Supplementary Material
Learning Across Domains and Devices: Style-Driven Source-Free Domain Adaptation in Clustered Federated Learning

Donald Shenaj[*,1], Eros Fanì[*,2], Marco Toldo[1], Debora Caldarola[2], Antonio Tavera[2], Umberto Michieli[†,1], Marco Ciccone[†,2], Pietro Zanuttigh[†,1], and Barbara Caputo[†,2]

[1]University of Padova, Italy [2]Politecnico di Torino, Italy

This document contains supporting material for the paper *Learning Across Domains and Devices: Style-driven Source-Free Domain Adaptation in Clustered Federated Learning*. Here, we include additional details on the federated splits employed in the paper along with analyses of the convergence stability of our approach when compared to competing strategies adapted to our federated setup. Finally, we show some qualitative segmentation maps.

1. Additional Details on Splits

In this section, we complete the description of how the federated splits used in our experiments are generated.

Cityscapes. We used the *heterogeneous* federated split of Cityscapes [3] proposed in [4]. The split comprises 144 clients, where each client has between 10 and 45 samples belonging to a single city from the dataset. Further details on the distribution of the number of images per client are shown in Figure 1.

CrossCity. We generated the federated split of the CrossCity [2] dataset by assigning 27 ± 10 images taken from the same city to each client, where the number of samples per client is uniformly sampled. The final distributions of the number of images per client are shown in Figure 2 both per city and overall. We observe how the distributions are balanced across the four cities.

Mapillary. We propose a novel split for the Mapillary Vistas [6] dataset via a clustering procedure based on the GPS coordinates of the images. We started from the original training set of 18000 images and discarded 31 of them missing the GPS coordinates. Then, we run the k-Means algorithm over the GPS coordinates six times, one per continent. The k-Means algorithm is constrained to assign every client a random number of images in the range 16 and 100. The procedure resulted in 357 clients, where each client observed samples from only one continent. The final distributions of the number of images per client are shown in Figure 3. Unlike the other scenarios, we observe a large variability across the distributions obtained in different continents due to the highly imbalanced nature of the dataset. Also, note that the two entries with higher values, 16 and 100, correspond to the extreme values of the constrained k-Means process.

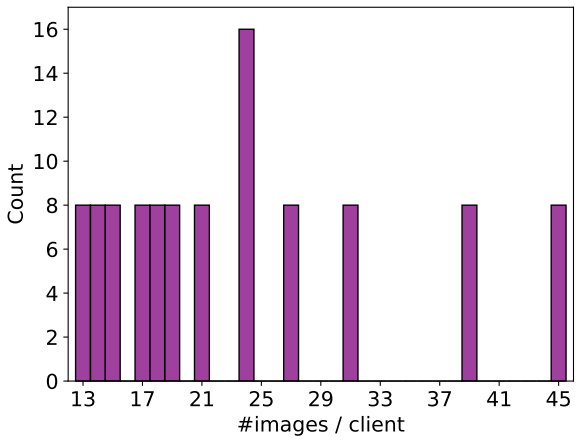

Figure 1: Histogram of images per client in the federated Cityscapes split.

2. Additional Details on the Style-Based Client Clustering

In a realistic FL setting, different clients may observe similar samples, *e.g.* self-driving cars in the same region are likely to collect similar images, thus they are not subject to statistical heterogeneity during the server aggregation. Therefore, we proposed a style-driven client clustering as one of the foundational parts of our algorithm. During the FL optimization stage, we employed the identified commu-

*: Equal contribution. †: Equal supervision.

