Transformer-Based Video-Structure Multi-Instance Learning for Whole Slide Image Classification

Yingfan Ma[1 2], **Xiaoyuan Luo**[1 2], **Kexue Fu**[3], **Manning Wang**[1 2 *]

[1]Digital Medical Research Center, School of Basic Medical Sciences, Fudan University, Shanghai 200032, China
[2]Shanghai Key Laboratory of Medical Imaging Computing and Computer Assisted Intervention, Shanghai 200032, China
[3]Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Jinan, China
22211010089@m.fudan.edu.cn, {19111010030, mnwang}@fudan.edu.cn, fukx@sdas.org

Abstract

Pathological images play a vital role in clinical cancer diagnosis. Computer-aided diagnosis utilized on digital Whole Slide Images (WSIs) has been widely studied. The major challenge of using deep learning models for WSI analysis is the huge size of WSI images and existing methods struggle between end-to-end learning and proper modeling of contextual information. Most state-of-the-art methods utilize a two-stage strategy, in which they use a pre-trained model to extract features of small patches cut from a WSI and then input these features into a classification model. These methods can not perform end-to-end learning and consider contextual information at the same time. To solve this problem, we propose a framework that models a WSI as a pathologist's observing video and utilizes Transformer to process video clips with a divide-and-conquer strategy, which helps achieve both context-awareness and end-to-end learning. Extensive experiments on three public WSI datasets show that our proposed method outperforms existing SOTA methods in both WSI classification and positive region detection.

Introduction

The primary challenge in analyzing Whole Slide Images (WSIs) using deep learning techniques is their huge size. Due to the memory constraint, neural networks cannot directly take the whole original WSI as input. To address this challenge, WSIs are usually tiled into small patches as network input (Lu et al. 2021a; Mahmood et al. 2019; Lu et al. 2022). However, in WSI classification tasks, labels are often only available for each slide instead of each individual patch, which makes fully supervised methods on the patch level infeasible. The lack of patch-level labels is due to the fact that a single slide can consist of hundreds to thousands of patches, and patch-level labeling is very costly and time-consuming. WSI classification is often solved as a weakly-supervised multiple instance learning (MIL) problem (Rony et al. 2019; Chen et al. 2020a; Lu et al. 2019, 2021b; Yuan et al. 2022), in which WSI is noted as 'bag' with the patches cut from it as 'instances', and only has slide-level label.

Due to the large number of patches in a WSI, directly inputting all patches of a bag into deep neural networks would lead to memory overload, making end-to-end learning on the slide-level difficult. Some researchers randomly sample a few patches from a slide as model input and train the model with slide labels. However, this approach may introduce a noise label issue, where for example an all-negative subset may be sampled from a positive slide (Chikontwe et al. 2020). As a result, mainstream MIL methods consist of two stages: 1) Extracting patch features using unsupervised pre-trained models and then inputting all the patch features of a slide into the classification model. 2) Performing global aggregation operations on the input unordered instance features for slide classification (Ilse, Tomczak, and Welling 2018; Li, Li, and Eliceiri 2021; Hashimoto et al. 2020; Lu et al. 2021c; Rymarczyk et al. 2022), as depicted in Figure 1 (a). These methods operate under the assumption that all instances within a bag are independently distributed, overlooking the inter-instance correlations present in the diagnostic process (Shao et al. 2021). However, pathologists often consider contextual information between different regions when

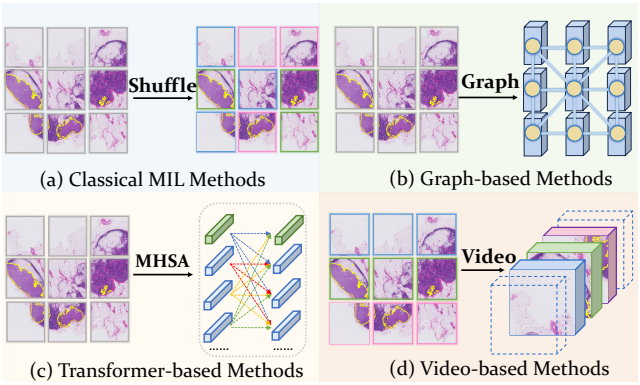

Figure 1: Motivation of our method. (a) Classical MIL Methods. Two-stage methods include encoding unordered instances into features and aggregating features. (b) Graph-based Methods. GNN models feature as nodes and establish edges between neighboring nodes to leverage graph convolution. (c) Transformer-based Methods. Calculate the relativity between features, retaining the spatial and morphological information of WSI. (d) Video-based Method. Construct a series of synthetic video clips from a WSI.

*Corresponding author.
Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

making diagnostic decisions. Some recent methods incorporate contextual information during the aggregation phase to enhance model performance, using Graph Neural Networks (GNN) (Zhao et al. 2020; Tu et al. 2019; Chen et al. 2021) or Transformers (Shao et al. 2021; Li et al. 2021; Chen et al. 2022; Chen and Krishnan 2022; Zhang et al. 2023; Xiang and Zhang 2022). These models utilize the power of graph networks and self-attention mechanisms to achieve context-aware capabilities. As illustrated in Figure 1 (b), GNN models features as nodes and establishes edges between neighboring nodes to leverage graph convolution, to aggregate contextual information. In Figure 1(c), the Transformer utilizes self-attention mechanism to explore the relationships between features in different regions.

The GNN and Transformer-based methods can explore contextual information for better classification. However, given the huge number of patches contained within a WSI, simultaneously extracting patch features and modeling the contextual relationships between all features also leads to memory overload, thus precluding the possibility of end-to-end training. To address this problem, these methods employ a pre-trained model to extract patch features pre-trained with tasks such as self-supervised tasks, which are different from the target task. These differences cause domain gaps and improper inductive bias, limiting the performance of these methods on downstream tasks.

