

Federated Learning with a Single Shared Image

Sunny Soni

Universiteit van Amsterdam
sunnysoni97@gmail.com

Aaqib Saeed

TU Eindhoven
a.saeed@tue.nl

Yuki M. Asano

Universiteit van Amsterdam
y.m.asano@uva.nl

Abstract

Federated Learning (FL) enables multiple machines to collaboratively train a machine learning model without sharing of private training data. Yet, especially for heterogeneous models, a key bottleneck remains the transfer of knowledge gained from each client model with the server. One popular method, FedDF, uses distillation to tackle this task with the use of a common, shared dataset on which predictions are exchanged. However, in many contexts such a dataset might be difficult to acquire due to privacy and the clients might not allow for storage of a large shared dataset. To this end, in this paper, we introduce a new method that improves this knowledge distillation method to only rely on a single shared image between clients and server. In particular, we propose a novel adaptive dataset pruning algorithm that selects the most informative crops generated from only a single image. With this, we show that federated learning with distillation under a limited shared dataset budget works better by using a single image compared to multiple individual ones. Finally, we extend our approach to allow for training heterogeneous client architectures by incorporating a non-uniform distillation schedule and client-model mirroring on the server side.

1. Introduction

Federated Learning (FL) is a paradigm in the field of distributed machine learning that enables multiple clients to collaboratively train powerful predictive models without the need to centralize the training data [20]. It comes with its own set of key challenges in terms of skewed non-IID distribution of data between the participating clients [4, 8, 12, 13, 21] and communication efficiency during training [9, 13] among others. These challenges are not directly answered with the classical approaches such as FedAvg [14], which rely primarily on a naive client network parameter sharing approach. Since the inclusion of clients with different data distributions has a factor of heterogeneity involved [8, 21], another well-known work [12] counteracts this heterogeneity directly during the client training.

This tries to solve one challenge related to non-iidness in private data distribution, but other key challenges related to network parameter sharing remain including concerns with privacy leakage during parameter sharing [16, 17], heterogeneity of client architectures [4, 13] and high bandwidth cost of parameter sharing [9]. To this end, along a second line of thought implementing a server-side training regime, approaches suggested in [11, 13, 15, 22] make use of the process of knowledge distillation (KD) [5, 7] to overcome these challenges without the exclusive need of network parameter sharing. To facilitate central network training with the help of KD, the sharing of public data is needed between the clients and the server.

In this work, we propose a novel approach of making use of a single datum source to act as the shared distillation dataset in ensembled distillation-based federated learning strategies. Our approach makes use of a novel adaptive dataset pruning algorithm on top of generating the distillation data from a single source image during the training. This combination of shared data generation and instance selection process not only allows us to train the central model effectively but also outperforms the other approaches which make use of multiple small-sized images in place of a single image under a limited shared dataset budget. The use of a single datum source has added benefits in domains, where publicly available data and client resources (e.g., network bandwidth and connectivity) are limited in nature. The use of a single datum source has been explored [1, 2] under the settings of self-supervised learning and understanding extrapolation capabilities of neural networks with knowledge distillation, but it has not yet been explored in federated setting for facilitating model training.

We perform a series of experiments to examine the viability of our proposed algorithm under varying conditions of heterogeneity in private client data, client-server model architectures, rate of pre-training network initializations before distillation, shared dataset storage budget, and real-world domain of the single images. We also extend our experiments to a use case of heterogeneous client architectures involved during a single federated training with the help of client-model mirroring on the server side. To fa-

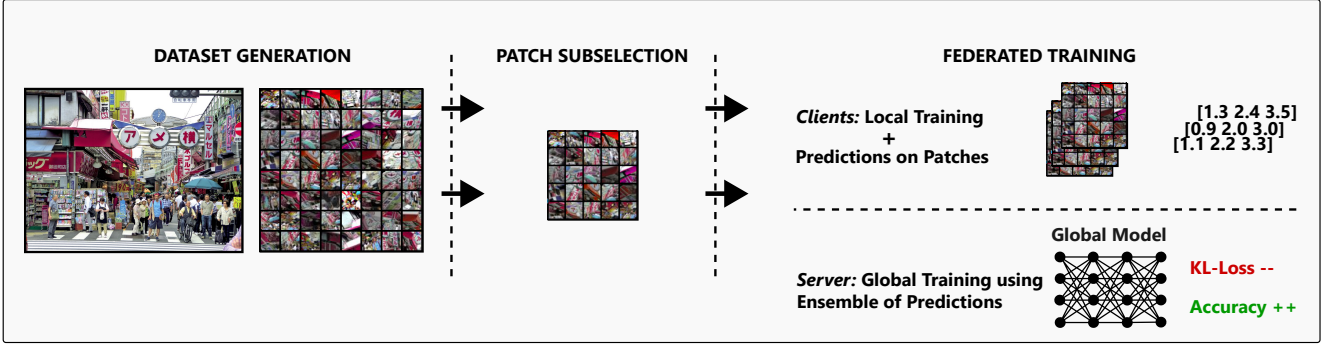


Figure 1. Illustration of our federated learning algorithm using a single image. Our algorithm works on the principle of generating a common distillation dataset from only one shared single image using deterministic augmentations. To this end, our method dynamically selects the best patches for the training of the global model in the next round using knowledge distillation.

Facilitate this, we keep one copy of the client model of each type on the server end, which acts as a global model for the clients that have the same network architecture. The global models are improved with knowledge distillation after each round of local client training with the help of shared logits over the single image dataset. The results we obtain during the aforementioned experiments demonstrate positive reinforcement towards reaching our goal of efficient federated training using Knowledge Distillation under a limited shared dataset budget.

The primary contributions of this work are :

1. Demonstrating the efficiency of a single image as a powerful medium for knowledge transfer in a federated learning setting using knowledge distillation.
2. Novel algorithm for dynamic data pruning which evolves with the current global model during federated learning.
3. Extensive evaluation of our proposed methods under a variety of conditions in a federated setting.

