PathFL: Multi-alignment Federated Learning for pathology image segmentation

Yuan Zhang [a], Feng Chen [b], Yaolei Qi [a], Guanyu Yang [a],*, Huazhu Fu [c],*

[a] Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, No. 2, Sipai Lou, Xuanwu District, Nanjing, 210096, China
[b] Department of Biostatistics, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, 211166, China
[c] Institute of High-Performance Computing, Agency for Science, Technology and Research, 138632, Singapore

ARTICLE INFO

ABSTRACT

Pathology image segmentation across multiple centers encounters significant challenges due to diverse sources of heterogeneity including imaging modalities, organs, and scanning equipment, whose variability brings representation bias and impedes the development of generalizable segmentation models. In this paper, we propose PathFL, a novel multi-alignment Federated Learning framework for pathology image segmentation that addresses these challenges through three-level alignment strategies of image, feature, and model aggregation. Firstly, at the image level, a collaborative style enhancement module aligns and diversifies local data by facilitating style information exchange across clients. Secondly, at the feature level, an adaptive feature alignment module ensures implicit alignment in the representation space by infusing local features with global insights, promoting consistency across heterogeneous client features learning. Finally, at the model aggregation level, a stratified similarity aggregation strategy hierarchically aligns and aggregates models on the server, using layer-specific similarity to account for client discrepancies and enhance global generalization. Comprehensive evaluations on four sets of heterogeneous pathology image datasets, encompassing cross-source, cross-modality, cross-organ, and cross-scanner variations, validate the effectiveness of our PathFL in achieving better performance and robustness against data heterogeneity. The code is available at https://github.com/yuanzhang7/PathFL.

1. Introduction

Pathology image segmentation plays a crucial role in clinical diagnostics, enabling the accurate identification and classification of various disease markers (Wang et al., 2019). In particular, accurate segmentation of nuclear, cellular, and tissue changes serves as a pivotal indicator for a pathologist in the evaluation of cancer malignancy degree and the identification of tumor regions (Irshad et al., 2013). However, the scarcity of annotated data in individual medical centers hinders its performance and generalizability. Furthermore, data sharing between different hospitals faces legal and practical hurdles in realistic clinical scenarios. Thus, the sensitive nature of pathology images, intertwined with patient privacy and ethical concerns, necessitates the adoption of multi-center collaborative training under privacy preservation techniques. The broader challenges of data ownership, regulatory restrictions, and institutional collaboration barriers remain significant obstacles in healthcare (Van Panhuis et al., 2014). Many

pathology datasets hold commercial and proprietary value, making institutions reluctant to share them freely (Rieke et al., 2020). Additionally, clinical pathology data are subject to strict governance, such as the US Health Insurance Portability and Accountability Act (Cheng and Hung, 2006) and the EU General Data Protection Regulation (Regulation, 2018), with concerns over data security, stewardship, and intellectual property rights. FL provides a viable pathway to overcome these challenges by allowing institutions to benefit from collaborative learning while maintaining control over their proprietary data. It also reduces the need for centralized data storage, mitigating risks associated with information leakage and data breaches. Moreover, FL supports large-scale, multi-institutional collaborations, allowing distributed learning across different centers while respecting data privacy and regulatory constraints (Sheller et al., 2020), thereby preserving confidentiality while promoting collaborative advancements in pathology image segmentation.

---

* Corresponding authors.

E-mail addresses: yuanzhang_@seu.edu.cn (Y. Zhang), fengchen@njmu.edu.cn (F. Chen), yaolei710@seu.edu.cn (Y. Qi), yang.list@seu.edu.cn (G. Yang), hzfu@ieee.org (H. Fu).
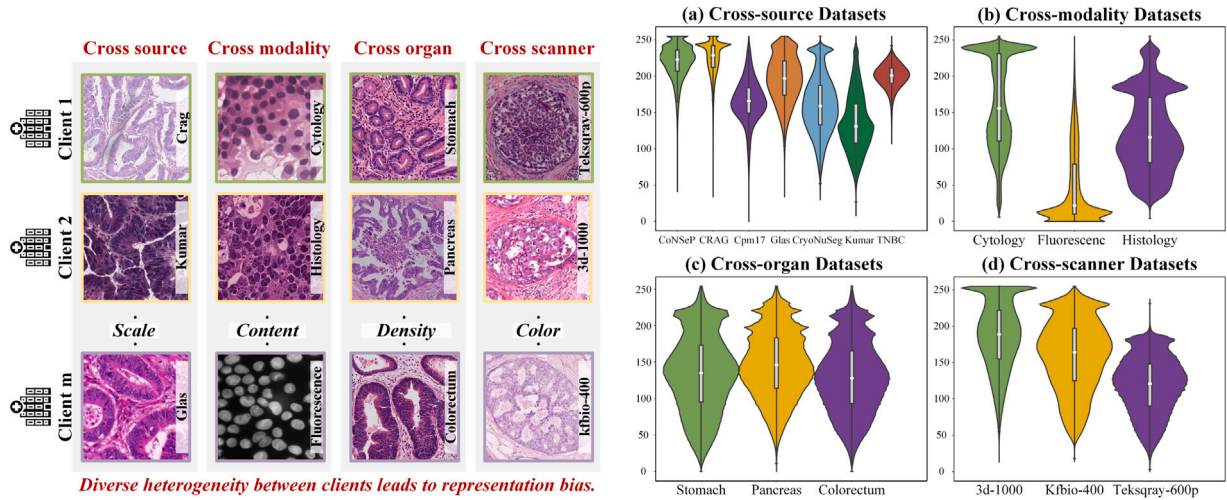
Fig. 1. Left: Illustration of heterogeneity for pathology images. The diverse sources of heterogeneity in multi-center pathology image segmentation highlight the significant representation bias across centers in four key aspects: cross-source, cross-modality, cross-organ, and cross-scanner. **Right:** Intensity distributions of four sets of pathology segmentation datasets. Violin plots reveal both intra-group and inter-group distribution differences, highlighting diverse heterogeneity.

The extensive heterogeneity inherent in pathology images from multiple centers presents a significant challenge for FL. Unlike gray-scale medical images, pathology images exhibit pronounced heterogeneity due to richer color variations, encompassing disparities in texture, scale, content, and density. These variations arise from multiple factors, including differences in diverse organs, imaging modalities, staining techniques, and image resolution, as depicted in Fig. 1. To comprehensively evaluate heterogeneity challenges in multi-center pathology segmentation, we constructed four datasets representing **cross-source**, **cross-modality**, **cross-organ**, and **cross-scanner** scenarios. The high diversity of morphological features amplifies representation bias, introducing substantial inductive bias that exacerbates divergence among local model updates and diminishes global generalization.

Although the effectiveness of FL for pathology segmentation has been demonstrated by the prior studies (Schoenpflug et al., 2024), it comes with limitations. There are three key factors that are greatly influenced by the heterogeneity: input images, latent features, and aggregation mechanisms. **First,** at the image level, the diversity of distribution differences in heterogeneous data from various sources is illustrated in Fig. 1. Variations in intensity distributions and visual characteristics of input images across different centers often lead to inconsistent learning patterns among clients, which may overfit to specific data distributions, leading to poor performance on unseen datasets. **Second,** at the feature level, the features extracted from heterogeneous inputs can vary significantly, leading to representation bias, where the global model disproportionately favors features from dominant or similar clients while neglecting essential features from distinct or smaller datasets. As a result, the model may not adequately capture the full range of characteristics present in the overall data population. **Lastly,** this heterogeneity amplifies imbalances in the aggregation process, as updates from clients with significantly divergent data can conflict or dominate, resulting in suboptimal updates to the global model and poor generalization performance. Most existing methods focus on only one aspect of heterogeneity, without integrating a comprehensive design that considers all three levels of heterogeneity simultaneously.

In this paper, we propose a multi-alignment framework to address heterogeneity by incorporating image-level style enhancement to diversify input distributions, representation-level feature alignment to unify latent feature spaces across clients, and model-level stratified similarity aggregation to balance and refine global model updates. We posit that by effectively leveraging statistical insights, the multi-layer exchange of shallow texture details and deep semantic features in segmentation

tasks can enhance the model's representational capability (Zhang et al., 2018). The richness of color in pathology images is fundamentally different from other medical images, and multi-center pathology data is naturally and specifically richer in stylistic information, yet this is often overlooked (Yamashita et al., 2021). This suggests that the structural and shape characteristics of cells captured in pathology images from different centers are fundamentally alike, even though they may be influenced by varying backgrounds, colors, and imaging conditions (Hoque et al., 2024). The underlying hypothesis is that cross-center pathology images exhibit similarities in cellular morphology. The style information of an image can be effectively characterized by statistical features, specifically the mean and variance (Huang and Belongie, 2017; Li et al., 2021a). The mean captures the overall brightness level of the image, while the variance provides insights into contrast and texture details. (1) By exchanging stylistic statistical information at the image level, we propose the **collaborative style enhancement** module to achieve cross-client style exchange, which not only enhances the model's adaptability to diverse data distributions but also aligns data across different centers. (2) By exchanging global mean and variance at the feature level, we propose the **adaptive feature alignment** module, which allows local models to explore the feature space more thoroughly, aligning the latent features learned by different clients. This collaborative mechanism helps unify the representations from multiple perspectives, reducing negative interference between clients. As a result, the global model's generalization performance is enhanced, with the capability to extract more universal features that improve segmentation accuracy. Moreover, only transmitting mean and variance without disclosing the actual feature content or raw data ensures compatibility with the privacy-preserving mechanism of federated learning, while providing a simple and efficient approach for aligning feature distributions across clients. (3) By evaluating layer-wise similarity, we proposed the **stratified similarity aggregation** strategy, which allows each client greater flexibility during model updates, aligns model parameters hierarchically at the server, effectively managing client discrepancies and improving generalization.

Our primary contributions are as follows:

- Framework: We propose PathFL, a multi-alignment framework to address cross-center heterogeneity in pathology image segmentation. To our knowledge, PathFL offers the first unified FL solution that aligns three levels of image, feature, and aggregation.
- Module design: At the image level, a collaborative style enhancement module diversifies local images by style exchange across clients, thereby enhancing the diversity of input and aligning

them across different centers. At the feature level, an adaptive feature alignment module ensures implicit alignment within the latent feature space, promoting consistency of representation. At the aggregation level, a stratified similarity aggregation strategy hierarchically aligns and aggregates models on the server, improving global generalization.

- Theoretical analysis: We provide a theoretical analysis of our algorithm, elucidating how cross-center style enhancement enhances model performance. Specifically, we demonstrate that exchanging statistical features, such as mean and variance, facilitates effective knowledge transfer, thereby improving model efficacy.
- Extensive validation: We conducted comprehensive experiments on four distinct pathology image segmentation datasets, exploring various types of heterogeneity commonly encountered. This demonstrates the effectiveness of our approach in handling cross-center heterogeneity.

2. Related works

2.1. Federated learning for pathology images

Collaborative training across centers risks privacy breaches and faces legal and logistical challenges in clinical practice, making data sharing between institutions impractical (Guan et al., 2024). FL is increasingly favored for pathology image analysis, offering a solution to leverage multi-institutional datasets without centralizing data, thus preserving privacy and addressing interoperability issues (Schoenpflug et al., 2024). Current Research on FL in pathology primarily focuses on three tasks: (1) Classification task. Many studies on pathology image classification (Shi et al., 2024) explore fairness (Hosseini et al., 2023), privacy (Lu et al., 2022), and communication cost (Deng et al., 2024), etc. (2) Staining normalization task. In pathological imaging, inconsistencies introduced by varying staining protocols increase the complexity of automated processing, driving considerable attention in the field of research (Wagner et al., 2022; Shen et al., 2022). (3) Segmentation task. Due to the scarcity of precise annotations for pathology images, research specifically focused on pathology image segmentation remains relatively limited. Ma et al. (2024) explores the issue of model heterogeneity in medical image segmentation, while Zhang et al. (2024) investigates the problem of catastrophic forgetting in medical image models. Additionally, Pan et al. (2024) emphasizes the importance of privacy protection and Chen et al. (2024a) evaluates the uncertainty through active learning. Given the significant clinical importance of pathology images, many FL methods for medical images incorporate pathology datasets to extend the applicability of their methods. However, these works often overlook the extensive and unique heterogeneity present in pathology images. In contrast, our approach is the first to conduct comprehensive experiments and validations specifically addressing the diverse heterogeneity in pathology image segmentation.