In this paper, we introduce a novel Transformer-based end-to-end framework for WSI classification. We construct a series of synthetic video clips from a WSI which makes it possible to realize end-to-end training and model contextual relationship between patches. Instead of simulating the pathologist's process of observing a WSI which may include a lot of switches of focusing area, zoom in and out, we select patches with the same scale to form video clips which is practical for computers. This formulation helps achieve end-to-end training and considers neighboring semantic information, resulting in promising performance. Specifically, as illustrated in Figure 2, we employ a divide-and-conquer strategy for all patches in a bag by organizing spatially consecutive patches into synthetic video clips. These clips are then input into Transformer-based models to explore spatial contextual information, and parameters are shared among all the Transformer-based models to reduce spatial complexity to realize end-to-end training. Within each video clip, the Transformer utilizes self-attention mechanisms exclusively to the internal patches, thus exploring inter-patch correlations. Each video clip employs class-specific class tokens to learn semantics for different categories, and the class tokens across all video clips are aggregated for bag-level classification. As depicted in Figure 3, by computing the dot product between the class-specific tokens of video clips and instance features, high-quality pseudo-labels can be obtained and then used to train an instance classifier, thereby achieving instance-level classification.

We extensively evaluated our method on three public WSI datasets: CAMELYON16, PANDA, and TCGA-NSCLC. Our method outperformed SOTA baseline methods in both bag and instance classification tasks in all datasets. The main contribution of this paper can be summarized as follows:

- We propose to model a WSI as synthetic video clips, making it simple for computers to simulate pathologists' observing process, which preserves the local spatial relationships between patches, facilitating context-aware learning. Furthermore, we input each video clip into Transformer-based model and share parameters across all these models to reduce the spatial complexity.
- To train instance classifier with high-quality pseudo-labels, we introduce the class-specific clip token to represent the semantic information within each video clip. We generate pseudo-labels for patches by utilizing the clip tokens to train an instance-level classifier based on the Transformer model.
- Extensive experiments on three public datasets show that our proposed method outperforms existing SOTA methods in both bag and instance classification.

Related Work

Two-stage MIL

Most existing WSI classification methods explore a two-stage strategy: first extracting features from original patches using a pre-trained feature extractor and then aggregating these patch features for bag classification. In the feature extraction stage, the feature extractor is either pre-trained on ImageNet (Shao et al. 2021) or unlabeled pathological images(Li, Li, and Eliceiri 2021; Zhang et al. 2022; Chen and Krishnan 2022; Cai et al. 2023). However, there is a big domain gap between natural images and pathological images, which makes the ImageNet pre-trained feature extractor ineffective in extracting patch features. Pre-trained feature extractor on unlabeled pathological images using self-supervising techniques introduce domain gap, but may inject improper inductive bias into the model, which also limit its performance on downstream classification tasks (Chen et al. 2022; Wang et al. 2022). In the feature aggregation stage, to incorporate global information, attention mechanisms (Ilse, Tomczak, and Welling 2018; Lu et al. 2021b; Li, Li, and Eliceiri 2021) have been predominantly implemented for the weighted aggregation of instances. The limitation of these methods is that the patches are perceived as isolated entities, losing the interactions with contextual information.

GNN is used to model contextual information in WSI (Hou et al. 2022; Zhao et al. 2020; Tu et al. 2019; Chen et al. 2021), where each patch is regarded as a node and edges are constructed between spatially adjacent patches. By using graph convolution, information exchange is realized between nodes in both local neighborhoods and within the whole graph. Some other studies utilize Transformer to explore pairwise correlations between each token in a sequence to enhance aggregation quality (Shao et al. 2021; Li et al. 2021; Chen et al. 2022; Chen and Krishnan 2022; Zhang et al. 2023; Xiang and Zhang 2022). Transformers can adaptively capture spatial characteristics and interaction between patches. The above methods rely on a frozen patch feature extractor, which is not optimal for a specific downstream task. In this paper, we proposed a novel Transformer-based framework that is not only end-to-end trainable but also capable of capturing contextual information.

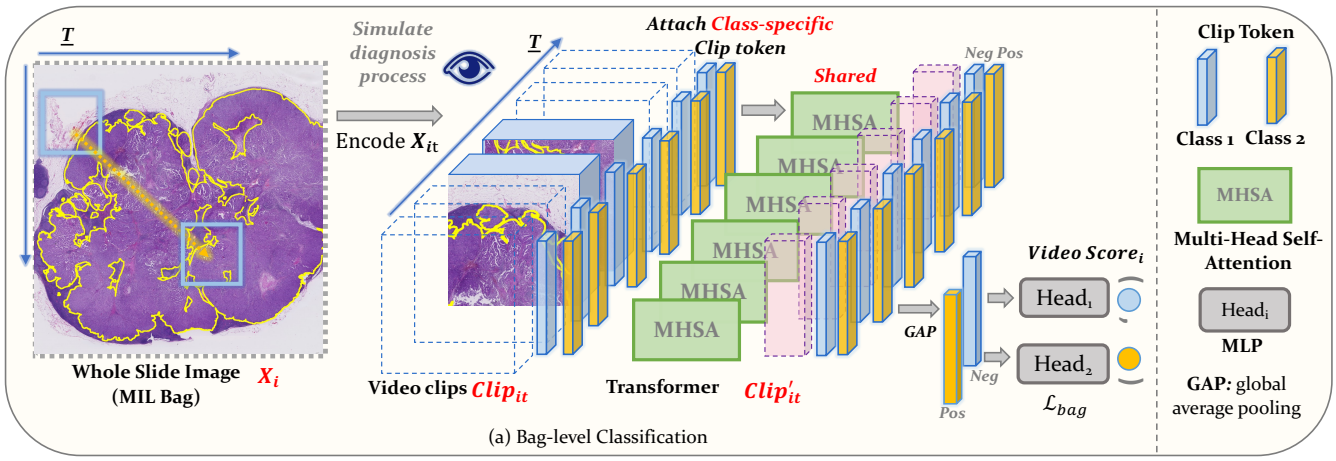

Figure 2: Overview of the proposed VINO framework for bag-level classification. We construct a series of synthetic video clips from a WSI which help achieve end-to-end training and modeling of contextual relationships between patches at the same time.