2. Related Work

Federated learning with knowledge distillation. Knowledge Distillation (KD) [7] has been shown to successfully transfer the knowledge of an ensemble of neural networks into a single network with the means of output logits over a shared dataset. KD has also been leveraged in federated settings, such as Federated Distillation Fusion (FedDF) [13] and Federated Ensemble Distillation (FedED) [15], where the respective authors make use of KD to allow robust and faster convergence on top of using other ensembling methods such as the ones suggested in Federated Averaging [14] for initializing the central network before the distillation training of the global model. On the other hand, authors of works such as Federated Model Distillation (FedMD) [11] have also successfully shown that KD can be used for knowledge transfer in a federated setting for the purpose of client model personalization. However, the application of

algorithms such as FedMD is targeted for personalization by client-side knowledge distillation rather than improvement of a central model, hence we have not delved into it in the scope of our research. In the case of ensembling methods, it has been shown in [13] that in the absence of an ensemble of local parameters before distillation training, the final test performance of the central network tends to suffer. As a result, these methods have been shown by the authors to significantly rely on parameter exchange every round similar to naive parameter exchange-based algorithms such as FedAvg [14] for robust performance on top of KD. Since the aforementioned KD-based federated algorithms also require significant regular ensembling using network parameter exchange, our approach focuses on improving this aspect by relying significantly on knowledge distillation with the help of data pruning and augmentations on the shared public dataset, which has not yet been explored in these works.

Communication-efficient federated learning. To solve the high bandwidth costs related to parameter sharing, authors of [3, 9] have shown that quantization of network parameters before the transfer can significantly reduce the bandwidth costs incurred during their transfer. However, with the application of the same low-bit quantization methods with the KD-based federated learning methods in [13], the authors have also shown a significant decrease in the overall test performance of models compared to their non-quantized counterparts. On the other hand to not rely on public data sources, authors of the work [22] have successfully shown that data-free approaches using a centrally trained generative network for producing the public shared dataset work robustly. However, this also requires an additional exchange of the generative network parameters before each round, which leads to an increase in the network bandwidth usage itself. In pursuit of reducing the bandwidth costs pertaining to network parameter exchanges as well as shared dataset sharing, these works have not yet

made an attempt to make use of a storage-efficient single data source, which can simultaneously generate a public distillation dataset alongside being used for dynamic selection without added bandwidth costs. We explore this in our work.

Single image representation learning. Asano et al. [2] have successfully made use of a single image to produce augmented patches for facilitating self-supervised learning of useful representations required for solving various downstream tasks. However, the focus of our work is not on the process of solving tasks with the help of self-supervised learning, but on the implications of making use of the single image patches in a federated setting as a medium of knowledge transfer for training robust classifiers. To this end, in a closely resembling work to our target task, the authors in [1] have shown to be able to successfully use KD with single image patches to transfer the knowledge between a pre-trained network and an untrained network to solve the classification task of ImageNet-1k. However, the experiments by the authors were all based in a non-federated setting. In our work, we explore the usage of single image patches in a federated setting as the shared public distillation dataset and its implications in limited shared dataset budget settings.

3. Methodology

Our methods focus on a dynamic procedure to utilize a single image to act as a proxy dataset for distillation in addition to a federated learning setup which is similar to existing ensemble-based knowledge distillation-based methods such as FedDF [13]. Alongside the generation of a distillation dataset from a single data source, we take care of dynamically selecting the best patches every round to improve the training. The two important parts of our federated strategy are: *Distillation Dataset Generation* (Sec. 3.1) and *Distillation Dataset Pruning* (Sec. 3.2).

3.1. Distillation Dataset Generation - Patchification

For generating meaningful image representations out of a single image, we make use of the *patchification* technique. Using this technique, we generate a large number of small-size crops from a single image by making combined use of augmentations such as 35-degree rotations, random horizontal flipping, color transformations, etc. We use the same order of augmentations as the one tested in [1] for knowledge distillation-based learning. The image generation procedure is controlled by a given seed to ensure consistent dataset generation when required. This method is communication bandwidth efficient because:

1. It requires the transfer of a single image only once across the network to all the clients during the entire training, which may also be present before the training itself.
2. The distillation dataset generation is done locally using a seed number, hence only index values need to be trans-

ferred when selecting a subset every round.

The generated dataset is used as the proxy dataset for improving the global model using Knowledge Distillation with clients' current predictions. Due to the flexibility provided by augmentations in combination with the subset selection procedure described in Sec. 3.2, one can make use of a single image to produce a varying desired number of patches for the fixed amount of single image data.

3.2. Distillation Dataset Pruning - Subset Selection

After obtaining an initial dataset for knowledge distillation using the method in Sec. 3.1, we apply dataset pruning methods to it to ensure the selection of information-rich patches for the current round of federated training. The dataset generation procedure makes use of a single information-rich image to generate the small patches, due to which it can produce bad representation patches such as: containing no entities, overlapping with others, being dissimilar to the target domain, and having similar problems arising due to heavy augmentations and presence of information-less regions of the single image. To prune the bad patches, we make use of the following two mechanisms: *KMeans Based Class Balancing* (Sec. 3.2) and *Entropy Based Pruning* (Sec. 3.2). These mechanisms depend on the current global model for their operation, which makes them dynamic in nature and improves their data-pruning ability with the improvement in the global model. Hence, better global models yield better representations.

Through the t-SNE visualization in Fig. 2 using a single image with our data pruning methods, we observe the formation of better identifiable boundary structures with a more accurate global model. It provides us with a glimpse of the positive correlation between the effectiveness of our patch subset selection methods with the evolution of the global model.

KMeans-based class balancing. KMeans-Based Class Balancing is based on the use of unsupervised KMeans clustering on the embedding layer representations of the training examples. To accomplish this, we establish a KMeans clustering model (based on Euclidean distance) with k cluster centeroids and try to fit the embedding representations (using the current global model) of the distillation training examples on it. Using a selection heuristic $H^K \in \{\text{Easy, Hard, Mixed}\}$ on their respective cluster distances, we can select the training examples for the next round of training. It is given by the equation 1. We also involve a lower-bound factor $F^K \in [0.0, 1.0]$ ensuring a fixed lower bound on the number of training examples from each of the predicted classes. The algorithm has been described in detail in the Appendix A.

