Lymph Node Metastasis Prediction From Whole Slide Images With Transformer-Guided Multiinstance Learning and Knowledge Transfer

Zhihua Wang, Lequan Yu, *Member, IEEE*, Xin Ding, Xuehong Liao, and Liansheng Wang, *Member, IEEE*

***Abstract*— The gold standard for diagnosing lymph node metastasis of papillary thyroid carcinoma is to analyze the whole slide histopathological images (WSIs). Due to the large size of WSIs, recent computer-aided diagnosis approaches adopt the multi-instance learning (MIL) strategy and the key part is how to effectively aggregate the information of different instances (patches). In this paper, a novel transformer-guided framework is proposed to predict lymph node metastasis from WSIs, where we incorporate the transformer mechanism to improve the accuracy from three different aspects. First, we propose an effective transformer-based module for discriminative patch feature extraction, including a lightweight feature extractor with a pruned transformer (Tiny-ViT) and a clustering-based instance selection scheme. Next, we propose a new Transformer-MIL module to capture the relationship of different discriminative patches with sparse distribution on WSIs and better nonlinearly aggregate patch-level features into the slide-level prediction. Considering that the slide-level annotation is relatively limited to training a robust Transformer-MIL, we utilize the pathological relationship between the primary tumor and its lymph node metastasis and develop an effective attention-based mutual knowledge distillation (AMKD) paradigm. Experimental results on our collected WSI dataset demonstrate the efficiency of the proposed Transformer-MIL and attention-based knowledge distillation. Our method outperforms the state-of-the-art methods by over 2.72% in AUC (area under the curve).**

***Index Terms*— Whole slide image analysis, multi-instance learning, transformer, knowledge distillation.**

Manuscript received 5 March 2022; revised 11 April 2022 and 18 April 2022; accepted 23 April 2022. Date of publication 29 April 2022; date of current version 30 September 2022. This work was supported by the Fundamental Research Funds for the Central Universities under Grant 20720190012 and Grant 20720210121. *(Zhihua Wang and Lequan Yu contributed equally to this work.) (Corresponding author: Liansheng Wang.)*

Zhihua Wang and Liansheng Wang are with the School of Informatics, Xiamen University, Xiamen 361005, China (e-mail: wzh04157181@163.com; lswang@xmu.edu.cn).

Lequan Yu is with the Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong, SAR, China (e-mail: lqyu@hku.hk).

Xin Ding and Xuehong Liao are with the Zhongshan Hospital, Xiamen University, Xiamen 361005, China (e-mail: xinding2014@gmail.com; lxh8881021@163.com).

Digital Object Identifier 10.1109/TMI.2022.3171418

I. Introduction

THYROID carcinoma is the eighth most common cancer and the fourth most common malignant tumor in women, of which papillary thyroid carcinoma (PTC) accounts for more than 90%. Medical research shows that the 5-year survival rate of papillary thyroid carcinoma is more than 95%, which is considered as "curable" cancer [1] with early diagnosis and treatment. However, the recurrence rate and mortality rate of patients with lymph node metastasis (LNM) are significantly increased [2]. The gold standard for diagnosing lymph node metastasis is to analyze the histopathological whole slide images (WSIs). However, manual diagnosis of lymph node metastasis from WSIs is laborious and time-consuming, which requires a lot of time for pathologists to locate cancer regions. Therefore, computer-aided WSI analysis approaches were proposed [3]–[8]. Considering that the WSIs are very large (*e.g.*, around 100000×50000 pixels in our case), it is common to divide WSIs into patches first, analyze these patches, and finally summarize the analysis result of WSIs. For these computer-aided methods, how to summarize the patch analysis results is a challenging problem. Multi-instance learning (MIL) [9] is thus proposed to meet the need for cancer diagnosis from the whole slide image. For those MIL-based WSI analysis methods, the slide is usually regarded as the bag and the patches as the instances.

