Pseudo-Data based Self-Supervised Federated Learning for Classification of Histopathological Images

Yuanming Zhang, Zheng Li, Xiangmin Han, Saisai Ding, Juncheng Li, Jun Wang, *Member*, *IEEE*, Shihui Ying, *Member*, *IEEE*, Jun Shi, *Member*, *IEEE*

***Abstract*—Computer-aided diagnosis (CAD) can help pathologists improve diagnostic accuracy together with consistency and repeatability for cancers. However, the CAD models trained with the histopathological images only from a single center (hospital) generally suffer from the generalization problem due to the straining inconsistencies among different centers. In this work, we propose a pseudo-data based self-supervised federated learning (FL) framework, named SSL-FT-BT, to improve both the diagnostic accuracy and generalization of CAD models. Specifically, the pseudo histopathological images are generated from each center, which contain both inherent and specific properties corresponding to the real images in this center, but do not include the privacy information. These pseudo images are then shared in the central server for self-supervised learning (SSL) to pre-train the backbone of global mode. A multi-task SSL is then designed to effectively learn both the center-specific information and common inherent representation according to the data characteristics. Moreover, a novel Barlow Twins based FL (FL-BT) algorithm is proposed to improve the local training for the CAD models in each center by conducting model contrastive learning, which benefits the optimization of the global model in the FL procedure. The experimental results on four public histopathological image datasets indicate the effectiveness of the proposed SSL-FL-BT on both diagnostic accuracy and generalization.**

***Index Terms*—Histopathological image, federated learning, multi-center learning, self-supervised learning, Barlow twins contrastive learning**

I. INTRODUCTION

Cancers seriously threaten human health. Histopathological diagnosis is the "gold standard" for the diagnosis of cancers in clinical practice [1]. However, it generally suffers from the issues of low efficiency, consistency and repeatability [2]. To this end, computer-aided diagnosis (CAD) for histopathological images has attracted considerable attention in recent years [3][4]. As a classical deep learning method, convolutional neural network (CNN) and its variants have proved their effectiveness as the backbone for the CAD models of histopathological images [3][4][5][6][7][8].

It is worth noting that even the most common and accessible type of stain, such as hematoxylin and eosin (H&E), will still produce different color intensities depending on the brand, storage time, and temperature [1]. It then results in inconsistencies in the stained histopathological images among different hospitals [9]. If the training samples are only acquired from one hospital, the generalization of a CAD model then will be degraded. To this end, a potential solution is to train the CAD model with the histopathological images from multiple hospitals (*i.e.*, multi-centers). Some pioneering works have indicated the feasibility and effectiveness of multi-center learning for improving the generalization of CNN models [10]. Moreover, this manner also can alleviate the problem of small sample size (SSS), which is a common issue in the field of CAD [4].

For multi-center learning, it is a popular way to gather data from all centers together to train a model [11][12]. However, this training strategy suffers from the issues of privacy protection, data security, and data ownership [13]. In fact, some hospitals strictly prohibit the use of medical data outside the hospital. Hence, the application of multi-center learning is greatly limited with shared data. Federated learning (FL) then emerges as a promising solution, which can jointly train the CAD models by sharing parameters of distributed local models instead of the local data in the conventional multi-center learning paradigm [14][15]. This new multi-center learning paradigm has gained considerable attention in the field of healthcare [13][16], and it has been successfully applied for the CAD tasks [17][18], including for histopathological images [19]. However, it still cannot guarantee that the distributed CAD models fully capture the specific properties of different centers' data, because FL only shares the model parameters instead of data, and the distributed local models do not contain enough specific information.

In recent years, image synthesis has achieved remarkable performances due to the fast development of generative adversarial network (GAN) and its variants [20][21][22]. Some works have adopted the synthesized images for data augmentation to train CNN models [23][24]. Thus, if some pseudo histopathological images are generated in each center,

This work is supported by National Natural Science Foundation of China (81830058, 11971296) and the 111 Project (D20031). (Corresponding authors: Jun Shi)

Y. Zhang, Z. Li, S. Ding, J. Li, J. Wang and J. Shi are with the Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Joint International Research Laboratory of Specialty Fiber Optics and Advanced Communication, Shanghai Institute for Advanced Communication and Data Science, School of Communication and Information Engineering, Shanghai University, China. (Email: junshi@shu.edu.cn)

X. Han is with the School of Software, Tsinghua University, China.

S. Ying is with the Department of Mathematics, School of Science, Shanghai University, China.

they can contain inherent and specific properties corresponding to the real histopathological images of the center, but do not include the privacy information. Thus, it is a feasible way to share these pseudo data to pre-train the backbone of CNN model in the central server, and then further conduct FL. This strategy can promote the CAD model to learn more specific properties of each center's data and further improve the generalization ability. However, the pseudo histopathological images do not have corresponding labels for cancers, and therefore, they cannot be directly used in the same classification task as the real images to pre-train the backbone network of a CAD model.

Self-supervised learning (SSL) then provides a feasible way to explore and learn inherent information from these pseudo histopathological images, because it generates supervision directly from the training samples themselves to design pretext tasks [25][26]. SSL can effectively improve the feature representation of a backbone network for the downstream task, and it has been successfully applied to various tasks in the field of medical image analysis [27][28]. Consequently, we can develop a multi-task SSL-based FL (SSL-FL) framework to make full use of the pseudo histopathological images. In particular, since the previous works have proved that the image restoration task can effectively learn the detailed contextual information [27][29], the inherent anatomical information [30], and other inherent knowledge from medical images, it can also be applied to learn these common properties of all multi-center pseudo histopathological images stored in the central server in this work. Moreover, since we know which center a pseudo image is generated from, it can be used as label information. Consequently, we specifically design a center classification task that discriminates the source of an image generated from. This SSL task can make the pre-trained backbone learn more individual data representation of different centers to improve model generalization. Overall, the abovementioned tasks perform simultaneously and effectively learn both inherently common representation across multiple centers and center-specific knowledge.

It is worth noting that although the proposed multi-task SSL with shared pseudo images under FL framework can effectively improve the generalization of CAD models, the FL still suffers from the issue of data heterogeneity due to the stain difference in different centers. In the conventional FL algorithms, the data heterogeneity will result in the drift of local models during training procedure, which then makes the objective functions of local models far from that of global model [16]. To this end, Model-Contrastive Federated Learning (MOON) has been proposed recently, which innovatively introduces contrastive SSL into FL for model-level contrast [31]. MOON adopts the similarity between model representations to correct the local training of individual centers, and it has achieved superior performance in handling the heterogeneity of local data distribution [31].

Although MOON has the potential to alleviate the heterogeneity of histopathological images, when MOON maximizes the representation agreement between the local and global models, the contrastive operation is still inefficient due to the requirement of negative samples similar to SimCLR [32]. In fact, insufficient negative sample pairs in contrastive SSL will result in insufficient clustering, and cannot distinguish the sample difference between groups [33]. On the contrary, excessive negative sample pairs may lead to over-clustering, which makes the model difficult to learn common features for samples of the same class [33]. Moreover, in MOON, the positive pair is formed by the local model being updated and the global model, and the negative pair is formed by the local model being updated and the local model from the previous round [31]. However, the definition of positive and negative samples in MOON is somewhat unclear and inexplicable. In fact, the prior local model might also possess well-suited parameters for FL. If both the model presentations of the current model and the previous model are considered as negative pairs, the performance of the previous model is denied and cannot converge in local data [34]. On the other hand, due to the requirement of historical local models, MOON also introduces overheads in memory. Therefore, we try to eliminate the definition of negative pairs in the model contrastive learning based FL framework.

As a new competitive contrastive SSL algorithm, Barlow Twins (BT) proposes a contrastive objective function based on the cross-correlation matrix by minimizing the redundancy between the components of the output vectors [35], and therefore, it eliminates the redundant information expression in the representation vector as much as possible. Compared to the contrastive operation in MOON, BT has the advantage of training without negative samples, and avoids the above-mentioned problems. In addition, it is more robust to the training batch size and avoids other complex implementations, such as asymmetric mechanisms and momentum encoders [36]. On the other hand, the previous contrastive learning algorithms, such as MoCo and SimCLR, build the similarity matrix in the batch dimension, while BT performs it in the feature dimension to learn a feature representation with more information, since the dimension of each feature has an independent meaning [35]. Therefore, the idea of BT has the potential to be integrated into FL to conduct model-level contrast for improving local training of individual centers, and also save memory overhead.

In this work, a novel SSL-FL framework is proposed to improve the performance of a CAD model for histopathological images. Specifically, the pseudo histopathological images are firstly generated in each center and then shared in the central server, which are fed to a specially designed multi-task SSL model to pre-train the backbone as the initial global model for further FL. The BT-based FL (FL-BT) algorithm is then developed to further effectively train this global CAD model with distributed data, which is finally applied for the diagnosis task in each center. The experimental results on four public histopathological image datasets indicate the effectiveness of the proposed SSL-based FL-BT (SSL-FL-BT).

The main contributions of this work are three-fold as follows:

1) A novel SSL-based FL framework is proposed to improve the diagnostic accuracy and generalization of a CAD model for histopathological images. Different from the conventional FL paradigm that only shares the parameters of local models in multi-center learning, we suggest to generate pseudo histopathological images from each center and then share these images to pre-train a backbone network on the central server. Thus, the specially designed

SSL can capture and learn both the inherent and specific properties of data from different centers, which is beneficial to the generalization of CAD models.

2) A dual-task SSL driven by properties of pseudo histopathological images is developed for pre-training the backbone of the CAD model. Specifically, the center classification task is designed to discriminate which center a pseudo image is generated from, and the image restoration task is applied to learn the common information of all centers. This strategy helps to capture both the specific and common inherent information from multi-center pseudo histopathological images.
3) A new FL-BT algorithm is proposed to improve the performance of the global CAD model, in which the idea of BT is innovatively integrated into the FL framework. Since FL-BT eliminates the definition of negative pairs in the model contrastive learning, it is more clear and interpretable than MOON. In particular, FL-BT compares the representations generated by the local and global models instead of the different images in the original BT, and it integrates a cross-correlation matrix-based contrastive objective function into FL to conduct the model-level contrastive learning. FL-BT minimizes the representation gaps between the local and global models to correct the local training, and therefore, it can alleviate the issue of data heterogeneity.

