A Federated Learning System for Histopathology Image Analysis with an Orchestral Stain-Normalization GAN

Yiqing Shen, Arcot Sowmya, Yulin Luo, Xiaoyao Liang, Dinggang Shen, Jing Ke

***Abstract*— Currently, data-driven based machine learning is considered one of the best choices in clinical pathology analysis, and its success is subject to the sufficiency of digitized slides, particularly those with deep annotations. Although centralized training on a large data set may be more reliable and more generalized, the slides to the examination are more often than not collected from many distributed medical institutes. This brings its own challenges, and the most important is the assurance of privacy and security of incoming data samples. In the discipline of histopathology image, the universal stain-variation issue adds to the difficulty of an automatic system as different clinical institutions provide distinct stain styles. To address these two important challenges in AI-based histopathology diagnoses, this work proposes a novel conditional Generative Adversarial Network (GAN) with one orchestration generator and multiple distributed discriminators, to cope with multiple-client based stain-style normalization. Implemented within a Federated Learning (FL) paradigm, this framework well preserves data privacy and security. Additionally, the training consistency and stability of the distributed system are further enhanced by a novel temporal self-distillation regularization scheme. Empirically, on large cohorts of histopathology datasets as a benchmark, the proposed model matches the performance of conventional centralized learning very closely. It also outperforms state-of-the-art stain-style transfer methods on the downstream Federated Learning image classification task, with an accuracy increase of over 20.0% in comparison to the baseline classification model.**

***Index Terms*— Histopathology, Federated Learning, Generative Adversarial Network, Stain Normalization.**

Yiqing Shen is with School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: shenyq@sjtu.edu.cn).

Arcot Sowmya is with School of Computer Science and Engineering, University of New South Wales, Sydney, NSW 2052, Australia (e-mail: a.sowmya@unsw.edu.au).

Yulin Luo is with School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: lyl010221@sjtu.edu.cn).

Xiaoyao Liang is with School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China, (e-mail: liang-xy@cs.sjtu.edu.cn).

Dinggang Shen is with School of Biomedical Engineering, ShanghaiTech University, Shanghai, 201210, China. He is also with Shanghai United Imaging Intelligence Co., Ltd., Shanghai, 200230, China, and Shanghai Clinical Research and Trial Center, Shanghai, 201210, China. (e-mail: dgshen@shanghaitech.edu.cn).

Jing Ke is with School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China. She is also with School of Computer Science and Engineering, University of New South Wales, Sydney, NSW 2052, Australia (e-mail: kejing@sjtu.edu.cn).

(Corresponding author: Jing Ke. E-mail: kejing@sjtu.edu.cn)

I. Introduction

OBJECTIVE and efficient diagnosis in cancer care are much in demand due to the rise in the global cancer burden. Clinically, histopathology examination is widely considered as the gold standard in cancer diagnosis, including but not limited to malignancy detection, tissue classification, and region of interest segmentation. Recently, much research and application in computational pathology have been powered by Artificial Intelligence (AI) [1]. Yet a huge amount of histopathology data remains a prerequisite for robust deep learning models, by and large from multiple clinical centers globally. However, the data-sharing requirement for the indispensable intra-tumor heterogeneity has largely hindered the co-training of cross-sectional data samples.

Fortunately, the advent of Federated Learning (FL), as a collaborative learning method, well meets this requirement. FL enables multiple decentralized data to be trained in a distributed fashion to satisfy the data-privacy regulations [2]. Typically, a central server aggregates local parameter updates while leaving all the training images to reside at their local sites without data exposure [2]. Conversely, traditional centralized machine learning refers to the learning paradigm that data instances from multiple clients are gathered together in one location to train a diagnostic model. Distinguished by its privacy preservation and legal guarantee, together with the competent model performance, FL has drawn increasing interest for its mitigation of the systemic privacy risks from traditional centralized learning. Originally proposed for mobile and edge device use cases, FL now has drawn attention to many medical image applications, including computed tomography (CT), or magnetic resonance imaging (MRI) [3]–[7].

However, apart from the scale and privacy problems as also in other medical disciplines, DL based tumor identification in histology images faces an additional challenge, namely stain heterogeneity. Staining is an indispensable pre-stage of histology analysis, e.g., Hematoxylin and Eosin (H&E), a most widely-used staining protocol by its superiority of morphological recognition of information between tissues and nuclei [8]. The equipment manufacturers and consequently the staining protocol used, may vary to each hospital, among pathologists, or even between batches of slides. As a result, stain variations are universal in histology. Empirically, histopathology diagnosis suffers from stain heterogeneity when the analysis is conducted by humans, and particularly severe for automatic

systems [9]. Literature reports [10] that without stain normalization, [1] a sharp decline is often observed in test accuracy, *i.e.*more than 20% [10].

Conventional stain normalization methods have given a boost in diagnostic performance by transferring heterogeneous stain styles into an aligned target style, but the defects are also observable and non-privacy-guaranteed [11]. Recently, the advent of Generative Adversarial Networks (GANs) [12] enable end-to-end pure learning-based stain normalization, which has gradually replaced traditional approaches [13], [14]. However, one shortcoming is that GANs are restricted to learning from a specific stain style distribution, which requires the collection of all samples in one location. Therefore, the data-privacy gap can not be closed in the context of FL, where data instances are characterized as distributed, diverse and private [11]. Another challenge that hinders the wide implementation of the cGAN is its difficulty in training. Specifically, the unstable training dynamics of GAN in the FL setting are characterized by strong sensitivity to hyper-parameters [15]. Moreover, the GAN training may cause model divergence and eventually mode collapse, and finally present poor visual quality and ruined tissue structures [11], [14], [16].

To address the problems discussed, a novel conditional GAN (cGAN) framework is proposed which targets to alleviate stain or eliminate variation issues in FL settings. The major contributions of this paper are summarized as follows:

1) A novel framework of a single orchestrating generator with multiple distributed discriminators is proposed for histopathology image normalization. The orchestrating server generator progressively transfers multiple heterogeneous stain styles to construct an optimized interpolated distribution mode, and eventually, achieves aligned stain normalization for all data centers. The multiple client sites, each taken as a distributed discriminator, are applied with non-independently-identically distributed (non-i.i.d.) data. [2]
2) A novel temporal self-distillation scheme is designed to improve the generalization and stabilization, which mitigates the parameter divergence of the orchestrating generator. Apart from valid prevention of potential mode collapse by imposing a consistent regularization, the model also lessens the complication in training a distributed normalization-based cGAN framework.

Empirically, the proposed FL framework reaches comparable performance with prevalent data-public centralized learning methods, evaluated with the downstream classification task for cancer diagnosis. Additionally, the temporal self-distillation regularization achieves guidance in stable training dynamics and higher training performance without a heavy workload in hyper-parameter tuning. Our work fills the gap in histopathology stain-style transfer within a data-private setting. As far as can be ascertained, it makes the first attempt at FL-based histopathology image classification, along with decentralized learning-based stain normalization pre-processing.

[1]The 'stain normalization' is also known as 'color normalization', hence the two terms are used interchangeably in some literature works.

[2]This paper is an extension of [11] in the temporal self-distillation scheme and supportive experiments.

II. Related Work

A. Stain Normalization Approaches

Literature has many approaches to overcoming the color-variation problem in histopathology. Conventional methods worked for some years before DL yet with many restrictions, such as color-matching based approaches [17] generally suffer from improper color mapping, stain-separation models [18], [19] lead to improper normalization due to the absence of tissue spatial feature extraction. Moreover, these traditional approaches require prior knowledge of the whole data set to select appropriate template images. Lately, machine learning models have been introduced to overcome these issues, among which the GAN-based transfer-style models are most prevalent in a line of research on adversarial training. For example, conditional GANs [20] transfer initial stains to target style of one particular set [21], 'pix2pix' cGAN follows to constrain 'non-biological variations' in histopathology images [22].

However, the discussed GANs often suffer from distressful mode collapse issues *i.e.*, generated images repeat very similar patterns that are not shown in input images. Consequently, more dedicated GAN architectures have been designed such as StainGAN, an improved version of CycleGan [23] [24], to preserve the original tissue structure in normalization. Nevertheless, it is restricted to learning from a single client among multiple clients, and besides, prone to encounter performance problems when datasets are many. Innovatively, in this work, an interpolated stain-style algorithm is developed to preserve accuracy in the decentralized setting [11] for FL-based scenarios.

B. Federated Learning in Biomedical Images

Although research on privacy-preserving data analysis occurred over fifty years ago, only in the past decade have explicitly and widely deployed solutions been applied to real-world applications. FL enables multiple distributed clients to collaboratively solve a machine learning problem, orchestrated by a central server. In this FL scenario, raw data is stored locally per client and well preserved privately, and these clients provide local parameters updates to the central server for model aggregation [25]. Notable attention has been paid to FL systems in the context of biomedical images *e.g.*, CT, MRI, or X-ray images. For example, FL frameworks to segment brain tumors on the BraTS dataset show the comparable performance of FL and centralized learning for MRI images, only at the cost of longer training time [3], [4] proposes. [26] provide empirical results on COVID-19 X-ray images with FL models, with very close performance compared with regular centralized training. However, conventional radiology images do not require the extra color style transfer procedures, so the gap remains unfilled in the histopathology image analysis with FL [27], [28]. Note that the extra extensive computational overheads come along with the regularization strategies [29], and their structures are not applicable for the FL setting. Consequently, a novel self-distillation method is designed in this paper to impose training consistency regularization and stabilize the training of the generator.

