mTREE: Multi-Level Text-Guided Representation End-to-End Learning for Whole Slide Image Analysis

Quan Liu, Department of Computer Science, Vanderbilt University, Nashville, TN

Ruining Deng, Department of Computer Science, Vanderbilt University, Nashville, TN

Can Cui, Department of Computer Science, Vanderbilt University, Nashville, TN

Tianyuan Yao, Department of Computer Science, Vanderbilt University, Nashville, TN

Yuechen Yang, Department of Computer Science, Vanderbilt University, Nashville, TN

Vishwesh Nath, NVIDIA, Santa Clara, CA

Bingshan Li, Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, TN

You Chen, Department of Biomedical Informatics, Vanderbilt University, Nashville, TN

Yucheng Tang, NVIDIA Cooperation, Redmond, WA

Yuankai Huo, Department of Computer Science, Vanderbilt University, Nashville, TN

Yuankai Huo is the corresponding author, e-mail: yuankai.huo@vanderbilt.edu

Abstract

Multi-modal learning adeptly integrates visual and textual data, but its application to histopathology image and text analysis remains challenging, particularly with large, high-resolution images like gigapixel Whole Slide Images (WSIs). Current methods typically rely on manual region labeling or multi-stage learning to assemble local representations (e.g., patch-level) into global features (e.g., slide-level). However, there is no effective way to integrate multi-scale image representations with text data in a seamless end-to-end process. In this study, we introduce Multi-Level Text-Guided Representation End-to-End Learning (mTREE). This novel text-guided approach effectively captures multi-scale WSI representations by utilizing information from accompanying textual pathology information. mTREE innovatively combines – the localization of key areas ("global-to-local") and the development of a WSI-level image-text representation ("local-to-global") – into a unified, end-to-end learning framework. In this model, textual information serves a dual purpose: firstly, functioning as an attention map to accurately identify key areas, and secondly, acting as a conduit for integrating textual features into the comprehensive representation of the image. Our study demonstrates the effectiveness of mTREE through quantitative analyses in two image-related tasks: classification and survival prediction, showcasing its remarkable superiority over baselines. Code and trained models are made available at https://github.com/hrlblab/mTREE.

Introduction

Analyzing Whole Slide Images (WSIs) is a critical aspect of medical imaging research. WSIs are digitized scans of histological samples captured at multiple magnifications, preserving both the overarching view and intricate microscopic details. From a broader perspective, WSIs offer a macroscopic overview of tumor distribution throughout the entire digital slide. This allows for the study of spatial relationships and general tumor traits. At the same time, WSIs enable detailed inspections of cell and tissue structures at the microscopic scale, significantly enhancing diagnostic accuracy and advancing the field of histopathology research [1].

Over recent years, computer vision has become increasingly crucial and successful in analyzing Whole Slide Images (WSIs), tackling the challenges of handling their super high-resolution ($> 10^9$ gigapixels) for tasks like image classification, object detection, and segmentation. Unlike these tasks, our study confronts the unique challenge of performing multi-modal representation learning (involving both image and textual data) for WSIs, focusing on modeling information across multiple scales in both images and text. Textual data in this context can describe both broad and detailed aspects of multi-scale WSIs. A key unresolved question is how to effectively learn representations that encompass both global and local features.

Current methods often rely on manually labeling areas of interest for local representation or using multi-stage learning to merge these local features into a global representation. Yet, these approaches typically fall short of seamlessly integrating multi-scale image representations with text data in an end-to-end process.

In this paper, we propose Multi-Level Text-Guided Representation End-to-End Learning (mTREE), an innovative text-guided method that effectively captures multi-scale image representations through the use of accompanying textual pathology data. mTREE uniquely blends two formerly separate processes – the identification of crucial areas (**"global-to-local"**) and the creation of a WSI-level image-text representation (**"local-to-global"**) – into a unified, end-to-end learning framework.

While text-based clinical records are consistently available, they haven't been fully utilized in multi-modal representation learning for high-resolution images. Our goal is to develop an algorithm that leverages these textual records to guide the selection of diagnostic patches and aggregate WSI representations without the need for manual annotations. We hypothesize that there is an inherent correlation between the text and image domains; for a given WSI, the clinical text can provide criteria for selecting WSI patches and extracting features.