II. Related Work

A. SSL for CAD of Histopathological Images

Over the last years, the fast development of deep learning has made breakthroughs in the field of CAD for histopathological images [3]. According to the size of histopathological images, the current works are developed for the whole-slide images (WSI) and patches from WSIs [4], respectively. Although lots of deep learning algorithms have been proposed in this field [37], they should be further improved due to the complexity of histopathological images and a variety of cancers.

Since it is time-consuming to annotate a large number of histopathological images for CAD, SSL is a promising approach to alleviate this problem by pre-training the model under the supervision of the data itself. For example, Hu *et al.* proposed a unified generative adversarial network to learn robust cell-level representation for classification of histopathological images [38]; Stack *et al.* applied the contrastive predictive coding to histopathology datasets, indicating that the low-level features were more effective for tumor classification [39]; Ciga *et al.* utilized SimCLR to pre-train the model on multiple histopathological datasets, which improved the performances on different downstream CAD tasks [40]. All these works demonstrate the effectiveness of SSL for CAD with limited histopathological images.

It is worth noting that the application of SSL should not only retain the center-specific information, but also mine more inherent common features from the data of all centers for FL in our task. However, the single pretext task generally cannot well explore this information. To this end, the multi-task SSL has the potential to learn more comprehensive features from training samples. In the pioneering work, Koohbanani *et al*. proposed a multi-task SSL algorithm Self-Path for histopathological images, which included three pathology-specific tasks, *i.e.*, magnification prediction, magnification Jigsaw puzzle and Hematoxylin channel prediction, to improve the model performance with limited annotations [41]. Since Self-Path can achieve superior performance over the single-task based approaches, we will also specifically design a multi-task SSL according to the data characteristics of multi-center histopathological images.

B. Federated Learning

FL is an emerging distributed learning method, which aims to share the local model parameters in a parallel manner instead of the conventional local data[13]. FedAvg is the first FL algorithm that aggregates the local models by averaging the model weights [42]. Thereafter, FL has been successfully applied to many fields [14], such as financial, smart retail and healthcare, due to the advantages of both privacy-preserving and distributed optimization.

Recently, some variants of FedAvg have been proposed, which mainly include the following two methods: 1) Local training method, such as FedProx [43], SCAFFOLD [44] and MOON [31]; 2) Aggregation method, such as FedNova [45], FedMA [46], FedAvgM [47] and Auto-FedAvg [48]. FedProx introduced a proximal term in local training, which was calculated based on the Euclidean norm between the output of both current global model and local model [43]. FedBN was proposed to locally keep batch normalization parameters in order to mitigate feature variation in non-IID data [49]. MOON developed a model-level contrast learning strategy, whose key idea was to use the similarity between model outputs to rectify individual local training [31]. All these works show the effectiveness of FL for multiple center learning.

FL is particularly attractive for CAD now [19]. It can not only improve the generalization of CAD models, but also alleviate the SSS issue by collecting multi-center data with privacy protection. Some pioneering works have been conducted. For example, Li *et al.* proposed an FL algorithm for diagnosing the autism spectrum disorders with multi-site fMRI data, in which decentralized iterative optimization and randomization mechanism were used [17]; Andreux *et al.* introduced a local statistical batch normalization layer in the model architecture of FL, which was applied to the diagnosis of breast tumor with multi-centric histopathology datasets [50]; Yang *et al.* proposed an FL algorithm using partial networks for COVID-19 diagnosis with multiple X-ray datasets [18]; Adnan *et al.* applied the FL framework to WSIs on the data from TCGA, and they adopted the multiple instance learning (MIL) method for classification of WSIs by extracting multiple patches from the WSIs [51]. These works indicate that FL can effectively improve the model performance in local servers together with privacy-preserving.

However, the existing FL methods cannot sufficiently handle the gap between the local models and the central model, resulting in limited learning performance.

C. GAN in Histopathological Images

Due to the success of GAN in computer vision, it has also been widely used in different medical image tasks, such as

image reconstruction, image segmentation and lesion detection [52][53][54]. Moreover, they can help alleviate the problems of small sample size and limited annotation in medical imaging applications. For example, Madani *et al.* applied the GAN-based data augmentation for the CAD model of Chest X-ray, and achieved superior performance over the traditional augmentation strategy [23]; Adar *et al*. used conditional GANs to generate synthetic CT images to improve the performance of liver lesion classification [24]; Elmas *et al.* proposed a FL-based MRI reconstruction algorithm, which implemented cross-cite learning with generative MRI prior and the following prior adaptation to improve reconstruction performance [52]. These works demonstrate the effectiveness of GAN in different medical image processing tasks.

Some works have applied GAN to the field of histopathological image processing, such as color normalization, image enhancement and data augmentation [55][56][57]. For example, CycleGAN was effectively used for color normalization of breast histopathological images, which then eliminated the selection of representative reference slides by pathologists [55]; SRGAN was applied to simultaneously increase image resolution and reduce image noise for breast histopathological images [56]; cGAN was adopted to synthesize realistic cervical histopathology images for augmenting the training dataset so as to improve the performance of trained model [57]. All these works indicate the emerging applications of GAN for histopathological images.

Different from these previous works, we propose to adopt GAN to generate pseudo histopathological images in different centers, and then share these images in the central server for training the global model under the FL framework. This idea breaks through the limitation of traditional FL that only shares model parameters. The specially designed SSL can capture and learn both the inherent and specific properties of data based on the shared pseudo images to further improve the generalization of the global CAD model.

III. METHODOLOGY

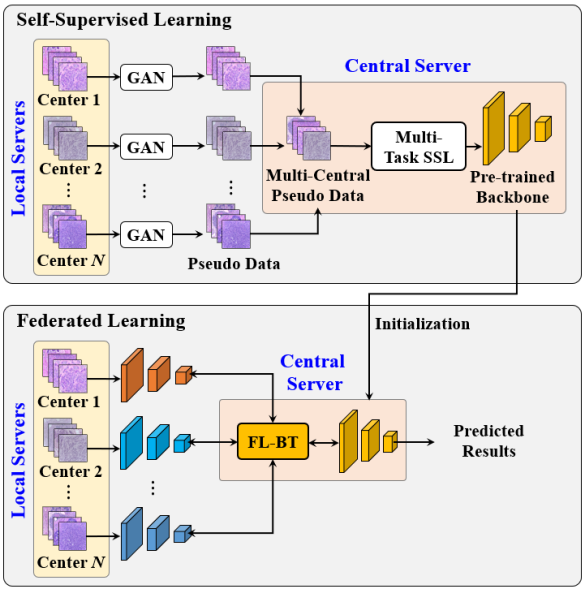


Fig. 1. The pipeline of the proposed SSL-FL-BT, which includes two stages, *i.e.*, SSL stage and FL stage. In the SSL stage, the specially designed multi-task SSL is performed on all the pseudo images to pre-train the backbone network. In the FL stage, the pre-trained backbone is used as the initialization network for the proposed FL-BT.

Fig. 1 shows the overall pipeline of the proposed SSL-FL-BT, which includes two stages, *i.e.*, SSL stage and FL stage. In the SSL stage, the pseudo histopathological images are firstly generated in each center with a GAN. The specially designed multi-task SSL is then performed on all the pseudo images to pre-train the backbone network. Here, the center classification and image restoration tasks are designed as the dual pretext tasks, and both tasks share the backbone. The pre-trained backbone is then used as the initialization network in the subsequent FL stage, and it is trained by the proposed FL-BT with multi-center real histopathological images. In the testing stage, a histopathological image is fed to the corresponding CAD model in a center for cancer diagnosis.

A. Multi-task SSL for FL

Since the stained histopathological images have inconsistencies among different centers, we propose to share the pseudo histopathological images without privacy information for FL, which can provide more heterogeneous center-specific information of each center for the CAD model, and further improve its generalization. Here, we specifically design a multi-task SSL to capture and learn both the center-specific information and common inherent representation according to the data characteristics of multi-center pseudo histopathological images. The overall pipeline of our proposed multi-task SSL is shown in Fig. 2.

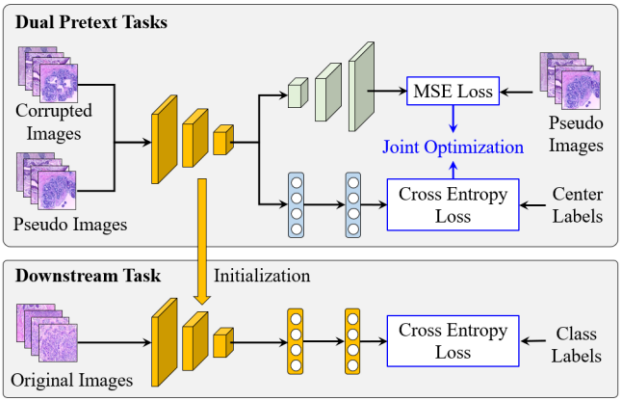


Fig. 2. The pipeline of the proposed multi-task SSL. The pseudo data are shared in the central server to pre-train the backbone network by multi-task SSL. Two pretext tasks are designed, *i.e.*, the center classification task and image restoration task.

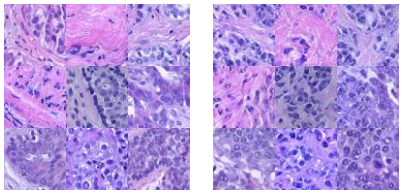
As shown in Fig. 2, the pseudo data are firstly generated in each center through the GAN model, which are then shared to the server for pre-training backbone network by SSL. Two pretext tasks are designed, *i.e.*, the center classification task and image restoration task. The former pretext task predicts which center the synthetic data belong to. It can explore more specific properties of data in each center. While the latter pretext task restores the corrupted images to their original pseudo images, which can learn more inherent information of the data collected from different centers.

Pseudo Image Generation: In order to generate high-fidelity pseudo histopathological images, the multi-scale gradient generative adversarial network (MSG-GAN)

algorithm is adopted in this work, which provides high-quality synthesized images for the following multi-task SSL [58]. In particular, each center individually trains an MSG-GAN.