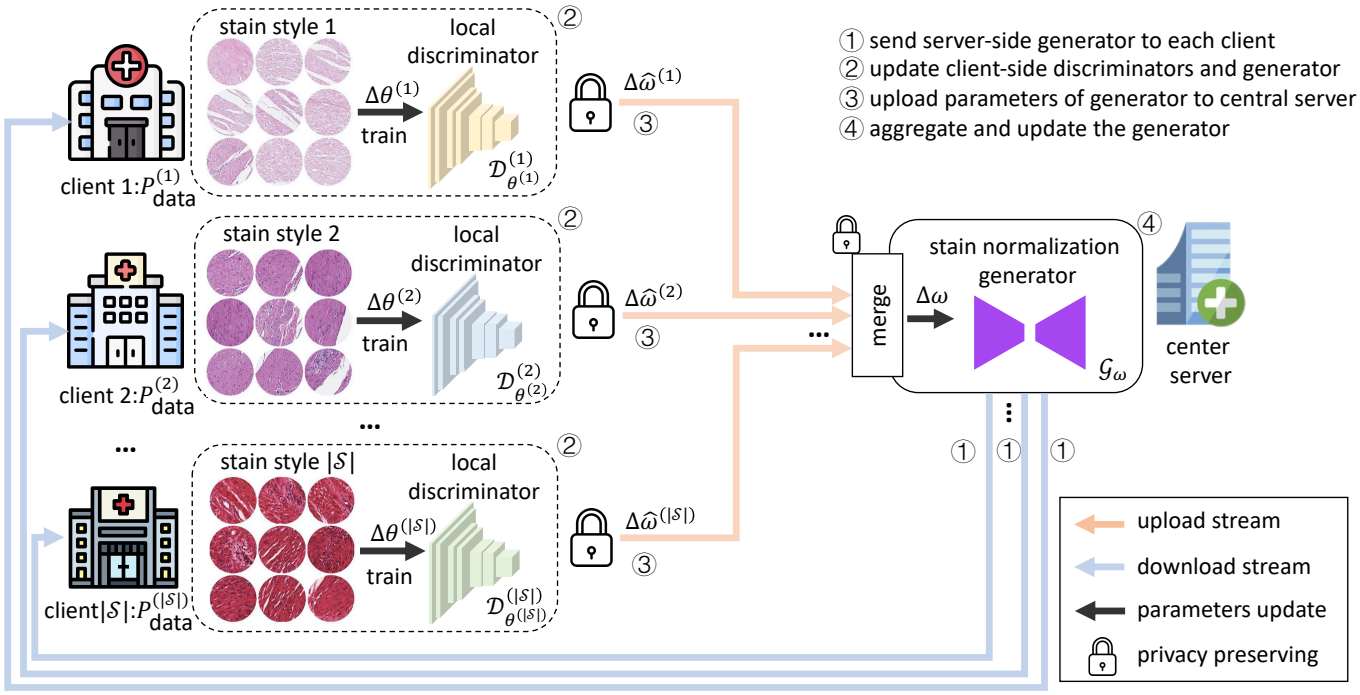

Fig. 1. The overall pipeline of the proposed one-orchestration-distributed-discriminator cGAN for stain normalization in a federated learning paradigm. Each client independently downloads the generator from the central server to be trained locally. The discriminators in each client remain locally while the generator is uploaded to the central server for aggregation.

III. Methodology

A. Problem Formulation and Assumptions

The work is under the assumption that the stain distribution of each client is independently and identically distributed (i.i.d.), and it is stain-heterogeneity or morphology-phenotype irrelevant. In other words, the coincidence of two individual clients with the same style is excluded. Meanwhile, images from different clients are non-i.i.d. Hence, a one-to-one mapping from a stain style to a client can be constructed straightforwardly i.e., stain style s can represent a unique style, as well as the client index. We write the collection of all the heterogeneous styles as $\mathcal{S}$, which is fixed for training and inference. We denote the data images curated in one client $s \in \mathcal{S}$ as $\{\mathbf{x}_i^{(s)}\}_{i=1}^{n_s}$, where n_s is the total number of images and its subscript stands for the client index. Notably, this assumption always holds as stain normalization can be performed locally at each client. In real practice, where one client (typically a hospital or one medical lab) usually follows the same staining protocols and operates on the same scanner, where variations can be alleviated to the largest extent and thus results in a very close effect to our i.i.d. assumption. In this case, stain normalization is not necessary, and our framework is not subject to the normalization process in clinical practice. We employ it to prove the effectiveness of the diagnostic task. Generally, local stain normalization requires light computation, e.g. 100K patches of 224×224 pixels taking less than two minutes in a middle-level CPU device, and even less in some recent work [30]. The construction of the federated learning framework is also based on another assumption that each individual client has its own computational capability to train both the shared and personalized models locally [31].

B. Decentralized Stain Normalization by cGAN

The major motivation is to normalize the heterogeneous styles of distributed and private histology data under the federated learning manner. The overall pipeline is depicted in Fig. 1.

The Proposed Decentralized cGAN. A drawback in the existing GAN-based stain normalization solutions is the restriction of a pre-defined target style in the adversarial training. Accordingly, the target can not be generated adaptively, nor can a different stain style be produced among a collection of styles. The proposed method in this work differs from conventional conditional GANs in the target style, that not a particular style is pre-targeted. It comprises one global shared generator and multiple client-side specific-personalized discriminators. The global generator $\mathcal{G}_\omega$ orchestrates multiple local discriminators to learn an adaptive interpolation of stain-style distributions s^* from the style collection $\mathcal{S}$. A schematic diagram of learning an interpolated stain style is illustrated in Fig. 2, along with a single fixed style in routine methods for comparison.

Instead of a straightforward optimization in centralized learning [24], we work out the target normalized stain style s^* in a data-private FL paradigm, which is not explicitly formulated as the training objective. Correspondingly, we train $\mathcal{G}_\omega$ to optimize a weighted averaged adversarial loss with the assistance of multiple private discriminators. Each discriminator $\mathcal{D}_{\theta^{(s)}}^{(s)}$ resides locally at client s, without being uploaded to central server. Thus, they contribute to the server independently. Additionally, the trainable parameters $\theta^{(s)}$ of the discriminator, where the superscript $s \in \mathcal{S}$ stands for its client index, are not necessarily identical.

Adversarial Loss. As the stain normalization problem is converted to a style transfer problem [11], the generator $\mathcal{G}_\omega$ processes input histopathology patches $\mathbf{x} \in \mathcal{X}$ as a condition input, where $\mathcal{X}$ writes for the collection of all the valid input. We follow the terminology in cGAN [20] by terming the input image as a condition. Correspondingly, distributed local discriminators $\mathcal{D}^{(s)}_{\theta^{(s)}} : \mathcal{X} \to [0,1]$ are trained separately to distinguish its local histology stain style s from the synthetic normalized images. The local paired network, *i.e.*shared $\mathcal{G}_\omega$ and private $\mathcal{D}^{(s)}_{\theta^{(s)}}$, is trained at its client s with the standard adversarial loss function, i.e.,

$$\begin{aligned} \mathcal{L}^{(s)}_{adv}\left(\theta^{(s)}, \omega\right) &= \mathbf{A}_{\theta^{(s)}} + \mathbf{B}_{\theta^{(s)},\omega}, \text{ with} \\ \mathbf{A}_{\theta^{(s)}} &= \mathbb{E}_{\mathbf{x}\sim P^{(s)}_{\text{data}}}\left[\log\left(\mathcal{D}^{(s)}_{\theta^{(s)}}(\mathbf{x})\right)\right] \text{ and} \\ \mathbf{B}_{\theta^{(s)},\omega} &= \mathbb{E}_{(\mathbf{x},\mathbf{z})\sim P^{(s)}_{\text{data}}\times P_{\text{noise}}}\left[\log\left(1-\mathcal{D}^{(s)}_{\theta^{(s)}}(\mathcal{G}_\omega(\mathbf{z}|\mathbf{x}))\right)\right]. \end{aligned} \tag{1}$$

Operator $\log(\cdot)$ and $\mathbb{E}(\cdot)$ denote the logarithm and expectation respectively.

In the training stage, the global generator tries to fool its local private discriminator with the least variation in tissue structures by optimizing $\mathcal{L}^{(s)}_{adv}$. Simultaneously, the discriminator aims to distinguish a synthetic stain style from its local style distribution. The synthetic images are generated with random noise input, extracted from a pre-defined distribution P_{noise} *e.g.*, a Gaussian, under the conditional input $\mathbf{x}$ characterized by distribution $P^{(s)}_{\text{data}}$. The generator applies a U-Net [32] as its backbone and feeds original images as the input. The noise is injected into the latent feature space by adding directly to the representations. Additionally, multiple real-world stain styles in the federation numbered $|\mathcal{S}|$ require an equivalent number of local discriminators. To be more specific, a client has a unique private discriminator, which iteratively approaches its local stain style during the training. In our system, a distribution-correlated interpolated style is progressively generated with the orchestral generator, which is trained with a number $|\mathcal{S}|$ of discriminators by optimizing the weighted average of the adversarial loss in Eq. (1) *i.e.*,

$$\mathcal{L}_{\text{adv}}\left(\theta^{(1)}, \cdots, \theta^{(|\mathcal{S}|)}, \omega\right) = \sum_{s\in\mathcal{S}} \lambda^{(s)} \cdot \mathcal{L}^{(s)}_{\text{adv}}\left(\theta^{(s)}, \omega\right). \tag{2}$$

The balancing coefficient $\lambda^{(s)}$ is an associated weight to each style s. Innovatively, to achieve faster convergence in training, softened weight coefficients [14], [33] are defined as follows:

$$\lambda^{(s)} = n_s \cdot \exp(\lambda_{adv}\mathcal{L}^{(s)}_{adv}) \cdot \left(\sum_{s'\in\mathcal{S}} n_{s'} \cdot \exp(\lambda_{adv}\mathcal{L}^{(s')}_{adv})\right)^{-1}, \tag{3}$$

where n_s is the number of data instances at client s and λ is a balancing coefficient. Compared with fixed weight coefficients, we adaptive assign the weights to each client based on the training performance.

Pattern Preserving Loss.

To a large extent, the morphological recognizability of input images $\mathbf{x}$ should be retained throughout the stain transfer processing [11]. Correspondingly, the proposed framework adopts a novel pattern preserving loss function by minimizing a L_1-norm between the extracted features from the original input and normalized one to achieve this goal, namely a pattern preserving loss $\mathcal{L}_{pp}$. It attempts to build a better generative model for a fusion of feature representations by explicitly combining a comprehensive range of client modes. Consequently, the target is not prone to omit some important styles for the generator.

Formally, $\mathcal{L}_{pp}$ is built upon $\mathcal{F} = \mathcal{F}_{\text{cnn}} \circ \mathcal{F}_{\text{grey}}$, *i.e.*a composition of the grey image transfer operator $\mathcal{F}_{\text{grey}}$ that maps a RGB image to a grey image, and the pre-trained feature extractor CNN $\mathcal{F}_{\text{cnn}}$ that is trained with grey images. The feature extractor $\mathcal{F}_{\text{cnn}}$ comprises two residual `BasicBlock` layers [36], followed by a max pooling after each block. We set the convolutional channel numbers to 8 for the 1st and 16 for the 2nd residual block layer to reduce parameters.

