Cross-Attention-Based Saliency Inference for Predicting Cancer Metastasis on Whole Slide Images

Ziyu Su, Mostafa Rezapour, Usama Sajjad, Shuo Niu, Metin Nafi Gurcan, and Muhammad Khalid Khan Niazi

***Abstract*—Although multiple instance learning (MIL) methods are widely used for automatic tumor detection on whole slide images (WSI), they suffer from the extreme class imbalance WSIs containing small tumors where the tumor may include only a few isolated cells. For early detection, it is important that MIL algorithms can identify small tumors. Existing studies have attempted to address this issue using attention-based architectures and instance selection-based methodologies but have not produced significant improvements. This paper proposes cross-attention-based salient instance inference MIL (CASi-iMIL), which involves a novel saliency-informed attention mechanism to identify small tumors (e.g., breast cancer lymph node micro-metastasis) on WSIs without needing any annotations. In addition to this new attention mechanism, we introduce a negative representation learning algorithm to facilitate the learning of saliency-informed attention weights for improved sensitivity on tumor WSIs. The proposed model outperforms the state-of-the-art MIL methods on two popular tumor metastasis detection datasets. The proposed approach demonstrates great cross-center generalizability, high accuracy in classifying WSIs with small tumor lesions, and excellent interpretability attributed to the saliency-informed attention weights. We expect that the proposed method will pave the way for training algorithms for early tumor detection on large datasets where acquiring fine-grained annotations is is not practical.**

***Index Terms*—Multiple instance learning, attention mechanism, whole slide images, digital pathology, breast cancer metastasis.**

Received 27 September 2023; revised 10 July 2024; accepted 30 July 2024. Date of publication 6 August 2024; date of current version 6 December 2024. This work was supported in part by National Cancer Institute Gurcan under Grant R21 CA273665, in part by National Cancer Institute Niazi, Chen under Grant R01 CA276301, in part by the National Institute of Biomedical Imaging and Bioengineering Niazi, Segal under Grant R21 EB029493, and in part by the Alliance Clinical Trials in Oncology Frankel, Niazi under Grant GR125886. *(Corresponding author: Ziyu Su.)*

Ziyu Su, Mostafa Rezapour, Usama Sajjad, Metin Nafi Gurcan, and Muhammad Khalid Khan Niazi are with the Center for Artificial Intelligence Research, Wake Forest University School of Medicine, Winston-Salem, NC 27101 USA (e-mail: zsu@wakehealth.edu; mrezapou@wakehealth.edu; usajjad@wakehealth.edu; mgurcan@wakehealth.edu; mniazi@wakehealth.edu).

Shuo Niu is with the Department of Pathology, Wake Forest University School of Medicine, Winston-Salem, NC 27101 USA (e-mail: sniu@wakehealth.edu).

Digital Object Identifier 10.1109/JBHI.2024.3439499

I. Introduction

DIGITAL pathology is playing an increasingly important role in cancer diagnosis and transforming how pathologists provide diagnostic information to patients and clinicians [1]. By digitizing cancer specimens as whole slide images (WSIs) with high resolution, pathologists can now view, share, and analyze them more easily and efficiently. One of the benefits of digital pathology is that it enables pathologists to seek a second opinion from other experts more quickly by sharing images. Moreover, it provides opportunities to develop machine learning-based computer-aided diagnosis technologies [1], [2].

In recent years, deep learning has become the preferred machine learning method for analyzing WSIs due to its remarkable learning capabilities [3]. However, two major challenges exist when using deep learning models to analyze WSIs. Firstly, WSIs are often extremely large (giga-pixel size) and stored in a multi-resolution format to imitate the light microscope, making them even larger. Secondly, accurate ground truth labels describing and annotating regions of interest (e.g., lesions) are often scarce. Pathologists usually prefer to provide overall diagnostic labels (e.g., cancer or normal) rather than to annotate lesions on WSIs or to draw their boundaries.

Unfortunately, MIL methods employed in detecting breast cancer micro-metastasis to the lymph nodes have not yielded satisfactory results. Identifying breast cancer metastasis to lymph nodes holds significant importance as it assists oncologists in determining the stage of breast cancer and devising treatment plans [4]. The presence of tumor cells in the lymph nodes indicates the spread of cancer beyond the breast tissue and can imply a higher stage of the disease. The size of the tumor involvement in a lymph node and the number of lymph nodes that are involved by the tumor are both essential in cancer staging [5], which directly affects the treatment plan and the disease prognosis. Moreover, *an early diagnosis* of lymph node metastasis in breast cancer is vital for improving treatment outcomes and overall prognosis [6], [7]. Hence, we aim to develop a reliable and accurate diagnostic tool for detecting tumors, including challenging-to-identify *micro-metastases*.

When dealing with WSIs with small tumor lesions, such as micro-metastasis to the lymph nodes, the tumor size may comprise only a few isolated cells. For this reason, an MIL bag may contain a large number of normal instances and only a few tumor instances. This leads to a severe class imbalance during

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

model development [8]. Some previous studies have tackled this issue by performing instance-level classification and predicting WSIs with a few high-confidence instances in the bag [2], [9]. Others have used attention-based MIL models to enable the model to focus on potential tumor instances [8], [10], [11], [12], [13]. However, in the case of early diagnosis, the tumor lesions typically occupy a very small portion (usually less than 5%) of the WSIs, resulting in only a few tumor instances in the corresponding tumor bags. Meanwhile, the current attention mechanism tries to learn proper attention weights solely based on each instance without guidance from tumor-related information. Therefore, the aforementioned MIL algorithms fail to pay attention to or classify the positive instances, leading to unsatisfactory sensitivity in predicting WSIs with small tumor lesions [14]. To address this problem, several MIL studies have explored the selection of highly salient instances within MIL bags. For instance, in our previous work [14], we achieved high accuracy by pre-training an instance-level tumor detection model and using the salient instances (i.e., possible tumor instances with high-confidence) in the bags for MIL prediction. However, this approach assumes that there are some large-tumor WSIs in the dataset. Other studies combined their tumor instance detection model with gradient flow to feed salient instances to the MIL model in a learnable manner [2], [9], but detecting tumor instances solely based on slide-level label is a radical approach. As a result, these methods either achieved moderate sensitivity or required large-scale training sets.

To address the above issues, we propose a novel MIL methodology, named *Cross-Attention-based Salient instance inference-MIL (CASiiMIL)*, that can infer possible tumor instances and mitigate the class imbalance between normal and tumor instances in an end-to-end neural network. Inspired by Transformer [15] and open-set learning methods [16], [17], [18], we propose CASii network that can automatically correlate the input instances with the representative normal instances in a more discriminative feature space (for tumor identification) and infer the salient instances dynamically by learning saliency-informed attention weights to highlight them. The contributions of this work are as follows:

- To mitigate the class-imbalance issue of small tumor WSIs encountered by existing MIL models, we propose a novel attention mechanism, named cross-attention-based salient instance inference (CASii), to learn saliency-informed attention weights for improved tumor WSI identification performance.
- We present a negative representation learning method that can learn representative normal instances to support salient instance inference.
- We introduce two instance-level loss functions to further improve the sparsity and saliency of the learned attention weights for our MIL model.

II. Related Works

A. MIL for WSI Analysis

MIL is one of the most extensively used deep learning methods for WSI analysis given its weakly supervised property. It assigns labels to groups of instances, known as "bags," instead of individual instances [19]. MIL is commonly used to determine whether a bag contain at least one positive instance, making it perfect fit for analyzing WSIs.

In 2018, Courtiol et al. [9] introduced MIL to identify tumor metastasis WSIs for the first time. They cropped each WSI into 224 $\times$ 224 patches, so a WSI can be considered as a bag containing more than 10,000 instances. To save GPU usage, they embedded patches into feature embeddings before applying MIL. In their MIL method, they predict confidence score for each instance, and then select top-K instances' scores as the final representation of a bag.

