A SUPPLEMENTARY MATERIAL

A.1 MODELS IN THE EXPERIMENTS

We adopt the model WideResNet-40 (Zagoruyko & Komodakis, 2016) for CIFAR10, CIFAR100, and STL10, and AlexNet extended with BN layers (Li et al., 2021) for DomainNet. The classifier denoted in the paper has the same architecture as the last linear layer of the training model. The generator $G_{\mathbf{w}}$ is a MLP based model, which takes a noise ϵ and a one-hot label vector as input. The MLP sequentially consists of a linear layer with hidden dimension 512, a batch normalization layer with RELU as its activation function, and a representation linear layer with 512 as input dimension and 128 as output dimension to generate virtual OoD samples z.

A.2 EXTENDED ABLATION STUDY

Effects of δ for random soft label strategy. Given an one-hot encoding label vector y of class c, we assign $1-\delta$ to the c-th entry, and a random value within $(0, \delta)$ to the rest of the positions, where $\delta \in (0, 0.5)$. We investigate the effect of δ for the random soft label strategy on STL10. According to the results in Table 8, a mild value of δ shows the best results, for which we fix $\delta = 0.2$ in Section 5. When δ vanishes, the soft label degrades to the vanilla one-hot label which lacks essential OoD hardness. When we further increase δ, all information from the specific external classes will be eliminated due to the zero condition value and the sample will be generated from other regions randomly, which may increase the overlap with the ID data.

δ	Test acc ↑	AUROC ↑	AUPR ↑
0	**0.8410**	0.7671	0.9425
0.1	0.8258	0.7768	0.9475
0.2	0.8294	**0.7872**	**0.9501**
0.3	0.8396	0.7751	0.9450

Table 8: Effects of δ for random soft label strategy. ↑ indicates larger value is better. **Bold** numbers are superior results.

Effects of the number of samples generated by the generator. Table 9 shows the effect of the number of samples generated by the generator per iteration on CIFAR-10. As the number of generated samples increases, we obtain 2.45% AUROC increase and 0.53% AUPR increase, respectively. With more generated samples, not only can we obtain more sufficient samples to choose from, but we can also achieve more precise Gaussian distribution estimations. Thus, we fix the number of samples to be 1000 in Section 5.

Number of samples	Test acc ↑	AUROC ↑	AUPR ↑
100	0.9424	0.8846	0.9732
500	0.9302	0.8979	0.9761
1000	**0.9432**	**0.9091**	**0.9785**

Table 9: Effects of the number of samples generated by the generator. ↑ indicates larger value is better. **Bold** numbers are superior results.

p.d.f. filter can increase sample diversity. Table 10 shows the variance of the ID p.d.f of selected samples for three different clients on CIFAR10. According to the results, for all clients, the variance of w/ p.d.f. filter is much larger than that of w/o p.d.f. filter , thus, the diversity for the selected samples after p.d.f. filter is much larger than w/o p.d.f. filter .

Client	w/o p.d.f. filter	w/ p.d.f. filter
0	3.8117e+09	4.8624e+09
1	2.2709e+09	9.0618e+09
2	6.0931e+09	1.0391e+10

Table 10: The variance of the ID p.d.f of selected samples for three different clients.

Effects of different FL algorithms Our proposed FOSTER is a general framework that can be directly applied to similar FL algorithms as FedAvg, such as Fedprox (Li et al., 2020b). We report the results for FOSTER and other baselines in Table 11 on STL10, and our FOSTER outperforms competitive baselines for both FedAvg and its variants Fedprox.

FL algorithm	Method	Acc ↑	AUROC ↑	AUPR ↑
FedAvg	Energy	0.8236	0.7529	0.9228
	MSP	0.8236	0.7410	0.9309
	ODIN	0.8236	0.7418	0.9306
	VOS	0.8264	0.7370	0.9126
	FOSTER	**0.8410**	**0.7671**	**0.9425**
Fedprox	Energy	0.8292	0.7840	0.9296
	MSP	0.8292	0.7740	0.9369
	ODIN	0.8292	0.7749	0.9366
	VOS	0.8262	0.7306	0.9057
	FOSTER	**0.8379**	**0.7990**	**0.9503**

Table 11: Our FOSTER outperforms competitive baselines for both FedAvg and its variants Fedprox.