End-to-end MIL

There are a few end-to-end trainable models for WSI classification and they mainly focus on training an instance-level classifier. For example, Chikontwe (Chikontwe et al. 2020) identified the top-k instances with the highest predicted positive probabilities within positive bags as positive instances and used them together with negative instances from negative bags to train an instance-level classifier. However, this kind of Top-k approach may introduce a significant amount of label noise. The Instance-MIL (Qu et al. 2022; Lin et al. 2022; Luo et al. 2023; Łukasz Struski et al. 2023) is particularly amenable to end-to-end training. By training an instance classifier, it facilitates the classification of individual instances. This approach can maintain a consistent spatial complexity. (Luo et al. 2023) proposed a knowledge distillation framework to generate more accurate positive pseudo labels to train an instance-level classifier in an end-to-end way. (Qu et al. 2022) train end-to-end instance classifier using the attention scores obtained from the teacher network as pseudo labels. While these methods achieve end-to-end training, they process each patch independently and overlook the contextual information of each patch.

Method

Problem Formulation

In the MIL classification problem setting, taking binary classification as an example, given N training WSIs $\boldsymbol{X} = \{X_1, X_2, \ldots, X_N\}$, each WSI is labeled as $Y_i \in \{0, 1\}$, where $i = \{1, 2, \ldots, N\}$. Our goal is to utilize a deep learning model to predict the label of each slide and locate the positive regions. WSI $\boldsymbol{X_i}$ is divided into M_i non-overlapping patches $\{x_{i,j}, | \ j = 1, 2, \ldots, M_i\}$, where M_i denotes the number of patches within $\boldsymbol{X_i}$. It is noted that their corresponding instance-level labels, $\{y_{i,j} \mid j = 1, 2, \ldots, M_i\}$, are unknown. The problem of positive region localization can turn into patch-level classification problems. All patches from $\boldsymbol{X_i}$ form a 'bag', where each patch serves as an 'instance' within this bag. If it contains at least one positive instance, the bag is positive. On the other hand, if all patches are negative, the bag is negative.

The classification of WSIs can be categorized into two tasks: 1) "Bag-level" classification, which focuses on accurately predicting the bag's label. 2) "Instance-level" classification, predicting instances' labels, subsequently enabling the localization of positive regions.

Framework Overview

In order to realize end-to-end training and model contextual relationships between patches at the same time, we propose a novel VINO framework for WSI classification and the localization of positive regions, and the framework is denoted as VINO. Figure 2 illustrates the pipeline of VINO. Given a slide $\boldsymbol{X_i}$, we initially segment it into a series of patches, which act as a series of frames in the video. Then, we select patches according to their positional relationships to form a video clip, where adjacent positions patches are in the same clip. Each video clip is independently inputted into a Multi-Head Self Attention (MHSA) that shares parameters. The MHSA computes the intra-patch correlations and outputs both the clip token representing different categories' semantic information and the interactively computed clip feature, where the clip token denotes the Transformer's class token.

Specifically, we append to each video clip class-specific clip tokens that represent negative and positive categories before inputting them into the MHSA(This is particularly relevant to binary classification problems. For multi-class tasks, there would be n category-specific clip tokens, where n is the number of categories). Subsequently, after processing through MHSA, all the class-specific clip tokens from the video clips are aggregated by global average pooling according to their classes. The aggregated clip tokens are then inputted into a Multi-Layer Perceptron (MLP) Head to obtain the bag-level prediction, which is used to calculate Cross Entropy loss with true bag label for model training.

To address the issues related to preserving contextual re-

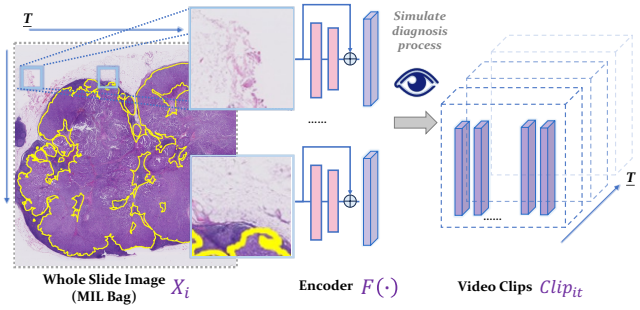

Figure 3: Construction of video clips. A WSI contains a series of video clips according to patch position. ResNet18 is used to extract features for each video clip.

lationships between patches, domain gap, and improper inductive bias, in this paper, we construct a series of synthetic video clips from a WSI that can achieve end-to-end training and the modeling of contextual relationships between patches at the same time. In the setting of the pathological video, such relationships are perceived as temporal sequences. In this study, these pathological instance sub-bags are referred to as "video clips". Clips preserve the local spatial relationships between patches, facilitating context-aware learning. By applying Transformers, we capture local spatial relationships within the WSIs. Additionally, a divide-and-conquer strategy is employed, where only a subset of WSIs are inputted into the parameter-shared Transformer-based model to reduce the spatial complexity of the proposed method to achieve end-to-end training.