2.2. Heterogeneity in federated learning

Existing research in FL for pathology image segmentation has primarily addressed heterogeneity by focusing on isolated factors, without adequately considering the integration of input image, feature representation, and model aggregation (Qi et al., 2024). Firstly, HarmoFL (Jiang et al., 2022) addresses input image heterogeneity by normalizing images' frequency domain characteristics on pathology images. Secondly, FedBN (Li et al., 2021b) targets batch normalization layers to mitigate feature distribution discrepancies, and FedFA (Zhou and Konukoglu, 2023) augments local feature statistics to address feature shifts. Thirdly, numerous methods primarily focus on improving model aggregation mechanisms to mitigate inter-client variations. FedAvg (McMahan et al., 2017) applies the average aggregation of

clients, which struggles with non-IID data distributions. FedProx (Li et al., 2020) introduces a proximal regularization term to constrain local model divergence, while HistoFL (Lu et al., 2022) injects random noise into model weights to preserve privacy. FedHeal (Chen et al., 2024b) refines aggregation by pruning unimportant parameters and reducing variance between local and global models. Despite notable advancements, existing methods remain constrained by their narrow scope, typically addressing only one dimension of heterogeneity while overlooking the intricate interdependence between input, feature representation, and model aggregation. These three factors are not only directly influenced by heterogeneity but also mutually interact, compounding their impact and imposing significant constraints on performance. Thus, we propose a comprehensive multi-alignment framework that systematically addresses heterogeneity across input, feature, and model aggregation levels, thereby substantially enhancing federated learning robustness and generalization capabilities.

3. Methodology

In this section, we first outline the preliminaries and provide a detailed formulation of the problem. We then introduce our approach to stylistic information exchange, leveraging the principles of Vicinal Risk Minimization (VRM) to mitigate the impact of cross-client data heterogeneity. Specifically, we present three modules: the collaborative style enhancement module, the adaptive feature alignment module, and the stratified-layer aggregation strategy. An overview of our framework PathFL is depicted in Fig. 2.

3.1. Problem formulation

FL involves training a server model across M distinct clients without transferring data directly. Each client retains its own distribution $D_m = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n_m}$ $(m = 1, \ldots, M)$, where $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ represent the image-label pairs and n_m is the number of samples in client m. Due to the inherent variability in pathology images across source, modality, organ, and scanner, the samples in different clients exhibit *non-IID* (non-independent and identically distributed) characteristics, leading to significant heterogeneity across clients. More formally, for any two clients m and $\tilde{m}$, the expectation of their local loss functions $f_i(\boldsymbol{w}; \boldsymbol{x}_i)$ generally differs:

$$\mathbb{E}_{\boldsymbol{x}_i \sim D_m}[f_i(\boldsymbol{w}; \boldsymbol{x}_i)] \neq \mathbb{E}_{\boldsymbol{x}_i \sim D_{\tilde{m}}}[f_i(\boldsymbol{w}; \boldsymbol{x}_i)], \tag{1}$$

where $\boldsymbol{w}$ represents the model parameters. As a result, local datasets from different clients may not be representative of the global data distribution, i.e.,

$$\mathbb{E}_{\boldsymbol{x}_i \sim D_k}[f_i(\boldsymbol{w}; \boldsymbol{x}_i)] \neq F_{\text{global}}(\boldsymbol{w}), \tag{2}$$

which poses challenges for model generalization because the aggregated global model may not effectively capture the diverse distributions present across the clients. Furthermore, this heterogeneity can lead to imbalanced aggregation of the global model, resulting in degraded performance and compromised stability of the deep learning algorithm.

Theoretically, as the number of participating clients M increases and data heterogeneity diminishes, the global expectation $F_{\text{global}}(\boldsymbol{w})$ can asymptotically approach the weighted average of client-specific expectations. In the limit of infinite clients with sufficiently similar data distributions, this global model approximates the true data distribution:

$$\lim_{M \to \infty} \frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_{\boldsymbol{x}_i \sim D_m}[f(\boldsymbol{w}; \boldsymbol{x}_i)] \approx \mathbb{E}_{\boldsymbol{x}_i \sim D_{\text{global}}}[f(\boldsymbol{w}; \boldsymbol{x}_i)], \tag{3}$$

where D_{global} represents the cumulative distribution of data from all clients.

3.2. Theory: Vicinal risk minimization

Vicinal Risk Minimization (VRM) (Chapelle et al., 2000) improves the approximation of the local distribution by introducing "vicinal
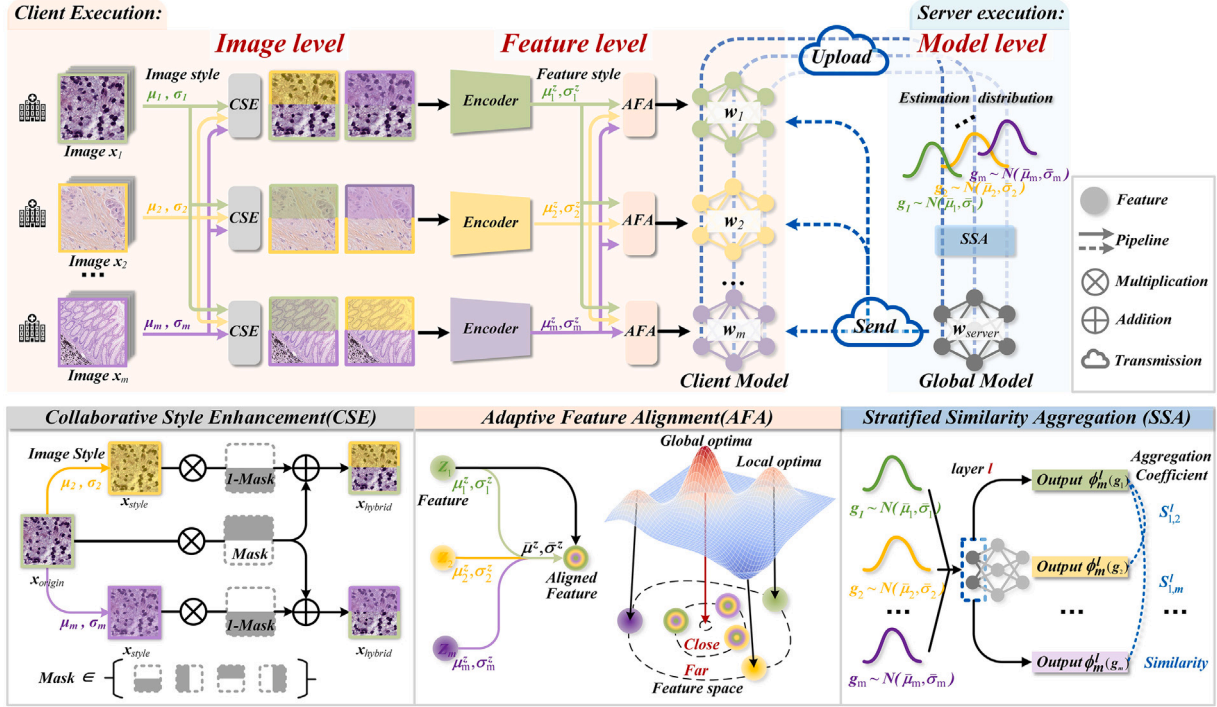
Fig. 2. Framework illustration of our PathFL, including three key modules: (1) Image-level collaborative style enhancement (CSE), which enriches the stylistic differences in images across clients by transferring essential statistical features; (2) Feature-level adaptive feature alignment (AFA), enabling the exchange of feature distributions between client models to improve alignment and mitigate heterogeneity; and (3) Model-level stratified similarity aggregation (SSA), which aligns and aggregates network layer outputs across clients to enhance model robustness and generalization.

distributions" of the data, which is achieved by incorporating statistical style information, such as mean and variance, from other clients to expand each client's local data distribution, forming a closer approximation of the global distribution D_{global}. VRM introduces flexibility into the optimization problem by broadening the scope of the local distribution. When clients receive style information from others, they effectively extend their data distribution to encompass patterns and features that were previously absent or unseen. Thus, each client can approximate a distribution $\mathcal{V}(D_m)$ that not only captures its unique data characteristics but also introduces variations aligned with the broader diverse patterns. This approximation provides robustness by mitigating local biases, as each client is less likely to overfit to its own potentially skewed data, resulting in improved model stability and performance across the federation. Specifically, the objective function of VRM is:

$$F_{\mathrm{VRM}}(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{V}(D)}\left[f(\boldsymbol{w}; \boldsymbol{x})\right], \tag{4}$$

where $\mathcal{V}(D)$ represents the vicinal distribution generated around the true data distribution D. In this manner, VRM not only accounts for the information of local samples but also integrates the variability of their neighboring samples, thereby capturing potential data features more comprehensively. By collaboratively aligning stylistic patterns across clients, our approach enriches sample representations, enabling the learning process to effectively transcend local data limitations and capture more generalized feature representations.

Mathematically, VRM can be viewed as an effective approximation of the global distribution. Specifically, as the number of clients M approaches infinity, the VRM theory can be expressed as:

$$F_{\mathrm{VRM}}(\boldsymbol{w}) \approx F_{\mathrm{global}}(\boldsymbol{w}), \tag{5}$$

which indicates VRM is better positioned to approximate the true data distribution, particularly in scenarios where the data distributions exhibit significant heterogeneity. By mitigating the discrepancy between local and global expectations through a VRM-based approach, we effectively enhance model generalization across diverse client distributions.

In the context of federated learning, where significant variability in client data distributions is prevalent, the proposed vicinal distribution methodology offers a more precise approximation of local data characteristics, consequently improving model generalization and robustness while systematically addressing the inherent challenges of *non-IID* data across heterogeneous client environments.

3.3. Collaborative style enhancement

Building upon the principles of VRM, we propose the collaborative style enhancement module that capitalizes on the variability introduced by neighboring distributions within the federated learning framework. For pathology image segmentation tasks, image-level style exchange is crucial as it contains the most detailed texture information necessary for effective analysis (Tang et al., 2024). By aligning style statistics between clients, we enrich local models with broader stylistic features. By integrating vicinal risk minimization (VRM) theory, our approach significantly enhances model generalization across heterogeneous datasets, effectively incorporating local variability to achieve a more comprehensive risk estimation. This methodological integration not only harmonizes with VRM principles but also fosters a robust model capable of seamlessly capturing diverse image representations in environments characterized by substantial data distribution variations. To quantify and align data distributions across different clients, we compute statistical features — specifically, the means and standard deviations — at two levels: (1) the image level, characterizing the global intensity distribution of pixel values within each client (Eqs. (6), (7)), and (2) the feature level, capturing the statistical properties of high-dimensional representations (Eqs. (10), (11)).

For client m at the tth training epoch in federated learning, we compute the mean $\mu_m^{(t)}$ and standard deviation $\sigma_m^{(t)}$ at each batch as the local stylistic statistics:

$$\mu_m^{(t)} = \frac{1}{UV} \sum_{u=1}^{U} \sum_{v=1}^{V} \boldsymbol{x}_m^{u,v(t)} \in \mathbb{R}, \tag{6}$$

$$\sigma_m^{(t)} = \sqrt{\frac{1}{UV} \sum_{u=1}^{U} \sum_{v=1}^{V} \left(\boldsymbol{x}_m^{u,v(t)} - \mu_m^{(t)} \right)^2} \in \mathbb{R}. \qquad (7)$$

Here, U and V represent the height and width of the images, and $\boldsymbol{x}_m^{u,v(t)}$ denotes the pixel intensity at position (u, v) for client m at epoch t. To facilitate style transfer, we aggregate these statistics from all clients at the tth epoch to form a global style pool, denoted by $\boldsymbol{\mu}_{\text{style}}^{(t)} = [\mu_1^{(t)}, \mu_2^{(t)}, \ldots, \mu_m^{(t)}], \boldsymbol{\sigma}_{\text{style}}^{(t)} = [\sigma_1^{(t)}, \sigma_2^{(t)}, \ldots, \sigma_m^{(t)}]$, where M is the total number of clients. These global style statistics are shared among clients, allowing for the collaborative alignment of stylistic patterns and the enhancement of local data into more comprehensive feature representations. In the $(t+1)$th epoch, each client m performs style transfer (Huang and Belongie, 2017) using the shared statistics from the global style pool, which is particularly well suited for our purposes due to its efficiency and compact representation. The transfer is mathematically expressed as:

$$\boldsymbol{x}_{m,style}^{(t+1)} = \left(\frac{\boldsymbol{x}_m^{(t+1)} - \mu_m^{(t+1)}}{\sigma_m^{(t+1)}} \right) \boldsymbol{\sigma}_{\text{style}}^{(t)} + \boldsymbol{\mu}_{\text{style}}^{(t)}, \qquad (8)$$

where $\mu_m^{(t)}$ and $\sigma_m^{(t)}$ are the mean and standard deviation of client m's images at the tth epoch. This exchange of style information among clients enhances the model's generalization ability in heterogeneous FL environments by introducing variability in image-level representations.