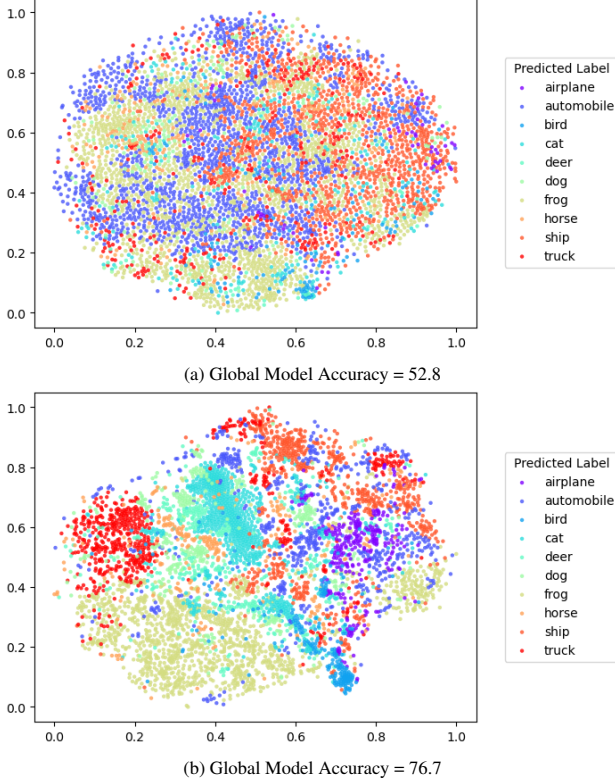


Figure 2. Scatter plot of t-SNE embeddings of pruned single image patches during different phases of training, using our method with FedAvg and ResNet-8 on CIFAR10.

Given : Distillation Dataset \mathbf{X} , Respective Embeddings $f(\mathbf{X}; \theta)$, KMeans Cluster Centroids k , Selection Heuristic H^K ;

$G(\mathbf{X}, f(\mathbf{X}; \theta), k, H^K) = \text{KMeans Based Class Balancing s.t.}$

$G(\mathbf{X}, f(\mathbf{X}; \theta), k, H^K) = \mathbb{Y} \subset \mathbf{X}$,

$$\text{where } \mathbb{Y} = \begin{cases} \text{Ex. with low } |f(\mathbf{X}; \theta) - k|, & \text{if } H^K = \text{Easy} \\ \text{Ex. with high } |f(\mathbf{X}; \theta) - k|, & \text{if } H^K = \text{Hard} \end{cases} \quad (1)$$

Entropy-based pruning. Entropy-Based Pruning makes use of randomness present in the output logits of the distillation training examples to select a subset of them. To achieve this, we examine the max softmax values of the logits obtained for every distillation training example using the current global model. On the basis of a removal heuristic $H^E \in \{\text{Top, Bottom, Random}\}$, we remove e percent of examples from each group (grouped on the basis of their predicted class using current global model). It is given by the equation 2. The algorithm has been described in detail in the Appendix A.

Given : Distillation Dataset \mathbf{X} , Classifier Output $g(\mathbf{X}; \theta)$,

Removal Heuristic H^E ;

$G(\mathbf{X}, g(\mathbf{X}; \theta), H^E) = \text{Entropy Based Pruning s.t.}$

$G(\mathbf{X}, g(\mathbf{X}; \theta), H^E) = \mathbb{W} \subset \mathbf{X}$,

$$\text{where } \mathbb{W} = \begin{cases} \text{Ex. with low } \arg \max(g(\mathbf{X}; \theta)), & \text{if } H^E = \text{Top} \\ \text{Ex. with high } \arg \max(g(\mathbf{X}; \theta)), & \text{if } H^E = \text{Bottom} \end{cases} \quad (2)$$

3.3. Federated Strategy with Subset Selection using Single Image Dataset

Using the mechanism described in Sec. 3.1 and Sec. 3.2, we can perform federated training using a single image according to the steps in Alg. 1. *FedAvg Initialisation* makes use of naive weighted averaging of parameters on top of KD training for global model updates, which can be toggled with different uniform schedules. We make use of uniform intervals on the basis of defined rates in this work.

Computation overhead. Since the clients using our method only do supervised training similar to FedAvg [14], there's no computation overhead on the client side during the training. The compute cost of forward inferencing a small distillation dataset is negligible compared to the supervised training the clients are subjected to.

4. Experiments

4.1. Experimental Setup

Dataset. We do our experiments across the following publicly available datasets: CIFAR10/100 [10] and MedMNIST (PathMNIST) [18]. For the distribution of private data among the clients from the collective training data, we use a similar strategy to the one suggested in [8], which allows us to control the degree of heterogeneity using the parameter α (lower α = higher degree of non-iidness and vice-versa). We use the full test sets corresponding to the private client datasets as evaluation sets for the global model (testing is only server-side). 10% of the training examples are held as a validation dataset.

For the shared public dataset, we generate patches out of a single image for all the experiments with our method. For the FedDF experiments, we have made use of CIFAR100 training set for CIFAR10 experiments, unless mentioned otherwise. The single images have been visualized in Appendix A alongside the patches and t-SNE visualizations.

Server-client model architecture. ResNets (trained from scratch) have been used for most of our experiments as the model of choice for the server and clients [6]. We have also used *heterogeneous network distributions* in Sec. 4.3, with clients having varied network architectures training together. WideResNets have also been used for some of the

Algorithm 1: Federated Learning with Single Image

```
begin
1  Generate distillation dataset  $\mathbf{X}$  on all clients and the server independently using a single image and fixed seed.
2  Initialise the global model  $\theta^G$  on the server. (More than one in case of heterogeneous architectures)
3  forall rounds of training do
4      Using data pruning, select indices of distillation training examples.
5      Send the indices and current global model  $\theta^G$  to local clients selected for the next round of training.
6      forall Local Clients (Selected) in Parallel do
7          Update local model parameters with the received global model.
8          Do supervised training using the private dataset to improve the local model.
9          Predict output on selected distillation training examples.
10         Send predictions to the server. If FedAvg Initialisation is enabled for the current round, also send the local model
            parameters.
        end
11     If FedAvg Initialisation is enabled, initialize the global model for KD training using the weighted average of local
        parameters.
12     Using KD training on the weighted average of predictions, update the global model.
end
end
```

experiments [19]. Our method should be directly applicable to all kinds of network architectures that have an intermediate embedding layer present. The models have been explicitly defined in the table descriptions for unambiguity.