Mathematically, Given a WSI X, considered as a set of image patches $\{x_i\}_{i=1}^N$ where $x_i \in X$, and a corresponding label Y for
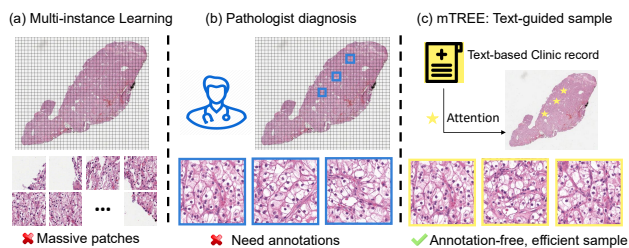
Figure 1. Comparison between multi-instance learning, pathologist diagnosis, and our proposed mTREE. *(a) Traditional multi-instance learning needs to process all patches without patch selection. (b) Pathologists in the diagnosis process focus on the most essential patches selected by manual efforts. (c) Our proposed mTREE generates text-guided attention to sample efficiently without manual annotation.*

WSI X, our approach involves training two mappings. The first is for patch selection $A : \{x_i\}_{i=1}^N \rightarrow \{x_i\}_{i=1}^K, (K \ll N)$, and the second is for feature abstraction $E : \{x_i\}_{i=1}^K \rightarrow Y$. Textual data provides guidance for both mappings: optimizing the selection from the original WSI set and consolidating patch-level features into a comprehensive WSI feature. This dual mapping requires a multi-level approach for text-guided analysis, executed in an end-to-end fashion. In the first mapping, patch selection from $\{x_i\}_{i=1}^N$ depends on each patch's relevance to the final prediction, determined by an attention map that scores each patch's importance. Due to the extensive size of WSIs, the attention map is initially learned on lower-resolution images and then mapped to high-resolution images based on coordinate relationships. This approach allows the model to process only a fraction of patches (K out of N) when $K \ll N$. In the second mapping, the model uses the features extracted from the selected patches, with the text feature identifying and amalgamating the most pertinent features (those with smaller feature distances) into a unified WSI representation.

In summary, our study introduces a text-guided representation learning method aimed at improving efficiency and extracting features from vital image regions, thereby eliminating the confusion caused by unnecessary image patches. Our approach does not require image annotations from pathologists. Instead, it leverages text descriptions from clinical records to guide the learning process of WSI representations at multiple levels in an integrated, end-to-end manner. We have applied our method to various applications, including image classification and survival prediction across multiple WSI datasets, and have compared it with previous approaches based on Multiple Instance Learning (MIL) models.

The key contributions of our work are fourfold:

- We present the first efficient visual-language model for gigapixel WSIs, operating in a seamless end-to-end fashion.
- We utilize text information to optimize learning strategies across multiple levels.
- Our pipeline is weakly supervised at the WSI level, eliminating the need for patch-level annotations from pathologists.
- Our model offers explainability by providing visualizations at different levels, such as attention maps and significant patches.

Related Work
Multi-instance learning

For pathology image analysis, Multi-Instance Learning (MIL) has emerged as a prominent paradigm, offering a robust framework to address challenges associated with the lack of patch-wise annotation of pathological images [17, 18, 19]. Different from the supervised learning method on patches, [20, 21] regard the pathology image as a collection of multiple instances or regions, each potentially harboring critical information for diagnostic or prognostic purposes. This approach allows the model to operate on bags of instances for weakly-supervised learning [22]. MIL has demonstrated its efficacy in capturing nuanced spatial relationships [23] and patterns [24] within pathology images, accommodating the inherently diverse nature of tissue structures [25, 26] and cellular compositions [27]. Upon the set-based concept, Ilse et al. [15] apply the attention mechanism to Whole Slide Images (WSIs). In a similar vein, Yao et al. [17] integrated attention-based MIL into clustered phenotypes, yielding promising outcomes. Furthermore, [28] validated the MIL performs well on large-scale WSI datasets.

Attention sampling

Performing analysis on large images, attention sampling [6] has emerged as a powerful technique to efficiently process extensive visual datasets by selectively focusing computational resources on regions of interest. Attention sampling aims to address the challenges posed by the sheer scale of high-resolution images, where the majority of the content may be irrelevant to the specific task at hand [29]. Attention sampling leverages mechanisms inspired by human visual attention [42], directing computational resources toward salient regions while bypassing less informative areas. Notable approaches include the integration of attention mechanisms within convolutional neural networks (CNNs) [30, 31, 32] to dynamically weigh the importance of different image regions. Additionally, attention sampling strategies, such as region-based methods and attention-guided sampling, have been proposed to enhance computational efficiency in tasks such as object detection [33, 34], image classification [35, 36], and segmentation [37, 38].