MSG-GAN introduces a multiscale gradient technique that allows the gradients flow to propagate from the discriminator to the generator at multiple scales. This technique improves the stability of training for image synthesis on data with different sizes, resolutions and domains. Compared to other GANs and their variants, MSG-GAN can boost the performance in most of cases. The detailed information about the MSG-GAN can be referred to [58].

In this work, the quality of the generated images is evaluated by frechet inception distance (FID) score, which is widely used in generative models for image quality evaluation by calculating the distance between the feature vectors of real and generated images [59]. The pseudo histopathological images with small FID score will be selected for the following SSL. The real histopathological images and generated pseudo examples are shown in Fig. 3.



(a) (b)

Fig. 3. (a) Real histopathological images and their corresponding (b) Pseudo histopathological images, where the images of three rows in (a) are acquired for Center 1, Center 2, and Center 3, respectively.

Dual Pretext Tasks: Two SSL pretext tasks are developed based on the characteristics of pseudo histopathological images for pre-training backbone, *i.e.*, the center classification task and image restoration task.

The center classification task tries to predict which center a pseudo image is generated from, and thus the center identity document (ID) is considered as the label. Since the pseudo images are generated based on each center's data, these images contain center-specific information extracted from the real histopathological images of the corresponding center. Thus, this pretext task can effectively learn the heterogeneous characteristics pseudo images generated from each center.

The SSL image restoration task is applied to learn detailed contextual information [27][29], the anatomical information [30], and other inherent knowledge in histopathological images, which contain the common characteristics across centers. Specifically, we randomly swap patches in the pseudo images to generate the corrupted ones [27]. These corrupted images are then fed to the backbone network to restore the original pseudo images as ground truths.

To conduct two pretext tasks in a unified framework with a single shared network, the hard parameter sharing is utilized to construct a multi-task learning architecture [60]. In our implementation, the commonly used ResNet50 is used as the shared backbone in Fig. 2, followed by a classification branch and a reconstruction head [61].

For the center classification task, the cross-entropy (CE) loss L_{CE} is utilized for the SSL classification task, which can be given as follows:

$$L_{CE} = -\frac{1}{M}\sum_{k} \sum_{n=1}^{N} a_{kn}\log(u_{kn}) \quad (1)$$

where $a_{kn} \in \{0,1\}$ is an indicator, which takes value 1 if and only if the label of k-th sample is n, u_{kn} denotes the probability of the k-th sample coming from the n-th center, and M denotes the number of pseudo images.

For the image restoration task, the mean squared error (MSE) is adopted as the objective function for the SSL image restoration task. Given a corrupted image Q_k and the reconstruction sub-network $G(\cdot)$, the MSE loss is formulated as:

$$L_{MSE} = \frac{1}{M}\sum_{k=1}^{M} \|G(Q_k) - V_k\|^2 \quad (2)$$

where $G(Q_k)$ and V_k denote the restored image and the corresponding ground truth, respectively.

In order to effectively promote the feature representation in the shared backbone, the two tasks are simultaneously optimized with the following overall loss L_{SSL}:

$$L_{SSL} = L_{CE} + L_{MSE} \quad (3)$$

Thus, this shared backbone naturally contains both the center-specific knowledge from the center classification task and the inherent information from the image restoration task. The pre-trained backbone is then used as the initialization for the followed FL stage. This multi-task SSL strategy can capture both the specific and common inherent information from the multi-center pseudo histopathological images.

B. FL-BT for Histopathological Images

The generalization of a CAD model for histopathological images is generally limited by the training samples only from a single center, because the stained images have different data distributions in different hospitals. FL can improve both the diagnostic accuracy and generalization ability of CAD models with multi-center histopathological images, while private information can also be protected [62].

The existing FL methods cannot well handle the heterogeneity of multiple local data distribution. The MOON algorithm has been proposed to address this issue, which adopts the similarity among model representations to correct the local training of individual centers [31]. However, the robustness of MOON should be further improved, since the trivial implementations are applied in the local training process. To this end, we propose a novel FL-BT to train the global CAD model with multi-center histopathological images.

Network Architecture of Local Model: The architecture of each local network for FL-BT is shown in Fig. 4, which consists of a base encoder, a projector network, and an output layer. The base encoder is the widely used ResNet50, which learns the feature representation from input real histopathological images. The projector network is then adopted to map the representation to a feature space with a fixed dimension. Finally, the output layer is used to predict the classification results for each cancer class.

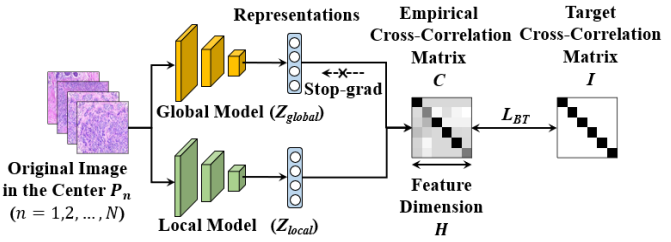


Fig. 4: Flowchart of the FL-BT algorithm. FL-BT feeds the same image into the local model and the global model, respectively, and then calculates the cross-correlation matrix of the corresponding two features. FL-BT optimizes the statistical properties that make the cross-correlation matrix tend to be the identity matrix.

Local Training: Suppose there are N centers, which are denoted as $P_1, \dots, P_N$. Center P_n has a local dataset $\mathcal{D}_n$ with histopathological images X_n and the corresponding label Y_n, $n = 1,2, \dots, N$. The proposed FL-BT aims to learn a global model W over the local dataset $\mathcal{D}_n$ with the help of a global model in the central server.

To train the model W_n in each center, the proposed FL-BT assumes that the pre-trained global model W trained by multi-task SSL is set as the initial model W_n in center P_n. The histopathological images X_n are fed to the base encoders of both global model and local model to generate representations z_{global} and z_{local}, respectively. The projector network then maps the z_{global} and z_{local} to the fixed feature dimension H. Finally, the classification result is predicted by the output layer.

We further define $F_w(\cdot)$ as the whole network, $R_w(\cdot)$ as the network before the output layer with model weight w, and t as the t-th communication round. We extract the representation of X_n from both the global model W^t (i.e., $z_{global} = R_{w^t}(X_n)$) and the local model being updated W_n^t (i.e., $z_{local} = R_{w_n^t}(X_n)$), respectively.

The local objective function contains two parts: L_{sup} and $L_{FL\text{-}BT}$. The former part L_{sup} is a cross-entropy loss in the supervised learning manner, while the second part $L_{FL\text{-}BT}$ is the contrastive loss in our proposed FL-BT.

Specifically, the supervision loss L_{sup} in FL-BT can be given as:

$$L_{sup} = CrossEntropyLoss(F_{w_n^t}(X_n), Y_n) \quad (4)$$

While the contrastive loss $L_{FL\text{-}BT}$ designed in FL-BT can be formulated as:

$$L_{FL-BT} = \sum_i (1 - C_{ii})^2 + \lambda \sum_i \sum_{j \neq i} C_{ij}^2 \quad (5)$$

where λ is the weight to trade off the importance of the first and second terms; C denotes the cross-correlation matrix computed between the outputs of the two branches along the batch dimension, which is a square matrix with size the dimensionality of the network's output; and both i and j are the vector dimensions of the network outputs. More specifically, C_{ij} can be calculated by:

$$C_{ij} = \frac{\sum_b z_{local}^{b,i} z_{global}^{b,j}}{\sqrt{\sum_b (z_{local}^{b,i})^2} \sqrt{\sum_b (z_{global}^{b,j})^2}} \quad (6)$$

where the superscript b denotes batch samples. When $i = j$, we can get C_{ii}. The loss of FL-BT includes two parts, i.e., the invariance term and the redundancy reduction term. Among them, the invariance term makes the positive examples closer to each other in the representation space, and the redundancy reduction term decorrelates the different components of the embedding vector by making the off-diagonal elements of the cross-correlation matrix to 0. This decorrelation reduces the redundancy between outputs, make the outputs only contain non-redundant information about the samples. Therefore, it eliminates the redundant information expression in the representation vector as much as possible, making FL-BT can effectively optimize the FL procedure.

The definition of the whole loss function can be given by:

$$L_n = L_{sup}(w_n^t; (X_n, Y_n)) + \mu L_{FL-BT}(w_n^t; w^t; X_n) \quad (7)$$

where μ is the factor to balance the weight of contrastive loss $L_{FL\text{-}BT}$.

The local objective is to minimize:

$$\min_{w_n^t} \mathbb{E}_{(X_n, Y_n) \sim \mathcal{D}_n} [L_{sup}(w_n^t; (X_n, Y_n)) + \mu L_{FL-BT}(w_n^t; w^t; X_n))] \quad (8)$$

In each round, the server sends the global model to the centers, receives the local model from the centers, and updates the global model using weighted averaging. In local training, each model uses stochastic gradient decent to update the parameters with the local data, the objective is shown in Eq. (8).

The model-contrastive loss compares representations learned by different models, and the contrastive loss compares representations of different images in FL-BT. It is worth noting that the conventional BT calculates the cross-correlation matrix of the two representations after inputting two views of an image into the same network, while it calculates the cross-correlation matrix between the two features of one image, which are generated from the local model and the global model, respectively.

Global Aggregation: After the local training in each center, the updated model parameters w_n, which $n = 1, \dots, N$, in local models are then sent to the central server to implement the model aggregation.

FL-BT seeks to minimize the following objective function for model training:

$$\min_w L(w) = \sum_{n=1}^N \alpha_n L_n \quad (9)$$

where N denotes the number of the centers, α_n represents the importance of the n-th center with $\sum_n \alpha_n = 1$.

In this work, we adopt the classical FedAvg as the aggregation method [42], in which w_n^t are averaged as the global model. In communication round t, the updated parameters for the global model can be formulated as the follows:

$$w^{t+1} \leftarrow \sum_{n=1}^N \frac{m_n}{M} w_n^t \quad (10)$$

where m_i denotes the number of images in center P_n, and M denotes the total number of images.

Then, the updated parameters of the global model are deployed to all the local servers for the local models, which can be formulated as:

$$\forall_n w_n^t \leftarrow w_n^t - \eta g_n \quad (11)$$

where η denotes the learning rate for model optimization and g_i denotes the gradients at each local model. The final model is obtained after several communication rounds.

