FD-MIL: Frequency-Guided Differential Attention for Whole-Slide Image Classification

Yixian Kong
Beijing University of Posts and Telecommunications
Beijing, China
kongyixian@bupt.edu.cn

***Abstract*—Multiple instance learning (MIL) has gained increasing adoption in the classification of histopathology whole slide images (WSIs). However, unlike supervised learning, MIL relies solely on bag-level labels, introducing interference from challenging negative samples, such as those morphologically similar to positive cancer tissues or affected by tissue staining variations. These interferences are difficult to characterize in the visual domain but can be effectively separated in the frequency domain. To address this challenge, this paper proposes a framework that leverages global frequency-domain information to extract critical features, uses inter-instance difference in frequency domain to realign the distribution distance between positive and negative samples, thereby mitigating the interference caused by sample preparation variations, and simultaneously fuses frequency-domain features with visual spatial features to collectively enhance the classification capability of the model. Experiments conducted on the TGA-BRCA, TCGA-NSCLC, and Camelyon16 datasets demonstrate substantial performance improvements, validating that the frequency-spatial fusion approach effectively addresses MIL challenges in WSI analysis by leveraging spectral feature separability.**

***Index Terms*—Whole Slide Image, Multiple Instance Learning, Frequency Domain Analysis**

I. INTRODUCTION

Whole Slide Image (WSI) classification is pivotal in computational pathology but remains computationally challenging due to gigapixel resolutions. To circumvent this, the field has converged on a two-stage Multiple Instance Learning (MIL) paradigm, where tissue slides are tessellated into instances for feature extraction followed by slide-level aggregation. Despite its widespread adoption, this paradigm faces dual impediments. First, models are highly sensitive to non-pathological variations—such as staining inconsistencies and acquisition artifacts—often misinterpreting these "nuisance" signals as pathological features, leading to biased aggregation [1]–[5]. Second, as bag sizes grow, the quadratic complexity of standard self-attention mechanisms prohibits efficient global context modeling, a bottleneck only partially alleviated by region-based methods like RRT-MIL [6].

Critically, these spatial challenges are inherently more tractable in the frequency domain. Spectral analysis not only offers linear separability for confounding artifacts like staining bias but also enables global information processing with significantly lower computational overhead than spatial self-attention. Motivated by these insights, we propose FD-MIL, a novel framework that leverages global frequency signatures to filter noise and exploits inter-instance spectral differences to realign feature distributions, thereby achieving robust and efficient WSI classification.

To leverage spectral properties at a macro level, we first propose a Global Token Frequency Fusion Module (GFFM). This module introduces a set of learnable spectral prompts to summarize bag-level frequency signatures, effectively filtering out widespread noise and identifying general characteristics of positive bags. However, relying solely on global frequency information risks diluting signals from small, localized lesions. To overcome this and harness inter-instance disparities, we introduce a refined approach–Differential Frequency Guidance Module (DFGM). Unlike global representations, DFGM utilizes instance-level scoring in the spectral domain to dynamically screen for key candidate instances most likely to be malignant. It computes a differential amplitude based on the frequency contrast between these key instances and the remaining bag, explicitly enhancing the subtle yet discriminative features that distinguish cancerous from normal tissue.

Finally, we recognize that spatial and frequency-domain information are complementary; while frequency analysis offers robustness against noise and efficiency, it may neglect critical tissue morphology. To bridge this gap, we introduce the Selective Multi-Cognitive (SMC) block designed to capture multi-scale spatial patterns. Crucially, these components are not trained in isolation. We propose a joint optimization framework that fuses the focused spatial features from the SMC with the robust frequency-domain signals from GFFM and DFGM. This optimization strategy ensures a synergistic representation that preserves both tissue morphology and cellular texture cues while enforcing semantic consistency across domains. Experiments conducted on the TCGA-BRCA, TCGA-NSCLC, and Camelyon16 datasets demonstrate that FD-MIL achieves substantial performance improvements over state-of-the-art methods, validating that this frequency-spatial fusion approach effectively addresses MIL challenges in WSI analysis. To enhance the reproducibility of the research, we will make all source codes publicly available once the paper is accepted.

The contributions of this paper are summarized as follows:

- We propose the Global Token Frequency Fusion Module, which utilizes a learnable token to encapsulate bag-level spectral signatures, effectively filtering high-frequency acquisition noise and staining artifacts.