In the collaborative style enhancement module, we also introduce a random mask to balance the fidelity of the original image $\boldsymbol{x}_{origin}$ with the integration of stylistic variations $\boldsymbol{x}_{style} \in \{\boldsymbol{x}_{m,j}\}_{j=1}^{M}$. The random mask $\mathcal{M}$ is generated as a binary matrix with equal halves of zeros and ones, followed by random rotations of 0°, 90°, 180°, and 270°, including horizontal and vertical flips, as illustrated in Fig. 3(a). This ensures a balanced hybrid ratio between original and style-transferred images to generate hybrid images in Fig. 3(b). Specifically, the hybrid image $\boldsymbol{x}_{hybrid}$ is achieved using the following formula:

$$\boldsymbol{x}_{hybrid} = \mathcal{M} \otimes \boldsymbol{x}_m + (1 - \mathcal{M}) \otimes \boldsymbol{x}_{style} \qquad (9)$$

where $\otimes$ denotes the pixel-wise multiplication operator. This masking allows the model to retain the original semantic features while integrating style information from different clients. The combined representation captures both the inherent content and diverse stylistic features, enabling the model to improve segmentation accuracy across heterogeneous datasets.

3.4. Adaptive feature alignment

At the feature level, we propose an adaptive feature alignment module to ensure implicit alignment within the representation space by effectively integrating local feature details with global contextual information. Feature-level representations capture high-level semantic information essential for pathology image segmentation, enabling a better understanding of the contextual relationships between different structures within the image (Song et al., 2022). Each client's feature representations retain the unique characteristics of their local data while being aligned with global mean style statistics, promoting semantic consistency across clients.

Specifically, we calculate the average mean $\mu^{z(t)}$ and standard deviation $\sigma^{z(t)}$ of the deepest level feature maps Z_m at the tth training round, which represents the global distribution of high-level features across all centers. The mean and standard deviation are computed as follows:

$$\mu^{z(t)} = \frac{1}{MN} \sum_{m=1}^{M} \sum_{i=1}^{N} Z_m^{(t)}(i), \qquad (10)$$

$$\sigma^{z(t)} = \frac{1}{M} \sum_{m=1}^{M} \sqrt{\frac{1}{N} \sum_{i=1}^{N} (Z_m^{(t)}(i) - \mu_m^{z(t)})^2}, \qquad (11)$$
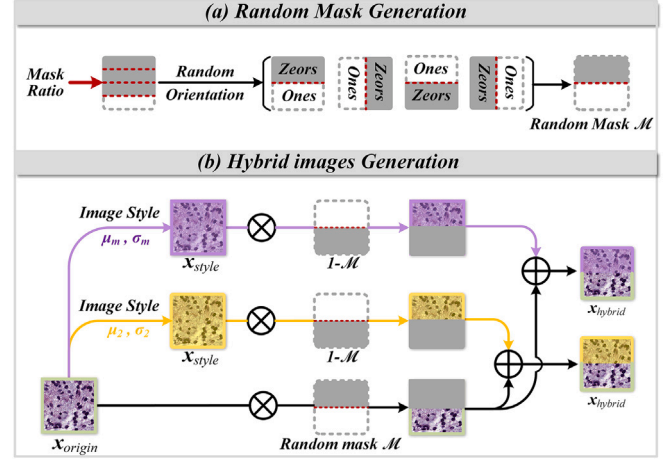

Fig. 3. The flowchart illustrates the generation process of random masks and hybrid images.

where N is the total number of feature elements in $Z_m^{(t)}$. The client m incorporates the statistics from the global feature style, thereby adapting its own feature representations of mean $\mu_m^{z(t)}$ and standard deviation $\sigma_m^{z(t)}$ and aligning high-level semantic representations by infusing local feature details with global contextual insights. The core formula for feature alignment is given by:

$$Z_{m,style}^{(t+1)} = \sigma^{z(t)} \left(\frac{Z_m^{(t+1)} - \mu_m^{z(t+1)}}{\sigma_m^{z(t+1)}} \right) + \mu^{z(t)}, \qquad (12)$$

where $Z_m^{(t)}$ and $Z_m^{(t+1)}$ represent the feature maps before and after the style alignment process, respectively. This implicit alignment harmonizes client-specific feature distributions within a shared semantic space, facilitating a cohesive understanding of structural information. By seamlessly integrating local features with global style information, the adaptive feature alignment module improves the model's ability to capture fine-grained boundary details and understand contextual relationships in multi-center heterogeneous data.

Overall, the collaborative style enhancement module enhances texture detail information at the image level, broadening the diversity of unseen visual patterns across different centers. Meanwhile, the adaptive feature alignment module aligns semantic representations at the feature level, promoting consistency and reducing representation bias. Together, these two modules effectively mitigate heterogeneity and enhance overall model performance in segmentation tasks.

3.5. Stratified similarity aggregation

We propose a stratified similarity aggregation strategy using Gaussian-distributed synthetic data to evaluate the similarity between layers of the network for better preservation of critical information while minimizing the impact of heterogeneity. For each client m, the mean $\overline{\mu}_m$ and standard deviation $\overline{\sigma}_m$ characterize the distribution of image intensities computed over its entire local dataset. During the t_{th} communication round, Gaussian synthetic data $g_m \sim \mathcal{N}(\overline{\mu}_m, \overline{\sigma}_m)$ is randomly generated and then uploaded to the server to facilitate similarity evaluation among clients. The entire client network is stratified into L layers, each exhibiting distinct feature representation capabilities. By employing this stratified similarity aggregation strategy, we achieve a more refined and flexible aggregation process that hierarchically aligns and dynamically integrates weights from both shallow and deep layers, enhancing size-specific representations amid heterogeneous data.

Let the feature output of client m at layer l be denoted as $\phi_m^l(g_m)$, obtained by passing the Gaussian synthetic data through the client's model up to layer l. For client m and j, the cosine similarity between

Table 1

Detailed information of four sets of heterogeneous pathology image datasets involved in this study.

Cross-source datasets							
Client	Dataset	Organ	Origin	Scanner	Magnification	Size	Number
C1	CoNSep	Colon	UHCW	Omnyx VL120 scanner	$40\times$	1000×1000	41
C2	CPM17	Multiple	TCGA	/	$20\times$ and $40\times$	500×500 and 600×600	32
C3	CRAG	Colon	UHCW	Omnyx VL120 scanner	$20\times$	1512×1516	213
C4	CryoNuSeg	Multiple	TCGA	/	$40\times$	512×512	30
C5	Glas	Colon	UHCW	Zeiss MIRAX MIDI scanner	$20\times$	775×522	165
C6	KUMAR	Multiple	TCGA	/	$40\times$	1000×1000	30
C7	TNBC	Breast	Curie Institute	Philips ultra fast scanner	$40\times$	512×512	50
Cross-modality datasets							
Client	Dataset	Organ	Modality	Scanner	Magnification	Size	Number
C1		Thyroid	Cytology	3CCD linear camera	$20\times$	224×224	487
C2	ClusterSeg	Embryos and archaea	Fluorescence	Zeiss confocal microscope	$63\times$	512×512	524
C3		Colorectal	Histology	/	$40\times$	512×512	462
Cross-organ datasets							
Client	Dataset	Organ	Modality	Scanner	Magnification	Size	Number
C1		Colorectum		TEKSQRAY SQS-600P	/	512×512	198
C2	COSAS2024 Task 1	Pancreas	Histology	TEKSQRAY SQS-600P	/	512×512	138
C3		Stomach		TEKSQRAY SQS-600P	/	512×512	141
Cross-scanner datasets							
Client	Dataset	Organ	Modality	Scanner	Magnification	Size	Number
C1		Breast		3d-1000	/	512×512	203
C2	COSAS2024 Task 2	Breast	Histology	Kfbio-400	/	512×512	193
C3		Breast		Teksqray-600p	/	512×512	205

layer l feature outputs is defined as $\text{CosSim}^l_{m,j} = \frac{\phi^l_m(g_m)\cdot\phi^l_j(g_j)}{|\phi^l_m(g_m)||\phi^l_j(g_j)|}$. Subsequently, the cosine similarity between each pair of layers is scaled by the total sum of all pairwise cosine similarities, denoted as $s^l_{m,j}$. This normalization ensures that the values are relative to the overall similarity distribution across all client pairs. Next, the weights of the server model are aggregated layer by layer based on the similarity evaluation interaction between clients' models. During the t communication round, the server model $w^{t,l}$ aggregates the weights from layer l of the $t-1$ round of client models with corresponding interaction weights as follows:

$$w^{l(t+1)}_{server} = \frac{1}{2M}\sum_{m=1}^{M}\sum_{j=1,j\neq m}^{M}[w^{(t)}_m + \sum_{j\neq k}s^l_{m,j}\cdot w^{(t)}_j]. \quad (13)$$

This hierarchical aggregation process culminates in the updated server model w^t_{server}, which is subsequently distributed to each client for the next training epoch, expressed as $w^{t+1}_m \leftarrow w^t_{server}$. The adoption of a stratified similarity aggregation strategy allows for the allocation of greater weight to layers that exhibit higher similarity during the model aggregation process, while assigning lower weight to layers that demonstrate lower similarity. By emphasizing layers with high similarity, the model enhances its ability to capture shared features, thereby improving overall performance and accuracy, while simultaneously mitigating the influence of low-similarity layers, which reduces noise and irrelevant information during aggregation. This improves the interactive flexibility of the server model, ensuring more effective model alignment. As a result, this strategy supports enhanced generalization capabilities and performance across diverse and heterogeneous data environments.

4. Results

4.1. Datasets

In our experiments, we validated the algorithm on four sets of multi-center datasets, each reflecting different dimensions of heterogeneity: cross-source, cross-modality, cross-organ, and cross-scanner. These four sets were selected to simulate the various heterogeneity challenges commonly encountered in real-world pathology clinical environments, where data sources (cross-source), imaging techniques

(cross-modality), anatomical regions (cross-organ), and scanning devices (cross-scanner) vary significantly. (a) **Cross-source datasets**: seven independent datasets of hematoxylin and eosin (HE) stained pathology images from different sources: CoNSeP (Graham et al., 2019b), CPM-17 (Vu et al., 2019), CRAG (Graham et al., 2019a), CryoNuSeg (Mahbod et al., 2021), Glas (Sirinukunwattana et al., 2017), Kumar (Kumar et al., 2017), and TNBC (Naylor et al., 2018). The cross-source datasets encompass different origins, disease types, organs, scanning devices, and resolutions, representing the most complex and realistic scenarios. (b) **Cross-modality datasets**: three datasets from ClusterSeg (Ke et al., 2023) with characteristic severely clustered nuclei across three image modalities: cytology, fluorescence, and histology. (c) **Cross-organ datasets**: three datasets from COSAS2024 Task 1 (Challenge, 2024) with the same adenocarcinoma across three organs: colorectum, pancreas, and stomach. (d) **Cross-scanner datasets**: three datasets from COSAS2024 Task 2 (Challenge, 2024) with the same organ scanned by three distinct WSI scanners. Samples from COSAS2024 are recropped into 512×512 patches, excluding empty background patches. (e) **Unified datasets**: a combination of all 16 datasets from Table 1. All datasets were then split into training and test sets following their original dataset partition protocol or with a ratio of $8:2$, maintaining this splitting strategy consistently across all experiments to ensure fair comparison. The detailed information for the four sets of datasets is summarized in Table 1, and each dataset is allocated to a single client.

4.2. Implementation details

In this work, we employ the U-Net (Ronneberger et al., 2015) as the basic client network, which is trained over 300 epochs with 5 local update epochs per communication round in all experiments involving cross-source, cross-organ, and cross-scanner data settings. For cross-modality datasets, training sets to 500 epochs, maintaining a 5-epoch communication round structure. To ensure fairness, all comparative experiments are initialized identically, with the same communication round configurations across experiments. For training, we consistently use a batch size of 4 and the Adam optimizer, configured with a learning rate of $1e^{-4}$ and momentum parameters set to 0.9 and 0.95. To ensure fairness in model training, we consistently apply Cross

Table 2
Cross-source results across different models and centers. The bold indicates the best value.