The detailed scheme of FL-BT is shown in Algorithm 1.

Algorithm 1: The FL-BT framework

Input: local datasets, number of communication rounds T, number of local epochs E,number of classes, number of centers N, learning rate η

Output: The final model W^T

Server executes:
initialize w^0
for $t = 0,1,\cdots,T-1$ **do**
  for $n = 1,2,\cdots,N$ **in parallel do**
    send the global model w^t to P_n
    $w_n^t \leftarrow$ **PartyLocalTraining**(n, w^t)
  $w^{t+1} \leftarrow \sum_{n=1}^{N} \frac{m_n}{M} w_n^t$
return w^T
PartyLocalTraining(n, w^t):
$w_n^t \leftarrow w^t$
for epoch $e = 1,2,\cdots,E$ **do**
  for each batch b $=\{\boldsymbol{X_n}, \boldsymbol{Y_n}\}$ **of** $\boldsymbol{D^n}$ **do**
    $L_{sup} \leftarrow CrossEntropyLoss(F_{w_n^t}(X_n), Y_n)$
    $z \leftarrow R_{w_n^t}(X_n)$
    $z_{glob} \leftarrow R_{w^t}(X_n)$
    $L_{FL-BT} \leftarrow \sum_j (1 - \mathcal{C}_{ii})^2 + \lambda \sum_i \sum_{j \neq i} \mathcal{C}_{ij}^2$
    $L_n \leftarrow L_{sup} + \mu L_{FL-BT}$
    $w_n^t \leftarrow w_n^t - \eta g_n$
return w_n^t **to server**

Comparisons with MOON: MOON is a simple and effective approach for FL. Inspired by MOON, we propose FL-BT that performs the model-level contrastive learning in FL. The FL-BT and MOON have the following differences:

1) MOON requires negative sample pairs for model-level contrastive learning, while our FL-BT does not require negative sample pairs for model contrastive learning. Therefore, FL-BT avoids under-clustering and over-clustering problems.
2) The mathematical principle of FL-BT is different from that of MOON. MOON directly optimizes the geometric properties of feature space. It pulls the positive samples closer and pushes the negative samples farther, allowing the feature space to be clustered by classes. While FL-BT optimizes statistical properties, and favors the cross-correlation towards the identity matrix rather than the geometric properties of the feature space.
3) The similarity matrix of MOON is developed based on the batch-wise, while that of FL-BT is developed based on the sample-wise. Since each sample has its own characteristics, the sample-wise learning criterion of FL-BT helps to learn superior representations for the classification.

IV. EXPERIMENTS

A. Datasets and Data Preprocessing

The proposed algorithm was evaluated on four public breast histopathological datasets: the 2015 Bioimaging Challenge Dataset [63], the 4th Symposium in Applied Bioimaging Dataset [64], the ICIAR 2018 Grand Challenge on Breast Cancer Histology Images Dataset [65] and the Databiox Dataset [66], which were introduced as follows:

1) The 2015 Bioimaging Challenge Dataset [63]

The 2015 Bioimaging Challenge Dataset (https://rdm.inesctec.pt/dataset/nis-2017-003) includes high-resolution (2048×1536 pixels), uncompressed, and annotated hematoxylin and eosin (H&E) stained images. All the images were digitized with the magnification of 200× and pixel size of 0.42μm×0.42μm. These images were labeled by two pathologists, and the disagreement cases between pathologists were discarded.

2) The 4th Symposium in Applied Bioimaging Dataset [64]

The 4th Symposium in Applied Bioimaging Dataset has 140 high-resolution (2048×1536 pixels) annotated HE-stained images. The images were all digitized under the same acquisition conditions with a magnification of 200x. The dataset has been assembled and annotated by two pathologists. The dataset is publicly available at http://www.bioimaging2015.ineb.up.pt/challenge_overview.html.

3) The ICIAR 2018 Grand Challenge on Breast Cancer Histology Images Dataset [65]

This dataset includes microscopy images annotated by two expert pathologists. The images with divergence between normal and benign classes were then discarded. The remaining doubtful cases were confirmed via imunohistochemical analysis. The provided images had the same size of 2048 × 1536 pixels and a pixel scale of 0.42 μm × 0.42 μm. The data is publicly available from the BACH challenge website: https://iciar2018- challenge.grand-challenge.org/.

4) The Databiox Dataset [66]

The Databiox dataset is a histopathological image dataset for grading breast invasive ductal carcinoma (IDC) into three categories: grade I, grade II, and grade III. It includes 922 images in four magnification levels, i.e. 4×, 10×, 20×, and 40×. We then selected the high-magnification 40× images (131 with grade I, 180 with grade II, and 143 with grade III) for experiments, since this subset had the largest number of samples with clearer structure and morphology about tissues. Finally, after removing the surrounding non-tissue regions, all images were then cropped into the size of 2048×1536.

The detailed information about the four datasets is given in Table I. All the histopathological images were stained by hematoxylin and eosin (H&E). Fig. 5 shows some example images from four datasets.

Since the former three datasets, namely the 2015 Bioimaging Challenge Dataset, the 4th Symposium in Applied Bioimaging Dataset, the ICIAR 2018 Grand Challenge on Breast Cancer Histology Images Dataset, had the same classes, *i.e.*, normal tissues, benign lesions, in situ carcinomas, and invasive carcinomas, these datasets were then used as three centers for FL in this work, which were denoted as Center 1 (C1), Center

2 (C2) and Center 3 (C3), respectively. The Databiox dataset was adopted as an Additional Center (AC) to verify the generalization of the model, and the final global model was used as the initialization model for training.

TABLE I
DETAILED INFORMATION ABOUT THE FOUR DATASETS

Datasets	Classes/Numbers	Center
2015 Bioimaging Challenge Dataset	Normal Tissues:64 Benign Lesions:78 In Situ Carcinomas:72 Invasive Carcinomas:71	Center 1
The 4th Symposium in Applied Bioimaging Dataset	Normal Tissues:30 Benign Lesions:30 In Situ Carcinomas:30 Invasive Carcinomas:30	Center 2
ICIAR 2018 Grand Challenge	Normal Tissues:100 Benign Lesions:100 In Situ Carcinomas:100 Invasive Carcinomas:100	Center 3
Databiox Dataset	Grade I: 131 Grade II:180 Grade III:143	Additional Center

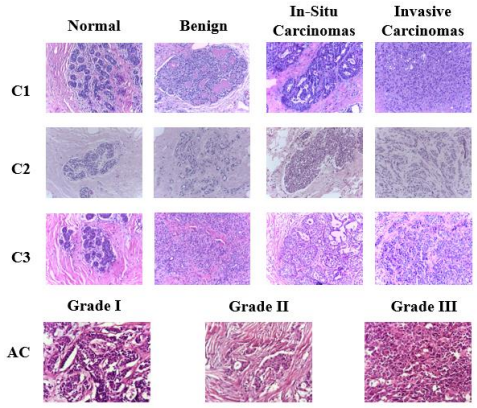


Fig. 5: The example histopathological images from four datasets.

B. Experimental Setup

To validate the effectiveness of the proposed SSL-FL-BT, the following related algorithms were compared:

1) ResNet50 [61]: The ResNet50 was directly trained only with the histopathological images from one center, which was a single-center based CAD without FL.
2) FedAvg [42]: FedAvg was selected for comparison as a classical FL algorithm, which utilized fixed weights to average the local models for the optimization of the global model.
3) FedProx [43]: FedProx was also selected for comparison as a classical FL algorithm, which added a proximal term in the aggregation method proposed by FedAvg to stabilize the convergence.
4) FedBN [49]: FedBN was compared as a stage-of-the-art FL algorithm, which aggregated the local models without sharing parameters in BN layers to obtain the global model.
5) MOON [31]: MOON was compared as the contrastive learning-based FL algorithm, which adopted contrastive learning to reduce the gaps between the local models and the global model.

It is worth noting that all the compared algorithms adopted ResNet50 as the backbone.

On the other hand, an ablation experiment was conducted to evaluate the effectiveness of the multi-task SSL in SSL-FL by comparing SSL-FL-BT with the following variants:

1) FL-BT: FL-BT directly trained the CAD models of all centers without the pre-trained backbone by multi-task SSL.
2) SSL-C-FL-BT: This variant conducted the SSL-based FL-BT, but it only designed the center classification task as the pretext task for SSL.
3) SSL-R-FL-BT: This variant also conducted the SSL-based FL-BT, but it only designed the restoration task as the pretext task for SSL.
4) SSL-Fed-R: This variant conducted a two-stage training strategy. In particular, it first performed the SSL on the original images in each center individually in the first stage, and then the pre-trained models from different centers were used as the initialized models for the followed FL training in the second stage. Here, we adopted the image restoration as the SSL pretext task.
5) SSL-Fed-S: This variant has the same training strategy as SSL-Fed-R, but we adopted a typical contrastive learning task, namely SimCLR, instead of the image restoration as the SSL pretext task.
6) FedSSL-R: This variant was conducted based on the newly proposed divergence-aware federated SSL algorithm [67]. It performs another two-stage training strategy that was different from the SSL-Fed-R. In [67], for the first stage, the federated SSL (FedSSL) was performed on the original images, which included three key steps: (1) pre-training local models in each center; (2) aggregating pre-trained model on the central server; and (3) communicating models (upload and update) between the server and centers. These three steps were iterated until the training was completed. In the second stage, these pre-trained models were used as the initialized modes for the following FL training. Here, we adopted image restoration task as the SSL pretext task.
7) FedSSL-S: This variant has the same training strategy as FedSSL-R, but we adopted a typical contrastive learning task, named SimCLR, instead of the image restoration as the SSL pretext task.
8) SSL-L-FT: This variant conducted the same self-supervised pretext task as the SSL-FL-BT in the central server with all the pseudo images, and then the pre-trained global model was fine-tuned on each center to generate the local model. It is worth noting that the local models in different centers were directly used for diagnosis, and the FL did not further perform on these local models.
9) SSL-FL: This variant conducted the same self-supervised pretext task as the SSL-FL-BT, but it performed the FedAvg instead of the FL-BT during the FL training.