Pathology Video Clip Construction

Inspired by the physician diagnosis process, in this paper, we transform the patches in a WSI into video, which preserves the local spatial relationships between patches, facilitating context-aware learning. To simplify this setup, as shown in Figure 3, we assume that the position of the instance within the slide (coordinate relationship) corresponds to the temporal information of the video. Firstly, the slide X_i is divided into non-overlapping size $P \times P$ patches $x_{i,j} \in \mathbb{R}^{H \times W \times C}$, with C being the number of channel and W, H being the dimensional sizes,. We take the feature extraction of patches for every time period T to form a video clip $Clip_{it} \in \mathbb{R}^{T \times dim}$, where dim represents the dimension of the feature output. $Clip$ ensures that internal patches are within a location range, but the internal order we assume is not sequential.

It is noted that since our framework is end-to-end, the entire encoder is trained. Here, we employ ResNet18 as the backbone to extract patch features, denoted as $F(\cdot)$.The resulting video clip, abbreviated as $Clip_{it} \in \mathbb{R}^{T \times dim}$.

$$Clip_{it} = \boldsymbol{F}\left(X_{it}\right), X_{it} \in \mathbb{R}^{T \times H \times W \times C} \tag{1}$$

Transformer-based Bag-level Classification

Each video clip, denoted as $Clip_{it} \in \mathbb{R}^{T \times dim}$, represents the region of focus in Slide X_i over a time span T. Each video clip is fed into a Transformer with shared parameters for exploring inter-patch correlations. Transformer consists of multi-head self-attention ($MHSA$). Furthermore, inspired by MCT-Former (Xu et al. 2022), to achieve multi-class classification, we attach n clip tokens $\in \mathbb{R}^{1 \times dim}$ to each video clip to learn the class-specific semantic information. Each clip token embodies different semantic information for a certain category.

The Transformer (Vaswani et al. 2017), employs three learnable matrices $W^q \in \mathbb{R}^{dim \times dim_k}$, $W^k \in \mathbb{R}^{dim \times dim_k}$, and $W^v \in \mathbb{R}^{dim \times dim_v}$ to get $Q \in \mathbb{R}^{(T+n) \times dim_k}$, $K \in \mathbb{R}^{(T+n) \times dim_k}$, and $V \in \mathbb{R}^{(T+n) \times dim_v}$, resulting in the generation of the attention output. The output vectors encompass not only the current instance's information but also the context surrounding it, calculated as,

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{\dim_k}}\right)V \tag{2}$$

The multi-head attention mechanism enables the model to capture a richer set of information. In the context of this framework, multiple clips are inputted into the $MHSA$ with shared parameters. Here, we denote the number of heads as $heads$, the dimensions as $dim_k = dim_v = \frac{dim}{heads}$. Thus, we train $heads$ of $W^q \in \mathbb{R}^{dim \times \frac{dim}{heads}}$, $W^k \in \mathbb{R}^{dim \times \frac{dim}{heads}}$, and $W^v \in \mathbb{R}^{dim \times \frac{dim}{heads}}$. After obtaining QKV, Transformer then generate $heads$ vectors. The concatenated output following the $MHSA$ process is depicted as follows:

$$Z_h = softmax\left(\frac{Q_h K_h^T}{\sqrt{\dim_k}}\right)V_h, \quad h = 1, \ldots, heads \tag{3}$$

$$MHSA(Q, K, V) = W_o \cdot \text{Concat}(Z_1, \cdots, Z_{heads}) \tag{4}$$

where W^o is the output weight matrix, and the Concat function is utilized to concatenate the attention output vectors from each head into a vector. By processing with the MHSA, the internal self-attention mechanism within the clip is computed, and the result includes clip tokens representing class semantic information and clip features that have undergone interactive computation, denoted as $Clip'_{it}$.

For an n-class classification task, $Clip_{it}$ has corresponding n clip tokens, $Clip\ token_{itc} \in \mathbb{R}^{n \times dim}$, where $c = 1, \ldots, n$, which can learn the semantic information of different categories in the video clip, and a global average pooling is performed on the clip token representing each category. This produces the class-specific feature $Class\ feat \in \mathbb{R}^{n \times dim}$. Then, the feature is then inputted into the classification head for the i^{th} category, denoted MLP_i, to generate video scores $\in \mathbb{R}^{n \times 1}$ as follows.

$$Class\ feat_{ic} = GAP(Clip\ token_{itc}) \tag{5}$$

$$Video\ score_i = softmax\left(MLP_i\left(Class\ feat_{ic}\right)\right) \tag{6}$$

where c is ranged from 1 to c, and GAP denotes the global average pooling. $Video\ score_i$ are used to calculate a cross-entropy loss as follows, providing full supervision information for the class-specific token.

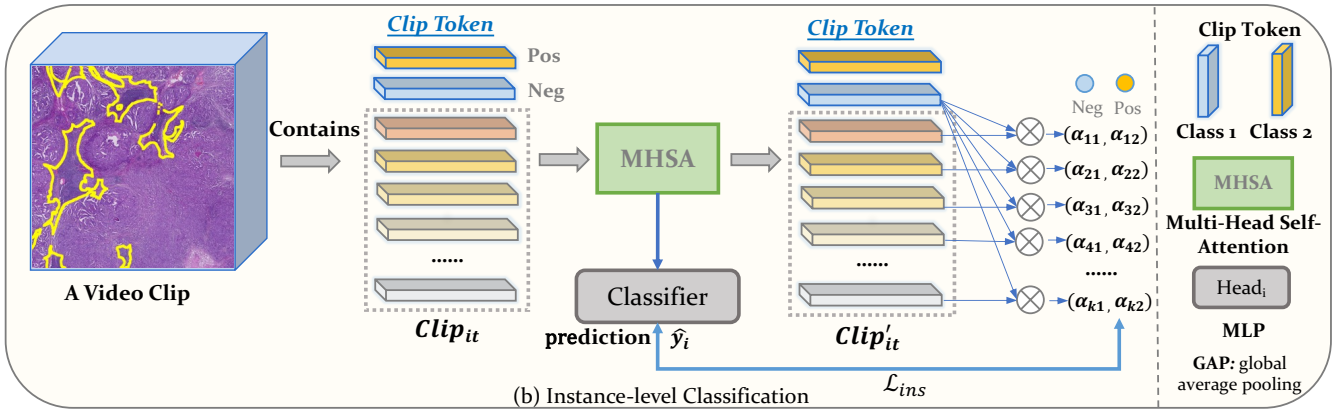

Figure 4: Instance-level classification of VINO.