	Dice (%) ↑										ASSD (pix.) ↓									
	C1	C2	C3	C4	C5	C6	C7	Average	p-value	95% CI	C1	C2	C3	C4	C5	C6	C7	Average	p-value	95% CI
FedAvg (McMahan et al., 2017)	$80.48_{\pm3.82}$	$87.01_{\pm3.25}$	$84.09_{\pm8.15}$	$81.81_{\pm2.26}$	$86.97_{\pm8.21}$	$79.99_{\pm3.29}$	$78.01_{\pm5.96}$	$82.62_{\pm4.99}$	0.0013	[80.21, 84.97]	$2.89_{\pm0.74}$	$2.33_{\pm0.93}$	$46.23_{\pm42.14}$	$2.64_{\pm0.22}$	$19.27_{\pm18.41}$	$3.42_{\pm0.74}$	$6.01_{\pm4.80}$	$11.83_{\pm9.71}$	0.0436	[3.26, 17.50]
FedBN (Li et al., 2021b)	$81.46_{\pm3.59}$	$86.94_{\pm3.94}$	$82.24_{\pm5.56}$	$80.92_{\pm2.03}$	$87.63_{\pm6.80}$	$79.78_{\pm4.13}$	$\mathbf{79.93_{\pm3.87}}$	$82.70_{\pm4.28}$	0.0524	[80.55, 84.79]	$2.95_{\pm0.85}$	$2.23_{\pm0.71}$	$39.18_{\pm23.39}$	$3.17_{\pm0.47}$	$17.89_{\pm11.88}$	$3.52_{\pm0.82}$	$\mathbf{5.61_{\pm4.43}}$	$10.65_{\pm6.08}$	0.0985	[3.54, 14.45]
FedProx (Li et al., 2020)	$79.82_{\pm4.09}$	$86.79_{\pm3.58}$	$83.37_{\pm7.56}$	$80.89_{\pm2.66}$	$87.28_{\pm7.93}$	$79.58_{\pm3.90}$	$77.73_{\pm5.53}$	$82.21_{\pm5.04}$	0.0017	[79.65, 84.62]	$3.32_{\pm1.51}$	$2.41_{\pm1.02}$	$41.38_{\pm30.96}$	$2.87_{\pm0.16}$	$15.43_{\pm10.66}$	$3.39_{\pm0.78}$	$6.06_{\pm4.02}$	$10.69_{\pm7.01}$	0.0224	[3.13, 14.73]
HarmoFL (Jiang et al., 2022)	$79.69_{\pm3.79}$	$86.65_{\pm3.39}$	$80.33_{\pm7.67}$	$81.12_{\pm4.15}$	$86.54_{\pm7.60}$	$80.88_{\pm2.35}$	$78.38_{\pm5.75}$	$81.94_{\pm4.96}$	0.0171	[79.86, 84.09]	$4.49_{\pm4.63}$	$2.42_{\pm0.96}$	$44.14_{\pm34.98}$	$2.55_{\pm0.55}$	$17.37_{\pm13.03}$	$3.21_{\pm0.61}$	$6.55_{\pm5.09}$	$11.53_{\pm8.55}$	0.0216	[3.24, 16.68]
HistoFL (Lu et al., 2022)	$80.32_{\pm3.26}$	$87.01_{\pm3.44}$	$83.68_{\pm7.34}$	$81.60_{\pm3.04}$	$87.69_{\pm7.38}$	$80.37_{\pm3.28}$	$78.37_{\pm5.95}$	$82.72_{\pm4.81}$	0.0049	[80.13, 85.21]	$3.23_{\pm1.27}$	$2.38_{\pm0.96}$	$44.19_{\pm33.38}$	$2.70_{\pm0.17}$	$16.96_{\pm13.14}$	$3.26_{\pm0.69}$	$6.80_{\pm5.23}$	$11.36_{\pm7.83}$	0.0290	[3.29, 15.84]
FedFA (Zhou and Konukoglu, 2023)	$80.11_{\pm3.75}$	$86.80_{\pm3.23}$	$83.99_{\pm6.20}$	$81.20_{\pm3.39}$	$86.21_{\pm7.92}$	$80.32_{\pm3.19}$	$78.07_{\pm6.70}$	$82.39_{\pm4.91}$	0.0027	[80.22, 84.59]	$3.34_{\pm1.16}$	$2.30_{\pm0.82}$	$42.21_{\pm33.12}$	$2.73_{\pm0.17}$	$18.79_{\pm18.52}$	$3.23_{\pm0.67}$	$8.06_{\pm8.02}$	$11.52_{\pm8.93}$	0.0114	[3.60, 15.67]
FedHEAL (Chen et al., 2024b)	$80.48_{\pm3.58}$	$86.89_{\pm3.45}$	$83.86_{\pm8.50}$	$81.75_{\pm3.43}$	$87.33_{\pm7.27}$	$80.19_{\pm3.87}$	$77.64_{\pm5.87}$	$82.59_{\pm5.14}$	0.0011	[80.25, 84.98]	$2.95_{\pm0.75}$	$2.29_{\pm0.84}$	$38.10_{\pm29.63}$	$2.65_{\pm0.15}$	$17.45_{\pm14.05}$	$3.25_{\pm0.76}$	$8.04_{\pm7.89}$	$10.68_{\pm7.73}$	0.0079	[3.23, 14.72]
PathFL (Ours)	$\mathbf{82.06_{\pm3.57}}$	$\mathbf{87.46_{\pm3.24}}$	$\mathbf{86.12_{\pm5.02}}$	$\mathbf{84.72_{\pm3.44}}$	$\mathbf{88.96_{\pm8.27}}$	$\mathbf{81.44_{\pm2.75}}$	$79.17_{\pm4.70}$	$\mathbf{84.27_{\pm4.43}}$	–	[81.72, 86.73]	$\mathbf{2.76_{\pm0.73}}$	$\mathbf{2.16_{\pm0.77}}$	$\mathbf{36.47_{\pm26.28}}$	$\mathbf{2.14_{\pm0.46}}$	$\mathbf{13.96_{\pm10.80}}$	$\mathbf{2.99_{\pm0.53}}$	$5.94_{\pm4.92}$	$\mathbf{9.49_{\pm6.36}}$	–	[2.53, 13.20]

Table 3

Cross-modality results across different models and centers. The bold indicates the best value.

	Dice (%) ↑						ASSD (pix.) ↓					
	C1	C2	C3	Average	p-value	95% CI	C1	C2	C3	Average	p-value	95% CI
FedAvg (McMahan et al., 2017)	$83.41_{\pm9.85}$	$91.46_{\pm4.64}$	$73.53_{\pm6.71}$	$82.80_{\pm7.07}$	0.0240	[76.00, 89.45]	$6.22_{\pm8.70}$	$6.72_{\pm4.42}$	$9.36_{\pm4.66}$	$7.43_{\pm5.93}$	0.0271	[6.47, 8.70]
FedBN (Li et al., 2021b)	$87.07_{\pm5.24}$	$90.45_{\pm4.66}$	$77.38_{\pm5.29}$	$84.97_{\pm5.06}$	0.0069	[79.80, 89.08]	$4.47_{\pm3.69}$	$12.49_{\pm11.89}$	$7.69_{\pm3.11}$	$8.21_{\pm6.23}$	0.0755	[5.40, 11.29]
FedProx (Li et al., 2020)	$73.20_{\pm11.98}$	$87.82_{\pm7.55}$	$63.00_{\pm7.67}$	$74.67_{\pm9.06}$	0.0173	[65.91, 84.53]	$13.07_{\pm7.73}$	$17.44_{\pm19.06}$	$15.00_{\pm6.08}$	$15.17_{\pm10.96}$	0.0002	[13.60, 16.87]
HarmoFL (Jiang et al., 2022)	$81.80_{\pm6.85}$	$89.04_{\pm5.75}$	$69.55_{\pm6.88}$	$80.13_{\pm6.49}$	0.0157	[72.61, 85.45]	$8.55_{\pm8.79}$	$11.12_{\pm6.74}$	$11.38_{\pm4.97}$	$10.35_{\pm6.83}$	0.0001	[9.19, 11.25]
HistoFL (Lu et al., 2022)	$83.01_{\pm9.33}$	$92.74_{\pm3.68}$	$75.59_{\pm6.91}$	$83.78_{\pm6.64}$	0.0363	[77.64, 90.31]	$7.15_{\pm8.91}$	$7.59_{\pm11.42}$	$8.48_{\pm4.50}$	$7.74_{\pm8.28}$	0.0052	[7.30, 8.26]
FedFA (Zhou and Konukoglu, 2023)	$83.32_{\pm7.31}$	$90.69_{\pm4.53}$	$73.49_{\pm6.15}$	$82.50_{\pm6.00}$	0.0156	[75.95, 88.65]	$7.58_{\pm8.93}$	$10.40_{\pm11.49}$	$9.75_{\pm4.71}$	$9.24_{\pm8.38}$	3.798E−05	[8.12, 10.07]
FedHEAL (Chen et al., 2024b)	$83.22_{\pm9.49}$	$91.68_{\pm4.30}$	$74.53_{\pm6.59}$	$83.14_{\pm6.79}$	0.0220	[76.70, 89.55]	$5.51_{\pm6.83}$	$9.91_{\pm13.10}$	$8.90_{\pm4.39}$	$8.11_{\pm8.11}$	0.0056	[6.36, 9.46]
PathFL (Ours)	$\mathbf{89.93_{\pm3.49}}$	$\mathbf{93.88_{\pm3.10}}$	$\mathbf{83.00_{\pm5.30}}$	$\mathbf{88.93_{\pm3.96}}$	–	[84.48, 91.92]	$\mathbf{3.54_{\pm4.62}}$	$\mathbf{5.81_{\pm6.69}}$	$\mathbf{5.44_{\pm3.60}}$	$\mathbf{4.93_{\pm4.97}}$	–	[4.01, 5.63]

Entropy Loss as the unified loss function to evaluate the discrepancy between predicted segmentation results and ground truth labels. Our experiments are executed on a cluster of four 40 GB NVIDIA A6000 GPUs. All implementations are completed using PyTorch.

To comprehensively evaluate segmentation performance, we employ two complementary metrics: the Dice similarity coefficient (Dice) and the Average Symmetric Surface Distance (ASSD) (Yeghiazaryan and Voiculescu, 2018). The Dice coefficient, widely used in medical image segmentation, quantifies the overlap between predictions and labels. By balancing sensitivity and precision, it provides a robust measure of overall segmentation similarity. ASSD focuses on boundary alignment, capturing the average discrepancy between predicted and ground truth surfaces. This makes it particularly valuable for assessing the precision of segmentation boundaries, especially in high-resolution medical images where accurate delineation is critical. Together, these metrics offer a holistic segmentation evaluation. We used paired t-tests and 95% confidence intervals to evaluate statistical significance, with p-value < 0.05 indicating significant differences between methods.

4.3. Comparison with the state-of-the-arts

For a comprehensive evaluation, we benchmark our approach against state-of-the-art (SOTA) FL methods designed to address heterogeneity. The comparison includes classic methods like FedAvg, pathology-specific approaches such as HistoFL, input-based techniques like HarmoFL, and feature-level strategies including FedBN and FedFA. Additionally, we evaluate aggregation-focused methods such as FedProx, HistoFL, and FedHEAL to assess their effectiveness in managing task heterogeneity within federated learning frameworks.

4.3.1. Performance on cross-source datasets

Our proposed model PathFL achieves the best performance on the cross-source datasets, attaining an average Dice of 84.27% and an average ASSD of 9.49, with notable improvements of 1.55% in the Dice and 1.65 in ASSD compared to the baseline (FedAvg). The quantitative segmentation results of Dice and ASSD on the cross-source datasets for seven different centers (C1 to C7) are summarized in Table 2. The cross-source dataset represents the most challenging clinical scenario, characterized by the largest number of centers, the highest degree of data imbalance, and significant inter-center heterogeneity. Importantly, our method demonstrates consistent performance across all centers without sacrificing the accuracy of any specific center for the benefit of another. This balanced performance validates that our approach can adapt effectively to heterogeneous clients, providing a tailored yet generalizable solution across diverse data distributions without compromising the segmentation quality of any center. Furthermore, this performance is particularly notable in most centers, especially in C3 and C5, with Dice of 86.12% and 88.96%, respectively. Both C3 and C5 correspond to datasets derived from colon organs, highlighting the capability of our method to not only mitigate the interference caused by heterogeneity but also to effectively identify and leverage clients with

representational similarities for more efficient aggregation. In conclusion, PathFL effectively addresses inter-center heterogeneity and data imbalance across cross-source datasets. By leveraging representational similarities, PathFL achieves balanced performance and demonstrates promising generalizability in diverse clinical settings.