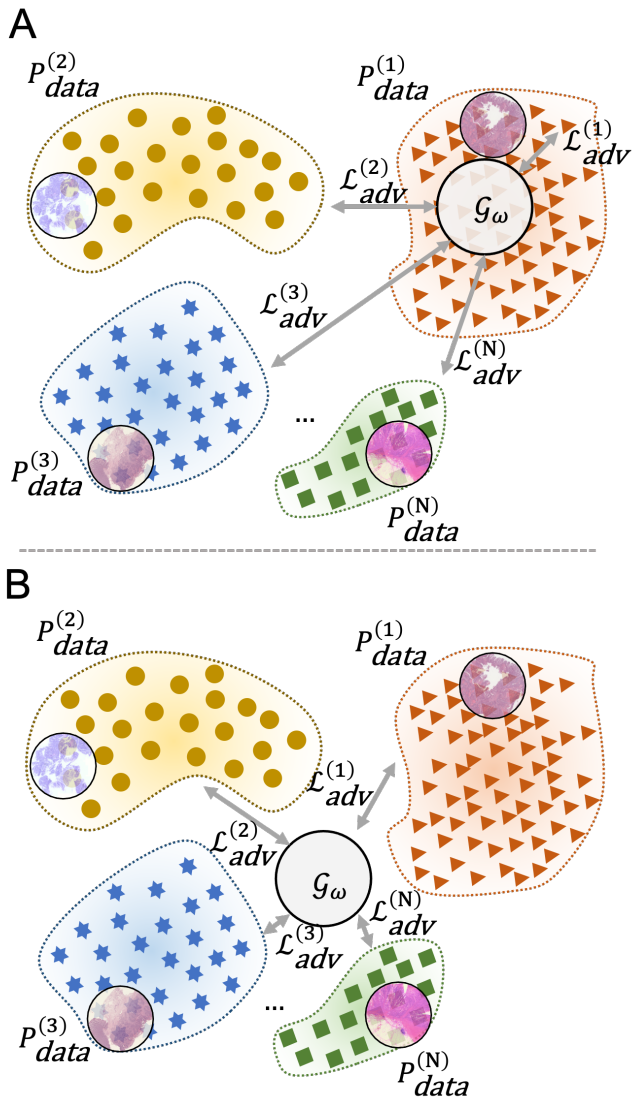

Fig. 2. The comparison between learning paradigms of stain distribution in the routine centralized learning mode and our proposed federated learning mode. A) Routine GAN-based approaches [34], [35] learn a known stain pattern in probability distribution space, *e.g.*the data sets with the maximum number of data instances, for centralized training. B) The adaptive interpolation mode from the entire collection of the distributed clients in the proposed model.

Consequently, an image of $H \times W \times 3$ comes to a feature map of $H/4 \times W/4 \times 16$ after one forwarding. We pretrain the extractor as the encoder in an AutoEncoder [37], where the decoder uses the same structure as the encoder and replaces all the pooling operators with an upsampling operator. The AutoEncoder is trained with L1-norm loss to construct images on a small-scale external public dataset [3]. We do not use any downstream dataset, or large centralized data to avoid potential bias. As we follow the conventional AutoEncoder training scheme, the classification label information from this dataset is totally unused.

The weights of $\mathcal{F}_{\text{cnn}}$ are frozen during the cGAN training. As a measurement of the distance between the generated normalized output image $\mathcal{G}_\omega(\mathbf{z}|\mathbf{x})$ with the original image $\mathbf{x}$, $\mathcal{L}_{pp}$ is defined as

$$\mathcal{L}_{pp}(\omega) = \mathbb{E}_{(\mathbf{x},\mathbf{z})} \Big\| \mathcal{F}[\mathbf{x}] - \mathcal{F}[\mathcal{G}_\omega(\mathbf{z}|\mathbf{x})] \Big\|_1. \tag{4}$$

Algorithm 1: Training the Stain Normalization cGAN with FL (TRAIN_NORMALIZATION_GAN)

Input : local dataset $D^{(s)} = \{\mathbf{x}_i^{(s)}\}_{i=1}^{n_s}$, loss function $\mathcal{L}$, local epochs weights, $\lambda_{\text{epo}}^{(s)}$, rate $\eta^{(s)}$, decay rate $\beta_1^{(s)}$ and $\beta_2^{(s)}$

while *cGAN not converges* **do**
  for *each local client* $s \in \mathcal{S}$ **do**
    Download generator $\mathcal{G}_{\omega_t}$ from central server;
    Train the model locally with LOCALTRAIN(s);
    Upload $\Delta\widehat{\omega}_t^{(s)}$ to central server;
  end
  Parameters aggregation at central server: $\omega_t \leftarrow \omega_{t-1} + \sum_{s\in\mathcal{S}} \lambda \cdot \Delta\widehat{\omega}_t^{(s)}$;
end

LOCALTRAIN(s) **begin**
  Set $N_t^{(s)} \leftarrow N_t \cdot \lambda_{\text{epo}}^{(s)}$ as local epochs ;
  Update the local discriminator by gradient ascent: $\theta_t^{(s)} \leftarrow \text{ADAM}(\mathcal{L}_{\text{adv}}^{(s)}, \theta_{t-1}^{(s)}, D^{(s)}, N_t^{(s)}, \eta^{(s)}, \beta_1^{(s)}, \beta_2^{(s)})$
  Update the local generator with $\mathcal{D}_{\theta_t^{(s)}}^{(s)}$ via gradient descent: $\omega_t^{(s)} \leftarrow \text{ADAM}(\mathcal{L}, \omega_{t-1}, P_{\text{noise}}, N_t^{(s)}, \eta^{(s)}, \beta_1^{(s)}, \beta_2^{(s)})$
  Compute the federated gradient on stain style s: $\Delta\omega_t^{(s)} \leftarrow \omega_t^{(s)} - \omega_{t-1}$
  Privacy preserving: $\widehat{\omega}_t^{(s)} \leftarrow \omega_t^{(s)}$
  Upload $\Delta\hat{\omega}_t^{(s)}$ to the central sever
end

Temporal Self-Distillation Loss.

[3]Dataset is available at https://doi.org/10.5281/zenodo.2530789, it has 11977 patches.

In addition to the tissue structure preservation issue, previous works in training stain normalization GAN [11] encountered another problem that the orchestral generator might heavily suffer from training divergence or mode collapse, resulting in slow convergence and poor performance. These difficulties hinder the real-world clinical implementations for downstream diagnostic tasks, as well as heavy reliance on the time-consuming and computational-costly hyper-parameter fine-tuning. To cope with these issues, a temporal self-distillation scheme is introduced to stabilize generator training by providing extra supervision signals. It stabilizes the training process by imposing a consistent regularization between the updated average with the current prediction. The temporal mean teacher model [38] is an exponential moving average (EMA) of successive weights of the global generator. Specifically, the EMA parameters ω^{EMA} at the t^{th} federated round is formulated as:

$$\omega^{EMA} = \lambda^{EMA} \cdot \omega + (1 - \lambda^{EMA}) \cdot \omega^{EMA} \tag{5}$$

with a fixed smoothing coefficient λ^{EMA} at 0.99, following the existing work [38]. Conclusively, the extra self-distillation loss is formulated as:

$$\mathcal{L}_{sd}(\omega) = \mathbb{E}_{(\mathbf{x},\mathbf{z})} \Big\| \mathcal{G}_{\omega^{EMA}}(\mathbf{z}|\mathbf{x}) - \mathcal{G}_\omega(\mathbf{z}|\mathbf{x}) \Big\|_1 \tag{6}$$

where $\mathcal{G}_{\omega^{EMA}}$ denotes the temporal mean model is parameterized by ω^{EMA}. We present the pseudo-code for training with temporal self distillation in Algorithm 2.

Algorithm 2: Train with temporal self-distillation

begin
  for *each local client* $s \in \mathcal{S}$ **do**
    Download student generator weights ω;
    Download teacher generator weights ω^{EMA};
    Train the student generator;
    Upload the trained weights ω to central server;
  end
  Aggregate weights ω;
  Update the teacher model with Eq. (5);
end

Overall Loss Functions. As a synergy of the three discussed loss functions, we aim to optimize the trainable parameters ω with the following objective:

$$\omega^* = \arg\max_{\omega} \min_{\theta^{(s)} \text{ for all } s\in\mathcal{S}} (\mathcal{L} = \mathcal{L}_{adv} + \gamma_1 \mathcal{L}_{pp} + \gamma_2 \mathcal{L}_{sd}) \tag{7}$$

where γ_1 and γ_2 balances different loss functions. In summary, $\mathcal{L}$ is a function of learnable weights ω and the collection of all $\theta^{(s)}$ for $\forall s \in \mathcal{S}$.

The gradient descent ascent (GDA) algorithm [12], widely used in practical applications such as GANs and adversarial training for natural generalization, is also employed in our work to train the proposed cGAN. Each local discriminator $\mathcal{D}_{\theta^{(s)}}^{(s)}$ is trained locally by its partial private data to maximize the loss $\mathcal{L}$; meanwhile, the orchestrating generator attempts to minimize $\mathcal{L}$.

