FedOSS: Federated Open Set Recognition via Inter-Client Discrepancy and Collaboration

Meilu Zhu, Jing Liao, *Member, IEEE*, Jun Liu, *Member, IEEE*, and Yixuan Yuan, *Member, IEEE*

***Abstract*— Open set recognition (OSR) aims to accurately classify known diseases and recognize unseen diseases as the unknown class in medical scenarios. However, in existing OSR approaches, gathering data from distributed sites to construct large-scale centralized training datasets usually leads to high privacy and security risk, which could be alleviated elegantly via the popular cross-site training paradigm, federated learning (FL). To this end, we represent the first effort to formulate federated open set recognition (FedOSR), and meanwhile propose a novel Federated Open Set Synthesis (FedOSS) framework to address the core challenge of FedOSR: the unavailability of unknown samples for all anticipated clients during the training phase. The proposed FedOSS framework mainly leverages two modules, i.e., Discrete Unknown Sample Synthesis (DUSS) and Federated Open Space Sampling (FOSS), to generate virtual unknown samples for learning decision boundaries between known and unknown classes. Specifically, DUSS exploits inter-client knowledge inconsistency to recognize known samples near decision boundaries and then pushes them beyond decision boundaries to synthesize discrete virtual unknown samples. FOSS unites these generated unknown samples from different clients to estimate the class-conditional distributions of open data space near decision boundaries and further samples open data, thereby improving the diversity of virtual unknown samples. Additionally, we conduct comprehensive ablation experiments to verify the effectiveness of DUSS and FOSS. FedOSS shows superior performance on public medical datasets in comparison with state-of-the-art approaches. The source code is available at https://github.com/CityU-AIM-Group/FedOSS.**

***Index Terms*— Open set recognition, federated learning, medical image classification.**

I. Introduction

With recent advancements in deep learning techniques, deep neural networks (DNNs)-based methods have achieved remarkable performance on various medical classification tasks [1], [2], [3]. However, these methods are usually evaluated in a closed-set setting, in which categories in test set are known and same as training set [4]. The closed-set setting is typically unreasonable in real-world medical scenarios, since rare or unknown diseases probably arise without any warning and would be misclassified into one of the known diseases [4], [5], [6], [7], [8], [9]. This problem poses an enormous risk to the public health and impedes the application of intelligent healthcare systems. Open-set recognition (OSR) [10], [11] is thus proposed to solve this issue, which aims to accurately classify known classes while at the same time identify open-set data as the unknown class.

Despite the impressing performance of existing OSR methods [4], [5], [6], [7], [8], [9] in open-set setting, they heavily depend on the availability of large-scale centralized datasets. Since the data in a single medical institution are usually limited, one common solution is to gather patient information from different hospitals [12]. Yet, due to growing privacy concerns or legal restrictions [13], distributed patient data are not able to be directly shared across institutions. Federated learning (FL) [14], [15] is a promising paradigm of decentralized machine learning to provide a feasible solution to this dilemma, which learns a global model by sharing the model parameters of clients (hospitals) instead of their raw data, under the orchestration of a trustworthy cloud server. Nevertheless, implementing OSR in such a decentralized FL framework is challenging and has not been investigated so far. Driven by above realistic issues, in this paper, we represent the first effort to formulate the problem of Federated Open Set Recognition (FedOSR). The purpose of this setting is to unite multiple distributed clients to learn a global model and reduce privacy as well as security risk, which can exactly classify known classes and recognize unknown classes in the testing stage, as illustrated in Fig. 1(a).

The core challenge for FedOSR to recognize unknown data is: **Unknown samples are not available for all anticipated clients during the training phase**. In FedOSR, client models will only focus on maximizing inter-known class distance and enhancing intra-known class compactness [5], failing to learn boundaries between known classes and unknown classes [16] due to lacking unknown data. After client model aggregation, the global model easily misclassifies an unknown sample into one given known class with a high confidence score since unknown classes might possess some similar features with known classes [17]. Therefore, it is very crucial to acquire unknown samples for FedOSR to learn the boundaries between known classes and unknown classes. An elegant idea is to synthesize virtual unknown data via generative adversarial

Manuscript received 21 May 2023; accepted 7 July 2023. Date of publication 10 July 2023; date of current version 2 January 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62001410; in part by the Hong Kong Research Grants Council under Grant 11212321, Grant 11217922, Grant ECS-21207420, Grant ECS-21212720, and Grant CRF-C4063-18G; in part by Hong Kong Special Administrative Region (HKSAR) Innovation and Technology Commission (ITC) under Innovation and Technology Fund (ITF) Project under Grant MHP/109/19; and in part by the Science, Technology and Innovation Committee of Shenzhen under Grant SGDX20210823104001011. *(Corresponding authors: Yixuan Yuan; Jun Liu.)*

Meilu Zhu and Jun Liu are with the Department of Mechanical Engineering, City University of Hong Kong, Hong Kong, SAR, China (e-mail: meiluzhu2-c@my.cityu.edu.hk; jun.liu@cityu.edu.hk).

Jing Liao is with the Department of Computer Science, City University of Hong Kong, Hong Kong (e-mail: jingliao@cityu.edu.hk).

Yixuan Yuan is with the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, SAR, China (e-mail: yxyuan@ee.cuhk.edu.hk).

Digital Object Identifier 10.1109/TMI.2023.3294014

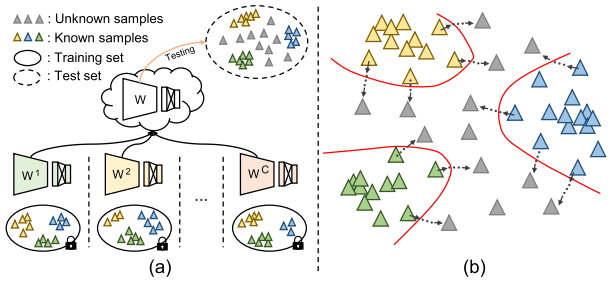


Fig. 1. (a) Federated open set recognition. This setting aims to unite multiple clients trained on known classes to learn a global model, which can accurately classify known classes and recognize unknown samples in test set. (b) Our solution to synthesize virtual unknown samples by transforming known samples near decision boundaries.

networks (GAN) [16], [17], [18], [19], [20], [21], [22], [23] or Mixup on known samples [7], [24], [25]. However, GAN not only undergoes convergence difficulty [26], [27] in the FL setting but also fails to exhaustively span the infinite open world [5], [9]. Meanwhile, Mixup [28], a data augmentation technique designed for centralized learning, would incur data privacy leakage and high bandwidth cost in the distributed setting. Hence, these methods [7], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25] are not applicable to the FedOSR scheme.

Differing from the above approaches [7], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], we conquer this intractable challenge from two new perspectives. (1) We are dedicated to generating unseen data near decision boundaries of known classes instead of fitting entire unknown space. As we all know, open-set data are from an infinite space and probably belong to multiple unknown classes. It is impossible to synthesize samples of all unknown classes for learning boundaries between known and unknown classes since the training set does not provide any prior information about unknown classes. By contrast, unknown samples near boundaries of known classes could provide rich information about the boundary of open data [29], [30]. (2) Considering the infeasibility of leveraging GAN and Mixup to generate open-set data in the FedOSR setting, we explore the recognition of boundary samples of known classes and then exploit these samples to synthesize open-set data. Intuitively, known and unknown samples lying near class boundaries have small geometric distance in latent feature space. They usually share some common patterns and thus have high similarities. This geometric view indicates that known boundary samples can be converted to unknown boundary data through appropriate transformation [31], [32], as shown in Fig. 1(b).

In this paper, we propose a novel framework, Federated Open Set Synthesis (FedOSS), to achieve open set recognition for medical classification tasks in the federated learning setting, which mainly contains two modules, i.e., Discrete Unknown Sample Synthesis (DUSS) and Federated Open Space Sampling (FOSS). Specifically, DUSS equips each client model with a personalized classifier to represent the personalization of these clients trained on Non-IID data, which is collected and distributed to all clients by the server. Based on the prediction inconsistency of these personalized classifiers, we can recognize the boundary samples of known classes in closed set. We push these known boundary samples beyond decision boundaries to generate discrete unknown samples via inversion updating. To further improve the diversity of virtual unknown samples, FOSS estimates local class-conditional distributions of the generated discrete unknown samples at each client and aggregates them into global distributions to span continuous open space near decision boundaries. Each client downloads the global distributions from the server and samples diverse unknown samples from the low-likelihood region of these distributions. With synthesized unknown samples from DUSS and FOSS modules, FedOSS can learn boundaries between known classes and unknown classes. The main contributions of our work are summarized as follows:

- We propose a novel FedOSS framework to tackle a new and realistic problem of Federated Open Set Recognition for medical classification tasks. To the best of our knowledge, this work represents the first effort to unite different institutions to achieve open set recognition in medical scenarios.
- Discrete Unknown Sample Synthesis (DUSS) is devised to recognize known samples near decision boundaries via inter-client inconsistency in knowledge and exploits these samples to generate discrete open data.
- Federated Open Space Sampling (FOSS) is designed to unite the generated unknown samples of all clients to fit global distributions of open space near decision boundaries and samples open data to further improve the diversity of virtual unknown samples.
- We conduct extensive experiments on public datasets to evaluate the proposed framework. The results demonstrate the superior performance of FedOSS against state-of-the-arts and the effectiveness of different modules.

Roadmap: The rest of the paper is organized as follows. We review previous OSR and FL methods in Section II. In Section III, the proposed FedOSS is introduced in detail. We describe the implementation details and verify the effectiveness of the proposed FedOSS in Section IV. Finally, the paper is closed with the conclusion in Section V.

II. Related Work

A. Open Set Recognition

To deploy the classification models to real scenarios with the high robustness, OSR is first proposed to improve the capability of detecting unknowns of models in [10]. Earlier works were mainly based on traditional machine learning methods, such as support vector machines (SVM) in [11], Extreme Value Theory (EVT) in [33], the nearest neighbor in [34], and so on. With the development of deep learning, deep neural networks based OSR approaches have received significant attention. As the earliest representative work, OpenMax [4] replaces the softmax layer in the network and exploits Weibull distribution to calibrate the output probability. Considering that feature representation is not discriminative enough in OpenMax, CROSR [35] further introduces an extra reconstruction task to improve the learning of network. Similar to

OpenMax, OVRN [9] exploits one-vs-rest units, i.e., sigmoid, to replace the softmax layer. In addition, some prototype based methods [5], [6] learn prototypes to represent known classes and identify open-set samples based on distances to the prototypes. These methods usually undergo the challenge of tuning an optimal threshold to separate known and unknown classes.

A recently popular solution for OSR is using the information of known classes to simulate unknown examples. PROSER [7] exploits the manifold mixup on the hidden representations of different known classes to mimic novel patterns. SN [24] regards the mixture of samples from the same class as known classes and the mixture of samples from the different classes as unknown classes to train a segregation network for novel class detection. Oza and Patel [25] found that activation maps of non-target classes present similar patterns to that of novel classes. Thus, they mimicked the novel data by combining features of known samples with activation maps of non-target classes.