Following this work, several MIL models are proposed to aggregate the instances and learn bag representations effectively. These models can be typically divided into two categories: *top-K instance based models* [2], [9], [14], [20] and *attention-based models* [8], [11], [12], [21]. The top-K instance based models usually require training instance-level classifier based on the corresponding slide-level labels [2], [14]. However, the training of this classifier needs either large-scale WSI dataset or WSIs with macro-tumor lesions, which are not readily available.

Attention-based models are currently the most popular type of MIL model. It employs an attention or self-attention [15] module to learn appropriate attention weights and aggregate the instances within the WSIs. Nevertheless, the learning of attention weight is based on slide-level labels only, making it difficult to identify tumor instances to achieve moderate slide-level sensitivity. To overcome this challenge, Zhang et al. [22] propose to use GradCAM [23] mechanism to learn the positive probability of instances to pre-select a set of possible tumor instances for the MIL model. Tang et al., [24] propose to highlight the "hard-to-classify" instances by using a hard example mining mechanism. Given these improvements, existing attention-based MIL models still suffer from undesirable sensitivities, especially for the micro-tumor lesion cases. This has been shown in a previous study [25] where they show several MIL models have only 11-46% accuracy in identifying tumor WSIs in the micro-metastasis cases (tumors no larger than 2 mm on a slide). Therefore, in this paper, we propose providing additional guidance to the MIL model's attention module to improve the highlighting of salient instances.

B. Open-Set Learning

In open-set learning, the task is to classify the categories that has been seen during the training, and identify the data from unseen categories in the meantime [17]. Typically, this is accomplished by comparing input data to example seen data [17], [18], [26]. Specifically, a previous work proposes an attention-based architecture to correlate the local regions of an input image with the local regions from a support image set [16]. Hence, their model can highlight the local image regions from the unseen categories. Inspired by open-set learning methods, the proposed method first learns a set of representative normal instances, and then utilizes a novel cross-attention mechanism to infer possible tumor instances.

C. Cross-Attention for Medical Image Analysis

Cross-attention is a type of attention mechanism that allows neural networks to incorporate external information into learned features. Various cross-attention networks have been applied to numerous medical image analysis tasks. In some studies, cross-attention is used to combine multimodal data, such as medical images and clinical reports. Typical applications include image generation [27], visual question answering [28], PET-CT fusion [29]. Other research involves performing cross-attention between a query feature map and a support feature map representing lesion features to create a more lesion-focused query feature map [30], [31], [32], [33]. However, most of these studies focus on relatively small-sized medical images, such as radiology and fundus images. In this study, we propose integrating a novel cross-attention module into the MIL model to highlight the lesions from the giga-pixel size WSIs. We anticipate this new MIL design will help the model identify tumors and differentiate tumor WSIs from normal WSIs, even when the tumors are extremely small.

III. Method

A. Datasets

Our study is based on three publicly available WSI datasets. The first two datasets, named Camelyon16 and Camelyon17 [34], [35] are well-known deep-learning benchmarks for the automated detection of breast cancer lymph node metastasis (BCLNM) in hematoxylin and eosin (H&E) stained WSIs of lymph node biopsies.

Camelyon16 contains a training set with 270 WSIs and a hold-out testing set with 129 WSIs that are divided into two categories: normal and tumor. Tumor WSIs consist of both macro- and micro-metastasis WSIs, with the latter containing tumor lesions no larger than 2 mm and more challenging to detect. In our MIL study, we perform binary classification to identify normal and tumor WSIs using only slide-level labels.

The Camelyon17 dataset contains a training set with 500 WSIs and a testing set with 500 WSIs divided into normal and tumor categories. The latter is more challenging as it includes WSIs collected from five hospitals. We utilize this dataset to assess our method's generalizability to WSIs from unseen hospitals during training. To this end, we conduct a cross-center cross-validation study based on the 500 WSIs from the Camelyon17 training set, where the source hospitals are labeled.

We also included TCGA-NSCLC dataset as a subtyping benchmark for our MIL model. TCGA-NSCLC comprises WSIs from two non-small cell lung cancer subtypes: lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC). Among the 1023 WSIs in this dataset, 526 are LUAD WSIs and 497 are LUSC WSIs. Our model is trained to classify the subtypes of these WSIs, demonstrating its versatility in handling different histopathological tasks.

B. Revisiting MIL for WSI Classification

In the MIL paradigm, a WSI is first cropped into small image patches, and then, via the feature extraction module, all patches are transformed into feature embeddings. Throughout this paper, we refer to a WSI and a patch as a bag and an instance, respectively. For the sake of representation, we also assumed that $X_i = \{x_{i1}, x_{i2}, \ldots, x_{in_i}\}$ denotes the set of all instances within the ith bag, where x_{ij} is the jth instance of the ith bag. If an instance comes from a tumor region, then a positive label (encoded by 1) will be assigned to the instance, and otherwise, iis called positive if and only if at least one positive instance exists in the bag. If all instances within a bag are negative, then a negative label will be assigned to the bag. Therefore, the true bag-level label, Y, can be defined by:

$$Y = \begin{cases} 0, & \text{iff } \sum_j y_j = 0, \\ 1, & \text{otherwise.} \end{cases} \tag{1}$$

where $y_j \in \{0,\ 1\}$, for $j = 1, \ldots,\ n_i$, denotes the jth instance of the bag [36]. To make a bag-level decision (bag-level label prediction) for the ith bag, X_i, all instances within a bag, $\{x_{i1}, x_{i2}, \ldots, x_{in_i}\}$, are first transformed into instance-level embeddings. Then all instance-level embeddings are aggregated into a bag-level embedding. Finally, a classifier is employed that takes the bag-level embedding as an input and produces a bag-level label prediction as an output. The process can be formulated as:

$$\tilde{Y} = g\left(\sigma\left(f\left(x_{i1}\right),\ \ldots,\ f\left(x_{in_i}\right)\right)\right) \tag{2}$$

where $\tilde{Y}$ is a predicted bag-level label, $f(\cdot)$ is an instance-level embedding transformation function, $\sigma(\cdot)$ is an aggregation function, and $g(\cdot)$ is a bag-level prediction classifier.

Currently, the common practice for the aggregation function in multiple instance learning (MIL) models is attention-based weighted summation, where the attention weights are learned for each instance by a neural network. The essential idea is to enable the MIL model to highlight potential tumor instances within a bag, thereby differentiating tumor WSIs from normal WSIs. While these models have achieved significant success, they often exhibit low sensitivity on WSIs with micro-metastasis [25], where the majority of regions in the WSIs are non-tumorous.

This limitation arises because the MIL model struggles to adequately focus on the minority tumor instances when it can only leverage bag-level labels (i.e., tumor or normal WSI) for differentiation during training. Without instance-level annotations, the model may not pay enough attention to the few but crucial tumor regions within a largely normal WSI.

Our objective is to address this issue by providing additional guidance to the MIL model, enabling it to better differentiate between tumor and normal WSIs without relying on instance-level annotations. We aim to achieve this by:

1) Learning representative negative instances, which helps the model understand and ignore normal tissue regions more effectively.
2) Employing our novel cross-attention-based salient instance inference architecture, which enhances the model's ability to identify and focus on potential tumor regions within a bag.