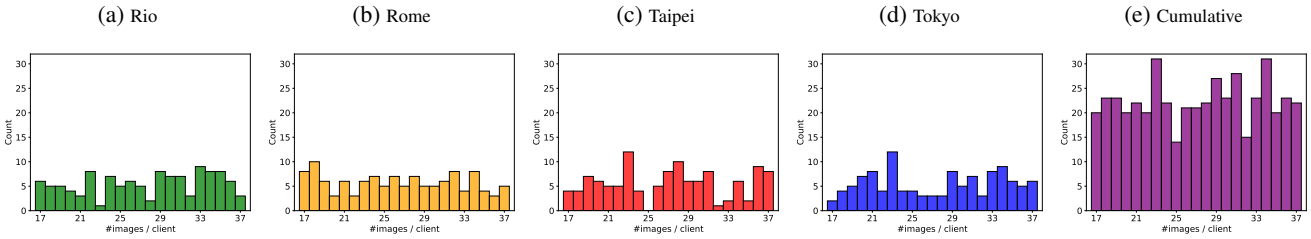

Figure 2: Histogram of images per clients in the proposed federated CrossCity split.

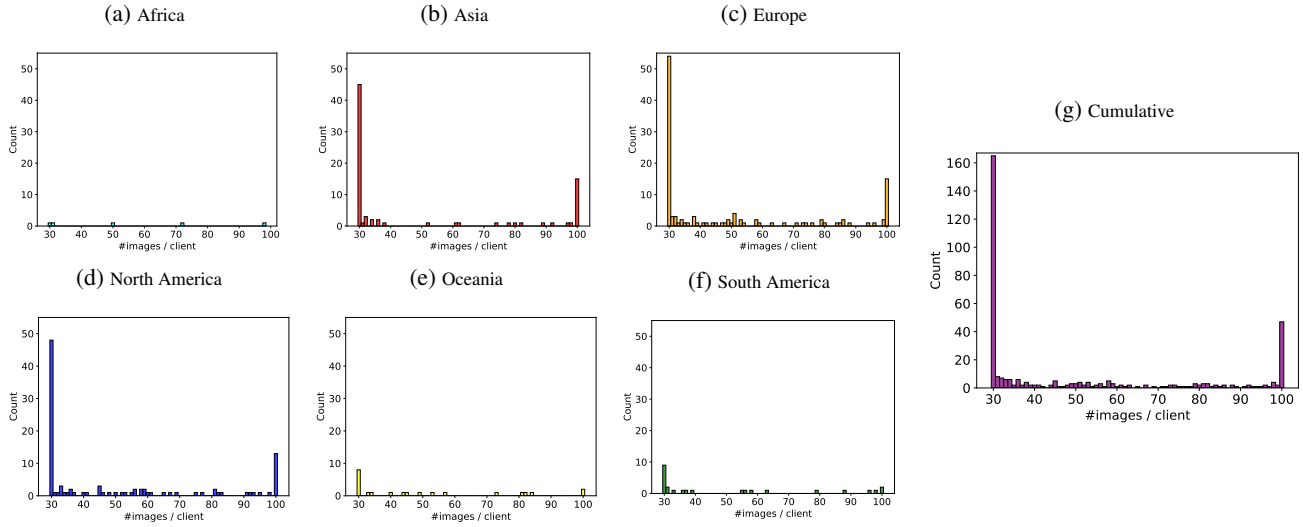

Figure 3: Histogram of images per clients in the proposed federated Mapillary Vistas split.

nities in a clustered and layer-aware aggregation policy on the server side.

First of all, we remark that the four clusters identified by the styles extracted from the images contain mostly clients belonging to one single geographical location (*i.e.*, city). Table 1 shows the number of clients belonging to a specific city assigned to each cluster for the federated CrossCity dataset. Overall, the clustering accuracy, considering each cluster a city, is equal to 68%. Therefore, there is not a one-to-one correspondence of the clusters with the cities.

Table 1: Number of clients belonging to a specific city assigned to each cluster for the federated CrossCity split.