$$\mathcal{L}_{bag} = CrossEntropy(Video\ score_i, Y_i) \tag{7}$$

Transformer-based Instance-level Classification

In WSI analysis tasks, apart from the bag-level classification task, another main task is the instance-level classification task. In cancer diagnostic tasks, classifying each instance into benign or malignant categories has significant clinical importance. For example, in prostate grading tasks, regions are required to be rated by Gleason scores. The final grading is then determined based on the overall area percentage of each score, and instance classification plays a vital role in clinical diagnoses of prostate cancer.

We introduce an instance classification method based on the VINO Transformer architecture. Each video clip has n clip tokens $\in \mathbb{R}^{1\times dim}$ designed to learn class-specific representations. Consequently, every video clip possesses class-specific semantic information. The similarity between every clip feature that has passed through MHSA and the class-specific clip token is computed. As depicted in Figure 4, this measures the similarity of the instance feature with the vectors representing benign and malignant categories. During the training process, the instance branch utilizes the MHSA parameters trained by the bag branch to conduct instance-level training.

Due to the lack of instance-level labels for instance classification, this paper proposes creating high-quality pseudo-labels for the instance classifier branch by computing the similarity between the class-specific clip token and the instance features. Specifically, the class-specific clip tokens that have learned category semantic information during bag classification training and the clip features encoded through the self-attention mechanism, denoted as $Clip'_{it}$, are used to compute the dot product. The resulting scores serve as the pseudo-labels $\hat{y}_{i,j}$. This is then used to train the Classifier Head $\mathcal{H}$, specifically by computing the cross-entropy loss. Notably, for binary classification problems, each clip attaches two clip tokens $\in \mathbb{R}^{2\times dim}$, corresponding to positive and negative categories. For multi-class pathological classification tasks, such as the Gleason grading in prostate cancer diagnosis, each clip is attached with n tokens $\in \mathbb{R}^{n\times dim}$, where n is the number of classes. As depicted in Figure 3, supposing $Clip'_{it}$ contains k instances, the instance-level prediction for calculating the i^{th} class score is as follows:

$$\hat{y}_{i,j} = dot\left(Clip\ token, Clip'_{itj}\right) \tag{8}$$

$$y_j = \mathcal{H}(Clip'_{itj}), \quad j = 1, \ldots, k \tag{9}$$

$$\mathcal{L}_{\text{ins}} = CrossEntropy(\hat{\boldsymbol{y}}, \boldsymbol{y}) \tag{10}$$

Experiment

Datasets

To demonstrate the performance of VINO and compare it to SOTA algorithms, various experiments were conducted over three public datasets: CAMELYON16 (Bejnordi et al. 2017), TCGA-NSCLC[1], PANDA (Bulten et al. 2022). Finally, analysis and ablation studies are conducted on the public dataset CAMELYON16. It's noted that CAMELYON and TCGA-NSCLC datasets are binary classification problems, while the PANDA dataset is a multi-class dataset contains 3 Gleason scores and instance-level annotations.

Evaluation Metrics

For the bag classification and positive patch localization tasks, we use the bag-level AUC and the instance-level AUC, correspondingly. The PANDA dataset comprises three Gleason Score grades, and individual slides may contain regions with different Gleason Score patterns, which is a multi-label task. we present both the bag-level and instance-level AUCs distinctively for each Gleason Score category. As for the datasets from TCGA, our experiments only focus on the task of bag classification.

[1]http://www.cancer.gov/tcga

Implementation Details and Comparison Methods

During the training process, we utilized the cross-entropy loss, with an Adam optimizer having a learning rate of 1e-4 and weight decay of 1e-4. The encoder for extracting video clips is implemented with ResNet18 (Chen et al. 2020b). All experiments are conducted using 2 A100s.

We compared the VINO framework with current SOTA methods, including ABMIL (Ilse, Tomczak, and Welling 2018), Loss-ABMIL (Shi et al. 2020), DSMIL (Li, Li, and Eliceiri 2021), TransMIL (Shao et al. 2021), DTFD-MIL (Zhang et al. 2022), IBMIL (Lin et al. 2023), and an instance-level approach proposed by Chikontwe (Maksoud et al. 2020). We reproduced these methods based on the codes they published and reported their instance-level AUC (if instance-level predictions were available). We also reproduced ABMIL and Loss-ABMIL with end-to-end training. When reproducing them, due to the memory constraint, we sampled 64 instances in each batch to represent the whole bag. In addition, we also constructed video clips using features extracted by the frozen encoder (denoted as VINO-Feature) to show the advantage of end-to-end training, while our proposed method is denoted as VINO-E2E. We compared VINO with all the other methods on the CAMELYON16, PANDA, and TCGA-NSCLC datasets. For the CAMELYON16 and PANDA datasets, we also compared VINO with fully supervised methods using instance-level labels.

Comparison results

Results with CAMELYON16 The instance-level and bag-level classification AUCs of our VINO and the other comparison methods on CAMELYON16 are shown in Table 1. The instance-level and bag-level AUC of VINO-E2E are 0.9213 and 0.9466, respectively, which surpass the other methods by a large margin. The results of VINO-Feature demonstrate that the VINO framework can also achieve SOTA results even without end-to-end training. This is attributed to the fact that each clip within the video structure retains local context awareness, and every clip token carries class-specific semantic information. Nevertheless, the bag-level AUC of VINO-E2E is much higher than VINO-Feature, which shows the importance of end-to-end training.