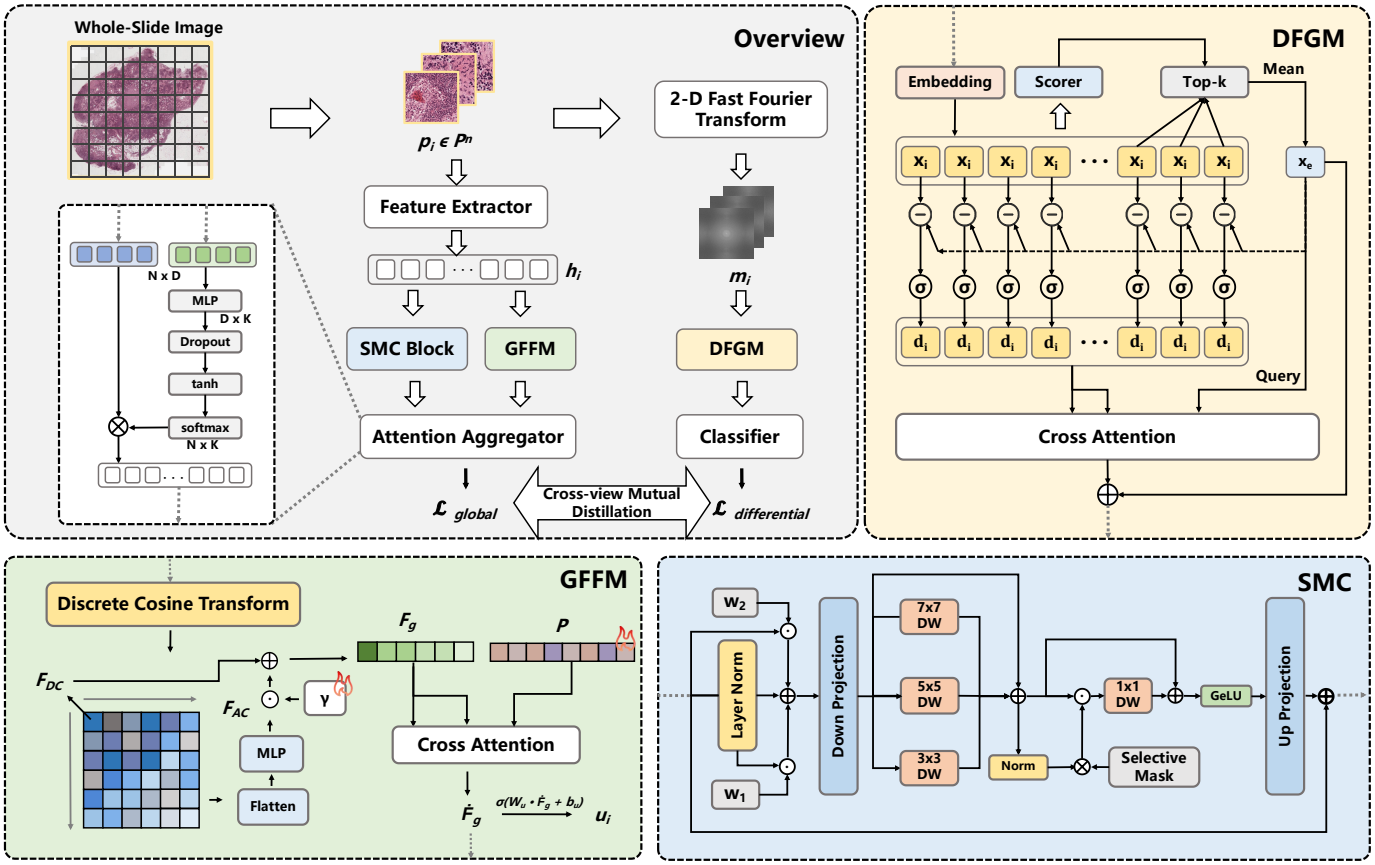


Fig. 1. Overview of the proposed FD-MIL framework. The framework consists of a Global Fusion Branch and a Differential Frequency Branch. (1) Global Fusion Branch: This branch synergistically integrates spatial and frequency domains. The Selective Multi-Cognitive (SMC) block extracts multi-scale morphological features from ResNet-encoded patch embeddings h_i. Simultaneously, the Global Frequency Fusion Module (GFFM) applies DCT and interacts with learnable spectral prompts P to capture bag-level frequency signatures F_g. The spatial features from SMC and frequency signals from GFFM are fused via an attention aggregator to yield the global prediction. Additionally, GFFM estimates an uncertainty scalar u_i to quantify sample reliability. (2) Differential Frequency Branch: The Differential Frequency Guidance Module (DFGM) processes the magnitude spectra m_i of raw patches derived via 2D FFT. It computes a rectified spectral residual d_i by comparing instances to a key malignant exemplar x_e, explicitly amplifying inter-class discrepancies to generate the differential prediction. Joint Optimization: The model is trained using a unified objective where the spectrum-guided uncertainty u_i dynamically weights the global loss to suppress noise, and Cross-View Mutual Distillation enforces semantic consistency between the global fusion and differential branches.