A.3 CONMMUNICATION COST FOR FOSTER

Communication cost includes per-round communication costs and the number of communication rounds.

Per-round communication costs: Due to the reason that sharing data would violate data confi-

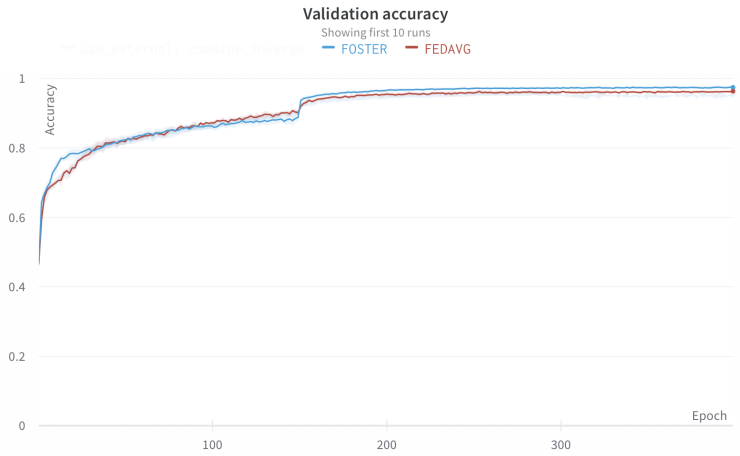

Figure 4: Validation accuracy for FedAvg and FOSTER. Generated OoD samples for FOSTER will not increase conmmunication rounds of FedAvg.

dentiality in FL, we choose to broadcast the central generator to allow each local client to generate their own virtual OoD samples. Thus, the communication cost is related to the model size of the generator. However, we note that a generator model usually has much smaller parameters than that of the main model that is learned, whereas the main model has to be transferred frequently between the server and clients in most FL paradigms. For example, the model size of the generator in this paper is only 5% of the FL global classification model. Thus, the communication cost largely depends on the model size of the global classification model and the cost of the generator is marginal. In other words, broadcasting generator only marginally increases the communication cost for each round compared with the standard FL setting.

The number of communication rounds: To investigate whether the generator will affect the communication rounds of FedAvg, we report the validation accuracy for STL10 in Fig. 4 in the appendix. We found that FOSTER will not increase the communication rounds compared with FedAvg, that is because we train the central generator without updating the global classifier by

optimizing Eq. (2). Thus, the global convergence is mainly determined by the main loss of the server instead of the loss of the generator.

A.4 ADDITIONAL BASELINES

We add one local training without FL method One-class SVM[1], and also two post hoc scoring methods KNN[2] (Sun et al., 2022) and ViM (Wang et al., 2022) to compare with the state-of-art in Table 12. One-class SVM can only achieve very low AUROC and AUPR, which is not comparable to other state-of-art methods. With limited local training set with partial classes for heterogeneous FL, KNN shows much lower AUROC and AUPR, and ViM also shows very low AUPR despite its comparable AUROC.

ID dataset	Method	Acc ↑	AUROC ↑	AUPR ↑
CIFAR-10	Energy	0.9431	0.7810	0.9262
	MSP	0.9431	0.8829	0.9691
	ODIN	0.9431	0.8842	0.9689
	KNN	0.9431	0.7349	0.6015
	ViM	0.9431	0.7938	0.6862
	VOS	0.9426	0.7970	0.9342
	One-class SVM	0.9431	0.7742	0.6481
	FOSTER	**0.9432**	**0.9091**	**0.9785**
CIFAR-100	Energy	0.8129	0.8056	0.9575
	MSP	0.8129	0.8606	0.9782
	ODIN	0.8129	0.8657	0.9789
	KNN	0.8129	0.3094	0.0605
	ViM	0.8129	0.7588	0.5493
	VOS	0.8063	0.8372	0.9666
	One-class SVM	0.8129	0.7688	0.5662
	FOSTER	**0.8218**	**0.8945**	**0.9838**
STL10	Energy	0.8236	0.7529	0.9228
	MSP	0.8236	0.7410	0.9309
	ODIN	0.8236	0.7418	0.9306
	KNN	0.8236	0.2586	0.0943
	ViM	0.8236	**0.7882**	0.5676
	VOS	0.8264	0.7370	0.9126
	One-class SVM	0.8236	0.7394	0.5417
	FOSTER	**0.8410**	0.7671	**0.9425**

Table 12: Our FOSTER outperforms competitive baselines. ↑ indicates larger value is better. **Bold** numbers are best performers.