In recent years, many advanced MIL models have been proposed for medical image analysis problems [10]–[16]. However, these models still meet several challenges when deployed to WSI analysis tasks: 1) For the histopathological image analysis task, due to the huge size of histopathological images, inputting all patches into the network will largely increase the burden of the model and affect the inference speed, so how to select the most distinctive patches is an important step for the multi-instance learning method; 2) The embedded space (ES) based MIL methods are not trainable or only partially trainable, which makes models difficult to effectively learn the relationship among different patches in whole slide images and the flexibility of these methods is limited; and 3) there are usually only a few hundred patholog-

ical slides for MIL, and thus the information that the model can utilize to learn the relationship among different patches is limited, which would lead to the overfitting problem in the training process. In this case, how to utilize the knowledge from other relevant datasets to alleviate this problem is also a challenge.

In this paper, we aim at developing an effective and accurate MIL-based framework for the lymph node metastasis prediction of papillary thyroid cancer from the whole slide image. To tackle the above challenges, we propose a unified transformer-guided framework to extract features of the discriminate patches, aggregate the patch-level information by comprehensively leveraging the relationship of different patches, and transfer such relationship knowledge from another related dataset to enhance the robustness of the framework. Specifically, we employ a transformer-based network, ViT [17] to extract the representative features of each patch by better exploiting the context information and then select discriminative patches with an effective clustering-based strategy. Also, we compress the ViT model to a lightweight feature extractor, called Tiny-ViT, to speed up the network inference. Exploiting the relationship of these discriminative patches is critical to aggregate them into a slide-level prediction, we thereby incorporate the multi-head self-attention mechanism proposed in Transformer [18] with our clustering-based embedding mechanism to form a novel Transformer-MIL module. For the discriminative patches with sparse distribution on WSI, the proposed Transformer-MIL can utilize the context relationship between different patches and thus learn how to nonlinearly aggregate them into a slide-level prediction. More importantly, considering that the slide-level label is usually limited to training a robust Transformer-MIL, we further design a novel attention-based mutual knowledge distillation (AMKD) paradigm to facilitate the Transformer-MIL to implicitly learn how to aggregate discriminative patches from a related PTC WSI classification dataset. The experiments on our collected WSI datasets show that our method outperforms the state-of-the-art methods for WSI prediction and we acquire additional improvement by incorporating another related PTC dataset. We also conducted an extensive analysis to demonstrate the effectiveness of each component in our proposed framework.

The main contributions are summarized as follows.

- We present an effective transformer-guided framework for lymph node metastasis prediction from WSIs, including transformer-based feature extraction and feature aggregation. Our framework largely outperforms other recent WSI prediction methods.
- We propose an innovative Transformer-MIL to learn the relationship among different discriminative patches through the multi-head self-attention mechanism and better embed the patch-level features into slide-level prediction.
- We develop a novel attention-based mutual knowledge distillation (AMKD) paradigm to leverage the pathological similarity of two related WSI datasets to improve the prediction performance of lymph node metastasis. We take the attention map as the medium to better transfer the relationship knowledge to improve the MIL performance.

The remainder of this paper is organized as follows. We discuss the related works in Section II and elaborate on the proposed framework in Section III. We present the experimental setting and results in Section IV, and further discuss the key points of our method in Section V. Then, we draw the conclusions in Section VI.

II. Related Works

A. Whole Slide Image Analysis

Conventional histopathological examination is the gold standard for the diagnosis and grading of various types of cancer. Tumor information, such as histological grading, mitosis rate, lymph node status, and tumor stage, can be extracted from pathological images. With the development of digital whole slide image (WSI) technology, there is more and more research on computer-aided pathological image analysis [19]–[22]. For example, pathologists often diagnose low-resolution WSI first and then high-resolution WSI when necessary. The computer-aided prognosis system of neuroblastoma (NB) developed by Sertel *et al.* [3] imitates this diagnostic strategy and uses the off-line feature selection step to determine the most distinctive features at each resolution level during the training process. Then, the improved k-nearest neighbor classifier is used to determine the confidence level of classification to make a decision at a specific resolution level. Li *et al.* [5] noted the important role of WSI's generalization feature in the analysis process. Taking WSI as the graphical modeling, they developed a graphical convolution neural network (graph CNN) with attention learning to better serve survival prediction by rendering the best graphical representation of WSI. Because WSI is very large, in the process of analysis, we usually cut WSI to patches, and then summarize the patch-level analysis to slide-level analysis results. This analysis method from patch to slide coincides with the idea of instance to bag in multi-instance learning, so the effective deployment of multi-instance learning to WSI analysis has become a key issue in the field of medical image analysis in recent years.