Visual language model in WSI analysis

The integration of visual language models (VLM) has emerged as a cutting-edge approach, revolutionizing the interpretation of large-scale pathological images. Visual language models combine the strengths of natural language processing (NLP) and computer vision, enabling a comprehensive understanding of complex visual information assisted by knowledge from multiple domains. Unlike fine-tuning, VLM is based on prompt prediction in the template, as seen in CLIP [14] and CoOp [40]. The trained language model has a strong capability in knowledge and zero-shot learning [41]. By leveraging pre-trained language models such as BERT [39] and adapting them to the unique challenges of WSI, researchers have achieved remarkable strides in capturing contextual relationships and hierarchical structures within pathology images [11]. These models empower the extraction of meaningful features and semantic understanding, enhancing the interpretability of WSIs for tasks such as image classification, tumor detection, and prognosis prediction [12].

Methods

As illustrated in 2, our goal is to learn the multi-level mapping from gigapixel WSI X to the representation R guided by text prompt T. To better describe the model learning on gigapixel images, we regard high-resolution images (base layer in image pyramid in 2) as a collection of image patches $\{x_i\}_{i=1}^N$ ($x_i \in X$). To efficiently learn the WSI representation, the global-level alignment learns an attention map M from WSI in low resolution W_{low}. The global alignment on M is between image attention map M_w and text attention map M_t. The image patch collection $\{x_i\}_{i=1}^N$ is ranked by the attention score in attention map M. Top K image patches $\{x_i\}_{i=1}^K$ with highest attention score are selected by the sampler. Image patch features $\{I_i\}_{i=1}^K$ is abstracted by image encoder E from top K image patches $E(x_i)$. The image feature I_i is selected by the cosine similarity with T_0. The WSI representation R is aggregated by the I_i weighted by distance $I_0 \times T_0$. At the core of our approach is the idea of multi-level alignment from supervision contained in natural language and the end-to-end training manner to coordinate the region localization and feature extraction.

Global alignment

The attention map can, in theory, learn the significance of the region playing a role in the final prediction. However, for the high-resolution WSI normally in size of $100,000 \times 100,000$, a gigapixel attention map is too large for attention model capacity to learning. Considering the histopathology image composed of regions of different tissue types, the region distribution of WSI makes attention map learning even harder to match the target prediction. Since learning the attention map for the high-resolution image is not valid, we can learn the attention map in low-resolution WSI to reduce the training burden and localize the rough essential region. To further optimize the attention map, the global-level alignment is between text feature T_0 and low-resolution WSI X_{low}. The attention mappings on the text side $A_t: T_0 \rightarrow M_t$ and WSI side $A_w: X_{low} \rightarrow M_w$ provides the WSI attention map M_w and the text attention map M_t. The mapping A_t is achieved by a projection head, composed of multiple fully connected layers. We use MSELoss as attention loss $\mathscr{L}_{Attn}$ to match the M_w with M_t as expressed in 1.

$$\mathscr{L}_{Attn} = \frac{1}{N}\sum_{i=1}^N (M_w^i - M_t^i)^2 \tag{1}$$

To prevent the attention from collapsing and to encourage non-zero values in the attention map, the sparsity-inducing regulation term $\mathscr{L}_{Sparse}$ is applied to both M_t and M_w, penalizing overly sparse attention maps following 2.

$$\mathscr{L}_{Sparse} = \frac{1}{N}\sum_{i=1}^N |M_t| + \frac{1}{N}\sum_{i=1}^N |M_w| \tag{2}$$

To keep the low-resolution WSI in the same size from the same magnification, the original image is center-cropped before being fed into the attention model. To prevent the attention from focusing only on the cropped image boundary, the $\mathscr{L}_{Boundary}$ is defined by gradients of the attention map with respect to the x-axis (G_x) and y-axis (G_y). The $\mathscr{L}_{Boundary}$ is expressed in 3 where $G_x(M_w)_i$ is the i-th element in the gradient of M_w with respect to x-axis and

$G_y(M_w)_i$ is the i-th element in the gradient of M_w with respect to y-axis.

$$\mathscr{L}_{Boundary} = \lambda \left(\sum_{i=1}^N |G_x(M_w)_i| + \sum_{i=1}^N |G_y(M_w)_i| \right) \tag{3}$$

The full loss function for attention map global alignment is:

$$\mathscr{L}_{Global} = \mathscr{L}_{Attn} + \mathscr{L}_{Sparse} + \mathscr{L}_{Boundary} \tag{4}$$

Local alignment

The attention map M provides the criterion score for image patch significance. The top K patch selected from $\{x_i\}_{i=1}^N$ provides the most essential patches for the prediction. As shown in 3 lower panel, essential patch collection $\{x_i\}_{i=1}^K$ selected by the sampler is fed into the image encoder to generate the image feature $\{I_i\}_{i=1}^K$. Because the image patches and text are both encoded into feature space, the similarity of text feature T_0 with each image feature candidate I_i can be computed and ranked based on 5

$$\text{Cosine Similarity}(\{I_i\}_{i=1}^K, T_0) = \left\{ \frac{I_i \cdot T_0}{\|I_i\|\|T_0\|} \right\}_{i=1}^K \tag{5}$$