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
Rio	7	1	**70**	38
Rome	**76**	6	22	17
Taipei	6	**103**	0	9
Tokyo	26	8	10	**73**

To investigate this aspect, we show in Figure 4 some samples taken from the clients belonging to each of the four clusters in the federated CrossCity dataset. Here, we observe an interesting finding: despite being generated via style information only, the clusters tend to show scenes with similar semantics. For instance, *Cluster 1* contains clients having images of large and trafficked streets, and grayish sky. *Cluster 2* contains clients having images of narrow streets with little to no vegetation, many buildings, a few parked cars and whitish sky. *Cluster 3* contains clients having images of empty roads with green surrounding vegetation. *Cluster 4* contains clients having images from sunny weather and blue sky, narrow streets with no traffic and green vegetation.

Finally, we show in Figure 5 some samples taken from the clients belonging to each of four clusters in the federated Mapillary dataset. Unlike as for CrossCity, here we do not appreciate a clear assignment as the number of clusters is different from the number of towns or continents. Therefore, we observe that here the clustering is much more appearance-related, according to the style of the images.

For instance, *Cluster 1* contains clients having cloudy and foggy images where the visual appearance is grayish. *Cluster 2* contains clients having grayish sky and yellowish buildings with some similar semantics across clients. *Cluster 3* contains clients having images at the sunset or sunrise where the light scatters yellow shadows. *Cluster 4* contains

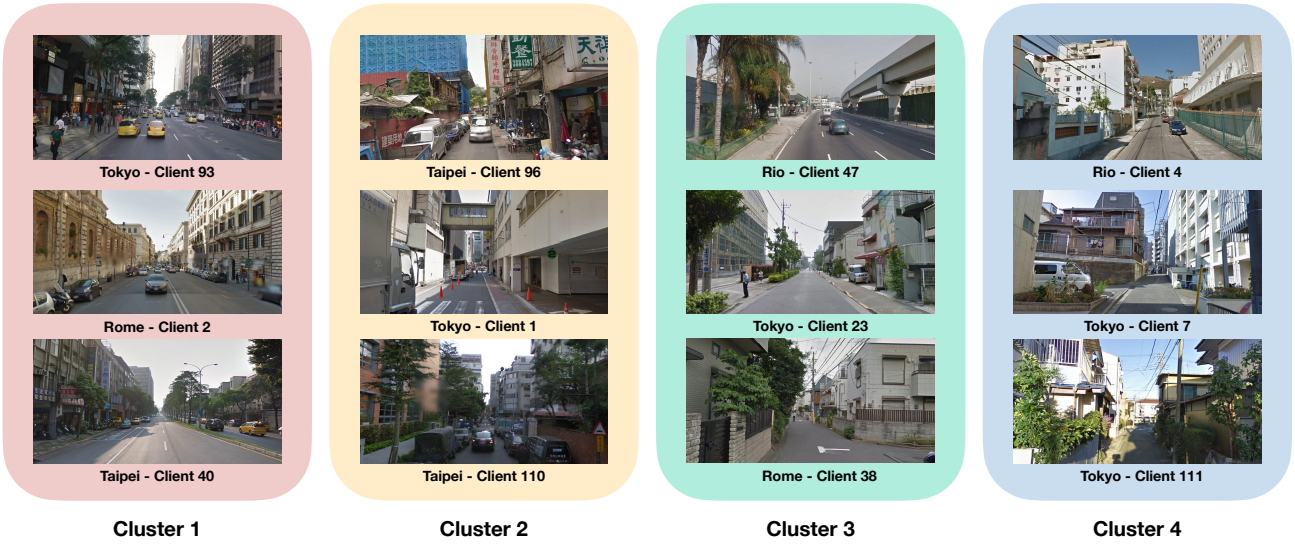

Figure 4: Sample images in each cluster for the federated CrossCity split.

Figure 5: Sample images in some clusters for the federated Mapillary split.

clients having images with predominant blue colors in the sky.

3. Implementation Details

The proposed method is implemented in PyTorch, the code and federated splits are available at `https://github.com/Erosinho13/LADD`.