In each round of FL, the updated local models of different centers were transferred to the server to further update the global model. After training the global model, each center adopted this global model for the diagnosis task. Therefore, for each algorithm, we reported the performance of the global model on the three centers as the final result.

The classification accuracy, precision, recall, and F1-score were used as evaluation indices, which were computed as follows:

$$\begin{cases} Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \\ Precision = \frac{TP}{TP+FP} \\ Recall = \frac{TP}{TP+FN} \\ F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall} \end{cases} \quad (12)$$

where TP is the number of true positive, TN is the number of true negative, FP is the number of false positive and FN is the number of false negative. The precision-recall (PR) curve and average precision (AP) were also used to evaluate the performance.

To verify the generalization of the global model, the value of global test average (GTA) was further used to quantitatively measure the generalization ability of the global model [48], which computed the average accuracy, precision, recall and F1-score values on the results of three centers, respectively.

In addition, the additional fourth dataset (Databiox Dataset) in AC was further used to verify the generalization of the global model. In particular, since the former three datasets have different disease classes from that of the fourth dataset, the global model trained through the FL paradigm under three centers was used as the pre-trained model for the last dataset. That is, the network structure before the last layer of the pre-trained model remained unchanged, and the output number of the final fully connected layer was changed from four to three. This backbone was then fine-tuned with the training set of the fourth dataset.

The resource usages, including communication and computation costs, were further evaluated for the proposed SSL-FL-BT based the Reference [69]. The number of communication rounds and the number of parameters communicated per round were calculated as the communication cost, and the number of parameters and the floating-point operations (FLOPs) were computed as the computation cost. It is worth noting that we independently calculated the resource usage of the proposed SSL-FL-BT algorithm according to the two training stages, namely the SSL stage (named SSL-FL-BT-S1) and the FL-BT stage (named SSL-FL-BT-S2).

The five-fold cross-validation strategy was applied to all algorithms. That is, three folds were used as the training set, one fold as the verification set, and the last fold as the testing set [68]. In particular, we first randomly divided the samples in each center into 5 groups, respectively, and then randomly selected one group from each center to form a new fold. In this way, we obtained the five-fold dataset across the centers, which was used for performing the 5-fold cross-validation across centers. The results were reported in the format of mean ± SD (standard deviation).

C. Implementation Details

All histopathological images were resized to 256×256 for model training. MSG-GAN was used for pseudo images generation. The learning rate for the generator and discriminator was 0.003, while the number of epochs for training was 100. Each center generated 1000 images with the size of 256×256. Data augmentation was conducted on all datasets for all algorithms, including rotation (90º, 180º, 270º) and horizontally flipping.

ResNet50 was adopted as the backbone for multi-task SSL and FL. For each FL algorithm, the model was trained for 300 rounds of 1 local epoch using a batch size of 4. The weight μ for contrastive loss in Eq. (4) was set to 0.01, while the trade-off weight λ in Eq. (7) was set to 0.005. The stochastic gradient descent (SGD) was used for the optimization of each algorithm with the learning rate of 0.001. All algorithms were implemented on Pytorch.

V. EXPERIMENTS RESULTS

A. Results on Multi-Center Datasets

Table II to IV show the results of different algorithms on C1, C2 and C3 datasets, respectively. It can be found that all the FL algorithms achieve superior performances to ResNet50, which is a single-center based approach, indicating that FL algorithms can effectively improve the performances of a CAD model with multi-center data. Moreover, the proposed SSL-FL-BT outperforms all the compared FL algorithms with statistical significance on all indices in all three datasets, while FL-BT also gets significantly improvements over the compared ResNet50, FedAvg, FedProx, FedBN, and MOON algorithms.

In particular, SSL-FL-BT achieves the best mean classification accuracy of 96.06±0.57%, precision of 96.21±0.46%, recall of 96.15±0.53%, and F1-score of 96.03±0.56% on C1. It improves at least 3.82%, 3.33%, 3.65%, and 3.73% on the corresponding indices compared to FedAvg, FedProx, FedBN, and MOON, suggesting that both multi-task SSL and FL-BT effectively improve the performance. On the other hand, it also can be observed that the proposed FL-BT achieves the second-best results, and gets the improvements of 1.18%, 1.16%, 1.14% and 1.17% on the corresponding indices, respectively, compared to other FL algorithms except for our proposed SSL-FL-BT, which demonstrates the effectiveness of BT in the proposed FL-BT.SSL-FL-BT also gains the best results of 96.66±1.86%, 97.14±1.60%, 96.66±1.86%, and 96.64±1.88% on the accuracy, precision, recall, and F1-score, respectively, on C2 dataset, which improves at least 4.16%, 3.03%, 4.16%, and 4.39%, on the corresponding indices, respectively, compared to other FL algorithms. The FL-BT again achieves the second-best performance by improving at least 1.67%, 1.25%, 1.67%, and 1.80% on the accuracy, precision, recall, and F1-score, respectively, over other compared FL algorithms. Moreover, SSL-FL-BT obtains the best results of 94.50±0.68%, 94.85±0.83%, 94.50±0.68%, and 94.46±0.69% on the accuracy, precision, recall, and F1-score, respectively, on C3 dataset, which improves 3.25%, 3.64%, 3.25% and 3.21% on the corresponding indices, respectively, compared to FedAvg, FedProx, FedBN, and MOON. Besides, The FL-BT is second to SSL-FL-BT and outperforms all the other compared FL algorithms. We can see that FL-BT improves of 1.25%, 1.81%, 1.25%, and 1.22% on the accuracy, precision, recall, and F1-score on C3. All these results suggest the effectiveness of our proposed SSL-FL-BT.

Fig. 6 shows the PR curves and the corresponding AP values for different algorithms. The proposed SSL-FL-BT achieves the best AP value of 0.9864 on C1, 0.9703 on C2, and 0.9591 on C3, which again indicates its effectiveness.

TABLE II

CLASSIFICATION RESULTS OF DIFFERENT ALGORITHMS ON C1 (UNIT: %)

Algorithms	Accuracy	Precision	Recall	F1-score
ResNet50	89.26±2.33†*	90.31±2.68†*	89.21±2.08†*	89.26±2.24†*
FedAvg	90.83±1.50†*	91.84±1.79†*	90.83±1.56†*	90.73±1.65†*
FedProx	91.50±1.65†*	92.14±1.86†*	91.75±1.55†*	91.58±1.78†*
FedBN	91.84±1.42†*	92.48±1.55†*	92.08±1.41†*	91.92±1.56†*
MOON	92.24±1.70†*	92.88±1.66†*	92.50±1.56†*	92.30±1.84†*
FL-BT	**93.42±1.33***	**94.04±1.52***	**93.64±1.22***	**93.47±1.46***
SSL-FL-BT	***96.06±0.57***	***96.21±0.46***	***96.15±0.53***	***96.03±0.56***

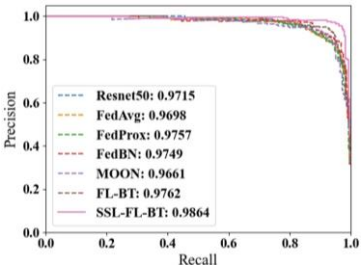
TABLE III
CLASSIFICATION RESULTS OF DIFFERENT ALGORITHMS ON C2 (UNIT: %)

Algorithms	Accuracy	Precision	Recall	F1-score
ResNet50	85.83±3.73†*	87.27±3.34†*	85.83±3.73†*	85.65±3.76†*
FedAvg	90.84±1.86†*	92.15±2.81†*	90.84±1.86†*	90.60±1.97†*
FedProx	91.67±2.95†*	93.39±2.24†*	91.67±2.95†*	91.41±3.03†*
FedBN	92.50±3.48†*	93.52±3.75†*	92.50±3.48†*	92.31±3.65†*
MOON	92.50±1.86†*	94.11±1.35†*	92.50±1.86†*	92.25±1.99†*
FL-BT	**94.17±1.86***	**95.36±1.47***	**94.17±2.28***	**94.05±2.39***
SSL-FL-BT	***96.66±1.86***	***97.14±1.60***	***96.66±1.86***	***96.64±1.88***

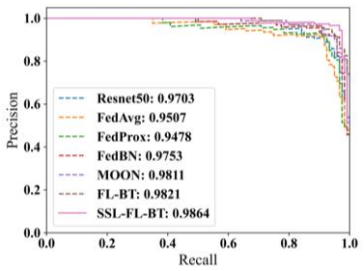
TABLE IV
CLASSIFICATION RESULTS OF DIFFERENT ALGORITHMS ON C3 (UNIT: %)

Algorithms	Accuracy	Precision	Recall	F1-score
ResNet50	89.75±2.05†*	90.45±1.88†*	89.75±2.05†*	89.74±2.14†*
FedAvg	90.00±2.65†*	91.11±2.46†*	90.00±2.65†*	90.00±2.58†*
FedProx	90.25±2.05†*	90.98±1.82†*	90.25±2.05†*	90.18±2.08†*
FedBN	90.75±2.88†*	91.54±2.50†*	90.75±2.88†*	90.78±2.77†*
MOON	91.25±1.53†*	91.21±1.37†*	91.25±1.53†*	91.25±1.41†*
FL-BT	**92.50±0.88***	**93.02±1.11***	**92.50±0.88***	**92.47±0.88***
SSL-FL-BT	***94.50±0.68***	***94.85±0.83***	***94.50±0.68***	***94.46±0.69***

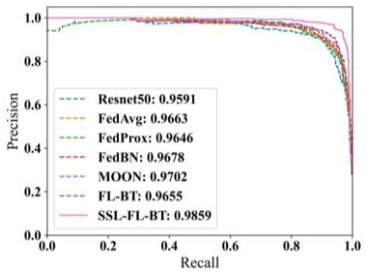
Noting: the † and * denote the improvements achieved by FL-BT and SSL-FL-BT, respectively, are statistically significant.



(a) C1



(b) C2



(c) C3

Fig. 6. PR curves of the compared algorithms with the corresponding AP values on the datasets of (a) C1, (b) C2 and (c) C3.

