HistoFS: Non-IID Histopathologic Whole Slide Image Classification via Federated Style Transfer with RoI-Preserving

Supplementary Material

In the supplementary material, we describe the attention-gated aggregation in details in Sec. 8, provide a detailed description of math symbols frequently used in the paper in Sec. 9, provide the pseudo code of HistoFS in Sec. 10, present an analysis of pseudo bag style transfer in Sec. 11, present an analysis of our authenticity module in Sec. 12 present the ablation studies in Sec. 13, and provide detailed experimental settings in Sec. 14.

8. Attention-Gated Aggregation MIL

According to Eq. (3), we choose attention-gated aggregation ($\theta_{agg}^{I_n}$) [7] as our MIL backbone. Recall from Sec. 3 that a WSI $X_m^{I_n}$ is treated as a bag of patch features $X_m^{I_n} = \{x_{m,1}, \ldots, x_{m,P}\} \in \mathbb{R}^{P \times d}$. $\theta_{agg}^{I_n}$ applies attention weights to obtain the final bag representation as:

$$\mathbf{F} = \sum_{m=1}^{M} a_{m,p} x_{m,p} \in \mathbb{R}^P, \tag{11}$$

where $a_{m,p}$ is the learnable attention weight for $x_{m,p}$ and P is the number of patches. The attention weight $a_{m,p}$, based on [7], is formulated as:

$$a_{m,p} = \frac{\exp\left\{\mathbf{w}^T \left(\tanh\left(\mathbf{V}_1 x_{m,p}\right) \odot \operatorname{sigm}\left(\mathbf{V}_2 x_{m,p}\right)\right)\right\}}{\sum_{m=1}^{M} \exp\left\{\mathbf{w}^T \left(\tanh\left(\mathbf{V}_1 x_{m,p}\right) \odot \operatorname{sigm}\left(\mathbf{V}_2 x_{m,p}\right)\right)\right\}}, \tag{12}$$

where $\mathbf{w}$, $\mathbf{V}_1$, and $\mathbf{V}_2$ are the learnable parameters. Lastly, the prediction layer f_{pred} maps the final bag representation $\mathbf{F}$ to slide label prediction: $\hat{Y}_m^{I_n} = \theta_{agg}^{I_n}(f_{\text{pred}}(\mathbf{F}))$.

9. Math Symbols

Table 7 depicts the mathematical notations frequently used in our proposed HistoFS.

10. Pseudo Code of HistoFS

Algorithms 1 and 2 describe the steps of HistoFS and local pseudo bag style transfers, respectively.

11. Analysis of Pseudo Bag Style Transfer

11.1. Qualitative Analysis

We provide the visualization results of patch features from four institutions for inspecting the behavior of pseudo bag styles transfer. The visualization was conducted by transforming the multi-dimensional patch features into a two-dimensional t-SNE (t-distributed Stochastic Neighbor Embedding) space [2]. As shown in Fig. 3(a), the patch feature distributions are initially non-independent and non-identically distributed. However, after applying the pseudo

Table 7. Mathematical Notations.

Symbol	Meaning/Definition
Federated MIL	
N	Number of participating client/institution in FL
I_n	i-th client/institution in a FL framework
T	Total communication rounds in FL
$\mathbf{W}$	Weight matrix of the MIL model
ℓ	Loss function of the MIL model
$a_{m,p}$	p-th attention weight corresponding to $x_{m,p}$
f_{pred}	Prediction layer of the MIL model
$\mathbf{F}$	Final bag representation
$\theta_{ext}^{I_n}$	Feature extractor at I_n
$\theta_{agg}^{I_n}$	Attention-gated aggregation at I_n
Whole Slide Image (WSI)	
M^{I_n}	Total number of WSIs in I_n
P	Total number of patches in a WSI
K	Number of pseudo styles in a WSI
$X_m^{I_n}$	m-th WSI (bag) at I_n
$X_m^{I_n(\text{aug})}$	Augmented WSI corresponding to $X_m^{I_n}$
$Y_m^{I_n}$	Slide label of $X_m^{I_n}$
$\hat{Y}_m^{I_n}$	Slide prediction of $X_m^{I_n}$
$x_{m,p}$	p-th patch feature from $X_m^{I_n}$
$x_{m,p}^{(\text{aug})}$	Augmented patch feature corresponding to $x_{m,p}$
Style Information	
$S_m^{I_n}$	Bag of styles corresponding to $X_m^{I_n}$
$\mathcal{P}_m$	Feature distribution corresponding to $X_m^{I_n}$
$\mu(\cdot)/\sigma(\cdot)$	Mean/Standard deviation
W_2^2	2-Wasserstein distance
$c_{m,k}$	k-th pseudo style of $X_m^{I_n}$
$\mathbf{C}_m^{I_n}$	Pseudo bag styles of $X_m^{I_n}$
AdaIN$(\cdot)$	Adaptive Instance Normalization
Authenticity Module	
$\text{Auth}_{\text{score}}^{I_n}$	Authenticity score
$\text{A}_m^{I_n}$	Attention weights of $X_m^{I_n}$
$\text{A}_m^{I_n(\text{align})}$	Aligned attention weights of $X_m^{I_n}$
λ	Tunable hyper parameter in the Authenticity Module