The semantic segmentation network used is DeepLab-V3 [1] with Mobilenet-V2 [7] as the backbone and width multiplier equal to 1, representing a good compromise in terms of performance and lightness, important aspects to consider for real-world applications, such as self-driving cars. On each communication round, the selected clients are trained sequentially, allowing to perform the complete simulation and reproduce the results on a single GPU with 32GB of VRAM (we used a NVIDIA RTX 3090).

4. Qualitative Results

We provide some qualitative results in the form of segmentation maps of target images generated by the segmentation model subject to different adaptation schemes. Figures 6, 7 and 8 refer to the 3 adaptation setups chosen

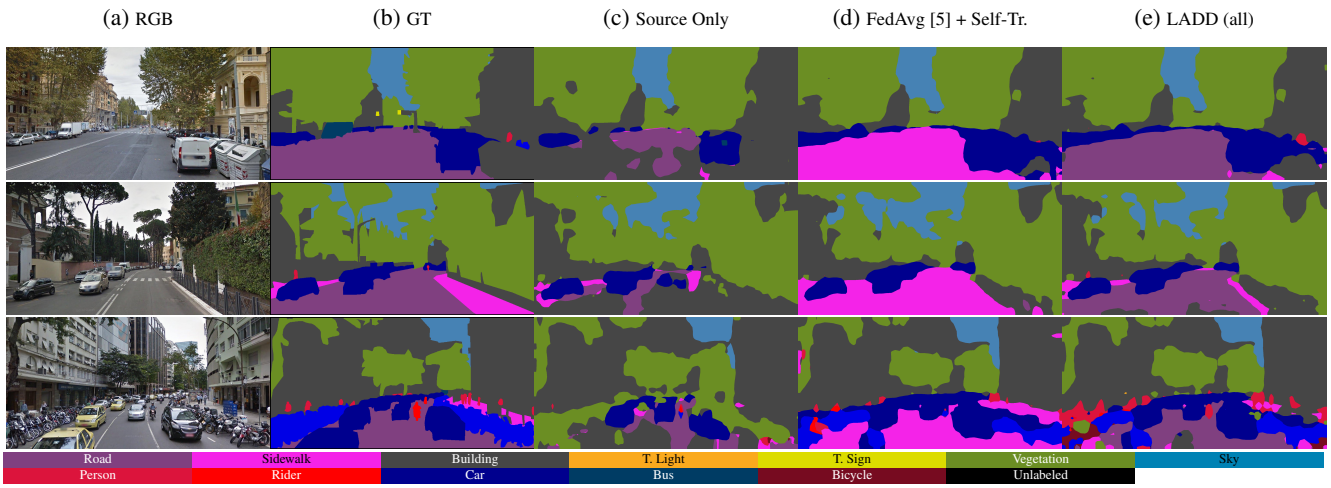

Figure 6: GTA5→CrossCity qualitative results.