Other approaches mostly fall into the types of employing generative models to generate unknown samples. G-OpenMax [22] improves OpenMax by adopting a conditional generative network to synthesize unknown instances. OSRCI [16], [18] develops encoder-decoder GAN architecture to generate counterfactual images as open data. OpenGAN [19] uses outlier data to select the appropriate GAN-discriminator for generating better fake open examples. To simulate the real-world open space, DIAS [17] generates virtual unknown samples with diverse difficulty levels via GAN. For this type of approach, it is difficult to apply them to the FL setting due to the training complexity of generative models.

B. Federated Learning in Medical Image Classification

Federated learning (FL) aims to coordinate multiple clients to learn a global model by transmitting local model parameters instead of raw patient data [14]. Currently, FL has been widely applied to a lot of medical image classification tasks. Existing approaches [13], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47] mainly focus on two intractable challenges: data heterogeneity and annotation scarcity.

Approaches [13], [36], [37], [38], [39], [40], [41], [42] concentrated on the data heterogeneity problem can be further divided into two classes. The first type of methods [36], [37], [38] try to improve the local training at the client side when the local data of clients are heterogeneous. For example, HarmoFL [36] collects amplitude information of images of clients and generates a global amplitude, which is used to normalize the frequency-space amplitude components of local data into a unified space. Zhu and Luo [37] exploited the global model and adversarial training to generate virtual data similar to samples of other clients to eliminate the discrepancy between clients. Another type of approaches [13], [39], [40], [41], [42] are devoted to improving server aggregation of local models trained on heterogeneous data. For instance, IDA [39] uses the inverse distance of local models to the average model as weights to aggregate these client models and thereby weighting less the out-of-distribution models. Considering that data heterogeneity causes the mismatching problem of parameters of local models, Chen et al. [13], [40] aggregated the low-frequency components of client parameters while preserving the remaining high-frequency components to achieve personalized federated learning. FedBN [41] handles the non-iid issue by keeping client BN layers updated locally and only fusing non-BN layers at the server.

Previous works [43], [44], [45], [46], [47] focusing on annotation scarcity problem introduce the semi-supervised setting into federated learning systems to train local models via labeled and unlabeled data. For instance, Yang et al. [43] united multi-national data with or without annotations from China, Italy and Japan to learn a global model for the detection of COVID-19. FedIRM [44] exploits the inter-class correlation matrix estimated at labeled clients to supervise the learning of local models at unlabeled clients. Besides, FedPerl [45] finds multiple similar peers for each client to help it generate pseudo labels for the unlabeled data. However, existing FL methods on medical image classification [13], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47] only consider the closed-set performance and ignore the possible occurrence of novel classes in realistic scenes. In this work, we represent the first effort to introduce open set recognition into FL to address this practical problem.

III. Federated Open Set Synthesis

A. Problem Formulation and Overview

1) Problem Formulation: Due to growing privacy concerns, in this paper, we develop the standard OSR to Federated Open Set Recognition (FedOSR) to perceive unknown diseases and reduce privacy as well as security risk. FedOSR follows the standard FL setting, FedAvg [14], to unite C distributed clients to learn a global model f with the parameters $\mathbf{W}$ under the orchestration of the server. Each client has a local cohort $\mathcal{D}_{tr}^{c} = \{(\mathbf{x}_i^c, \mathbf{y}_i^c)\}$ with K classes of known diseases, where $\mathbf{x}_i^c$ is a training instance with the label $\mathbf{y}_i^c = \{1, \ldots, K\}$. The client model f^c with the parameters $\mathbf{W}^c$ can only access its local dataset $\mathcal{D}_{tr}^{c}$ and is not allowed to share it with other clients, where $\mathbf{W}^c$ contains the parameters $\mathbf{W}_{fe}^c$ of the feature extractor and the parameters $\mathbf{W}_{mc}^c$ of the main classifier. After training, the global model is evaluated with a real-world test set $\mathcal{D}_{te} = \{(\mathbf{x}_i, \mathbf{y}_i)\}$, where $\mathbf{y}_i = \{1, \ldots, K, K+1\}$, and the class $K+1$ is a group of unknown diseases and may contain more than one class. FedOSR aims to accurately classify K known diseases and recognize unknown diseases as the class $K+1$. However, since local datasets $\{\mathcal{D}_{tr}^{c}\}_{c=1}^{C}$ do not provide patient samples and prior information of unknown diseases, local models fail to learn a boundary between known classes and the class $K+1$. As a result, the global model obtained by client model aggregation would always misclassify unknown samples in $\mathcal{D}_{te}$ into known classes [4], [7].

2) Overview of FedOSS: To solve this problem, we propose a Federated Open Set Synthesis (FedOSS) framework to synthesize virtual open data for learning a boundary between known classes and the class $K+1$, as illustrated in Fig. 2. The training of FedOSS proceeds through multiple rounds of communication between clients and the server. Specifically,

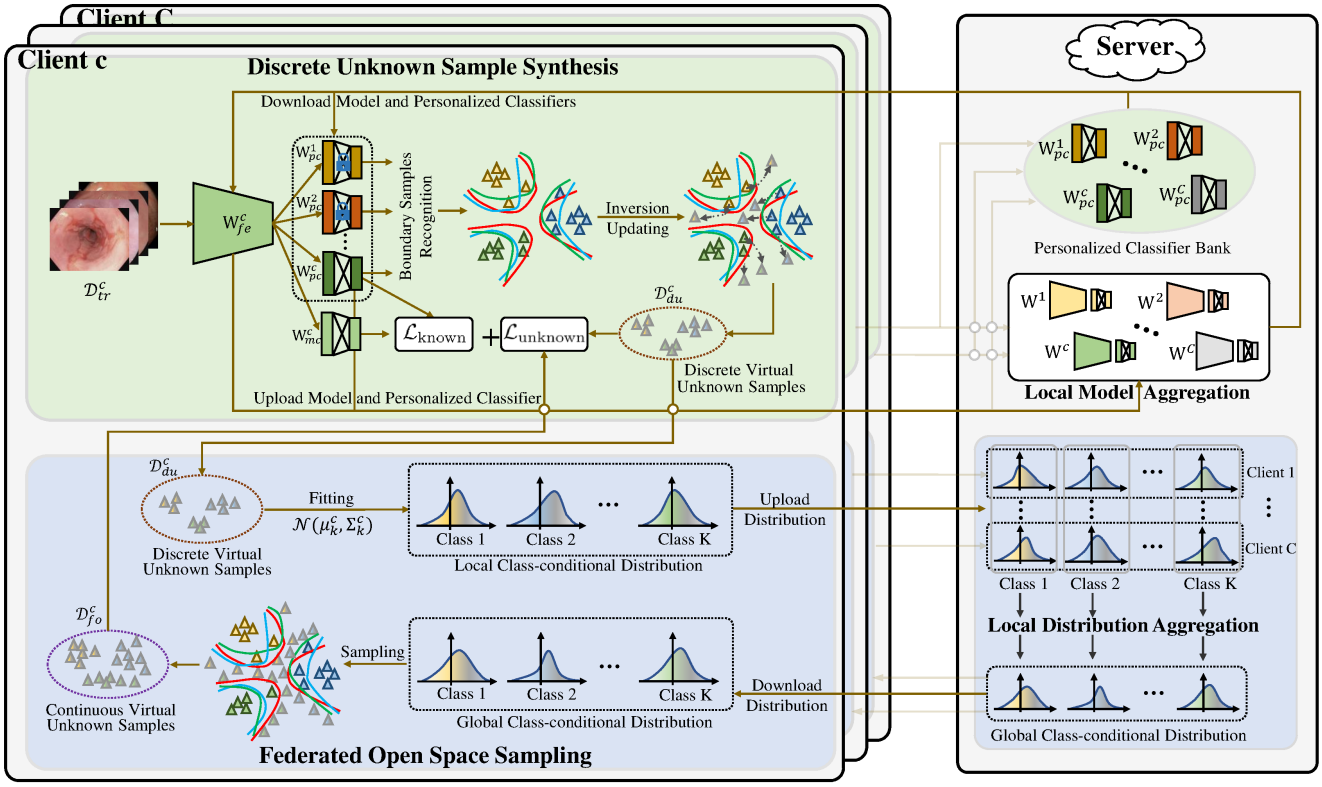


Fig. 2. The overview of the proposed FedOSS framework to federated open set recognition (Best view in color). FedOSS overcomes the challenge in unavailability of open data via discrete unknown sample synthesis (DUSS) and federated open space sampling (FOSS). DUSS utilizes the client discrepancy to recognize boundary samples, which are used to synthesize unknown samples via inversion updating. The generated unknown samples from all clients are inputted into FOSS to estimate the distribution of open space for sampling diverse unknown data.

in each round, all clients download the parameters $\mathbf{W}$ of the global model f as the initialization of local models $\{f^c\}_{c=1}^C$. Then, the c-th client performs local training on its local data $\mathcal{D}_{tr}^c$ to update the initial parameters $\mathbf{W}$. During local training, the discrete unknown sample synthesis (DUSS) module first recognizes known samples near decision boundaries, and then exploits them to synthesize discrete virtual unknown samples. Next, these generated unknown samples are input into the federated open space sampling (FOSS) module to estimate the distributions of open data space near decision boundaries. We further sample more open data from these distributions. These synthesized open data of DUSS and FOSS are used to help client models to learn decision boundaries for separating known and unknown classes. In the end of each round, the server gathers these updated local models and aggregate them to update the global model f for the next iteration.

B. Discrete Unknown Sample Synthesis

The unavailability of unknown samples is the core barrier to learn boundaries to separate known and unknown classes in FedOSR. Although previous studies [7], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [48] on the standard OSR have explored the feasibility of leveraging GAN [49] and Mixup [28] to synthesize virtual unknown samples, they possibly undergo the convergence difficulty and incur the high risk of privacy leakage in the decentralized setting. To beak this dilemma, we propose a discrete unknown sample synthesis (DUSS) module to generate virtual discrete unknown data using known samples for FedOSR. As shown in Fig. 2, DUSS first leverages the knowledge discrepancy of clients to recognize known samples near decision boundaries. Then, these boundary samples are pushed beyond decision boundaries via inversion updating, and thereby transformed into discrete virtual unknown samples.

1) Boundary Sample Recognition via Client Discrepancy: A boundary sample is more likely to be misclassified due to its small geometric distance to the decision boundary, compared with an instance located at the cluster center. Such an intrinsic characteristic inspires us to leverage the knowledge discrepancy of client models trained on Non-IID local data to recognize known samples near decision boundaries. Specifically, besides the main classifier ($\mathbf{W}_{mc}^c$), we equip each client model f^c with an extra personalized classifier of K classes with parameters $\mathbf{W}_{pc}^c$, following the feature extractor ($\mathbf{W}_{fe}^c$). Here, the personalized classifier only focuses on K known classes while the main classifier not only predicts K known classes but also recognizes unknown samples as the class $K+1$. In each round of communication, each client uploads three parts of parameters, i.e., $\mathbf{W}_{fe}^c$, $\mathbf{W}_{mc}^c$, and $\mathbf{W}_{pc}^c$, to the server. The server aggregates the parameters $\{\mathbf{W}_{fe}^c\}_{c=1}^C$ and $\{\mathbf{W}_{mc}^c\}_{c=1}^C$ to update the global model f. We store the parameters $\{\mathbf{W}_{pc}^c\}_{c=1}^C$ of the personalized classifiers in a bank $\mathbf{\Omega}$. Then, the global model f and the bank $\mathbf{\Omega}$ are sent back to each client. Noticeably, the personalized classifiers are usually

lightweight, and thus transmitting the bank $\mathbf{\Omega}$ does not incur the high communication overhead.