By implementing these strategies, we aim to improve the sensitivity and overall performance of MIL models, particularly

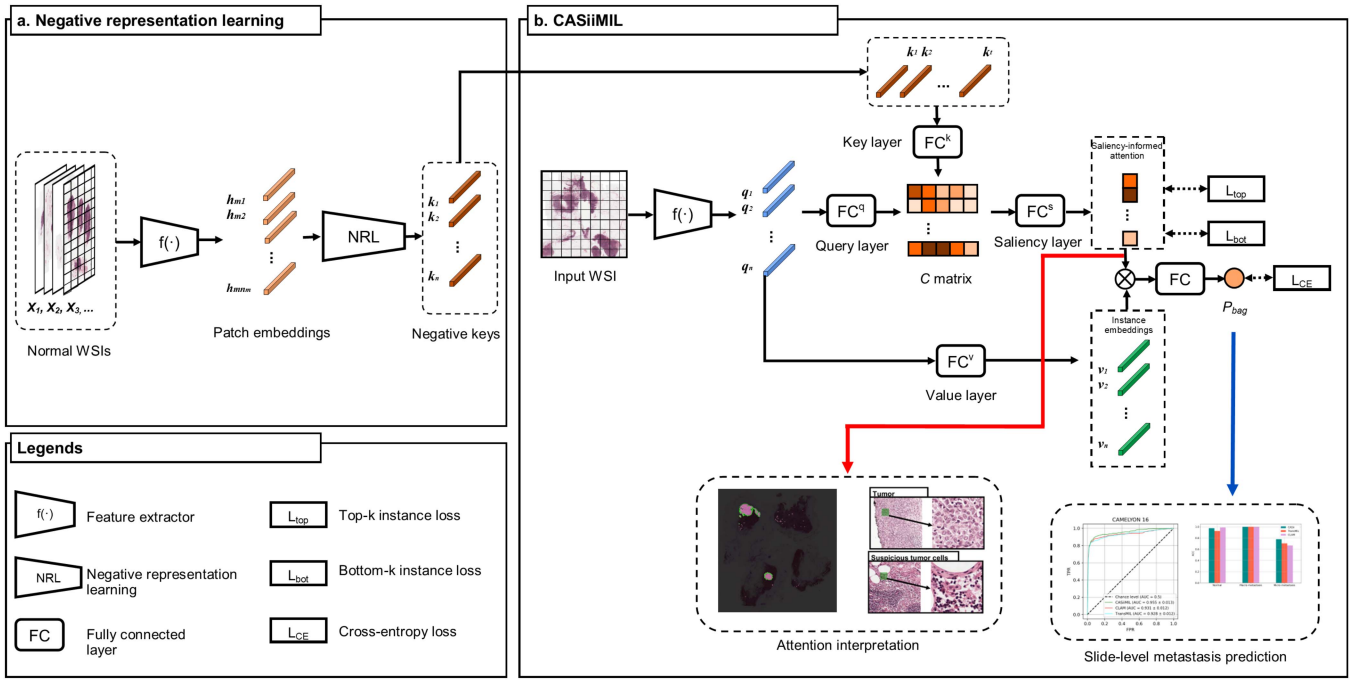

Fig. 1. Overview of CASiiMIL. (a) Negative Representation Learning (NRL): We extract negative keys from training normal WSIs using a CUR-based method to identify the top representative patch embeddings from each normal WSI. These embeddings form a negative key set for the subsequent CASiiMIL step. (b) CASiiMIL for Bag-Level Classification: Our cross-attention-based model correlates the input instances with the negative keys in the C matrix and then learns saliency-informed attention scores through the saliency layer FC^s. Using the learned attention weights, we aggregate all instances to make the final slide-level prediction. To enhance model performance, we incorporate a regular cross-entropy loss along with two instance-level loss functions, L_{bot} and L_{top}, to encourage attention sparsity.

in challenging cases involving micro-metastasis, ultimately contributing to more accurate and reliable diagnostic tools in clinical practice.

C. The Proposed CASiiMIL

The proposed method is composed of three main steps: (i) forming a bag by cropping an input WSI into patches and embedding the patches into feature embeddings using pretrained model (see Section III-E and III-F for details); (ii) learning negative keys (i.e., representative negative instances) from normal WSIs in the training set using the proposed negative representation learning (NRL) module; (iii) bag-level classification using CASiiMIL based on an input bag and negative keys. The overview of our method is shown in Fig. 1.

1) Negative Representation Learning: Redundancy in histopathology images refers to the presence of repetitive and overlapping visual information, commonly observed in both pathological and normal histology. In normal histology, the structured nature of tissue architecture and the consistency of cellular patterns across different sections contribute to this redundancy. While such redundancy can enhance the robustness of image analysis by providing consistent references, it also necessitates more sophisticated algorithms to discern subtle pathological changes amidst the repetitive normal patterns. Effectively addressing redundancy is crucial for improving the accuracy and efficiency of computational pathology and enhancing diagnostic precision. Since pathological slides often contain normal histology and we lack precise tumor locations, we decided to learn the normal histology representation in a succinct manner from normal WSI. We approached this normal histology learning as a redundancy minimization problem through NRL.

The Negative Representation Learning (NRL) module is a key component of our methodology, designed to enhance the specificity and accuracy of tumor instance detection in WSIs by focusing on learning discriminative features from normal (non-tumor) instances. The NRL module operates by identifying and representing the most informative normal patches from WSIs, which serve as a baseline for distinguishing subtle tumor features.

Given a training set that contains P negative bags (normal WSIs) denoted by $X = \{X_1, X_2, \ldots, X_P\}$, where X_m is the m^{th} negative bag that contains n_m negative instances (normal patches), $X_m = \{x_{m1}, x_{m2}, \ldots, x_{mn_m}\}$, we first apply a feature extraction neural network $f(\cdot)$ on $x_{mq} \in X_m$, for $q = 1, 2, \ldots, n_m$ and $m = 1, 2, \ldots, P$, to construct feature embeddings $h_{mq} \in \mathbb{R}^D$. For $m = 1, 2, \ldots, P$, we then construct $A_m = [h_{m1}, h_{m2}, \ldots, h_{mn_m}] \in \mathbb{R}^{D \times n_m}$, apply CUR decomposition [37] on A_m to identify the most representative normal instances. The CUR decomposition technique specifically targets the selection of columns from A_m that best capture the inherent variability and statistical significance of the features within the data, without a significant loss of information.

This selection is based on the statistical importance of each column, assessed through measures such as leverage scores.

Algorithm 1: Pseudo-Code for Negative Representation Learning (NRL).

Input: The set of embedded normal WSIs, $A = \{A_1, A_2, \ldots, A_P\}$.

Step 1: for $m = 1, 2, \ldots, P$ do

- **Step 1.1:** Compute the rank of A_m, and set $k = rank(A_m)$.
- **Step 1.2:** Construct matrix V whose rows are the eigenvector of $A_m^T A_m$.
- **Step 1.3:** Compute the importance score of the j^{th} column of A_m by $s_j = \frac{1}{k}\sum_{h=1}^{k} V_{hj}^2$, where V_{hj} is the element in the h^{th} row and j^{th} column of V, for $j = 1, 2, \ldots, n_m$.
- **Step 1.4:** Sort columns of A_{b_m} based on the scores s_j's.
- **Step 1.5:** Construct $\tilde{A}_{b_m} \in \mathbb{R}^{D\times t_i}$ whose columns are the first t_m columns of sorted A_{b_m} in Step 1.4.

End (for).

Output: Construct a key matrix K by concatenating all $\tilde{A}_{b_m}$, for $m = 1, 2, \ldots, P$, $K = \tilde{A}_{b_1} \oplus \tilde{A}_{b_2} \ldots \oplus \tilde{A}_{b_P} = [k_1, \ldots, k_\tau] \in \mathbb{R}^{D\times \tau}$

These scores reflect how much each column contributes to the overall variance and structure of the data, thereby allowing us to maintain the most informative features. The columns selected through this process are then used to form a submatrix $\tilde{A}_{b_m} \in \mathbb{R}^{D\times t_m}$. This submatrix contains columns that are a subset of A_m and represent the high statistical leverage of the m^{th} negative bag's feature embeddings. These columns will capture the key characteristics of normal tissue architecture and provide a robust baseline for distinguishing tumor features.

Finally, we construct a key matrix K of representative negative instances by concatenating $\tilde{A}_{b_m}$ for all $m = 1, 2, \ldots, P$,

$$K = \tilde{A}_{b_1} \oplus \tilde{A}_{b_2} \ldots \oplus \tilde{A}_{b_P} = [k_1, \ldots, k_\tau] \in \mathbb{R}^{D\times \tau} \quad (3)$$

where $\oplus$ denotes the concatenation operation and $\tau = \sum_{m=1}^{P} t_m$. This process is depicted in Algorithm 1.

2) Cross-Attention-Based Salient Instance Inference MIL: In this section, we introduce a new cross-attention-based salient instance inference MIL (CASiiMIL) model that can efficiently highlight salient instances of positive bags. Unlike existing attention-based MIL methods [2], [8], [11], [12], which learn attention weights solely from input instances, our cross-attention-based architecture can automatically correlate the input instances and negative keys, enabling the learning of high attention weights for instances that have low semantic relevance to normal tissues.