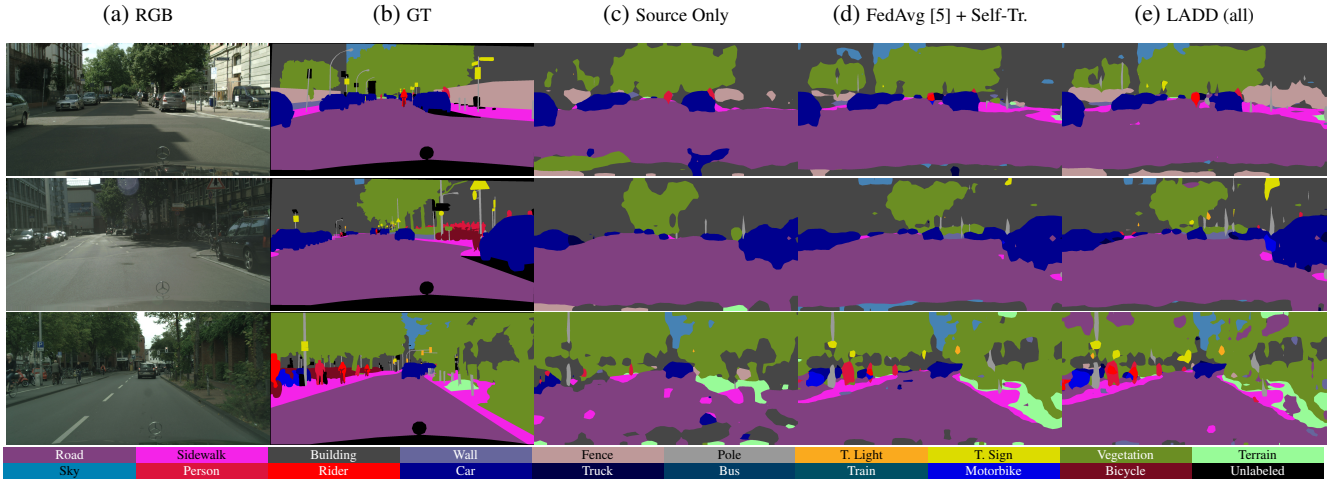

Figure 7: GTA5→Cityscapes qualitative results.

for experimental evaluations, with respectively CrossCity, Cityscapes and Mapillary as target datasets. We compare the naïve source only training (3rd columns in all the aforementioned figures) and the baseline federated adaptation strategy (4th columns), based on FedAvg[5] aggregation and local self-training, with the proposed LADD (when cluster-specific aggregation is extended to all the segmentation network layers) (last columns). For fair comparison we employ the same pretraining for FedAvg and LADD. By inspecting the segmentation maps produced by the different adaptation strategies, we notice how the *source only* maps show inconsistent and noisy predictions, where semantically similar classes are confused, such as *sidewalk* and *road* or *terrain* in all the reported samples. Local self-training and standard FedAvg aggregation at server-side partially mitigate the prediction accuracy drop caused by domain shift between source and target data. Nonetheless, we observe that the adapted model still tends to mistake semantically-similar classes such as sidewalk and road in the first sample of Figure 6. The proposed regularized local training leads to more robust local optimization, which otherwise tends to suffer from unsteady behavior, due to the small amount of available training data and the lack of any form of supervision (even from the source domain) at the client side. This, along with the cluster-specific semantically aware aggregation mechanism, results into less noisy and more accurate predictions as we can see in the last columns of the figures.

5. Additional Quantitative Results

Finally, we report additional results in the form of per-class IoUs achieved when different modules of our framework are enabled. Once more, results are reported with