B. Multi-Instance Learning

The existing MIL methods include three main paradigms: bag space (BS), instance space (IS), and embedded space (ES) MIL [23]. The BS paradigm directly analyzes the bags and classifies them by the distance from bag to bag. Because of the huge size of WSI, it is not accurate to detect lesions by analyzing slides directly. In the IS paradigm, the learning process is mainly carried out at the instance-level, focusing on the instance analysis stage, and simply summarizing the instance-level analysis results into the bag-level classification results. The performance of this paradigm is usually poor. The recent multi-instance learning is mainly based on embedded space (ES) [23]. ES-MIL first embeds instances into low-dimensional features and then uses these low-dimensional features to get the bag representation. Therefore, an effective bag embedding method is the key to the ES-MIL method. As for bag embedding methods, some methods based on fixed

pooling are proposed firstly, such as max pooling, average pooling [23], and so on. Later, the method based on parameter pooling was proposed, such as dynamic pooling [24], adaptive pooling [25], and so on. This kind of method is partially trainable, but it is not flexible enough. Until Ilse *et al.* [12] introduced the attention mechanism into multi-instance learning and used the attention module as a bag embedding method to get a fully trainable multi-instance model. Based on this research direction, this paper proposes a multi-instance learning method based on Transformer [18], Transformer-MIL, to predict lymph node metastasis of papillary thyroid carcinoma by analyzing WSI.

C. Transformer in Medical Image Analysis

In recent years, more and more research on transformer-based network architecture for medical image analysis has been carried out. For example, Valanarasu *et al.* [26] combined the gated axial-attention mechanism with the self-attention mechanism in Transformer [18] to learn global and local features through slide and patch branches to realize image segmentation. Gao *et al.* [27] combined convolution and the self-attention mechanism to get a new self-attention module and used the module in each encoder and decoder to capture remote dependencies at different scales. Concurrently, Shao *et al.* [28] applied the Transformer framework to the WSI classification task and proposed an innovative ES-MIL method: TransMIL. However, due to the lack of discriminative patch selection, TransMIL cannot perform the classification task well in the face of high-resolution WSI containing tens of thousands of patches. In this work, we propose a clustering-based instance selection method to reduce the number of patches and further design an attention-based knowledge distillation scheme to utilize other related datasets.

D. Knowledge Distillation

Due to the lack of data for medical image analysis tasks, the knowledge that the model can learn in the training process is very limited. To alleviate this problem, Li *et al.* [29] proposed a multi-modal mutual knowledge distillation method,where they firstly reduced the appearance gap between the two modes and then deployed two segmentation networks to segment the images of two modes respectively. Each network not only obtained explicit knowledge from its mode but also obtained implicit knowledge from another mode through knowledge distillation. Since the morphology of papillary thyroid carcinoma and its lymph node metastasis is the same [2], this method of mutual knowledge distillation can also be applied to the multi-instance learning task of predicting lymph node metastasis. The Transformer-MIL proposed in this paper is a multi-instance learning method based on the multi-head self-attention mechanism. The knowledge distillation based on model prediction cannot give full play to the role of the multi-head self-attention module [18]. Komodakis and Zagoruyko [30] proposed a method to guide the simple model by extracting the attention map of the complex model. Through this attention-based distillation method, a simple model cannot only learn feature information but also understand how to extract feature information. It makes the features generated by simple models more flexible and not limited to complex models. Therefore, we propose an attention-based mutual knowledge distillation to better utilize the knowledge of existing data.