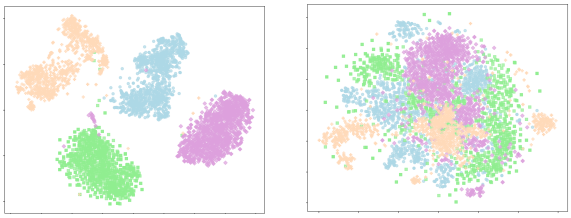
Algorithm 1 HistoFS

Input: Initialized global weight matrix $\mathbf{W}$
Output: Final global weight matrix $\mathbf{W}$
for global round $t = 1, \ldots, T$ **do**
    Stage 1: Federated Learning Process
    The server selects participating institutions N and sends $\mathbf{W}$ to each institution $I_n \in N$.
    for each $I_n \in N$ **do**
        Compute pseudo bag styles $\mathbf{C}^{I_n}$ in Sec. 4.1.1.
        Transmit $\mathbf{C}^{I_n}$ to the server in Sec. 4.1.2.
    end for
    The server shares $\{\mathbf{C}^{I_n}\}_{n \in N}$ with all institutions.
    Stage 2: Local Updates
    for each $I_n \in N$ **do**
        (i) Pseudo bag styles transfer in Sec. 4.1.3.
        (ii) Authenticity module in Sec. 4.1.4.
        (iii) MIL prediction in Eq. (3).
        Compute local cost function (Eq. (2)).
    end for
    The server updates the $\mathbf{W}$ by minimizing a global cost function (Eq. (1)).
end for
return Final global weights $\mathbf{W}$ for evaluation.

Algorithm 2 Local Pseudo Bag Styles Transfer

Input:
Training set $\{X_m^{I_n}\}_{m=1}^{M^{I_n}}$
Style bag $\mathbf{C}^{I_{n'}}$ with subsets $\mathbf{C}_j^{I_{n'}} = \{c_{j,1}, \ldots, c_{j,k}\}$ for $j = 1, \ldots, J$
Parameter: Augmentation ratio $\gamma \in [0, 1]$
Output: Augmented WSIs $\{X_m^{I_n(\text{aug})}\}_{m=1}^{M^{I_n}}$
Stage 1: Selected WSIs to augment
Sample $\gamma \times M^{I_n}$ WSIs: $\mathcal{S} \subset \{1, \ldots, M^{I_n}\}$
Stage 2: Iterate over the selected WSIs
for $m \in \mathcal{S}$ **do**
    Randomly select $\mathbf{C}_j^{I_{n'}}$ from $\mathbf{C}^{I_{n'}}$
    Stage 3:Iterate over patches of $X_m^{I_n}$
    for $p = 1, \ldots, P$ **do**
        Randomly select $c_{j,k} \in \mathbf{C}_j^{I_{n'}}$
        $x_{m,p}^{\text{aug}} = \text{AdaIN}(x_{m,p}, c_{j,k})$
        Append $x_{m,p}^{\text{aug}}$ to $X_m^{I_n(\text{aug})}$
    end for
end for
Stage 4:Assign augmented label
Assign label $Y_m^{I_n}$ to $X_m^{I_n(\text{aug})}$
return $\{X_m^{I_n(\text{aug})}, Y_m^{I_n}\}_{m=1}^{M^{I_n}}$

bag styles transfer (Sec. 4.1.3), the distributions become more identically distributed and aligned (Fig. 3(b)). This obviously demonstrates that our pseudo bag styles transfer effectively introduces different style properties from distinct institutions, enabling each institution to train a local MIL model without the risk of weight bias arising from a focus on institution-specific styles.