Results with PANDA The instance-level and bag-level classification of our VINO and the other comparison methods for each Gleason score is shown in Table 2, where the Gleason score i is denoted as $Grade\ i$. Comparing VINO with end-to-end and two-stage methods, we can see that VINO surpasses these methods in both instance-level and bag-level AUC, only slightly below the fully supervised methods. The outcomes suggest that our VINO can also be applied to multi-class prediction tasks. The end-to-end VINO also outperforms the VINO with pre-extracted features in this task.

Results with TCGA-NSCLC Table 3 displays the performance of VINO and other methods on the TCGA-NSCLC Dataset. Since there are no precise instance labels available for this dataset, we only evaluate the performance of bag classification. This is a relatively easy task and the patch features pre-extracted by SimCLR(Yan et al. 2018) are sufficient, so all methods work with the pre-extracted features instead of in an end-to-end way. We randomly select features to construct the video clip for MHSA for further processing. VINO-Feature, utilizing the pre-extracted features, achieves a bag-level AUC of 0.9834 and outperforms other methods. The results indicate that our VINO can also be employed in a two-stage training approach with the pre-extracted features, and the advantage of VINO lies in using class-specific clip tokens to obtain robust class-specific representation.

Methods	End to End	Instance AUC	Bag AUC
Fully Supervised	**TRUE**	**0.9644**	**0.8621**
ABMIL(18'ICML)	**TRUE**	0.4739	0.6612
Loss-Attention(20'AAAI)	**TRUE**	0.6173	0.7024
Chikontwe(20'MICCAI)	**TRUE**	0.7880	0.7024
ABMIL(18'ICML)	**FALSE**	0.8480	0.8379
Loss-Attention(20'AAAI)	**FALSE**	0.8995	0.7965
DSMIL(21'CVPR)	**FALSE**	0.8858	0.7826
TransMIL(21'NeurIPS)	**FALSE**	-	0.8360
DTFD-MIL(22'CVPR)	**FALSE**	0.7411	0.8638
IBMIL(23'CVPR)	**FALSE**	-	0.8991
VINO-Feature	**FALSE**	**0.9185**	**0.9085**
VINO-E2E	**TRUE**	**0.9213**	**0.9466**

Table 1: Comparison of classification performance on the CAMELYON16 Dataset.

Ablation Study

We conducted ablation studies on the key components of VINO using the CAMELYON16 dataset. It should be noted that we already conducted ablation studies on each dataset to compare using pre-extracted features and end-to-end training. The additional ablation study results are shown in Table 5. Among them, "w/o contextual information" means randomly picking patches to construct each video clip. End-to-end training, modeling contextual information, and designing class-specific clip tokens can all effectively enhance VINO performance.

Moreover, we employed the traditional video anomaly detection method, RTFM (Tian et al. 2021), to perform anomaly detection on pathological videos, examining the anomalies in each frame (instance-level) and the entire video (bag-level), which corresponds to benign or malignant. Experiments show traditional video anomaly detection methods are not suitable for WSI classification. One significant issue is that in traditional videos, each frame is sequential and has overlapping relationships, whereas in our pathological videos, frames are independent with only positional relationships. To facilitate computer processing, we treated adjacent non-overlapping patches without natural temporal continuity as temporally consecutive video frames. We also experimented with overlapping patches to construct video clips. We found enhanced continuity can be more feasible for video formulation. Compared with VINO-Feature

Methods	End to End	Instance-level AUC			Bag-level AUC		
		Grade 3	Grade 4	Grade 5	Grade 3	Grade 4	Grade 5
Fully Supervised	**TRUE**	**0.9862**	**0.9800**	**0.9784**	**0.9405**	**0.9420**	**0.9244**
ABMIL(18'ICML)	**TRUE**	0.8976	0.8959	0.8423	0.8269	0.8522	0.7988
Chikontwe(20'MICCAI)	**TRUE**	0.9260	0.9183	0.9411	0.8438	0.8269	0.8561
ABMIL(18'ICML)	**FALSE**	0.8614	0.9144	0.9142	0.9013	0.8782	0.8874
Loss-Attention(20'AAAI)	**FALSE**	0.9398	0.9284	0.9245	0.8708	0.8608	0.8584
DSMIL(21'CVPR)	**FALSE**	0.9525	0.9363	0.9263	0.8783	0.8650	0.8714
TransMIL(21'NeurIPS)	**FALSE**	-	-	-	0.8089	0.8260	0.7396
DTFD-MIL(22'CVPR)	**FALSE**	0.9361	0.9290	0.9159	0.8875	0.8813	0.8546
IBMIL(23'CVPR)	**FALSE**	-	-	-	0.9015	0.8923	**0.9010**
VINO-Feature	**FALSE**	**0.9626**	**0.9370**	**0.9602**	**0.9100**	**0.8991**	0.8978
VINO-E2E	**TRUE**	**0.9659**	**0.9406**	**0.9603**	**0.9173**	**0.9048**	**0.9140**

Table 2: Comparison of classification performance on the PANDA.

Methods	End to End	Bag AUC
Mean-pooling	**FALSE**	0.9369
Max-pooling	**FALSE**	0.9014
ABMIL(18'ICML)	**FALSE**	0.9488
Loss-Attention(20'AAAI)	**FALSE**	0.9517
Chikontwe(20'MICCAI)	**FALSE**	0.9523
DSMIL(21'CVPR)	**FALSE**	0.9633
TransMIL(21'NeurIPS)	**FALSE**	0.9830
DTFD-MIL(22'CVPR)	**FALSE**	0.9808
IBMIL(23'CVPR)	**FALSE**	0.9789
VINO-Feature	**FALSE**	**0.9853**

Table 3: Comparison of classification performance on the TCGA-NSCLC.