III. Methodology

The overall framework of the proposed method is illustrated in Fig. 1, which consists of three parts: discriminative patch selection, Transformer-MIL aggregation, and attention-based mutual knowledge distillation (AMKD). In the discriminative patch selection, the features of overlapped patches in the WSI are extracted with a light-weight ViT network and then we adopt a clustering-based strategy to select the discriminative patches. An effective Transformer-MIL is designed to learn the relationship between instances from multiple aspects to aggregate the discriminative patch-level features. Finally, we use AMKD to facilitate the Transfomer-MIL to learn the knowledge of attention allocation implicitly from the datasets of papillary thyroid carcinoma and its lymph node metastasis to obtain a better patch aggregation effect. It is worth noting that we first perform discriminative patch selection on all WSIs and then train the Transformer-MIL aggregation module and AMKD module in an end-to-end manner.

A. Discriminative Patch Selection

For histopathological image analysis, each WSI is often divided into tens of thousands of patches, so how to select the most distinctive patches is an important step for the multi-instance learning method. Different from the previous approaches that usually select several instances with the maximum and minimum scores, we propose a new strategy to select discriminative patches: a lightweight feature extractor with Tiny-ViT and a clustering-based instance selection. We first crop the slides into 512×512 non-overlapping patches and remove the patches without tissue cells. Then we use the state-of-the-art feature extractor, Vision Transformer [17], to extract the features of each patch. Compared with other convolutional neural networks, ViT can better extract the context information of patch images. To train the ViT, the cancer regions of a few lymph node metastasis WSIs are labeled by pathologists. Due to the huge amount of model parameters, the direct application of ViT in the feature extraction stage will not only consume lots of computing resources but also extract instances excessively. To reduce the computation burden and accelerate the training and inference time, we reduce the depth and dimension of ViT and acquire a lightweight feature extractor, Tiny-ViT. The patches with the size of 512×512 are extracted as 312-dimensional feature vectors by the Tiny-ViT. The patch-level features in the WSI are further clustered into 10 categories by traditional K-means [31]. Finally, a total of 200 patch-level features are extracted according to the proportion as the final discriminative patch features, in which the proportion is the number of patches in each category to the total number of patches to represent the slide-level features. Different from previous prediction-based instance selection,

