FR-MIL: Distribution Re-Calibration-Based Multiple Instance Learning With Transformer for Whole Slide Image Classification

Philip Chikontwe, Meejeong Kim, Jaehoon Jeong, Hyun Jung Sung, Heounjeong Go, Soo Jeong Nam, and Sang Hyun Park, *Member, IEEE*

***Abstract*— In digital pathology, whole slide images (WSI) are crucial for cancer prognostication and treatment planning. WSI classification is generally addressed using multiple instance learning (MIL), alleviating the challenge of processing billions of pixels and curating rich annotations. Though recent MIL approaches leverage variants of the attention mechanism to learn better representations, they scarcely study the properties of the data distribution itself *i.e.*, different staining and acquisition protocols resulting in intra-patch and inter-slide variations. In this work, we first introduce a distribution re-calibration strategy to shift the feature distribution of a WSI bag (instances) using the statistics of the max-instance (critical) feature. Second, we enforce class (bag) separation via a metric loss assuming that positive bags exhibit larger magnitudes than negatives. We also introduce a generative process leveraging Vector Quantization (VQ) for improved instance discrimination *i.e.*, VQ helps model bag latent factors for improved classification. To model spatial and context information, a position encoding module (PEM) is employed with transformer-based pooling by multi-head self-attention (PMSA). Evaluation of popular WSI benchmark datasets reveals our approach improves over state-of-the-art MIL methods. Further, we validate the general applicability of our method on classic MIL benchmark tasks and for point cloud classification with limited points. https://github.com/PhilipChicco/FRMIL**

***Index Terms*— Histopathology, multiple instance learning, whole slide images, weakly supervised learning.**

Manuscript received 20 March 2024; revised 18 June 2024 and 27 July 2024; accepted 14 August 2024. Date of publication 20 August 2024; date of current version 2 January 2025. This work was supported in part by the Smart HealthCare Program funded by the Korean National Police Agency under Grant 220222M01 and in part by the Institute for Information and Communication Technology Planning and Evaluation (IITP) funded by Korean Government [Ministry of Science and ICT (MSIT)] through the Artificial Intelligence Innovation Hub under Grant 2021-0-02068. *(Philip Chikontwe and Meejeong Kim are co-first authors.) (Corresponding authors: Sang Hyun Park; Soo Jeong Nam.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Asan Medical Center with Institutional Review Board (IRE) under Approval No. 2019-1192.

Philip Chikontwe, Jaehoon Jeong, and Sang Hyun Park are with Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu 42988, South Korea (e-mail: Philip_Chikontwe@hms.harvard.edu; j.hoon@dgist.ac.kr; shpark13135@dgist.ac.kr).

Meejeong Kim is with St. Mary's Hospital and the College of Medicine, The Catholic University of Korea, Seoul 03083, South Korea (e-mail: altec0372@gmail.com).

Hyun Jung Sung and Soo Jeong Nam are with the Asan Medical Center, Seoul 05505, South Korea (e-mail: shj78730@naver.com; soojeong_nam@amc.seoul.kr).

Heounjeong Go is with the Asan Medical Center, Seoul 05505, South Korea, and also with the College of Medicine, University of Ulsan, Ulsan 44610, South Korea (e-mail: damul37@amc.seoul.kr).

Digital Object Identifier 10.1109/TMI.2024.3446716

I. Introduction

IN MODERN clinical workflows, the adoption of digital pathology has emerged as a vital tool in modern medicine for cancer diagnosis and treatment planning [1], [2]. This can be attributed to the ever growing popularity of digitized tissue samples with scanners *i.e.*, whole slide images (WSI), facilitating global distribution for research, teaching, and diagnostics. Digital slides also afford the possibility to apply several image analysis techniques for classification, segmentation, and detection applications. Note that algorithmic solutions are beneficial in many contexts. They can reduce the tedious and laborious nature of providing quantification. Additionally, they can act as a second reader to reduce expert inter-reader variability. At the same time, advanced techniques using machine learning [3], [4] with convolutional neural networks (CNN) for WSI analysis are gaining popularity but are still hindered by the very nature of WSIs *i.e.*, large and extremely high-resolution images, lack of precise labeling, and staining (color) variations [4]. This motivates the need for memory-efficient methods that mitigate the need for fine-grained labels and are fairly interpretable.

To circumvent these challenges, a popular approach is to employ multiple instance learning (MIL) [5], [6], a weakly supervised learning [7] paradigm where the complexities of computational training can be alleviated by (1) extracting patches from WSI as independent instances, and (2) pool over the set of unordered instances via global aggregation operators for classification tasks (*e.g.*, cancer grading or subtyping) when learning with inexact or incomplete labels. Owing to recent advances in deep learning [8], [9], MIL-based histopathology [10], [11], [12], [13], [14] analysis has achieved notable success [1], [15], [16], [17]. In particular, existing works [15], [18], [19], [20], [21] leverage attention-based variants with Transformer models [8], [22] to model long-range instance correlations, including context-aware modules. However, the use of Transformers with several

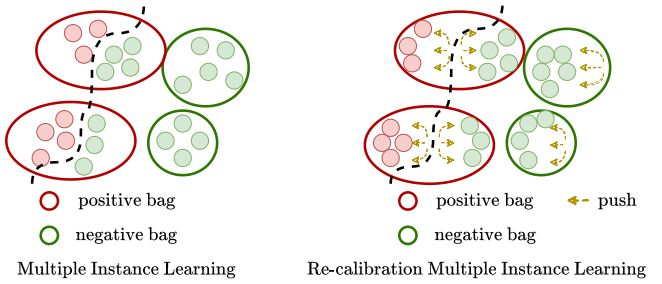


Fig. 1. Intuition of the proposed framework. Compared to standard MIL, our method encourages better feature separation within bags via re-calibration. Positive bags may contain at least one positive instance (light red), with negative bags having only similar instances (green).

Multi-head Self-Attention (MSA) blocks can be computationally prohibitive, and less over-parameterized designs are more desirable. Also, while attention mechanisms are effective, prior works [18] only focus on instance re-weighting, and are sensitive to both the choice of instance encoders and scales of patches employed - albeit, they rarely pay attention to the data distribution itself *i.e.*, different staining and acquisition protocols resulting in intra-patch and inter-slide variations.

In this work, we propose a **F**eature **R**e-calibration based MIL framework (FR-MIL), building upon prior MIL approaches [18], [19], [20] (Fig. 1). We hypothesize that features from positive and negative bags (binary MIL) exhibit larger and smaller feature magnitudes, respectively, and this prior can be used for improved representation learning [23]. Second, general vision tasks such as few-shot learning and anomaly detection [24], [25] leverage feature/distribution re-calibration for improved generalization from few (limited) examples by transferring statistics from classes with sufficient examples [26]. Note that for MIL, instances are not always i.i.d. [19], and certain WSI may have highly isolated positive instances (diseased regions) *i.e.*, ($\leq$ 10%) making learning difficult. Recent works [27], [28], [29] employing generative models based on variational autoencoders (VAE) [30] in the MIL framework can capture key instance characteristics per class while reducing the uncertainty of positive instance labels. Inspired by this, we introduce an extension of FR-MIL coined FR-MIL++ *i.e.*, introduces a generative process in FR-MIL based on Vector Quantization (VQ) [31] to identify latent factors that generate observed instances in a bag. This approach encourages disentanglement and class separation without explicit use of a distance objective (FR-MIL).

Herein, we first establish a simple non-parametric MIL baseline to highlight the phenomena (Fig. 2) of re-calibrating instance features *i.e.*, shift the original distribution after critical instance selection to obtain improved scores comparable to classic MIL operators (*i.e.*, , max/mean-pooling) [5]. Further, we incorporate re-calibration in our framework to encourage maximal class/bag separation via feature embedding loss, alongside a context positional encoding module (PEM) [19] followed by a single Pooling Multi-head Self-Attention block (PMSA) [32] for bag classification. Finally, as opposed to prior generative MIL methods [27], [28], [29] that focus on instance-level tasks, FR-MIL++ uses VQ as a memory bank for critical instances during learning, while also regularizing learning via reconstruction of instance latent features.

The main contributions of this work are as follows:

- We show that feature re-calibration using the max instance embedding is a simple yet powerful technique for MIL, and introduce a feature magnitude loss to learn better bag separation.
- To obtain robust bag embeddings, we leverage a positional encoder and a single self-attention block for instance aggregation (pooling), and improve over prior state-of-the-art MIL methods with considerable margins on benchmark datasets.

Part of this work's results were published in a preliminary work [33]. The current work substantially extends it with the following contributions:

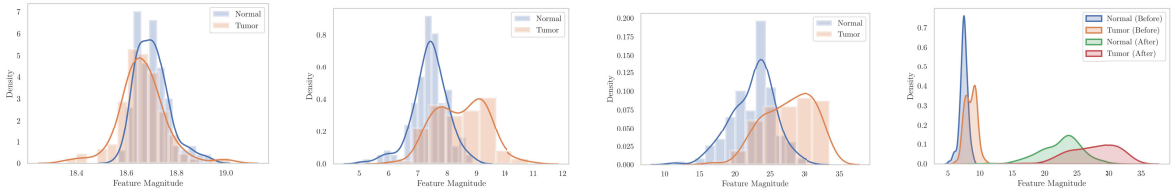
- We further apply FR-MIL to a public WSI Lung cancer subtyping dataset and validate the general applicability of FR-MIL *i.e.*, evaluation on classic MIL benchmark tasks and verify its potential for Point Cloud classification with limited points.
- We further extend FR-MIL to FR-MIL++, a novel framework that incorporates a vector quantized (VQ) generative model for the first time to improve instance-level learning with reduced ambiguity. FR-MIL++ helps capture meaningful representations in the observed instances, and avoids using distance-based objectives with fixed hyper-parameters for class separation.
- Extensive quantitative and qualitative evaluations with additional ablations across several datasets reveal the benefit of the proposed method.

II. Related Works

A. Multiple Instance Learning for Pathology

In computational pathology, MIL [3], [5], [6], [34] is a popular strategy to address weakly supervised learning problems such as segmentation [35] and tumor detection [1], [10], [17], [21], [36], [37], [38], [39], [40]. Note that conventional MIL methods consider handcrafted feature aggregators such as mean- and max-pooling [5], whereas recent works have shown parameterizing aggregation with neural networks is better [17], [18]. For instance, Ilse et al. [17] first introduced a learnable attention-based aggregation operator that re-weighted the contribution of each instance in a bag (set of instances). Follow-up methods [18], [19] consider more efficient and complex strategies to model instance interactions inspired by the self-attention mechanisms in Transformer models [8].