Suppose a fixed key matrix $K = [k_1, k_2, \ldots, k_\tau] \in \mathbb{R}^{D\times\tau}$ is constructed from all negative bags (see Section III-C-1), and a random input bag (WSI) containing n instances, $Q = [q_1, q_2, \ldots, q_n] \in \mathbb{R}^{D\times n}$, is given. The keys $k_1, k_2, \ldots, k_\tau$, and the queries $q_1, q_2, \ldots, q_n$ are first transformed into three latent feature spaces: key, query and value spaces, via three different fully connection layers:

$$\tilde{k}_j = FC^k(k_j) = \tanh\left(W_k^T k_j + b_k\right) \in \mathbb{R}^{D_h}, \quad (4)$$

$$\tilde{q}_i = FC^q(q_i) = \tanh\left(W_q^T q_i + b_q\right) \in \mathbb{R}^{D_h}, \quad (5)$$

$$\tilde{v}_i = FC^v(q_i) = \mathrm{ReLU}\left(W_v^T q_i + b_v\right) \in \mathbb{R}^{D_h}, \quad (6)$$

where $W_k, W_q, W_v \in \mathbb{R}^{D\times D_h}$, and $b_k, b_q, b_v \in \mathbb{R}^{D_h}$ are learnable parameters. We then construct a cross-attention matrix C,

$$C = [c_{ij}] = \tilde{Q}^T \tilde{K} \in \mathbb{R}^{n\times\tau}, \quad (7)$$

where $\tilde{K} = [\frac{\tilde{k}_1}{||\tilde{k}_1||}, \frac{\tilde{k}_2}{||\tilde{k}_2||}, \ldots, \frac{\tilde{k}_\tau}{||\tilde{k}_\tau||}] \in \mathbb{R}^{D_h\times\tau}$ and $\tilde{Q} = [\frac{\tilde{q}_1}{||\tilde{q}_1||}, \frac{\tilde{q}_2}{||\tilde{q}_2||}, \ldots, \frac{\tilde{q}_n}{||\tilde{q}_n||}] \in \mathbb{R}^{D_h\times n}$. In matrix C, each row is a correlation vector between a query, q_i, to all the keys, $k_1, k_2, \ldots, k_\tau$.

Then, we construct a saliency layer, FC^s, whose inputs are rows of cross-attention matrix, $C_i \in \mathbb{R}^\tau$, and outputs are the saliency logits, $s_i \in \mathbb{R}$, for the queries:

$$s_i = FC^s(C_i) = W_s^T C_i + b_s \in \mathbb{R}. \quad (8)$$

Here, $W_s \in \mathbb{R}^{\tau\times 1}$ and $b_s \in \mathbb{R}$ are learnable parameters. The functionality of this saliency layer is to assign low attention scores to instances highly correlated with negative keys (likely normal) and high scores to those with low correlation (likely tumorous). By visualizing the attention map (see Fig. 4) in the Results section, we show that the saliency layer learns this reverse logic after training.

Finally, we compute a bag-level embedding for the input bag by aggregating the latent queries in value space as follows:

$$z = \sum_{i}^{n} a_i \tilde{v}_i, \quad (9)$$

where:

$$a_i = \frac{\exp(s_i)}{\sum_i^n \exp(s_i)}, \quad (10)$$

where a_i are the saliency informed attention weights. Finally, we feed the bag-level embedding z into a fully connected layer to classify the bag-level label (i.e., normal or tumor WSI) of the input bag in a supervised manner. The overall CASiiMIL architecture is depicted in Fig. 1(b).

D. Model Training

The proposed model is primarily trained with a binary cross-entropy loss function:

$$L_{CE} = -Y log(P_{bag}) - (1-Y)\log(1-P_{bag}), \quad (11)$$

where $Y \in \{0, 1\}$ is bag-level (slide-level) label of a WSI, and $P_{bag} \in [0, 1]$ is the bag-level probability of being positive predicted by CASiiMIL.

Moreover, we introduce two instance-level loss functions for the instances with bottom-r and top-r attention weights within each WSI. We propose these two instance-level loss functions to encourage the sparsity of the attention weights and guide CASiiMIL model to learn appropriate attention weights for each WSIs. The loss functions are as follows:

$$L_{bot} = -(1-Y_0)\log\left(1 - P_s^{n-r, \ldots, n}\right), \quad (12)$$

$$L_{top} = -Y_1 \log\left(P_s^{1,\cdots,r}\right), \tag{13}$$

where:

$$P_s = \sigma(s), \tag{14}$$

where $Y_0 = 0$ and $Y_1 = 1$ are the pseudo-labels assigned to the bottom- and top-r instances, s is the saliency logit (see (8)) of one of the bottom or top-r instances, and $\sigma(\cdot)$ is the sigmoid activation function. In practice, L_{bot} is applied for all training WSIs and L_{top} is applied only for positive training WSIs. We train the model solely via L_{CE} for 5 epochs to allow the model to warm-up. Each loss is averaged across the bottom- or top-r instances before optimization.

As a result, the proposed model is trained via:

$$L = L_{CE} + \lambda_1 L_{bot} + \lambda_2 L_{top}, \tag{15}$$

where λ_1 and λ_2 are constant coefficients for the instance-level losses.

E. Histopathology Specific Feature Extractor

In this study, we apply two different pretrained CNN feature extractors (i.e., $f(\cdot)$), namely ResNet50 [38] and CTransPath [39], for patch feature extraction. ResNet50 (truncated at the third residue block) is the most common feature extractor for MIL-based WSI analysis studies and pretrained on the ImageNet dataset [40]. The dataset contains more than one million natural images divided into 1000 categories, and this model has an output dimension of $D = 1024$. Despite its widespread use and successful applications [8], [12], [41], [42], the domain shift issue between natural images and histopathology images remains. Thus, we propose to use CTransPath, which is a transformer-based histopathology specific feature extraction model [39]. CTransPath is pretrained in a self-supervised learning manner using around 15 million histopathology image patches collected from the cancer genome atlas (TCGA) and pathology AI platform (PAIP) datasets. Its output dimension is $D = 768$.

F. Implementation Details

For preprocessing, all WSIs were cropped into patches in the size of 224×224 under $20\times$ magnification (same as the settings of some recent studies [8], [11]). Patches from foreground tissue regions were extracted using color thresholding.

Our model was optimized by Adam optimizer [43] with 0.0002 learning rate and 0.00001 weight decay. The training was carried out with 10 epochs warm-up steps and halted if the validation AUC didn't improve for over 10 epochs. For the instance-level loss functions (14) and (15), we empirically chose instances with bottom and top-5 attention weights since some WSIs have small tissue regions or micro-metastasis. Our code is available at https://github.com/cialab/CASiiMIL.

G. Experimental Design

To evaluate the performance of the proposed model, we run our model five times where we randomly split the training set of Camelyon16 into training and validation sets in a ratio of 9:1. Then, we selected the model with the best validation AUC from the five models and tested the model on the official testing set of Camelyon16.

To evaluate the cross-center generalizability of the proposed model, we also conducted five-fold cross-center cross-validation on the Camelyon17 dataset. Namely, in each fold, we employed WSIs from one center as the testing set, and combined the WSIs from the rest four centers and the Camelyon16 training WSIs as the training set. The training set was also randomly split into training and validation set in a ratio of 9:1. For comparison, we conducted the same experiments on the state-of-the-arts MIL methods [8], [11], [12], [22], [24], [44]. These two methods represent two different types of MIL frameworks: attention-based and self-attention-based methods. Both these methods have been reported to outperform other MIL frameworks. The experimental results are reported in Section IV.