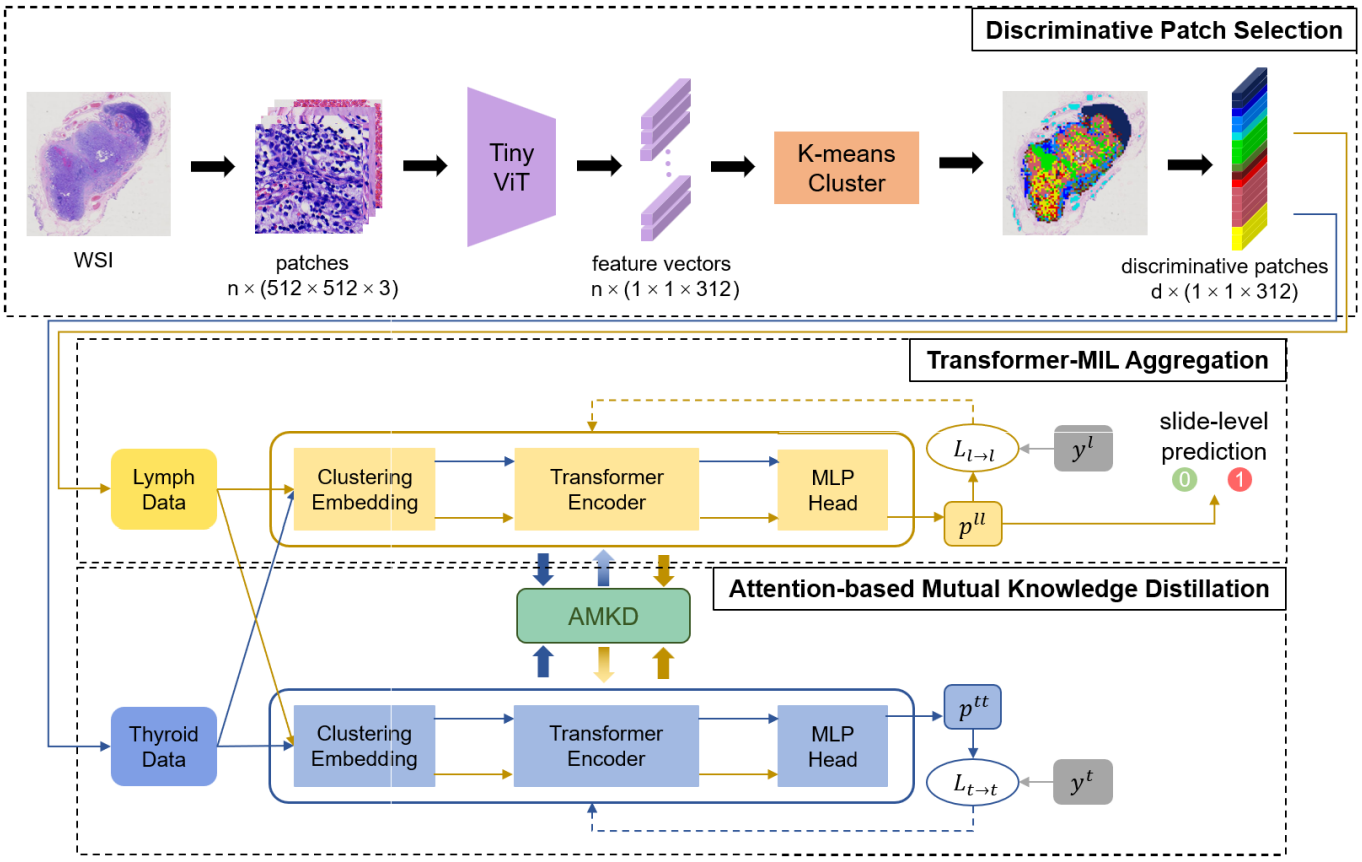


Fig. 1. The whole pipeline of our framework equipped with Transformer-MIL and AMKD for WSI analysis. The pipeline is composed of three parts, *i.e.*, discriminative patch selection, Transformer-MIL aggregation, and attention-based mutual knowledge distillation. *n* is the number of patches extracted from a WSI and *d* is the number of discriminative patch-level features with a value of 200.

clustering-based instance selection can cluster the patches from many aspects to ensure that the selected patches are more diversified and representative.

B. Transformer-MIL Aggregation

ES-MIL aims to acquire the bag-level representation from instance-level features. In this process, it is worth studying the attention mechanism to calculate the relationship between different patches in WSI, so that the model can assign appropriate weights to the patches in WSIs more flexibly. Each selected patch in the WSI has its specific location information, and the simple attention mechanism is easy to ignore the relationship between long-range instances. Therefore, we propose to incorporate the Transformer [18] mechanism into the MIL to alleviate the aforementioned issue. The multi-head self-attention module proposed in Transformer mines information by calculating the similarity between every two words, rather than only focusing on adjacent words, so there is no such information loss. We thus adopt this module to our WSI prediction problem.

Fig. 2 illustrates the architecture of the proposed Transformer-MIL for instance aggregation, which consists of three parts: clustering embedding, transformer encoder, and classification head. One important consideration of employing the Transformer is the sequence of input patches. As the distribution of selected patches on WSI is very sparse, the position embedding based on 2D location information cannot well reflect the relative relationship between different patches. Therefore, we design a clustering embedding submodule to make Transfomer more proper for the task of histopathological image analysis. As shown in Fig. 2, the clustering embedding submodule rearranges the patches of the same cluster together according to the clustering results of patches and then combines all clusters to form a patch sequence, where the patches of the same cluster are closer. Therefore, each patch is given embedding information according to the patch order in the sequence. For the transformer encoder, we use N repeated encoder blocks ($N = 4$ in our experiments) to embed 200 instance-level features into the whole slide representation. Each block is composed of a multi-head self-attention module (MHA), a feed-forward network (FFN) with the residual connection, and two Layer Normalization (LN) layers [32], which can be represented as:

$$\begin{aligned} Block(\boldsymbol{X_i}) &= \boldsymbol{A} + FFN(LN(\boldsymbol{A})), \\ \boldsymbol{A} &= \boldsymbol{X_{i-1}} + MHA(LN(\boldsymbol{X_{i-1}})), \end{aligned} \tag{1}$$

where $\boldsymbol{X_i}$ is the feature map after the i^{th} encoder block and $\boldsymbol{X_0}$ is the original 200 instance-level feature vectors after clustering embedding. It is worth noting that we set a learnable embedding as the classification token to perform classification. The adopted multi-head attention allows the network to learn the relationship between each instance and