Hyper-parameter configuration. The values of the learning rate (local and global) have been motivated by the experiments described in Appendix D. We use a client learning rate of 0.01 for ResNet and WResNet, while the distillation training learning rate is 0.005. For KMeans Balancing, we use a KMeans model with 1000 clusters, a class balancing factor of 1.0, and the 'Hard' selection heuristic. For Entropy selection, we remove 90% of the training examples using the 'Top' removal heuristic (Appendix C). For the experiment in Table 5, we do local client training for 10 epochs and server-side distillation for 250 steps, while 40 epochs and 500 distillation steps have been our choice for other experiments unless mentioned otherwise. We prefer to keep the use of FedAvg initializations to 20% in our experiments unless mentioned otherwise. For all the experiments, we simulate **20 private clients**, with a selection probability (C) of 0.4 per training round.

FedAvg Initialisation Rate. FedAvg Initialisation Rate is responsible for controlling the schedule of FedAvg Initialisations (Sec. 3.3), where the percentage decides the round intervals in which we initialise the central network with them before distillation training. For reference, 20% means using FedAvg Initialisations every 5th round and 100% means using them every round.

4.2. Network and Storage Conditions

Comparative analysis of single image performance under limited shared dataset budget.

This is our most significant experiment in terms of exhibiting the viability of federated learning under limited shared dataset budget settings using a single image. Going through the results in Table 1, we see that for the same amount of storage budget, a single image with patch selection outperforms similarly sized individual samples. If we also lower the network budget and the rate of exchange of initialization parameters, it is able to hold at par with individual training samples 10 times its size. This shows a promising future for our work in the scenario where there is limited availability of publicly shared datasets as well as storage budget would be low on participating clients.

Performance evaluation in limited network bandwidth settings against heterogeneous data distributions. To test the impact of high data distribution heterogeneity on our FL strategy against an existing SOTA federated learning strategy based on knowledge distillation, we show the performance gains in Table 2. We also vary the network initialization rate to test our method in high and low-bandwidth situations. We notice that with the help of patch subset selection, our methods outperform the fed strategy which doesn't make use of this process. This trend is constant across all bandwidth scenarios and local client training expenditures. We have also extended our approach to incorporate FedProx local client training regime, which shows better results than naive local client training. This extendability makes our method unique and viable to more approaches than just one kind of local training which can have added performance

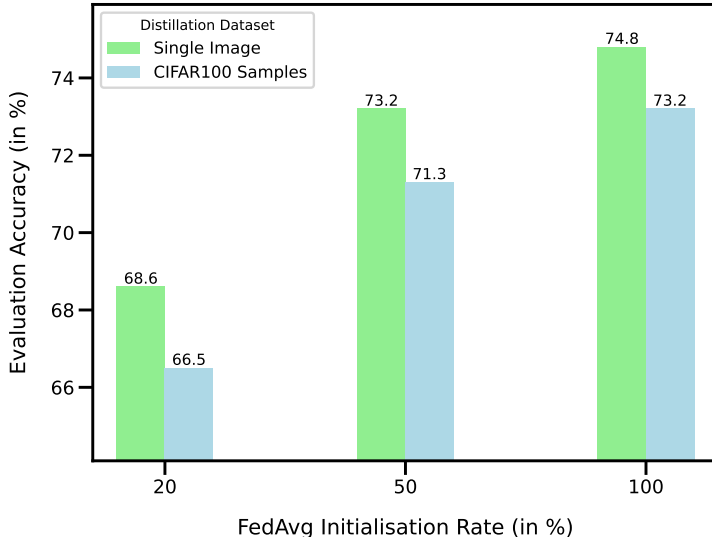


Figure 3. Comparison of test performance in federated setting using a single image with patch selection compared to the equivalent size of multiple independent training samples from a labeled dataset as shared distillation dataset. We use different rates of FedAvg. initializations to emulate different network bandwidth conditions. Detailed result in Table 1.

Distillation Dataset	Number of Pixels	FedAvg Initialisation Rate (in %)		
		100	50	20
5K CIF100 Samples	5M	76.4 ± 1.4	74.1 ± 1.6	68.9 ± 1.4
Single Image with Patch Selection	0.5M	74.8 ± 2.6	73.2 ± 3.2	68.6 ± 0.8
500 CIF100 Samples	0.5M	73.2 ± 1.7	71.3 ± 2.0	66.5 ± 0.9

Table 1. Best test performance during 30 rounds of training with CIFAR10 Private Data with Distribution $\alpha = 1.0$ using ResNet-8 with **different distillation datasets** and rate of using FedAvg initialisation.

benefits with our algorithm.

4.3. Client-Server Neural Network Architectures

Evaluating our strategy under homogeneous network Architecture. We perform all the experiments in the earlier sections using ResNet-8 as the client and server models. To make sure our federated strategy works equally well among other homogeneous network distributions, we put it to the test against FedDF using ResNet-20 as well as W-ResNet-16-4 in Table 3. We see that under the same distillation set storage budget, our method works better under all the tested network architectures. As per nominal expectations, network architectures with more parameters show better results than the ones with less number of parameters which enables us to achieve better test performance with more complex networks. Irrespective of the network architecture, the trend is constant when it comes to our FL strategy outperforming other strategies making use of a labelled distillation dataset in a limited storage budget scenario.

Evaluating our strategy under heterogeneous network architecture. In the final experimental section, we test our

federated strategy in the presence of heterogeneity in the client model architectures. The results present in Table 4 show the success of our method in training the global models when pitted against a strategy not utilizing a single image. It also exhibits the importance of constant distillation training for the success of our methods, as our non-uniform approach gives subpar results with less training time. However, when going from 15k to 11k steps, we also save about 1/3 of the training time and computation resources used on the server side. It can be an interesting point of extension to our work to improve upon this non-uniform scheduling to allow for more robust training of heterogeneous models with less computation time.