- We introduce the Differential Frequency Guidance Module to explicitly amplify the discriminability between hard negative and positive instances. By calculating spectral residuals and leveraging inter-instance frequency contrasts, DFGM sharpens the decision boundary and accurately localizes key malignant regions.
- We design the Selective Multi-Cognitive block to extract multi-scale spatial patterns, which are fused with frequency signals to yield a synergistic representation preserving both tissue morphology and cellular texture.
- We develop a unified joint optimization framework incorporating Spectrum-Guided Uncertainty Loss and Cross-View Mutual Distillation, which dynamically regulates optimization based on spectral noise and enforces semantic consistency across domains.

II. RELATED WORK

The application of Deep Learning to WSI analysis is predominantly governed by the Multiple Instance Learning (MIL) paradigm. Early approaches, such as AB-MIL [7], introduced attention-based pooling to weigh instances based on their contribution to the bag label, moving beyond simple max-pooling or mean-pooling strategies. Subsequent advancements focused on enhancing the expressiveness of the aggregation stage. For instance, methods like CLAM [8] introduced clustering-constrained attention to better separate positive and negative instances, while TransMIL [9] leveraged Transformer architectures to model long-range dependencies between instances. Despite their efficacy, these spatial-attention methods face inherent limitations regarding computational efficiency. The computational cost of standard self-attention mechanisms grows quadratically with the sequence length, posing severe scalability issues for WSIs containing tens of thousands of patches. While recent works such as RRT-MIL [6] have attempted to alleviate this by grouping instances into regions to model local correlations, they essentially retain the spatial processing paradigm. Furthermore, these spatial-based models remain susceptible to domain shifts caused by staining variations, often requiring complex augmentation or normalization techniques to generalize effectively.

Frequency Domain Learning in Vision Frequency domain

analysis has long been a staple in signal processing and has recently seen a resurgence in deep learning for computer vision. By transforming signals into the spectral domain using operations like the Fast Fourier Transform (FFT) or Discrete Cosine Transform (DCT), models can separate information based on frequency bands—often isolating high-frequency noise or texture from low-frequency structural content [10]. In general vision tasks, this property has been exploited for domain adaptation and generating robust feature representations that are less sensitive to superficial visual changes [11]. While some recent studies have begun to explore spectral features for texture analysis, few have successfully integrated them into the MIL framework to solve the specific challenges of WSI classification. Specifically, the utility of spectral differences for distinguishing between hard negative instances and true positives, and the computational advantage of frequency-domain global modeling over quadratic spatial attention, have not been fully exploited. Our work bridges this gap by introducing a unified framework that combines the efficiency and robustness of spectral analysis with the morphological precision of spatial features.

III. METHOD

A. Problem Formulation

We formulate WSI classification as a Multiple Instance Learning (MIL) problem. Each WSI is pre-processed into a bag of N instances. Each instance is represented by a feature embedding extracted from a pre-trained network. The bag is assigned a label $Y \in \{0, 1\}$ based on the standard MIL assumption: it is positive if and only if it contains at least one positive instance. Formally:

$$Y = \begin{cases} 0, & \text{if } \sum_{i=1}^{N} y_i = 0, \\ 1, & \text{otherwise,} \end{cases} \tag{1}$$

where $y_i \in \{0, 1\}$ is the unobserved label for instance p_i. Our goal is to predict the bag label Y from the set of instance embeddings.

B. Global Frequency Fusion Module

Given the spatial feature embedding $h_i \in \mathbb{R}^{H \times W \times C}$ of the i-th instance derived from a pre-trained feature extractor, we note that these local instance features, residing in the abstract feature domain rather than the original RGB space, lack periodic boundary conditions. Direct application of the FFT for frequency conversion would inevitably trigger the Gibbs phenomenon due to signal discontinuities at the boundaries, resulting in significant high-frequency noise caused by spectral leakage. Consequently, this module employs the Two-Dimensional Discrete Cosine Transform (2D-DCT) for frequency feature extraction. Its implicit even symmetric extension property ensures signal continuity at the edges, rendering it inherently more suitable for non-periodic global frequency feature extraction.