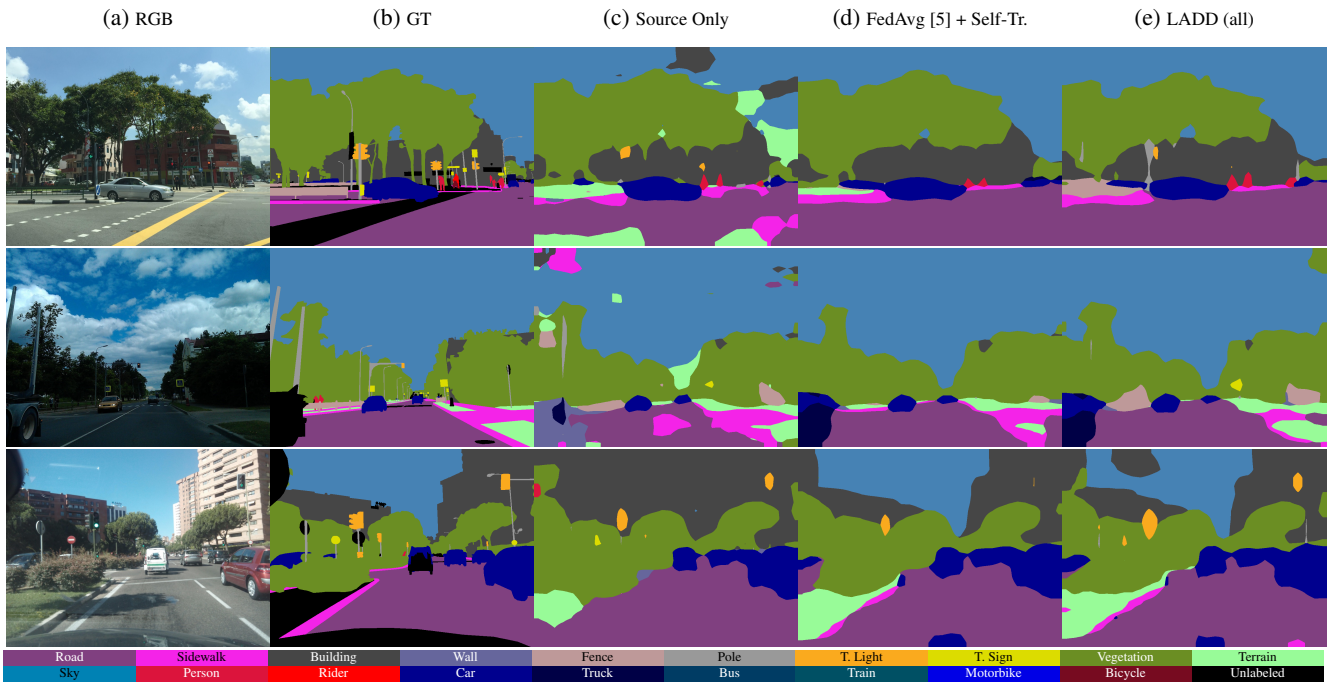

Figure 8: GTA5→Mapillary qualitative results.

CrossCity (Table 2), Cityscapes (Table 3) and Mapillary (Table 4) as target datasets, in terms of mean and standard deviation computed over the last 10% rounds.

When enabled, we observe that each module improves the overall mIoU score, which is also generally shared by the individual IoU scores of the semantic classes in the different experimental setups.

In addition, in Figure 9 we report the learning curves as a result of federated optimization under different configurations of the proposed LADD method in the GTA→CrossCity setup. When only ST is employed in the client-side optimization, the training is extremely unstable, showing a small initial burst of performance followed by a rapid decrease after few rounds. When adding KD and then SWAt, the training curves become progressively more robust and stable, achieving the best results when KD and SWAt are joined by the cluster-specific aggregation, in either classifier-exclusive or full model configuration of cluster-specific parameters. We finally remark how LADD in its complete configuration is characterized by steady and converging learning curves, unaffected by diverging phenomena.

References

[1] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[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 *Proceedings of International Conference on Computer Vision (ICCV)*, pages 1992–2001, 2017.

[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 *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.

[4] Lidia Fantauzzo, Eros Fanì, Debora Caldarola, Antonio Tavera, Fabio Cermelli, Marco Ciccone, and Barbara Caputo. Feddrive: Generalizing federated learning to semantic segmentation in autonomous driving. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022.

[5] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics (AISTATS)*, pages 1273–1282. PMLR, 2017.

[6] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 4990–4999, 2017.

[7] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520, 2018.

Table 2: CrossCity IoU by class and mIoU (%).

FDA	ST	KD	SWAt	Cl Aggr	road	sidewalk	building	traffic light	traffic sign	vegetation	sky	person	rider	car	bus	motorcycle	bicycle	**mIoU**
					25.6	21.6	65.9	3.9	8.6	67.5	73.5	33.1	2.1	43.0	6.6	0.3	0.2	26.5 ± 1.5
✓					38.2	24.0	74.8	7.0	8.9	70.5	80.9	37.0	4.0	63.6	12.0	3.5	0.0	32.4 ± 0.6
✓	✓				21.9	17.9	81.3	9.5	14.5	77.4	85.2	41.0	2.3	66.1	10.7	8.0	0.9	33.6 ± 1.3
✓	✓	✓			49.5	26.7	81.2	11.7	12.4	77.6	87.0	40.4	1.0	68.9	16.3	11.3	3.4	37.5 ± 0.1
✓	✓	✓	✓		53.3	28.6	81.1	12.1	12.0	77.5	87.1	42.1	1.9	68.9	17.2	16.7	4.9	38.8 ± 0.1
✓	✓	✓		✓	63.4	32.3	81.5	12.1	12.2	77.5	86.9	41.0	1.1	69.0	16.0	12.5	3.8	39.2 ± 0.2
✓	✓	✓	✓	✓	64.3	33.7	81.0	12.5	14.4	77.2	86.8	42.1	1.4	69.1	18.1	15.6	4.8	40.1 ± 0.2