B. Results of Ablation Experiments

Table V to VII show the results of ablation experiments on C1, C2 and C3 datasets, respectively. These experiments evaluate the effectiveness of the performance of multi-task SSL in SSL-FL-BT compared to SSL-C-FL-BT and SSL-R-FL-BT. Both the SSL-C-FL-BT and SSL-R-FL-BT improve their performances compared to FL-BT, indicating that the single pretext task, namely center classification task or image restoration task, effectively promotes the feature representation of backbone for final classification task. Moreover, SSL-C-FL-BT achieves a little better improvement compared to SSL-R-FL-BT, suggesting that the specific and heterogeneous information provided by the center-source classification task can further assist the diagnosis. Besides, our proposed SSL-FL-BT improvs at least 1.65%, 1.22%, 1.62%, and 1.61% on the accuracy, precision, recall, and F1-score, respectively, on C1 dataset, 1.66%, 1.78%, 1.66%, and 1.72% on C2 dataset, respectively, on C2 dataset, and 0.75%, 0.55%, 0.75%, and 0.77%, respectively, on C3 dataset on the corresponding indices over other compared algorithms. It demonstrates the effectiveness of the multi-task SSL framework in FL.

Tables VIII to X show the results of SSL-FL-BT and SSL-FL compared to other SSL algorithms, including SSL-L-FT, SSL-Fed-R, SSL-Fed-S, FedSSL-R, and FedSSL-R. SSL-FL improves of 1.06%, 0.68%, 1.06%, and 1.08% on accuracy, precision, recall, and F1-score on CI in Table VIII. It promotes the classification accuracy, precision, recall, and F1-score by at least 1.67%, 0.89%, 1.67%, and 1.69% on C2, respectively, and also gets at least 1.00%, 0.93%, 1.00%, and 1.01% improvements over FedSSL-S on C3. Moreover, the SSL-FL-BT still achieves significantly improvements over SSL-FL, suggesting the effectiveness of the proposed FL-BT algorithm.

TABLE V
ABLATION EXPERIMENT RESULTS ON C1 (UNIT: %)

Algorithms	Accuracy	Precision	Recall	F1-score
FL-BT	93.42±1.33*	94.04±1.52*	93.64±1.22*	93.47±1.46*
SSL-C-FL-BT	94.41±1.55*	94.99±1.32*	94.53±1.22*	94.42±1.58*
SSL-R-FL-BT	94.16±2.29*	94.75±2.19*	94.36±2.18*	94.20±2.36*
SSL-FL-BT	**96.06±0.57**	**96.21±0.46**	**96.15±0.53**	**96.03±0.56**

TABLE VI
ABLATION EXPERIMENT RESULTS ON C2 (UNIT: %)

Algorithms	Accuracy	Precision	Recall	F1-score
FL-BT	94.17±1.86*	95.36±1.47*	94.17±2.28*	94.05±2.39*
SSL-C-FL-BT	95.00±1.86*	95.36±1.20*	95.00±1.86*	94.92±1.95*
SSL-R-FL-BT	95.00±3.48*	96.07±2.57*	95.00±3.48*	94.89±3.60*
SSL-FL-BT	**96.66±1.86**	**97.14±1.60**	**96.66±1.86**	**96.64±1.88**

TABLE VII
ABLATION EXPERIMENT RESULTS ON C3 (UNIT: %)

Algorithms	Accuracy	Precision	Recall	F1-score
FL-BT	92.50±0.88*	93.02±1.11*	92.50±0.88*	92.47±0.88*
SSL-C-FL-BT	93.75±1.53*	94.30±1.37*	93.75±1.53*	93.69±1.49*
SSL-R-FL-BT	93.50±1.85*	93.91±1.94*	93.50±1.85*	93.42±1.84*
SSL-FL-BT	**94.50±0.68**	**94.85±0.83**	**94.50±0.68**	**94.46±0.69**

TABLE VIII
ABLATION EXPERIMENT RESULTS ON C1 (UNIT: %)

Algorithms	Accuracy	Precision	Recall	F1-score
ResNet50	89.26±2.33†*	90.31±2.68†*	89.21±2.08†*	89.26±2.24†*
SSL-L-FT	92.28±2.10†*	92.81±1.73†*	92.31±2.25†*	92.06±1.90†*
FedAvg	90.83±1.50†*	91.84±1.79†*	90.83±1.56†*	90.73±1.65†*
SSL-Fed-R	91.84±1.42†*	92.48±1.55†*	92.08±1.41†*	91.92±1.56†*
SSL-Fed-S	92.58±1.26†*	93.20±1.11†*	92.83±1.19†*	92.65±1.41†*
FedSSL-R	93.55±2.29†*	94.20±2.13†*	93.75±2.19†*	93.59±2.42†*
FedSSL-S	93.94±2.04†*	94.70±1.70†*	94.03±2.03†*	93.99±2.18†*
SSL-FL	**94.94±0.98***	**95.38±0.98***	**95.09±1.02***	**95.07±1.01***
SSL-FL-BT	**96.06±0.57**	**96.21±0.46**	**96.15±0.53**	**96.03±0.56**

TABLE IX
ABLATION EXPERIMENT RESULTS ON C2 (UNIT: %)

Algorithms	Accuracy	Precision	Recall	F1-score
ResNet50	85.83±3.73†*	87.27±3.34†*	85.83±3.73†*	85.65±3.76†*
SSL-L-FT	91.67±2.95†*	93.39±2.24†*	91.67±2.95†*	91.50±3.04†*
FedAvg	90.84±1.86†*	92.15±2.81†*	90.84±1.86†*	90.60±1.97†*
SSL-Fed-R	91.67±2.95†*	92.99±3.41†*	91.67±2.95†*	91.44±3.08†*
SSL-Fed-S	92.50±3.48†*	93.52±3.75†*	92.50±3.48†*	92.31±3.65†*
FedSSL-R	92.50±1.86†*	94.29±1.20†*	92.50±1.86†*	92.30±1.95†*
FedSSL-S	93.33±2.28†*	94.82±1.47†*	93.33±2.28†*	93.18±2.39†*
SSL-FL	**95.00±1.86***	**95.71±1.61***	**95.00±1.86***	**94.87±2.08***
SSL-FL-BT	**96.66±1.86**	**97.14±1.60**	**96.66±1.86**	**96.64±1.88**

TABLE X
ABLATION EXPERIMENT RESULTS ON C3 (UNIT: %)

Algorithms	Accuracy	Precision	Recall	F1-score
ResNet50	89.75±2.05†*	90.45±1.88†*	89.75±2.05†*	89.74±2.14†*
SSL-L-FT	91.50±1.37†*	92.03±1.40†*	91.50±1.37†*	91.58±1.34†*
FedAvg	90.00±2.65†*	91.11±2.46†*	90.00±2.65†*	90.00±2.58†*
SSL-Fed-R	91.00±1.63†*	91.72±1.50†*	91.00±1.63†*	90.98±1.72†*
SSL-Fed-S	91.75±1.43†*	92.27±1.41†*	91.75±1.43†*	91.74±1.35†*
FedSSL-R	92.50±0.88†*	93.03±1.09†*	92.50±0.88†*	92.48±0.87†*
FedSSL-S	92.75±1.05†*	93.25±1.23†*	92.75±1.05†*	92.70±1.01†*
SSL-FL	**93.75±0.88***	**94.18±0.87***	**93.75±0.88***	**93.71±0.91***
SSL-FL-BT	**94.50±0.68**	**94.85±0.83**	**94.50±0.68**	**94.46±0.69**

Noting: the * denotes that SSL-FL-BT gets statistically significant improvement on this result and the † denotes that SSL-FL gets statistically significant improvement on this result.

C. Generalization and Robustness Analysis

Fig. 7 shows the GTA results of different algorithms. It can be found that all the FL algorithms achieve superior GTA values to ResNet50, suggesting that the FL strategy can effectively improve the generalization of CAD models with multi-center data. Moreover, the proposed SSL-FL-BT algorithm achieves the best GTA values with the mean accuracy of 95.74±1.43%, precision of 96.06±0.96%, recall of 95.77±1.02%, and F1-score of 95.71±1.44%. It improves at least 3.75%, 3.33%, 3.69%, and 3.78% on the corresponding indices, respectively, over FedAvg, FedProx, FedBN, and MOON. Moreover, as shown in Fig. 7, the proposed SSL-FL-BT achieves the lowest standard deviation on all indices, indicating SSL-FL-BT has better robustness and generalization.

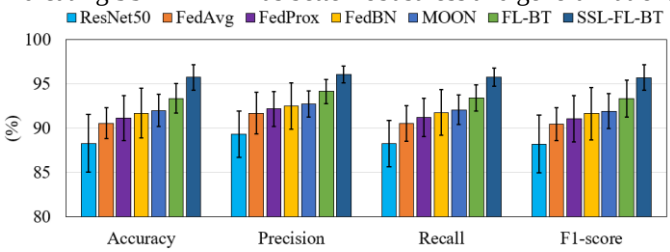


Fig. 7. Histogram chart of the GTA for the compared algorithms.

We performed additional generalization experiments, which trained the global model on C1, C2 and C3, and then fine-tuned the backbone using the Databiox Dataset on AC for generalization verification.

In Table XI, our proposed SSL-FL-BT achieves the best results by improving at least 1.27%, 0.78%, 1.10%, and 1.23% on the accuracy, precision, recall, and F1-score, respectively.

TABLE XI
CLASSIFICATION RESULTS OF DIFFERENT ALGORITHMS ON AC (UNIT: %)

Algorithms	Accuracy	Precision	Recall	F1-score
ResNet50	77.68±3.36†*	78.11±3.28†*	78.03±3.28†*	77.56±3.58†*
FedAvg	78.00±3.21†*	78.15±3.20†*	78.03±3.21†*	77.86±3.22†*
FedProx	78.62±3.48†*	79.26±3.34†*	78.68±4.18†*	78.52±3.86†*
FedBN	78.77±2.66†*	78.94±3.21†*	78.92±2.81†*	78.76±2.78†*
MOON	79.29±2.95†*	79.08±3.62†*	78.75±2.63†*	78.52±3.04†*
FL-BT	**80.21±2.59***	**80.86±3.02***	**80.42±2.46***	**80.22±2.48***
SSL-FL-BT	***81.48±2.51***	***81.64±2.41***	***81.52±2.62***	***81.45±2.53***

D. Communication and Computation Costs

As shown in Table XII, we set the number of communication rounds to 300 for all the FL-based algorithms. Since the FedAvg, FedProx, and FedBN have the same backbone network of global CAD model, the number of parameters and FLOPs of these algorithms are the same for each center. Similarly, the MOON, FL-BT, and SSL-FL-BT-S2 algorithm also have the same numbers of parameters and FLOPs for each center, which are slightly more than those of FedAvg, FedProx, and FedBN. On the hand, the number of communication parameters of MOON, FL-BT, and SSL-FL-BT-S2 are also slightly more than those of FedAvg, FedProx, and FedBN. However, the proposed FL-BT and SSL-FL-BT achieve superior performance over other compared algorithms.