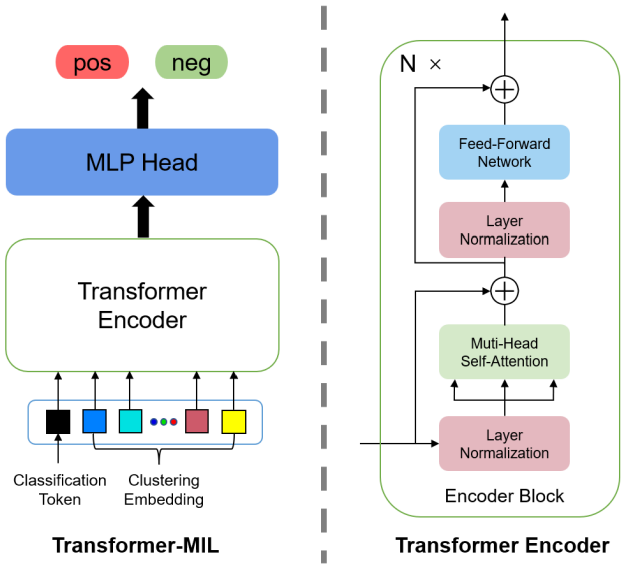


Fig. 2. The architecture of the proposed Transformer-MIL and the encoder block. N is the number of encoder blocks with a value of 4.

other instances from multiple aspects. Specifically, the multi-head self-attention module in the encoder block employs the key-value attention multiple times for better embedding, which can be defined as follows:

$$\begin{aligned} MHA(\boldsymbol{X}) &= Concat(\boldsymbol{H_1}, \ldots, \boldsymbol{H_M})\boldsymbol{W^O}, \\ \boldsymbol{H_i} &= Att(\boldsymbol{X}\boldsymbol{W_i^Q}, \boldsymbol{X}\boldsymbol{W_i^K}, \boldsymbol{X}\boldsymbol{W_i^V}), \\ Att(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) &= softmax(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}})\boldsymbol{V}, \end{aligned} \tag{2}$$

where the parameter matrices $\boldsymbol{W_i^Q} \in \mathbb{R}^{d_{model} \times d_k}$, $\boldsymbol{W_i^K} \in \mathbb{R}^{d_{model} \times d_k}$, $\boldsymbol{W_i^V} \in \mathbb{R}^{d_{model} \times d_v}$, $\boldsymbol{W^O} \in \mathbb{R}^{d_{model} \times Md_v}$, d_{model} is the feature embedding dimensions of the model. $\boldsymbol{X}$ is the input feature map. $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are matrices packing a set of queries, keys, and values. $\boldsymbol{H}_i$ is the i^{th} attention head and M is the number of attention heads. d_k is the dimension of keys and queries and d_v is the dimension of the values. After getting the representation of the bag, we map it to the final prediction through a multi-layer perceptron (MLP) composed of two hidden layers.

C. Attention-Based Mutual Knowledge Distillation

To train the proposed Transformer-MIL, we use one WSI as one training sample. As the number of WSIs is relatively scarce, the knowledge that the model can learn is limited and it would lead to the over-fitting problem. To alleviate this problem, we propose to exploit other related datasets to assist the training of Transformer-MIL. Specifically, we notice that the morphology of cancer cells in papillary thyroid carcinoma (PTC) and its lymph node metastasis are similar [2], showing that the knowledge in the PTC dataset would be helpful for the model to predict lymph node metastasis. Based on this observation, we propose to use PTC classification datasets to assist the training of Transformer-MIL for lymph node metastasis prediction via knowledge distillation [29], [33]. Different from the traditional knowledge distillation scheme, which lets the student model mimic the final prediction probability of the teacher model, we propose a novel attention-based knowledge distillation scheme to effectively transfer the knowledge of patch relation via mimicking the attention map knowledge in the transformer encoder.

