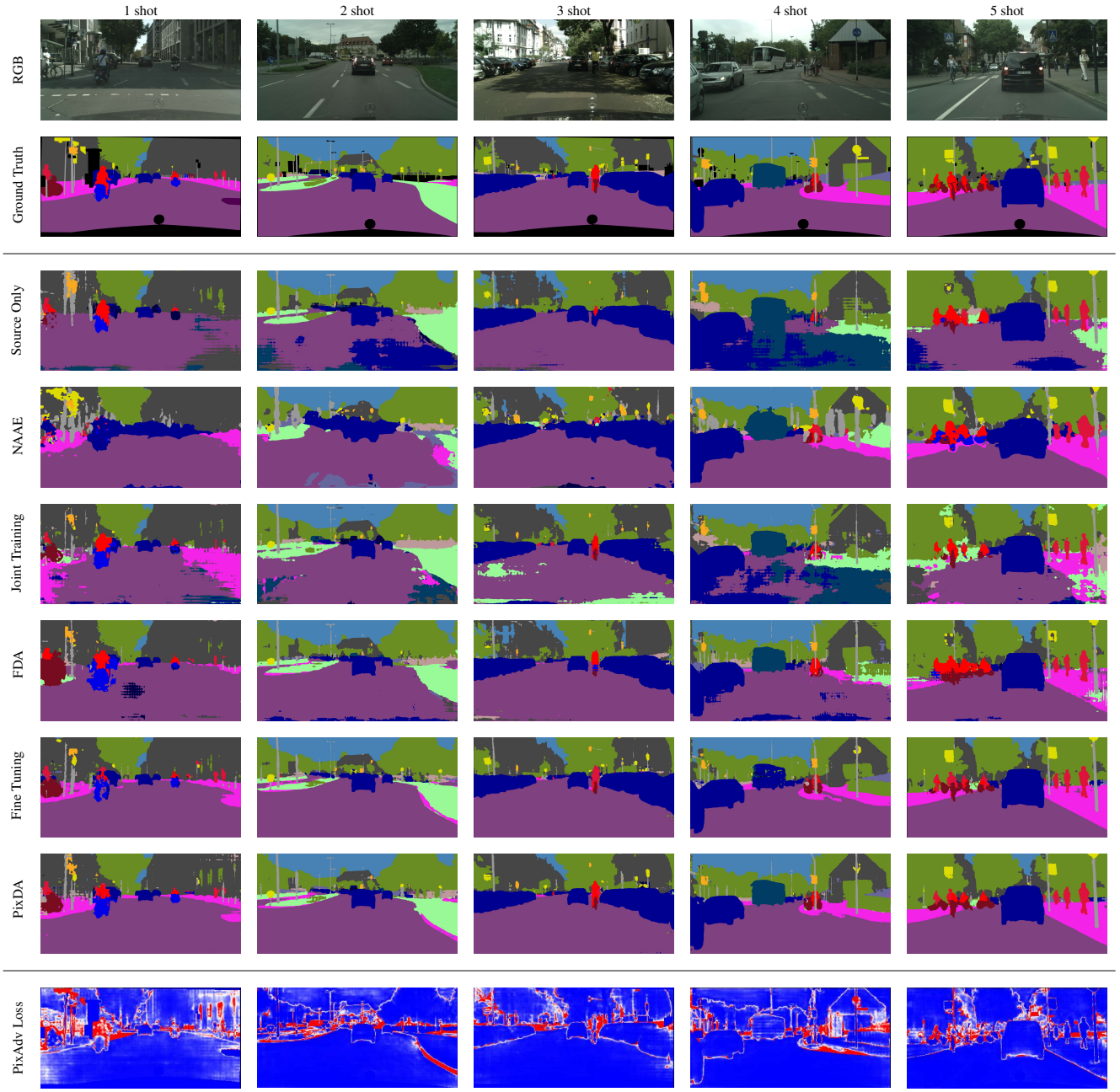


Figure 2. Qualitative results for the GTA→Cityscapes 1to5-shot scenarios. The columns span different few-shot settings, from 1-shot (left) to 5-shot (right). The first row shows the RGB image and the second row illustrates the ground truth. All other row show the predictions achieved by the different methods, except for the last row that visualizes the normalized levels of the PixAdv loss $L_{\text{PixAdv}}$ used by PixDA: blue indicates a low value whereas a darker red indicates a higher value. It's worth remembering that the loss image level is taken at a particular moment in time and varies with each iteration of the training.

