Supplementary Material

A1. FEDA - relugarisation scalers tuning

Our proposed solution for FDG introduces two local regularisations $\mathcal{L}_{L2}$ and $\mathcal{L}_{CMI}$ and one global optimisation loss $\mathcal{L}_{fea}$. These regularisations are attenuated by their corresponding scaling factors λ_{L2}, λ_{CMI}, and λ_{fea}. When increasing the local scalers, we noticed a smoothness in the local representations from the first local rounds, indicating a lower source client performance. In this way, both the local and global accuracies (after the communication round) are affected by this degradation when the models are not learning more complex domain-invariant representations. Although, as in [32], the scalers can be checked with random search, we propose for future work to learn them locally. Even with the same local architecture, each domain may overlap more or less with the others; thus, the local scalers should be adjusted dynamically.

Regarding the global scaler λ_{fea}, we observed that with higher values, the global model pulls the sources towards the target. Therefore, this effect affects local training at the client level while yielding greater source losses after communications. Together with the previous intuition from the local scalers, over several FL rounds, we could adapt the global one with a slow increment. This behaviour will be tested in our future work.

A2. Source accuracy evaluation

A2.1. FEDA - Federated Domain Generalisation experiments

FEDA aims for generalisation on an unseen target domain. Despite this, we also overview the impact on the source local domains. For all three benchmarks evaluated, we confirm a fairly larger performance in average source accuracy when compared to the target domain. This shows that the model can generalise well to all the clients where the data is inferred during the FL training. In PACS, FEDA achieves the closest performance source accuracy to the target with a gap of approximately 8%.

Dataset	FEDA - Domain/ Source Accuracy									
PACS [20]	A	94.96	C	92.67	P	93.69	S	96.66	Avg	94.49
OfficeHome [36]	A	91.2	C	80.21	P	84.4	R	86.26	Avg	85.52
OfficeCaltech [48]	A	75.87	C	87.1	D	80.68	W	81.55	Avg	81.30

Table A1. FEDA - Federated Domain Generalisation **source** performance [average over 3 trials]

A2.2. FEDALV - Federated Active Learning with Domain Generalisation experiments

Before delving into the quantitative FEDALV evaluation of sources, it is necessary to mention the target accuracies of FEDA on OfficeCaltech when trained with domain per client and ResNet-18 architecture. Table 1(c) displays the metrics according to the FDG settings of FPL [15], while the evaluation of FEDALV follows the configurations on PACS. Therefore, for each domain taken as a target, FEDA obtains the following accuracies: Amazon 90.45, Caltech 61.51, DSLR 77.08, Webcam 78.335, and an average of 76.84. According to Table 3(c), FEDALV attains a close target average performance of 72.34 (-4.5%) with 50% of the entire source dataset.

FAL Method / % of labelled data	1%	2%	3%	4%	5%
Random	59.07	72.09	74.8	78.88	80.62
CoreSet [38]	61.75	73.24	76.66	79.4	80.82
LoGo [18]	61.95	73.39	77.45	79.57	81.13
EADA [43]	59.81	71.89	77.74	82.24	<u>84.63</u>
FEDAL (ours)	62.56	69.21	74.61	78.16	79.43
FEDALV (ours)	62.5	70.04	76.56	79.01	79.92

(a) **PACS**

FAL Method / % of labelled data	10%	20%	30%	40%	50%
Random	62.56	69.21	74.61	78.16	79.43
CoreSet [38]	37.83	45.7	56.46	65.61	73.31
LoGo [18]	35.29	49.43	53.89	70.67	77.29
EADA [43]	37.75	53.21	58.8	70.74	78.69
FEDAL (ours)	36.08	52.94	59.61	71.16	77.76
FEDALV (ours)	36.86	61.66	73.59	74.69	**80.11**

(b) **OfficeHome**

FAL Method / % of labelled data	1%	2%	3%	4%	5%
Random	27.98	37.63	42.92	47.65	50.99
CoreSet [38]	27.78	38.52	44.49	50.95	<u>53.58</u>
LoGo [18]	28.2	36.75	40.96	46.8	50.63
EADA [43]	28.39	38.31	43.42	49.23	52.92
FEDAL (ours)	28.48	33.53	39.07	44.81	47.35
FEDALV (ours)	28.12	36.38	43.49	48.81	53.48

(c) **OfficeCaltech**

Table A2. FEDALV - FAL **source** performance [average over 3 trials]

Shifting the performance measurements of FEDALV on the source datasets, we gathered the results in Tables A2 (a, b, c), for the three tackled datasets. We can observe that FEDALV yields competitive results against the other active learning (AL) baselines, especially for the OfficeCaltech dataset. The lack of consistent AL performance on source domains can be attributed to the uneven favouring of some clients over others during sampling to improve the target domain. Nevertheless, the decrease in performance happens at low budgets (1%) and can be adjusted depending on the objective.