C. Train Proposed cGAN with Federated Learning

Existing stain normalization approaches [35] remain a data-sharing centralized training mode and restrict stain style learned from a single dataset [34], whereas the proposed architecture is data-privacy and data-safety guaranteed for multiple clients normalization in a federated learning mode. In this section, we present training on the proposed decentralized cGAN in FL paradigm. Orchestrated by the central generator, each and every discriminator at the client-side is one-to-one associated with its stain style s on their local datasets $D^{(s)} = \{\mathbf{x}_i^{(s)}\}_{i=1}^{n_s}$. In the initial of each local epoch, the parameters ω of generator $\mathcal{G}_\omega$ are downloaded from the central server and fed to the distributed clients, as shown in Fig. 1, indicated with stage ①. The subsequent local client-side adversarial training process is depicted in Algorithm 1, where the subscript t of parameters denotes the t^{th} global federated round. The generator conditioned with all the styles is then optimized regarding the corresponding conditioned discriminator. Notably, every discriminator is responsible for its local parameter updates during every global federated round, allowing the absence of weight sharing between the clients. And coordinately, the global generator is updated once every global federated round. The iterative update of local training is illustrated by stage ② in Fig. 1. In parallel, the federated gradient $\Delta\widehat{\omega}_t^{(s)}$ with respect to $s \in \mathcal{S}$ at each client is progressively updated. Afterwards, they are uploaded back to the central server which has never seen any raw data during the parameters aggregation, as demonstrated by stage ③ in Fig. 1. Subsequently, a weighted averaging algorithm is used by computing the weighted balancing coefficient in Eq. (3). The server aggregates the client updates when federated weights from every individual client are gathered up. The aggregation processing is shown as stage ④ in Fig. 1. The iterative training is terminated, when the model converges. The convergence criteria in the training procedure are determined by the decline rate of the overall training loss function or by the maximum training epochs. For example, in our paper, the model converges when the loss function does not decline.

D. Training Diagnostic Model with Federated Learning

To show the applicability of the proposed stain normalization framework on downstream tasks, the performance of the FL classifier is validated on the normalized images. First, a well-trained optimal style normalization generator $\mathcal{G}^*$ is applied to each client for local stain normalization. Then, a normalized set $\{\mathcal{G}^*(\mathbf{x}_i^{(s)})\}_{i=1}^{n_s}$ is derived at each client $s \in \mathcal{S}$, where the stain variation is alleviated by normalization. Without any data sharing, all normalized images are then utilized in the federation to train a diagnostic model *e.g.*, a classification neural network such as ResNet in our case. The target model FL training framework is depicted in Algorithm 3, in which we adopt the widely-used FedAvg [31].

Algorithm 3: Training target neural network

Input : datasets $\{\mathbf{x}_i^{(s)}\}_{i=1}^{(s)}$

begin
  $\mathcal{G}^* \leftarrow$ TRAIN_NORMALIZATION_GAN();
  for *each local client* $s \in \mathcal{S}$ **do**
    Download $\mathcal{G}^*$ from central server;
    Normalize the images in $\widehat{D}^{(s)} \leftarrow \{\mathcal{G}^*(\mathbf{x}_i^{(s)})\}_{i=1}^{n_s}$ locally;
  end
  Train FL model with normalized images $\mathcal{M}^* \leftarrow$ FEDAVG$(\widehat{D}^{(s)})$;
end

IV. Experiments

A. Dataset and Setting Ups

Stain Normalization. Notably, the proposed normalization method is downstream task-agnostic and tissue-structure agnostic. In our experiment for performance evaluation, we use histopathology images of Colorectal cancer (CRC), the third most common cancer worldwide [39]. Three public data sets are employed for the classification training and test [40]: 1) Sub-data sets from Cancer Genome Atlas (TCGA) COAD, where a total number of 100 histopathology slides are freely available [4], 2) CRC-VAL-HE-7K data set containing 25 WSIs, 3) NCT-CRC-HE-100K data set of 86 WSIs. The whole-slide images are divided into patches sized 224×224 pixels at the magnitude of $20\times$, as have been demonstrated to be effective in previous works [41]. The very same dataset is applied to the federated stain normalization cGAN and classification neural network training. To be more specific, the entire set of image patches was firstly used to train our proposed orchestration-generator-distributed-discriminators cGAN for distributed stain normalization. Afterwards, the data instances were normalized with the well-trained cGAN for the subsequent FL classification task, in order to further evaluate the performance of stain normalization. In the data partition for network evaluation, 70% of the whole set is used to train a federated classifier, 15% to validate and the left 15% to test. Moreover, we perform dataset partition at the client level to evaluate the performance to evaluate the model with multiple distributions. For reproducibility, we fixed the random seed to 0 at this stage.

Patch-level classification. In the downstream FL patch-level classification task, each individual patch is categorized into one of nine classes, namely i) adipose (ADI) ii) background (BACK) iii) debris (DEB) iv) lymphocytes (LYM) v) mucus (MUC) vi) smooth muscle (MUS) vii) normal colon mucosa (NORM) viii) cancer-associated stroma (STR) ix) colorectal adenocarcinoma epithelium (TUM) for training and testing a tissue phenotype classification neural network. The patch-level annotations for CRC-VAL-HE-7K and NCT-CRC-HE-100K are freely available [5], and the TCGA-COAD subset was manually annotated by three experienced pathologists. To test the robustness with different distributions, the whole set was divided into subsets of $\mathcal{S} = 5, 6, 7$ and 8 without data overlap, and their frequency distributions are along shown in

[4] https://portal.gdc.cancer.gov

[5] The original patches and labels are available at https://zenodo.org/record/1214456#.YbyoJi8RpQI.

the chart of Fig.4. We follow the rule that images in one client share exactly the same stain style. For a clear notation, we write Dataset A for $|\mathcal{S}| = 5$, Dataset B for $|\mathcal{S}| = 6$, C for $|\mathcal{S}| = 7$, D for $|\mathcal{S}| = 8$. To ensure that patches in each site are i.i.d., distinctive stain normalization by Macenko's method [42] is applied to each client. Each client picks up a local private template [11] for normalization. Consequently, after the local normalization, the non-i.i.d. properties of data samples from the different sites are still retained, while samples from the same site are i.i.d. This well complies with our previous assumption. Notably, as current conventional normalization schemes [13], [17], [19], [42], [43] work only in the centralized setting, they are not capable of federated learning settings. Consequently, their impact on FL performance is not discussed in this paper.

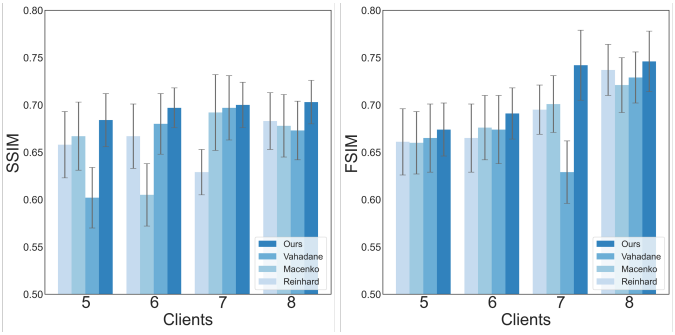

Fig. 3. Comparison of the averaged SSIM and FSIM across different tissue types between our proposed cGAN and conventional stain normalization methods.

WSI-level classification. Considering the clinical significance of tumor status identification at the WSI-level, we also adopt a WSI-level downstream evaluation task with Camelyon17 [44] [6]. It contains 899 WSIs for training and 500 WSIs for testing. We separate Camelyon17 WSIs into 5 individual clients by their different institution sources.

B. Evaluation Metrics

1) Stain Normalization: The following two metrics are used for evaluating the stain normalization effect: 1) Structural Similarity Index (SSIM) [45], and 2) Feature Similarity Index for Image Quality Assessment (FSIM) [46]. SSIM aims to quantify the discrepancy in the normalized data set [45], and FSIM uses computational models to measure image quality consistency [46]. The higher value of both two metrics indicates the better performance of stain normalization outcome. Three baseline normalization methods [17]–[19], [47] were tested to evaluate our normalization method in centralized training paradigm. However, the pre-requisite of selecting a template in stain normalization approaches violates the data decentralized setting in FL. And because of this infeasibility of privacy conversion, the other existing normalization algorithms [11] are not compared in the federated setting. Therefore, the proposed method in the FL setting is compared straightforwardly with the baselines in the centralized setting. Likewise, neither can decentralized learning be compared with the routine centralized learning based model in terms of normalized quality.

2) Classification: In addition to the two metrics, a nine-class classification neural network is also trained in an FL setting, which further evaluates the stain normalization effect on downstream tasks with the methods described above. ResNet [36], DenseNet [48] and EfficientNet [49] are selected as the backbone architectures for their outstanding performance in pathology analysis and their compactness in the literature works. The performance was evaluated by the classification accuracy on the nine-class histopathology test set. The accuracy is computed by

$$\text{Acc} = \frac{\sum_{j=1}^{N} \frac{tp_j + tn_j}{tp_j + tn_j + fp_j + fn_j}}{N}, \tag{8}$$

where tp_j, fp_j, tn_j, and fn_j are the true positives, false positives, true negatives, and false negatives for the j-th class respectively [54]. N counts for the total number of classes, which in this research is set to $N = 9$ of tissue type. Additionally, the area under the receiver operating characteristic curve (AUC) is also computed on the test set as another metric for the federated classifier, for its precision in terms of informativeness [54]. Specifically, the individual AUC per class (one-vs-all) is computed and the resulting values are averaged. The AUC for each class is computed by

$$\text{AUC} = \frac{S_p - n_p(n_p + 1)/2}{n_p n_n}, \tag{9}$$

where S_p is the sum of all positive examples ranked, and n_p, and n_n are the number of positive and negative examples respectively [54].

C. Implementation Details

The benchmark tests were performed in a Python (version 3.6) environment on NVIDIA Tesla V100 GPUs with 32GB GPU memory. The proposed orchestration-generator-distributed-discriminator structure was implemented in Pytorch, and it was likewise the federated ResNet, DenseNet and EfficientNet [36], [48], [49] classifiers. Adam optimizer was adopted to train the patch-level classifier by optimizing a cross-entropy loss function. The remaining hyper-parameters in the experiments were set as follows: local epoch number $= 10$ for each client, learning rate $= 1 \times 10^{-2}$, decay rate $= 5 \times 10^{-4}$, $\lambda_{adv} = 1$ in Eq. (3), and $\gamma_1 = \gamma_2 = 1$ in Eq. (7). Thse training configurations for pretraining the $\mathcal{F}_{\text{cnn}}$ are set as follows: a total epoch number of 100, a fixed learning rate of 5×10^{-4}, a cosine annealing scheduler with T=10, and an Adam optimizer. During the training stage, random flipping is adopted for image augmentation. The proposed network was implemented with Automatic Mixed Precision (AMP) to improve training efficiency. About three hours were taken to train one global epoch of the proposed stain normalization cGAN in the applied deep learning platform.

[6]WSIs and annotations are available at `https://camelyon17.grand-challenge.org/Data/`.