4.3.2. Performance on cross-modality datasets

Our PathFL achieves superior performance on the cross-modality dataset with the highest average Dice of 88.93% and the lowest average ASSD of 4.93, greatly outperforming the baseline (FedAvg) by 6.13% in Dice and 2.50 in ASSD. The quantitative segmentation results on the cross-modality datasets for three centers (C1 to C3) are summarized in Table 3. Unlike the cross-source dataset, which consists of single modality H&E-stained pathology images, the cross-modality dataset addresses challenges arising from diverse imaging modalities of cytology, fluorescence, and histology. Our model effectively handles cross-modality heterogeneity by leveraging similarities in cellular patterns, which remain consistent across different modalities. Other methods designed for single-modality heterogeneity often fail to adequately address these complexities, resulting in mutual degradation among client models. While our PathFL effectively enhances performance through its triple heterogeneity balancing strategy, which takes into account variations in images, features, and model aggregation. Specifically, PathFL shows substantial enhancements over FedAvg, with Dice increases of 6.52% in C1, 2.42% in C2, and 9.47% in C3. These results highlight PathFL's capability to effectively address the challenges presented by diverse imaging modalities and maintain consistent performance.

4.3.3. Performance on cross-organ datasets

Our PathFL achieves superior performance on the cross-organ dataset with the highest average Dice of 83.05% and the lowest average ASSD of 41.54, greatly outperforming FedAvg by 7.94% in Dice and 40.12 in ASSD. The quantitative segmentation results on the cross-organ datasets for three centers (C1 to C3) are summarized in Table 4. The cross-organ dataset focuses on a consistent adenocarcinoma segmentation task across various organs, featuring three distinct adenocarcinomas, all acquired from the same scanner. This uniformity allows for a clearer comparison of model performance across different anatomical contexts. The results demonstrate that our method consistently outperforms other SOTAs, achieving Dice of 85.65% at C1, 83.36% at C2, and 80.16% at C3. Moreover, PathFL significantly reduces ASSD compared to the baseline, with decreases of 27.50 at C1, 43.77 at C2, and 49.09 at C3, highlighting its enhanced boundary accuracy and segmentation consistency. In contrast, other approaches like FedProx and FedFA, exhibit limitations in effectively addressing this cross-organ variability, with average ASSD of 109.10 and 97.37, respectively. These methods often struggle to fully account for substantial density variations and anatomical differences across diverse datasets, consequently compromising segmentation performance. PathFL uniquely mitigates these heterogeneity challenges through a comprehensive, multi-perspective approach, substantially enhancing segmentation robustness and accuracy.

Table 4

Cross-organ results across different models and centers. The bold indicates the best value.

	Dice (%) ↑						ASSD (pix.) ↓					
	C1	C2	C3	Average	p-value	95% CI	C1	C2	C3	Average	p-value	95% CI
FedAvg (McMahan et al., 2017)	$79.08_{\pm16.72}$	$72.79_{\pm20.83}$	$73.47_{\pm22.67}$	$75.11_{\pm20.07}$	0.0264	[72.79, 79.08]	$65.39_{\pm56.70}$	$87.61_{\pm69.36}$	$91.98_{\pm69.44}$	$81.66_{\pm65.17}$	0.0488	[74.70, 153.42]
FedBN (Li et al., 2021b)	$83.15_{\pm15.58}$	$77.08_{\pm15.90}$	$73.64_{\pm22.26}$	$77.95_{\pm17.92}$	0.0116	[74.71, 81.63]	$45.93_{\pm45.05}$	$65.68_{\pm63.65}$	$82.80_{\pm62.37}$	$64.81_{\pm57.02}$	0.0269	[61.65, 113.04]
FedProx (Li et al., 2020)	$75.54_{\pm18.42}$	$70.17_{\pm21.48}$	$70.10_{\pm25.08}$	$71.94_{\pm21.66}$	0.0086	[70.10, 75.54]	$101.98_{\pm55.69}$	$111.89_{\pm66.33}$	$113.44_{\pm68.01}$	$109.10_{\pm63.34}$	0.0008	[111.29, 153.78]
HarmoFL (Jiang et al., 2022)	$74.23_{\pm18.86}$	$69.23_{\pm21.78}$	$69.66_{\pm25.21}$	$71.04_{\pm21.95}$	0.0081	[69.23, 72.70]	$67.20_{\pm32.22}$	$73.58_{\pm40.26}$	$73.18_{\pm43.77}$	$71.32_{\pm38.75}$	0.0001	[73.58, 122.39]
HistoFL (Lu et al., 2022)	$78.06_{\pm17.39}$	$72.90_{\pm19.85}$	$75.19_{\pm21.96}$	$75.38_{\pm19.73}$	0.0402	[72.90, 78.06]	$71.05_{\pm41.71}$	$88.74_{\pm70.51}$	$82.14_{\pm66.59}$	$80.64_{\pm59.60}$	0.0172	[80.36, 143.66]
FedFA (Zhou and Konukoglu, 2023)	$77.16_{\pm17.92}$	$71.54_{\pm20.61}$	$71.89_{\pm23.82}$	$73.53_{\pm20.78}$	0.0142	[71.54, 77.16]	$80.55_{\pm49.43}$	$101.44_{\pm70.07}$	$110.13_{\pm74.09}$	$97.37_{\pm64.53}$	0.0105	[89.86, 150.48]
FedHEAL (Chen et al., 2024b)	$78.81_{\pm17.02}$	$74.28_{\pm19.57}$	$74.84_{\pm22.01}$	$75.98_{\pm19.53}$	0.0230	[74.28, 78.81]	$60.96_{\pm39.53}$	$82.73_{\pm74.09}$	$85.62_{\pm72.96}$	$76.43_{\pm62.19}$	0.0352	[70.27, 136.87]
PathFL (Ours)	**$85.65_{\pm10.86}$**	**$83.36_{\pm9.21}$**	**$80.16_{\pm17.62}$**	**$83.05_{\pm12.57}$**	–	[80.16, 85.65]	**$37.89_{\pm23.65}$**	**$43.84_{\pm44.77}$**	**$42.89_{\pm33.30}$**	**$41.54_{\pm33.91}$**	–	[43.84, 91.78]

Table 5

Cross-scanner results across different models and centers. The bold indicates the best value.

	Dice (%) ↑						ASSD (pix.) ↓					
	C1	C2	C3	Average	p-value	95% CI	C1	C2	C3	Average	p-value	95% CI
FedAvg (McMahan et al., 2017)	$77.59_{\pm18.65}$	$80.31_{\pm20.72}$	$72.78_{\pm24.25}$	$76.90_{\pm21.21}$	0.0127	[74.66, 79.46]	$59.23_{\pm38.42}$	$80.53_{\pm76.31}$	$89.79_{\pm71.98}$	$76.51_{\pm62.24}$	0.0088	[123.58, 170.89]
FedBN (Li et al., 2021b)	$77.60_{\pm17.73}$	$84.52_{\pm14.18}$	$77.35_{\pm18.61}$	$79.82_{\pm16.84}$	0.0232	[77.47, 82.79]	$52.05_{\pm35.54}$	$28.90_{\pm23.91}$	$60.80_{\pm64.70}$	$47.25_{\pm41.38}$	0.0784	[109.92, 128.69]
FedProx (Li et al., 2020)	$75.33_{\pm20.21}$	$79.26_{\pm22.62}$	$69.46_{\pm26.26}$	$74.68_{\pm23.03}$	0.0114	[70.93, 78.12]	$86.48_{\pm51.52}$	$97.69_{\pm80.05}$	$126.61_{\pm69.25}$	$103.59_{\pm66.94}$	0.0066	[144.37, 194.54]
HarmoFL (Jiang et al., 2022)	$74.79_{\pm20.27}$	$79.10_{\pm22.87}$	$69.25_{\pm26.32}$	$74.38_{\pm23.15}$	0.0098	[70.64, 77.92]	$54.73_{\pm31.15}$	$50.54_{\pm46.21}$	$73.82_{\pm45.23}$	$59.69_{\pm40.86}$	0.0025	[124.49, 162.58]
HistoFL (Lu et al., 2022)	$76.11_{\pm19.55}$	$80.36_{\pm21.38}$	$72.35_{\pm24.73}$	$76.27_{\pm21.88}$	0.0075	[73.33, 79.30]	$76.15_{\pm51.59}$	$72.34_{\pm69.53}$	$75.01_{\pm54.81}$	$74.50_{\pm58.64}$	0.0003	[125.48, 164.95]
FedFA (Zhou and Konukoglu, 2023)	$75.45_{\pm19.84}$	$79.89_{\pm22.14}$	$70.43_{\pm25.76}$	$75.25_{\pm22.58}$	0.0106	[71.68, 78.73]	$85.36_{\pm52.58}$	$75.31_{\pm70.48}$	$93.52_{\pm65.66}$	$84.73_{\pm62.91}$	0.0009	[138.71, 167.64]
FedHEAL (Chen et al., 2024b)	$76.62_{\pm19.19}$	$80.62_{\pm21.00}$	$73.28_{\pm24.15}$	$76.84_{\pm21.45}$	0.0098	[73.00, 79.45]	$74.48_{\pm51.68}$	$68.47_{\pm71.45}$	$88.56_{\pm74.99}$	$77.17_{\pm66.04}$	0.0019	[137.69, 170.04]
PathFL (Ours)	**$81.59_{\pm14.06}$**	**$85.77_{\pm11.34}$**	**$82.38_{\pm14.90}$**	**$83.25_{\pm13.43}$**	–	[81.99, 85.14]	**$32.59_{\pm22.79}$**	**$26.57_{\pm23.31}$**	**$42.06_{\pm33.82}$**	**$33.74_{\pm26.64}$**	–	[87.76, 124.40]

Table 6

Unified Datasets Results across different models and clients. The bold indicates the best value.

	Dice (%) ↑																		
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	Average	p-value	95% CI
FedAvg	$79.95_{\pm3.51}$	$86.33_{\pm3.47}$	$81.32_{\pm10.40}$	$80.11_{\pm1.35}$	$79.80_{\pm9.80}$	$80.33_{\pm3.30}$	$75.90_{\pm8.56}$	$77.04_{\pm10.26}$	$89.86_{\pm5.64}$	$67.82_{\pm7.97}$	$78.71_{\pm16.61}$	$71.25_{\pm19.97}$	$71.49_{\pm23.30}$	$76.46_{\pm18.68}$	$81.03_{\pm20.27}$	$73.87_{\pm23.18}$	$78.49_{\pm11.64}$	1.123E−05	[75.85, 81.30]
FedBN	$81.31_{\pm2.72}$	$86.87_{\pm3.64}$	$82.06_{\pm6.33}$	$81.74_{\pm4.17}$	**$87.07_{\pm7.01}$**	$79.49_{\pm4.70}$	$79.63_{\pm4.86}$	$83.45_{\pm13.27}$	$88.53_{\pm6.00}$	$70.57_{\pm6.77}$	$76.87_{\pm18.20}$	$73.12_{\pm19.33}$	$70.95_{\pm24.27}$	$75.54_{\pm19.75}$	$79.88_{\pm21.26}$	$72.51_{\pm23.61}$	$79.35_{\pm11.04}$	2.501E−04	[76.55, 81.95]
FedProx	$80.11_{\pm3.39}$	$86.33_{\pm3.35}$	$81.22_{\pm9.30}$	$80.59_{\pm2.85}$	$84.69_{\pm2.85}$	$80.67_{\pm3.36}$	$77.00_{\pm4.81}$	$70.84_{\pm13.27}$	$88.55_{\pm5.71}$	$66.13_{\pm8.41}$	$81.07_{\pm15.92}$	$70.57_{\pm20.63}$	$72.21_{\pm22.01}$	$75.72_{\pm19.42}$	$80.42_{\pm20.88}$	$73.61_{\pm23.45}$	$78.11_{\pm11.04}$	5.084E−04	[75.27, 80.98]
HarmoFL	$78.78_{\pm3.98}$	$86.16_{\pm3.13}$	$78.82_{\pm9.04}$	$81.17_{\pm5.36}$	$85.02_{\pm8.77}$	$80.67_{\pm2.23}$	$75.95_{\pm4.89}$	$73.69_{\pm8.98}$	$87.02_{\pm6.99}$	$64.49_{\pm8.82}$	$74.35_{\pm18.82}$	$69.24_{\pm21.77}$	$69.66_{\pm20.17}$	$74.90_{\pm20.42}$	$79.49_{\pm26.27}$	$69.34_{\pm26.27}$	$76.80_{\pm12.47}$	3.382E−05	[73.44, 79.97]
HistoFL	$81.19_{\pm2.55}$	$86.78_{\pm3.94}$	$81.73_{\pm6.03}$	$81.76_{\pm1.08}$	$86.37_{\pm7.15}$	$78.87_{\pm4.90}$	**$80.65_{\pm4.75}$**	$83.09_{\pm4.51}$	$88.10_{\pm6.45}$	$72.11_{\pm6.43}$	$78.48_{\pm17.13}$	$72.68_{\pm19.47}$	$71.05_{\pm24.60}$	$75.40_{\pm19.44}$	$79.98_{\pm20.51}$	$72.03_{\pm23.86}$	$79.39_{\pm10.80}$	2.648E−04	[76.74, 81.91]
FedFA	$79.86_{\pm3.16}$	$86.51_{\pm3.36}$	$81.27_{\pm9.22}$	$80.97_{\pm3.83}$	$84.94_{\pm8.74}$	$80.24_{\pm3.73}$	$74.47_{\pm12.32}$	$73.04_{\pm13.23}$	$89.05_{\pm6.80}$	$66.95_{\pm7.65}$	$77.92_{\pm16.29}$	$69.90_{\pm20.80}$	$72.16_{\pm22.44}$	$75.89_{\pm19.04}$	$80.63_{\pm20.87}$	$72.49_{\pm24.14}$	$78.08_{\pm12.23}$	8.762E−05	[75.25, 80.81]
FedHEAL	$76.33_{\pm4.36}$	$84.96_{\pm3.70}$	$76.06_{\pm20.61}$	$79.67_{\pm3.50}$	$80.25_{\pm9.49}$	$70.49_{\pm4.63}$	$63.55_{\pm11.89}$	$84.02_{\pm5.88}$	$89.51_{\pm5.99}$	$73.27_{\pm6.66}$	$74.23_{\pm18.87}$	$69.22_{\pm21.79}$	$69.66_{\pm25.21}$	$74.80_{\pm20.27}$	$79.17_{\pm22.85}$	$69.25_{\pm26.32}$	$74.65_{\pm13.25}$	4.127E−04	[70.82, 78.63]
PathFL (Ours)	**$81.86_{\pm3.41}$**	**$87.08_{\pm3.69}$**	**$85.72_{\pm7.48}$**	**$84.02_{\pm3.91}$**	$86.71_{\pm2.27}$	**$81.09_{\pm5.69}$**	$78.48_{\pm3.88}$	**$87.25_{\pm4.28}$**	**$93.00_{\pm6.65}$**	**$74.67_{\pm12.44}$**	**$84.00_{\pm18.24}$**	**$75.33_{\pm22.10}$**	**$73.79_{\pm16.40}$**	**$79.85_{\pm15.43}$**	**$85.33_{\pm20.35}$**	**$80.50_{\pm9.72}$**	**$82.42_{\pm9.72}$**	–	[79.95, 84.97]