TABLE XII
COMMUNICATION AND COMPUTATION COST OF DIFFERENT ALGORITHMS

Algorithms	# of Comm Rounds	# of Comm Params	Params	FLOPs
ResNet50	0	0	23.52M	5.397G
FedAvg	300	70.55M	23.52M	5.397G
FedProx	300	70.55M	23.52M	5.397G
FedBN	300	70.39M	23.52M	5.397G
MOON	300	83.89M	27.97M	5.401G
FL-BT	300	83.89M	27.97M	5.401G
SSL-FL-BT-S1	0	0	37.64M	13.932G
SSL-FL-BT-S2	300	83.89M	27.97M	5.401G

VI. Discussion

In this work, a novel SSL-FL-BT framework is proposed to promote both the diagnostic accuracy and generalization ability of the CAD model for histopathological images. The experimental results on four public datasets have validated the effectiveness of the proposed SSL-FL-BT.

The immense diversities of straining result in the inconsistencies of histopathological images across different hospitals. It then degrades the generalization ability of the CAD model, if this model is only trained with the data acquired from a single center. On the other hand, it is generally time-consuming and expensive to collect large amounts of annotated data in one center, and therefore, the SSS problem is common in the field of CAD. The FL-based multi-center learning is an effective way to alleviate both issues, and it is more feasible to meet the clinical requirement than the single-center based approach for improving both the diagnostic accuracy and generalization ability.

However, existing FL paradigm only shares the model parameters of different centers, and it cannot guarantee that the distributed CAD models well capture the specific properties of data from different centers. Therefore, we break through this limitation by sharing not only the model parameters in the central server, but also the pseudo histopathological images generated from each center, because they contain inherent and specific properties corresponding to the real images in this center, but do not include the privacy information. Therefore, these pseudo images can be shared in the central server.

For the pseudo histopathological images without disease labels, the SSL is a feasible and effective solution to explore the inherent feature representation from the unlabeled data. Thus, two data-driven SSL pretext tasks are then designed based on the characteristics of pseudo images for pre-training the backbone. Since the pretext dual tasks are optimized simultaneously to improve the overall performance of the model, the shared backbone of the model naturally contains both the center-specific knowledge generated by the center classification task and the inherent comment information generated by the image restoration task. In the ablation experiments, the SSL-L-FT outperforms the ResNet50 and FedAvg, indicating that the pre-trained network by SSL can effectively improve the performance of the followed downstream classification task. Moreover, the classification task can better promote the performance compared to the restoration task. It seems that the center-specific information is more helpful for the generalization of a CAD model. Besides, the combination of two tasks achieves significant improvement compared with the single task, demonstrating the effectiveness of dual-task SSL driven by the properties of multi-center data themselves.

As shown in Table VIII to X, both the FedSSL-R and FedSSL-S algorithms achieve superior performance over the corresponding SSL-Fed-R and SSL-Fed-S. In the first stage, the FedSSL-based variants conduct the SSL in the FL framework, and thus well train the initial models of different centers for the following FL, while the SSL-Fed-based variants only implement SSL once to initialize the model of each center for the following FL. Therefore, the FedSSL-based approaches can learn more information from other centers than the SSL-Fed-based variants to initialize the model. Moreover, although the proposed SSL-FL and SSL-FL-BT also only implement the SSL once, they still outperform both the FedSSL- and SSL-Fed-based variants, because of the following two reasons: 1) the SSL is directly conducted in the central server on all the pseudo images from different centers; and 2) the proposed dual-task SSL further promotes the model to learn the common representations and specific properties. On the other hand, we conduct the centralized SSL on the pseudo images before the proposed FL-BT process, since this manner is more efficient. In particular, the SSL is only conducted once on the server before the following FL process. Therefore, our proposed SSL strategy can effectively reduce the computational complexity, but still achieves superior performance.

In addition, we propose an effective algorithm, namely FL-BT, to improve the classification performance of local training. FL-BT utilizes the similarity between model representations to minimize the representation redundancy of the local model, which benefits the optimization of the global model in the FL procedure. Since our FL-BT does not require negative samples, it not only performs more clear and interpretable model contrastive learning than MOON, but also avoids these problems caused by negative pairs in MOON. Moreover, the learning criterion of FL-BT tries to obtain a feature representation that contains more information, so that the dimension of each feature preferably has an independent meaning. On the other hand, as shown in Eq. (5), the loss function in FL-BT consists of two parts, *i.e.*, the invariance term and redundancy reduction term. The former plays a role in bringing the positive examples closer to each other in the representation space, while the latter enhances the independence of each element of the vector. Compared with the conventional contrastive SSL algorithms, which require positive and negative samples to conduct contrastive learning, the proposed FL-BT eliminates the redundant information expression in the representation vector as much as possible. Therefore, FL-BT achieves superior performance.

As shown in Table XII, the proposed SSL-FL-BT has an additional computation overhead compared to FL-BT without SSL. We think that this overhead is acceptable, since the SSL training only performs once on the central server, which has sufficient computing resources. Moreover, during testing, the trained ResNet50 backbone network in SSL-FL-BT has a similar computation cost to other compared algorithms. In fact, we can use a more lightweight network as the backbone of the CAD model in practical applications, which can further reduce the computational cost.

Although the results on four public datasets indicate the superior performance of the proposed SSL-FL-BT framework, it still has room for improvement. For example, the proposed framework performs the patch-level diagnosis for histopathological images in this work, while the WSI-based CAD has attracted considerable attention in recent years, which is more difficult due to the SSS problem. In fact, MIL is a commonly used method for the classification task of WSIs. Specifically, each WSI is regarded as a bag and the numerous cropped patches in this WSI are used as instances. Therefore, MIL is also the patch-based method for WSIs [51]. Consequently, it is also feasible to further extend the proposed

framework to the MIL-based CAD for WSIs, which is our future work. On the other hand, the current pseudo-data based SSL is performed on the central server, and it can be further improved with the SSL training manner by combining both the central server and distributed centers.

VII. Conclusion

In this work, a novel pseudo-data based SSL-FL-BT framework is proposed to improve both the diagnostic accuracy and generalization of the CAD model for histopathological images. The self-generated pseudo images contain inherent and center-specific properties corresponding to the real histopathological images of each center without privacy information, while the self-designed multi-task SSL captures both the representation from these pseudo images for the pre-trained backbone network. The experimental results on four public histopathological image datasets indicate the effectiveness of the proposed SSL-FL-BT.

References

[1] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE Rev. Biomed. Eng.*, vol. 2, pp. 147-171, 2009.

[2] M. Veta, J. P. Pluim, P. J. Van Diest, and M. A. Viergever, "Breast cancer histopathology image analysis: A review," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 5, pp. 1400-1411, 2014.

[3] S. Deng, X. Zhang, W. Yan, E. I.-C. Chang, Y. Fan, M. Lai, and Y. Xu, "Deep learning in digital pathology image analysis: A survey," *Front. Med.*, vol. 14, no. 4, pp. 470-487, 2020.

[4] C. L. Srinidhi, O. Ciga, and A. L. Martel, "Deep neural network models for computational histopathology: A survey," *Med. Image Anal.*, vol. 67, p. 101813, 2021.

[5] J. Shi, X. Zheng, J. Wu, B. Gong, Q. Zhang, and S. Ying, "Quaternion Grassmann average network for learning representation of histopathological image," *Pattern Recognit.*, vol. 89, pp. 67-76, 2019.

[6] J. Noorbakhsh, S. Farahmand, A. Foroughi pour, S. Namburi, D. Caruana, D. Rimm, M. Soltanieh-ha, K. Zarringhalam, and J. H. Chuang, "Deep learning-based cross-classifications reveal conserved spatial behaviors within tumor histological images," *Nat. Commun.*, vol. 11, no. 1, p. 6367, 2020.

[7] Z. Gao, J. Shi, and J. Wang, "GQ-GCN: Group quadratic graph cnvolutional network for classification of histopathological images," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv. (MICCAI)*, pp. 121-131, 2021.

[8] Z. Gao, Z. Lu, J. Wang, S. Ying, and J. Shi, "A convolutional neural network and graph convolutional network based framework for classification of breast histopathological images," *IEEE J. Biomed. Health. Inf.*, pp. 1-11, 2022.

[9] F. M. Howard, J. Dolezal, S. Kochanny, J. Schulte, H. Chen, L. Heij, *et al.*, "The impact of site-specific digital histology signatures on deep learning model accuracy and bias," *Nat. Commun.*, vol. 12, no. 1, p. 4423, 2021.

[10] Z. Huang, H. Lei, G. Chen, H. Li, C. Li, W. Gao, Y. Chen, Y. Wang, H. Xu, G. Ma, and B. Lei, "Multi-center sparse learning and decision fusion for automatic COVID-19 diagnosis," *Appl. Soft Comput.*, vol. 115, p. 108088, 2022.

[11] J. N. Kather, J. Krisam, P. Charoentong, T. Luedde, E. Herpel, C.-A. Weis, T. Gaiser, A. Marx, N. A. Valous, D. Ferber, L. Jansen, C. C. Reyes-Aldasoro, I. Zörnig, D. Jäger, H. Brenner, J. Chang-Claude, M. Hoffmeister, and N. Halama, "Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multi-center study," *PLoS Med.*, vol. 16, no. 1, p. e1002730, 2019.

[12] H. Pinckaers, B. van Ginneken, and G. Litjens, "Streaming convolutional neural networks for end-to-end learning with multi-megapixel images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1581-1590, 2022.