Given the personalized classifier bank $\mathbf{\Omega} = \{\mathbf{W}_{pc}^c\}_{c=1}^C$, we define a score $\boldsymbol{s}$ for a known sample $(\mathbf{x}_i^c, \mathbf{y}_i^c)$ at the c-th client, which indicates the *geometric distance* from the sample $\mathbf{x}_i^c$ to the cluster boundary of class $\mathbf{y}_i^c$:

$$\boldsymbol{s} = \frac{1}{C} \sum_{\mathbf{W}_{pc}^c \in \mathbf{\Omega}} \mathbf{I}(\mathbf{y}_i^c, f^c(\mathbf{x}_i^c, \mathbf{W}_{fe}^c, \mathbf{W}_{pc}^c)), \mathbf{\Omega} = \{\mathbf{W}_{pc}^c\}_{c=1}^C, \quad (1)$$

where $\mathbf{y}_i^c$ is the label of the i-th sample of the c-th client. $f^c(\mathbf{x}_i^c, \mathbf{W}_{fe}^c, \mathbf{W}_{pc}^c)$ is the prediction of the sample $\mathbf{x}_i^c$ for the c-th personalized classifier $\mathbf{W}_{pc}^c$ in the bank $\mathbf{\Omega}$. $\mathbf{I}(\cdot)$ is a function to check whether $\mathbf{y}_i^c$ and $f^c(\mathbf{x}_i^c, \mathbf{W}_{fe}^c, \mathbf{W}_{pc}^c)$ are equal. In Eq. (1), personalized classifiers of clients are trained on Non-IID data, and thus their corresponding class boundaries are significantly different. Samples with different distances to class center regions will be misclassified by the different numbers of personalized classifiers. Therefore, we can divide known samples of the c-th client into three types based on the score $\boldsymbol{s}$. (1) If the score $\boldsymbol{s}$ of a sample $\mathbf{x}_i^c$ is equal to 1, the sample is classified accurately by the personalized classifiers of all clients. This means that it is pretty close to the center of class $\mathbf{y}_i^c$. (2) On the contrary, this sample is considered as a hard one and outside the cluster of class $\mathbf{y}_i^c$ when its score $\boldsymbol{s}$ is equal to 0. (3) If the score $\boldsymbol{s}$ is between 0 and 1, the sample $\mathbf{x}_i^c$ would be located in decision boundary areas of class $\mathbf{y}_i^c$. Meanwhile, the smaller score $\boldsymbol{s}$ indicates that the sample $\mathbf{x}_i^c$ is farther away from the center of class $\mathbf{y}_i^c$. Hence, relying on the inter-client knowledge discrepancy, we can obtain a boundary sample set $\mathcal{D}_{bs}^c = \{x_i^c \mid 0 < \boldsymbol{s}(x_i^c) < 1\}$ at the client c.

We use personalized classifiers of all clients to compute the geometric distance, instead of main classifiers. Personalized classifiers are uploaded to the server but are not aggregated, and thus can maintain the discrepancy for recognizing boundary samples. In contrast, main classifiers of all clients interact with each other via server aggregation, leading to the high similarity. If using the main classifiers to compute Eq. (1), we will fail to recognize boundary samples since the score $\boldsymbol{s}$ of one sample would always be 0 or 1.

2) Unknown Sample Synthesis via Inversion Updating: Known boundary samples in $\mathcal{D}_{bs}^c$, serving as the intermediary between known and unknown classes, have high similarity to open data near the decision boundary. Therefore, we propose transforming these boundary samples into open data via inversion updating. Given a fixed client model $f_\star^c$, the goal of inversion updating is to iteratively optimize boundary samples $\mathbf{x}_i^c$ and push them beyond the decision boundary. This can be implemented by updating $\mathbf{x}_i^c$ to maximize the following empirical classification risk over known classes,

$$\min_{\mathbf{x}_i^c} \; -\mathcal{L}(f_\star^c(\mathbf{W}_{fe}^c, \mathbf{W}_{mc}^c, \mathbf{x}_i^c), \mathbf{y}_i^c); \quad (\mathbf{x}_i^c, \mathbf{y}_i^c) \sim \mathcal{D}_{bs}^c, \quad (2)$$

where $\mathcal{L}(\cdot)$ is the standard cross-entropy loss. Eq.(2) provides a fastest updating direction to push $\mathbf{x}_i^c$ across the decision boundaries, i.e., reversed gradient descent. Similar to adversarial attack [50], we apply the inversion updating to $\mathbf{x}_i^c$ multiple times, and the update process of $\mathbf{x}_i^c$ is formulated:

$$\begin{aligned} \overline{\mathbf{x}}_i^c(0) &= \mathbf{x}_i^c, \\ \overline{\mathbf{x}}_i^c(T+1) &= \overline{\mathbf{x}}_i^c(T) + \lambda \nabla_{\overline{\mathbf{x}}_i^c} \mathcal{L}(f_\star^c(\mathbf{W}_{fe}^c, \mathbf{W}_{mc}^c, \overline{\mathbf{x}}_i^c(T)), \mathbf{y}_i^c), \end{aligned} \quad (3)$$

where T is the times of updating and λ is the step size. With inversion updating, we can transform boundary samples in $\mathcal{D}_{bs}^c$ to obtain a discrete unknown samples set $\mathcal{D}_{du}^c = \{\overline{\mathbf{x}}_i^c\}_{i=1}^{|\mathcal{D}_{du}^c|}$. To reduce the computation cost caused by multiple times of back-propagation in Eq (3), we perform inversion updating in feature space instead of image space.

Compared with previous methods [16], [17], [18], [19], [20], [21], [22], [23], [48] that exploit GAN to generate virtual unknown samples, the proposed DUSS module does not require a large amount of extra training time to learn the generator. Besides, in contrast to Mixup based approaches [7], [24], [25], synthesized unknown samples in $\mathcal{D}_{du}$ of DUSS lie near decision boundary areas of known classes and thus are able to effectively enhance the compactness of known classes, leaving more space for unknown classes. Furthermore, DUSS emphasizes unknown boundary data near known classes, which facilitates boundary learning, and separate known and unknown classes.

C. Federated Open Space Sampling

Learning reliable boundaries between known and unknown classes tends to rely highly on the diversity of open set [17], [51]. Additionally, the unreliable boundaries of one client may influence other clients via model aggregation and even harm the performance of the overall federation [52] in the FedOSR setting. To improve the diversity of virtual unknown samples of clients, we propose a federated open space sampling (FOSS) module to integrate the knowledge on the open space of all clients to further generate diverse unknown data. As shown in Fig. 2, each client uploads local statistics of synthesized unknown samples in $\mathcal{D}_{du}^c$ to the server. These statistics are aggregated to estimate global distributions for open space near the decision boundaries. With the global distributions, each client can sample diverse high-quality unknown samples to improve the local training.

It is reported that the features learned by deep neural networks can be theoretically approximated with a mixture of Gaussian distribution [53], [54]. Therefore, the virtual unknown samples of all clients near the decision boundary of the known class k in $\{\mathcal{D}_{du}^c\}_{c=1}^C$ follow a global class-conditional multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)$, where $\boldsymbol{\mu}_k$ is the mean of these unknown samples and $\mathbf{\Sigma}_k$ is the covariance matrix. To estimate the global $\boldsymbol{\mu}_k$ and $\mathbf{\Sigma}_k$, we first compute local means $\{\boldsymbol{\mu}_k^c\}_{c=1}^C$ and covariance matrixes $\{\mathbf{\Sigma}_k^c\}_{c=1}^C$ of all clients. Let $\{\overline{\mathbf{x}}_{k,i}^c\}_{i=0}^{\bar{N}_k^c}$ be the generated unknown samples near the k-th known class at the client c, where $\bar{N}_k^c$ is the number of samples, the local mean $\boldsymbol{\mu}_k^c$ and covariance matrix $\mathbf{\Sigma}_k^c$ can be computed as follows:

$$\boldsymbol{\mu}_k^c = \frac{1}{\bar{N}_k^c} \sum_{i=1}^{\bar{N}_k^c} \overline{\mathbf{x}}_{k,i}^c, \quad \mathbf{\Sigma}_k^c = \frac{1}{\bar{N}_k^c - 1} \sum_{i=1}^{\bar{N}_k^c} (\overline{\mathbf{x}}_{k,i}^c - \boldsymbol{\mu}_k^c)(\overline{\mathbf{x}}_{k,i}^c - \boldsymbol{\mu}_k^c)^{\mathrm{T}}. \quad (4)$$

Then, the local means $\{\boldsymbol{\mu}_k^c\}_{c=1}^C$ and covariance matrixes $\{\boldsymbol{\Sigma}_k^c\}_{c=1}^C$ of all clients are transmitted to the server. With these local statistics, the server estimates the global mean $\boldsymbol{\mu}_k$ and the global covariance matrix $\boldsymbol{\Sigma}_k$ straightforwardly according to the following theorem:

Theorem 1: With the local means $\{\boldsymbol{\mu}_k^c\}_{c=1}^C$ and covariance matrixes $\{\boldsymbol{\Sigma}_k^c\}_{c=1}^C$ of virtual unknown samples from C clients near the known class k, the global mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$ can be calculated:

$$\boldsymbol{\mu}_k = \sum_{c=1}^{C} \frac{\bar{N}_k^c}{\bar{N}_k} \boldsymbol{\mu}_k^c,$$

$$\boldsymbol{\Sigma}_k = \sum_{c=1}^{C} \frac{\bar{N}_k^c - 1}{\bar{N}_k - 1} \boldsymbol{\Sigma}_k^c + \sum_{c=1}^{C} \frac{\bar{N}_k^c}{\bar{N}_k - 1} \boldsymbol{\mu}_k^c \boldsymbol{\mu}_k^{cT} - \frac{\bar{N}_k}{\bar{N}_k - 1} \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T,$$

where $\bar{N}_k$ is the total number of unknown samples near the known class k, $\bar{N}_k = \sum_{c=1}^C \bar{N}_k^c$.