	ASSD (pix.) ↓																		
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	Average	p-value	95% CI
FedAvg	$2.88_{\pm0.67}$	$2.46_{\pm0.93}$	$51.45_{\pm40.47}$	$2.91_{\pm0.22}$	$19.72_{\pm18.56}$	$3.29_{\pm0.77}$	$7.20_{\pm6.38}$	$10.53_{\pm8.51}$	$14.32_{\pm4.09}$	$12.48_{\pm6.00}$	$88.15_{\pm88.44}$	$109.28_{\pm68.32}$	$153.95_{\pm95.85}$	$128.16_{\pm98.02}$	$178.28_{\pm131.17}$	$132.72_{\pm119.74}$	$57.36_{\pm43.78}$	2.921E−03	[30.21, 89.83]
FedBN	$2.97_{\pm0.70}$	**$2.40_{\pm0.84}$**	$39.75_{\pm22.45}$	$2.39_{\pm0.50}$	$16.35_{\pm9.86}$	$3.53_{\pm3.94}$	**$4.74_{\pm2.43}$**	$3.52_{\pm1.51}$	$15.34_{\pm13.82}$	$11.13_{\pm4.37}$	$95.47_{\pm92.16}$	$104.57_{\pm69.03}$	$130.96_{\pm119.72}$	$146.06_{\pm121.46}$	$192.55_{\pm133.90}$	$143.66_{\pm116.79}$	$57.21_{\pm44.55}$	3.136E−02	[27.33, 87.52]
FedProx	$2.94_{\pm0.68}$	$2.45_{\pm0.92}$	$49.45_{\pm45.68}$	$2.72_{\pm0.20}$	$21.03_{\pm19.28}$	$3.21_{\pm0.74}$	$7.56_{\pm6.97}$	$11.92_{\pm8.02}$	$18.89_{\pm19.15}$	$13.85_{\pm6.03}$	$75.03_{\pm91.60}$	$109.38_{\pm59.95}$	$192.30_{\pm174.70}$	$139.77_{\pm111.52}$	$175.17_{\pm126.68}$	$133.71_{\pm121.20}$	$59.96_{\pm49.58}$	7.463E−03	[29.36, 92.36]
HarmoFL	$3.86_{\pm2.61}$	$2.49_{\pm0.94}$	$40.80_{\pm32.44}$	$2.59_{\pm0.64}$	$18.43_{\pm16.51}$	**$3.15_{\pm0.48}$**	$9.69_{\pm11.78}$	$10.02_{\pm7.83}$	$19.48_{\pm15.39}$	$14.33_{\pm6.62}$	$74.42_{\pm66.09}$	**$75.25_{\pm39.94}$**	**$119.92_{\pm114.21}$**	$118.67_{\pm131.32}$	$179.01_{\pm171.61}$	$131.76_{\pm121.47}$	$51.49_{\pm46.24}$	4.856E−01	[26.58, 79.96]
HistoFL	$2.91_{\pm0.67}$	$2.45_{\pm0.86}$	$46.42_{\pm35.51}$	$2.89_{\pm0.54}$	$17.38_{\pm9.31}$	$3.69_{\pm0.80}$	$4.78_{\pm3.72}$	$4.21_{\pm3.29}$	$17.20_{\pm15.28}$	$10.18_{\pm3.85}$	$85.04_{\pm90.75}$	$86.83_{\pm68.73}$	$147.00_{\pm148.42}$	$151.82_{\pm118.01}$	$189.73_{\pm132.76}$	$143.72_{\pm116.87}$	$57.27_{\pm46.84}$	6.309E−03	[28.07, 88.12]
FedFA	$3.15_{\pm0.77}$	$2.44_{\pm0.89}$	$48.53_{\pm39.26}$	$2.64_{\pm0.39}$	$20.68_{\pm19.31}$	$3.30_{\pm0.80}$	$13.21_{\pm21.93}$	$11.08_{\pm8.67}$	$13.77_{\pm15.74}$	$13.19_{\pm6.05}$	$97.10_{\pm70.57}$	$120.36_{\pm189.27}$	$189.23_{\pm98.70}$	$133.46_{\pm132.82}$	$173.08_{\pm116.22}$	$140.20_{\pm48.90}$	$61.59_{\pm48.90}$	6.177E−03	[33.02, 94.01]
FedHEAL	$4.05_{\pm1.68}$	$2.67_{\pm0.95}$	$126.38_{\pm47.86}$	$2.83_{\pm0.21}$	$21.67_{\pm17.41}$	$5.66_{\pm1.45}$	$8.03_{\pm4.49}$	$6.19_{\pm2.38}$	$16.23_{\pm4.57}$	$9.40_{\pm3.36}$	$151.34_{\pm54.43}$	$145.48_{\pm55.26}$	$175.42_{\pm81.10}$	$166.40_{\pm90.73}$	$196.62_{\pm110.23}$	$180.55_{\pm80.90}$	$76.18_{\pm35.91}$	4.615E−03	[38.44, 116.08]
PathFL (Ours)	**$2.80_{\pm0.71}$**	$2.44_{\pm1.17}$	$39.91_{\pm40.21}$	**$2.16_{\pm0.40}$**	**$16.01_{\pm13.53}$**	$3.31_{\pm0.80}$	$6.84_{\pm6.23}$	**$2.81_{\pm2.08}$**	**$6.68_{\pm9.49}$**	**$9.31_{\pm4.53}$**	**$68.51_{\pm107.13}$**	$79.32_{\pm70.44}$	$143.34_{\pm131.87}$	**$126.61_{\pm92.20}$**	$169.98_{\pm121.82}$	$119.12_{\pm78.13}$	$49.95_{\pm30.05}$	–	[22.11, 77.47]

4.3.4. Performance on cross-scanner datasets

Our PathFL achieves superior performance on the cross-scanner dataset with the highest average Dice of 83.25% and the lowest average ASSD of 33.74, greatly outperforming FedAvg by 6.35% in Dice and 42.77 in ASSD. The quantitative segmentation results on the cross-scanner datasets for three centers (C1 to C3) are summarized in Table 5. The cross-scanner dataset illustrates the complexities inherent in segmentation tasks when utilizing images acquired from various scanners, which may differ in resolution, contrast, and noise characteristics. Our method not only surpasses other SOTAs but also demonstrates consistent performance across clients, achieving Dice of 81.59% at C1, 85.77% at C2, and 82.28% at C3. The findings from the cross-scanner dataset affirm the generalization of our method, illustrating its capacity to maintain high segmentation accuracy over scanner diversity. This reinforces the potential of our method for real-world applications, where the integration of multi-source imaging data is increasingly prevalent, and emphasizes its applicability in enhancing segmentation precision in clinical settings.

4.3.5. Performance on unified datasets

Our PathFL achieves the best segmentation performance, yielding a substantial and consistent improvement across all 16 clients as Table 6 shown. Specifically, our PathFL significantly outperforms other methods on the unified dataset, attaining an average Dice of 82.42% and
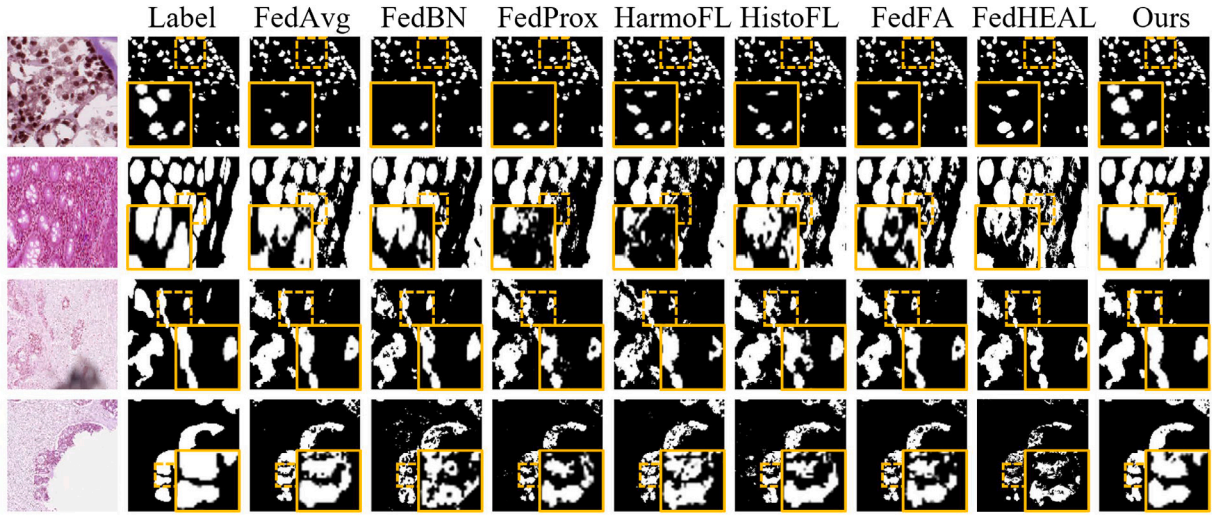
Fig. 4. Qualitative visualization comparison on segmentation results with our method and other state-of-the-art methods. From left to right: (a) Original Image, (b) Label, (c) FedAvg, (d) FedBN, (e) FedProx, (f) HarmoFL, (g) HistoFL, (h) FedFA, (i) FedHEAL, and (j) PathFL. Our PathFL achieves improved boundary accuracy and structure preservation compared to other methods.

an average ASSD of 49.95. PathFL also exhibits significant superiority over the baseline (FedAvg), with an average improvement of 3.93% in Dice and 7.41 in ASSD. Other methods like HistoFL perform well on few clients but lack generalization, leading to reduced overall performance. In contrast, PathFL achieves a superior balance between accuracy and stability, consistently maintaining high Dice with lower variance. Furthermore, statistical tests confirm the statistical significance of our PathFL's segmentation improvements, with most p-values below 0.01. The 95% CIs for Dice ([79.95, 84.97]) and ASSD ([22.11, 77.47]) demonstrate superior consistency and robustness across clients compared to other methods. In summary, these findings collectively confirm that PathFL not only advances federated segmentation in highly heterogeneous scenarios but also ensures stable and reliable performance across diverse datasets, making it a promising solution for real-world medical applications.