TABLE I

PERFORMANCE COMPARISON OF DIFFERENT STAIN NORMALIZATION METHODS, WHERE † DENOTES CONVENTIONAL CENTRALIZED LEARNING AND § DENOTES FEDERATED LEARNING.

Mode	Metric	SSIM				FSIM			
	Methods	Reinhard et al.†	Macenko et al.†	Vahadane et al.†	**Ours**§	Reinhard et al.†	Macenko et al.†	Vahadane et al.†	**Ours**§
5 clients	ADI	0.651±0.021	0.632±0.044	0.652±0.051	**0.665**±0.021	0.614±0.041	0.628±0.043	0.618±0.043	**0.634**±0.020
	BACK	0.800±0.020	0.833±0.031	0.841±0.021	**0.858**±0.010	0.803±0.019	0.841±0.052	0.831±0.024	**0.840**±0.010
	DEB	0.651±0.044	0.670±0.043	0.682±0.030	**0.691**±0.033	0.650±0.032	0.665±0.042	0.678±0.038	**0.680**±0.021
	LYM	0.645±0.023	0.654±0.039	0.670±0.026	**0.672**±0.031	0.659±0.025	0.670±0.031	0.671±0.041	**0.673**±0.023
	MUC	0.655±0.031	0.650±0.028	0.641±0.010	**0.663**±0.021	0.632±0.031	0.633±0.024	0.623±0.031	**0.640**±0.035
	MUS	0.662±0.037	0.671±0.036	0.689±0.030	**0.694**±0.035	0.671±0.043	0.641±0.026	0.670±0.036	**0.674**±0.029
	NORM	0.651±0.039	0.649±0.044	0.642±0.041	**0.667**±0.040	0.661±0.031	0.637±0.048	0.652±0.035	**0.670**±0.033
	STR	0.602±0.043	0.621±0.031	0.616±0.034	**0.628**±0.031	0.598±0.034	0.609±0.037	**0.628**±0.044	0.627±0.040
	TUM	0.611±0.057	0.619±0.028	0.610±0.043	**0.620**±0.032	0.611±0.057	0.623±0.034	0.616±0.035	**0.628**±0.038
	Average	0.658±0.035	0.667±0.036	0.602±0.032	**0.684**±0.028	0.661±0.035	0.660±0.033	0.665±0.036	**0.674**±0.028
6 clients	ADI	0.643±0.023	0.645±0.041	0.641±0.025	**0.689**±0.010	0.603±0.034	0.647±0.031	0.619±0.031	**0.651**±0.025
	BACK	0.803±0.019	0.853±0.013	0.877±0.019	**0.891**±0.008	0.870±0.023	0.885±0.015	0.878±0.014	**0.894**±0.011
	DEB	0.674±0.033	0.683±0.031	0.693±0.031	**0.704**±0.021	0.688±0.036	0.710±0.042	0.709±0.031	**0.714**±0.032
	LYM	0.693±0.031	0.685±0.043	0.703±0.012	**0.704**±0.013	0.666±0.027	0.678±0.035	0.667±0.042	**0.682**±0.013
	MUC	0.652±0.020	0.641±0.032	0.634±0.019	**0.669**±0.016	0.623±0.033	0.634±0.028	0.617±0.036	**0.633**±0.024
	MUS	0.689±0.044	0.677±0.031	**0.704**±0.031	0.701±0.020	0.668±0.047	0.673±0.035	0.669±0.041	**0.708**±0.031
	NORM	0.651±0.039	0.649±0.044	0.642±0.041	**0.667**±0.040	0.661±0.031	0.637±0.048	0.652±0.035	**0.670**±0.033
	STR	0.598±0.055	0.612±0.047	0.621±0.061	**0.629**±0.023	0.604±0.052	0.611±0.043	0.636±0.056	**0.645**±0.041
	TUM	0.600±0.041	0.620±0.019	0.613±0.047	**0.621**±0.034	0.610±0.043	0.610±0.025	0.621±0.041	**0.629**±0.036
	Average	0.667±0.034	0.605±0.033	0.680±0.032	**0.697**±0.021	0.665±0.036	0.676±0.034	0.674±0.036	**0.691**±0.027
7 clients	ADI	0.687±0.012	0.688±0.032	0.634±0.042	**0.690**±0.018	**0.728**±0.021	0.693±0.023	0.693±0.013	0.723±0.042
	BACK	0.894±0.008	0.888±0.030	0.883±0.004	**0.925**±0.010	0.911±0.006	0.903±0.012	0.900±0.010	**0.918**±0.007
	DEB	0.713±0.022	0.711±0.018	0.719±0.024	**0.731**±0.010	0.734±0.021	0.743±0.013	0.758±0.019	**0.767**±0.061
	LYM	0.691±0.010	0.704±0.031	**0.713**±0.043	0.628±0.033	0.694±0.018	0.703±0.023	0.678±0.033	**0.709**±0.017
	MUC	0.641±0.022	0.633±0.081	0.662±0.023	**0.671**±0.021	0.604±0.021	0.638±0.031	0.659±0.049	**0.677**±0.020
	MUS	0.721±0.021	0.731±0.030	0.724±0.042	**0.730**±0.022	0.701±0.029	0.738±0.046	0.754±0.042	**0.768**±0.023
	NORM	0.647±0.042	0.647±0.052	0.630±0.033	**0.670**±0.042	0.678±0.022	0.669±0.054	0.674±0.042	**0.681**±0.055
	STR	0.598±0.055	0.612±0.047	0.621±0.061	**0.629**±0.023	0.604±0.052	0.611±0.043	0.636±0.056	**0.645**±0.041
	TUM	0.611±0.021	0.610±0.036	0.599±0.059	**0.634**±0.038	0.603±0.051	0.611±0.021	0.613±0.032	**0.634**±0.071
	Average	0.629±0.024	0.692±0.040	0.687±0.034	**0.700**±0.024	0.695±0.026	0.701±0.030	0.629±0.033	**0.742**±0.037
8 clients	ADI	0.621±0.023	0.632±0.012	0.582±0.034	**0.634**±0.020	**0.721**±0.032	0.688±0.028	0.689±0.026	0.719±0.028
	BACK	0.892±0.008	0.883±0.010	0.888±0.009	**0.933**±0.003	0.902±0.004	0.890±0.004	0.901±0.004	**0.911**±0.005
	DEB	0.701±0.031	0.702±0.021	0.703±0.032	**0.703**±0.018	0.756±0.032	0.754±0.021	**0.757**±0.022	0.755±0.026
	LYM	0.683±0.024	**0.711**±0.023	0.693±0.025	0.698±0.033	0.688±0.035	0.690±0.034	0.650±0.040	**0.691**±0.036
	MUC	0.632±0.048	0.621±0.051	0.641±0.044	**0.644**±0.032	0.703±0.044	0.688±0.023	0.699±0.029	**0.707**±0.027
	MUS	**0.742**±0.051	0.722±0.031	0.731±0.030	0.732±0.022	0.798±0.019	0.801±0.037	0.804±0.030	**0.811**±0.022
	NORM	0.688±0.011	0.633±0.033	0.632±0.021	**0.693**±0.018	0.743±0.012	0.688±0.041	0.733±0.034	**0.755**±0.043
	STR	0.563±0.042	0.602±0.027	0.608±0.032	**0.630**±0.034	**0.650**±0.044	0.603±0.055	0.639±0.037	0.643±0.059
	TUM	0.621±0.032	0.599±0.061	0.582±0.055	**0.656**±0.025	0.670±0.025	0.688±0.049	0.692±0.055	**0.699**±0.044
	Average	0.683±0.030	0.678±0.033	0.673±0.031	**0.703**±0.023	0.737±0.027	0.721±0.029	0.729±0.027	**0.746**±0.032

TABLE II

THE P-VALUE COMPUTED WITH STUDENT'S T TEST BETWEEN OUR APPROACH FOR THE HIGHEST CONVENTIONAL STAIN NORMALIZATION METHOD IN TERMS OF AVERAGE SSIM AND FSIM IN TABLE I.

Mode	SSIM	FSIM
5 clients	3.1×10^{-6}	7.8×10^{-7}
6 clients	4.9×10^{-7}	1.7×10^{-6}
7 clients	2.7×10^{-5}	8.0×10^{-6}
8 clients	4.4×10^{-4}	3.9×10^{-5}

D. Experimental Results

1) Performance comparison between the proposed FL mode and conventional methods in data-sharing mode: Three conventional and widely-used color transfer methods were employed in a centralized learning setting for performance comparison [17], [18], [47] to our FL normalization scheme in terms of SSIM and FSIM. Image styles transfer is performed across the labeled groups collaboratively. Concretely, for each labeled category, three experienced pathologists manually curated the template images collaboratively. In the proposed framework, histopathology images are kept at local clients at both the GAN training and color transferring stages to preserve data privacy. As demonstrated in Table I and Fig. 3, the proposed model significantly outperforms the previous collaborative stain normalization methods, with higher SSIM and FSIM values. The advantages are consistent under four different data distribution schemes with a varying number of clients from 5 to 8. Furthermore, the statistical significance is confirmed by the p-values computed from the student's t-test in Table II between our approach with the conventional approach that achieved the best performance. All p-values are consistently smaller than 0.05, yielding the statistical significance of the performance improvement achieved by our approach. Without the exposure of private data, the proposed method achieves higher stain homogeneity in the distinctive histopathology images, compared with previous collaborative normalization methods. A couple of patch examples are listed in Fig. 4, where a better visual normalization quality can be observed.

2) Effectiveness of patch-level classification: In addition to stain normalization, the power of stain augmentation (also known as 'stain randomization') as an alter for stain normalization is already investigated in [52]. Hence, we employ its settings by taking in two popular stain augmentation methods. (1) Noise is added to each channel with multiplication rule

TABLE III

COMPARISON OF CLASSIFICATION ACCURACY ON THE TEST SET BETWEEN OUR PROPOSED NORMALIZATION METHODS WITH FEDERATED LEARNING SETTING (FL) AND 1) NORMALIZATION METHODS WITH CONVENTIONAL CENTRALIZED LEARNING SETTING, 2) STAIN AUGMENTATION METHODS WITH BOTH CENTRALIZED LEARNING AND FEDERATED LEARNING SETTINGS, 3) NON-NORMALIZATION METHODS WITH FEDERATED LEARNING SETTING. THREE NETWORK BACKBONE ARCHITECTURES ARE USED: RESNET [36], DENSENET [48], AND EFFICIENTNET [49]. THE PROPOSED METHOD CONSISTENTLY ACHIEVES ABOUT 20% TEST ACCURACY IMPROVEMENT, WHICH PARALLELS WITH TRADITIONAL CENTRALIZED TRAINING.