We project the spatial features into the frequency domain to generate the spectral feature map F^i_{freq}. Leveraging the energy compaction property of the DCT, the majority of energy associated with macro-structural tissue patterns is concentrated in the low-frequency components, whereas fine-grained textures and noise are distributed across the high-frequency bands. To further decouple redundancy across feature channels, we introduce an orthogonal basis decomposition strategy along the channel dimension. This process extracts the Direct Current (DC) component, representing the global context, and the Alternating Current (AC) components, capturing detailed variations, to construct the initial global frequency token F_g.

$$F_{DC} = F^i_{freq}[:, :, 0, 0]$$

$$F_{AC} = \text{MLP}(\text{Flatten}(F^i_{freq}[:, :, 0 : K, 0 : K]))$$

$$F_g = F_{DC} + \boldsymbol{\gamma} \odot F_{AC}$$

where $\boldsymbol{\gamma} \in \mathbb{R}^C$ denotes a learnable spectral modulation vector. It imposes a soft gating mechanism on the AC component F_{AC} via a channel-wise Hadamard product, thereby dynamically filtering and preserving features across different frequency bands. To mitigate domain shifts that manifest as spectral energy discrepancies (e.g., staining variations), we introduce Learnable Spectral Prompts $P \in \mathbb{R}^{K \times D}$. These prompts facilitate Spectral-Channel Decoupling, dynamically suppressing acquisition noise while enhancing diagnostic mid-frequency textures. Unlike filters with fixed weights, these prompts interact dynamically with the frequency features via a Multi-Head Attention mechanism:

$$\hat{F}_g = \text{LayerNorm}(F_g + \text{MultiHeadAttn}(Q = F_g, K = V = P)) \tag{2}$$

To fundamentally mitigate the interference of noisy samples on the discrimination between positive and negative instance features, this module incorporates an uncertainty estimation mechanism. Specifically, an Uncertainty Head projects the frequency features onto a scalar u_i that quantifies the "unreliability" of the sample:

$$u_i = \sigma(\mathcal{W}_u \cdot \hat{F}_g + b_u) \in (0, 1) \tag{3}$$

where σ denotes the Sigmoid activation function. A value of u_i approaching 1 indicates disordered spectral characteristics, identifying the instance as a high-noise sample, whereas a lower value indicates a credible sample. This uncertainty parameter functions as a dynamic weighting factor in the subsequent loss function to adaptively regulate the model's optimization trajectory. The detailed derivation of the objective function is elaborated in Section Joint Optimization Objectives.

C. Differential Frequency Guidance Module

In contrast to the GFFM, which employs DCT on abstract feature maps, the DFGM module applies the FFT directly to raw pathological patches to conduct microscopic textural analysis. This design explicitly exploits the inherent Spatial Quasi-periodicity of biological tissues in the RGB domain, where repetitive cellular arrangements form periodic signals that are mathematically optimal for FFT decomposition.

This module starts by converting every patch p_i into the frequency domain via a two-dimensional FFT applied to the RGB components separately of the patch:

$$\mathbf{m}_i(u,v) = \sum_{x=0}^{H-1}\sum_{y=0}^{W-1} \mathbf{p}_i(x,y)\cdot e^{-j2\pi\left(\frac{ux}{H}+\frac{vy}{W}\right)} \tag{4}$$

The resulting magnitude spectrum $\mathbf{m}_i$ is subsequently flattened and projected into a d-dimensional latent space as x_i by a learnable linear mapping. Then, a lightweight spectral saliency encoder assigns each processed feature a scalar score $s_i = f(\mathbf{x}_i;\theta_e)$. The k highest-scoring indices are selected, and their embeddings are averaged to form a single key malignant exemplar $\mathbf{x}_e = \frac{1}{k}\sum_{j\in\mathcal{K}}\mathbf{x}_j$.