Image feature $\{I_i\}_{i=1}^J$ are selected based on the cosine similarity with the text feature T_0. The final representation for WSI is aggregated with the selected image features, as shown in 6.

$$R = \sum_{i=1}^J \left(\frac{I_i \cdot T_0}{\|I_i\|\|T_0\|} \cdot I_i \right) \tag{6}$$

End-to-end training with image sampler

The multi-level alignment under text guidance is built in an end-to-end manner to provide the advantages from two aspects: (1) Coordinate the optimization on the attention map and feature selection and aggregation. (2) Provide the backpropagation path from the final prediction to the attention map for the sampler. Representation learning with attention maps often involves a two-stage training process, where the first stage focuses on learning a representation, and the second stage incorporates attention mechanisms as in [2, 7, 8, 9]. Insufficient end-to-end optimization potentially leads to suboptimal integration of attention and representation learning. To better describe the end-to-end manner, the pseudocode of mTREE is provided.

Algorithm 1 Pseudocode for mTREE implementation

1: **Input:** Gigapixel WSI X, Text Prompt T
2: **Output:** WSI Representation R
3: Learn attention map M_w from low-resolution X_{low}
4: Align image attention map M_w and text attention map M_t with attention loss
5: Rank image patches $\{x_i\}_{i=1}^N$ using attention score in M_w
6: Select top K image patches $\{x_i\}_{i=1}^K$ with highest attention score
7: Extract image features $\{I_i\}_{i=1}^K$ with image encoder E
8: Select image feature I_i based on cosine similarity with T_0
9: Aggregate WSI representation R using I_i weighted by distance $I_0 \times T_0$
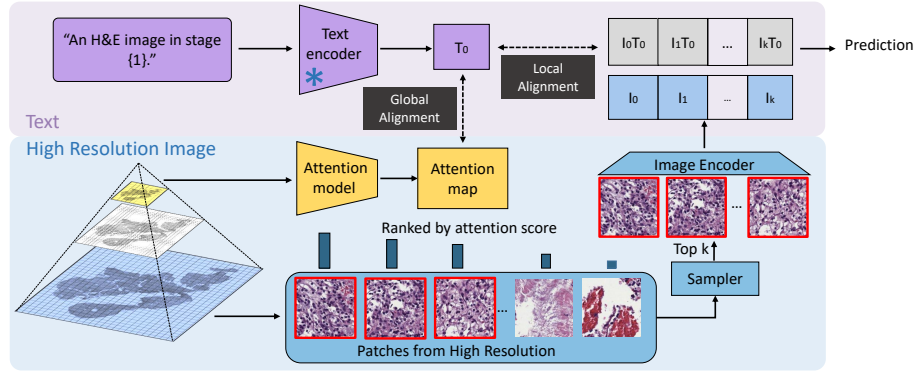

Figure 2. *This figure demonstrates the proposed mTREE pipeline. The upper panel shows the text process flow. The text encoder is frozen with pre-trained weights. Text feature T_0 is used for global alignment and alignment of image patch features. The lower panel shows the WSI analytic flow. The attention model learns an attention map from the WSI in low resolution. The attention map aligns with the text feature T_0. The image patches tiled up from high-resolution WSI are ranked by attention score. The image features $I_0, I_1...I_k$ abstracted from image patches with higher attention scores are aggregated with text feature T_0.*

Method	Inputs	Patch #	Image encoder	Acc (grade)	C-Index (survival)
PathomicFusion [5]	diagnostic regions (**manual**)	20	ResNet-50	N/A	63.1
AttenDeepMIL [15]	diagnostic regions (**manual**)	20	ResNet-50	60.9	61.5
AttenSample [13]	raw WSI (**automatic**)	1	ConvNet	49.1	55.4
AttenDeepMIL [15]	raw WSI (**automatic**)	100	ResNet-50	51.0	58.8
CLAM [16]	raw WSI (**automatic**)	>5000	ResNet-50	57.5	60.1
mTREE (Ours)	raw WSI (**automatic**)	10	ResNet-50	63.1	63.2
mTREE (Ours)	raw WSI (**automatic**)	20	ResNet-50	**64.7**	**65.1**