TABLE I
SLIDE-LEVEL CLASSIFICATION RESULTS ON CAMELYON16 BASED ON CTRANSPATH FEATURE EXTRACTOR. 95% CI REPORTED IN []

	AUC	F1	PRECISION	RECALL
CLAM [8]	0.9227 [0.9187, 0.9248]	0.8913 [0.8838, 0.8967]	0.9535 [0.9496, 0.9559]	0.8367 [0.8296, 0.9400]
TransMIL [12]	0.9313 [0.9286, 0.9331]	0.8541 [0.8470, 0.8553]	0.8723 [0.8667, 0.8745]	0.8367 [0.8310, 0.8398]
DTFD-MIL [22]	0.9408 [0.9382, 0.9422]	0.8444 [0.8363, 0.8445]	0.9268 [0.9238, 0.9313]	0.7755 [0.7690, 0.7799]
SSM-MIL [44]	0.9393 [0.9382, 0.9432]	0.8723 [0.8688, 0.8762]	0.9111 [0.9094, 0.9176]	0.8367 [0.8348, 0.8762]
MHIM-MIL [24]	0.9349 [0.9333, 0.9396]	0.8400 [0.8352, 0.8473]	0.8235 [0.8219, 0.8374]	0.8571 [0.8526, 0.8668]
CASiiMIL	**0.9679** **[0.9663, 0.9688]**	**0.9149** **[0.9101, 0.9151]**	**0.9556** **[0.9528, 0.9578]**	**0.8776** **[0.8722, 0.8799]**

IV. RESULTS

A. BCLNM Identification

Tables I, II, and Fig. 2 report the slide-level classification results on the Camelyon16 dataset based on two different feature extractors. In comparison with state-of-the-art MIL models, CASiiMIL achieves the best overall performance. Although some methods achieve better precision than CASiiMIL, the proposed method achieves the best recall while maintaining an outstanding precision. As a result, CASiiMIL maintains a good balance between precision and recall. Fig. 2 reports the slide-level classification accuracies on the WSIs grouped in normal, macro-metastasis, and micro-metastasis. Moreover, Fig. 3(a) reports the averaged ROC curve of the five runs based on the Camelyon16 dataset.

Table III reports the cross-center slide-level classification results on the Camelyon17 dataset based on CTransPath feature extractor. In comparison with state-of-the-art MIL models, CASiiMIL exhibits better cross-center generalizability.

TABLE II
SLIDE-LEVEL CLASSIFICATION RESULTS ON CAMELYON16 BASED ON RESNET FEATURE EXTRACTOR. 95% CI REPORTED IN []

	AUC	F1	PRECISION	RECALL
DSMIL [11]	0.8265 [0.8219, 0.8281]	0.6197 [0.6086, 0.6193]	**1.0000** [1.0000, 1.0000]	0.4490 [0.4429, 0.4540]
CLAM [8]	0.8319 [0.8298, 0.8367]	0.7500 [0.7440, 0.7524]	0.8462 [0.8380, 0.8475]	0.6735 [0.6682, 0.6791]
TransMIL [12]	0.8520 [0.8463, 0.8537]	0.7907 [0.7867, 0.7946]	0.9189 [0.9187, 0.9262]	0.6939 [0.6915, 0.7022]
DTFD-MIL [22]	0.8452 [0.8439, 0.8513]	0.7579 [0.7508, 0.7598]	0.7826 [0.7783, 0.7897]	0.7334 [0.7311, 0.7425]
SSM-MIL [44]	0.8668 [0.8639, 0.8709]	0.7407 [0.7365, 0.7470]	0.9375 [0.9342, 0.9423]	0.6122 [0.6060, 0.6193]
CASiiMIL	**0.8842** **[0.8804, 0.8860]**	**0.8222** **[0.8133, 0.8205]**	0.9024 [0.8960, 0.9036]	**0.7551** **[0.7473, 0.7571]**

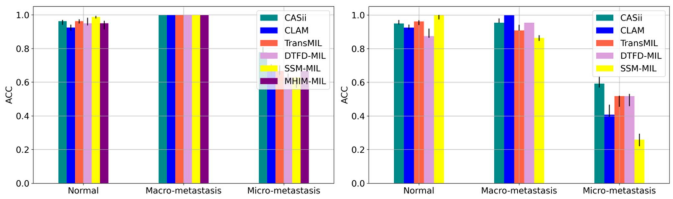

Fig. 2. Slide-level classification accuracies with 95% CIs on Camelyon16 dataset that grouped in normal, macro-metastasis, and micro-metastasis. (a) Results based on CTransPath feature extractor. (b) Results based on ResNet50 feature extractor.

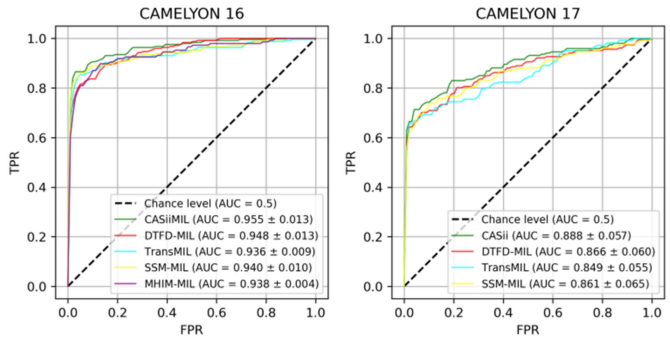

Fig. 3. (a) Averaged ROC curve of the five runnings on Camelyon16 dataset. (b) Averaged ROC curve of five-fold cross-center cross-validation on Camelyon17 dataset. All results are based on CTransPath feature extractor.

B. Lung Cancer Subtyping

We also adapted CASiiMIL for cancer subtyping in NSCLC (non-small cell lung cancer) to demonstrate its versatility. We trained and evaluated our model on the TCGA-NSCLC dataset, which includes WSIs categorized as LUAD (lung adenocarcinoma) and LUSC (lung squamous cell carcinoma). Cancer subtyping is different from traditional MIL formulation that we discussed in Section III-B. Instead of predicting the presence of tumor instance in a bag, we need to comprehensively analyze whether the instances are from one tumor type or another tumor type. Thus, we perform NRL on two subtypes separately to extract LUAD keys and LUSC keys. Then, we will employ two cross-attention branches to predict LUAD and LUSC. The logits predicted from the two branches are then normalized by softmax function for final prediction.

Since the TCGA-NSCLC dataset lacks an official testing set, we conducted a four-fold cross-validation as per Tang et al. [24]. Table IV reports our classification results, where CASiiMIL outperforms state-of-the-art MIL models.

C. Ablation Studies and Computation Analysis

To demonstrate the effect of different loss functions, we conduct an ablation study on different combination of our loss function terms. The results are summarized in Table V. We find that combining L_{CE}, L_{bot}, and L_{top} lead to the best overall performance. Besides, both L_{bot} and L_{top} were helpful individually, indicating that encouraging the sparsity of the attention weights using instance-level loss functions leads to improved classification performance.

To investigate the effect of the number of keys extracted using NRL, we varied the number of keys extracted from each WSI (t_m). We set t_m to 20, 50, 100, 200, and 300, naming the resulting key matrices as: (1) NRL-20, (2) NRL-50, (3) NRL-100, (4) NRL-200, and (5) NRL-300. Additionally, we also applied a random selection strategy to select 50 keys per WSI, resulting in a key matrix labeled as Rand-50. This key matrix was used to compare with our NRL method to verify its effectiveness. We applied these key matrices separately in our CASiiMIL model and reported the resulting testing AUCs and F1-scores in Table VI. The results indicate that the performance of CASiiMIL is influenced by the number of keys extracted. The highest AUC and F1-scores were achieved with NRL-50, suggesting that an optimal number of keys exist for maximizing model performance. When the number of keys is set too high (NRL-200 and NRL-300), there is a notable drop in both AUC and F1-scores, indicating potential overfitting or noise introduced by an excessive number of keys. Furthermore, CASiiMIL's performance based on NRL-selected keys is noticeably better than its performance based on randomly selected keys, highlighting the representativeness and effectiveness of our NRL-selected keys.

We also report the FLOPs and model size (Params) in the Table VII. The computation is based on an input bag with shape (1, 120, 1024) under the evaluation mode following the strategy of DTFD-MIL's [22]. This analysis demonstrates that CASiiMIL has comparable complexity to other MIL models, except for TransMIL [12], which has a significantly larger size due to its use of multiple transformer blocks.