The central idea of DFGM rests on the empirical observation that positive bags, slides harboring at least one malignant focus, tend to exhibit consistently larger spectral amplitudes than purely negative bags. By subtracting the exemplar component $\mathbf{x}_e$ from every instance embedding and rectifying the residual, we explicitly magnify the inter-class discrepancy. Formally, the differential frequency feature for patch i is computed as:

$$\mathbf{d}_i = \max\left(0, \mathbf{x}_i - \mathbf{x}_e\right) \tag{5}$$

In positive bags, instances that share spectral signatures similar to the exemplar retain substantial activation after non-negative operation, whereas most negative-bag instances—whose amplitudes cluster around lower values—are suppressed mostly. This selective suppression sharpens the decision boundary at the bag level.

To exploit this refined representation, $\mathbf{x}_e$ is treated as a query vector and $\{\mathbf{d}_i\}_{i=1}^{N}$ as keys and values in a cross attention mechanism. The resulting attention sequences is further combined with the exemplar vector $\mathbf{x}_e$ to integrate the discriminative frequency differences with the key malignant signatures.

D. Selective Multi-Cognition Block

To effectively fuse global and local information in spatial space, we propose the Selective Multi-Cognition Block, a compact module designed to capture multi-scale local patterns and selectively integrate them with global context. The SMC Block processes input features through a streamlined pipeline, as detailed below:

First, feature re-weighting stabilizes the input distribution using LayerNorm (LN) and learnable scalars to balance normalized and original features. the preprocessed feature $h_{\text{pre}} = w_1\cdot|h_0|_{\text{LN}} + w_2\cdot h_0$ is compressed via a down-projection layer to a lower dimension c, resulting in $h_d \in \mathbb{R}^{H\times W\times c}$. A multi-scale local perception module then extracts diverse local patterns using three parallel depth-wise convolutions [12] (DWConv) with kernel sizes 3×3, 5×5, and 7×7, which capture fine to coarse local structures. Their outputs are averaged and aggregated with a residual connection to preserve initial information.

Subsequently, a selective masking module identifies task-relevant local features from the multi-scale output f_{dw}. These features are normalized to obtain a selective mask α (summing to 1), which weights the sequence s to retain critical local features $s' = \alpha \odot s$, where s_i is the i-th element of s, and $\odot$ denotes element-wise multiplication.

Finally, a global-local fusion step refines and integrates the features. The masked feature is processed with a 1×1 point-wise convolution to model channel interactions, followed by GeLU [13] activation and up-projection to restore the original dimension. This refined local feature is fused with the initial global feature h_0 via a residual connection, ensuring global context is preserved while local patterns are adaptively integrated.

E. Joint Optimization Objectives

To mitigate non-pathological noise in WSI and enforce semantic consistency across feature domains, we propose a joint optimization framework rooted in Bayesian deep learning. Unlike standard linear weighting, our approach integrates a Spectrum-Guided Global Uncertainty Loss with Cross-View Mutual Distillation to simultaneously enhance robustness and discriminability.

1) Spectrum-Guided Global Uncertainty Loss: WSI analysis is inherently plagued by aleatoric uncertainty arising from artifacts such as staining variations, bubbles, and blurring. Standard Cross-Entropy (CE) loss treats all samples uniformly, often leading to overfitting on high-frequency noise. To address this, we leverage the uncertainty scalar u_i derived from the GFFM module to dynamically modulate the optimization trajectory. Formulating the problem as heteroscedastic regression, we model observations as Gaussian variables with variance controlled by u_i. The resulting Maximum Likelihood Estimation (MLE) objective is defined as:

$$\mathcal{L}_{Global} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{1+u_i}\mathcal{L}_{CE}(p_i^{fusion}, y_i) + \lambda_{reg}\cdot\log(1+u_i)\right) \tag{6}$$

This formulation enforces two adversarial mechanisms:

Risk Attenuation: The coefficient $\frac{1}{1+u_i}$ acts as an adaptive gate. As $u_i \to 1$ (indicating high spectral noise), the classification loss is attenuated, effectively blocking gradients from noisy samples to prevent backbone contamination.

Uncertainty Regularization: The term $\lambda_{reg}\cdot\log(1+u_i)$ penalizes trivial uncertainty estimation. It forces the model to increase u_i only when definitive spectral evidence of noise exists (e.g., abnormal high-frequency energy spikes), preventing the model from minimizing loss by simply ignoring hard samples.