A.5 CENTRAL EXPERIMENTS

To better verify our intuition that external classes can serve as effective OoD samples during training for OoD detection, we also conduct central experiments without FL training on CIFAR-10. We compare training with *External-class data* with VOS and energy score.

Effects of the number of ID classes. We fix the ID training data size to be 15000, and vary ID classes number to be 7, 5 and 3. According to the results shown in Table 13, with limited ID samples from each class, *External-class data* outperforms other baselines for OoD detection without hurting ID Acc. Although VOS shows better performance than Energy, it shows worse performance compared with the results reported in Du et al. (2022) where the entire training set is utilized, since VOS cannot get an accurate estimation of ID class-conditional Gaussian with limited samples for each class. When the number of ID classes drops from 7 to 3, for energy, AUROC and AUPR drop by 4.52% and 3.86% respectively, for VOS, AUROC and AUPR drop by 1.76% and 3.10% respectively, while *External-class data* improve AUROC and AUPR by 7.73% and 0.42% respectively. The decrease of ID classes number will make baselines produce worse OoD detection results, but it is not the case

[1]Since One-class SVM cannot produce ID accuracy, we use the accuracy from FedAvg for One-class SVM in the table.

[2]Following the original paper setting, we use $k = 50$ for CIFAR-10 and STL10, and $k = 200$ for CIFAR-100.

for *External-class data*. On the contrary, the decrease of ID classes will give *External-class data* a chance to get access to more diverse external class data, which can serve as real OoD samples for training. These results give an explicit explanation for why existing post-hoc and synthesized based OoD detection methods do not perform well under the FL setting, and verify our intuition that external classes can serve as effective OoD samples during training for OoD detection.

ID Classes	Method	Acc ↑	AUROC ↑	AUPR ↑
7	Energy	0.8455	0.7997	0.9642
	VOS	0.8530	0.8073	0.9668
	External-class data	**0.8605**	**0.8469**	**0.9751**
5	Energy	**0.9157**	0.8348	0.9539
	VOS	0.9147	0.8384	0.9536
	External-class data	0.9150	**0.8923**	**0.9750**
3	Energy	**0.9647**	0.7545	0.9256
	VOS	0.9627	0.7897	0.9358
	External-class data	0.9617	**0.9242**	**0.9793**

Table 13: Without hurting Acc, *External-class data* outperforms other baselines for OoD detection, especially when ID classes number is small. ↑ indicates larger value is better. **Bold** numbers are superior results.

Effects of the number of external classes. Based on the observation from Table 13, we also investigate how the number of external classes will affect the OoD performance of *External-class data*. We fix training data size to be 15000, ID classes to be 3, and vary the number of external classes from 7 to 2. The OoD performance for *External-class data* with different number of external classes is shown in Table 14. According to the results, diversity plays a key role for *External-class data* performance. The more external classes, the better OoD performance *External-class data* can achieve. Thus, for the proposed FOSTER we utilize all the external classes knowledge for training.

External Classes	Acc ↑	AUROC ↑	AUPR ↑
7	**0.9617**	**0.9242**	**0.9793**
5	0.9593	0.9062	0.9773
2	0.9377	0.8554	0.9610

Table 14: The more external classes, the better OoD performance exclass can achieve. ↑ indicates larger value is better. **Bold** numbers are superior results.

A.6 ADDITIONAL BENCHMARK

ID dataset	Method	Acc ↑	AUROC ↑	AUPR ↑
ImageNet-12	Energy	0.8552	0.7267	0.9174
	MSP	0.8552	0.7447	0.9286
	ODIN	0.8552	0.7399	0.9277
	VOS	0.8605	0.7207	0.9171
	FOSTER	0.8663	0.7526	0.9312

Table 15: Our FOSTER outperforms competitive baselines. ↑ indicates larger value is better. **Bold** numbers are best performers.