After computing the global means $\{\boldsymbol{\mu}_k\}_{k=1}^K$ and covariance matrixes $\{\boldsymbol{\Sigma}_k\}_{k=1}^K$, the server constructs a global bank $\boldsymbol{\Phi} = \{\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^K$ and distributes it to all clients. With the global Gaussian distributions $\{\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^K$, the client c can randomly sample virtual unknown instances from the ϵ-likelihood region of each distribution $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, yielding a new unknown sample set $\mathcal{D}_{fo}^c = \{\mathcal{D}_{fo}^{c,k}\}_{k=1}^K$:

$$\mathcal{D}_{fo}^{c,k} = \{\widehat{\mathbf{x}}_{k,i}^c \mid p(\widehat{\mathbf{x}}_{k,i}^c) < \epsilon\}_{i=1}^{B_k},$$
$$p(\widehat{\mathbf{x}}_{k,i}^c) = \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\widehat{\mathbf{x}}_{k,i}^c - \boldsymbol{\mu}_k)^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} (\widehat{\mathbf{x}}_{k,i}^c - \boldsymbol{\mu}_k)\right), \tag{5}$$

where $\mathcal{D}_{fo}^{c,k}$ denotes the unknown sample subset near the class k at the c-th client. D is the dimension of the instance $\widehat{\mathbf{x}}_{k,i}^c$ and B_k is the number of sampled virtual instances, $|\mathcal{D}_{fo}^c| = \sum_{k=1}^K B_k$. To obtain high-quality unknown samples $\widehat{\mathbf{x}}_{k,i}^c$ in experiments, we employ a small ϵ to guarantee the certain distance between $\widehat{\mathbf{x}}_{k,i}^c$ and decision boundaries. The global distributions $\{\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^K$ in the bank $\boldsymbol{\Phi}$ contain global statistics and can represent the knowledge on open space near the decision boundaries from all clients. By sampling unknown samples from the estimated continuous space, each client can have enough high-quality unknown samples to learn reliable boundaries between unknown and known classes.

Proof of Theorem 1. We first demonstrate the derivation of global mean $\boldsymbol{\mu}_k$, which is estimated by aggregating $\{\boldsymbol{\mu}_k^c\}_{c=1}^C$:

$$\boldsymbol{\mu}_k = \frac{1}{\bar{N}_k} \sum_{c=1}^{C} \sum_{i=1}^{\bar{N}_k^c} \overline{\mathbf{x}}_{k,i}^c = \sum_{c=1}^{C} \frac{\bar{N}_k^c}{\bar{N}_k} \cdot \frac{1}{\bar{N}_k^c} \sum_{i=1}^{\bar{N}_k^c} \overline{\mathbf{x}}_{k,i}^c = \sum_{c=1}^{C} \frac{\bar{N}_k^c}{\bar{N}_k} \boldsymbol{\mu}_k^c. \tag{6}$$

Next, we calculate the global covariance matrix $\boldsymbol{\Sigma}_k$:

$$\begin{aligned}\boldsymbol{\Sigma}_k &= \frac{1}{\bar{N}_k - 1} \sum_{c=1}^{C} \sum_{i=1}^{\bar{N}_k^c} (\overline{\mathbf{x}}_{k,i}^c - \boldsymbol{\mu}_k)(\overline{\mathbf{x}}_{k,i}^c - \boldsymbol{\mu}_k)^{\mathrm{T}} \\ &= \frac{1}{\bar{N}_k - 1} \sum_{c=1}^{C} \sum_{i=1}^{\bar{N}_k^c} \overline{\mathbf{x}}_{k,i}^c (\overline{\mathbf{x}}_{k,i}^c)^{\mathrm{T}} - \frac{1}{\bar{N}_k - 1} \sum_{c=1}^{C} \sum_{i=1}^{\bar{N}_k^c} \overline{\mathbf{x}}_{k,i}^c \boldsymbol{\mu}_k^{\mathrm{T}} \\ &\quad - \frac{1}{\bar{N}_k - 1} \sum_{c=1}^{C} \sum_{i=1}^{\bar{N}_k^c} \boldsymbol{\mu}_k (\overline{\mathbf{x}}_{k,i}^c)^{\mathrm{T}} + \frac{1}{\bar{N}_k - 1} \sum_{c=1}^{C} \sum_{i=1}^{\bar{N}_k^c} \boldsymbol{\mu}_k \boldsymbol{\mu}_k^{\mathrm{T}}.\end{aligned} \tag{7}$$

By applying the identity $\boldsymbol{\mu}_k^{\mathrm{T}} = \frac{1}{\bar{N}_k} \sum_{c=1}^C \sum_{i=1}^{\bar{N}_k^c} (\overline{\mathbf{x}}_{k,i}^c)^{\mathrm{T}}$ and Eq. (6), both the second and third terms in Eq. (7) can be rewritten as $\frac{\bar{N}_k}{\bar{N}_k - 1} \boldsymbol{\mu}_k \boldsymbol{\mu}_k^{\mathrm{T}}$. With the identity $\sum_{c=1}^C \sum_{i=1}^{\bar{N}_k^c} \boldsymbol{\mu}_k = \bar{N}_k \boldsymbol{\mu}_k$, the last term can be also simplified as $\frac{\bar{N}_k}{\bar{N}_k - 1} \boldsymbol{\mu}_k \boldsymbol{\mu}_k^{\mathrm{T}}$. Therefore, $\boldsymbol{\Sigma}_k$ in Eq. (7) can be rewritten as follows:

$$\boldsymbol{\Sigma}_k = \frac{1}{\bar{N}_k - 1} \sum_{c=1}^{C} \sum_{i=1}^{\bar{N}_k^c} \overline{\mathbf{x}}_{k,i}^c (\overline{\mathbf{x}}_{k,i}^c)^{\mathrm{T}} - \frac{\bar{N}_k}{\bar{N}_k - 1} \boldsymbol{\mu}_k \boldsymbol{\mu}_k^{\mathrm{T}}. \tag{8}$$

Similar to the global covariance matrix $\boldsymbol{\Sigma}_k$, we also rewrite the local covariance matrix $\boldsymbol{\Sigma}_k^c$ in Eq. (4) as

$$\boldsymbol{\Sigma}_k^c = \frac{1}{\bar{N}_k^c - 1} \sum_{i=1}^{\bar{N}_k^c} \overline{\mathbf{x}}_{k,i}^c (\overline{\mathbf{x}}_{k,i}^c)^{\mathrm{T}} - \frac{\bar{N}_k^c}{\bar{N}_k^c - 1} \boldsymbol{\mu}_k^c (\boldsymbol{\mu}_k^c)^{\mathrm{T}}. \tag{9}$$

By rearranging Eq. (9), we can further acquire the identity $\sum_{i=1}^{\bar{N}_k^c} \overline{\mathbf{x}}_{k,i}^c (\overline{\mathbf{x}}_{k,i}^c)^{\mathrm{T}} = (\bar{N}_k^c - 1)\boldsymbol{\Sigma}_k^c + \bar{N}_k^c \boldsymbol{\mu}_k^c (\boldsymbol{\mu}_k^c)^{\mathrm{T}}$, and then apply it into Eq. (8) to rewrite $\boldsymbol{\Sigma}_k$:

$$\boldsymbol{\Sigma}_k = \sum_{c=1}^{C} \frac{\bar{N}_k^c - 1}{\bar{N}_k - 1} \boldsymbol{\Sigma}_k^c + \sum_{c=1}^{C} \frac{\bar{N}_k^c}{\bar{N}_k - 1} \boldsymbol{\mu}_k^c \boldsymbol{\mu}_k^{c\mathrm{T}} - \frac{\bar{N}_k}{\bar{N}_k - 1} \boldsymbol{\mu}_k \boldsymbol{\mu}_k^{\mathrm{T}}. \tag{10}$$

From Theorem 1, we can observe that the global mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$ are only related to local means $\{\boldsymbol{\mu}_k^c\}_{c=1}^C$ and covariance matrixes $\{\boldsymbol{\Sigma}_k^c\}_{c=1}^C$, and do not depend on the original generated unknown samples at the client side. Hence, FOSS does not incur the heavy communication overhead and privacy leakage risk.

D. Overall Training

The proposed FedOSS framework aims to unite the private data $\{\mathcal{D}_{tr}^c\}_{c=1}^C$ of C clients to reduce known and unknown space risk and learn a shared global model. The overall training procedure of FedOSS includes pre-training and fine-tuning stages, as shown in Algorithm III-C. We first pre-train the feature extractor, the main and personalized classifiers via the standard cross-entropy loss on known classes. Then, FedOSS is fine-tuned by minimizing the following hybrid loss function $\mathcal{L}_{total}$:

$$\mathcal{L}_{\text{total}} = \frac{1}{C} \sum_{c=1}^{C} \mathcal{L}_{\text{known}}^c + \mathcal{L}_{\text{unknown}}^c + \mathcal{L}_{\text{extra}}^c, \tag{11}$$

where $\mathcal{L}_{\text{known}}^c$ reduces the empirical classification risk on the local data $\mathcal{D}_{tr}^c$ to guarantee the performance of known classes, which is defined as follows:

$$\mathcal{L}_{\text{known}}^c = \mathbb{E}_{(\mathbf{x}_i^c, \mathbf{y}_i^c) \sim \mathcal{D}_{tr}^c} \mathcal{L}_{\text{CE}}(f^c(\mathbf{W}_{fe}^c, \mathbf{W}_{mc}^c, \mathbf{x}_i^c), \mathbf{y}_i^c), \tag{12}$$

where $\mathcal{L}_{\text{CE}}$ is the standard cross-entropy loss. If the main classifier $\mathbf{W}_{mc}^c$ of $K+1$ classes is only supervised by $\mathcal{L}_{\text{known}}^c$, it is able to distinguish K known classes and incapable of perceiving unknown samples during testing since the training

Algorithm 1 The Proposed FedOSS Algorithm for Federated Open Set Recognition

Input: The number client C, the class number K, the training datasets $\{\mathcal{D}_{tr}^c\}_{c=1}^C$, the number E of local epochs.

// Pre-training stage

Pre-training parameters $\mathbf{W} = \{\mathbf{W}_{fe}, \mathbf{W}_{mc}\}$, $\{\mathbf{W}_{pc}^c\}_{c=1}^C$ using known samples in the FL setting.

// Fine-tuning stage

Server executes:

Initializing the bank $\mathbf{\Omega} = \{\mathbf{W}_{pc}^c\}_{c=1}^C$, the global set: $\mathbf{\Phi} = \{\varnothing\}$.

for each communication round **do**

  for each client $c = 1, 2, \ldots, C$ **do**

    $\mathbf{W}^c, \mathbf{W}_{pc}^c, \{(\boldsymbol{\mu}_k^c, \mathbf{\Sigma}_k^c)\}_{k=1}^K \leftarrow$ ClientUpdate$(\mathbf{W}, \mathbf{\Omega}, \mathbf{\Phi})$

  end for

  Aggregating models: $\mathbf{W} \leftarrow \sum_{c=1}^C \frac{|\mathcal{D}_{tr}^c|}{|\mathcal{D}_{tr}|}\mathbf{W}^c$, $|\mathcal{D}_{tr}| = \sum_{c=1}^C |\mathcal{D}_{tr}^c|$.

  Aggregating local distributions $\{(\boldsymbol{\mu}_k^c, \mathbf{\Sigma}_k^c)\}_{k=1}^K$ of clients via Theorem 1 to update the global distribution set: $\mathbf{\Phi} \leftarrow \{\mathcal{N}(\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)\}_{k=1}^K$.

