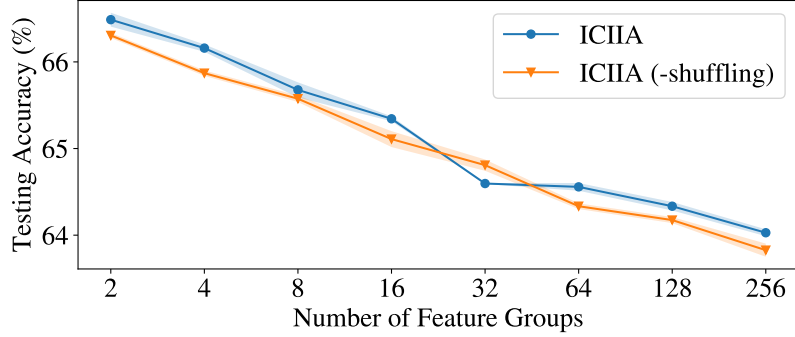
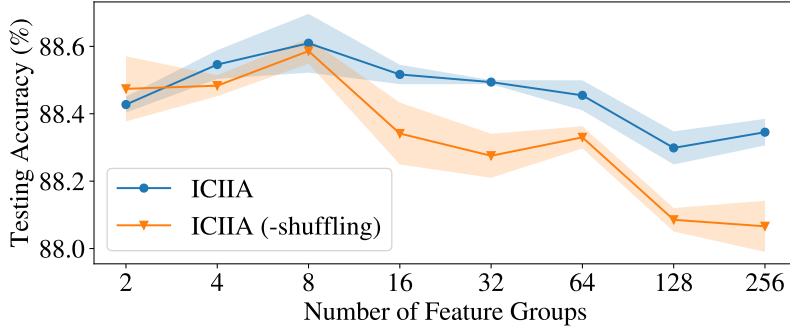


Appendix A. Supplementary Notes on the Evaluation

A.1. Ablation Study



(a) *iNaturalist 2019* with EfficientNet-B0



(b) *ImageNet-1K* with EfficientNet-B4

Figure 1: ICIIA with and without feature shuffling, by varying the number of feature groups P .

In the ablation studies, we consistently study the image classification task over *iNaturalist 2019* and *ImageNet-1K*. For each dataset, we choose a representative backbone with middle-level complexity (i.e., EfficientNet-B0 for *iNaturalist 2019* and EfficientNet-B4 for *ImageNet-1K*). As a supplement to the main document, in Figure 1, we compare ICIIA with and without feature shuffling by varying the number of feature groups P . We can observe that the accuracy drop generally becomes larger for a larger P , especially on *ImageNet-1K*.

A.2. Implementation Details

Regarding the detailed experimental settings, we use stochastic gradient descent (SGD) as the optimizer. For pretraining the original backbone model, we consistently set the batch size to 32 and the weight decay to 0.0005. Regarding the learning rate, we have tried different recipes and chosen the one which achieves the highest accuracy on the validation set for each dataset:

iNaturalist 2019 We use a momentum factor of 0.9 and set the initial learning rate to 0.01 for the last layer and 0.001 for the other layers, and let it decay by 0.1 every 10 epochs.

CelebA We set a constant learning rate of 0.01 with a momentum factor of 0.9.

FEMNIST* and *UCF101 We set a constant learning rate of 0.01 without momentum.

For the proposed ICIIA and the two baselines of fine-tuning and prompt tuning, we consistently set the batch size to 16, which is the best choice among 2, 4, 8, 16, and 32, and set the learning rate to a constant value of 0.01. A special case is *CelebA*, where the number of samples per client is small, often below the default value of 16, and we make each client one batch. For ICIIA, we try and take the optimal number of layers for each task ($N = 1$ for *FEMNIST* and *CelebA*, $N = 2$ for *iNaturalist 2019* and *UCF101*, and $N = 3$ for *ImageNet-1K*). Regarding the number of attention heads H in ICIIA, we have tried the values of 1, 2, 4 and 8, and choose $H = 4$ which achieves the highest accuracy on the validation dataset. For the dimension of the prompt tokens in prompt tuning, we have tried, $D/4$, $D/2$, D , $2D$ and $4D$, and choose $D/2$ which achieves the highest accuracy on the validation dataset.

We conduct evaluation on machines with operating system Ubuntu 18.04.3, CUDA version 11.4, python version 3.7.13, and two NVIDIA GeForce RTX 2080Ti GPUs. For each run, we train the model for at most 100 epochs, stop early if the accuracy on the validation dataset has not improved for 10 epochs, and adopt the model with the highest accuracy on the validation dataset. Each run takes roughly ten hours to complete and requires approximately 2GB of GPU memory. For each result that requires random weight initialization, we repeat with three random seeds and report the average accuracy on the testing dataset.

We implement all the models and learning algorithms with PyTorch (torch 1.12.1, torchvision 0.13.1). Source code and the detailed instructions to reproduce the results are also available in our supplementary materials.