4.4. Patch Selection Mechanism

Optimal patch sub selection strategies across private dataset. To find the effectiveness of patch subset selection mechanisms, we test it under different private datasets from different real-world domains (General and Medical). Through Table 5, it is evident that the single image patches work best in the presence of a selection strategy in our fed-

Strategy	Local Epochs	FedAvg Initialisation Rate (in %)					
		100		50		20	
		$\alpha = 1.0$	$\alpha = 0.1$	$\alpha = 1.0$	$\alpha = 0.1$	$\alpha = 1.0$	$\alpha = 0.1$
FedDF	20	75.7 \pm 1.2	48.2 \pm 2.6	73.9 \pm 0.8	47.3 \pm 5.2	71.1 \pm 0.5	42.2 \pm 9.4
	40	75.7 \pm 0.9	49.5 \pm 3.1	74.9 \pm 1.9	49.3 \pm 1.1	72.5 \pm 0.5	46.1 \pm 6.6
Ours w/ FedAvg	20	76.9 \pm 0.6	47.8 \pm 5.3	75.8 \pm 0.3	47.3 \pm 5.5	73.7 \pm 1.0	45.5 \pm 5.1
	40	77.0 \pm 0.6	47.8 \pm 5.4	76.2 \pm 1.4	49.5 \pm 2.2	74.3 \pm 0.6	46.6 \pm 6.7
Ours w/ FedProx	20	77.2 \pm 0.8	47.2 \pm 7.0	74.5 \pm 1.3	44.6 \pm 7.9	73.1 \pm 0.2	46.9 \pm 4.5
	40	77.7 \pm 0.8	47.7 \pm 3.8	76.3 \pm 0.4	46.0 \pm 5.3	74.3 \pm 1.1	45.1 \pm 6.0

Table 2. Comparison of best test performance under different settings (FedAvg Initialisation Rate, **Degree of Heterogeneity** (α), Local Training Epochs) using different federated learning strategies with ResNet-8 on CIFAR10 during 30 rounds of training (2 seeds). 5000 single image patches have been used as distillation proxy set (w/o selection mechanism for FedDF).

Fed Strategy	Network Architecture		
	ResNet-8	ResNet-20	W-ResNet-16-4
FedDF	67.3 \pm 1.9	73.0 \pm 0.6	75.3 \pm 1.2
Ours	70.2 \pm 0.8	74.1 \pm 0.9	75.7 \pm 0.9

Table 3. Best test performance during 30 rounds of training using CIF10 Private Data with Distribution $\alpha = 1.0$ using different Fed strategies and **homogeneous client-server network architectures** with 20% rate of FedAvg. initialisation. FedDF uses 500 CIF100 samples as distillation proxy, while our method makes use of a single image of equivalent size with patch subset selection.

Fed Strategy	Total Distillation Steps	Macro-Avg Accuracy (Hetero Server Models)
FedDF	15K	67.4 \pm 0.6
Ours	15K	68.5 \pm 1.1
Ours with Scheduling	11.3K	65.2 \pm 1.3

Table 4. Best test performance across during 30 rounds of training using CIF10 Private Data with Distribution $\alpha = 1.0$ using different Fed strategies and distillation step scheduling, under a **heterogenous client architecture distribution (6 ResNet-8, 7 ResNet-20, 7 W-ResNet-16-4)** with 20% rate of FedAvg. Initialisation. 500 CIF100 samples have been used as distillation proxy for FedDF, while our method makes use of a Single Image of equivalent size with patch selection.

erated algorithm. On their own, both KMeans Balancing (3.2) and Entropy Selection (3.2) strategy works better than employing no selection for the same number of patches. Together, they perform best across all the datasets we have tested which is what we use in our other experiments in this work. Both of the selection strategies and their combination significantly impact the final performance. We have done our primitive analysis with them in light of this work to find

an optimal setting (Appendix C, but there might be a correlation between their settings which we have not delved into. We can propose this detailed analysis of their combinative work as future work for improving the test performance of our federated strategy with the means of better data pruning.

Selection Strategy	Private Dataset		
	CIFAR10	CIFAR100	PathMNIST
No Selection	63.4 \pm 1.4	24.2 \pm 1.1	64.5 \pm 4.7
KMeans	66.2 \pm 0.8	21.8 \pm 2.1	67.9 \pm 8.4
Entropy	65.9 \pm 1.0	26.3 \pm 1.0	76.4 \pm 2.8
KMeans + Entropy (Ours)	67.0 \pm 1.1	26.4 \pm 1.2	77.1 \pm 3.0

Table 5. Best test performance achieved during 30 rounds of training with **different selection mechanisms** (Distillation Set Size = 5000 patches) across different private datasets ($\alpha = 1.0$) using our federated strategy with ResNet-8 while using 20% rate of FedAvg. initialisations. (2 seeds)

Impact of selection mechanism with manually labelled distillation set. We test the viability of our selection mechanism in case of extending it to the use cases where we already have a shared public dataset at hand in Table 6. During the regular exchange of initialization parameters, the application of our selection mechanism exhibits no advantage. However, when we reduce the exchange of initialization parameters to emulate low bandwidth conditions, it shows significant gains. This shows that even with the availability of standard distillation sets at hand in ensembled distillation methods, the subset selection mechanism can play an important role in low bandwidth cost federated training.

4.5. Selecting the Best Source of Single Image

We conduct cross-dataset single-image testing using our algorithm across 3 private training datasets and 3 images, with two of them corresponding to one of the dataset domains and the third one being a random noise. The results in Table 7 exhibit that it is necessary to use a single image that

Selection Mechanism Applied	FedAvg Initialisation Rate (in %)		
	100	50	20
✗	75.0 ± 0.5	73.1 ± 1.1	67.4 ± 0.5
✓	73.8 ± 1.9	72.3 ± 0.6	70.7 ± 1.2

Table 6. Comparison of best test performance during 30 rounds of training with CIFAR10 Private Data with Distribution $\alpha = 1.0$ using FedDF (with ResNet-8) between **use/non-use of selection mechanism** across varying rate of using FedAvg initialisation. 1000 samples from CIFAR100 train split make the distillation proxy dataset.

is similar to the domain of the target task for optimal performance. In the case of using a single random noise image as the distillation proxy, we get the lowest test performance as it is hard for random augmentations to convert random noise patches into a knowledge transfer medium. Hence, care must be taken in choosing a single image with similar patch representations as the target task for optimal learning with our algorithm. There can be an interesting area to explore with more augmentation experiments and generative algorithms if it is possible to use a random noise image viably as a single image with our method. We leave this as future work.

Private Dataset	Single Image Source		
	City Street	Pathology Samples	Random Noise
CIFAR10	75.3	69.0	39.4
CIFAR100	32.0	12.0	6.8
PathMNIST	69.7	71.6	33.0

Table 7. Best test performance during 30 rounds of training using our federated method with **varying Private Datasets** (Distribution $\alpha = 100.0$) and **varying Single Image Sources** (5k Patches) (Distillation Proxy Set) on ResNet-8 architecture with 20% rate of FedAvg. initialization.