  Updating the personalized classifier bank: $\mathbf{\Omega} \leftarrow \{\mathbf{W}_{pc}^c\}_{c=1}^C$.

end for

ClientUpdate$(\mathbf{W}, \mathbf{\Omega}, \mathbf{\Phi})$: *// Running on client c*

for each epoch $e = 1, 2, \ldots, E$ **do**

  Randomly selecting a batch of known samples from $\mathcal{D}_{tr}^c$ for computing the loss terms Eq. (12) and Eq. (14).

  Synthesizing the discrete unknown samples set $\mathcal{D}_{du}^c$ via DUSS.

  Sampling the unknown samples $\mathcal{D}_{fo}^c$ from the non-empty $\mathbf{\Phi}$ via FOSS.

  Exploiting $\mathcal{D}_{du}^c$ and $\mathcal{D}_{fo}^c$ to calculate the loss term Eq. (13).

  Minimizing the total loss in Eq. (11) to update $\mathbf{W}^c$ and $\mathbf{W}_{pc}^c$.

end for

Collecting samples in $\mathcal{D}_{du}^c$ to estimate local distributions $\{\boldsymbol{\mu}_k^c, \mathbf{\Sigma}_k^c)\}_{k=1}^K$.

Uploading $\mathbf{W}^c$, $\mathbf{W}_{pc}^c$ and $\{\boldsymbol{\mu}_k^c, \mathbf{\Sigma}_k^c)\}_{k=1}^K$ to the server.

Output: The global model $\mathbf{W} = \{\mathbf{W}_{fe}, \mathbf{W}_{mc}\}$.

set $\mathcal{D}_{tr}^c$ does not contain unknown classes. The proposed DUSS and FOSS can utilize boundary samples in $\mathcal{D}_{bs}^c$ to obtain two virtual unknown samples sets $\mathcal{D}_{du}^c = \{\overline{\mathbf{x}}_i^c\}_{i=1}^{|\mathcal{D}_{du}^c|}$ and $\mathcal{D}_{fo}^c = \{\widehat{\mathbf{x}}_i^c\}_{i=1}^{|\mathcal{D}_{fo}^c|}$. We use these unknown samples to minimize $\mathcal{L}_{\text{unknown}}^c$ to reduce open space risk, which is formulated as:

$$\begin{aligned}\mathcal{L}_{\text{unknown}}^c = &\mathbb{E}_{(\overline{\mathbf{x}}_i^c, \mathbf{y}_i^c)\sim\mathcal{D}_{du}^c}\mathcal{L}_{\text{CE}}(f^c(\mathbf{W}_{fe}^c, \mathbf{W}_{mc}^c, \overline{\mathbf{x}}_i^c)\backslash\mathbf{y}_i^c, K+1),\\ &+\mathbb{E}_{(\widehat{\mathbf{x}}_i^c, \mathbf{y}_i^c)\sim\mathcal{D}_{fo}^c}\mathcal{L}_{\text{CE}}(f^c(\mathbf{W}_{fe}^c, \mathbf{W}_{mc}^c, \widehat{\mathbf{x}}_i^c)\backslash\mathbf{y}_i^c, K+1),\end{aligned} \tag{13}$$

where $\mathbf{y}_i^c$ is the corresponding original known class of $\overline{\mathbf{x}}_i^c/\widehat{\mathbf{x}}_i^c$. $f^c(\mathbf{W}_{fe}^c, \mathbf{W}_{mc}^c, \widehat{\mathbf{x}}_i^c)\backslash\mathbf{y}_i^c$ denotes removing the prediction probability of the class $\mathbf{y}_i^c$, which reduces the effect of some low-quality unknown samples on the performance of known classes. $\mathcal{L}_{\text{extra}}^c$ is used to optimize the parameters $\mathbf{W}_{pc}^c$ of the personalized classifier:

$$\mathcal{L}_{\text{extra}}^c = \mathbb{E}_{(\mathbf{x}_i^c, \mathbf{y}_i^c)\sim\mathcal{D}_{tr}^c}\mathcal{L}_{\text{CE}}(f^c(\mathbf{W}_{fe}^c, \mathbf{W}_{pc}^c, \mathbf{x}_i^c), \mathbf{y}_i^c). \tag{14}$$

Considering that the personalized classifier $\mathbf{W}_{pc}^c$ does not participate in the model aggregation, we do not allow the gradients from it to update the parameters $\mathbf{W}_{fe}^c$ in experiments, thereby reducing its effect on the global model.

IV. Experiments

A. Datasets

To investigate the effectiveness of our FedOSS framework, we evaluate it on two different types of medical datasets.

1) Microscopic Peripheral Blood Cell Dataset: The PBC [55] dataset contains 17,092 microscopic images of peripheral normal blood cells, which can be divided into 8 categories. These images are acquired using the analyser CellaVision DM96 and labelled by expert clinical pathologists at the Hospital Clinic of Barcelona. We follow the work [56] to split all images into training, validation, and test sets using a ratio of 7 : 1 : 2.

2) Gastrointestinal Endoscopy Dataset: We collect 10,662 endoscopic images of the gastrointestinal tract from HyperKvasir dataset [57] and divide these samples into 15 classes according to the location in the gastrointestinal tract and the types of findings. Among these classes, 10 classes belong to the lower gastrointestinal tract, and the rest is from the upper gastrointestinal tract. We randomly partition all samples into training, validation, and test sets with a ratio of 7 : 1 : 2.

3) 3D Organ Classification Dataset: We also perform experiments on a 3D organ classification dataset from the liver tumor segmentation benchmark [58], which contains CT scans of 201 patients. Based on the bounding-box annotations in the work [59], we crop 11 classes of 3D body organs and finally obtain 1743 volumes to perform multi-class classification. All samples are randomly divided into training, validation, and test sets with a ratio of 7 : 1 : 2.

B. Experiment Setup

1) Implementation Details: The proposed FedOSS and comparison methods [4], [5], [6], [7], [8], [9] are implemented with PyTorch library. We adopt the ResNet-18 [61] as the backbone network of all methods for two 2D datasets, which is converted to 3D networks for the organ dataset via ACSConv [62]. The number of clients is set to 8, 8 and 16 for 3D organ, PBC and HyperKvasir datasets, respectively. During the pre-training stage, client models are trained using the Adam [63] optimizer with the initial learning rate of 5×10^{-4} for 100 epochs on both the PBC and 3D organ datasets, and for 200 epochs on the HyperKvasir dataset. Their batch sizes are set to 4, 8 and 8, respectively. The learning rate is divided by 2 every 25 epochs for the PBC and 3D organ datasets, and every 50 epochs for the HyperKvasir dataset. During the fine-tuning stage, we utilize the Adam optimizer with a fixed learning rate of 1×10^{-4} to finetune FedOSS for 30 epochs on the three datasets. The step sizes λ and updating times T of inversion updating are set to 0.1, 1.0, 1.0 and 1, 5, 1 for PBC, HyperKvasir and 3D organ datasets, respectively. Similar to existing FL works [64], [65], we use Dirichlet distribution on label ratios to simulate the

TABLE I
THE PERFORMANCE COMPARISON OF THE PROPOSED METHOD AND EXISTING METHODS ON BLOOD CELL DATASETS

Methods	U = 3				U = 5			
	ACC (%)	Recall (%)	Precision (%)	F1-score (%)	ACC (%)	Recall (%)	Precision (%)	F1-score (%)
Softmax	70.42±3.52	81.03±1.19	72.09±2.76	59.56±4.07	44.20±2.11	74.48±0.51	62.48±5.17	48.78±1.99
OpenMax [4]	73.96±7.32	76.72±4.11	74.98±6.35	72.52±6.70	65.78±3.72	71.88±3.63	72.05±2.13	65.04±1.91
CPN [5]	74.72±5.58	82.18±4.18	75.12±4.43	74.81±5.41	70.89±2.01	83.17±1.70	73.77±2.25	73.07±1.10
CAC [6]	75.50±6.42	82.67±4.54	75.52±5.94	75.75±6.23	74.85±2.54	85.86±1.25	74.86±1.58	75.54±1.21
PROSER [7]	77.48±3.39	82.99±4.21	77.56±2.26	77.64±2.99	61.01±7.36	81.42±1.97	69.29±2.19	64.74±4.38
FedMix [60]	72.18±4.83	73.90±7.12	72.82±5.91	71.92±7.50	57.26±10.56	74.74±7.10	65.87±3.64	61.97±6.54
SSB [8]	78.71±7.92	83.38±4.73	79.46±6.56	79.05±7.19	73.46±10.86	84.41±4.97	73.92±6.25	74.28±9.04
OVRN [9]	77.35±8.02	**83.45±5.54**	78.87±5.36	77.65±7.50	77.95±4.16	**87.49±1.31**	76.20±2.08	78.00±2.90
Ours	**80.04±5.38**	82.46±5.31	**80.73±4.02**	**81.06±4.59**	**82.91±4.27**	84.54±3.80	**79.97±1.96**	**80.56±3.58**

Non-IID data distribution among clients. We set the Dirichlet parameter as 0.5 to ensure the high data heterogeneity.

2) Configuration of Closed and Open Set: To investigate the effect of the number U of unknown classes, we compare FedOSS with previous works [4], [5], [6], [7], [8], [9] in two cases for each dataset. In the first case (U = 3) of the PBC dataset, we randomly sample 5 classes as known classes and the remaining 3 classes as unknown classes. In another case (U = 5), we randomly sample 3 classes as known classes and the remaining 5 classes as unknown classes. Considering that the HyperKvasir dataset is extremely unbalanced, we randomly sample 6 classes from the top 9 categories in number as known classes. In the first case (U = 3), we select 3 classes from the remaining as unknown classes. All remaining 9 classes are regarded as unknown classes in the second case (U = 9). For the 3D organ dataset, we randomly sample 7 classes as known classes and the remaining 4 classes as unknown classes in the first case (U = 4). In another case (U = 7), four classes are randomly sampled as known classes and the remaining 7 classes as unknown classes. For all datasets, the unknown classes are removed from training and validation sets and the training set is divided into local clients based on Dirichlet distribution. The test set contains known classes and unknown class and is used to verify the performance of the global model.

3) Evaluation Metrics: To measure the classification performance of the proposed FedOSS framework in open set scenarios, we adopt commonly-used classification metrics, including accuracy (ACC), F1 score (F1-score), recall score (Recall), and precision score (Precision) on the known classes and the unknown class.

C. Comparisons With State-of-the-Art Methods

We compare our FedOSS framework with Softmax and the state-of-the-art OSR approaches [4], [5], [6], [7], [8], [9] on both microscopic and endoscopic datasets. For a fair comparison, these OSR approaches are implemented in the standard FL framework using the same dataset splits as FedOSS for each dataset.