D. Interpretability

To demonstrate the interpretability of the proposed model, we visualize the attention outputs (see (10)) of CASiiMIL overlaid on the original WSIs in Fig. 4. Specifically, in the top row, we highlight the patches that receive the largest 10% attention weights and dim the remaining regions. In the bottom row, we visualize the overall attentions in a heatmap format. The visualization results demonstrate CASiiMIL's sensitivity on tumor lesions in different sizes. Additionally, we invited a

TABLE III
CROSS-CENTER SLIDE-LEVEL CLASSIFICATION RESULTS ON CAMELYON17 DATASET BASED ON CTRANSPATH FEATURE EXTRACTOR. 95% CI REPORTED IN []

	Center0		Center1		Center2		Center3		Center4	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1
TransMIL [12]	0.8251 [0.8184, 0.8257]	0.6415 [0.6308, 0.6434]	0.7672 [0.7656, 0.7733]	0.6250 [0.6179, 0.6278]	0.8453 [0.8448, 0.8538]	0.6552 [0.6484, 0.6598]	0.9098 [0.9074, 0.9122]	0.8108 [0.8021, 0.8103]	0.9162 [0.9116, 0.9168]	0.8696 [0.8679, 0.8746]
CLAM [8]	0.8316 [0.8275, 0.8348]	0.7018 [0.6928, 0.7018]	**0.8239 [0.8221, 0.8279]**	0.6290 [0.6237, 0.6301]	0.9131 [0.9100, 0.9166]	0.6667 [0.6537, 0.6683]	0.9258 [0.9234, 0.9275]	0.8421 [0.8357, 0.8428]	0.9088 [0.9085, 0.9130]	0.7792 [0.7742, 0.7824]
DTFD-MIL [22]	0.8468 [0.8414, 0.8490]	**0.7500 [0.7398, 0.7498]**	0.7674 [0.7639, 0.7720]	0.6667 [0.6572, 0.6667]	0.8811 [0.8769, 0.8838]	0.6897 [0.6778, 0.6892]	0.9379 [0.9348, 0.9394]	**0.9041 [0.8975, 0.9034]**	0.9138 [0.9117, 0.9169]	0.8732 [0.8667, 0.8733]
SSM-MIL [44]	0.8303 [0.8264, 0.8340]	0.6786 [0.6678, 0.6791]	0.7574 [0.7542, 0.7662]	0.6512 [0.6425, 0.6558]	0.8731 [0.8699, 0.8786]	0.6087 [0.5950, 0.6105]	0.9304 [0.9270, 0.9331]	0.8219 [0.8128, 0.8237]	0.9302 [0.9257, 0.9317]	0.8158 [0.8056, 0.8164]
CASiiMIL	**0.8529 [0.8498, 0.8567]**	0.7368 [0.7281, 0.7370]	0.7976 [0.7915, 0.7991]	**0.7000 [0.6927, 0.7017]**	**0.9296 [0.9248, 0.9302]**	**0.7600 [0.7527, 0.7639]**	**0.9446 [0.9422, 0.9457]**	0.8493 [0.8439, 0.8509]	**0.9340 [0.9317, 0.9365]**	**0.8889 [0.8738, 0.8890]**

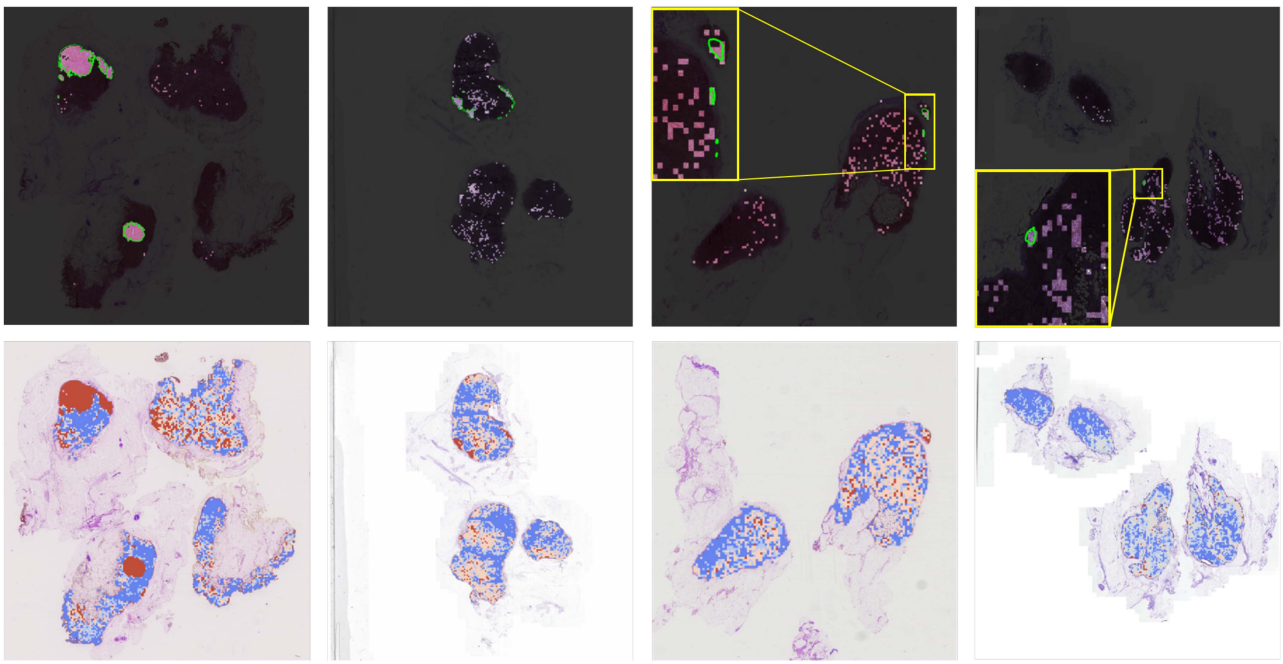

Fig. 4. Attention visualization of example tumor WSIs. In the top row, the bright areas correspond to the highest 10% attention weights and the dark areas correspond to regions receiving low attention weights. Tumor regions are annotated in green color. From left to right, we visualize the example WSIs with different tumor sizes ranges from large to small. In the bottom row, we visualize the attention heatmaps of the corresponding WSIs. The results demonstrate the sensitivity of CASiiMIL on tumor lesions of different sizes.

TABLE IV
SLIDE-LEVEL CLASSIFICATION RESULTS ON TCGA-NSCLC BASED ON CTRANSPATH FEATURE EXTRACTOR. 95% CI REPORTED IN []

	AUC	Accuracy	F1
CLAM [8]	0.9643 [0.9618,0.9655]	0.9091 [0.9055,0.9113]	0.9103 [0.9047,0.9107]
TransMIL [12]	0.9597 [0.9575,0.9613]	0.8944 [0.8926,0.8986]	0.8975 [0.8936,0.8998]
DTFD-MIL [22]	0.9583 [0.9555,0.9595]	0.9032 [0.9004,0.9063]	0.9053 [0.9003,0.9065]
SSM-MIL [44]	0.9618 [0.9604,0.9640]	0.9081 [0.9046,0.9104]	0.9106 [0.9039,0.9100]
MHIM-MIL [24]	0.9426 [0.9406,0.9451]	0.8514 [0.8489,0.8557]	0.8595 [0.8535,0.8607]
CASiiMIL	**0.9690 [0.9680,0.9712]**	**0.9101 [0.9067,0.9123]**	**0.9122 [0.9063,0.9124]**

TABLE V
ABLATION STUDY ON LOSS FUNCTIONS. RESULTS ARE AVERAGED TESTING RESULTS ON CAMELYON16 DATASET

	AUC	F1
L_{CE}	0.9456	0.8736
$L_{CE} + L_{bot}$	0.9514	0.8657
$L_{CE} + L_{top}$	0.9527	0.8855
$L_{CE} + L_{bot} + L_{top}$	**0.9545**	**0.8919**

pathologist to interpret the false positive patches that received the largest 10% attention weights but were not from the annotated tumor regions, according to Camelyon16's official annotation.