5. Conclusion

Through this work, we present a novel approach for federated learning using ensembled knowledge distillation with the use of augmented image patches from a single image with patch subset selection. We successfully exhibit the performance gains with our approach in a limited shared dataset budget scenario as well as low network bandwidth requiring scenarios with less exchange of network parameters. Alongside low resource usage, the use of a single image also enables our federated strategy to be applicable to scenarios where we have a lack of public datasets required during federated training of multiple clients.

Future Work. We mention a few specialized avenues of extension to our work during the discussion of results in Section 4. Some of the key points that were not mentioned in it include: Application of the single datum-based federated learning to other modalities and machine learning tasks; Application of our work to other knowledge distillation-based algorithms in federated learning other than ensembled methods, such as FedMD [11]; Analysis of different kind of augmentations to improve the robustness of our method. With the aforementioned points, significant work can be done to improve the viability of our novel approach presented in this work to incorporate more real-world challenges.

References

- [1] Yuki M. Asano and Aaqib Saeed. The augmented image prior: Distilling 1000 classes by extrapolating from a single image, 2023. 1, 3
- [2] Yuki M. Asano, Christian Rupprecht, and Andrea Vedaldi. A critical analysis of self-supervision, or what we can learn from a single image, 2020. 1, 3
- [3] Sebastian Caldas, Jakub Konečný, H Brendan McMahan, and Ameet Talwalkar. Expanding the reach of federated learning by reducing client resource requirements. *arXiv preprint arXiv:1812.07210*, 2018. 2
- [4] Zheng Chai, Hannan Fayyaz, Zeshan Fayyaz, Ali Anwar, Yi Zhou, Nathalie Baracaldo, Heiko Ludwig, and Yue Cheng. Towards taming the resource and data heterogeneity in federated learning. In *2019 USENIX conference on operational machine learning (OpML 19)*, pages 19–21, 2019. 1
- [5] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021. 1
- [6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4
- [7] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1, 2
- [8] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019. 1, 4
- [9] Jakub Konečný, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016. 1, 2

- [10] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. [4](#)
- [11] Daliang Li and Junpu Wang. Fedmd: Heterogenous federated learning via model distillation, 2019. [1](#), [2](#), [8](#)
- [12] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020. [1](#)
- [13] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. *Advances in Neural Information Processing Systems*, 33:2351–2363, 2020. [1](#), [2](#), [3](#)
- [14] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data, 2023. [1](#), [2](#), [4](#)
- [15] Dianbo Sui, Yubo Chen, Jun Zhao, Yantao Jia, Yuan-tao Xie, and Weijian Sun. FedED: Federated learning via ensemble distillation for medical relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2118–2128, Online, 2020. Association for Computational Linguistics. [1](#), [2](#)
- [16] Jingwei Sun, Ang Li, Binghui Wang, Huanrui Yang, Hai Li, and Yiran Chen. Soteria: Provable defense against privacy leakage in federated learning from representation perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9311–9319, 2021. [1](#)
- [17] Zhibo Wang, Mengkai Song, Zhifei Zhang, Yang Song, Qian Wang, and Hairong Qi. Beyond inferring class representatives: User-level privacy leakage from federated learning. In *IEEE INFOCOM 2019-IEEE conference on computer communications*, pages 2512–2520. IEEE, 2019. [1](#)
- [18] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023. [4](#)
- [19] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. [5](#)
- [20] Chen Zhang, Yu Xie, Hang Bai, Bin Yu, Weihong Li, and Yuan Gao. A survey on federated learning. *Knowledge-Based Systems*, 216:106775, 2021. [1](#)
- [21] Hangyu Zhu, Jinjin Xu, Shiqing Liu, and Yaochu Jin. Federated learning on non-iid data: A survey. *Neurocomputing*, 465:371–390, 2021. [1](#)
- [22] Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In *International conference on machine learning*, pages 12878–12889. PMLR, 2021. [1](#), [2](#)

Federated Learning with a Single Shared Image

Appendix

Table of Contents

A Detailed Algorithm : Dataset Pruning	1
A.1. KMeans Based Class Balancing	1
A.2. Entropy Based Pruning	1
B Image and Patch Visualisations	1
B.1. Visualisation of Single Images	1
B.2. Visualisation of Image Representations from Single Images	1
C Finding Optimal Hyper-Parameters for Selection Algorithms	3
C.1. KMeans Balancing	3
C.2. Entropy Selection	3
D Learning Rate Optimisation	3
D.1. Client Side LR	3
D.2. Server Side LR	4

A. Detailed Algorithm : Dataset Pruning

A.1. KMeans Based Class Balancing

Alg. 2 describes the implementation of our KMeans-based balancing strategy used for data pruning.

A.2. Entropy Based Pruning

Alg. 3 describes the implementation of our Entropy-based pruning strategy used for data pruning.

B. Image and Patch Visualisations

B.1. Visualisation of Single Images

We make use of the images depicted in Fig. 4 as the sources for generating our distillation dataset. Kindly note, that we only used the images for non-profit educational research purposes and we do not hold any rights over their commercial use. These images have been selected in correspondence to the domains of datasets we have tested in our work.

Algorithm 2: K-Means Balancing

Input: Distillation Training Dataset (X), Current Global Model (M^G)
Parameters: Number of Clusters (K), Size of New Dataset (s), Balancing Factor (F^K), Selection Heuristic (H^K)
Output: Pruned Distillation Training Dataset with KMeans Selection (X^K)

```

begin
1  For all  $x_n \in X : n \in [1..S]$  where  $S = \text{size of } X$ ,
   find  $Z = \{z_n : z_n = \text{Embedding Representation } (M^G, x_n)\}$  and
    $Y = \{y_n : y_n = \text{Max-Index (Classifier Output } (M^G, x_n))\}$ .
2  Define  $C^P = \{\text{Set of unique classes in } Y\}$  and
   Number of unique classes  $C = |C^P|$ .
3  Initialise an independent unsupervised KMeans
   Clustering Model ( $M^C$ ) using  $K$  number of
   cluster centers. Fit  $M^C$  on  $Z$  and find
    $D = \{d_n : d_n = \{\text{Shortest euclidean distance of } z_n \text{ to its cluster center}\}$ .
4  Define the minimum number of examples
   (balancing lower bound) to be selected from each
   class  $c_i \in C^P$ , as  $LB = \lceil \frac{s}{C} * F^K \rceil$ .
5  forall  $c_i \in C^P : i \in [1..C]$  do
   Find the indices of examples belonging to the
    $c_i$  using  $y_n \in Y : y_n = c_i$ .
   Select indices of the new training examples on
   the basis of selection heuristic
    $H^K \in \{\text{Easy, Hard, Mixed}\}$  with their
   corresponding cluster distance values  $D$ .
   Push the training examples from  $X$  with
   selected indices in the new dataset ( $X^K$ ).
   Remove the selected training examples from
    $X$  and  $D$ .
   end
6  Remaining number of examples to be selected can
   be calculated as given by :  $s - \text{size of } X^K$ .
7  Using selection heuristic =  $H^K$  on the cluster
   distance values in  $D$ , find the indices of the
   remaining examples to be selected. Push the
   training examples with selected indices in the
   new dataset ( $X^K$ )
end