2) Cross-View Mutual Distillation: To align semantic representations across the Difference Frequency and Global Fusion domains, we introduce Cross-View Mutual Distillation. We employ Bidirectional Kullback-Leibler (KL) Divergence to enforce consistency between the prediction distributions p^{freq} and $p^{spatial}$, fostering symmetric mutual learning:

$$\mathcal{L}_{KD} = \frac{1}{2}\left(D_{KL}(p^{fusion}||p^{diff}) + D_{KL}(p^{diff}||p^{fusion})\right) \tag{7}$$

TABLE I
COMPARISON OF FD-MIL WITH OTHER MIL METHODS ON THREE DATASETS WITH RESNET-50 AS FEATURE EXTRACTOR. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD. THE BACC REFERS TO BALANCED ACCURACY.

	FLOPs	Camelyon 16			TCGA-BRCA			TCGA-NSCLC		
	(M)	F1	AUC	BACC	F1	AUC	BACC	F1	AUC	BACC
AB-MIL [7]	131.2	0.827±0.024	0.851±0.015	0.831±0.021	0.799±0.035	0.892±0.022	0.744±0.031	0.858±0.028	0.904±0.019	0.821±0.030
CLAM-MB [8]	786.7	0.790±0.019	0.845±0.012	0.802±0.018	0.784±0.028	0.882±0.015	0.746±0.025	0.824±0.022	0.887±0.014	0.819±0.021
DS-MIL [14]	1090.9	0.884±0.015	0.917±0.011	0.871±0.014	0.829±0.026	0.882±0.021	0.781±0.024	0.882±0.020	0.946±0.015	0.867±0.018
DTFD-MIL [15]	656.2	0.893±0.010	0.932±0.008	0.881±0.011	0.792±0.024	0.873±0.019	0.763±0.022	0.859±0.018	0.917±0.012	0.843±0.016
TransMIL [9]	2876.0	0.863±0.022	0.867±0.018	0.855±0.020	0.827±0.031	0.906±0.025	0.795±0.029	0.867±0.025	0.936±0.017	0.858±0.024
RRT-MIL [6]	1593.8	0.902±0.009	0.967±0.006	0.889±0.010	0.844±0.018	0.914±0.014	0.807±0.019	0.901±0.015	**0.954±0.009**	0.884±0.012
FD-MIL(Ours)	**978.0**	**0.929±0.005**	**0.976±0.003**	**0.928±0.006**	**0.856±0.012**	**0.925±0.010**	**0.821±0.015**	**0.912±0.008**	0.948±0.006	**0.901±0.009**

This constraint ensures multi-view consensus on the feature manifold: the noise-robust fusion branch guides the difference branch to ignore artifacts, while the morphology-sensitive difference branch calibrates the fusion branch on fine-grained details.

Consequently, the total objective function is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{global} + \mathcal{L}_{differential} + \beta \cdot \mathcal{L}_{KD} \tag{8}$$

where $\mathcal{L}_{differential}$ represent the standard classification loss for the DFGM branch, and β balances the distillation weight.

IV. EXPERIMENT

A. Datasets and Evaluation Metrics

Camelyon16 [16] is a public WSI dataset for detecting breast cancer metastases. We adopted the three-fold cross-validation protocol to mitigate selection bias, ensuring that each slide appears in both training and validation splits across folds. TCGA-BRCA and TCGA-NSCLC are large-scale WSI datasets from The Cancer Genome Atlas. Following prior work, we applied five-fold cross-validation with patient-level fold isolation to prevent data leakage.

Note that all three datasets exhibit pronounced class imbalance. Therefore, we report balanced accuracy—the average of recall obtained on each class—to provide a more reliable estimate of model performance under class-skewed conditions.

B. Main Results

1) Comparison with State-of-the-Art: As shown in Table 1, FD-MIL consistently outperforms leading methods (including TransMIL and RRT-MIL) across all benchmarks. On Camelyon16, FD-MIL achieves an AUC of 0.976 and F1 of 0.929, surpassing RRT-MIL by +3.9% in BACC. This gain highlights the efficacy of the DFGM, which sharpens decision boundaries via spectral residuals. On the heterogeneous TCGA-BRCA and TCGA-NSCLC datasets, FD-MIL demonstrates superior robustness, achieving BACC scores of 0.821 (+1.4%) and 0.901 (+1.7%), respectively. This stability is attributed to the GFFM, which effectively decouples stain-related spectral noise from pathological features, preventing overfitting to artifacts.

2) Computational Efficiency Analysis: FD-MIL achieves a compelling balance between precision and cost. With 978M FLOPs, it reduces computational burden by 66% and 38% compared to Trans-MIL and RRT-MIL, respectively, by avoiding quadratic self-attention complexity. While slightly heavier than simple pooling methods, the substantial diagnostic gains justify this trade-off for precision-critical clinical applications.