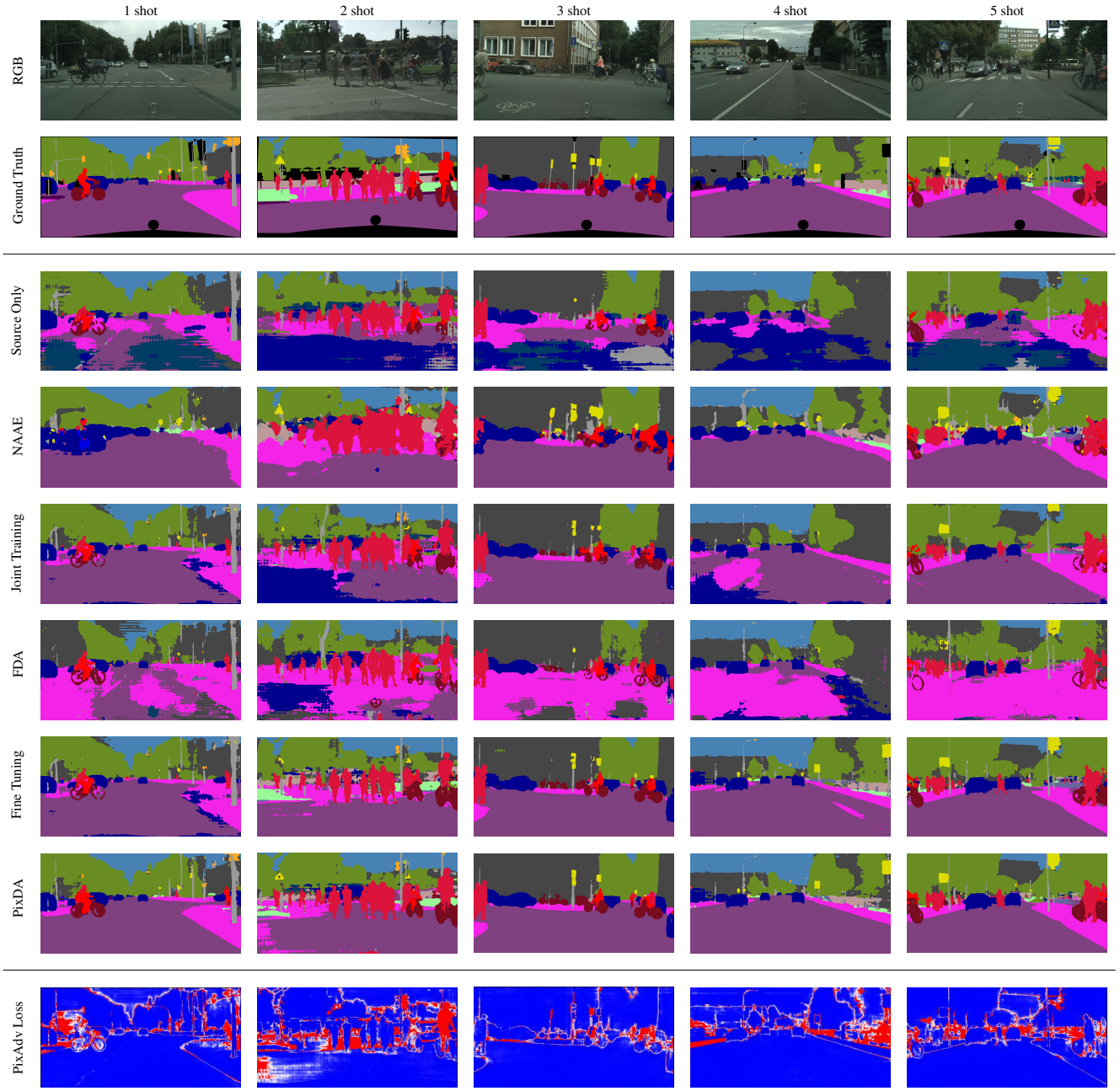


Figure 3. Qualitative results for the SYNTHIA→Cityscapes 1to5-shot scenarios. The columns span different few-shot settings, from 1-shot (left) to 5-shot (right). The first row shows the RGB image and the second row illustrates the ground truth. All other row show the predictions achieved by the different methods, except for the last row that visualizes the normalized levels of the PixAdv loss $L_{\text{PixAdv}}$ used by PixDA: blue indicates a low value whereas a darker red indicates a higher value. It's worth remembering that the loss image level is taken at a particular moment in time and varies with each iteration of the training.