	Methods	Settings	ResNet-18 [36]	ResNet-50 [36]	DenseNet-121 [48]	DenseNet-169 [48]	EfficientNet [49]
Normalization Performance	Reinhard et al. [17]	Centralized Training	94.793%	94.671%	95.512%	95.912%	96.341%
	Macenko et al. [18]		96.374%	96.381%	96.102%	96.008%	96.410%
	Vahadane et al. [47]		91.689%	92.849%	94.212%	94.295%	94.203%
	Khan et al. [19]		90.031%	91.012%	93.571%	94.531%	95.002%
	StaNoSA [50]		95.217%	96.007%	95.721%	95.879%	96.120%
	StainGAN [24]		96.119%	96.569%	95.874%	96.104%	96.708%
	MultiPathGAN [51]		96.487%	**96.982%**	96.577%	**96.893%**	96.774%
	The proposed method		**96.539**%	96.784%	**96.810**%	96.811%	**96.893**%
Augment Performance	SA1-L [52]	Centralized Training	81.294%	82.031%	81.199%	81.037%	74.232%
	SA1-S [52]		85.211%	85.789%	83.280%	84.432%	87.217%
	SA2-L [52]		91.102%	90.456%	91.578%	91.788%	92.563%
	SA2-S [52]		91.213%	90.573%	91.942%	92.031%	92.674%
Federated Learning	Classification without normalization	5 clients	74.871%	74.789%	74.981%	74.983%	75.021%
		6 clients	75.122%	75.124%	75.245%	75.312%	76.004%
		7 clients	74.313%	76.237%	75.128%	75.120%	76.215%
		8 clients	78.131%	75.110%	77.231%	76.342%	78.102%
	Classification with the stain augmentation (SA2-S)	5 clients	89.021%	88.237%	87.321%	88.228%	89.208%
		6 clients	89.120%	88.782%	87.983%	88.310%	89.418%
		7 clients	89.135%	88.705%	88.598%	88.984%	90.032%
		8 clients	89.130%	89.102%	88.781%	89.044%	90.047%
	Classification with the proposed normalization strategy	5 clients	94.318%	94.551%	94.573%	94.672%	95.021%
		6 clients	94.213%	94.434%	94.553%	93.983%	95.550%
		7 clients	95.230%	95.412%	94.973%	95.312%	95.117%
		8 clients	96.024%	96.031%	96.112%	96.129%	96.210%

TABLE IV

THE P-VALUE COMPUTED FROM FRIEDMAN STATISTIC TEST [53] FOR THE RESULTS PRESENTED IN TABLE III. ALL P-VALUES ARE CONSISTENTLY SMALLER THAN 0.05, WHICH CAN CONFIDENTLY CONFIRM THE RELIABILITY AND STATISTICAL SIGNIFICANCE OF THE PERFORMANCE IMPROVEMENT ACHIEVED BY OUR APPROACH.

Settings	ResNet-18 [36]	ResNet-50 [36]	DenseNet-121 [48]	DenseNet-169 [48]	EfficientNet [49]
5 clients	5.2×10^{-4}	6.1×10^{-5}	3.8×10^{-6}	3.8×10^{-6}	7.2×10^{-7}
6 clients	1.4×10^{-5}	4.3×10^{-6}	1.0×10^{-6}	9.9×10^{-8}	7.5×10^{-5}
7 clients	3.9×10^{-4}	2.9×10^{-4}	8.5×10^{-8}	6.7×10^{-9}	3.5×10^{-4}
8 clients	5.3×10^{-4}	1.9×10^{-5}	4.9×10^{-4}	9.3×10^{-6}	8.3×10^{-7}

in LAB color space *i.e.*, $p' = p * \varepsilon_1 + \varepsilon_2$, where p' is the augmented pixel value, p is the associated original pixel value, ε_1, ε_2 are the uniform random noises. This method is termed stain augmentation scheme #1 (SA1). (2) Noise is added to each channel with addition rule in HSV color space *i.e.*, $p' = p + p * \varepsilon$, which is termed as stain augmentation scheme #1 (SA2). For both stain augmentation methods, we follow [52] by adopting two augmentation configurations *i.e.*, light (L) and strong (S), which is determined by the degree of distortion for the random noise. In Table III, the performance was compared in terms of classification accuracy, with baselines for both the collaborative and federated settings. Stain augmentations are compared for both collaborative and federated settings. We only adopt SA2-S in a federated setting, because it achieves the best performance in a collaborative setting. The gap in the distributed and data-private color transfer of histopathology can be narrowed with our normalization method, at an obvious accuracy improvement of over 20.0% in comparison with the absence of a stain normalization pre-processing. Notably, our method was supposed to be reasonably compared with other GANs in the FL settings, yet hindered by the restrictions of current GANs with centralized training. As a result, comparisons with StainGAN [24] and MultiPathGAN [51] in the centralized training settings are performed. The marginal improvement for patch-level classification in a centralized setting is a by-product to demonstrate the superiority to non-FL settings.

Comparisons between stain normalization and stain augmentation yield that the former consistently outperforms the latter in both collaborative and federated learning settings. The advantage of stain normalization is an effective alignment of various stain styles, while stain augmentation is a simulation of reliable stain styles. In FL, stain normalization outperforms stain augmentation due to the restricted data visibility of each client, hence stain augmentation fails to estimate the whole distribution to simulate potential valid styles. This finding

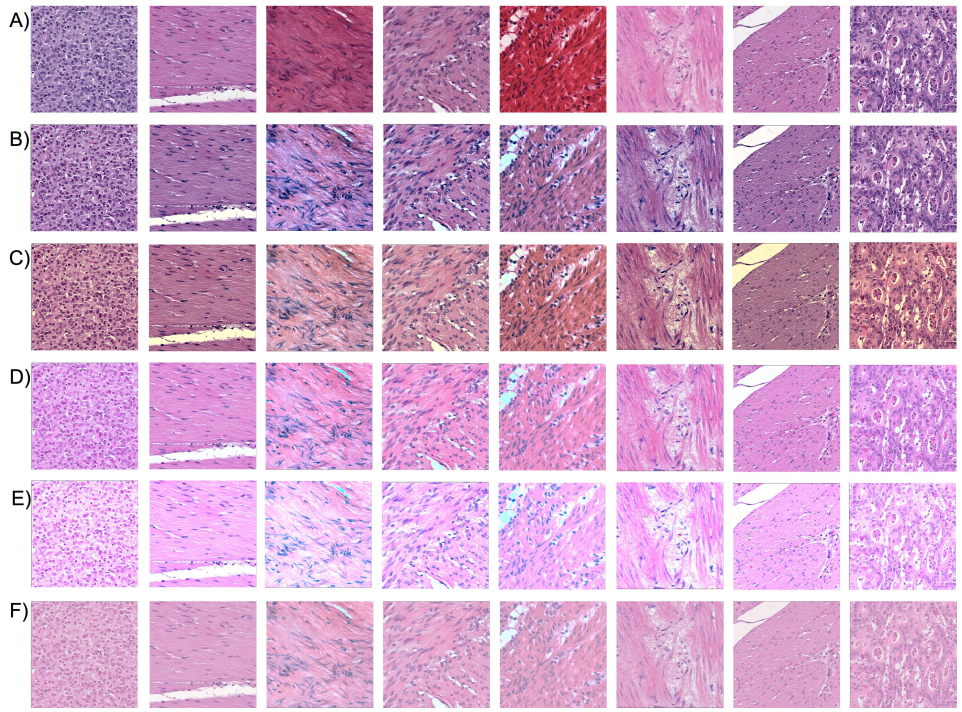

Fig. 4. The comparisons of sample patches after stain normalization with different methods. A) Original patches. B) Reinhard et al. [17]. C) Macenko et al. [18]. D) Vahadane et al. [47]. E) Khan et al. [19]. F) Proposed method trained with 8-client mode. A noticeable contribution is that the recognition ability is mostly preserved even with drastic changes in color, compared to state-of-the-art normalization methods. Other GAN-based normalization methods in the literature are not useful for comparison, due to the limitation of data-sharing between only two domains.

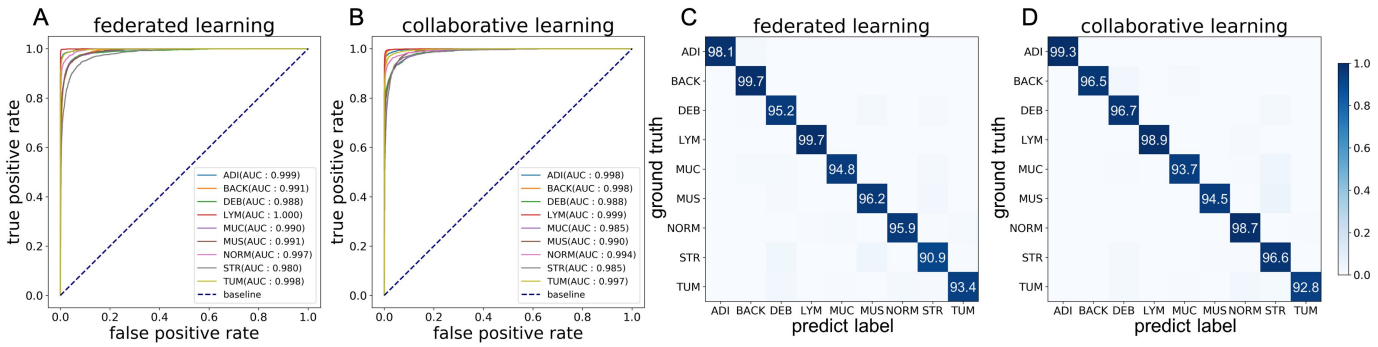

Fig. 5. Performance comparison of the federated learning (FL) system and the current centralized learning system on the classification task after different stain normalization approaches. A-B) The area under the receiver operating characteristic curves (AUCs) on the nine categories of FL and centralized learning respectively. C-D) The confusion matrix. All evaluation was performed in the 8-client mode, where we can draw parallel results within other client numbers.

suggests that stain normalization is a powerful solution for addressing stain variation under FL settings. The associated p-values to the Table III are computed with Friedman statistic test [53], and are presented in Table IV. Empirically, all p-values are consistently smaller than 0.05, which thus confidently confirms the reliability and statistical significance of our performance improvement.