(a) Before pseudo bag styles transfer.

(b) After pseudo bag styles transfer.

Figure 3. The 2D t-SNE plot represents the distribution of patch features of WSIs from four institutions. (a) Without style transfer, patch features are non-i.i.d. (b) After HistoFS with pseudo bag styles transfer, patch features become more i.i.d.

11.2. Quantitative Analysis

As shown in Table 8, Wasserstein Distance (WD), Mutual Information (MI), FID, and KID were used to quantify either the similarity or dependency between any two distributions of patch features. Table 8 shows our method increase the similarity and dependency of patch features across different clients. Since our method leads to a smaller WD/FID/KID and a larger MI for any two distributions of features, it's an indicator of tending to i.i.d.

Table 8. Quantitative measure of pseudo bag style transfer.

Metric	WD (↓)	MI (↑)	FID (↓)	KID (↓)
Before (Fig. 3(a))	2.28	2.27	11.36	0.32
After (Fig. 3(b))	**0.58**	**4.82**	**3.68**	**0.06**

12. Analysis of Authenticity Module

As discussed in Sec. 4.1.4, augmenting local WSIs via style transfer risks omitting RoIs. We analyze the effectiveness of our authenticity module from the aspects of qualitative analysis (in terms of visualization inspection) and quantitative analysis.

12.1. Qualitative Analysis

To verify the effectiveness of our authenticity module (Sec. 4.1.4) in aligning RoIs after augmenting WSIs through pseudo bag styles transfer, we provide the color map visualization for WSI images, as an example shown in Fig. 4. Style transfer computes the statistical properties globally without considering the local context and correlation information between patch features. Indeed, this information is important for localizing the region of interests (RoIs)