Cancer grade classification and survival prediction results on KIRC dataset. Acc represents the accuracy of grade classification, while the C-Index evaluates the survival prediction performance.
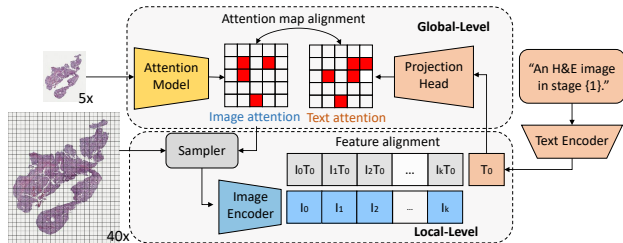

Figure 3. *This figure presents the principle of multi-level text guidance. Global-level text guidance (upper panel) aligns the attention map from images and text. Image attention map is learned from low-resolution WSI, while text attention is projected from the text feature T_0. Local-level text guidance (lower panel) performs patch selection by computing the cosine similarity distance to the text feature T_0 and aggregates features from both image and text.*

Experiments
Data description

To substantiate our proposed text-guided representation learning approach, integrating histological and text features, we sourced glioma and clear cell renal cell carcinoma data from the TCGA, a comprehensive cancer data consortium housing paired high-throughput text in clinic records and diagnostic whole slide images. This dataset is enriched with ground-truth survival outcomes and histologic grade labels. For both the TCGA-KIRC (519 WSIs) and TCGA-GBMLGG (1589 WSIs) projects, region-of-interests (ROIs) from diagnostic slides are provided by [5].

For clear cell renal cell carcinoma in the TCGA-KIRC project, 512×512 ROIs from diagnostic whole slide images are provided as the diagnostic region. This yielded 3 ROIs per patient (512×512 at 40× magnification) for 417 patients, resulting in a total of 1251 images. For the TCGA-GBMLGG project, 1024×1024 region-of-interests (ROIs) from diagnostic slides are leveraged. The WSI data is publicly available on the TCGA database [10].

Data preprocessing

Both WSI image and text data require preprocessing before feature extraction. The preprocessing for both datasets follows the same strategy.

WSI image data. The input image data for our pipeline is provided from two levels: low-resolution images and high-resolution images. The low-resolution images are from 5x magnifications in the WSI pyramid structure. To ensure low-resolution images in the same size and scale, we center-crop the $5,000 \times 5,000$ patches from the low-resolution images. All $5,000 \times 5,000$ patches are then resized to 500×500.

Text data. Follow the design in [11, 12], text information is composed of templates and prompts curated from the clinical records. In our experiments, the paragraph related to "survival time" and the "cancer grade" are used as text information.

Network architectures

We adopt the representation learning flow from the "low-resolution" to "high-resolution", as proposed in Attention-

Sampling [13], which has shown impressive results on megapixel image analysis. Based on the "low-resolution" to "high-resolution" strategy, our mTREE pipeline is composed of three parts: image feature analysis models, text feature analysis models, and alignment blocks between image and text.

Image feature analysis model. Attention model for low-resolution WSI is a Convolution Network (ConvNet) with four convolution layers. We use 3×3 convolution kernel and the channel number of four convolution layers is [8, 16, 32, 1]. The Sampler ranks image patches and selects the top K patches. No learnable parameters in the sampler. ResNet-50 with ImageNet pretrained weight is used as the image encoder.

Text feature analysis model. The text encoder is the pretrained ViT-B/32 used in CLIP [14]. The text encoder is frozen in the training process.

Alignment block. The global alignment block includes a projection head composed of two sequential convolution layers to project text feature to text attention map. The local alignment between text features and image features is based on the cosine similarity matrix in the shape of $1 \times K$.

Training details

We apply our mTREE to two TCGA datasets (KIRC and GBMLGG) on two downstream tasks: grade classification and survival prediction.

Tasks. The KIRC dataset has three grades (Stage I, Stage II, and Stage III) for grade classification. The patient's overall survival time in month is used as the label for the survival prediction task. The GBMLGG dataset also has three grades (2, 3, 4) for grade classification. The "Time to last follow-up or death (Month)" is used as the label for survival prediction in the GBMLGG dataset.

Hyper-parameters. The three most important parameters are evaluated for our proposed mTREE. The first one is the size of the attention map learned in global alignment. Tuned by the projection head channel number and attention model structure, we evaluated attention map size in 123×123 and 246×246. The second parameter is the sample number from the attention map (K in 5). Based on the number of patches from the diagnostic region provided by [5], approximately 20 patches for each WSI, we evaluated sample number in set 5, 10 20, 50. The third parameter is the sample number from the image patch features (J in 6). According to $K \in \{5, 10, 20, 50\}$, we sampled $J \in \{2, 5, 10, 20\}$.