3) Ablation Study: To validate the proposed framework, we analyzed the contribution of each stream on the Camelyon16 dataset in Table 2. Using the SMC block as the Spatial Baseline (86.9% BACC), the integration of GFFM forms the Global Fusion Branch, improving BACC to 88.4%. This gain confirms that global frequency tokens effectively filter high-frequency acquisition noise, complementing morphological features. Separately, the Differential Branch (SMC+DFGM) yields a significant boost to 91.5% BACC, validating that exploiting spectral residuals amplifies the discriminability of hard-negative instances. Finally, the Full Synergy model achieves peak performance (92.8% BACC, 0.976 AUC). By employing joint optimization with spectrum-guided uncertainty and cross-view distillation, FD-MIL harmonizes the noise-robust global context with discriminative differential signals, proving the two branches are mutually reinforcing.

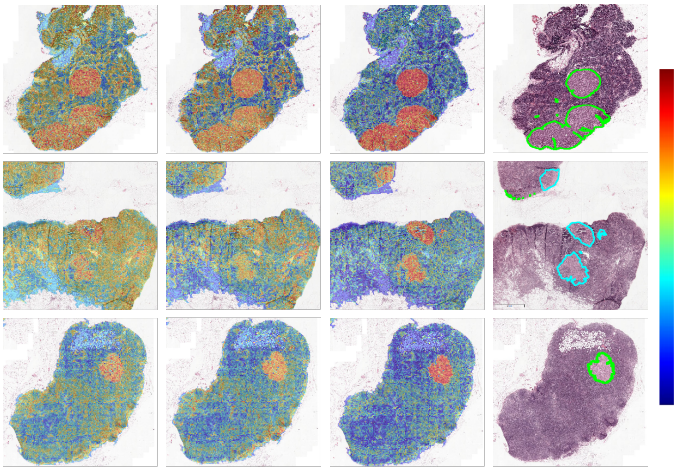
TABLE II
ABLATION EXPERIMENTS ON THE IMPROVEMENT CONTRIBUTIONS OF DIFFERENT MODULES IN FD-MIL, CONDUCTED ON THE CAMELYON 16 DATASET.

Method Configuration	**Components**			**Camelyon 16**		
	SMC	**GFFM**	**DFGM**	**F1**	**AUC**	**BACC**
Spatial Baseline	✓			0.891	0.949	0.869
Global Fusion Branch	✓	✓		0.905	0.952	0.884
Differential Branch Integration	✓		✓	0.918	0.961	0.915
FD-MIL (Full Synergy)	✓	✓	✓	**0.929**	**0.976**	**0.928**

The full-module FD-MIL shows a 2.7% improvement in AUC compared to the visual baseline with the SMC block, confirming the effectiveness of integrating the frequency domain with the spatial module. Incorporating DFGM increases the BACC from 86.5% (pure baseline) to 91.5%, demonstrating the efficacy of using frequency differences between positive and negative instances to re-calibrate the characteristic distribution of the overall samples. This also highlights its notable robustness in positive-negative imbalanced datasets.

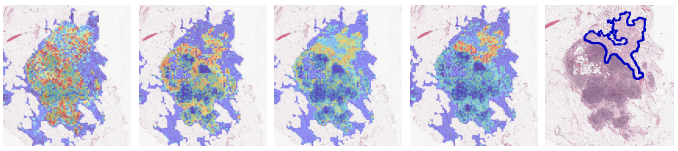
C. Visualization

To visualize the effectiveness of our proposed modules, we conduct a qualitative analysis on the Camelyon16 dataset. We select a representative case and visualize the attention maps generated by our model and other models. The attention maps highlight the regions of interest that the model focuses on for classification. The results are shown in Figure.2. In contrast to existing models whose attention is often dispersed and lacks focus, our approach precisely localizes fine-grained textural cues. The generated attention maps exhibit sharp boundaries



(a) AB-MIL (b) TRANS-MIL (c) **our FD-MIL** (d) Ground Truth

Fig. 2. Comparison of attention map quality across different models aligned with ground truth contours of positive tissue on representative Camelyon16 tumor slides. Red indicates high positive instance probability, and blue indicates low.



(a) Baseline (b) w/SMC (c) w/GFFM (d) w/DFGM (e) Ground Truth

Fig. 3. Ablation of attention distribution over positive instance probabilities by internal modules in FD-MIL.

between normal and neoplastic tissues, indicating the effectiveness of our frequency-domain extraction of pathological details and thereby enhancing classification performance.