TABLE VI
ABLATION STUDY ON KEY SET SIZE. RESULTS ARE AVERAGED TESTING RESULTS ON CAMELYON16 DATASET

	AUC	F1
NRL-20	0.9537	0.8889
NRL-50	**0.9545**	**0.8919**
NRL-100	0.9329	0.8483
NRL-200	0.8135	0.6460
NRL-300	0.8542	0.6840
Rand-50	0.9258	0.8669

TABLE VII
COMPUTATION ANALYSIS OF DIFFERENT MIL MODELS. THE COMPUTATION IS BASED ON AN INPUT BAG WITH SHAPE (1, 120, 1024) UNDER THE EVALUATION MODE

	FLOPS	PARAMS
CLAM [8]	118.9M	790K
TransMIL [12]	613.8M	2.66M
DTFD-MIL [22]	79.4M	986K
SSM-MIL [44]	252.2M	1.05M
MHIM-MIL [24]	134.1M	1.18M
CASiiMIL	178.2M	1.57M

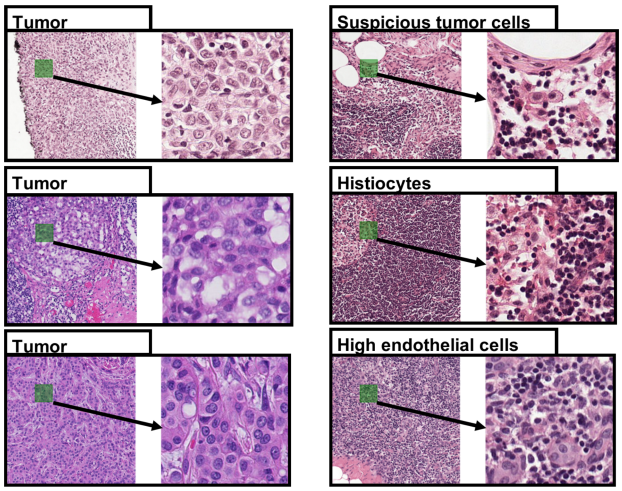

Fig. 5. Interpretation of patches with high attention weights. Here, we show example patches that receive the largest 10% attentions weights within their corresponding WSIs. (a) True positive patches that are from annotated tumor regions. (b) False positive patches that are not from annotated tumor regions.

We find that most false positive patches are normal cells such as histiocytes (tissue macrophages) and high endothelia cells that shared similar morphology with tumor cells, such as large cell size, large but less blue nucleus, profound nucleolus, etc. Some example patches are shown in Fig. 5.

V. DISCUSSIONS AND CONCLUSION

This manuscript makes a significant contribution to the field of pathology image analysis by proposing a novel method for automatically computing a saliency score for each patch in an image, which can be used to arrive at a slide-level decision. The motivation behind this study is to develop a reliable and accurate diagnostic tool for the detection of tumors, including challenging-to-identify micro-metastases. Accurate diagnosis and treatment planning are crucial in clinical practice, and current methods often struggle with balancing recall and precision. Our objective is to address this challenge by leveraging CASiiMIL's capabilities. CASiiMIL is designed to enhance the sensitivity of multiple instance learning (MIL) to tumor instances, ensuring high recall rates. High recall is essential to minimize the risk of under-diagnosis by reducing missed tumor instances. Additionally, CASiiMIL aims to maintain excellent precision to minimize false positives, which can lead to unnecessary medical interventions. By achieving this balance, CASiiMIL aims to improve the accuracy and reliability of histopathological diagnoses, ultimately contributing to better patient outcomes and advancing the standards of clinical practice.

CASiiMIL has several advantages over existing approaches, including the ability to produce more reliable and interpretable attention maps, as well as the ability to correctly identify cases with extremely small lesions that may be missed by other methods. By computing a saliency score for each patch, our method is able to identify regions of interest within an image that are most likely to contain pathological features. This allows us to focus our analysis on these regions and improve the accuracy of our predictions. Additionally, by combining these scores across all patches in an image, we are able to arrive at a slide-level decision that takes into account all available information. One of the key advantages of our method is its ability to identify cases with extremely small lesions accurately. This is particularly important in pathology image analysis, where small lesions can be easily missed or overlooked by human observers. By automatically computing saliency scores for each patch, our method is able to detect even very small lesions and incorporate this information into the slide-level decision. Overall, our method represents a significant advance in the field of pathology image analysis and has the potential to improve the accuracy and reliability of diagnostic and prognostic predictions.

We innovatively developed a cross-attention architecture to integrate a salient instance inference module into the gradient-flow of MIL classification. There are two main advantages associated with our model. First, it transforms the key and queries to a latent space to correlate query instances with the negative keys automatically. Second, it enables the learning of saliency informed attention weights to highlight the possible positive instances in the bags. In Section IV-A, we show that the proposed CASiiMIL achieves outstanding slide-level classification performance on Camelyon16 dataset with both natural image pretrained and histopathology-specific feature extractors (see Tables I and II). It is noted that some comparison methods achieve better precision than CASiiMIL when utilizing the ResNet encoder. We attribute this to our cross-attention module, designed to enhance the MIL's sensitivity to tumor instances, potentially leading to some overcalls. Despite this, the proposed CASiiMIL maintains the best recall rate while also achieving excellent precision. Consequently, our overall performance surpasses that of all comparison methods. This finding highlights CASiiMIL's superior capability to accurately

identify tumor slides without compromising precision, a crucial aspect for diagnostic tools used in clinical practice.

Clinically, this means that CASiiMIL can reliably detect the presence of tumors, including difficult-to-identify micro-metastases, which are critical for accurate diagnosis and treatment planning. High recall ensures that fewer tumor instances are missed, reducing the risk of under-diagnosis, while maintaining excellent precision minimizes false positives, which can lead to unnecessary interventions. Therefore, CASiiMIL represents a valuable tool in improving the accuracy and reliability of histopathological diagnoses, ultimately contributing to better patient outcomes.

In Fig. 2, we compare different MIL models' slide-level classification accuracies on the WSIs grouped in normal, macro-metastasis, and micro-metastasis. The proposed CASiiMIL achieves the best accuracies on the micro-metastasis WSIs compared with all other MIL models. This result demonstrates that our cross-attention architecture and saliency-informed attention weights are helpful in identifying tumor WSIs, even when the tumor lesions are small.

In Table III, we report the cross-center slide-level classification results of CASiiMIL compared with other MIL methods based on Camelyon17 dataset. The results reveal that the proposed model can greatly generalize to WSIs from unseen hospitals during training. We noticed that MIL models generally achieved lower performance on the Camelyon17 dataset than Camelyon16. The primary reason lies in the differences in data collection and validation methodologies. Unlike Camleyon16, whose training and testing slides were collected from the same hospitals, Camelyon17 presents a more challenging scenario as the slides were collected from five different hospitals, introducing significant variability in the appearance, scanning parameters, and staining of the histopathology slides. This variability can affect the model's ability to generalize across different centers. Moreover, for Camelyon17, we performed cross-center cross-validation, ensuring that the training and testing sets are from different hospitals. This approach is more reflective of real-world scenarios where models must generalize to data from various sources. However, it also increases the task's difficulty, leading to lower ROC/AUC performance than Camelyon16.

In Table IV, we report the cross-validation results of CASiiMIL and comparison methods on the TCGA-NSCLC dataset. We found that CASiiMIL outperforms state-of-the-art MIL models on this subtyping benchmark. Cancer subtyping is a common yet unique MIL task that varies from the traditional positive/negative MIL setting. Our results exhibit CASiiMIL's versatility and robustness across different histopathological tasks, indicating its potential for broader clinical applications.

In the visualizations of attention heatmaps (see Fig. 4), we further show that our model is able to identify tumor lesions and predict high attention weights on the tumor instances. This outcome demonstrates the outstanding sensitivity and interpretability of the proposed saliency informed attention layer regardless of the size of tumor lesions. Besides tumor lesions, this model is also identifying several groups of benign cells (except for lymphocytes) that are normal components of lymph nodes. These cell groups are mostly high endothelial venules and histiocytes (tissue macrophages). They may share some similar morphological features with tumor cells in comparison to background lymphocytes, such as larger cell size, larger nucleus with profound nucleoli, and abundant cytoplasm. Notably, the model also identified a few isolated cells that were highly suspicious for tumor cells, based on the morphology comparing to the confirmed main tumor lesion in the same lymph node (see Fig. 5).