Metrics. The metric used for grade classification tasks is accuracy (ACC). The ACC evaluates the WSI representation performance on class prediction tasks with discrete labels.

The metric for the survival prediction task is the C-Index. It quantifies the concordance between predicted and observed survival times, with a higher C-index indicating improved predictive accuracy.

Baseline experiments for comparison

To validate the advantages of the proposed mTREE pipeline, the baseline experiments comparisons are compared from three aspects: (1) MIL-based model: we use AttenDeepMIL [15] as a general MIL-based model and CLAM [16], designed specifically for WSI analysis. (2) Attention Sampling [13], and (3) PathomicFusion [5].

AttenDeepMIL. The implementation of AttenDeepMIL follows the settings in [15]. ResNet-50 with ImageNet-pretrained weights is used as the image encoder. For both the KIRC dataset and the GBMLGG dataset, two patch selection strategies are evaluated: (1) diagnostic region (DR in), and (2) tiled-up image patches from 40x WSI (Origin in).

CLAM. In [16], CLAM provides the implementation of MIL for classification. For a fair comparison, the image encoder is ResNet-50, similar to other baseline methods. CLAM processes all image patches from WSI, except the background patches, normally more than 5,000 patches for a WSI.

AttenSample. Attention sampling input set has an image in high resolution ($1,500 \times 1,500$) and a low-resolution image rescaled by a ratio of 0.1 (150×150). For the TCGA dataset, the high-resolution image is from a 5x magnification WSI and center-cropped in size of $5,000 \times 5,000$. The low-resolution image is rescaled by a ratio of 0.1 to 500×500.

PathomicFusion. In [5], PathomicFusion is a multi-modal fusion method incorporated with image, genomics, and cell graph data. In our experiments, only image data is used for performance comparison.

Results

Following the experiment settings, two datasets TCGA-KIRC and TCGA-GBMLGG on grade classification and survival prediction are discussed. To optimize the hyperparameter settings, an ablation study of different parameters is evaluated on the survival prediction task on the GBMLGG dataset. To provide the model with explainability, the visualization of the attention map and selected diagnostic region by mTREE is presented.

KIRC

In 1, we compare the performance of the proposed mTREE with baselines on the TCGA-KIRC dataset for the grade classification task. It is observed that the performance of MIL-based methods improves with an increasing sample number from the WSI patch collection. When processing all patches from the WSI, CLAM achieves an accuracy of 57.5%. Remarkably, our proposed mTREE outperforms, achieving a superior accuracy of 64.7% with just 20 sampled patches from the WSI, surpassing even AttenDeepMIL with a diagnostic region.

Moving on to the survival prediction task in 1, we observe a performance trend similar to the grade classification task. MIL-based methods exhibit better performance with a diagnostic region compared to original image patches. The best performance from the MIL baseline achieves a C-Index of 0.631 when trained on the diagnostic region. Notably, our proposed mTREE demonstrates superior performance (C-Index 0.651) over the baselines, utilizing both original image patches and diagnostic regions.

GBMLGG

In 2, we present the prediction performance of the proposed mTREE and baselines on the grade classification task. Similar to the performance comparison in the grade classification task, MIL-based methods exhibit better performance with a diagnostic region compared to original image patches. The best performance from the MIL baseline achieves an accuracy of 78.8% when trained on the diagnostic region. Notably, our proposed mTREE outperforms the baselines, achieving a superior accuracy

Method	Inputs	Patch #	Image encoder	Acc (grade)	C-Index (survival)
PathomicFusion [5]	diagnostic regions (**manual**)	20	ResNet-50	N/A	**72.4**
AttenDeepMIL [15]	diagnostic regions (**manual**)	20	ResNet-50	78.8	71.4
AttenSample [13]	raw WSI (**automatic**)	1	ConvNet	70.4	65.4
AttenDeepMIL [15]	raw WSI (**automatic**)	100	ResNet-50	70.6	63.6
CLAM [16]	raw WSI (**automatic**)	>5000	ResNet-50	75.3	65.7
mTREE (Ours)	raw WSI (**automatic**)	10	ResNet-50	76.5	69.0
mTREE (Ours)	raw WSI (**automatic**)	20	ResNet-50	**79.6**	70.1

Cancer grade classification and survival prediction results on GBMLGG dataset. 'Acc' represents the accuracy of grade classification, while the C-Index evaluates the survival prediction performance.