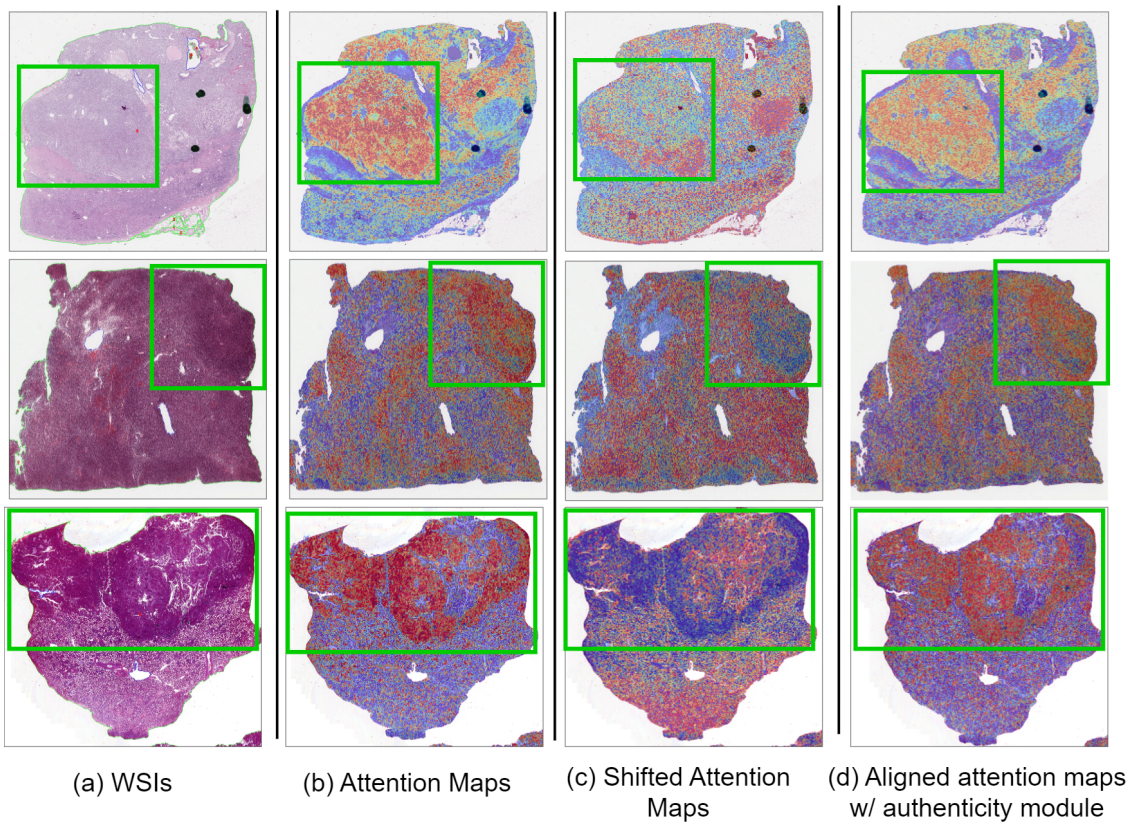


Figure 4. Visualization of three WSI examples from the TCGA-RCC dataset. For each WSI, the attention weights for all the patch features in the slide are normalized to a range of [0,1]. The normalized scores are then mapped to their spatial locations in the slide, where the RGB color map is applied (red: high attention, blue: low attention). The green bounding box represents the RoIs' locations generated from attention weights (Eq. (12)).

in WSIs to make a slide-level prediction. We utilized the configuration of CLAM [8] to produce an attention maps for WSI. Fig. 4 (a) and Fig. 4 (b) show three WSIs from the TCGA-RCC dataset and their attention maps, respectively. Regions with high attention (in red) in the attention maps can be recognized as ROIs; otherwise, it is a normal tissue. Notably, the MIL model can localize RoIs that contribute to the WSI classification task as well. In Fig. 4(c), it indicates that RoI shift occurs [21] when the MIL model cannot pay attention to the RoIs after style transfer and produce many false highly-attended regions outside the green bounding box. However, when our proposed authenticity module is applied, Fig. 4(d) shows that the RoIs can be refined and aligned well, indicating RoIs can be maintained and patch features can be learned with diverse styles from other institutions.

12.2. Quantitative Analysis

Attention weights can serve as an indicator to identify RoIs in a WSI. We combined the test sets from three different institutions (see Table 2) to calculate the authenticity score (Eq. (9)) between the attention weights of both local and augmented WSIs. We then apply kernel density estimation (KDE) to assess the probability distribution of the authenticity scores. Fig. 5 shows KDE comparison between our proposed authenticity module and the baseline methods, including HistoFL [3] and CCST [21], where the x-axis denotes the authenticity score in that the values close to 0 indicate the minimal changes in RoIs and those near 1 suggest significant deviations. Our module effectively preserves attention weight consistency between the local and augmented RoIs, maintaining RoIs as also shown in Fig. 5.

13. Ablation Studies

13.1. Comparison via bACC and F1-Score

Table 9 presents the comparison results under the RCC and HER2 datasets in terms of bACC and F1-score. Our proposed method outperforms the competitors, which validates the advantage of using pseudo bag styles transfer and authenticity module. Specifically, the average bACC and F1 scores are improved by $2 \sim 5\%$ on RCC and HER2, re-

Table 9. Classification accuracy comparison in terms of bACC and F1 score.