In particular, Li et al. [18] use non-local attention to re-weight instances relative to the highest scoring instance (critical) in a bag - a simple yet effective approach. On the other hand, to incorporate context-aware learning Shao et al. [19] proposed spatial positional encoding (PE) modules with multi-scale learning strategies using a Transformer encoder. However, this approach is relatively sensitive to the depth of PE layers (*i.e.*, x3), does not explicitly pool all instances to a single bag representation, and is computationally prohibitive. Our work differs in that we employ single-scale learning with fewer positional encoding modules,



44.96% (a) 83.72% (+38.76%) (b) 89.10% (+44.14%) (c) (d)

Fig. 2. Normalized density plots of the mean feature magnitudes on the CAMELYON16 [36] train-set, with test-set accuracy and improvements (red color). (a) Original feature magnitudes. (b) Max-instance calibrated-based features. (c) Features learned by our FR-MIL model, and (d) overlay of (b) & (c) *i.e.*, before and after. The legend(s) refers to slide-level labels.

and explicitly shift features based on selected key instances before performing pooling via multi-head self-attention.

B. Multiple Instance Learning for Computer Vision

Deep MIL has been used for weakly supervised vision tasks such as object localization [23], [41] and detection [24], [25], [42]. For instance, Lee et al. [23] leveraged MIL for weakly-supervised uncertainty-based action localization where background/normal actions were modeled as out-of-distribution while considering their inconsistency in a sequence with video-level labels only. Along this line of work, Feng et al. [25] introduced MIST, a framework for weakly supervised anomaly detection in videos. MIL was used to learn discriminative representations for identifying anomalous events and also enable initial pseudo-label generation for subsequent re-training. It is worth noting that unlike standard MIL approaches for WSI that use single batches when training, our work uses both positive and negative bags jointly to learn more discriminative features inspired by MIST.

C. Generative Multiple Instance Learning

To learn meaningful low-dimensional feature representations in the MIL framework, recent studies [27], [28], [29] have combined discriminative and generative-based models, with an emphasis on improving instance-level classification. Notably, an early work by Ghaffarzadegan et al. [28] leveraged concepts behind maximum-likelihood and margin methods [43], [44] with deep learning. They introduced a novel approach that learned instance distributions using a Variational Autoencoder (VAE) [30], using it to estimate the similarity between positive and negative instances based on the reconstruction error of a VAE trained with negative instances only. Consequently, positive label uncertainty could be resolved for improved instance classification. Building on this work, Zhang et al. [29] introduced MI-VAE, a framework able to simultaneously infer both bag and instance latent factors, while also explicitly incorporating dependencies among instances.

More recently, subsequent research has delved into identifying instance-level causal representations [45], [46], [47] derived from bag-level weak supervision, aiming to further enhance robustness in instance-level prediction tasks through the use of an identifiable VAE in CausalMIL [27]. Inspired by these works, we instead extend our approach in FR-MIL to employ a Vector Quantized Variational Autoencoder (VQ-VAE) [31] for instance-level representation learning. A key motivation is that by modeling latent as discrete representations, the VQ-VAE overcomes issues such as posterior collapse and variance observed in traditional VAEs, and ensures faster convergence. The use of generative instance learning in our framework addresses some key drawbacks of our initial framework, further detailed in Section III-E.

III. Method

A. Overview

1) Background: In the multiple instance learning (MIL) paradigm, a set of training instances is considered a *bag*, and each bag is associated with a global label *i.e.*, instance level-labels are unknown. Let $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{N}$ denote a dataset, where $X_i = \{x_{i,1:N_i}\}$ denotes the i-th bag with N instances and Y is the bag label. In the standard MIL scenario *i.e.*, binary task, a bag is considered positive if at least one instance is positive, and negative if all instances are negative [5]. Note the number of instance per bag vary significantly, and existing MIL formulations follow:

$$\mathbf{F}(X) = \zeta(\kappa(\{\phi(\mathbf{x_i}) : \mathbf{x_i} \in X\})), \tag{1}$$

where $\phi : \mathbb{R}^{d_{in}} \rightarrow \mathbb{R}^{d_{out}}$ is an instance level operator, $\kappa : \mathbb{R}^{n \times d_{out}} \rightarrow \mathbb{R}^{d_{out}}$ is a permutation-invariant pooling function over bags and $\zeta : \mathbb{R}^{d_{out}} \rightarrow \mathbb{R}^{class}$ is a bag-level classifier [48]. MIL can be modeled differently depending on the choices of ϕ and κ *i.e.*, (1) instance-based approach: ϕ is an instance classifier and κ a pooling aggregator over instance scores, and (2) embedding-based: ϕ is a feature extractor mapping each instance to an embedding and κ pools embeddings to a single embedding to produce a bag score.

2) Baseline: In this work, a key hypothesis is that normal and positive bags may exhibit different magnitudes *i.e.*, negative instances have similar values whereas positives may show larger values as motivated by the findings of Lee et al. [23] on uncertainty estimation of background actions in video sequences. Thus, we design MIL baseline for the binary classification scenario *i.e.*, given instance features $\mathbf{H}_i = \{h_1, h_2, \ldots, h_n\}$ we obtain the mean feature magnitude per WSI as $\mu_i^c = \frac{1}{n}\sum ||\mathbf{H}_i||_2^2$, where n denotes the number of instances in a bag and c the WSI class. The probability $\mathbf{P}(y = 1.)$ of a bag is:

$$\mathbf{P}(y = 1.|\mu_i^c) = \frac{\min(\tau, \mu_i^c)}{\tau}, \tag{2}$$

where τ is the pre-defined maximum feature magnitude determined on the train-set only *i.e.*, point at which the

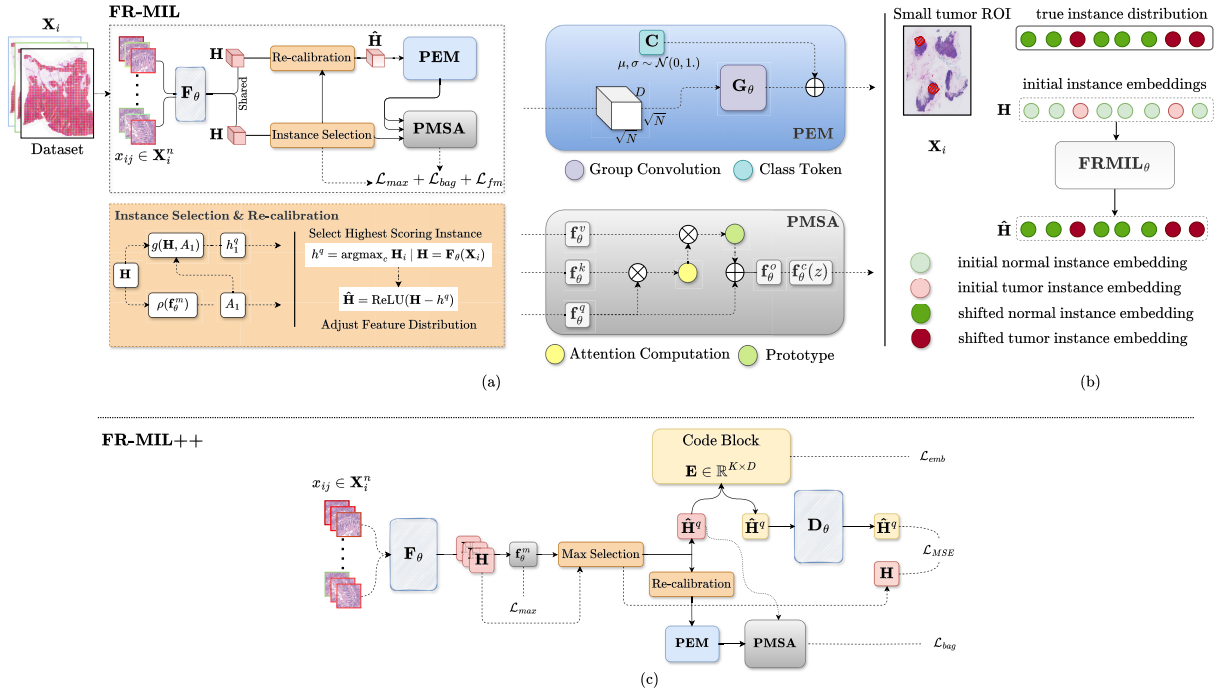


Fig. 3. (a) Overview of the proposed FR-MIL framework, and (b) Intuition of our proposed re-calibration mechanism in WSI with small tumor regions *i.e.*, the model maximizes instance features of otherwise isolated tumors and reduces the influence of normal/background features for improved representation learning. (c) Proposed FR-MIL++ that includes VQ-VAE based learning on key instances to better capture key latent instance representations for improved bag prediction without distance based objectives. Here, different colored *H* denoted the new updated features after the codeblock (yellow) and the selected to top-instance relating the original bag (red).

distributions first meet. Note that Eq. 2 is ensured to fall between 0 and 1, *i.e.*, $0 \leq \mathbf{P}(y = 1.|\cdot) \leq 1$. Moreover, the baseline is non-parametric and does not require training; we directly use extracted features $\mathbf{H}$. In Fig. 2, we present a density plot of obtained magnitudes on a given train-set *i.e.*, Camelyon16 [36]. We observe that while both normal and tumor bag curves appear to follow the Gaussian distribution (Fig. 2(a)), separation is non-trivial due to the presence of more normal than tumor instances ($\leq 10\%$) *i.e.*, low accuracy (44%) with $\tau = 18.8$. On the other hand, Fig. 2(b) shows that re-calibrating the initial distribution by subtracting the max-instance feature before computing the magnitudes creates better separability. To achieve re-calibration, we obtain a new mean denoted $\hat{\mu}_i^c = \frac{1}{n}\sum ||\hat{\mathbf{H}}_i||_2^2$, where $\hat{\mathbf{H}}_i = \{\mathbf{H} - h_{\max}\}$ given $h_{\max} = \text{argmax}_c\ \mathbf{H}_i$. This improves test accuracy by +39 with $\tau = 8.2$. Finally, Fig. 2(c) presents further improvements (+44) obtained when the intuitions of the baseline are used in our framework *i.e.*, the learned distribution of FR-MIL when trained with a feature magnitude loss $\mathcal{L}_{fm}$ and re-calibration - validating our hypothesis.