1) Experimental Results on Microscopic Dataset: In Table I, we present the classification performance of different methods on the microscopic dataset to validate the proposed FedOSS. It can be observed that prototype based approaches [5], [6] outperform the baseline Softmax with pretty large margins, which indicate that thresholding distances between prototypes of known classes and unknown samples to reject open data is more effective than thresholding softmax based confidence scores. Our FedOSS suppresses the best prototype based approach, i.e., CAC [6], with significant performance increments for two cases, such as 5.31% in F1-score for U = 3 and 8.06% in ACC for U = 5. Compared with CAC, FedOSS directly learns the boundary between known classes and the unknown class by synthesizing unknown data instead of depending on thresholding. Additionally, compared with SSB [8] that utilizes various strategies to learn more tight clusters of known classes, such as more augmentation, better learning rate schedules, and label smoothing, FedOSS obtains better performance with significant increments, such as 2.01% in F1-score (U = 3) and 9.45% in ACC (U = 5), since it utilizes synthesized unknown data to improve the tightness of known classes and learn boundaries between known and unknown classes. Furthermore, we can also find that most of previous methods [4], [5], [6], [7], [8] undergo performance degradation when the number U of unknown classes increases. In contrast, the performance of FedOSS is stable with the number U of unknown classes. These experimental results on the microscopic dataset can confirm that the proposed FedOSS is superior to previous methods in recognizing unknown samples.

2) Experimental Results on Endoscopic Dataset: In Table III, we further compare the proposed FedOSS framework with previous approaches [4], [5], [6], [7], [8], [9] on the endoscopic dataset. In contrast to the baseline Softmax, the best prototype based method, CAC [6], only yields limited performance improvements, especially in the case of U = 9, such as merely 1.56% in Precision and 1.71% in F1-score. By comparison, FedOSS outperforms the baseline Softmax with enormous increments in the case of U = 9, such as 11.08% in Precision and 13.55% in F1-score. The performance advantages can indicate that directly learning the boundary between known classes and the unknown class to recognize open data is more effective than thresholding distances between prototypes of known classes and unknown samples when the number of unknown classes increases. PROSER [7] achieves the second-best classification performance in the case of U = 3 but obtains inferior results in the case of U = 9. On the contrary, OVRN [9] performs better in the case of U = 9

TABLE II
THE PERFORMANCE COMPARISON OF THE PROPOSED METHOD AND EXISTING METHODS ON ENDOSCOPY DATASET

Methods	U = 3				U = 9			
	ACC (%)	Recall (%)	Precision (%)	F1-score (%)	ACC (%)	Recall (%)	Precision (%)	F1-score (%)
Softmax	77.41±8.19	84.75±0.37	81.64±4.39	76.60±3.75	60.62±3.67	84.43±0.96	66.81±2.50	65.86±1.46
OpenMax [4]	80.27±5.41	81.80±7.74	84.91±1.87	79.79±7.80	65.99±1.88	82.09±3.24	66.89±7.96	69.80±3.39
CPN [5]	80.35±5.19	84.36±4.86	80.30±7.90	80.25±7.06	61.03±9.31	82.21±3.47	64.81±8.68	67.57±7.55
CAC [6]	83.34±4.65	87.52±1.73	84.67±2.59	84.67±2.60	64.73±6.51	83.51±1.11	68.37±4.40	70.11±4.82
PROSER [7]	85.70±2.95	**89.05±0.67**	86.28±2.65	85.56±1.25	69.29±3.05	85.66±2.79	72.44±2.08	73.16±1.23
FedMix [60]	83.62±3.27	86.27±2.18	80.86±5.71	82.52±4.23	69.92±3.38	82.85±1.46	72.76±1.71	73.33±3.32
SSB [8]	80.89±4.74	84.92±3.13	83.83±2.53	81.89±3.48	66.45±5.59	83.55±3.29	71.70±2.93	72.06±4.73
OVRN [9]	78.49±9.29	84.47±2.63	83.09±2.95	81.58±4.86	71.45±7.78	**86.72±2.86**	74.84±3.84	76.06±5.35
Ours	**88.29±1.80**	88.29±3.07	**89.07±1.46**	**87.78±3.23**	**77.17±2.05**	86.23±2.01	**77.89±0.94**	**79.41±2.46**

TABLE III
THE PERFORMANCE COMPARISON OF THE PROPOSED METHOD AND EXISTING METHODS ON 3D ORGAN CLASSIFICATION DATASET

Methods	U = 4				U = 7			
	ACC (%)	Recall (%)	Precision (%)	F1-score (%)	ACC (%)	Recall (%)	Precision (%)	F1-score (%)
Softmax	67.72±1.29	82.20±2.81	71.45±1.47	72.27±1.73	44.89±3.24	75.57±1.96	56.82±0.62	52.79±2.29
OpenMax [4]	49.95±3.19	56.94±9.58	52.36±1.72	49.44±6.66	66.38±4.80	54.04±6.62	68.20±10.92	52.45±5.71
CPN [5]	59.12±3.12	74.25±1.84	65.84±3.83	65.20±1.70	40.40±7.13	63.81±6.59	51.54±13.07	48.43±6.95
CAC [6]	71.92±5.36	82.90±3.91	75.34±6.58	76.38±5.59	44.40±5.37	73.45±2.80	53.72±3.71	54.99±2.66
PROSER [7]	78.22±1.64	86.76±2.53	79.96±0.89	81.92±1.11	51.77±11.59	69.19±10.41	64.51±7.98	56.60±9.92
FedMix [60]	77.75±2.89	**87.06±3.09**	80.78±1.86	81.79±2.86	43.84±2.97	72.45±3.52	55.38±6.49	53.05±4.74
SSB [8]	67.81±3.98	81.51±2.04	75.21±5.45	74.19±4.19	46.23±7.03	70.88±5.60	58.16±5.33	52.66±6.69
OVRN [9]	64.18±1.30	78.59±2.64	73.77±4.56	71.17±2.10	57.69±3.52	**79.82±2.09**	61.59±4.30	62.55±4.25
Ours	**80.42±2.92**	81.77±1.81	**85.36±3.87**	**82.84±2.28**	**78.32±3.41**	75.44±4.02	**80.86±1.59**	**75.81±3.30**

TABLE IV
THE PERFORMANCE OF THE PROPOSED FEDERATED OPEN SET RECOGNITION FRAMEWORK WITH DIFFERENT MODULES

Configurations			U = 3				U = 9			
Baseline	DUSS	FOSS	ACC (%)	Recall (%)	Precision (%)	F1-score (%)	ACC (%)	Recall (%)	Precision (%)	F1-score (%)
√			77.41±8.19	84.75±0.37	81.64±4.39	76.60±3.75	60.62±3.67	84.43±0.96	66.81±2.50	65.86±1.46
√	√		87.22±2.64	**89.73±1.26**	87.48±1.91	87.51±1.38	72.45±2.24	86.21±1.34	74.08±2.66	76.29±2.25
√	√	√	**88.29±1.80**	88.29±3.07	**89.07±1.46**	**87.78±3.23**	**77.17±2.05**	**86.23±2.01**	**77.89±0.94**	**79.41±2.46**

TABLE V
THE PERFORMANCE OF THE PROPOSED DISCRETE UNKNOWN SAMPLE SYNTHESIS (DUSS) MODULE WITH DIFFERENT CONFIGURATIONS

Configurations	U = 3				U = 9			
	ACC (%)	Recall (%)	Precision (%)	F1-score (%)	ACC (%)	Recall (%)	Precision (%)	F1-score (%)
DUSS (w/o BSR)	84.50±4.71	88.74±0.23	85.71±3.58	84.67±2.47	69.87±1.38	86.08±2.26	72.09±1.73	73.68±0.84
DUSS (IU = 0)	85.75±3.51	89.31±0.65	86.30±2.70	85.92±1.81	70.40±0.36	86.02±1.60	72.54±1.17	74.23±1.08
DUSS (IU = 1)	86.23±2.95	89.42±0.74	86.54±2.25	86.36±1.41	70.49±3.59	86.38±1.91	72.52±1.32	74.29±0.98
DUSS (IU = 3)	**87.36±2.16**	**89.80±0.62**	**87.66±1.51**	87.48±0.58	71.53±1.70	**86.69±1.67**	73.25±1.86	75.27±1.27
DUSS (IU = 5)	87.22±2.64	89.73±1.26	87.48±1.91	**87.51±1.38**	72.45±2.24	86.21±1.34	74.08±2.66	76.29±2.25
DUSS (IU = 7)	86.72±2.66	89.17±1.36	87.25±1.43	87.19±1.28	**72.57±2.38**	86.05±1.77	**74.40±2.57**	**76.46±2.08**

than PROSER [7] but shows bad performance in the case of U = 3. Compared with PROSER [7], and OVRN [9], our FedOSS framework demonstrates the best performance in both two cases and exceeds them by a large margin in Precision and F1-score, such as 2.79% and 2.22% in contrast to PROSER [7] for the case of U = 3, and 3.05% and 3.35% in contrast to OVRN [9] for the case of U = 9. These experimental results further prove that the proposed FedOSS can recognize open-set classes and achieve better performance than existing methods.

3) Experimental Results on 3D Organ Dataset: We further compare the proposed FedOSS framework with previous approaches [4], [5], [6], [7], [8], [9] on the 3D organ classification dataset, as shown in Table III. It can be observed that, for the case of U = 4, FedMix [60] yields the excellent performance among existing methods, such as 77.75% in ACC and 80.78% in Precision. Noticeably, the proposed FedOSS outperforms FedMix with considerable margins, such as 2.67% in ACC and 4.58% in Precision. Additionally, the performance of existing methods [5], [6], [7], [8], [9] is sensitive to the

number of unknown classes. For instance, PROSER undergoes an enormous performance drop when the number of unknown classes increases to 7, achieving 51.77% in ACC and 64.51% in Precision with decrements of 26.45% and 15.45%, respectively. By comparison, our method does suffer from a slight drop and outperforms all previous methods with the highest performance, such as 78.32% in ACC and 80.86% in Precision with decrements of merely 2.10% and 4.50%, respectively. These experimental results prove that the proposed FedOSS can also perform well on 3D medical data and achieve the superior performance in contrast to the state-of-the-art methods.

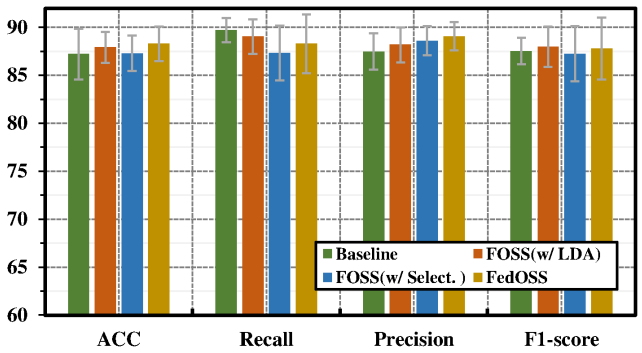
D. Ablation Study

DUSS and FOSS are pivotal components for FedOSS to recognize open data. We conduct ablation experiments on the gastrointestinal dataset to investigate the effectiveness of these modules and provide some visualization results.