	Methods	Inst. A		Inst. B		Inst. C		Avg		Entropy Score	
		bACC	F1	bACC	F1	bACC	F1	bACC	F1	bACC	F1
(RCC Dataset)											
MIL	HistoFL [38]	95.51	91.28	86.59	91.20	94.97	92.60	92.36	91.69	0.218	0.238
	DTFD-MIL [11]	95.80	93.65	88.67	89.92	93.60	93.82	92.60	92.46	0.209	0.216
	FRMIL [10]	96.98	95.62	89.44	86.89	93.39	93.82	93.27	92.11	0.193	0.225
Local Style Transfer	MixStyle [30]	96.97	93.65	89.55	91.40	96.83	96.58	94.45	93.88	0.160	0.177
	DSU [31]	96.97	93.65	88.57	88.92	96.58	96.58	94.04	93.05	0.171	0.199
Federated Style Transfer	CCST [21]	96.97	93.65	89.55	91.40	96.83	96.58	94.45	93.88	0.160	0.177
	DACS [26]	95.51	91.28	88.57	91.20	96.83	95.61	93.64	92.03	0.183	0.210
HistoFS	Pseudo Bag Styles Only	96.97	93.65	90.05	91.80	96.83	96.58	94.61	94.01	0.155	0.173
	+ Authenticity Module	97.81	95.63	**90.95**	**93.21**	**96.83**	**96.58**	**95.86**	95.14	**0.139**	0.142
HistoFS	w/ DTFD-MIL	97.20	94.05	89.55	91.40	96.83	96.58	94.52	94.00	0.157	0.173
	w/ FRMIL	**98.00**	**96.05**	90.50	93.00	96.83	96.58	95.11	**95.21**	0.141	**0.139**
(HER2 Dataset)											
MIL	HistoFL [38]	65.56	62.66	81.04	81.07	75.00	74.18	73.87	72.64	0.663	0.685
	DTFD-MIL [11]	68.61	68.72	79.60	83.42	76.21	74.38	74.80	75.50	0.647	0.629
	FRMIL [10]	72.22	72.22	81.32	81.38	78.21	71.43	77.25	75.01	0.595	0.643
Local Style Transfer	MixStyle [30]	73.33	72.75	88.46	88.49	78.57	78.12	82.12	79.96	0.525	0.533
	DSU [31]	76.76	75.65	88.44	88.21	78.30	78.05	78.92	78.47	0.503	0.515
Federated Style Transfer	CCST [21]	67.78	63.90	84.88	85.01	79.84	79.23	77.50	76.04	0.582	0.609
	DACS [26]	72.22	72.73	81.87	81.83	75.00	74.71	76.36	76.42	0.615	0.613
HistoFS	Pseudo Bag Styles Only	76.76	75.65	88.89	87.60	82.14	81.84	82.59	81.69	0.469	0.491
	+ Authenticity Module	**76.80**	**77.51**	**89.01**	**88.89**	82.14	81.84	82.65	82.08	0.468	**0.466**
HistoFS	w/ DTFD-MIL	74.80	75.00	88.44	88.21	79.84	79.23	81.03	80.81	0.505	0.510
	w/ FRMIL	76.76	76.65	88.46	88.49	**83.00**	**82.20**	**82.74**	**82.44**	**0.466**	0.473

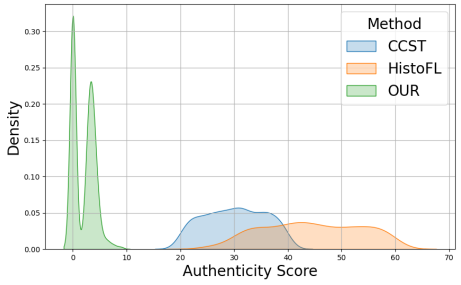


Figure 5. Kernel Density Estimation (KDE) of authenticity score for the RCC dataset.

spectively. We further revealed the performance divergence in terms of bACC and F1 score by calculating the entropy scores across all institutions. As shown in Table 9, our method has the smallest entropy scores, indicating performance divergence is reduced effectively.

13.2. Comparison with Aggregation-Based FL

Table 10 shows that our proposed HistoFS achieves the highest average AUC score among aggregation-based FL methods, such as FedBN [39] and FedProx [40]. Specifically, FedBN [39] preserves batch normalization statistics locally, while FedProx [40] incorporates a proximal term to regularize the gap between local and global models. In contrast, our proposed HistoFS shares style statistics across clients to minimize the gap.

Table 10. Comparison with aggregation-based FL methods.

Dataset	Methods	Inst. A	Inst. B	Inst. C	Avg.	Entropy
RCC	FedBN [39]	98.62	95.20	97.90	97.24	0.081
	FedProx [40]	97.80	96.65	98.00	97.48	0.074
	HistoFS	**99.48**	**95.39**	**99.46**	**98.11**	**0.056**
HER2	FedBN [39]	70.20	88.74	79.21	79.38	0.539
	FedProx [40]	74.35	89.12	78.40	80.62	0.513
	HistoFS	**80.66**	**92.85**	**84.69**	**86.07**	**0.383**

13.3. Comparison with Shorter Local Iteration

We conducted experiments using local iterations that are four times shorter while increasing the total number of global rounds that are four times longer than usual. As shown in Table 11, the results suggest that shorter local iterations are insufficient to achieve convergence for each client. Our proposed method, however, can maintain good performance regardless of the local iteration length.

13.4. Effect of Feature Extractors

In an MIL setup, the feature extractor transforms patch images into features. We investigate the performance of proposed method under different backbones, including ResNet50 [49] and SSL-ViT [1], for feature extraction,

Table 11. Shorter Local Epochs and Longer Aggregations.