B. FR-MIL: Feature Re-Calibration & Instance Selection

Figure 3 shows an overview of the proposed framework for WSI classification. We consider a set of WSIs $\mathbf{X} = \{X_i\}$, each associated with slide-level labels $Y_i = \{0, 1\}$. Given input X_i we extract instance features $\mathbf{H} \in \mathbb{R}^D$ using a neural network $\mathbf{F}_\theta$ *i.e.*, $\mathbf{H_i} = \mathbf{F}_\theta(\mathbf{X_i})$, $\mathbf{F}$ can an ImageNet pre-trained or self-supervised mdoel [49], [50]. To re-calibrate instance features in a bag, we first select the critical instance in a bag based on probability and then use it to shift the features of all instances. We believe this has connections to feature normalization (mean-std, *e.g.*, Batch Normalization) to improve model learning stability. We assume that for negative bags, instance features will be i.i.d across bags, thus subtracting the critical instance per instance will produce a representation that is normally (Gaussian) distributed with the mean of the shifted bags features highlighting better separation via the critical instance (Fig. 2).

1) Instance Selection: Given the set of instance features $\mathbf{H}$, we select the max-instance (h^q) from the instances with the maximum probability (A) by max-pooling via classifier $\mathbf{f}_\theta^m$. Formally, $A = \rho(\mathbf{f}_\theta^m(\mathbf{H}))$ where ρ denotes the sigmoid function; ρ used here is analogous to the instance-based paradigm introduced earlier (Sec. III-A), and scores per instance are sorted and subsequently used to index the max-instance via operator $g(\cdot)$. The max score A is used to train the instance classifier $\mathbf{f}_\theta^m$ using the loss $\mathcal{L}_{max}$.

2) Feature Re-Calibration: Given h^q, this step follows the intuition of the baseline directly. Formally, initial instance features are shifted as follows:

$$\hat{\mathbf{H}} = \text{ReLU}(\mathbf{H} - h^q), \tag{3}$$

with ReLU applied to ensure non-negative values and stabilize training. While we select the max instance here, one can alternatively employ the *min* or mean instance instead (we provide evidence in Sec.V-B & Table V for different tasks). Nevertheless, we found *max* more beneficial in our evaluated WSI tasks. Herein, $\hat{\mathbf{H}}$ is employed in a spatially aware module (PEM) before final global embedding aggregation in PMSA.

To further incorporate the concept of distribution re-calibration in our framework, we draw connections

to prior work [23] by introducing a learning objective that enforces feature separation between bags. First, note that to effectively model the ambiguous normal/background features, the training procedure should employ both positive and negative bags simultaneously. To achieve this, unlike existing methods that use a single bag when training, we sample balanced bags fixed-sized per epoch (+/- bags) *i.e.*, we initialize a zero-tensor with the maximum bag size per dataset during training, and pad bag instance features. Formally, the feature magnitude loss $\mathcal{L}_{fm}$:

$$\mathcal{L}_{fm}(\hat{\mathbf{H}}_i^{pos}, \hat{\mathbf{H}}_i^{neg}, \tau) = \frac{1}{N}\sum_{n=1}^{N}(\max(0, \tau - ||\hat{\mathbf{H}}_i^{pos}||) + ||\hat{\mathbf{H}}_i^{neg}||), \quad (4)$$

where $\hat{\mathbf{H}}^{pos}$, and $\hat{\mathbf{H}}^{neg}$ are the positive- and negative-bag instance features, and τ is the pre-defined margin, respectively. Intuitively, $\mathcal{L}_{fm}$ enables us to model uncertainty by learning to produce large feature magnitudes for positive bags and smaller ones for negative bags.

C. Positional Encoding Module (PEM)

Our goal is to encourage the model to be spatially and context-aware. Due to the large size of WSI and the presence of thousands of instances per bag, achieving this was recently non-trivial and computationally infeasible. Note that in computer vision, encoding spatial information has proved useful for recognition tasks - mainly owed to the advances in transformer architectures [8], [22]. Existing methods applied to WSI employ either graph-based learning or multi-scale multi-stage pre-training to model the spatial relationships among different entities [20], [51], [52], [53], [54], [55]. In contrast, the positional encoding module (PEM) [19] is shown beneficial as no pre-training or graph construction is necessary. Inspired by this, we employ PEM to implicitly recover spatial organization from a set of ordered instances. It takes re-calibrated features $\hat{\mathbf{H}}$ as input, performs zero-padding to provide an absolute position for a convolution $\mathbf{G}$, and then concatenates the output with a class token $\mathbf{C}$ initialized from the normal distribution. Formally,

$$\begin{cases} \hat{\mathbf{H}} \in \mathbb{R}^{D} \rightarrow \hat{\mathbf{H}} \in \mathbb{R}^{B\times D\times H\times W}, \\ \acute{\mathbf{H}} = \text{concat}(\mathbf{C}, \mathbf{G}(\hat{\mathbf{H}})), \ \acute{\mathbf{H}} \in \mathbb{R}^{(N+1)\times D}. \end{cases} \quad (5)$$

In particular, features $\hat{\mathbf{H}}$ are re-shaped into a 2D image by first computing $\{H, W\}$ *i.e.*, $H = \sqrt{N} = \sqrt{n}$, where n is the number of instances in a bag similar to Shao et al. [19]; instances are assumed to follow square-grid (top→bottom, left→right) co-ordinate arrangement. B is the batch size, D is the instance feature dimension, and $\mathbf{G}$ is the 2D convolution layer that performs group convolution with kernel size 3×3, and 1×1 zero padding. $\acute{\mathbf{H}}$ denotes flattened restored features later employed for bag-level pooling. Note that PEM in prior work [19] used different sized convolutions in a hierarchical fashion - resulting in huge compute costs. For simplicity, we instead leverage a single layer to maintain computational feasibility.

D. Pooling With Multi-Head Self-Attention (PMSA)

In contrast to standard MIL aggregators [5], [21], [34] such as max- or mean-pooling, FR-MIL learns to create a single bag representation via a self-attention mechanism [8] agnostic to the order of instances. Formally, an attention function $\varphi(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ measures similarity between a query vector $\mathbf{Q}$ with key-value pairs $\mathbf{K} \in \mathbb{R}^{n\times m}$ and $\mathbf{V} \in \mathbb{R}^{n\times m}$ as:

$$\varphi(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{m}})\mathbf{V}. \quad (6)$$

where $\{d, m\}$ is the instance feature dimension. This can easily be extended to the multi-head setting where vectors $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ are first projected onto k different dimensional vectors *i.e.*, $\varphi(\mathbf{Q}W_j^{\mathbf{Q}}, \mathbf{K}W_j^{\mathbf{K}}, \mathbf{V}W_j^{\mathbf{V}})$, with each transformation having its own learnable parameters W.

To pool instance features $\acute{\mathbf{H}}$ to a single bag feature, we employ a single multi-head Transformer encoder following Lee et al. [32]. We consider the selected max-instance feature h^q as a query and $\acute{\mathbf{H}}$ as key-value pairs [18] *i.e.*, PMSA($h^q, \acute{\mathbf{H}}$). Note that this differs from prior works that self-attention to re-weight instances only, and the original formulation introduced in [32] where a learnable token is used as the query for pooling. In our module, the encoder consists of feed-forward networks $\{\mathbf{f}_\theta^q, \mathbf{f}_\theta^k, \mathbf{f}_\theta^v\}$, where $\mathbf{f}_\theta^o$ is fed the output of φ (prototype) together with residual connections and optional Layer Normalization [56] (LN), with learnable parameters θ per layer. Formally, let $\hat{\varphi} = \varphi(\mathbf{Q}, \mathbf{K}, \mathbf{V}) + \mathbf{Q}$, then:

$$z = \text{PMSA}(h^q, \acute{\mathbf{H}}, \acute{\mathbf{H}}) = \text{LN}(\hat{\varphi} + \text{ReLU}(\mathbf{f}_\theta^o(\hat{\varphi}))), \quad (7)$$

to produce a bag feature $z \in \mathbb{R}^d$ that is later fed to a bag classifier $\mathbf{f}_\theta^c$ for WSI classification (see Fig. 2). PMSA encourages the model to produce a more discriminative bag embedding by finding more correlations between the critical instance and spatially related features.

1) Learning Objectives: Finally, FR-MIL is trained end-end to minimize the bag-, max-pooling, and feature losses:

$$\mathcal{L} = \gamma_1\mathcal{L}_{bag}(\hat{y}, y) + \gamma_2\mathcal{L}_{max}(A^c, y) + \gamma_3 L_{fm}(\hat{\mathbf{H}}_i^{pos}, \hat{\mathbf{H}}_i^{neg}, \tau), \quad (8)$$

where $\{\gamma_i\}$ are balancing weights, $\mathcal{L}_{\{bag\}}$ is cross-entropy loss, and $\mathcal{L}_{\{max\}}$ is a binary cross-entropy loss over the true WSI labels y given $\hat{y} = \mathbf{f}_\theta^c(z)$, respectively. Algorithm 1 summarizes the overall learning strategy.

E. FR-MIL++: Generative MIL With VQ-VAE

We now formally describe our extended model depicted in Fig. 3(c). Two practical issues stem from the design of FR-MIL; (i) lack of a mechanism to select the optimal critical instance *i.e.*, {*max*,*mean*,*min*} for recalibration, and (ii) potential overfitting on easy critical instances without a mechanism for inferring the best per class, as well as heuristic feature loss margin $\mathcal{L}_{fm}$ requiring hyper-parameter (τ) selection before training. To address this, we introduce a generative process in FR-MIL by leveraging VQ-VAE on critical instances with a decoder $\mathbf{D}$ for reconstruction. This design does not require using the objective L_{fm} and acts as a regularizer by updating

Algorithm 1 FR-MIL

Input: parameters $\mathbf{F}_\theta$, PEM, PMSA epochs T, threshold τ, and dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{N}$
Pre-process and extract patches $x_{ij} \in \{\mathcal{D}\}$
Obtain instance features $\mathbf{H}_i = \mathbf{F}_\theta(X_i)$
Determine threshold τ
for $t = 1, 2, \ldots, T$ **do**
  get critical instance and scores: h^q, A
  re-calibrate features $\mathbf{H}_i$ with Eq. 3 to get $\hat{\mathbf{H}}_i$
  learn spatial context: $\acute{\mathbf{H}} = \text{PEM}(\hat{\mathbf{H}}_i)$ Eq. 8
  pool and get bag feature: $z = \text{PMSA}(h^q, \acute{\mathbf{H}}_i, \hat{\mathbf{H}}_i)$
  bag prediction: $\hat{y} = \mathbf{f}_\theta^c(z)$
  minimize cost in Eq. 8: $\mathcal{L}(\tau, \hat{y}, A, \hat{\mathbf{H}}_i^{pos}, \hat{\mathbf{H}}_i^{neg})$
endfor
output: PEM, PMSA

an embedding table of key instance characteristics across classes and dimensions in the VQ module. Our intuition is generative instance modeling alleviates the need for explicit distance learning and can better capture key instance features for latent representation learning. Overall, FR-MIL++ still employs recalibration in the same fashion as FR-MIL (before PEM), with the key difference being that the critical instance is fed to the VQ module at the same time. The combination aims to enforce learning of better key instance representations.