Key Point	Instance AUC	Bag AUC
VINO end-to-end	**0.9213**	**0.9466**
w/o contextual information	0.8925	0.9276
w/o end-to-end training	0.9185	0.9085
RTFM (Tian et al. 2021)	0.8749	0.7500

Table 4: Main results of ablation study on the CAMELYON16 Dataset.

(overlapping rate=0%), when we tiled CAMELYON16 with a 10% overlapping rate, VINO-Feature outperformed by 0.36% and 0.17% on both Instance-level and bag-level classification.

Evaluation of VINO end-to-end training. At present, most methods use pre-trained encoder to obtain instance features before training. The reasons why VINO is effective lie in VINO achieving end-to-end training by video formulation, which avoids the inductive bias in two-stage methods, in which a feature extractor is trained by tasks different from the target tasks. Second, our way of constructing video clips from neighboring patches considers context information. We analyze the quality of the patch features extracted by VINO in this section.

This analysis is also conducted on the CAMELYON16 dataset. Firstly, we utilized both ImageNet pre-trained and SimCLR pre-trained models as performed in DSMIL (Li, Li, and Eliceiri 2021) to train feature extractors and then used them to extract all patch features, as well as using the feature extractor trained with our method. Following this, using the true labels of each patch, we trained a simple SVM classifier and a linear classifier on the training set. We then evaluated these classifiers on the test set, keeping in mind that all methods adopted the ResNet-18. Table 5 indicates that the features extracted by VINO consistently achieved the highest scores, suggesting that the VINO method can extract better features than the unsupervised SimCLR.

Methods	Instance-level AUC	
	SVM	Linear
Fully Supervised	0.9644	0.9642
ImageNet Pretrained	0.7963	0.7959
SimCLR	0.9365	0.9344
VINO (ours)	**0.9521**	**0.9403**

Table 5: SVM and linear evaluation of pre-extracted features on the Camelyon16 Dataset.

Conclusion

To address issues related to preserving contextual relationship between patches, domain gap and improper inductive bias, in this paper, we construct synthetic video clips from a WSI that can achieve end-to-end training and the modeling of contextual relationship between patches. A divide-and-conquer strategy is employed, where only a subset of WSIs input into the Transformer-based model which can reduce the spatial complexity of the proposed method to achieve end-to-end training. In the future, we will apply the model to real pathological diagnosis videos such as eye-tracking videos, helping pathologists reduce the misdiagnosis rate.

Acknowledgments

This work was supported by the National Key Research and Development Program of China under Grant No. 2022ZD0116800.

References

Bejnordi, B. E.; Veta, M.; Van Diest, P. J.; Van Ginneken, B.; Karssemeijer, N.; Litjens, G.; Van Der Laak, J. A.; Hermsen, M.; Manson, Q. F.; Balkenhol, M.; et al. 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22): 2199–2210.

Bulten, W.; Kartasalo, K.; Chen, P.-H. C.; Ström, P.; Pinckaers, H.; Nagpal, K.; Cai, Y.; Steiner, D. F.; van Boven, H.; Vink, R.; et al. 2022. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nature medicine*, 28(1): 154–163.

Cai, H.; Feng, X.; Yin, R.; Zhao, Y.; Guo, L.; Fan, X.; and Liao, J. 2023. MIST: multiple instance learning network based on Swin Transformer for whole slide image classification of colorectal adenomas. *The Journal of Pathology*, 259(2): 125–135.

Chen, R. J.; Chen, C.; Li, Y.; Chen, T. Y.; Trister, A. D.; Krishnan, R. G.; and Mahmood, F. 2022. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16144–16155.

Chen, R. J.; and Krishnan, R. G. 2022. Self-supervised vision transformers learn visual concepts in histopathology. *arXiv preprint arXiv:2203.00585*.

Chen, R. J.; Lu, M. Y.; Shaban, M.; Chen, C.; Chen, T. Y.; Williamson, D. F.; and Mahmood, F. 2021. Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, 339–349. Springer.

Chen, R. J.; Lu, M. Y.; Wang, J.; Williamson, D. F.; Rodig, S. J.; Lindeman, N. I.; and Mahmood, F. 2020a. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, 41(4): 757–770.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020b. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 1597–1607. PMLR.

Chikontwe, P.; Kim, M.; Nam, S. J.; Go, H.; and Park, S. H. 2020. Multiple instance learning with center embeddings for histopathology classification. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 519–528. Springer.

Hashimoto, N.; Fukushima, D.; Koga, R.; Takagi, Y.; Ko, K.; Kohno, K.; Nakaguro, M.; Nakamura, S.; Hontani, H.; and Takeuchi, I. 2020. Multi-scale domain-adversarial multiple-instance CNN for cancer subtype classification with unannotated histopathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3852–3861.

Hou, W.; Yu, L.; Lin, C.; Huang, H.; Yu, R.; Qin, J.; and Wang, L. 2022. Hˆ 2-MIL: Exploring Hierarchical Representation with Heterogeneous Multiple Instance Learning for Whole Slide Image Analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 933–941.

Ilse, M.; Tomczak, J.; and Welling, M. 2018. Attention-based deep multiple instance learning. In *International Conference on Machine Learning (ICML)*, 2127–2136. PMLR.

Li, B.; Li, Y.; and Eliceiri, K. W. 2021. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14318–14328.

Li, H.; Yang, F.; Zhao, Y.; Xing, X.; Zhang, J.; Gao, M.; Huang, J.; Wang, L.; and Yao, J. 2021. DT-MIL: deformable transformer for multi-instance learning on histopathological image. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 206–216. Springer.