Further visualization, as shown in the Fig.3, based on ablation experiments demonstrates that the SMC baseline can perceive local positive instances but is accompanied by substantial noise interference. After incorporating GFFM, the noise in negative instance regions within the attention map is significantly reduced; as shown in Figure 3c, the attention is only directed toward obvious positive instances, indicating that global key frequency-domain features can effectively eliminate false positive noise. After adding DFGM, the boundaries between positive and negative instances become clearer, as illustrated in Figure 3d, which well align with the segmentation boundaries in the ground truth. This demonstrates that frequency-domain differentiation can amplify the feature space distance between positive and negative instances, thereby facilitating segmentation.

V. Conclusion

This study addresses the challenges inherent to multiple instance learning (MIL) arising from its bag-level labeling paradigm, demonstrating the effectiveness of integrating frequency-domain processing into MIL frameworks. We propose a framework that fuses frequency-domain and spatial feature extraction, leveraging global frequency information and inter-instance frequency differences to refine sample distributions while integrating spatial context. Experimental results confirm significant performance improvements over baselines, validating the framework's ability to enhance MIL-based classification by combining frequency and spatial information.

References

[1] K. Thandiackal, B. Chen, P. Pati, G. Jaume, D. F. Williamson, M. Gabrani, and O. Goksel, "Differentiable zooming for multiple instance learning on whole-slide images," in *European Conference on Computer Vision*. Springer, 2022, pp. 699–715.

[2] X. Wang, J. Xiang, J. Zhang, S. Yang, Z. Yang, M.-H. Wang, J. Zhang, W. Yang, J. Huang, and X. Han, "Scl-wc: Cross-slide contrastive learning for weakly-supervised whole-slide image classification," *Advances in neural information processing systems*, vol. 35, pp. 18 009–18 021, 2022.

[3] Z. Zhu, L. Yu, W. Wu, R. Yu, D. Zhang, and L. Wang, "Murcl: Multi-instance reinforcement contrastive learning for whole slide image classification," *IEEE Transactions on Medical Imaging*, vol. 42, no. 5, pp. 1337–1348, 2023.

[4] P. Chikontwe, M. Kim, J. Jeong, H. J. Sung, H. Go, S. J. Nam, and S. H. Park, "Fr-mil: Distribution re-calibration based multiple instance learning with transformer for whole slide image classification," *IEEE Transactions on Medical Imaging*, 2024.

[5] K. Han and A. M. Y. Koay, "Hits: Hierarchical interpretable time series classification via multiple instance learning," in *2025 International Joint Conference on Neural Networks (IJCNN)*, 2025, pp. 1–8.

[6] W. Tang, F. Zhou, S. Huang, X. Zhu, Y. Zhang, and B. Liu, "Feature re-embedding: Towards foundation model-level performance in computational pathology," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 11 343–11 352.

[7] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *International conference on machine learning*. PMLR, 2018, pp. 2127–2136.

[8] G. Xu, Z. Song, Z. Sun, C. Ku, Z. Yang, C. Liu, S. Wang, J. Ma, and W. Xu, "Camel: A weakly supervised learning framework for histopathology image segmentation," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10 681–10 690, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:201651201

[9] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji *et al.*, "Transmil: Transformer based correlated multiple instance learning for whole slide image classification," *Advances in neural information processing systems*, vol. 34, pp. 2136–2147, 2021.

[10] K. Yi, Q. Zhang, W. Fan, L. Cao, S. Wang, G. Long, L. Hu, H. He, Q. Wen, and H. Xiong, "A survey on deep learning based time series analysis with frequency transformation," 2025. [Online]. Available: https://arxiv.org/abs/2302.02173

[11] L. Jiang, B. Dai, W. Wu, and C. C. Loy, "Focal frequency loss for image reconstruction and synthesis," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 13 899–13 909.

[12] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[13] M. Lee, "Gelu activation function in deep learning: a comprehensive mathematical analysis and performance," *arXiv preprint arXiv:2305.12073*, 2023.

[14] B. Li, Y. Li, and K. W. Eliceiri, "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 318–14 328.

[15] H. Zhang, Y. Meng, Y. Zhao, Y. Qiao, X. Yang, S. E. Coupland, and Y. Zheng, "Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 802–18 812.

[16] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol *et al.*, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *Jama*, vol. 318, no. 22, pp. 2199–2210, 2017.