4.3.6. Visualization results

The visualization results (see Fig. 4) demonstrate the segmentation performance of our model, showcasing its ability to accurately delineate regions of interest across diverse pathology images. Our segmentation results reveal the ability to identify cell nuclei that other models fail to recognize, effectively reducing segmentation voids and minimizing fragmented regions. This improvement enhances the completeness of the segmentation and significantly increases its accuracy. The results highlight the effectiveness of our PathFL in handling multi-center heterogeneity, with clear segmentation boundaries and improved consistency across different datasets. These visualizations emphasize the model's robustness in capturing both structural and stylistic features, reinforcing its exceptional segmentation.

4.3.7. Ablation study

We conducted a comprehensive ablation study on cross-source datasets to systematically evaluate the performance contributions of our proposed modules, as detailed in Table 7. We explored various combinations of the three modules: the image-level module (Collaborative Style Enhancement, CSE), the feature-level module (Adaptive Feature Alignment, AFA), and the model aggregation-level module (Stratified Similarity Aggregation, SSA), with FedAvg serving as the baseline. Notably, our experimental results reveal that in the most challenging cross-source pathology segmentation scenarios, the image-level heterogeneity handling module demonstrates the most significant performance improvements. The ablation analysis further substantiates the synergistic interactions between these multi-level modules,

Table 7

Ablation analysis of modules on cross-source datasets. Δ denotes the Dice improvement compared to the baseline.

CSE	AFA	SSA	Average Dice (%) ↑	Δ
			$82.62_{\pm 4.99}$	–
✓			$83.67_{\pm 4.25}$	+1.05
	✓		$83.43_{\pm 4.61}$	+0.81
		✓	$83.33_{\pm 4.67}$	+0.71
✓	✓		$83.94_{\pm 4.84}$	+1.32
✓		✓	$83.81_{\pm 4.91}$	+1.19
	✓	✓	$83.48_{\pm 4.39}$	+0.86
✓	✓	✓	$\mathbf{84.27_{\pm 4.43}}$	**+1.65**

underscoring the holistic design and effectiveness of our proposed framework.

4.3.8. Comparison with different backbones

To further explore the impact of different SOTA backbones within our PathFL framework, we conducted additional experiments using Vision Transformer (ViT) (Dosovitskiy et al., 2020) and two pathology foundation models (CONCH Lu et al., 2024 and CHIEF Wang et al., 2024) alongside U-Net. Specifically, ViT was configured with six attention heads, eight depths, and an MLP dimension of 1536. For foundation models, we leveraged their pretrained encoders as frozen feature extractors coupled with an upsampling convolutional decoder. The results, summarized in Tables 8 and 9, show that our PathFL framework consistently outperforms the baseline (FedAvg) across all four backbones, confirming its broad applicability across different network architectures. PathFL with the CONCH backbone achieves the best performance across cross-source, cross-modality, cross-scanner, and cross-organ datasets, with average Dice scores of 84.38%, 90.52%, 86.10%, and 84.93%, respectively. The suboptimal performance of standalone foundation models with standard FedAvg in addressing cross-client heterogeneity stems from their inherent design prioritizing generalized feature extraction over local adaptability. Moreover, vision-language models such as CONCH and CHIEF predominantly focus on global semantic representations rather than local spatial details and limit their segmentation performance when confined to visual pathways alone. Our PathFL framework effectively bridges the gap between the global generalization capabilities of foundation models and the local adaptability required in FL. The integration of PathFL with foundation

Table 8

Comparison of four backbones on PathFL and FedAvg for cross-source datasets. The bold indicates the best value. Δ indicates the improvement of PathFL over FedAvg with the same backbone.

	Dice (%) ↑								
	C1	C2	C3	C4	C5	C6	C7	Average	Δ
FedAvg+U-Net	$80.48_{\pm3.82}$	$87.01_{\pm3.25}$	$84.09_{\pm8.15}$	$81.81_{\pm2.26}$	$86.97_{\pm8.21}$	$79.99_{\pm3.29}$	$78.01_{\pm5.96}$	$82.62_{\pm4.99}$	
PathFL+U-Net	$82.06_{\pm3.57}$	$\mathbf{87.46_{\pm3.24}}$	$86.12_{\pm5.02}$	$84.72_{\pm3.44}$	$88.96_{\pm8.27}$	$81.44_{\pm2.75}$	$\mathbf{79.17_{\pm4.70}}$	$84.27_{\pm4.43}$	1.65↑
FedAvg+ViT	$69.24_{\pm6.30}$	$83.15_{\pm3.97}$	$72.94_{\pm12.20}$	$75.16_{\pm4.71}$	$75.07_{\pm8.43}$	$74.14_{\pm4.91}$	$63.44_{\pm10.59}$	$73.31_{\pm7.30}$	
PathFL+ViT	$74.10_{\pm4.88}$	$83.37_{\pm4.08}$	$68.35_{\pm14.61}$	$78.13_{\pm3.99}$	$74.33_{\pm10.57}$	$75.88_{\pm3.39}$	$68.16_{\pm9.02}$	$74.62_{\pm7.22}$	1.31↑
FedAvg+CHIEF	$78.22_{\pm4.07}$	$85.35_{\pm4.02}$	$76.38_{\pm10.45}$	$80.20_{\pm4.27}$	$83.19_{\pm10.59}$	$79.29_{\pm3.55}$	$75.48_{\pm7.68}$	$79.73_{\pm6.38}$	
PathFL+CHIEF	$80.15_{\pm3.13}$	$86.64_{\pm4.06}$	$77.27_{\pm9.95}$	$82.05_{\pm5.40}$	$88.63_{\pm7.02}$	$77.55_{\pm5.71}$	$77.73_{\pm5.78}$	$81.43_{\pm5.86}$	1.7↑
FedAvg+CONCH	$80.13_{\pm3.57}$	$80.08_{\pm4.07}$	$90.39_{\pm4.77}$	$82.79_{\pm3.83}$	$91.74_{\pm6.07}$	$77.22_{\pm3.25}$	$72.37_{\pm4.87}$	$82.10_{\pm4.35}$	
PathFL+CONCH	$\mathbf{82.45_{\pm3.03}}$	$83.33_{\pm3.64}$	$\mathbf{91.00_{\pm4.77}}$	$\mathbf{85.09_{\pm3.55}}$	$\mathbf{92.16_{\pm6.07}}$	$\mathbf{81.63_{\pm2.42}}$	$75.01_{\pm3.74}$	$\mathbf{84.38_{\pm3.89}}$	2.28↑

Table 9

Comparison of four backbones on PathFL and FedAvg for three datasets. The bold indicates the best value. Δ indicates the improvement of PathFL over FedAvg with the same backbone.

	Cross-modality Datasets Dice (%) ↑				
	C1	C2	C3	Average	Δ
FedAvg+U-Net	$83.41_{\pm9.85}$	$91.46_{\pm4.64}$	$73.53_{\pm6.71}$	$82.80_{\pm7.07}$	
PathFL+U-Net	$89.93_{\pm3.49}$	$93.88_{\pm3.10}$	$83.00_{\pm5.30}$	$88.93_{\pm3.96}$	6.13↑
FedAvg+ViT	$68.21_{\pm10.26}$	$89.59_{\pm6.01}$	$60.13_{\pm8.23}$	$72.65_{\pm8.17}$	
PathFL+ViT	$89.32_{\pm3.08}$	$95.00_{\pm3.74}$	$81.56_{\pm5.24}$	$88.63_{\pm4.02}$	15.98↑
FedAvg+CHIEF	$82.58_{\pm5.01}$	$88.15_{\pm6.03}$	$66.86_{\pm6.88}$	$79.20_{\pm5.97}$	
PathFL+CHIEF	$\mathbf{92.90_{\pm2.96}}$	$\mathbf{94.45_{\pm4.45}}$	$83.71_{\pm5.48}$	$90.35_{\pm4.30}$	11.15↑
FedAvg+CONCH	$84.04_{\pm4.15}$	$91.61_{\pm4.40}$	$72.93_{\pm5.63}$	$82.86_{\pm4.72}$	
PathFL+CONCH	$91.60_{\pm3.88}$	$94.20_{\pm3.38}$	$\mathbf{85.75_{\pm3.54}}$	$\mathbf{90.52_{\pm3.60}}$	7.56↑

	Cross-oragn Datasets Dice (%) ↑				
	C1	C2	C3	Average	Δ
FedAvg+U-Net	$79.08_{\pm16.72}$	$72.79_{\pm20.83}$	$73.47_{\pm22.67}$	$75.11_{\pm20.07}$	
PathFL+U-Net	$\mathbf{85.65_{\pm10.86}}$	$83.36_{\pm9.21}$	$80.16_{\pm17.62}$	$83.05_{\pm12.57}$	7.94↑
FedAvg+ViT	$74.87_{\pm18.16}$	$70.98_{\pm20.80}$	$71.56_{\pm23.66}$	$72.47_{\pm20.88}$	
PathFL+ViT	$75.18_{\pm9.31}$	$74.15_{\pm9.58}$	$77.54_{\pm12.97}$	$75.62_{\pm10.62}$	3.15↑
FedAvg+CHIEF	$76.89_{\pm18.41}$	$74.65_{\pm19.65}$	$72.45_{\pm23.29}$	$74.66_{\pm20.45}$	
PathFL+CHIEF	$78.51_{\pm18.18}$	$69.22_{\pm21.79}$	$79.51_{\pm18.11}$	$75.74_{\pm19.36}$	1.08↑
FedAvg+CONCH	$85.18_{\pm13.34}$	$84.18_{\pm9.74}$	$84.73_{\pm13.95}$	$84.70_{\pm12.34}$	
PathFL+CONCH	$85.28_{\pm12.36}$	$\mathbf{86.96_{\pm8.27}}$	$\mathbf{86.06_{\pm13.08}}$	$\mathbf{86.10_{\pm11.24}}$	1.40↑

	Cross-scanner Datasets Dice (%) ↑				
	C1	C2	C3	Average	Δ
FedAvg+U-Net	$77.59_{\pm18.65}$	$80.31_{\pm20.72}$	$72.78_{\pm24.25}$	$76.90_{\pm21.21}$	
PathFL+U-Net	$81.59_{\pm14.06}$	$85.77_{\pm11.34}$	$82.38_{\pm14.90}$	$83.25_{\pm13.43}$	6.35↑
FedAvg+ViT	$75.10_{\pm19.69}$	$79.26_{\pm22.25}$	$70.23_{\pm25.91}$	$74.86_{\pm22.62}$	
PathFL+ViT	$75.52_{\pm19.00}$	$79.10_{\pm22.87}$	$70.81_{\pm25.14}$	$75.14_{\pm22.34}$	0.28↑
FedAvg+CHIEF	$75.83_{\pm19.54}$	$79.42_{\pm21.93}$	$72.18_{\pm24.84}$	$75.81_{\pm22.10}$	
PathFL+CHIEF	$78.04_{\pm17.49}$	$79.03_{\pm22.81}$	$74.93_{\pm22.03}$	$77.33_{\pm20.78}$	1.52↑
FedAvg+CONCH	$80.92_{\pm13.91}$	$85.51_{\pm14.77}$	$82.44_{\pm16.81}$	$82.96_{\pm15.17}$	
PathFL+CONCH	$\mathbf{83.79_{\pm11.28}}$	$\mathbf{86.21_{\pm14.07}}$	$\mathbf{84.80_{\pm15.13}}$	$\mathbf{84.93_{\pm13.49}}$	1.97↑

models like CONCH enables substantial performance gains, underscoring the effectiveness of our multi-alignment strategy in enhancing federated segmentation across heterogeneous data distributions.

4.3.9. Computational cost

To evaluate the computational cost of our PathFL framework when applied to larger backbones, such as ViT and foundation models like CONCH and CHIEF, we present key metrics in Table 10. These include FLOPs (billions, G), model parameters (millions, M), average training time per sample (milliseconds, ms), and communication overhead per client (MB). Compared with the baseline, the computation cost of our PathFL framework is comparable in all metrics. While the inclusion of the multi-alignment strategy slightly increases communication cost due to the exchange of style information between clients, the substantial improvements in segmentation performance justify this trade-off. As the backbone becomes more complex, the computational cost of PathFL increases. Thus, U-Net remains the most computationally efficient backbone, balancing segmentation accuracy and efficiency.

To further alleviate computational cost, especially when scaling to larger backbones, we evaluate a distributed similarity computation strategy. It shifts layer-wise similarity calculations from the server to clients, enabling parallel local computation based on the transmission of lightweight features between clients. This reduces local FLOPs and training time with a slight increase in communication overhead, thereby facilitating the scalability of PathFL to larger backbones.