1) Vector Quantization (VQ)-VAE: Vector Quantization (VQ) is a pivotal dictionary learning algorithm within the domain of neural networks for lossy image compression, as extensively discussed by Theis et al. [57] and Agustsson et al. [58]. This approach is instrumental in the compression of neural network activations before arithmetic operations, enabling more efficient processing. Building upon this fundamental concept, Oord et al. [31] innovatively integrated the principles of VQ into Variational Autoencoders (VAEs). By leveraging VQ, the VAE framework aims to learn discrete latent representations, effectively addressing challenges such as "*posterior collapse*" and high variance. Notably, this integration significantly enhanced the quality of image generation.

The key aspect of this integration lies in representing posterior and prior distributions as categorical distributions, wherein samples are extracted through the process of indexing an embedding table denoted $\mathbf{E}$. This embedding table acts as a discrete latent space, facilitating the acquisition of latent codes that can represent intricate features of the input data while mitigating issues associated with continuous latent spaces.

Inspired by this work, we define an embedding space $\mathbf{E} \in \mathbb{R}^{K \times D}$ otherwise referred to as the VQ-module, where K is the size of latent space (K-way or codebook size) and D is the feature dimension as shown in Fig. 3(c). To enable discrete latent modeling of instance features, initial embedding weights are initialized as $\mathbf{E} \sim \mathcal{N}(-1/K, 1/K)$ following [31]. In addition, $\mathbf{E}$ is used to capture factors related to instances that are causal to the bag label and model robust key features per class in the dataset. In contrast to FR-MIL, this new design introduces the codebook and a decoder $\mathbf{D}$, respectively. More specifically, given instance embeddings $\mathbf{H} = \mathbf{F}_\theta(X_i)$ from a frozen pre-trained model, we first obtain the critical instance h^q per bag, otherwise denoted $\mathbf{H}^q$ to be subsequently used as a reconstruction target. Secondly, instead of feeding all available instance embeddings to the VQ-module, only the critical instance $\mathbf{H}^q$ is passed, and then we employ a nearest neighbor look-up in $\mathbf{E}$ to obtain the posterior distribution of probabilities $q(\hat{\mathbf{H}}^q|X)$ (defined as one-hot) following:

$$q(\hat{\mathbf{H}}^q = k|X) = \begin{cases} 1, & \text{for } k = \text{argmin}_j ||\mathbf{H}^q - \mathbf{E}_j||_2 \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

with the final representation obtained as the nearest element of embedding $\mathbf{E}$ *i.e.*, $\hat{\mathbf{H}}^q = \mathbf{E}_k$ where $k = \text{argmin}_j ||\mathbf{H}^q - \mathbf{E}_j||_2$. Similar to standard generative frameworks, we also feed the obtained latent vector to a decoder $\mathbf{D}(\hat{\mathbf{H}}^q)$ to obtain the reconstructed input (feature).

2) Learning and Optimization: To update the embedding table $\mathbf{E}$, three components are employed: (i) a reconstruction loss to optimize the decoder $\mathcal{L}_{MSE}$, (ii) l_2 loss to move embedding vectors $\mathbf{E}_i$ towards the critical instance vector $\mathbf{H}^q$, and (iii) a commitment loss (l_2) to ensure encoded instance features commit to the embedding representation. Formally,

$$\mathcal{L}_{emb} = ||sg[\hat{\mathbf{H}}^q] - \mathbf{E}||_2^2 + \beta||\hat{\mathbf{H}}^q - sg[\mathbf{E}]||_2^2, \tag{10}$$

where sg is a stop-gradient operation defined as an identity during the forward pass with zero partial derivatives. β is a hyper-parameter on the loss and can vary depending on the scale of the reconstruction error ($\beta = 0.01$ in this work). It is worth noting that embeddings $\mathbf{E}$ receive no gradients from $\mathcal{L}_{MSE}$, instead the first term in $\mathcal{L}_{emb}$ enforces the obtained latent to be similar to $\mathbf{H}^q$, and the second term is the commitment loss for consistent training with other modules. Also, as we have no learnable encoder *i.e.*, instance vectors are from frozen pre-trained model, $\mathcal{L}_{MSE}$ optimizes the decoder via $\mathcal{L}_{MSE} = ||\hat{\mathbf{H}}^q - \mathbf{H}^q||_2^2$. Herein, the overall learning objective follows FR-MIL without $\mathcal{L}_{fm}$;

$$\mathcal{L} = \gamma_1 \mathcal{L}_{bag}(\hat{y}, y) + \gamma_2 \mathcal{L}_{max}(A^c, y) + \gamma_4 \mathcal{L}_{emb}(\hat{\mathbf{H}}^q, \mathbf{H}^q, \beta), \tag{11}$$

where γ_4 balances the contribution of $\mathcal{L}_{emb}$. Note that $\mathcal{L}_{MSE}$ is considered part of the VQ learning objective and is omitted for brevity. Algorithm 2 summarizes the procedure.

IV. Experiments

A. Datasets

To demonstrate the effectiveness of our approach, we conducted experiments on both clinical and computer vision datasets. For clinical evaluation, we employ publicly available datasets CAMELYON16 [36] and TCGA lung cancer, as well as an in-house colon cancer dataset termed COLON-MSI. On the other hand, for vision data - we employ classic MIL benchmark datasets [44], [59], and a point cloud dataset *i.e.*, ModelNet40 [48], [60].

Algorithm 2 FR-MIL++

Input: parameters $\mathbf{F}_\theta$, $\mathbf{E}$, $\mathbf{D}$, PEM, PMSA epochs T, hyper-parameters K, β, and dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{N}$
Pre-process and extract patches $x_{ij} \in \{\mathcal{D}\}$
Obtain instance features $\mathbf{H}_i = \mathbf{F}_\theta(X_i)$
Initialize $\mathbf{E}$ with $K = 2$ ($K \geq Y^n$)
for $t = 1, 2, \ldots, T$ **do**
  get critical instance and scores: $\{h^q, \mathbf{H}^q\}$, A
  re-calibrate $\mathbf{H}_i$ (Eq. 3) & obtain PEM($\hat{\mathbf{H}}_i$) (Eq. 8)
  perform lookup for $\hat{\mathbf{H}}^q$ (Eq. 9) and decode $\mathbf{D}(\hat{\mathbf{H}}^q)$
  pool and get bag feature: $z = \text{PMSA}(h^q, \acute{\mathbf{H}}_i, \acute{\mathbf{H}}_i)$
  bag prediction: $\hat{y} = \mathbf{f}_\theta^c(z)$
  minimize cost in Eq. 11: $\mathcal{L}(\beta, \hat{y}, A, \hat{\mathbf{H}}^q, \mathbf{H}^q)$
endfor
output: $\mathbf{E}$, $\mathbf{D}$, PEM, PMSA

1) CAMELYON16: This is a public dataset proposed for metastasis detection in breast cancer. It consists of 271 training sets and 129 testing sets with diseased and benign slides. Tumor regions are fully labeled with pixel-level annotations. After pre-processing, we obtained a total of 3.2 million patches at $\times$20 magnification, with an average of 8,800 patches per bag, resulting in a maximum of 30,000 patches per bag in the training set. Note that positive bags in this set are highly unbalanced with some slides having small tumor regions *i.e.*, $\leq$ 10% with large normal (benign) regions.

2) TCGA LUAD-LUSC: A public dataset from the National Cancer Institute Data portal for lung sub-type classification *i.e.*, Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC). This set consists of 1054 diagnostic slides with slide-level labels only, and a split of 80-20% was employed in our setup following Li et al. [18]. Pre-processing yields 5.2 million patches at x20 magnification, with an average of 5000 patches per bag. Note that compared to Camelyon16, tumor regions have larger portions per slide.

3) COLON-MSI: This in-house dataset consists 625 colorectal (adenocarcinoma) cancer slides involving microsatellite instable (MSI) molecular phenotypes [61] *i.e.*, a sub-typing task with both microsatellite-stable (MSS) and microsatellite-instable (MSI) slides curated at Asan Medical Center with Institutional Review Board (IRB) approval (No.2019−1192). We randomly split the WSI as follows: 360 training, 92 validation, and 173 testing sets. Note that expert pathologists detected the presence of tumors with Immunohistochemical analysis (IHC) and PCR-based amplification and collectively agreed on the final slide-level label. Moreover, tumor ROIs are not used in this work. After pre-processing, a total of 3.5 million patches at $\times$20 magnification, an average of 6,000 patches/bag were obtained, and a maximum of 8900 patches in the train-set, respectively.

4) MIL Benchmark: This benchmark consists of 5 datasets (*i.e.*, MUSK1, MUSK2, FOX, TIGER, and ELEPHANT) with instance feature vectors - thus, it is not required to learn a feature extractor. Overall, MUSK1 and MUSK2 are for drug effect prediction based on molecule conformations (*i.e.*, not all conformations are useful as only a subset are effective). Thus, each bag is labeled positive if at least one conformation is effective, negative otherwise, and each bag may have multiple conformations of the same molecule. On the other hand, FOX, TIGER, and ELEPHANT are for the classification of images based on image segments. Each bag is a group of segments of an image and is labeled positive if at least one segment contains the animal of interest, and negative otherwise.

5) ModelNet40: This is a point cloud subset of ShapeNet [60] objects. A point cloud is a set of low dimensional 3D vectors (x,y,z-coordinates), and such data is frequently encountered in applications such as robotics, vision, and cosmology, areas in which the application of deep learning is highly desired. In this work, the subset consists of 3D representations of 9,843 training and 2,468 testing instances for 40 object classes. We follow prior work Zaheer et al. [48] and employ point clouds with few particles (*e.g.*, 64, 100 to 1000) after pre-processing a given mesh representation.