A limitation of the proposed method is that CASiiMIL is computationally expensive due to feeding the entire key set into the model during the training. Approximately, our model has 1.57 M learnable parameters and 6M non-learnable parameters (i.e., the entire negative key set). This manner requires key sets of limited size given certain computational memories. To this end, a more robust negative representation learning method is needed to extract a more representative and less redundant negative key set.

In summary, we propose a novel MIL model called CASiiMIL, which achieved excellent accuracy on the tumor WSI classification task. We innovatively developed cross-attention-based architecture that enables the learning of saliency informed attention weights for MIL aggregation. The proposed CASiiMIL is of great sensitivity and interpretability in classifying tumor WSIs regardless the how small the tumor regions are, which makes it a reliable automatic diagnostic tool.

Acknowledgment

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, National Institute of Biomedical Imaging and Bioengineering, Alliance Clinical Trials in Oncology, and National Cancer Institute.

References

[1] M. K. K. Niazi, A. V. Parwani, and M. N. Gurcan, "Digital pathology and artificial intelligence," *Lancet Oncol.*, vol. 20, no. 5, pp. e253–e261, 2019.
[2] 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.
[3] G. Litjens et al., "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.
[4] A. G. Waks and E. P. Winer, "Breast cancer treatment: A review," *Jama*, vol. 321, no. 3, pp. 288–300, 2019.
[5] M. B. Amin et al., "The eighth edition AJCC cancer staging manual: Continuing to build a bridge from a population-based to a more 'personalized' approach to cancer staging," *CA: A Cancer J. Clinicians*, vol. 67, no. 2, pp. 93–99, Mar. 2017.
[6] R. F. van la Parra, P. G. Peer, M. F. Ernst, and K. Bosscha, "Meta-analysis of predictive factors for non-sentinel lymph node metastases in breast cancer patients with a positive SLN," *Eur. J. Surg. Oncol.*, vol. 37, no. 4, pp. 290–299, Apr. 2011.
[7] A. Nottegar et al., "Extra-nodal extension of sentinel lymph node metastasis is a marker of poor prognosis in breast cancer patients: A systematic review and an exploratory meta-analysis," *Eur. J. Surg. Oncol.*, vol. 42, no. 7, pp. 919–925, Jul. 2016.
[8] M. Y. Lu, D. F. 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, 2021.
[9] P. Courtiol, E. W. Tramel, M. Sanselme, and G. Wainrib, "Classification and disease localization in histopathology using only global labels: A weakly-supervised approach," 2018, *arXiv:1802.02212*.

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

[11] 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.*, 2021, pp. 14318–14328.

[12] 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.*, 2021, vol. 34, pp. 2136–2147.

[13] L. Qu, M. Wang, and Z. Song, "Bi-directional weakly supervised knowledge distillation for whole slide image classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 15368–15381.

[14] Z. Su, T. E. Tavolara, G. Carreno-Galeano, S. J. Lee, M. N. Gurcan, and M. Niazi, "Attention2majority: Weak multiple instance learning for regenerative kidney grading on whole slide images," *Med. Image Anal.*, vol. 79, 2022, Art. no. 102462.

[15] A. Vaswani et al., "Attention is all you need," 2017, *arXiv:1706.03762*.

[16] R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen, "Cross attention network for few-shot classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 4003–4014.

[17] K. Cao, M. Brbic, and J. Leskovec, "Open-world semi-supervised learning," 2021, *arXiv:2102.03526*.

[18] C. Ding, G. Pang, and C. Shen, "Catching both gray and black swans: Open-set supervised anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7388–7398.

[19] T. E. Tavolara, Z. Su, M. N. Gurcan, and M. K. K. Niazi, "One label is all you need: Interpretable AI-enhanced histopathology for oncology," in *Seminars in Cancer Biology*, vol. 97, New York, NY, USA: Academic, 2023, pp. 70–85.

[20] P. Chikontwe, M. Kim, S. J. Nam, H. Go, and S. H. Park, "Multiple instance learning with center embeddings for histopathology classification," in *Proc. Med. Image Comput. Comput. Assist. Intervention: 23rd Int. Conf.*, 2020, pp. 519–528.

[21] A. Myronenko, Z. Xu, D. Yang, H. R. Roth, and D. Xu, "Accounting for dependencies in deep learning based multiple instance learning for whole slide imaging," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2021, pp. 329–338.

[22] 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.*, 2022, pp. 18802–18812.

[23] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[24] W. Tang, S. Huang, X. Zhang, F. Zhou, Y. Zhang, and B. Liu, "Multiple instance learning framework with masked hard instance mining for whole slide image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4078–4087.

[25] J.-G. Yu et al., "Prototypical multiple instance learning for predicting lymph node metastasis of breast cancer from whole-slide pathological images," *Med. Image Anal.*, vol. 85, Apr. 2023, Art. no. 102748.

[26] Y. Wen, W. Liu, A. Weller, B. Raj, and R. Singh, "SphereFace2: Binary classification is all you need for deep face recognition," in *Proc. Int. Conf. Learn. Representations*, 2021.

[27] J. Ye, H. Ni, P. Jin, S. X. Huang, and Y. Xue, "Synthetic augmentation with large-scale unconditional pre-training," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2023, pp. 754–764.

[28] M. Moor et al., "Med-Flamingo: A multimodal medical few-shot learner," *Mach. Learn. Health*, pp. 353–367, 2023.

[29] Q. Zhu, H. Wang, B. Xu, Z. Zhang, W. Shao, and D. Zhang, "Multimodal triplet attention network for brain disease diagnosis," *IEEE Trans. Med. Imag.*, vol. 41, no. 12, pp. 3884–3894, Dec. 2022.

[30] W. Huang, B. Xiao, J. Hu, and X. Bi, "Location-aware transformer network for few-shot medical image segmentation," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2023, pp. 1150–1157.

[31] J. Wu et al., "SeATrans: Learning segmentation-assisted diagnosis model via transformer," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2022, pp. 677–687.

[32] O. Petit, N. Thome, C. Rambour, L. Themyr, T. Collins, and L. Soler, "U-net transformer: Self and cross attention for medical image segmentation," in *Proc. Mach. Learn. Med. Imag.: 12th Int. Workshop, MLMI, Held Conjunction MICCAI*, 2021, pp. 267–276.

[33] J. Zheng, H. Liu, Y. Feng, J. Xu, and L. Zhao, "CASF-Net: Cross-attention and cross-scale fusion network for medical image segmentation," *Comput. Methods Programs Biomed.*, vol. 229, 2023, Art. no. 107307.

[34] 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.

[35] P. Bandi et al., "From detection of individual metastases to classification of lymph node status at the patient level: The camelyon17 challenge," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 550–560, Feb. 2019.

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

[37] M. W. Mahoney and P. Drineas, "CUR matrix decompositions for improved data analysis," *Proc. Nat. Acad. Sci.*, vol. 106, no. 3, pp. 697–702, 2009.

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

[39] X. Wang et al., "Transformer-based unsupervised contrastive learning for histopathological image classification," *Med. Image Anal.*, vol. 81, 2022, Art. no. 102559.

[40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[41] R. J. Chen et al., "Multimodal co-attention transformer for survival prediction in gigapixel whole slide images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 4015–4025.

[42] T. E. Tavolara, M. K. K. Niazi, A. C. Gower, M. Ginese, G. Beamer, and M. N. Gurcan, "Deep learning predicts gene expression as an intermediate data modality to identify susceptibility patterns in Mycobacterium tuberculosis infected Diversity Outbred mice," *EBioMedicine*, vol. 67, May 2021.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[44] L. Fillioux, J. Boyd, M. Vakalopoulou, P.-H. Cournède, and S. Christodoulidis, "Structured state space models for multiple instance learning in digital pathology," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2023, pp. 594–604.