[13] M. J. Sheller, B. Edwards, G. A. Reina, J. Martin, S. Pati, A. Kotrotsou, *et al.*, "Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data," *Sci. Rep.*, vol. 10, no. 1, pp. 1-12, 2020.

[14] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, pp. 1-19, 2019.

[15] M. Alazab, S. P. RM, P. M, P. K. R. Maddikunta, T. R. Gadekallu, and Q.-V. Pham, "Federated learning for cybersecurity: Concepts, challenges, and future directions," *IEEE Trans. Ind. Inf.*, vol. 18, no. 5, pp. 3501-3509, 2022.

[16] D. C. Nguyen, Q.-V. Pham, P. N. Pathirana, M. Ding, A. Seneviratne, Z. Lin, O. Dobre, and W.-J. Hwang, "Federated learning for smart healthcare: A survey," *ACM Comput. Surv.*, vol. 55, no. 3, pp.1-37, 2022.

[17] X. Li, Y. Gu, N. Dvornek, L. H. Staib, P. Ventola, and J. S. Duncan, "Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results," *Med. Image Anal.*, vol. 65, p. 101765, 2020.

[18] Q. Yang, J. Zhang, W. Hao, G. P. Spell, and L. Carin, "FLOP: Federated learning on medical datasets using partial networks," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 3845-3853.

[19] M. Y. Lu, R. J. Chen, D. Kong, J. Lipkova, R. Singh, D. F. K. Williamson, *et al.*, "Federated learning for computational pathology on gigapixel whole slide images," *Med. Image Anal.*, vol. 76, p. 102298, 2022.

[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2014, vol. 27.

[21] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 4401-4410.

[22] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 8110-8119.

[23] M. Moradi, A. Madani, A. Karargyris, and T. F. Syeda-Mahmood, "Chest x-ray generation and data augmentation for cardiovascular abnormality classification," in *Medical Imaging 2018: Image Processing*, 2018, vol.10574, p.105741M.

[24] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018.

[25] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4037-4058, 2021.

[26] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE Trans. Knowl. Data Eng.*, pp. 1-20, 2021.

[27] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Self-supervised learning for medical image analysis using image context restoration," *Med. Image Anal.*, vol. 58, p. 101539, 2019.

[28] X. Chen, L. Yao, T. Zhou, J. Dong, and Y. Zhang, "Momentum contrastive learning for few-shot COVID-19 diagnosis from chest CT images," *Pattern Recognit.*, vol. 113, p. 107826, 2021.

[29] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: feature learning by inpainting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2016, pp. 2536–2544.

[30] X. Tao, Y. Li, W. Zhou, K. Ma, and Y. Zheng, "Revisiting Rubik's cube: self-supervised learning with volume-wise transformation for 3D medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv. (MICCAI)*, pp. 238-248, 2020.

[31] Q. Li, B. He, and D. Song, "Model-contrastive federated learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 10713-10722.

[32] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 1597-1607.

[33] G. Wang, K. Wang, G. Wang, P. H. S. Torr, and L. Lin, "Solving Inefficiency of Self-supervised Representation Learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9485–9495.

[34] X. Zeng, T. Zhou, Z. Bao, H. Zhao, L. Chen, X. Wang, and F. Wang, "Feature-contrastive graph federated learning: responsible AI in graph information analysis," *IEEE Trans.Comput. Soc. Syst.*, 2022.

[35] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 12310-12320.

[36] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 9729-9738.

[37] V. Cheplygina, M. de Bruijne, and J. P. Pluim, "Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Med. Image Anal.*, vol. 54, pp. 280-296, 2019.

[38] B. Hu, Y. Tang, I. Eric, C. Chang, Y. Fan, M. Lai, *et al*., "Unsupervised learning for cell-level visual representation in histopathology images with generative adversarial networks," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 3, pp. 1316-1328, 2018.

[39] K. Stacke, C. Lundström, and G. Eilertsen, "Evaluation of contrastive predictive coding for histopathology applications," *in Proceedings of the Machine Learning for Health, NeurIPS Workshop*, 2020, pp. 328-340.

[40] O. Ciga, T. Xu, and A. L. Martel, "Self supervised contrastive learning for digital histopathology," *Mach. Learn. Appl.*, vol. 7, p. 100198, 2022.

[41] N. A. Koohbanani, B. Unnikrishnan, S. A. Khurram, P. Krishnaswamy, and N. Rajpoot, "Self-path: Self-supervision for classification of pathology images with limited annotations," *IEEE Trans. Med. Imaging*, vol. 40, no. 10, pp. 2845-2856, 2021.

[42] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017, vol. 54, pp. 1273-1282.

[43] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proceedings of Machine Learning and Systems (MLSys)*, vol. 2, pp. 429-450, 2020.

[44] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic controlled averaging for federated learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 5132-5143.

[45] J. Wang, Q. Liu, H. Liang, and G. Joshi, "Tackling the objective inconsistency problem in heterogeneous federated optimization," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020, vol. 33, pp. 7611–7623.

[46] H. Wang, M. Yurochkin, Y. Sun, D. Papailiopoulos, and Y. Khazaeni, "Federated learning with matched averaging," in *International Conference on Learning Representations (ICLR)*, 2020.

[47] T.-M. H. Hsu, H. Qi, and M. Brown, "Measuring the effects of non-identical data distribution for federated visual classification," *arXiv preprint, arXiv:* 1909.06335, 2019.

[48] Y. Xia, D. Yang, W. Li, A. Myronenko, D. Xu, H. Obinata, *et al*., "Auto-FedAvg: Learnable federated averaging for multi-institutional medical image segmentation," *arXiv preprint, arXiv:* 2104.10195, 2021.

[49] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "FedBN: Federated learning on non-iid features via local batch normalization," in *International Conference on Learning Representations (ICLR)*, 2021.

[50] M. Andreux, J. O. du Terrail, C. Beguier, and E. W. Tramel, "Siloed federated learning for multi-centric histopathology datasets," in *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, Cham, 2020, pp. 129-139.

[51] M. Adnan, S. Kalra, J. C. Cresswell, G. W. Taylor, and H. R. Tizhoosh, "Federated learning and differential privacy for medical image analysis," *Sci Rep*, vol. 12, no. 1, p. 1953, 2022.

[52] G. Elmas, S. U. Dar, Y. Korkmaz, E. Ceyani, B. Susam, M. Ozbey, S. Avestimehr, and T. Cukur, "Federated learning of generative image priors for MRI reconstruction," *IEEE Trans. Med. Imaging*, 2022.

[53] F. Mahmood, D. Borders, R. J. Chen, G. N. Mckay, K. J. Salimian, A. Baras, and N. J. Durr, "Deep adversarial training for multi-organ nuclei segmentation in histopathology images," *IEEE Trans. Med. Imaging*, vol. 39, no. 11, pp. 3257–3267, 2020.

[54] A. Ben-Cohen, E. Klang, S. P. Raskin, S. Soffer, S. Ben-Haim, E. Konen, M. M. Amitai, and H. Greenspan, "Cross-modality synthesis from CT to PET using FCN and GAN networks for improved automated lesion detection," *Eng. Appl. Artif. Intell.*, vol. 78, pp. 186–194, 2019.

[55] M. T. Shaban, C. Baur, N. Navab, and S. Albarqouni, "Staingan: stain style transfer for digital histological images," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI)*, 2019, pp. 953–956.

[56] G. Çelik and M. F. Talu, "Resizing and cleaning of histopathological images using generative adversarial networks," *Physica A*, vol. 554, p. 122652, 2020.

[57] Y. Xue, Q. Zhou, J. Ye, L. R. Long, S. Antani, C. Cornwell, *et al*., "Synthetic augmentation and feature-based filtering for improved cervical histopathology image classification," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv. (MICCAI)*, pp. 387–396, 2019.

[58] A. Karnewar and O. Wang, "Multi-scale gradients for generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 7799-7808.

[59] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in Advances in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, vol. 30.

[60] Y. Zhang and Q. Yang, "A survey on multi-task learning" *IEEE Trans. Knowl. Data Eng.*, pp. 1-20, 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)*, 2016, pp. 770-778.

[62] M. Adnan, S. Kalra, J. C. Cresswell, G. W. Taylor, and H. R. Tizhoosh, "Federated learning and differential privacy for medical image analysis," *Sci. Rep.*, vol. 12, no. 1, p. 1953, 2022.

[63] T. Araújo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polónia, *et al*., "Classification of breast cancer histology images using convolutional Neural Networks," *PLoS One*, vol. 12, no. 6, p. e0177544, 2017.

[64] I. Fondón, A. Sarmiento, A. I. García, M. Silvestre, C. Eloy, A. Polónia, *et al*., "Automatic classification of tissue malignancy for breast carcinoma diagnosis," *Comput. Biol. Med.*, vol. 96, pp. 41-51, 2018.

[65] G. Aresta, T. Araújo, S. Kwok, S. S. Chennamsetty, M. Safwan, V. Alex, *et al*., "BACH: Grand challenge on breast cancer histology images," *Med. Image Anal.*, vol. 56, pp. 122-139, 2019.

[66] H. Bolhasani, E. Amjadi, M. Tabatabaeian, and S. J. Jassbi, "A histopathological image dataset for grading breast invasive ductal carcinomas," *Inf. Med. Unlocked*, vol. 19, p. 100341, 2020.

[67] W. Zhuang, Y. Wen, and S. Zhang, "Divergence-aware federated self-supervised learning," in *International Conference on Learning Representations (ICLR)*, 2022.

[68] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nat. Biomed. Eng.*, vol. 5, no. 6, pp. 555–570, 2021.

[69] Y. Tan, G. Long, L. Liu, T. Zhou, Q. Lu, J. Jiang, and C. Zhang, "FedProto: Federated Prototype Learning across Heterogeneous Clients," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, pp. 8432–8440.

[70] J. Nguyen, J. Wang, K. Malik, M. Sanjabi, and M. Rabbat, "Where to begin? on the impact of pre-training and initialization in federated learning," in *International Conference on Learning Representations (ICLR)*, 2023.

[71] H.Y. Chen, C.H. Tu, Z. Li, H.W. Shen, and W.L. Chao, "On the importance and applicability of pre-training for federated learning," in *International Conference on Learning Representations (ICLR)*, 2023.