Method	Global align	Local align	KIRC ACC	KIRC C-Index	GBMLGG ACC	GBMLGG C-Index
			49.1	55.4	70.4	65.4
mTREE	✓		51.0	56.3	70.7	65.2
	✓	✓	**64.7**	**65.1**	**79.6**	**70.1**

Ablation study for multi-level text alignments is shown in this table. The accuracy of grade classification (ACC) and the survival prediction performance (C-Index) are presented.
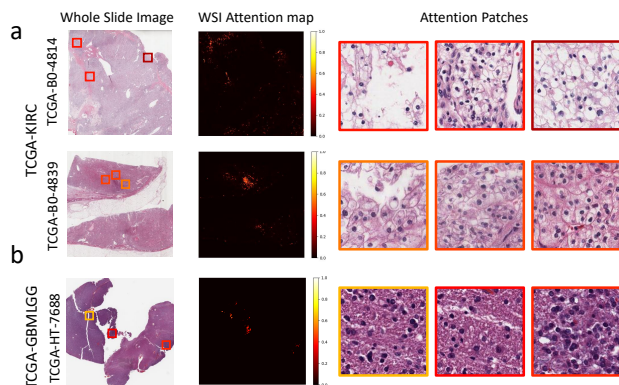

Figure 4. *This figure presents the visualization of WSI-level attention and the automatically derived diagnosis patches. For WSIs in the TCGA-KIRC dataset and TCGA-GBMLGG dataset, the attention map (middle panels) is learned from WSI (left panels), highlighting essential tissue regions. Essential image patches (right panels) are selected according to the attention score. The image boundary color indicates the according attention score.*

of 79.6% with both original image patches and the diagnostic region.

In 2, we present the prediction performance of the proposed mTREE and baselines for the survival prediction task. The MIL-based method exhibits better performance with a diagnostic region than with original image patches. The best performance from the MIL baseline achieves a C-Index of 0.724 when trained with the diagnostic region. Notably, our proposed mTREE outperforms the baselines, achieving a better performance (C-Index 0.701) with original image patches.

Evaluation for multi-level text alignments

In this section, we compare the performance of the proposed mTREE with and without global and local alignment. The results are presented in 3. From the performance shown in 3, both global and local alignment contribute to performance improvement in all four tasks. However, global alignment provides a limited contribution (as shown in the second row of 3). From another

perspective, the coordination between local and global alignment underscores the advantages of an end-to-end training approach.

Visualization

The attention map (4 middle) obtained through global alignment serves as a crucial tool for improving the interpretability of mTREE in the context of whole-slide image (WSI) analysis. The heightened intensity in the attention map accentuates key regions within the WSI that significantly contribute to the final prediction. In the domain of weakly-supervised learning, these bright regions indicate areas of essential diagnostic relevance. To enhance human understanding, image patches identified as having high diagnostic importance are presented in a zoomed-in view (4 right). The color of the image boundary indicates the corresponding attention score. Patches with higher attention scores are deemed more important for the final prediction.

Conclusion

This paper introduces a novel text-guided representation learning pipeline designed for the efficient processing of Whole-Slide Images (WSIs). Our proposed model, mTREE, seamlessly integrates textual pathology information with WSI features on multiple levels, enabling a comprehensive understanding of the underlying data. Trained in an end-to-end manner, mTREE demonstrates superior performance in both classification and survival prediction across two distinct WSI datasets. Notably, the model exhibits explainability, as evidenced by its capability to visualize attention maps at both the WSI level and specific patches with high diagnostic importance. This fusion of accuracy and interpretability underscores the effectiveness of mTREE in the domain of WSI analysis.

References

[1] L. Pantanowitz *et al.*, "Review of the current state of whole slide imaging in pathology," *Journal of pathology informatics*, vol. 2, no. 1, p. 36, 2011.

[2] Y. Xu *et al.*, "Deep learning of feature representation with multiple instance learning for medical image analysis," in *ICASSP*, 2014, pp. 1626–1630.

[3] K. He *et al.*, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[4] J. Fan *et al.*, "Speed Up Object Detection on Gigapixel-Level Images With Patch Arrangement," in *CVPR*, 2022, pp. 4653–4661.

[5] R. J. Chen *et al.*, "Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis," *IEEE Transactions on Medical Imaging*, vol. 41, no. 4, pp. 757–770, 2020.

[6] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015, pp. 2048–2057.

[7] A. Vaswani *et al.*, "Attention is all you need," in *NIPS*, 2017, pp. 5998–6008.

[8] V. Mnih *et al.*, "Learning to focus: A machine learning approach to visual attention," *arXiv preprint arXiv:1409.1259*, 2014.

[9] J. Yue-Hei Ng *et al.*, "Beyond short snippets: Deep networks for video classification," in *CVPR*, 2015, pp. 4694–4702.

[10] K. Tomczak *et al.*, "Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge," *Contemporary Oncology/Wspolczesna Onkologia*, vol. 2015, no. 1, pp. 68–77, 2015.