The accuracy improvement is consistent with all the backbone CNNs [36], [48], [49], which suggests that the proposed normalization method is structure-independent in a downstream CNN classification task. Amongst the literature works [14], Macenko's method in the centralized training stood out for its high accuracy, hence we compare our federated-learning based system with it in terms of the area under the receiver operating characteristic curve (ROC) and the confusion matrix in Fig. 5. We observe that our FL method achieves very competitive performance as centralized training after stain normalization. The learning curve of the training dynamics comparison is shown in Fig. 6, where the proposed FL-based system shows more stable training dynamics on the normalized images for training deep neural networks, together with better matching of test accuracy. The learning curves for CL are different because we set the data inputs slightly different for these four settings by keeping the data and the pre-processing for CL and FL identical i.e., the same local stain normalization.

Interestingly, most previous research on federated learning in the other discipline of biomedical images like CT or MRI [3] reported the superiority of conventional centralized learn-

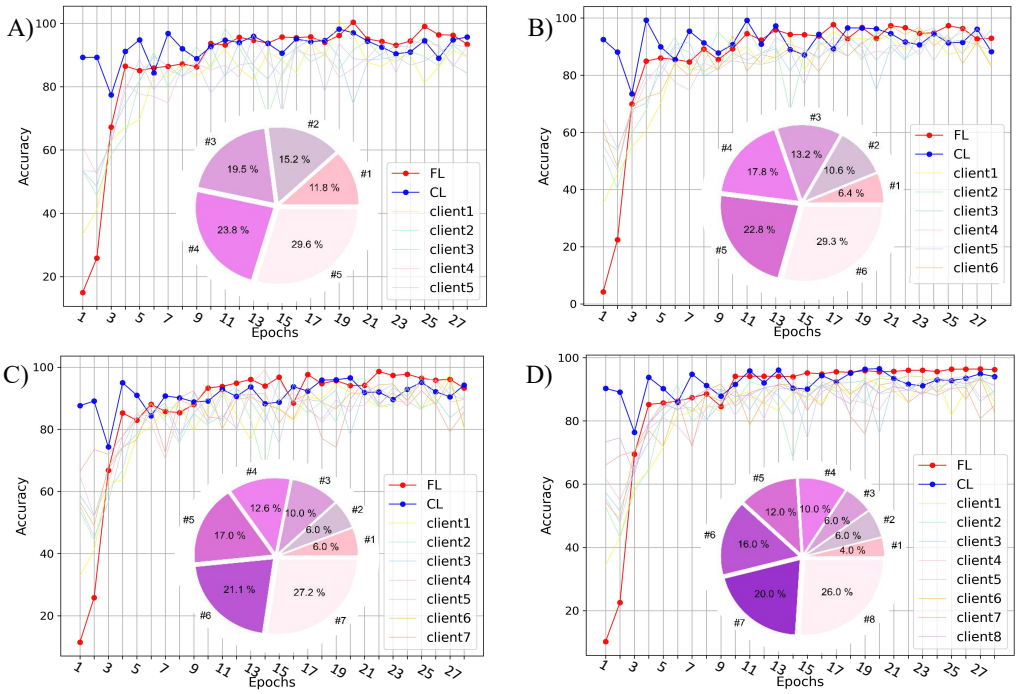

Fig. 6. The learning curves during the training for individual clients, in the conventional centralized learning (CL) mode [40] and the proposed histopathology stain normalization with federated learning (FL). The latter shows comparable convergence between the regular data-sharing centralized mode and the proposed data-private federated learning framework. The client number rises from 5 to 8 in A)-D). The pie chart in each sub-figure shows the data proportion of each client involved in the federated training.

ing over federated learning. Yet, in the context of histopathology image analysis, FL trains the model with a faster convergence *e.g.*, as shown in Fig. 6 (d), FL converged after the ninth federated epoch, when the centralized learning could not make the convergence. The distribution of histology images across different clients is highly non-i.i.d. partly due to the stain heterogeneity characteristic [55]. After stain normalization, the heterogeneity is remarkably mitigated *i.e.*achieving higher SSIM and FSIM scores as shown in Table I, but still is non-i.i.d. The work yields that a distributed training scheme that training a model locally on different stain styles can contribute to a more efficient training process.

TABLE V

COMPARISON WITH PERSONALIZED FEDERATED LEARNING APPROACHES [56]–[58] WHICH TARGET THE DATA HETEROGENEITY PROBLEM IN FL. WE USE RESNET-18 AS THE BACKBONE. BASELINE REFERS TO FEDAVG WITH NO STAIN NORMALIZATION.

Method	Acc (%)
Baseline	74.871
LG-FedAvg [57]	75.312
FedPer [58]	89.997
FedBN [56]	84.213
Ours	**94.318**

We also compare with personalized federated learning (PFL) methods specifically for heterogeneous and non-i.i.d. data, including LG-FedAvg [57], FedPer [58] and FedBN [56]. LG-FedAvg personalizes the lower layers and keeps the top layers globally shared [57]. Conversely, FedPer personalizes the top layers and keeps the bottom layers globally shared [58]. FedBN personalizes the batch normalization layers [56]. We use ResNet-18 as the CNN backbone on the patch-level classification dataset with $\mathcal{S} = 5$. We set the number of personalized layers to one residual block for LG-FedAvg and FedPer. The comparison results are shown in Table V, where our stain normalization approach outperforms these PFL methods. Additionally, it is expected to gain further improvement by combining PFL with our stain normalization method.

TABLE VI

WSI-LEVEL EVALUATION ON CAMEYLON 17 IN FEDERATED LEARNING. WE USE RESNET-18 AS THE CNN BACKBONE.

Method	Accuracy(%)	AUC
baseline	83.220	0.903
stain augmentation	86.601	0.920
ours	**87.208**	**0.932**

3) Effectiveness of WSI classification: As a binary classification task, we use accuracy and AUC on the test set as the evaluation metric. The comparisons with baseline (i.e., without stain normalization) and stain augmentation (with SA2-S) are presented in Table VI.

E. Ablation Study

The ablations on the pattern preserving loss, temporal self-distillation, and softened weighted coefficients are shown in Table VIII. We use a patch-level classification task and set the client number to 5, with ResNet-18 as the backbone. As suggested, the softened weighted coefficient contributes to a marginal improvement of 1.139%. While pattern preserving loss is the most important component, where an absence will drastically degrade the performance. Empirically, the ablation demonstrates the effectiveness of different components.

TABLE VII
THE AVERAGED SSIM QUANTIFICATION ON IMAGES BEFORE AND AFTER STAIN NORMALIZATION IN FEDERATED LEARNING. WE REPORT THE RESULT WHEN THE PATTERN PRESERVING LOSS INCLUDED OR EXCLUDED. ($\mathcal{L}_{pp}$).

Mode	w/o $\mathcal{L}_{pp}$	w/ $\mathcal{L}_{pp}$
5 clients	0.711	0.951
6 clients	0.712	0.955
7 clients	0.723	0.954
8 clients	0.717	0.952

The pattern preserving loss ($\mathcal{L}_pp$) is designed specifically to address the structure-preserving issue. Accordingly, we first perform ablations to show the effectiveness of $\mathcal{L}_pp$ on the semantic structure information preservation, where the extent is evaluated with the SSIM between an original image and the normalized image. In Table VII, we observe a drastic decrease in SSIM where $\mathcal{L}_pp$ is absent.

TABLE VIII
ABLATIONS ON THE LOSS IN TRAINING STAIN NORMALIZATION GAN. WE MEASURE THE DOWNSTREAM PATCH-LEVEL CLASSIFICATION ACCURACY WITH GAN TRAINED BY DIFFERENT CONFIGURATIONS, WITH THE BACKBONE OF RESNET-18 AND 5-CLIENT SETTING.

$\mathcal{L}_{pp}$	$\mathcal{L}_{sd}$	Softened weighted coefficients	Acc (%)
x	x	x	82.071
x	x	✓	83.210
x	✓	✓	85.879
✓	x	✓	91.121
✓	✓	✓	94.318

V. LIMITATIONS

One major limitation is the restriction on the dynamic involvement of new clients, that extra client may fail to achieve the best performance when participating in the inference stage. Empirically, thoroughly different styles may slightly affect the performance or the interpolation of our GAN, when directly taking new clients in the inference stage without training. Hence our further work will close the gap by incorporating meta learning based personalized federated learning [59]. The extra meta test to personalize the trained generator for each client may further advance the model.

Another limitation is the risk that the local client can attack the globally shared global generator. For example, [60] attacks the generator in FL settings on the assumption that generated representations may share similar distributions with the training data. One potential privacy-preserving solution to avoid the attack is PRIVACYPRESERVING algorithm in [4] (Alg.2). More importantly, the proposed framework is complementary to all privacy-preserving methods, thus researchers may adopt any recent method to avoid attacks.

VI. CONCLUSION

The advent of Federate Learning well satisfies the requirement of data-privacy training across multiple centers, which is particularly important for the data-driven models of pathology image diagnosis and its stain normalization. Before FL, conventional normalization methods are stuck in routine centralized settings without a guarantee of data privacy. In this work, an innovative orchestration-generator, multiple-discriminator cGAN is proposed, which generates an adaptive stain style for decentralized institutions to deploy federated learning. The experimental results demonstrate the effectiveness of the proposed color transformation and privacy preservation methodology for the style-heterogeneous problem, as well as comparable full learning capacity and a higher convergence speed compared with prior routine centralized learning methods.

Moreover, the framework is applicable to a variety of histology-interpretation systems without organ-specific knowledge, and also to the other distributed settings feasible for a federated setting transfer. It paves a promising way to shift the centralized collaborative paradigm for histology to a distributed and data-private training mode. The motivation to mitigate the highly non-i.i.d. property in histopathology yields one more promising future direction, namely personalized federated learning, to solve the stain heterogeneity problem in cancer image analysis. Our future work will cover more large-scale data sets and more cancer types.