B. Implementation Settings

1) Whole Slide Data: In the pre-processing step, we extracted valid patches of 256 $\times$ 256 after tissue detection and discarded patches with $\leq$ 15% tissue entropy. For the instance encoder $\mathbf{F}_\theta$, we employed the SimCLR [49] ResNet18 [9] encoder trained by Li et al. [18] for the CAMELYON16 and TCGA datasets. On the other hand, we used an ImageNet pre-trained ResNet18 on the COLON-MSI dataset. Consequently, each instance feature is represented as $\mathbf{H}_i \in \mathbb{R}^{n\times 512}$. FR-MIL is trained with balanced batch sampling (B = 2), and a learning rate of $1e-4$ with Adam optimizer for 100 epochs with 20% dropout as regularization, and PSMA has heads $k = 8$. Hyper-parameters $\{\gamma_{1,2,3}\} = 0.33$, with $\tau = 8.48$ for CAMELYON16, $\tau = 22.5$ on TCGA LUAD-LUSC, and $\tau = 57.5$ on COLON-MSI, respectively. FR-MIL++ training follows FR-MIL with similar settings and schedules (*i.e.*, batches, epochs & learning rate) except hyper-parameters $\{\gamma_{1,2}\} = 0.5$, $\gamma_4 = 0.25$, and $\beta = 0.01$. The embedding table $\mathbf{E} \in \mathbb{R}^{K\times 512}$ has $K = 2$ based on the number of classes and $\mathbf{D} \in \mathbb{R}^{512}$ is a single fully-connected layer.

2) Vision Datasets: For the classic MIL benchmark dataset, we directly use the provided feature vectors as $\mathbf{H}$ *i.e.*, instance features $\mathbf{H}_i \in \mathbb{R}^{n\times 166}$ for MUSK-1,2 and $\mathbf{H}_i \in \mathbb{R}^{n\times 230}$ for the rest. We trained for 40 epochs with a learning rate of $1e-4$, and performed 10-fold cross-validation per experiment repeated 5 times. Hyper-parameter $\tau = \{11.14, 11.10, 7.3, 10.69, 8.64\}$ for Musk1, Musk2, Fox, Tiger and Elephant, respectively. Also, FR-MIL++ has $\mathbf{D} \in \mathbb{R}^{\{166,230\}}$ with the same hyper-parameters as those used on WSI datasets. On the other hand, we also evaluate FR-MIL on point cloud data with varying numbers of instances. For this setting only, PEM was not employed - as it is designed to recover spatial context in WSI instances. In all cases, FR-MIL was trained for 250 epochs with a learning rate of $1e-3$ and a batch size of 256. Each point (x, y, z) was first projected to 256-d ($\mathbf{H}_i \in \mathbb{R}^{n\times 256}$) before applying our proposed frameworks.

TABLE I
EVALUATION OF THE PROPOSED METHOD ON CAMELYON16 (CM16), COLON-MSI AND TCGA LUAD-LUSC. METRICS ACCURACY (ACC) AND AREA UNDER THE CURVE (AUC) WERE EMPLOYED. CM16 & TCGA EMPLOY SIMCLR LEARNED REPRESENTATIONS, AND IMAGENET PRE-TRAINED FEATURES FOR COLON-MSI. †:*denotes Scores Reported in the Paper Using ResNet50 as* $\mathbf{F}_\theta$ WITH IMAGENET FEATURES. RED/**BOLD** ARE BEST/SECOND RESPECTIVELY

	CM16		**COLON-MSI**		**TCGA LUSC-LUAD**	
Method	ACC	AUC	ACC	AUC	ACC	AUC
Mean-pooling [5]	0.7984	0.7620	0.6240	0.8305	0.8857	0.9369
Max-pooling [5]	0.8295	0.8641	0.7603	0.8590	0.8088	0.9014
AbMIL [17]	0.8450	0.8653	0.7400	0.7790	0.9000	0.9488
MIL-RNN [21]	0.8062	0.8064	0.6300	0.6310	0.8619	0.9107
CLAM-SB [15]	0.8450	0.8940	0.7860	0.8200	0.9327	**0.9876**
DSMIL [18]	0.8682	0.8944	0.7340	0.8110	0.9190	0.9633
TransMIL [19]	0.7910	0.8130	0.6760	0.6170	0.8990	0.9642
TransMIL† [19]	0.8837	**0.9309**	-	-	-	-
AbMIL [17] + Re-calibration	0.8605	0.8888	0.7861	0.7775	0.9231	0.9838
Δ (Gain)	(+1.55%)	(+2.35%)	(+4.61%)	(−0.15%)	(+2.31%)	(+3.50%)
DSMIL [18] + Re-calibration	0.8840	0.8960	0.7920	0.8550	**0.9420**	**0.9880**
Δ (Gain)	(+1.58%)	(+0.16%)	(+5.80%)	(+4.40%)	(+2.30%)	(+2.47%)
FR-MIL ($w/\ \mathcal{L}_{bag}$)	0.8600	0.8990	0.8090	0.8800	0.9230	0.9650
FR-MIL ($w/\ \mathcal{L}_{bag} + \mathcal{L}_{fm}$)	0.8760	0.8990	0.7750	0.8420	0.9183	0.9664
FR-MIL ($w/\ \mathcal{L}_{bag} + \mathcal{L}_{max}$)	0.8840	0.8940	0.7800	0.8310	0.9039	0.9528
FR-MIL ($w/\ \mathcal{L}_{bag} + \mathcal{L}_{max} + \mathcal{L}_{fm}$)	0.8910	0.8950	0.8090	**0.9010**	0.9380	0.9770
FR-MIL++ ($w/\ \mathcal{L}_{bag} + \mathcal{L}_{emb}$)	**0.8992**	**0.9158**	**0.8092**	0.8329	0.9327	0.9466
FR-MIL++ ($w/\ \mathcal{L}_{bag} + \mathcal{L}_{max} + \mathcal{L}_{emb}$)	**0.9147**	0.9094	**0.8150**	**0.8892**	**0.9471**	0.9764

C. Comparison Methods

We compare FR-MIL/FR-MIL++ to traditional MIL methods max- and mean-pooling [5], as well as existing state-of-the-art methods: ABMIL [17], DSMIL [18], CLAM-SB [15], MIL-RNN [21], and TransMIL [19]. All compared methods are trained for 200 epochs on COLON-MSI with similar settings as the other WSI datasets.

On the other hand, for evaluation on classic MIL benchmark data, we compare with state-of-the-art methods, multiple instance networks (MI [5]), AbMIL [17], Graph neural network based MIL [62], and other attention based variants [18], [63]. In the case of point cloud classification, we employ the ideas of FR-MIL in the state-of-the-art Set-Transformer [32] method.

V. RESULTS

A. Whole Slide Image Classification

In Table I, we report performance against prior state-of-the-art methods on two public and an in-house dataset(s). First, compared to the highly related method DSMIL that equally uses max-instance selection with attention pooling, FR-MIL shows (+3%) gains for metrics. Secondly, FR-MIL++ had more significant gains over most compared methods *i.e.*, 91.47(+2%) (ACC) over the best method TransMIL and FR-MIL, respectively. The benefit of our method(s) on this dataset is more pronounced as learning on slides with small tumor regions of interest (ROI) in each positive bag is challenging. While TransMIL reports the best AUC (93.09%), we attribute this to the use of a larger feature extractor; suggesting a dependency on feature size (backbones). Note that our approach(es) used a smaller patch encoder and still report comparable performance via a single PEM module (further discussed in Sec.V-D).

On the publicly available TCGA Lung cancer data, FR-MIL/FR-MIL++ equally improves over prior methods *i.e.*, +2% and +3% gains in accuracy. Compared to Camelyon16, the majority of methods show relatively higher scores due to the presence of more cancerous patches per positive bag *i.e.*, $\geq$ 75%; thus baseline methods mean and max-pooling achieve $\geq$ 90% AUC. From a clinical perspective, lung cancer accounts for more deaths in both women and men than any other cancer. Early prognosis is crucial as most are diagnosed at advanced stages, resulting in lower overall survival rates. While re-calibration and generative latent representation learning of key instances aim to highlight otherwise difficult to detect tumors from benign tissues, the results validate our method(s) applicability on sub-typing tasks.

Finally, we report performance on the in-house colorectal cancer set (COLON-MSI) *i.e.*, a sub-typing task on tumors with more aggressive hypermutations in which the rate of genetic mutations is abnormally high because DNA repair mechanisms are disrupted. Evaluation on this set has clinical relevance as only 11% of patients survive when cancer spreads to other parts of the body. Here, the proposed method(s) show high accuracy across all metrics *i.e.*, +2% (ACC) and +5% (AUC). Note the majority of slides also contain relatively large tumor regions and as such max- and mean-pooling show high AUC but have inconsistent results (ACC). CLAM-SB had the best ACC among the compared methods *i.e.*, 78.6% whereas TransMIL performed poorly on this set, possibly due to over-parameterization, and the morphological similarities between MSS and MSI instances.

B. Effect of Re-Calibration Strategies

In Figure 4, we illustrate different re-calibration strategies based on the observed changes in feature space before and after employing FR-MIL. Note that by default,

TABLE II
EVALUATION OF THE EFFECT OF PEM AND MODEL LEARNING STABILITY ON CAMELYON16 (CM16), COLON-MSI AND TCGA LUAD-LUSC. RED/BLUE INDICATE IMPROVEMENT/REDUCTION, AND (MEAN±STD) ARE REPORTED FOR 5 RUNS, RESPECTIVELY

	CM16		COLON-MSI		TCGA LUSC-LUAD	
Method	ACC	AUC	ACC	AUC	ACC	AUC
FR-MIL (w/o PEM)	0.8062	**0.8989**	**0.8150**	0.8883	0.8942	0.9768
FR-MIL ($w/$ PEM)	**0.8910** (+8.48%)	0.8950 (−0.39%)	0.8090 (−0.60%)	**0.9010** (+1.27%)	**0.9380** (+4.38%)	**0.9770** (+0.02%)
FR-MIL++ (w/o PEM)	0.9070	**0.9105**	0.7977	0.8484	0.9423	**0.9784**
FR-MIL++ ($w/$ PEM)	**0.9147** (+0.77%)	0.9094 (−0.11%)	**0.8150** (+1.73%)	**0.8892** (+4.08%)	**0.9471** (+0.48%)	0.9764 (−0.20%)
FR-MIL	0.8898±0.0076	**0.9027±0.0075**	**0.8034±0.0115**	**0.8714±0.0182**	0.9318±0.0099	**0.9779±0.0008**
FR-MIL (*p-value*)	9.15e-03	8.69e-01	6.16e-01	4.29e-01	2.89e-01	1.11e-01
FR-MIL++	**0.9039±0.0105**	0.9015±0.0069	0.7988±0.0205	0.8598±0.0312	**0.9375±0.0053**	0.9723±0.0048