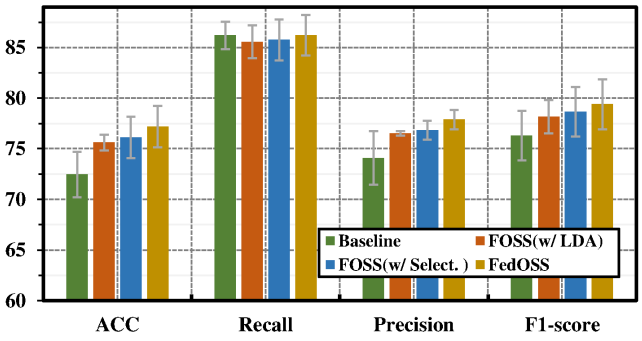
1) Evaluation of Different Modules: FedOSS is implemented based on the standard FedAvg [14] with softmax, which is regarded as the baseline. To evaluate the efficacy of individual modules, we combine the baseline with different modules. As shown in Table IV, DUSS promotes the baseline with remarkable performance improvements in two cases, such as 9.81% in ACC and 5.84% in Precision for U = 3, 11.83% in ACC and 7.27% in Precision for U = 9, which highlight the importance of synthesized unknown samples. The proposed FedOSS, i.e., the baseline equipped with DUSS and FOSS simultaneously, achieves the best performance in two cases, especially in the case of U = 9, with the increments of 4.72% in ACC and 3.81% in Precision in comparison with 'Baseline + DUSS'. This indicates that FOSS can help FedOSS to learn the higher quality of boundary between known classes and the unknown class and still obtain a high performance even when the number of unknown classes increases. These results confirm the effectiveness of the DUSS and FOSS modules.

2) Ablative Experiments on DUSS Module: Boundary sample recognition (BSR) and inversion updating (IU) are two crucial steps for DUSS to synthesize unknown samples. We investigate their efficacy by comparing the following settings. 1) DUSS (w/o BSR): DUSS regards all known samples as the virtual unknown class and does not conduct inversion updating; 2) DUSS (IU = T): DUSS selects out boundary samples via the client discrepancy and then conducts T times of inversion updating. From Table V, we observe that the setting 'DUSS (IU = 0)' obtains the better performance than the setting 'DUSS (w/o BSR)'. This result verifies the importance of boundary sample recognition. In addition, the settings DUSS (IU > 0) outperform the setting 'DUSS (IU = 0)'. Meanwhile, the model can yield the better classification performance as the times T of inversion updating increases. The results manifest that the sufficient inversion updating can transform boundary samples into virtual unknown samples.

The proposed strategy of boundary sample recognition can be viewed as a kind of uncertainty estimation method for measuring the distance between a sample and the decision boundary. To verify its superiority, we compare it with existing methods (including Method A [66] and Method B [67]) on the endoscopy dataset. In Method A [66], a training instance would have high uncertainty if its prediction changes after being imposed a strong perturbation. Method B [67] determines the distance from the decision boundary by computing the difference between the probability of the true label and the probability of the most confusing label (with the second highest probability). As shown in Table VI, the proposed strategy outperforms Method B [67] with large margins in the two cases, such as 9.03% in ACC and 9.36% in F1-score in the case of U = 3, and 15.25% in Precision and 11.35% in F1-score in the case of U = 9. The performance advantages

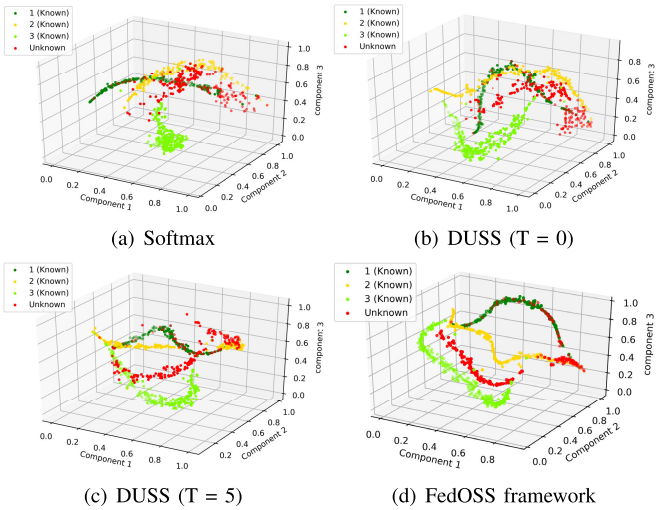


(a) U = 3



(b) U = 9

Fig. 3. The performance of FOSS module with different configurations on endoscopy dataset.



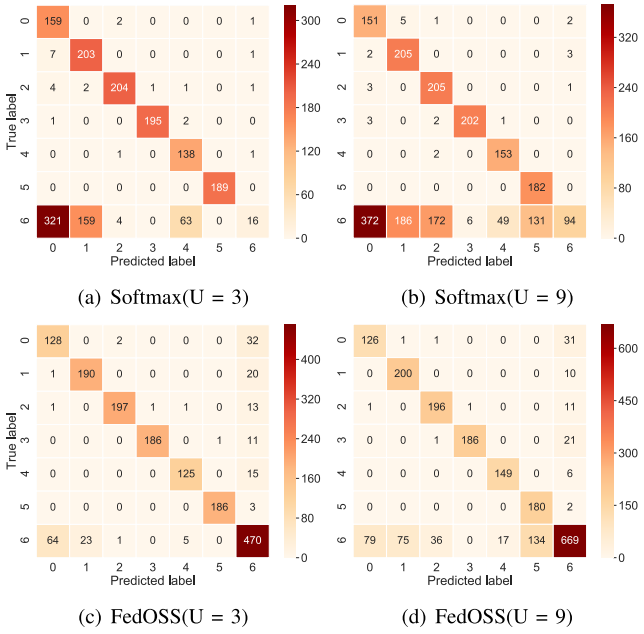
(a) Softmax (b) DUSS (T = 0)

(c) DUSS (T = 5) (d) FedOSS framework

Fig. 4. Visualization of the learned feature space of different approaches on endoscopy dataset. Best viewed in color.

TABLE VI
THE PERFORMANCE COMPARISON OF DIFFERENT STRATEGIES FOR BOUNDARY SAMPLE RECOGNITION ON ENDOSCOPY DATASET

Methods	U = 3				U = 9			
	ACC (%)	Recall (%)	Precision (%)	F1-score (%)	ACC (%)	Recall (%)	Precision (%)	F1-score (%)
Method A [66]	76.55±8.01	84.08±0.93	73.38±8.40	75.25±3.56	58.65±5.19	83.99±1.15	58.36±5.78	64.21±2.27
Method B [67]	78.19±9.36	85.37±0.94	79.26±10.75	78.15±5.91	59.07±5.30	83.49±1.47	58.83±6.14	64.94±2.35
Ours	**87.22±2.64**	**89.73±1.26**	**87.48±1.91**	**87.51±1.38**	**72.45±2.24**	**86.21±1.34**	**74.08±2.66**	**76.29±2.25**



(a) Softmax(U = 3) (b) Softmax(U = 9)

(c) FedOSS(U = 3) (d) FedOSS(U = 9)

Fig. 5. The confusion matrixes of the baseline (Softmax) and our FedOSS on endoscopy dataset.

highlight that exploiting inter-client discrepancy to estimate the uncertainty of samples is more accurate. In contrast, the existing methods can only use information of single client and are unsuitable for the distributed setting.

3) Ablative Experiments on FOSS Module: We further verify the effectiveness of the local distribution aggregation (LDA) (FOSS w/ LAD) and the selection of high-quality unknown samples from ϵ-likelihood region of global distributions (FOSS w/ Select.) in FOSS. The baseline setting is our FedOSS framework equipped with DUSS. In Fig. 3, it can be seen that both settings 'FOSS w/ LAD' and 'FOSS w/ Select.' outperform the baseline. Additionally, only when FOSS is configured with the two components simultaneously, FedOSS obtains the highest performance in the two cases. These results prove that FOSS can aggregate the knowledge of different clients in the open space and enhance the diversity of virtual unknown samples to improve the performance of FedOSS.

4) Visualization: We visualize the learned feature space of different approaches on the endoscopy dataset to observe the distribution of known classes and the unknown class, as shown in Fig. 4. The unknown class actually contains three categories of unknown samples. For the clear observation, we randomly select and visualize three classes of known samples. From Fig. 4(a), we can see that most of unknown samples are scattered and mixed with the known classes 1 and 2 when the model is trained only using softmax. After we select boundary samples as virtual unknown samples to update the model (DUSS (T = 0)) Fig. 4(b), the boundaries between known classes and the unknown class become slightly clear, which are clearer after imposing inversion updating on these selected boundary samples in Fig. 4(c). Relying on DUSS and FOSS modules, FedOSS learns compact feature representations and clearly separates known classes and the unknown class in Fig. 4(d). These visualization results are able to verify the effectiveness of the proposed FOSS framework.

We further visualize the confusion matrixes of the baseline (Softmax) and our FedOSS on the endoscopy dataset, as shown in Fig. 5. We observe that the baseline method nearly classifies all unknown samples into the known classes, with the accuracy of merely 2.84% (U=3) in Fig. 5(a) and 9.31% (U=9) in Fig. 5(b). In contrast, our FedOSS shows a superior ability in rejecting unknown classes, with an accuracy of 83.48% (U=3) and 66.24% (U=9). Meanwhile, our method still demonstrates the excellent performance in known classes. The experimental results confirm the effectiveness of the generated virtual unknown samples in our FedOSS framework for learning boundaries between known and unknown classes.

V. CONCLUSION AND DISCUSSION

In this paper, we propose a novel framework for federated open set recognition, Federated Open Set Synthesis (FedOSS), which contains two modules, i.e., Discrete Unknown Sample Synthesis (DUSS) and Federated Open Space Sampling (FOSS). DUSS leverages the knowledge discrepancy between clients to recognize known samples near boundaries and then transforms them into discrete virtual unknown samples via inversion updating. FOSS unites these virtual unknown samples of all clients to estimate global distributions of open space near boundaries, which is used to sample diverse unknown samples. Relying on these synthesized open data of DUSS and FOSS modules, FedOSS can learn decision boundaries to separate known classes and unknown classes. The comprehensive experiments on microscopic and endoscopic datasets validate the effectiveness of FedOSS. The results on both two datasets show the superior performance of FedOSS in contrast to state-of-the-art methods. In ablation experiments, we first verify the importance of the DUSS and FOSS modules of FedOSS, and then deeply analyze the impact of vital components in the two modules. Finally, visualization results further confirm the ability of our FedOSS framework to separate unknown and known classes in the feature space.

There are two limitations when directly applying our FedOSS framework into the realistic medical scenes.

(1) FedOSS suffers from the data security risk during communication process. Although it follows the basic rule of federated learning [68] and does not share the raw data of clients. However, local client models might be stolen by intruders for the reconstruction of original data. For this problem, we are able to apply existing homomorphic encryption techniques [68], [69] to encrypt client models and the global model. (2) FedOSS performs well when label distributions of clients are heterogenous, but it will face challenges for heterogeneous feature distributions. Feature distribution heterogeneity indicates that feature distributions among clients have the overlapping area and specific areas [70]. The samples lying in the overlapping area are more likely to be accurately classified by all clients, and the samples located in specific areas are usually misclassified. However, the misclassified samples are not necessarily boundary samples. Therefore, our method might fail to recognize boundary samples. For this issue, we can introduce existing methods [36], [71] to align feature spaces of different clients, thus ensuring that our method can still achieve the good performance.