```

B.2. Visualisation of Image Representations from Single Images

In Fig. 2, we visualize the TSNE embeddings of the training examples in our single image distillation dataset during different phases of training on a scatterplot. Its implications have been discussed in the main paper in brief detail. To add to the earlier analysis, by looking at the difference be-

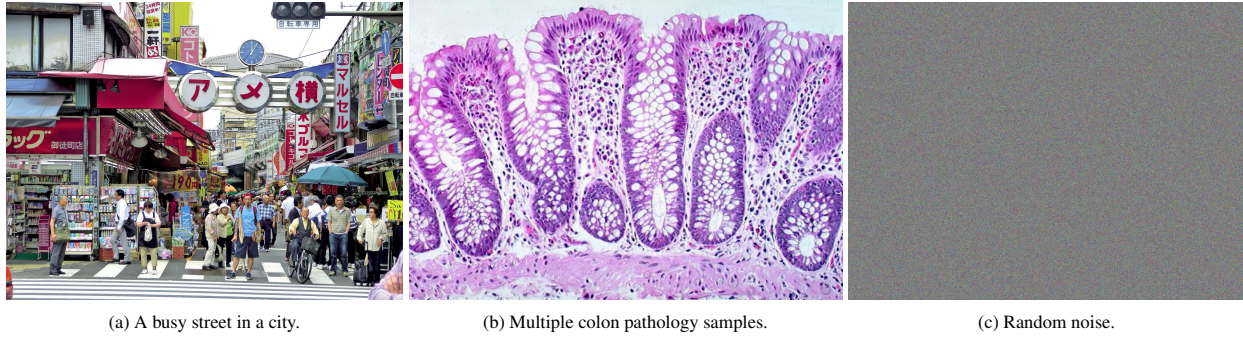


Figure 4. Single Image sources used for our experiments for distillation dataset generation.

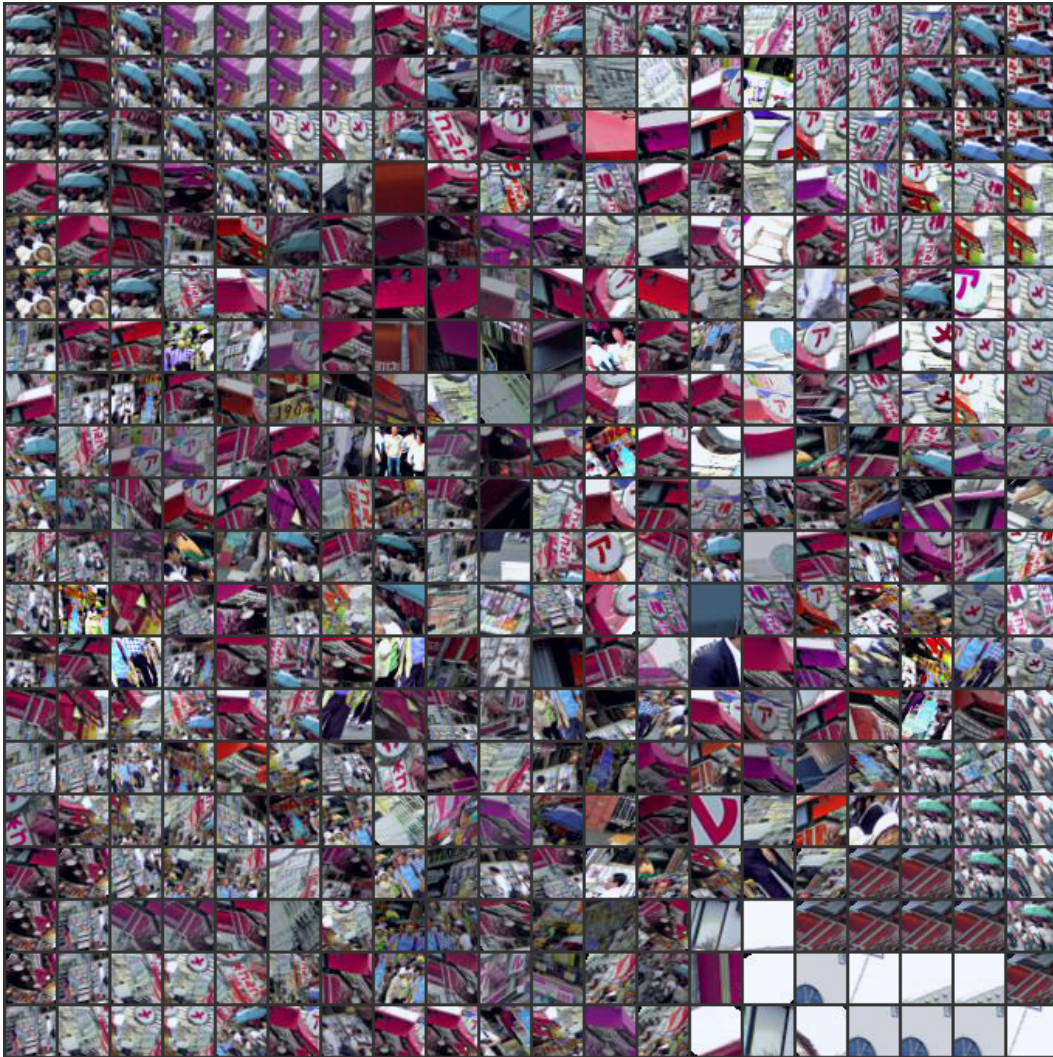


Figure 5. t-SNE Manifold visualization of distillation dataset corresponding to t-SNE scatter plot in Fig. 2b).