Dataset	Methods	Inst. A	Inst. B	Inst. C	Avg.	Entropy
RCC	HistoFL [3]	95.80	93.61	97.00	95.47	0.132
	CCST [21]	98.43	92.40	97.28	96.04	0.115
	HistoFS	**99.20**	**94.38**	**99.16**	**97.58**	**0.071**
HER2	HistoFL [3]	60.30	78.95	74.20	71.14	0.713
	CCST [21]	67.80	88.48	80.60	78.96	0.545
	HistoFS	**79.80**	**90.65**	**83.28**	**84.57**	**0.421**

where the dimension (d) of patch features is 1024. In Table 12, it can be observed that our HistoFS outperforms MixStyle [30] and CCST [21] in terms of AUC and entropy score.

Table 12. Effect of Feature Extractor.

Methods		Inst. A	Inst. B	Inst. C	Avg	Entropy
		AUC	AUC	AUC	AUC	AUC
(RCC Dataset)						
ResNet50	MixStyle	96.16	89.81	94.09	93.35	0.075
	CCST	96.36	89.49	96.02	93.96	0.074
	Proposed HistoFS	**96.65**	**90.02**	**96.14**	**94.27**	**0.069**
SS-ViT	MixStyle	99.02	94.44	98.21	97.22	0.081
	CCST	99.07	94.25	98.45	97.26	0.080
	Proposed HistoFS	**99.48**	**95.39**	**99.46**	**98.11**	**0.056**
(HER2 Dataset)						
ResNet50	MixStyle	59.71	82.96	65.81	69.49	0.7382
	CCST	60.65	81.31	64.79	68.91	0.752
	Proposed HistoFS	**67.50**	**84.06**	**70.69**	**74.05**	**0.657**
SS-ViT	MixStyle	75.70	89.01	78.06	80.92	0.507
	CCST	69.25	91.63	82.65	81.18	0.492
	Proposed HistoFS	**80.66**	**92.85**	**84.69**	**86.07**	**0.383**

13.5. Different strategies for Pseudo Bag Styles

As described in Sec. 4.1.1, one of the motivations for constructing pseudo bag styles $\mathbf{C}_m^{I_n} = \{c_{m,1}, \ldots, c_{m,K}\}$ is to avoid the high communication burden caused by possible transmitting thousands of styles in each FL round. In addition to the one discussed in Sec. 4.1.1, we provide three alternative designs for constructing pseudo bag styles as follows:

- **L-Bag styles**: We sort and select the J styles with the lowest values from the bag of styles $S_m^{I_n} = \{s_{m,1}, \ldots, s_{m,P}\}$, then concatenate them into L-Bag styles $\mathbf{L}_m^{I_n} = \{l_{m,1}, \ldots, l_{m,J}\}, 1 \leq J \leq P$.
- **R-Bag styles**: : Randomness is beneficial to promote diversity. We select J random styles and concatenate them into R-Bag styles $\mathbf{R}_m^{I_n} = \{r_{m,1}, \ldots, r_{m,J}\}$.
- **H-Bag styles**: We can also obtain J style with the highest values after sorting and concatenating them into H-Bag styles $\mathbf{H}_m^{I_n} = \{h_{m,1}, \ldots, h_{m,J}\}$.

Table 13 shows that pseudo bag styles is a sophisticated design tailored to the characteristics of WSI, as described in Section 1.1, and generally performs better than the other three strategies.

Table 13. Different Strategies for Constructing Pseudo Bag Styles.

Methods	Inst. A	Inst. B	Inst. C	Avg	Entropy
	AUC	AUC	AUC	AUC	AUC
(RCC Dataset)					
L-Bag Styles	98.93	94.79	98.32	97.35	0.078
R-Bag Styles	99.16	94.89	98.43	97.49	0.074
H-Bag Styles	99.09	94.63	98.38	97.36	0.077
Pseudo Bag Styles	**99.48**	**95.39**	**99.46**	**98.11**	**0.056**
(HER2 Dataset)					
L-Bag Styles	76.21	91.24	82.43	83.29	0.449
R-Bag Styles	77.08	91.69	82.78	83.85	0.436
H-Bag Styles	76.34	91.38	82.67	83.46	0.445
Pseudo Bag Styles	**80.66**	**92.85**	**84.69**	**86.07**	**0.383**