REFERENCES

[1] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017.

[2] M. Zhu, Z. Chen, and Y. Yuan, "DSI-Net: Deep synergistic interaction network for joint classification and segmentation with endoscope images," *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3315–3325, Dec. 2021.

[3] Z. Chen, J. Liu, M. Zhu, P. Y. M. Woo, and Y. Yuan, "Instance importance-aware graph convolutional network for 3D medical diagnosis," *Med. Image Anal.*, vol. 78, May 2022, Art. no. 102421.

[4] A. Bendale and T. E. Boult, "Towards open set deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1563–1572.

[5] H. Yang, X. Zhang, F. Yin, Q. Yang, and C. Liu, "Convolutional prototype network for open set recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2358–2370, May 2022.

[6] D. Miller, N. Sünderhauf, M. Milford, and F. Dayoub, "Class anchor clustering: A loss for distance-based open set recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3569–3577.

[7] D. Zhou, H. Ye, and D. Zhan, "Learning placeholders for open-set recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4399–4408.

[8] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman, "Open-set recognition: A good closed-set classifier is all you need," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[9] J. Jang and C. O. Kim, "Collective decision of one-vs-rest networks for open-set recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jul. 14, 2022, doi: 10.1109/TNNLS.2022.3189996.

[10] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1757–1772, Jul. 2013.

[11] W. J. Scheirer, L. P. Jain, and T. E. Boult, "Probability models for open set recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2317–2324, Nov. 2014.

[12] A. Xu et al., "Closing the generalization gap of cross-silo federated medical image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20834–20843.

[13] Z. Chen, M. Zhu, C. Yang, and Y. Yuan, "Personalized retrogress-resilient framework for real-world medical federated learning," in *Proc. MICCAI*. Cham, Switzerland: Springer, 2021, pp. 347–356.

[14] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. AISTATS*, 2017, pp. 1273–1282.

[15] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. MLSys*, vol. 2, 2020, pp. 429–450.

[16] L. Neal, M. Olson, X. Fern, W.-K. Wong, and F. Li, "Open set learning with counterfactual images," in *Proc. ECCV*, 2018, pp. 613–628.

[17] W. Moon, J. Park, H. S. Seong, C.-H. Cho, and J.-P. Heo, "Difficulty-aware simulator for open set recognition," in *Proc. ECCV*, 2022, pp. 365–381.

[18] Z. Yue, T. Wang, Q. Sun, X. Hua, and H. Zhang, "Counterfactual zero-shot and open-set visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15399–15409.

[19] S. Kong and D. Ramanan, "OpenGAN: Open-set recognition via open data generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 793–802.

[20] S. Girish, S. Suri, S. Rambhatla, and A. Shrivastava, "Towards discovery and attribution of open-world GAN generated images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14074–14083.

[21] G. Chen, P. Peng, X. Wang, and Y. Tian, "Adversarial reciprocal points learning for open set recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8065–8081, Nov. 2022.

[22] Z. Ge, S. Demyanov, and R. Garnavi, "Generative OpenMax for multi-class open set classification," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 42.1–42.12.

[23] Y. Hsu, Y. Shen, H. Jin, and Z. Kira, "Generalized ODIN: Detecting out-of-distribution image without learning from out-of-distribution data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10948–10957.

[24] S. Bhattacharjee, D. Mandal, and S. Biswas, "Multi-class novelty detection using mix-up technique," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1389–1398.

[25] P. Oza and V. M. Patel, "Utilizing patch-level category activation patterns for multiple class novelty detection," in *Proc. ECCV*, 2020, pp. 421–437.

[26] W. Li, J. Chen, Z. Wang, Z. Shen, C. Ma, and X. Cui, "IFL-GAN: Improved federated learning generative adversarial network with maximum mean discrepancy model aggregation," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Apr. 26, 2022, doi: 10.1109/TNNLS.2022.3167482.

[27] X. Wu, H. Huang, H. Wang, Y. Wang, and Q. Xu, "EP-GAN: Unsupervised federated learning with expectation-propagation prior GAN," 2022. [Online]. Available: https://openreview.net/forum?id=djwnKXz1B2

[28] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *Proc. ICLR*, 2018.

[29] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, Apr. 1995.

[30] K. Sun, Z. Zhu, and Z. Lin, "Enhancing the robustness of deep neural networks by boundary conditional GAN," 2019, *arXiv:1902.11029*.

[31] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014.

[32] J. Zhang, J. Zhu, G. Niu, B. Han, M. Sugiyama, and M. Kankanhalli, "Geometry-aware instance-reweighted adversarial training," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[33] E. M. Rudd, L. P. Jain, W. J. Scheirer, and T. E. Boult, "The extreme value machine," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 762–768, Mar. 2018.

[34] P. R. M. Júnior, T. E. Boult, J. Wainer, and A. Rocha, "Open-set support vector machines," 2016, *arXiv:1606.03802*.

[35] R. Yoshihashi, W. Shao, R. Kawakami, S. You, M. Iida, and T. Naemura, "Classification-reconstruction learning for open-set recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4011–4020.

[36] M. Jiang, Z. Wang, and Q. Dou, "HarmoFL: Harmonizing local and global drifts in federated learning on heterogeneous medical images," in *Proc. AAAI*, 2022, vol. 36, no. 1, pp. 1087–1095.

[37] W. Zhu and J. Luo, "Federated medical image analysis with virtual sample synthesis," in *Proc. MICCAI*. Cham, Switzerland: Springer, 2022, pp. 728–738.

[38] J. Wicaksana, Z. Yan, X. Yang, Y. Liu, L. Fan, and K. Cheng, "Customized federated learning for multi-source decentralized medical image classification," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 11, pp. 5596–5607, Nov. 2022.

[39] Y. Yeganeh, A. Farshad, N. Navab, and S. Albarqouni, "Inverse distance aggregation for federated learning with non-IID data," in *Proc. MICCAI Workshop Distrib. Collaborative Learn.* Cham, Switzerland: Springer, 2020, pp. 150–159.

[40] Z. Chen, C. Yang, M. Zhu, Z. Peng, and Y. Yuan, "Personalized retrogress-resilient federated learning toward imbalanced medical data," *IEEE Trans. Med. Imag.*, vol. 41, no. 12, pp. 3663–3674, Dec. 2022.

[41] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "FedBN: Federated learning on non-IID features via local batch normalization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.
[42] W. Lu et al., "Personalized federated learning with adaptive batchnorm for healthcare," *IEEE Trans. Big Data*, early access, May 23, 2022, doi: 10.1109/TBDATA.2022.3177197.
[43] D. Yang et al., "Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan," *Med. Image Anal.*, vol. 70, May 2021, Art. no. 101992.
[44] Q. Liu, H. Yang, Q. Dou, and P.-A. Heng, "Federated semi-supervised medical image classification via inter-client relation matching," in *Proc. MICCAI*. Cham, Switzerland: Springer, 2021, pp. 325–335.
[45] T. Bdair, N. Navab, and S. Albarqouni, "FedPerl: Semi-supervised peer learning for skin lesion classification," in *Proc. MICCAI*, 2021, pp. 336–346.
[46] M. Jiang, H. Yang, X. Li, Q. Liu, P.-A. Heng, and Q. Dou, "Dynamic bank learning for semi-supervised federated image diagnosis with class imbalance," in *Proc. MICCAI*, 2022, pp. 196–206.
[47] X. Liang, Y. Lin, H. Fu, L. Zhu, and X. Li, "RSCFed: Random sampling consensus federated semi-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10144–10153.
[48] Y. Yu, W.-Y. Qu, N. Li, and Z. Guo, "Open-category classification by adversarial sample generation," in *Proc. IJCAI*, 2017, pp. 3357–3363.
[49] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.
[50] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Proc. Artif. Intell. Saf. Secur.*, 2018, pp. 99–112.
[51] S. Cui, S. Wang, J. Zhuo, L. Li, Q. Huang, and Q. Tian, "Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3940–3949.
[52] X.-C. Li and D.-C. Zhan, "FedRS: Federated learning with restricted softmax for label distribution non-IID data," in *Proc. 27th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2021, pp. 995–1005.
[53] X. Sun, Z. Yang, C. Zhang, K. Ling, and G. Peng, "Conditional Gaussian distribution learning for open set recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13477–13486.
[54] Q. Wang, P. Li, and L. Zhang, "G2DeNet: Global Gaussian distribution embedding network and its application to visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6507–6516.
[55] A. Acevedo, A. Merino, S. Alférez, Á. Molina, L. Boldú, and J. Rodellar, "A dataset of microscopic peripheral blood cell images for development of automatic recognition systems," *Data Brief*, vol. 30, Jun. 2020, Art. no. 105474.
[56] J. Yang et al., "MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification," 2021, *arXiv:2110.14795*.
[57] H. Borgli et al., "HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy," *Sci. Data*, vol. 7, no. 1, pp. 1–14, Aug. 2020.
[58] P. Bilic et al., "The liver tumor segmentation benchmark (LiTS)," *Med. Image Anal.*, vol. 84, Feb. 2023, Art. no. 102680.
[59] X. Xu, F. Zhou, B. Liu, D. Fu, and X. Bai, "Efficient multiple organ localization in CT image using 3D region proposal network," *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1885–1898, Aug. 2019.
[60] T. Yoon, S. Shin, S. J. Hwang, and E. Yang, "FedMix: Approximation of mixup under mean augmented federated learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.
[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[62] J. Yang et al., "Reinventing 2D convolutions for 3D images," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 8, pp. 3009–3018, Aug. 2021.
[63] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[64] D. A. E. Acar, Y. Zhao, R. M. Navarro, M. Mattina, P. N. Whatmough, and V. Saligrama, "Federated learning based on dynamic regularization," 2021, *arXiv:2111.04263*.
[65] L. Gao, H. Fu, L. Li, Y. Chen, M. Xu, and C. Xu, "FedDC: Federated learning with non-IID data via local drift decoupling and correction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10102–10111.
[66] W. Hua, Y. Zhang, C. Guo, Z. Zhang, and G. E. Suh, "BulletTrain: Accelerating robust neural network training via boundary example mining," in *Proc. NIPS*, vol. 34, 2021, pp. 18527–18538.
[67] V. Koltchinskii and D. Panchenko, "Empirical margin distributions and bounding the generalization error of combined classifiers," *Ann. Statist.*, vol. 30, no. 1, pp. 1–50, Feb. 2002.
[68] P. Kairouz et al., "Advances and open problems in federated learning," *Found. Trends Mach. Learn.*, vol. 14, nos. 1–2, pp. 1–210, 2019.
[69] C. Zhang, S. Li, J. Xia, W. Wang, F. Yan, and Y. Liu, "BatchCrypt: Efficient homomorphic encryption for cross-silo federated learning," in *Proc. USENIX ATC*, 2020, pp. 493–506.
[70] S. Cui et al., "Decentralized federated learning via overlapping data augmentation," in *Proc. ICLR*, 2023, pp. 1–18.
[71] F. Yu et al., "Fed2: Feature-aligned federated learning," in *Proc. KDD*, 2021, pp. 2066–2074.