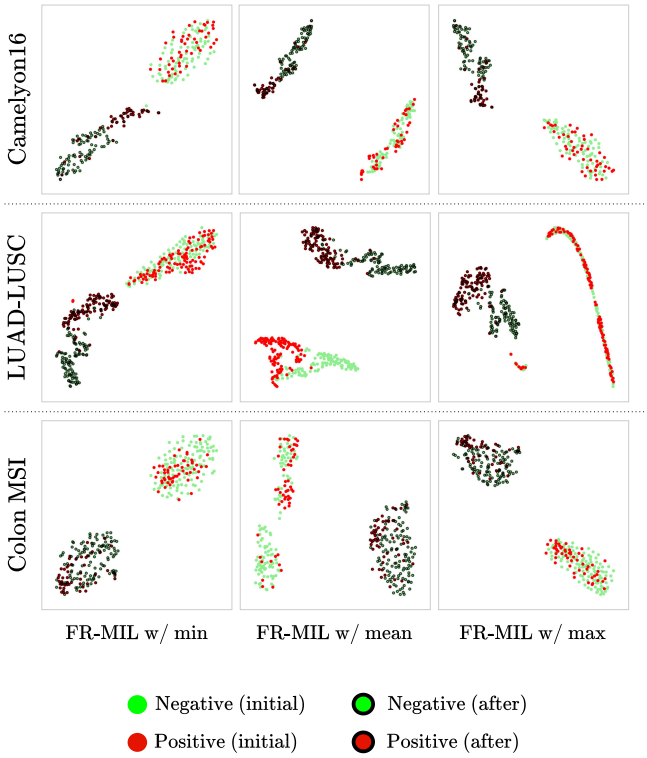


Fig. 4. TSNE feature plot to show the effect of re-calibration on whole slide datasets *i.e.*, CM16, TCGA Lung, and Colon MSI, respectively. For each dataset, we plot *before* and *after* re-calibration (*i.e.*, min, max, or mean) using FR-MIL. We observe significantly improved feature separability on all datasets. (Best viewed in color while zoomed in).

FR-MIL employs the max instance per bag as the critical feature for re-calibration; as such, we explore the impact of using either the mean or max can influence the distribution of features *i.e.*, *min* (minimum scoring instance in a bag) and the mean of all instances. We present a plot of bag-level features before and after re-calibration with the aforementioned strategies. Here, FR-MIL results in significantly separable and distinct clusters per class on each dataset. In some cases *e.g.*, on TCGA data, using the mean of all bag features revealed better separability since the majority of patches in positive bags are tumors, whereas on CM16 using max instance feature results produced better clusters.

In addition, to show that re-calibration can boost prior method(s) performance, we report the performance gains of DSMIL [18] and AbMIL [17] using our strategy in Table I (middle row) *i.e.*, DSMIL/AbMIL + Re-calibration. Specifically, as both FR-MIL and DSMIL include a critical instance selection step, we performed re-calibration before self-attention and bag-level feature pooling in DSMIL, and after computing attention scores (employed for critical instance selection) in AbMIL, incurring no overhead and facilitating improved learning in the baselines. While our extension is simple, including re-calibration in TransMIL or CLAM would require introducing extra modules. Overall, DSMIL +*re-calib* improves scores on all datasets, especially on TCGA where using our strategy reports the best results against comparison methods. The most significant gains were noted on COLON MSI (+4%), though still marginally lower than our proposed method.

C. Effect of Learning Objectives

To validate the effectiveness of the proposed objectives during learning, we present ablations on loss function permutations in Table I. FR-MIL using a bag-level loss alone reports comparable performance to the baselines, whereas the addition of either the max-pooling binary cross entropy loss or feature magnitude losses reveals more improvements *i.e.*, +2% on Camelyon16 and +1% on TCGA. Moreover, the omission of the max-pooling loss in FR-MIL *i.e.*, without re-calibration, reports lower performance across all datasets and equally serves to verify the benefit of our re-calibration strategy since instance selection is crucial to improve learning. Note that on COLON-MSI, performance drops when $\mathcal{L}_{fm}$ was not used, and the best scores were achieved with a combination of all losses. The use of L_{fm} encourages FR-MIL to reduce the uncertainty in bags by modeling positive instances as out-of-distribution via the distance measure of feature magnitudes *i.e.*, the learning ambiguity mainly lies in the positive bags, whereas normal bags generally have similar features. This enables better separation between classes in the MIL setting.

On the other hand, FR-MIL++ with only a reconstruction-based objective ($\mathcal{L}_{emb}$) shows marginal gains vs. FR-MIL, achieving the best scores overall. More importantly, when all losses were employed; the benefit of both the distance-based objective and generative VQ learning was more pronounced.

D. Effect of Positional Encoding Module (PEM) and Model Stability

In Table II, we present results on the benefit of using the PEM module including model learning stability with

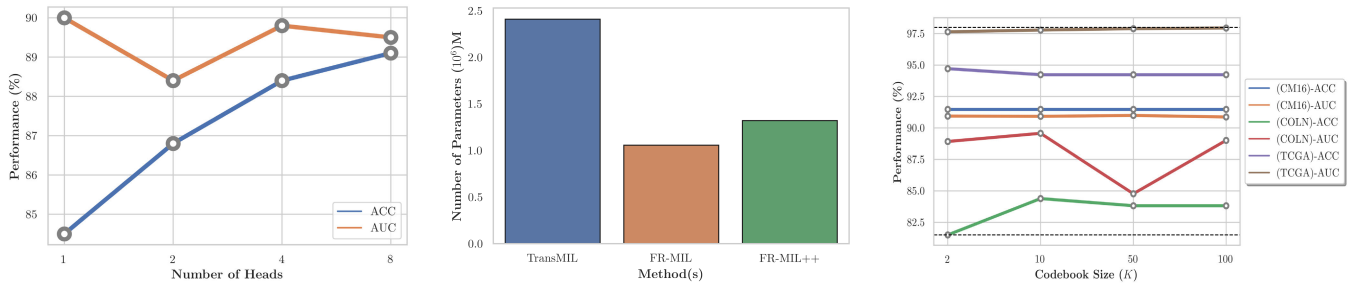


Fig. 5. Ablations on (a) varying the number of heads in PMSA on Camelyon16 dataset, (b) comparison on model (parameter) size between TransMIL and the proposed method(s), and (c) impact of varying the dimensions in FR-MIL++ VQ embedding space.

different weight initializations as a form of cross-validation, and statistical paired t-tests, respectively. Overall, in all settings, we found our method performs better with positional encoding, though on COLON-MSI, not using the module reports marginal improvements using FR-MIL. In addition, FR-MIL also shows consistent improvements despite including an additional generative component. On the CM16 dataset, using PEM reports more gains compared to the baseline methods. This serves to suggest that encoding spatial arrangements is more beneficial on the challenging CM16 dataset with small tumor ROIs. Moreover, our method is on par with TransMIL (Table I) despite using a single PEM module compared to several stacked layers.

Despite using a limited sample size WSI dataset, PEM enabled faster convergence in our framework by learning implicit correlations among instances. Compared to natural images of fixed size, WSIs have varying sizes and the model can observe a large pool of data even from few samples at the slide level *i.e.*, the average bag size on Camelyon16 was ∼9000. The results highlight that improvements do not solely stem from the use of this module, but rather the combination of others. In fact, Shao et al. [19] discuss the benefit of conditional position encoding over existing absolute encoding for square images and provided empirical evidence to support the claims *i.e.*, while extracting instances (during pre-processing) in standard ordering (top–bottom, left-right) is generally better than using random ordering, PEM can still learn meaningful correlations on unordered instances.

Furthermore, we report on model learning stability with different weight initializations (seeds) in the lower section of Table II. Notably, FR-MIL had more consistent performance across different seeds compared to FR-MIL++, though the differences were marginal as both faired well. The performance of FR-MIL++ may be attributed to the use of additional modules that learn instance latent features with the VQ-VAE model. When compared to prior methods, our approach still reports the best results on all datasets, validating stability. In addition, we also include statistical scores based on the p-values obtained by paired t-tests of our methods. We confirmed that performance differences on FR-MIL were significant on the Camelyon16 dataset (p-values $\leq$ 0.05), though relatively larger on the other datasets. This may be attributed to architectural design choices. Also, the results verify that our strategy is more beneficial on CM16.

E. Impact of Transformer Heads

We explore the impact of using several heads in the Pooling by Multi-head Self-Attention (PMSA) module (Figure 5(a)) and compare the number of parameters incurred in Figure 5(b). In general, prior works employ several heads to learn robust representations for classification ($\geq$ 4), with more computation required as heads increase. Here, we show that the proposed method is not the strictly sensitive number of heads *i.e.*, a reasonable trade-off was achieved with 8 heads. On the other hand, our approach also incurs fewer parameters than TransMIL *i.e.*, (−50%), a related method that employs PEM with several blocks compared to ours.

F. VQ-VAE Embedding Size Sensitivity

To further validate whether FR-MIL++ with VQ-VAE identifies key latent instance representations per slide, we evaluate our method by varying the dimensionality of the discrete space K in the VQ embedding table as shown in Figure 5(c). We believe it is reasonable to set $K = ||\mathcal{Y}||$ *i.e.*, a total number of classes (malignant/benign tissue). In contrast, by increasing the space, the model may capture other factors related to the instances that are causal to the bag label *i.e.*, slides may vary in staining, acquisition protocols, and come from different institutions. It is interesting to observe that using a larger space had minimal impact on Camelyon16 and TCGA datasets. However, $K \geq 2$ performs significantly better on the Colon dataset. This can be attributed to the high morphological similarities between subtypes. Therefore, using more dimensions could capture key characteristics for improved learning.

G. Qualitative Results

In Figure 6, we present a normalized heatmap of instance scores using the max-instance module to qualitatively illustrate localization performance. Results show that scores per instance correspond to actual tumor ROI (Fig.6(b)). Note that for some challenging cases (rows 1 & 2), the model fails to highlight all relevant regions but still finds key patches. On WSIs with fairly larger ROIs - our approach had better segmentation *i.e.*, Fig.6 (rows 3 & 4).