[11] Z. Huang *et al.*, "A visual–language foundation model for pathology image analysis using medical Twitter," *Nature Medicine*, vol. 29, no. 9, pp. 2307–2316, 2023.

[12] M. Y. Lu *et al.*, "Towards a visual-language foundation model for computational pathology," *arXiv preprint arXiv:2307.12914*, 2023.

[13] A. Katharopoulos and F. Fleuret, "Processing megapixel images with deep attention-sampling models," in *ICML*, 2019, pp. 3282–3291.

[14] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[15] M. Ilse *et al.*, "Attention-based deep multiple instance learning," in *ICML*, 2018, pp. 2127–2136.

[16] M. Y. Lu *et al.*, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 555–570, 2021.

[17] J. Yao *et al.*, "Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks," *Medical Image Analysis*, vol. 65, p. 101789, 2020.

[18] L. Hou *et al.*, "Patch-based convolutional neural network for whole slide tissue image classification," in *CVPR*, 2016, pp. 2424–2433.

[19] M. Y. Lu *et al.*, "AI-based pathology predicts origins for cancers of unknown primary," *Nature*, vol. 594, no. 7861, pp. 106–110, 2021.

[20] H. Edwards and A. Storkey, "Towards a neural statistician," *arXiv preprint arXiv:1606.02185*, 2016.

[21] M. Zaheer *et al.*, "Deep sets," *NIPS*, vol. 30, 2017.

[22] M.-A. Carbonneau *et al.*, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognition*, vol. 77, pp. 329–353, 2018.

[23] Y. Zhao *et al.*, "Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution," in *CVPR*, 2020, pp. 4837–4846.

[24] F. Wu *et al.*, "DeepGCNMIL: Multi-head Attention Guided Multi-Instance Learning Approach for Whole-Slide Images Survival Analysis Using Graph Convolutional Networks," in *ICMLC*, 2022, pp. 67–73.

[25] Z. Shao *et al.*, "LNPL-MIL: Learning from Noisy Pseudo Labels for Promoting Multiple Instance Learning in Whole Slide Image," in *ICCV*, 2023, pp. 21495–21505.

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

[27] O. Z. Kraus *et al.*, "Classifying and segmenting microscopy images with deep multiple instance learning," *Bioinformatics*, vol. 32, no. 12, pp. i52–i59, 2016.

[28] G. Campanella *et al.*, "Terabyte-scale deep multiple instance learning for classification and localization in pathology," *arXiv preprint arXiv:1805.06983*, 2018.

[29] H. Zheng *et al.*, "Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition," in *CVPR*, 2019, pp. 5012–5021.

[30] C. Wang *et al.*, "Attention Weighted Local Descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[31] Z. Xue *et al.*, "Progress prediction of Parkinson's disease based on graph wavelet transform and attention weighted random forest," *Expert Systems with Applications*, vol. 203, p. 117483, 2022.

[32] H. Wang *et al.*, "Wide weighted attention multi-scale network for accurate MR image super-resolution," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 962–975, 2021.

[33] Y. Cheng *et al.*, "Weighted feature fusion and attention mechanism for object detection," *Journal of Electronic Imaging*, vol. 30, no. 2, p. 023015, 2021.

[34] H. Li *et al.*, "Motion guided attention for video salient object detection," in *ICCV*, 2019, pp. 7274–7283.

[35] Y. Dong *et al.*, "Weighted feature fusion of convolutional neural network and graph attention network for hyperspectral image classification," *IEEE Transactions on Image Processing*, vol. 31, pp. 1559–1572, 2022.

[36] F. Wang *et al.*, "Residual attention network for image classification," in *CVPR*, 2017, pp. 3156–3164.

[37] V. Kulharia *et al.*, "Box2seg: Attention weighted loss and discriminative feature learning for weakly supervised segmentation," in *ECCV*, 2020, pp. 290–308.

[38] X. Shi *et al.*, "Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation," in *ECCV*, 2022, pp. 151–168.

[39] J. Devlin *et al.*, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[40] K. Zhou *et al.*, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[41] T. Brown *et al.*, "Language models are few-shot learners," *NIPS*, vol. 33, pp. 1877–1901, 2020.

[42] M. Hassanin *et al.*, "Visual attention methods in deep learning: An in-depth survey," *arXiv preprint arXiv:2204.07756*, 2022.

Author Biography

Quan Liu is a PhD in Computer Science at Vanderbilt University, working with Dr. Yuankai Huo. He received his Bachelor's degree from Huazhong University of Science and Technology. His work has focused on the medical image analysis with computer vision techniques.