Table 3: Cityscapes IoU by class and mIoU (%).

FDA	ST	KD	SWAt	Cl Aggr	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle	**mIoU**
✓	✓				84.5	36.8	77.3	23.9	11.3	20.5	29.1	22.6	76.9	26.5	68.9	53.4	13.7	79.0	15.2	14.0	1.4	11.0	5.1	35.1 ± 0.7
✓	✓	✓			79.3	34.0	73.6	22.0	16.4	24.6	30.3	31.3	61.7	23.2	70.1	51.2	19.3	73.7	13.6	17.9	7.3	12.1	15.3	35.6 ± 0.1
✓	✓	✓	✓	✓	80.0	36.1	74.1	22.8	18.3	26.3	30.6	33.0	65.2	25.4	69.4	52.3	19.1	74.5	13.4	18.0	7.2	12.6	14.2	36.5 ± 0.1

Table 4: Mapillary Vistas IoU by class and mIoU (%).

FDA	ST	KD	SWAt	Cl Aggr	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle	**mIoU**
✓	✓				67.4	36.9	74.7	24.8	25.4	10.9	21.0	33.3	72.8	40.8	91.2	46.1	23.1	73.7	31.1	22.7	3.1	30.6	11.9	39.0 ± 0.2
✓	✓	✓	✓		75.4	37.7	73.4	25.2	25.2	18.3	26.6	37.1	73.5	38.1	91.4	45.5	13.8	71.3	30.9	22.0	3.0	29.9	19.1	40.0 ± 0.1
✓	✓	✓	✓	✓	75.5	37.0	69.1	24.6	25.6	18.9	26.7	38.2	72.5	36.4	89.4	46.3	17.2	70.7	32.6	20.2	4.1	31.4	21.0	40.2 ± 1.0

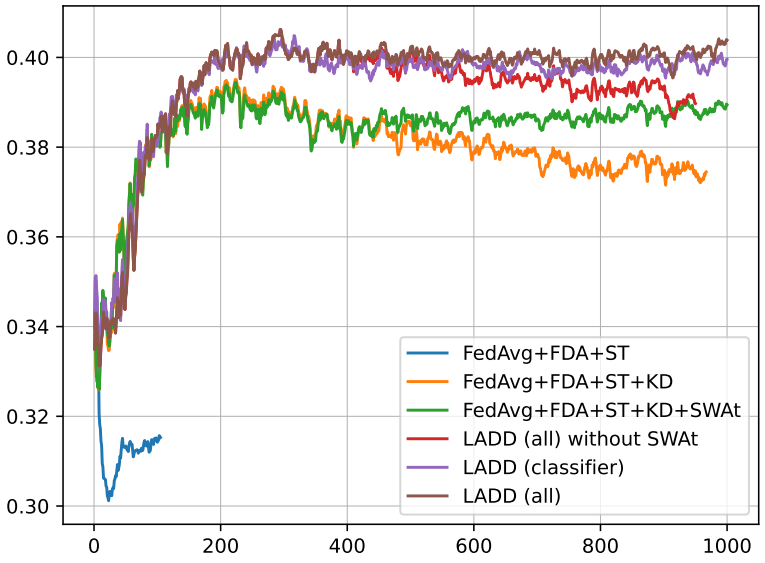

Figure 9: Comparison of learning curves in the CrossCity federated split.