In addition, we report localization performance based on Free Receiver Operating Characteristic (FROC) [36] in Table III *i.e.*, average sensitivity at 6 predefined false positive rates (FPR): {0.25, 0.5, 1, 2, 4, 8} per WSI. The results

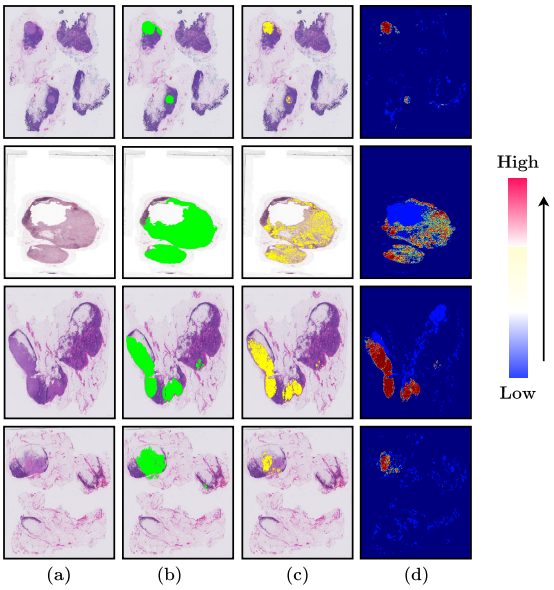


Fig. 6. Illustration of FR-MIL's critical (top) instance scores on Camelyon16. Left to right: (a) Input, (b) ground-truth (green), (c) prediction overlay (yellow), and (d) probability map, respectively. The color bar is relative to the heatmap in (d).

TABLE III
LOCALIZATION PERFORMANCE COMPARISON ON CAMELYON16 DATASET. RESULTS TAKEN FROM [18]. RED/**BOLD** ARE BEST/SECOND RESPECTIVELY

Method	Localization FROC
Fully-supervised	0.5254
Mean-pooling [5]	0.1162
Max-pooling [5]	0.3313
MIL-RNN [21]	0.3048
AbMIL [17]	0.4056
DSMIL [18]	**0.4296**
DSMIL-LC (*multi-scale*) [18]	**0.4371**
FR-MIL	0.4110
FR-MIL++	0.4201

show that our method(s) are comparable to prior works with marginal gains (vs. AbMIL), including multi-scale DSMIL-LC. While FR-MIL++ had better average scores than FR-MIL, we observed a slightly higher FPR due to using more modules.

H. FR-MIL for Vision Applications

We report the performance of FR-MIL for different computer vision tasks *i.e.*, on the classic MIL benchmark dataset and point cloud classification. While our method shows significant gains in medical data, we hypothesize it can be more general. To validate this, we report results on the MIL benchmark dataset in Table IV and include existing WSI-specific methods AbMIL [17] and DSMIL [18]. Note that PEM was not used for this setting as recovering spatial context is not applicable. Overall, the proposed method achieves state-of-the-art performance across all evaluated settings, and different ablations on the proposed losses further validated the benefit of our strategy.

In Table V, we present results on point cloud classification. For this scenario, we employ our proposed techniques as an extension of the state-of-the-art method by Lee et al. [32] to predict an object class (40 classes) *i.e.*, +FR-MIL. Note that, unlike WSI data, point clouds are homogeneous for any given set - whereas in WSI both normal and other tissue exist per slide. Thus, re-calibration was modified based on this observation as follows: $\hat{\mathbf{H}} = \frac{\mathbf{H}-h^q}{\sigma(\mathbf{H})}$, where $h^q \in \{\max, \text{mean}, \min\}$ and σ is the standard deviation of all instances in a set (point cloud).

Intuitively, the latter corresponds to Gaussian-style normalization to maintain homogeneity in a set. Interestingly, FR-MIL ($w/$ min) shows the best performance over the baseline (+1.3%) with only 64 points sampled per object with a linear trend of improvement when more points are added. This is counter to the phenomena observed in medical data where using max was better. Moreover, Set-Transformer + FR-MIL was trained with the proposed losses *i.e.*, $\mathcal{L}_{fm}$ with $\mathcal{L}_{max}$ replaced with cross-entropy, validating its potential for the multi-class setting.

VI. DISCUSSION

While existing MIL methods report good performance on several tasks, a general trend among prior art is the focus on (i) improved bag pooling strategies with Transformers and self-attention mechanisms, and (ii) leveraging multi-scale information for improved context-aware learning. In this work, our empirical results provide strong evidence to support the idea that paying attention to the data distribution is beneficial, especially on multi-site or otherwise highly similar pathology data. Though simple, our proposed strategy can improve the performance of prior methods as highlighted in Sec.V-A and Table I. Furthermore, explicit distance-based learning can be replaced with a generative MIL component using VQ-VAE optimization to further boost performance.

In the context of this study, it is worth highlighting some limiting factors of our methodology. First, while bag re-calibration with *max* works well on slides with distinct positive instances (*e.g.*, CM16 dataset), leveraging the *mean* strategy is better suited to sub-typing tasks, whereas *min* is validated to be better in the point cloud classification task. This implies selecting the strategy depends on the data employed with prior assumptions. Ideally, the model should automatically infer the best strategy, though this may require multi-stage training. Secondly, spatial positional encoding (PEM) is key in WSI tasks but omitted in others, a generalization of PEM to non-spatially data arranged is more desirable *e.g.*, via graph-based representation learning or using compressed WSI tensor representations for faster processing. In regards to the metric loss, expanding on different objectives to enforce separation would be more beneficial *e.g.*, contrastive losses. However, note that this requires sampling multiple augmented views/negatives, which is challenging for WSIs, and better suited to instance-level learning. In contrast, our approach is much simpler as it only requires sampling two bags per training cycle, without augmentation. For this work, we instead juxtapose the proposed metric loss with a generative approach that does not require sampling.

Finally, leveraging a generative model in MIL has connections to causal reasoning [45] and interventions [27], [65], [66]. Further exploring and quantifying the true hidden

TABLE IV

PERFORMANCE COMPARISON ON CLASSICAL MIL DATASET. EXPERIMENTS WERE RUN 5 TIMES EACH WITH A 10-FOLD CROSS-VALIDATION. THE MEAN AND STANDARD DEVIATION OF THE CLASSIFICATION ACCURACY ARE REPORTED (MEAN ± STD). RED/**BOLD** ARE BEST/SECOND RESPECTIVELY

Methods	MUSK1	MUSK2	FOX	TIGER	ELEPHANT
mi-Graph [5]	0.889±0.033	0.903±0.039	0.620±0.044	0.860±0.037	0.869±0.035
MI-Net [5]	0.887±0.041	0.859±0.046	0.622±0.038	0.830±0.032	0.862±0.034
MI-Net $w/$ DS [5]	0.894±0.042	0.874±0.043	0.630±0.037	0.845±0.039	0.872±0.032
AbMIL [17]	0.892±0.040	0.858±0.048	0.615±0.043	0.839±0.022	0.868±0.022
AbMIL-Gated [17]	0.900±0.050	0.863±0.042	0.603±0.029	0.845±0.018	0.857±0.027
GNN-MIL [62]	0.917±0.048	0.892±0.011	0.679±0.007	0.876±0.015	0.903±0.010
NL-MIL [63]	0.921±0.017	0.910±0.009	0.703±0.035	0.857±0.013	0.876±0.011
DP-MINN [64]	0.907±0.036	0.926±0.043	0.655±0.052	**0.897±0.028**	0.894±0.030
DSMIL [18]	0.932±0.023	0.930±0.020	0.729±0.018	0.869±0.008	0.925±0.007
FR-MIL ($w/\ \mathcal{L}_{bag}$)	0.960±0.013	0.922±0.013	**0.764±0.019**	0.863±0.014	0.915±0.008
FR-MIL ($w/\ \mathcal{L}_{bag} + \mathcal{L}_{fm}$)	0.960±0.013	0.922±0.013	**0.763±0.019**	0.866±0.011	0.915±0.008
FR-MIL ($w/\ \mathcal{L}_{bag} + \mathcal{L}_{max}$)	**0.973±0.015**	0.960±0.016	0.760±0.017	**0.904±0.004**	**0.944±0.005**
FR-MIL ($w/\ \mathcal{L}_{bag} + \mathcal{L}_{max} + \mathcal{L}_{fm}$)	0.964±0.018	0.950±0.016	0.760±0.005	0.901±0.009	**0.938±0.006**
FR-MIL++ ($w/\ \mathcal{L}_{bag} + \mathcal{L}_{emb}$)	**0.978±0.011**	**0.964±0.011**	0.757±0.011	0.887±0.008	0.934±0.008
FR-MIL++ ($w/\ \mathcal{L}_{bag} + \mathcal{L}_{max} + \mathcal{L}_{emb}$)	0.956±0.011	**0.956±0.011**	0.744±0.019	0.885±0.006	0.936±0.004

TABLE V

TEST PERFORMANCE OF DIFFERENT METHODS FOR POINT CLOUD CLASSIFICATION ON MODELNET40 [48], [60] DATASET WITH VARYING *N* (NUMBER OF POINTS). EXPERIMENTS WERE RUN 5 TIMES EACH WITH THE MEAN AND STANDARD DEVIATION OF THE CLASSIFICATION ACCURACY REPORTED (MEAN ± STD)

Method / Number of Points	64	100	256	500	1000
Set-Transformer [32]	0.813±0.005	0.825±0.006	0.856±0.002	0.866±0.005	0.873±0.003
+ FR-MIL w/ mean	0.814±0.004 (+0.1%)	0.832±0.001 (+0.7%)	0.861±0.002 (+0.5%)	0.871±0.002 (+0.5%)	0.879±0.003 (+0.6%)
+ FR-MIL w/ max	0.818±0.003 (+0.5%)	0.836±0.002 (+1.1%)	0.862±0.002 (+0.6%)	0.875±0.002 (+0.9%)	0.879±0.004 (+0.6%)
+ FR-MIL w/ min	**0.826±0.003** (+1.3%)	**0.839±0.005** (+1.4%)	**0.864±0.003** (+0.8%)	**0.875±0.003** (+0.9%)	**0.881±0.002** (+0.8%)

latent factors captured by the model can aid clinical workflows *i.e.*, unexpected spurious features (ROI) may influence the model predictions as instances are highly correlated. Qualitatively validating which instances are causal to the bag labels is an interesting future research topic.