13.6. Effect of Hyper-Parameters

We evaluate the impact of two key hyper-parameters, including K (the number of clusters for generating pseudo bag styles) and λ (the tunable parameter in the authenticity module). In the upper plot of Fig. 6, when K is small enough, the entropy score is high because the model lacks sufficient exploration of statistical properties beyond the client's WSIs. As K increases, the entropy scores stabilize around 0.005, indicating reduced divergence. Overall, HistoFS outperforms CCST [21]. The lower plot of Fig. 6 shows the effect of λ on the entropy score. With a smaller λ, HistoFS outperforms CCST [21], showing that the pseudo bag styles strategy is effective. However, as λ increases significantly, the entropy score rises, reflecting larger performance divergence across institutions and making fine-tuning λ challenging.

14. More details about Experimental Settings

We use HistoFL's code base for implementation and build other state-of-the-art methods based on their officially released codes. Specifically:

- The official code for HistoFL [3] can be found at https://github.com/mahmoodlab/HistoFL. We deployed its attention-gated aggregation and differential privacy settings for our use.
- The official codes for FRMIL [10] and DTFD-MIL [11] can be found at https://github.com/PhilipChicco/FRMIL and https://github.com/hrzhang1123/DTFD-MIL, respectively. We deployed their MIL backbones in a federated setting.
- The official codes for MixStyle [30] and DSU [31] can be found at https://github.com/KaiyangZhou/mixstyle-release and https://github.com/lixiaotong97/DSU, respectively. We deployed their style augmentation methods during the local update process.
- The official code for CCST [21] can be found at https://github.com/JeremyCJM/CCST. We deployed

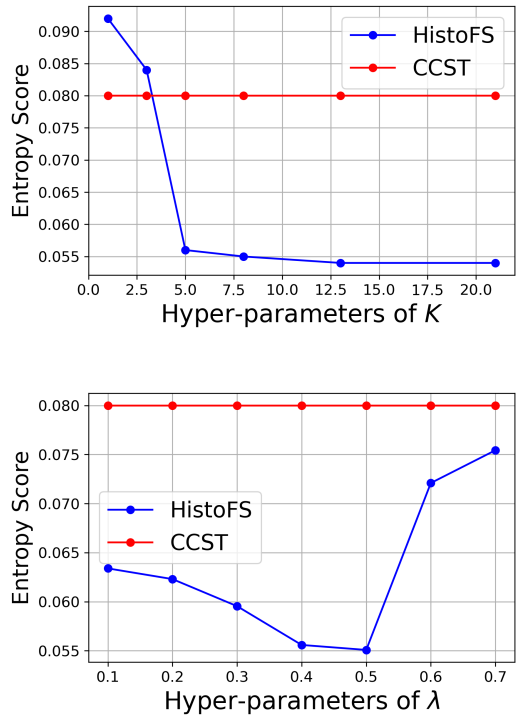


Figure 6. Effect of hyper-parameters, K and λ, on HistoFS.

its cross-client style transfer along with our federated process.

- The official code for DACS [26] can be found at https://github.com/FlyingRoastDuck/DACS_official. We deployed its style transformation model (STM) and authenticity constraints along with our MIL backbone choice.

In particular, these methods [21, 26, 30, 31] were employed to transform the statistical properties of each image feature from one client using those from others. However, this is not the case for WSIs that consists of hundreds to hundreds of thousands of patch image features. So, we modified their codes according to our use. For CCST [21], the server was employed to receive all the styles from each client's patch image features and broadcast them back to all clients. For DACS [26], we trained the STM according to style information of each patch and transformed the style via AdaIN [18]. For local style transfer [30, 31], we randomly selected patches according to the probability augmentation level that was 0.5 [30] within a local WSI.

On the other hand, we adopted SSL-ViT [1] and ResNet50 [49] as the feature extractors, respectively. Specifically, SSL-ViT is a transformer-based model. We thus built upon the DINO framework and used the pre-trained model released in the `timm` library (https://huggingface.co/timm/vit_small_patch16_224.dino). For each patch image, SSL-ViT outputs a feature with 384 dimensions. ResNet50, on the other hand, is a widely used CNN model in the WSI community, utilizing the ImageNet pre-trained model officially released by PyTorch (https://download.pytorch.org/models/resnet50-0676ba61.pth). For each patch image, ResNet50 outputs a feature with 1024 dimensions.