Lin, T.; Xu, H.; Yang, C.; and Xu, Y. 2022. Interventional multi-instance learning with deconfounded instance-level prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1601–1609.

Lin, T.; Yu, Z.; Hu, H.; Xu, Y.; and Chen, C.-W. 2023. Interventional bag multi-instance learning on whole-slide pathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19830–19839.

Lu, M. Y.; Chen, R. J.; Kong, D.; Lipkova, J.; Singh, R.; Williamson, D. F.; Chen, T. Y.; and Mahmood, F. 2022. Federated learning for computational pathology on gigapixel whole slide images. *Medical Image Analysis*, 76: 102298.

Lu, M. Y.; Chen, R. J.; Wang, J.; Dillon, D.; and Mahmood, F. 2019. Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding. *arXiv preprint arXiv:1910.10825*.

Lu, M. Y.; Chen, T. Y.; Williamson, D. F.; Zhao, M.; Shady, M.; Lipkova, J.; and Mahmood, F. 2021a. AI-based pathology predicts origins for cancers of unknown primary. *Nature*, 594(7861): 106–110.

Lu, M. Y.; Williamson, D. F.; Chen, T. Y.; Chen, R. J.; Barbieri, M.; and Mahmood, F. 2021b. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6): 555–570.

Lu, M. Y.; Williamson, D. F.; Chen, T. Y.; Chen, R. J.; Barbieri, M.; and Mahmood, F. 2021c. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6): 555–570.

Luo, X.; Qu, L.; Guo, Q.; Song, Z.; and Wang, M. 2023. Negative Instance Guided Self-Distillation Framework for Whole Slide Image Analysis. *IEEE Journal of Biomedical and Health Informatics*.

Mahmood, F.; Borders, D.; Chen, R. J.; McKay, G. N.; Salimian, K. J.; Baras, A.; and Durr, N. J. 2019. Deep adversarial training for multi-organ nuclei segmentation in histopathology images. *IEEE Transactions on Medical Imaging*, 39(11): 3257–3267.

Maksoud, S.; Zhao, K.; Hobson, P.; Jennings, A.; and Lovell, B. C. 2020. Sos: Selective objective switch for rapid immunofluorescence whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3862–3871.

Qu, L.; luo, x.; Wang, M.; and Song, Z. 2022. Bi-directional Weakly Supervised Knowledge Distillation for Whole Slide Image Classification. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 15368–15381. Curran Associates, Inc.

Rony, J.; Belharbi, S.; Dolz, J.; Ayed, I. B.; McCaffrey, L.; and Granger, E. 2019. Deep weakly-supervised learning methods for classification and localization in histology images: a survey. *arXiv preprint arXiv:1909.03354*.

Rymarczyk, D.; Pardyl, A.; Kraus, J.; Kaczyńska, A.; Skomorowski, M.; and Zieliński, B. 2022. Protomil: Multiple instance learning with prototypical parts for whole-slide image classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 421–436. Springer.

Shao, Z.; Bian, H.; Chen, Y.; Wang, Y.; Zhang, J.; Ji, X.; et al. 2021. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems (NeurIPS)*, 34: 2136–2147.

Shi, X.; Xing, F.; Xie, Y.; Zhang, Z.; Cui, L.; and Yang, L. 2020. Loss-based attention for deep multiple instance learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, 5742–5749.

Tian, Y.; Pang, G.; Chen, Y.; Singh, R.; Verjans, J. W.; and Carneiro, G. 2021. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4975–4986.

Tu, M.; Huang, J.; He, X.; and Zhou, B. 2019. Multiple instance learning with graph neural networks. *arXiv preprint arXiv:1906.04881*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, X.; Xiang, J.; Zhang, J.; Yang, S.; Yang, Z.; Wang, M.-H.; Zhang, J.; Yang, W.; Huang, J.; and Han, X. 2022. SCL-WC: Cross-Slide Contrastive Learning for Weakly-Supervised Whole-Slide Image Classification. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 18009–18021. Curran Associates, Inc.

Xiang, J.; and Zhang, J. 2022. Exploring low-rank property in multiple instance learning for whole slide image classification. In *The Eleventh International Conference on Learning Representations*.

Xu, L.; Ouyang, W.; Bennamoun, M.; Boussaid, F.; and Xu, D. 2022. Multi-class token transformer for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4310–4319.

Yan, Y.; Wang, X.; Guo, X.; Fang, J.; Liu, W.; and Huang, J. 2018. Deep multi-instance learning with dynamic pooling. In *Asian Conference on Machine Learning (ACML)*, 662–677. PMLR.

Yuan, M.; Li, Z.; Jin, Q.; Chen, X.; and Wang, M. 2022. PointCLM: A Contrastive Learning-based Framework for Multi-instance Point Cloud Registration. In *European Conference on Computer Vision*, 595–611. Springer.

Zhang, H.; Meng, Y.; Zhao, Y.; Qiao, Y.; Yang, X.; Coupland, S. E.; and Zheng, Y. 2022. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18802–18812.

Zhang, R.; Zhang, Q.; Liu, Y.; Xin, H.; Liu, Y.; and Wang, X. 2023. Multi-level Multiple Instance Learning with Transformer for Whole Slide Image Classification. *arXiv preprint arXiv:2306.05029*.

Zhao, Y.; Yang, F.; Fang, Y.; Liu, H.; Zhou, N.; Zhang, J.; Sun, J.; Yang, S.; Menze, B.; Fan, X.; et al. 2020. Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4837–4846.

Łukasz Struski; Rymarczyk, D.; Lewicki, A.; Sabiniewicz, R.; Tabor, J.; and Zieliński, B. 2023. ProMIL: Probabilistic Multiple Instance Learning for Medical Imaging. In *ECAI 2023*, 2210–2217. IOS Press.