VII. CONCLUSION

In this work, we introduced a multiple instance learning framework for Whole Slide Image classification that efficiently leverages feature/distribution statistics per input for improved performance. Our method applies not only to binary and sub-typing tasks but also to multi-class problems in both medical and computer vision datasets. Based on an extensive evaluation of our approach on several WSI datasets, our method reveals that enforcing feature discrepancy and explicitly reducing the influence of isolated features can significantly improve performance. Finally, we show that an extended strategy based on generative instance modeling can alleviate the need for explicit distance-based learning and capture key instance features using latent representation learning. Aside from the performance gains, we believe FR-MIL could further benefit from learning a self-supervised policy to select the re-calibration strategy automatically as a follow-up work.

REFERENCES

[1] N. Dimitriou, O. Arandjelović, and P. D. Caie, "Deep learning for whole slide image analysis: An overview," *Frontiers Med.*, vol. 6, p. 264, Nov. 2019.

[2] L. He, L. R. Long, S. Antani, and G. R. Thoma, "Histology image analysis for carcinoma detection and grading," *Comput. Methods Programs Biomed.*, vol. 107, no. 3, pp. 538–556, Sep. 2012.

[3] S. Banerji and S. Mitra, "Deep learning in histopathology: A review," *WIREs Data Mining Knowl. Discovery*, vol. 12, no. 1, Jan. 2022, Art. no. e1439.

[4] C. Li et al., "A comprehensive review of computer-aided whole-slide image analysis: From datasets to feature extraction, segmentation, classification, and detection approaches," 2021, *arXiv:2102.10553*.

[5] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," *Pattern Recognit.*, vol. 74, pp. 15–24, Feb. 2018.

[6] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artif. Intell.*, vol. 201, pp. 81–105, Aug. 2013.

[7] C. L. Srinidhi, O. Ciga, and A. L. Martel, "Deep neural network models for computational histopathology: A survey," *Med. Image Anal.*, vol. 67, Jan. 2021, Art. no. 101813.

[8] A. Vaswani et al., "Attention is all you need," in *Proc. NeurIPS*, vol. 30, 2017, pp. 1–24.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[10] H. Chen et al., "From pixel to whole slide: Automatic detection of microvascular invasion in hepatocellular carcinoma on histopathological image via cascaded networks," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2021, pp. 196–205.

[11] L. Fan, A. Sowmya, E. Meijering, and Y. Song, "Learning visual features by colorization for slide-consistent survival prediction from whole slide images," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2021, pp. 592–601.

[12] Y. Sharma, A. Shrivastava, L. Ehsan, C. A. Moskaluk, S. Syed, and D. Brown, "Cluster-to-conquer: A framework for end-to-end multi-instance learning for whole slide image classification," in *Medical Imaging With Deep Learning*. Breckenridge, CO, USA: PMLR, 2021, pp. 682–698.

[13] D. Rymarczyk, A. Borowa, J. Tabor, and B. Zielinski, "Kernel self-attention for weakly-supervised image classification using deep multiple instance learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1720–1729.

[14] P. Chikontwe, M. Kim, S. J. Nam, H. Go, and S. H. Park, "Multiple instance learning with center embeddings for histopathology classification," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2020, pp. 519–528.

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

[16] X. Shi, F. Xing, Y. Xie, Z. Zhang, L. Cui, and L. Yang, "Loss-based attention for deep multiple instance learning," in *Proc. AAAI*, 2020, vol. 34, no. 4, pp. 5742–5749.

[17] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2127–2136.

[18] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14318–14328.

[19] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, and X. Ji, "Transmil: Transformer based correlated multiple instance learning for whole slide image classification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 1–20.

[20] H. Li et al., "DT-MIL: Deformable transformer for multi-instance learning on histopathological image," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2021, pp. 206–216.

[21] G. Campanella et al., "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Med.*, vol. 25, no. 8, pp. 1301–1309, Aug. 2019.

[22] A. Dosovitskiy, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021, pp. 1–26.

[23] P. Lee, J. Wang, Y. Lu, and H. Byun, "Weakly-supervised temporal action localization by uncertainty modeling," in *Proc. AAAI*, vol. 2, 2021, pp. 1–29.

[24] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," 2019, *arXiv:1901.03407*.

[25] J.-C. Feng, F.-T. Hong, and W.-S. Zheng, "MIST: Multiple instance self-training framework for video anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14009–14018.

[26] S. Yang, L. Liu, and M. Xu, "Free lunch for few-shot learning: Distribution calibration," in *Proc. ICLR*, 2020, pp. 1–18.

[27] W. Zhang et al., "Multi-instance causal representation learning for instance label prediction and out-of-distribution generalization," in *Proc. NeurIPS*, vol. 35, 2022, pp. 34940–34953.

[28] S. Ghaffarzadegan, "Deep multiple instance feature learning via variational autoencoder," in *Proc. 32nd AAAI Conf. Artif. Intell.*, Jun. 2018, pp. 1–24.

[29] W. Zhang, "Non-IID multi-instance learning for predicting instance and bag labels using variational auto-encoder," 2021, *arXiv:2105.01276*.

[30] Y. Chen, J. Liu, L. Peng, Y. Wu, Y. Xu, and Z. Zhang, "Auto-encoding variational Bayes," *Cambridge Explorations Arts Sci.*, vol. 2, no. 1, pp. 1–29, Feb. 2024.

[31] A. Van Den Oord and O. Vinyals, "Neural discrete representation learning," in *Proc. NeurIPS*, vol. 30, 2017, pp. 1–11.

[32] J. Lee, Y. Lee, J. Kim, A. R. Kosiorek, S. Choi, and Y. W. Teho, "Set Transformer: A framework for attention-based permutation-invariant neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 3744–3753.

[33] P. Chikontwe, S. J. Nam, H. Go, M. Kim, H. J. Sung, and S. H. Park, "Feature re-calibration based multiple instance learning for whole slide image classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist.*, 2022, pp. 420–430.

[34] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognit.*, vol. 77, pp. 329–353, May 2018.

[35] P. Chikontwe et al., "Weakly supervised segmentation on neural compressed histopathology with self-equivariant regularization," *Med. Image Anal.*, vol. 80, Aug. 2022, Art. no. 102482.

[36] B. E. Bejnordi 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.

[37] J. Yang et al., "ReMix: A general and efficient framework for multiple instance learning based whole slide image classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist.*, 2022, pp. 35–45.

[38] Z. Wang, L. Yu, X. Ding, X. Liao, and L. Wang, "Lymph node metastasis prediction from whole slide images with transformer-guided multiinstance learning and knowledge transfer," *IEEE Trans. Med. Imag.*, vol. 41, no. 10, pp. 2777–2787, Oct. 2022.

[39] H. Zhang et al., "DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18802–18812.

[40] P. Chikontwe, M. Luna, M. Kang, K. S. Hong, J. H. Ahn, and S. H. Park, "Dual attention multiple instance learning with unsupervised complementary loss for COVID-19 screening," *Med. Image Anal.*, vol. 72, Aug. 2021, Art. no. 102105.

[41] R. G. Cinbis, J. Verbeek, and C. Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 189–203, Jan. 2017.

[42] N. Gonthier, S. Ladjal, and Y. Gousseau, "Multiple instance learning on deep features for weakly supervised object detection with extreme domain shifts," *Comput. Vis. Image Understand.*, vol. 214, Jan. 2022, Art. no. 103299.

[43] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Proc. NeurIPS*, vol. 10, 1997, pp. 1–24.

[44] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Proc. NeurIPS*, vol. 15, 2002, pp. 1–29.

[45] P. Sanchez, J. P. Voisey, T. Xia, H. I. Watson, A. Q. O'Neil, and S. A. Tsaftaris, "Causal machine learning for healthcare and precision medicine," *Roy. Soc. Open Sci.*, vol. 9, no. 8, Aug. 2022, Art. no. 220638.

[46] J. Brehmer, P. De Haan, P. Lippe, and T. S. Cohen, "Weakly supervised causal representation learning," in *Proc. NeurIPS*, vol. 35, 2022, pp. 38319–38331.

[47] B. Schölkopf et al., "Toward causal representation learning," *Proc. IEEE*, vol. 109, no. 5, pp. 612–634, May 2021.

[48] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola, "Deep sets," in *Proc. NeurIPS*, vol. 30, 2017, pp. 1–18.

[49] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn.*, vol. 119, 2020, pp. 1597–1607.

[50] J.-B. Grill et al., "Bootstrap your own latent—A new approach to self-supervised learning," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 21271–21284.

[51] R. J. Chen et al., "Scaling vision transformers to gigapixel images via hierarchical self-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16144–16155.

[52] Q. Zhang and Z. Chen, "Weakly supervised segmentation by tensor graph learning for whole slide images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, 2022, pp. 253–262.

[53] M. Aryal and N. Yahya Soltani, "Context-aware self-supervised learning of whole slide images," *IEEE Trans. Artif. Intell.*, vol. 5, no. 8, pp. 4111–4120, Aug. 2024.

[54] T. H. Chan, F. J. Cendra, L. Ma, G. Yin, and L. Yu, "Histopathology whole slide image analysis with heterogeneous graph representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15661–15670.

[55] Y. Zheng et al., "A graph-transformer for whole slide image classification," *IEEE Trans. Med. Imag.*, vol. 41, no. 11, pp. 3003–3015, Nov. 2022.

[56] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[57] L. Theis, W. Shi, A. Cunningham, and F. Huszár, "Lossy image compression with compressive autoencoders," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–22.

[58] E. Agustsson et al., "Soft-to-hard vector quantization for end-to-end learned compression of images and neural networks," 2017, *arXiv:1704.00648*.

[59] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, nos. 1–2, pp. 31–71, Jan. 1997.

[60] A. X. Chang et al., "Shapenet: An information-rich 3D model repository," 2015, *arXiv:1512.03012*.

[61] C. R. Boland and A. Goel, "Microsatellite instability in colorectal cancer," *Gastroenterology*, vol. 138, no. 6, pp. 2073–2087, 2010.

[62] M. Tu, J. Huang, X. He, and B. Zhou, "Multiple instance learning with graph neural networks," 2019, *arXiv:1906.04881*.

[63] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[64] Y. Yan, X. Wang, X. Guo, J. Fang, W. Liu, and J. Huang, "Deep multi-instance learning with dynamic pooling," in *Proc. Asian Conf. Mach. Learn.*, 2018, pp. 662–677.

[65] T. Lin, H. Xu, C. Yang, and Y. Xu, "Interventional multi-instance learning with deconfounded instance-level prediction," in *Proc. AAAI*, 2022, pp. 1601–1609.

[66] T. Lin, Z. Yu, H. Hu, Y. Xu, and C. W. Chen, "Interventional bag multi-instance learning on whole-slide pathological images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19830–19839.