5. Discussion and conclusion

Heterogeneity remains a fundamental challenge in federated learning for pathology segmentation, significantly impacting model performance and the feasibility of real-world deployment. To simulate multi-center collaborations, we explored four representative heterogeneous datasets capturing real-world pathology variations: data sources (cross-source), imaging techniques (cross-modality), anatomical regions (cross-organ), and scanning devices (cross-scanner). For instance, scenarios involving the same disease affecting different organs exemplify

Table 10

Computational cost per client of PathFL and FedAvg with different backbones on cross-source datasets. (distributed) indicates distributed similarity computation, while the default is centralized; Δ denotes the difference between the distributed and the centralized computation.

	FLOPs (G)	Params (M)	Average time (ms)	Communication overhead (MB)	Δ FLOPs (G)	Δ Average time (ms)	Δ Communication overhead (MB)
FedAvg+U-Net	8.21	1.95	3.34	14.88			
PathFL+U-Net	10.89	1.95	4.33	26.88			
PathFL+UNet (distributed)	8.59	1.95	3.77	30.30	−2.30	−0.56	+3.42
FedAvg+ViT	7.98	15.12	4.31	115.39			
PathFL+ViT	13.79	15.12	8.72	127.39			
PathFL+VIT (distributed)	8.81	15.12	7.75	132.08	−4.98	−0.97	+4.69
FedAvg+CHIEF	15.87	91.84	14.59	702.86			
PathFL+CHIEF	25.39	91.84	20.10	714.86			
PathFL+CHIEF (distributed)	17.23	91.84	19.35	723.46	−8.16	−0.75	+8.60
FedAvg+CONCH	26.30	67.10	12.55	3527.56			
PathFL+CONCH	38.96	67.10	21.47	3539.56			
PathFL+CONCH (distributed)	28.11	67.10	18.36	3553.34	−10.85	−3.11	+13.78

cross-organ heterogeneity, while variations in scanning technology for the same disease within the same organ illustrate cross-scanner heterogeneity. Our multi-alignment strategy demonstrates remarkable adaptability across heterogeneous scenarios. Experimental results demonstrate that PathFL with the U-Net backbone achieves an optimal balance between performance and efficiency, making it a lightweight yet effective choice for practical applications. Meanwhile, PathFL with the CONCH foundation model provides superior performance for all heterogeneous datasets, highlighting the potential of integrating foundation models into FL. Our findings suggest a promising direction for developing federated foundation models, capable of unlocking the potential of large-scale collaborative learning while preserving data privacy.

In conclusion, this study introduces PathFL, a novel multi-alignment federated learning framework that comprehensively addresses the critical challenge of heterogeneity in pathology image segmentation. By implementing a sophisticated three-tier alignment strategy—encompassing image-level collaborative style enhancement, feature-level adaptive feature alignment, and model-level stratified similarity aggregation—PathFL effectively mitigates representation bias and enhances cross-center generalizability. The framework systematically tackles complex heterogeneity across four heterogeneous datasets of sources, modalities, organs, and scanners, demonstrating superior segmentation accuracy and robustness. Our PathFL significantly outperforms the baseline, with average Dice improvements of 1.65% for cross-source datasets, 6.13% for cross-modality datasets, 7.94% for cross-organ datasets, and 6.35% for cross-scanner datasets. Experimental validation across four heterogeneous datasets substantiates PathFL's exceptional performance, highlighting its transformative potential for federated learning in medical image analysis. Future research directions include exploring PathFL's adaptability to additional medical imaging domains and practical medical implementation. Overall, this work lays a solid foundation for future research in federated learning and pathology image segmentation, paving the way for more effective and equitable applications in clinical diagnosis.

CRediT authorship contribution statement

Yuan Zhang: Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Feng Chen:** Writing – review & editing, Supervision, Resources. **Yaolei Qi:** Visualization, Software. **Guanyu Yang:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition. **Huazhu Fu:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Data availability

All the datasets used in this study are publicly available, and the code will be open-sourced upon acceptance.

References

Challenge, G., 2024. Grand challenge dataset. URL https://cosas.grand-challenge.org/ (Online; accessed 13-October-2024).

Chapelle, O., Weston, J., Bottou, L., Vapnik, V., 2000. Vicinal risk minimization. Adv. Neural Inf. Process. Syst. 13.

Chen, Y., Huang, W., Ye, M., 2024b. Fair federated learning under domain skew with local consistency and domain diversity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12077–12086.

Chen, J., Ma, B., Cui, H., Xia, Y., 2024a. Think twice before selection: Federated evidential active learning for medical image analysis with domain shifts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11439–11449.

Cheng, V.S., Hung, P.C., 2006. Health insurance portability and accountability act (HIPPA) compliant access control model for web services. Int. J. Heal. Inf. Syst. Informatics (IJHISI) 1 (1), 22–39.

Deng, T., Huang, Y., Han, G., Shi, Z., Lin, J., Dou, Q., Liu, Z., Guo, X.-j., Philip Chen, C.L., Han, C., 2024. Feddbl: Communication and data efficient federated deep-broad learning for histopathological tissue classification. IEEE Trans. Cybern. 54 (12), 7851–7864.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Graham, S., Chen, H., Gamper, J., Dou, Q., Heng, P.-A., Snead, D., Tsang, Y.W., Rajpoot, N., 2019a. MILD-net: Minimal information loss dilated network for gland instance segmentation in colon histology images. Med. Image Anal. 52, 199–211.

Graham, S., Vu, Q.D., Raza, S.E.A., Azam, A., Tsang, Y.W., Kwak, J.T., Rajpoot, N., 2019b. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Med. Image Anal. 58, 101563.

Guan, H., Yap, P.-T., Bozoki, A., Liu, M., 2024. Federated learning for medical image analysis: A survey. Pattern Recognit. 110424.

Hoque, M.Z., Keskinarkaus, A., Nyberg, P., Seppänen, T., 2024. Stain normalization methods for histopathology image analysis: A comprehensive review and experimental comparison. Inf. Fusion 102, 101997.

Hosseini, S.M., Sikaroudi, M., Babaie, M., Tizhoosh, H.R., 2023. Proportionally fair hospital collaborations in federated learning of histopathology images. IEEE Trans. Med. Imaging 42 (7), 1982–1995.

Huang, X., Belongie, S., 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1501–1510.

Irshad, H., et al., 2013. Methods for nuclei detection, segmentation, and classification in digital histopathology: a review—current status and future potential. IEEE Rev. Biomed. Eng. 7, 97–114.

Jiang, M., Wang, Z., Dou, Q., 2022. Harmofl: Harmonizing local and global drifts in federated learning on heterogeneous medical images. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36, pp. 1087–1095.

Ke, J., Lu, Y., Shen, Y., Zhu, J., Zhou, Y., Huang, J., Yao, J., Liang, X., Guo, Y., Wei, Z., et al., 2023. ClusterSeg: A crowd cluster pinpointed nucleus segmentation framework with cross-modality datasets. Med. Image Anal. 85, 102758.

Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., Sethi, A., 2017. A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE Trans. Med. Imaging 36 (7), 1550–1560.

Li, X., Jiang, M., Zhang, X., Kamp, M., Dou, Q., 2021b. Fedbn: Federated learning on non-iid features via local batch normalization. arXiv preprint arXiv:2102.07623.

Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V., 2020. Federated optimization in heterogeneous networks. Proc. Mach. Learn. Syst. 2, 429–450.

Li, B., Wu, F., Lim, S.-N., Belongie, S., Weinberger, K.Q., 2021a. On feature normalization and data augmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12383–12392.

Lu, M.Y., Chen, R.J., Kong, D., Lipkova, J., Singh, R., Williamson, D.F., Chen, T.Y., Mahmood, F., 2022. Federated learning for computational pathology on gigapixel whole slide images. Med. Image Anal. 76, 102298.

Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al., 2024. A visual-language foundation model for computational pathology. Nature Med. 30 (3), 863–874.

Ma, Y., Wang, J., Yang, J., Wang, L., 2024. Model-heterogeneous semi-supervised federated learning for medical image segmentation. IEEE Trans. Med. Imaging 43 (5), 1804–1815.

Mahbod, A., Schaefer, G., Bancher, B., Löw, C., Dorffner, G., Ecker, R., Ellinger, I., 2021. CryoNuSeg: A dataset for nuclei instance segmentation of cryosectioned h&e-stained histological images. Comput. Biol. Med. 132, 104349.

McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A., 2017. Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics. PMLR, pp. 1273–1282.

Naylor, P., Laé, M., Reyal, F., Walter, T., 2018. Segmentation of nuclei in histopathology images by deep regression of the distance map. IEEE Trans. Med. Imaging 38 (2), 448–459.

Pan, L., Huang, M., Wang, L., Qin, P., Peng, S., 2024. FedDP: Privacy-preserving method based on federated learning for histopathology image segmentation. arXiv preprint arXiv:2411.04509.

Qi, P., Chiaro, D., Guzzo, A., Ianni, M., Fortino, G., Piccialli, F., 2024. Model aggregation techniques in federated learning: A comprehensive survey. Future Gener. Comput. Syst. 150, 272–293.

Regulation, P., 2018. General data protection regulation. Intouch 25, 1–5.

Rieke, N., et al., 2020. The future of digital health with federated learning. NPJ Digit. Med. 3 (1), 119.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, pp. 234–241.

Schoenpflug, L.A., et al., 2024. A review on federated learning in computational pathology. Comput. Struct. Biotechnol. J. 23, 3938–3945.

Sheller, M.J., et al., 2020. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci. Rep. 10 (1), 12598.

Shen, Y., Sowmya, A., Luo, Y., Liang, X., Shen, D., Ke, J., 2022. A federated learning system for histopathology image analysis with an orchestral stain-normalization GAN. IEEE Trans. Med. Imaging 42 (7), 1969–1981.

Shi, J., Li, C., Gong, T., Zheng, Y., Fu, H., 2024. Vila-MIL: Dual-scale vision-language multiple instance learning for whole slide image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11248–11258.

Sirinukunwattana, K., Pluim, J.P., Chen, H., Qi, X., Heng, P.-A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U., et al., 2017. Gland segmentation in colon histology images: The glas challenge contest. Med. Image Anal. 35, 489–502.

Song, J., Chen, X., Zhu, Q., Shi, F., Xiang, D., Chen, Z., Fan, Y., Pan, L., Zhu, W., 2022. Global and local feature reconstruction for medical image segmentation. IEEE Trans. Med. Imaging 41 (9), 2273–2284.

Tang, Y., Lyu, T., Jin, H., Du, Q., Wang, J., Li, Y., Li, M., Chen, Y., Zheng, J., 2024. Domain adaptive noise reduction with iterative knowledge transfer and style generalization learning. Med. Image Anal. 98, 103327.

Van Panhuis, W.G., et al., 2014. A systematic review of barriers to data sharing in public health. BMC Public Health 14, 1–9.

Vu, Q.D., Graham, S., Kurc, T., To, M.N.N., Shaban, M., Qaiser, T., Koohbanani, N.A., Khurram, S.A., Kalpathy-Cramer, J., Zhao, T., et al., 2019. Methods for segmentation and classification of digital microscopy tissue images. Front. Bioeng. Biotechnol. 7, 433738.

Wagner, N., Fuchs, M., Tolkach, Y., Mukhopadhyay, A., 2022. Federated stain normalization for computational pathology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 14–23.

Wang, X., Zhao, J., Marostica, E., Yuan, W., Jin, J., Zhang, J., Li, R., Tang, H., Wang, K., Li, Y., et al., 2024. A pathology foundation model for cancer diagnosis and prognosis prediction. Nature 634 (8035), 970–978.

Wang, S., et al., 2019. Pathology image analysis using segmentation deep learning algorithms. Am. J. Pathol. 189 (9), 1686–1698.

Yamashita, R., Long, J., Banda, S., Shen, J., Rubin, D.L., 2021. Learning domain-agnostic visual representation for computational pathology using medically-irrelevant style transfer augmentation. IEEE Trans. Med. Imaging 40 (12), 3945–3954.

Yeghiazaryan, V., Voiculescu, I., 2018. Family of boundary overlap metrics for the evaluation of medical image segmentation. J. Med. Imaging 5 (1), 015006–015006.

Zhang, F., Liu, H., Cai, Q., Feng, C.-M., Wang, B., Wang, S., Dong, J., Zhang, D., 2024. Federated cross-incremental self-supervised learning for medical image segmentation. IEEE Trans. Neural Networks Learn. Syst. Early Access.

Zhang, Z., Zhang, X., Peng, C., Xue, X., Sun, J., 2018. Exfuse: Enhancing feature fusion for semantic segmentation. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 269–284.

Zhou, T., Konukoglu, E., 2023. FedFA: Federated feature augmentation. In: The Eleventh International Conference on Learning Representations. ICLR.