tween Fig. 2a and Fig. 2b, one can see how the embedding layer in the neural network starts forming particular regions belonging to certain classes with the increase in classifica-

tion accuracy. Even though the image representations do not contain many of the classes in the classifier as is prevalent from visual inspection of Fig. 5, the pseudo labels are

Algorithm 3: Entropy Selection

Input: Distillation Training Dataset (X), Current Global Model (M^G)

Parameters: Percentage of Examples to Prune (e), Removal Heuristic (H^E)

Output: Pruned Distillation Training Dataset with Entropy Selection (X^E)

begin

```
1   For all  $x_n \in X : n \in [1..S]$  where S = size of X,
    find  $Y = \{y_n : y_n = \text{Max}(\text{Softmax}(\text{Classifier Output}(M^G, x_n)))\}$ .
2   Select indices of the  $(100 - e)\%$  training examples
    using the removal heuristic
     $H^E \in \{\text{Top, Bottom, Random}\}$  with their
    corresponding values in  $Y$ .
3   Push the selected indices in the new dataset ( $X^E$ ).
end
```

powerful enough to enable the learning through knowledge distillation in a federated setting.

C. Finding Optimal Hyper-Parameters for Selection Algorithms

C.1. KMeans Balancing

To examine the impact of KMeans Balancing and its corresponding settings on the federated training, we conduct experiments while varying their values on multiple datasets having different numbers of prediction labels. The results have been presented in Table 8. Although values do not differ by a large margin for CIFAR10, CIFAR100 results provide us more assurance for the optimal values. We find that the KMeans selection strategy works best with a high number of clusters (K) compared to the number of classes in the corresponding classification problem (Table 8a), while forcing maximum class balancing while pruning (Table 8b) as well as selecting the hard examples (Table 8c) for distillation. Since the KMeans model is an independent model working on giving the pseudo-labels to the distillation training examples, there could be a play of correlation between the 3 hyper-parameters for this mechanism. It can be an interesting point for future research on this.

C.2. Entropy Selection

For the entropy selection mechanism, we only have 2 settings to vary: Removal Percentage and Heuristics (Section 3.2). The results obtained during ablation studies with these settings have been presented in Table 9. From the results in Table 9a, it is clear that removing training examples with high confidence (less entropy) provides us best results. Removing a high percentage of training examples from a large

Number of Clusters	Private Dataset	
	CIFAR10	CIFAR100
5	64.3 \pm 0.7	19.9 \pm 1.9
10	65.2 \pm 0.9	21.2 \pm 0.1
50	65.3 \pm 0.3	22.9 \pm 0.9
100	65.5 \pm 1.1	18.9 \pm 2.9
1000	66.3 \pm 1.1	22.9 \pm 0.1

(a) Varying cluster number with **selection heuristic = easy** and **balancing factor = 0.1**.

Balancing Factor	Private Dataset	
	CIFAR10	CIFAR100
0.0	64.9 \pm 1.3	22.5 \pm 0.2
0.05	65.0 \pm 1.7	22.5 \pm 0.4
0.1	65.2 \pm 2.0	21.3 \pm 0.1
0.5	66.6 \pm 1.4	22.2 \pm 0.1
1.0	66.0 \pm 1.2	24.0 \pm 1.4

(b) Varying balancing factor with **number of clusters (K) = 1000** and **selection heuristic = easy**.

Selection Heuristics	Private Dataset	
	CIFAR10	CIFAR100
Easy	65.8 \pm 1.3	22.1 \pm 0.4
Hard	66.7 \pm 1.1	23.8 \pm 1.8
Mixed (50-50)	65.5 \pm 2.1	21.8 \pm 1.0

(c) Varying selection heuristics with **number of clusters (K) = 1000** and **balancing factor = 1.0**.

Table 8. Best test set accuracy achieved during 30 rounds of training with KMeans balancing under different settings with different private datasets (Distribution $\alpha = 1.0$) using single image patches as the distillation dataset with 20% FedAvg initialisation rate on ResNet-8 (across 2 seeds).

initial set using this mechanism also provided us with more robust training, compared to removing a smaller number of training examples from a small initial set (Table 9b). Similar to our last experiments with the KMeans mechanism, the results are more clearly pronounced in the presence of a more difficult dataset (100 classes compared to 10 in the case of CIFAR100 and CIFAR10).

D. Learning Rate Optimisation

D.1. Client Side LR

We find the optimal learning rate for local model training by using vanilla FedAvg as represented in Table 10 on ResNet-8. We do a grid search across a certain set of values for it, which is described in Table 10.

Removal Heuristic	Private. Dataset	
	CIFAR10	CIFAR100
Top	67.0 \pm 0.9	26.2 \pm 1.9
Bottom	61.9 \pm 2.4	15.4 \pm 2.4
Random	67.4 \pm 1.5	23.2 \pm 0.9

(a) Varying Removal Heuristic with **Removal Percentage = 90%**.

Removal Percentage (%)	Private Dataset	
	CIFAR10	CIFAR100
10	65.0 \pm 0.9	24.4 \pm 1.2
50	67.5 \pm 1.2	25.8 \pm 1.4
90	66.2 \pm 1.5	26.7 \pm 1.5

(b) Varying Removal Percentage with **Removal Heuristic = Top**.

Table 9. Best test set accuracy achieved during 30 rounds of training with Entropy selection under different settings with different private datasets (Distribution $\alpha = 1.0$) using single image patches as the distillation dataset with 20% FedAvg initialisation rate on ResNet-8 (across 2 seeds).

Local L.R.	Accuracy
0.1	74.6
0.05	80.1
0.01	80.9
0.005	80.6
0.001	74.6

Table 10. Highest test accuracy achieved during 30 rounds of training with FedAvg (no distillation) on CIF10 with ResNet-8 (Distribution $\alpha = 100.0$).

D.2. Server Side LR

To find the optimal learning rate for the distillation training, we use FedDF as depicted in Table 11. This is accomplished with the help of a grid search across certain values for the distillation training learning rate, as mentioned in Table 11 itself.

Global L.R.	Accuracy
0.1	44.8 \pm 1.4
0.05	62.7 \pm 1.0
0.01	76.6 \pm 0.3
0.005	78.1 \pm 0.4
0.001	75.2 \pm 1.0

Table 11. Highest test accuracy achieved during 30 rounds of training with client learning rate = 0.01 using FedDF on CIF10 with ResNet-8 on 2 different seeds. (Distribution $\alpha = 100.0$)