ACKNOWLEDGMENTS

This work has been supported by NSFC grant 62102247.

REFERENCES

[1] S. M. McKinney, M. Sieniek, V. Godbole *et al.*, "International evaluation of an ai system for breast cancer screening," *Nature*, vol. 577, no. 7788, pp. 89–94, 2020.
[2] B. McMahan, E. Moore, D. Ramage *et al.*, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*, 2017, pp. 1273–1282.
[3] M. Sheller, B. Edwards, G. Reina *et al.*, "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data," *Scientific Reports*, vol. 10, no. 1, pp. 1–12, 2020.
[4] W. Li, F. Milletarì, D. Xu *et al.*, "Privacy-preserving federated brain tumour segmentation," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2019, pp. 133–141.
[5] K. Chang, N. Balachandar, C. Lam, D. Yi, J. Brown, A. Beers, B. Rosen, D. L. Rubin, and J. Kalpathy-Cramer, "Distributed deep learning networks among institutions for medical imaging," *Journal of the American Medical Informatics Association*, vol. 25, no. 8, pp. 945–954, 2018.
[6] M. J. Sheller, G. A. Reina, B. Edwards, J. Martin, and S. Bakas, "Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation," in *International MICCAI Brainlesion Workshop*. Springer, 2018, pp. 92–104.

[7] Q. Liu, C. Chen, J. Qin, Q. Dou, and P.-A. Heng, "Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1013–1023.

[8] D. Onder, S. Zengin, and S. Sarioglu, "A review on color normalization and color deconvolution methods in histopathology," *Appl. Immunohistochem.*, vol. 22, no. 10, pp. 713–719, 2014.

[9] N. Coudray, P. Ocampo, T. Sakellaropoulos *et al.*, "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," *Nat. Med.*, vol. 24, no. 10, pp. 1559–1567, 2018.

[10] F. Ciompi, O. Geessink, B. Bejnordi *et al.*, "The importance of stain normalization in colorectal tissue classification with convolutional networks," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE, 2017, pp. 160–163.

[11] J. Ke, Y. Shen, and Y. Lu, "Style normalization in histology with federated learning," in *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021, pp. 953–956.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza *et al.*, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[13] A. Rana, G. Yauney, A. Lowe *et al.*, "Computational histological staining and destaining of prostate core biopsy rgb images with generative adversarial neural networks," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2018, pp. 828–834.

[14] J. Ke, Y. Shen, X. Jiang *et al.*, "Multiple-datasets and multiple-label based color normalization in histopathology with cgan," in *Medical Imaging 2021: Digital Pathology*, vol. 11603. International Society for Optics and Photonics, 2021, p. 1160310.

[15] T. Salimans, I. Goodfellow, W. Zaremba *et al.*, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, pp. 2234–2242, 2016.

[16] J. Ke, Y. Shen, X. Liang *et al.*, "Contrastive learning based stain normalization across multiple tumor histopathology," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 571–580.

[17] E. Reinhard, M. Adhikhmin, B. Gooch *et al.*, "Color transfer between images," *IEEE Computer graphics and applications*, vol. 21, no. 5, pp. 34–41, 2001.

[18] M. Macenko, M. Niethammer, J. Marron *et al.*, "A method for normalizing histology slides for quantitative analysis," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE, 2009, pp. 1107–1110.

[19] A. Khan, N. Rajpoot, D. Treanor *et al.*, "A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 6, pp. 1729–1738, 2014.

[20] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[21] H. Cho, S. Lim, G. Choi *et al.*, "Neural stain-style transfer learning using gan for histopathological images," *arXiv preprint arXiv:1710.08543*, 2017.

[22] P. Salehi and A. Chalechale, "Pix2pix-based stain-to-stain translation: A solution for robust stain normalization in histopathology images analysis," *arXiv preprint arXiv:2002.00647*, 2020.

[23] J.-Y. Zhu, T. Park, P. Isola *et al.*, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[24] M. Shaban, C. Baur, N. Navab *et al.*, "Staingan: Stain style transfer for digital histological images," in *IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 953–956.

[25] P. Kairouz, H. McMahan, B. Avent *et al.*, "Advances and open problems in federated learning," *arXiv preprint arXiv:1912.04977*, 2019.

[26] B. Liu, B. Yan, Y. Zhou *et al.*, "Experiments of federated learning for covid-19 chest x-ray images," *arXiv preprint arXiv:2007.05592*, 2020.

[27] M. Y. Lu, R. J. Chen, D. Kong, J. Lipkova, R. Singh, D. F. Williamson, T. Y. Chen, and F. Mahmood, "Federated learning for computational pathology on gigapixel whole slide images," *Medical image analysis*, vol. 76, p. 102298, 2022.

[28] M. Andreux, J. O. d. 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*. Springer, 2020, pp. 129–139.

[29] T. Chen, X. Zhai, M. Ritter *et al.*, "Self-supervised gans via auxiliary rotation loss," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 154–12 163.

[30] A. Anghel, M. Stanisavljevic, S. Andani, N. Papandreou, J. H. Rüschoff, P. Wild, M. Gabrani, and H. Pozidis, "A high-performance system for robust stain normalization of whole-slide images in histopathology," *Frontiers in medicine*, vol. 6, p. 193, 2019.

[31] S. Silva, B. Gutman, E. Romero *et al.*, "Federated learning in distributed medical databases: Meta-analysis of large-scale subcortical brain data," in *IEEE 16th international symposium on biomedical imaging (ISBI 2019)*. IEEE, 2019, pp. 270–274.

[32] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[33] I. Durugkar, I. Gemp, and S. Mahadevan, "Generative multi-adversarial networks," in *International Conference on Learning Representations*, 2017, pp. 1–14.

[34] C. Fan and P. Liu, "Federated generative adversarial learning," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2020, pp. 3–15.

[35] C. Hardy, E. Le Merrer, and B. Sericola, "Md-gan: Multi-discriminator generative adversarial networks for distributed datasets," in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, 2019, pp. 866–877.

[36] K. He, X. Zhang, S. Ren *et al.*, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[37] C. Doersch, "Tutorial on variational autoencoders," *arXiv preprint arXiv:1606.05908*, 2016.

[38] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 1195–1204.

[39] I. Mármol, C. Sánchez-de Diego, A. Pradilla D. *et al.*, "Colorectal carcinoma: a general overview and future perspectives in colorectal cancer," *International journal of molecular sciences*, vol. 18, no. 1, p. 197, 2017.

[40] J. Kather, J. Krisam, P. Charoentong *et al.*, "Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study," *PLoS medicine*, vol. 16, no. 1, p. e1002730, 2019.

[41] J. Ke, Y. Shen, Y. Guo *et al.*, "A prediction model of microsatellite status from histology images," in *Proceedings of the 2020 10th International Conference on Biomedical Engineering and Technology*, 2020, pp. 334–338.

[42] M. Macenko, M. Niethammer, J. Marron *et al.*, "A method for normalizing histology slides for quantitative analysis," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE, 2009, pp. 1107–1110.

[43] B. Bejnordi, G. Litjens, N. Timofeeva *et al.*, "Stain specific standardization of whole-slide histopathological images," *IEEE transactions on medical imaging*, vol. 35, no. 2, pp. 404–415, 2015.

[44] P. Bandi, O. Geessink, Q. Manson, M. Van Dijk, M. Balkenhol, M. Hermsen, B. E. Bejnordi, B. Lee, K. Paeng, A. Zhong *et al.*, "From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge," *IEEE transactions on medical imaging*, vol. 38, no. 2, pp. 550–560, 2018.

[45] Z. Wang, A. Bovik, H. Sheikh *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[46] L. Zhang, L. Zhang, X. Mou *et al.*, "Fsim: A feature similarity index for image quality assessment," *IEEE transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[47] A. Vahadane, T. Peng, A. Sethi *et al.*, "Structure-preserving color normalization and sparse stain separation for histological images," *IEEE transactions on medical imaging*, vol. 35, no. 8, pp. 1962–1971, 2016.

[48] G. Huang, Z. Liu, L. Van D. M. *et al.*, "Densely connected convolutional networks," in *Proceedings of CVPR*, 2017, pp. 4700–4708.

[49] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[50] A. Janowczyk, A. Basavanhally, and A. Madabhushi, "Stain normalization using sparse autoencoders (stanosa): application to digital pathology," *Computerized Medical Imaging and Graphics*, vol. 57, pp. 50–61, 2017.

[51] H. Nazki, O. Arandjelović, I. Um, and D. Harrison, "Multipathgan: Structure preserving stain normalization using unsupervised multi-domain adversarial network with perception loss," *arXiv preprint arXiv:2204.09782*, 2022.

[52] D. Tellez, G. Litjens, P. Bándi, W. Bulten, J.-M. Bokhorst, F. Ciompi, and J. Van Der Laak, "Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology," *Medical image analysis*, vol. 58, p. 101544, 2019.

[53] D. W. Zimmerman and B. D. Zumbo, "Relative power of the wilcoxon test, the friedman test, and repeated-measures anova on ranks," *The Journal of Experimental Education*, vol. 62, no. 1, pp. 75–86, 1993.

[54] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International journal of data mining & knowledge management process*, vol. 5, no. 2, p. 1, 2015.

[55] Y. Yeganeh, A. Farshad, N. Navab *et al.*, "Inverse distance aggregation for federated learning with non-iid data," in *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*. Springer, 2020, pp. 150–159.

[56] 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*, 2020.

[57] P. P. Liang, T. Liu, L. Ziyin, N. B. Allen, R. P. Auerbach, D. Brent, R. Salakhutdinov, and L.-P. Morency, "Think locally, act globally: Federated learning with local and global representations," *arXiv preprint arXiv:2001.01523*, 2020.

[58] M. G. Arivazhagan, V. Aggarwal, A. K. Singh, and S. Choudhary, "Federated learning with personalization layers," *arXiv preprint arXiv:1912.00818*, 2019.

[59] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning: A meta-learning approach," *arXiv preprint arXiv:2002.07948*, 2020.

[60] B. Hitaj, G. Ateniese, and F. Perez-Cruz, "Deep models under the gan: information leakage from collaborative deep learning," in *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, 2017, pp. 603–618.