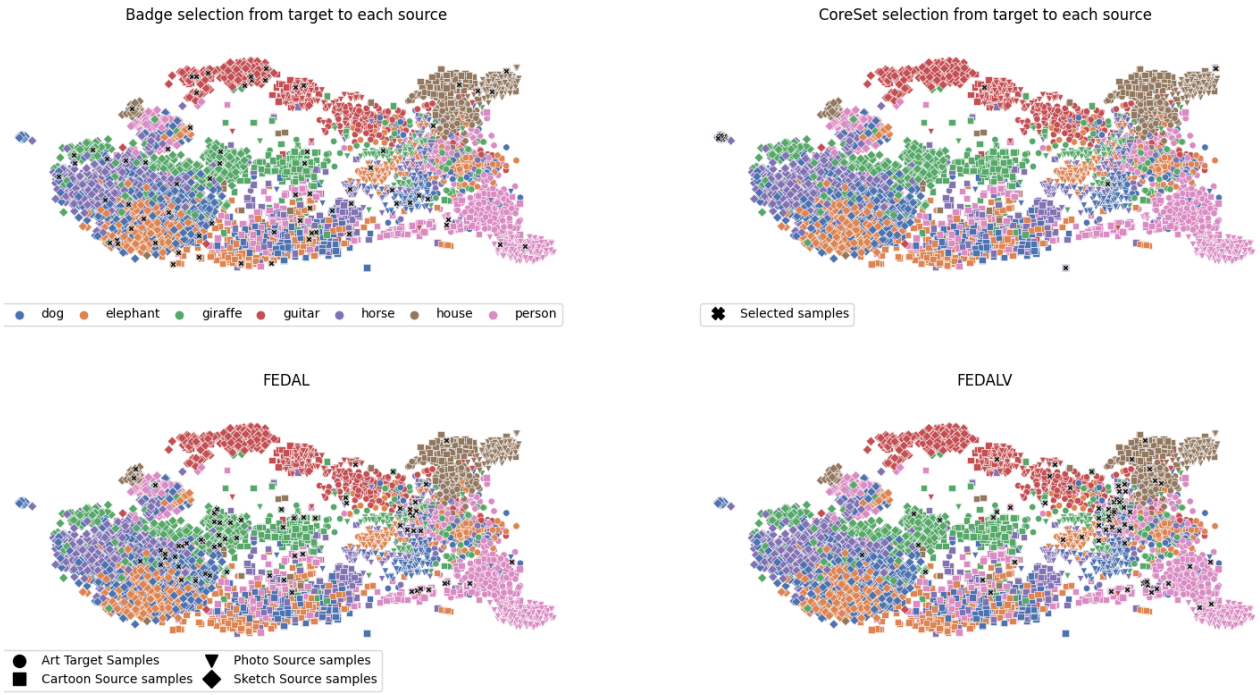

Figure A1. FEDALV: Different Selection functions [Zoom in for a better view]

A3. FEDALV - Extensive FAL Selection Analysis

A3.1. EMD measurements

In our analysis of the selection function, we investigated several approaches for sampling informative data for the target domain. Additionally, we identified, as shown in Figure 3, that target samples have higher free energies. To simplify our selection, we consider a fixed number of target samples to compare against the joint source distribution.

Therefore, given the target samples as centers, we deploy three metrics for selection: Badge [2], CoreSet [38], and the minimum Euclidean distance from FEDALV. Once selected within the defined budget, we calculate the Earth Moving Distance (EMD) between these new samples and the target ones. As expected, the source representations selected by FEDALV obtained the lowest EMD at 9.51 (compared to CoreSet 18.05 and 13.34 for Badge).

A3.2. FAL - variable metric and budget for all clients

In the FAL experiments that we conducted, we applied the AL baselines with their core selection principles in Table 3. Moreover, we extended the free energy alignment metric to other selection principles in Fig. 3 and measured the EMD in Sec. A3.1. We acknowledge that FEDALV may have an advantage due to its per-client variable selection, however this would not be possible without a global metric for sampling.

PACS/ Average over target domains	1%	2%	3%	4%	5%
Random	56.21	68.16	71.28	75.38	76.5
Entropy on all sources	54.36	62.14	67.47	75.58	73.07
CoreSet on all sources	53.64	60.54	71.28	74.54	76
FEDALV (Ours)	55.24	**71.72**	**78.07**	**81.47**	**83.63**

Table A3. FEDALV - **PACS** dataset - FAL baselines with variable client budget

Despite this, we included in Table A3 an uncertainty-based FAL method that uses class entropy and the data diversity method of CoreSet [38] under the similar budget variation. The class entropy uncertainty checks all the unlabelled images with the global model and ranks the top uncertain from all sources. For CoreSet, we initialise the centroids with the labelled samples from all sources and select from all of unlabelled. Both methods do not have a selection metric dependant of the target domain samples.

As shown in Table A3, even with these methods, FEDALV is still dominant in FAL for DG. The performance of these two baselines are as good as random sampling. This proves once again the relevance of our selection metric.

A3.3. Qualitative analysis

In Figure A1, we marked with a cross the selected samples for each criterion after the first FAL cycle. The CoreSet methodology groups the furthest samples in regard to the

target, while Badge selection has an even distribution over all the sources. However, both methods prove sub-optimal selection when aiming to reduce the misalignment with the target domain. On the other hand, FEDALV not only groups the selection closer to the high-energy target samples, but also where the classes are poorly clustered by the global model.