As shown in Fig. 1, we employ two Transformer-MIL networks with the same structure to learn the patch relation knowledge. These two networks are trained on the papillary thyroid carcinoma dataset and lymph node metastasis dataset of papillary thyroid carcinoma at the same time, but each network only receives the direct supervision of one dataset label. We also adopt the mutual learning strategy to train the two networks. In the training process, each network not only obtains knowledge explicitly from the current dataset but also implicitly absorbs knowledge from another dataset. Through this mutual learning paradigm, Transformer-MIL can better explore the patch relation knowledge of two datasets in both explicit and implicit ways. M_t is the model for predicting papillary thyroid carcinoma (t denotes thyroid) and M_l is the model for predicting lymph node metastasis (l denotes lymph).

For explicit learning, the update gradient is obtained by calculating the class-wise weighted cross-entropy loss of the prediction generated by the model and the input label,

$$\begin{aligned} \mathcal{L}_{t\to t} &= -\beta_0(1-y^t)log(1-p^{tt}) + \beta_1 y^t log(p^{tt}) \\ \mathcal{L}_{l\to l} &= -\beta_0(1-y^l)log(1-p^{ll}) + \beta_1 y^l log(p^{ll}) \end{aligned} \tag{3}$$

where y^t and y^l are the labels of thyroid data and lymph data respectively. p^{tt} is the prediction probability of M_t for input thyroid data and p^{ll} is the prediction probability of M_l for input lymph data. Since the number of negative samples is more than that of positive samples in both datasets, we add the class-wise weight [34] into the loss calculation, $[\beta_0, \beta_1] = [1.43, 0.77]$.

For implicit learning, we aim to each network can learn from another dataset to better map instance-level features to bag-level features. We thus propose the attention-based knowledge distillation strategy. Considering that the shallow layers of the network extract common features while the deep layers extract the corresponding refined features based on different datasets, we only apply knowledge distillation to the first two encoder blocks of Transformer-MIL to mine the commonly shared knowledge, as shown in Fig. 3. Formally, this strategy can be represented as:

$$\begin{aligned} \mathcal{L}_{l\to t} &= \sum_{i=1}^{2} \mathcal{L}_{mse}(\boldsymbol{A_i^{tl}}, \boldsymbol{A_i^{ll}}), \\ \mathcal{L}_{t\to l} &= \sum_{i=1}^{2} \mathcal{L}_{mse}(\boldsymbol{A_i^{lt}}, \boldsymbol{A_i^{tt}}), \end{aligned} \tag{4}$$

where $\boldsymbol{A_i^{tl}}$ is the attention map generated by the i^{th} encoder block in M_t for the input lymph data, $\boldsymbol{A_i^{ll}}$ is the attention map generated by the i^{th} encoder block in M_l for the input lymph data. We calculate the mean squared error (MSE) loss from the attention maps of the two models to get the knowledge $\mathcal{L}_{l\to t}$ learned by M_t from the lymph data. $\mathcal{L}_{t\to l}$ is the same.

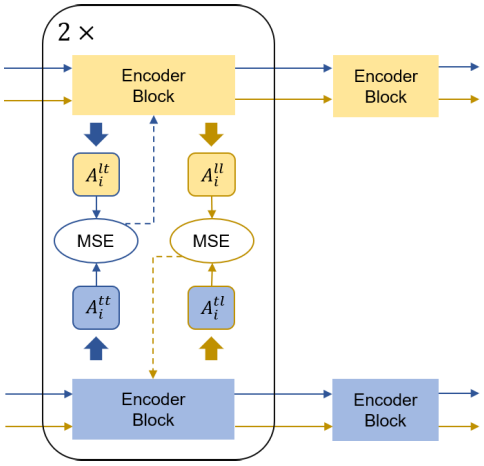


Fig. 3. The detailed process of the implicit learning in the proposed attention-based mutual knowledge distillation. Each encoder block obtains implicit knowledge by calculating the MSE loss of attention maps.

Overall, the total objective function to train the whole framework can be formulated as:

$$\mathcal{L}_t = (1-\alpha)\mathcal{L}_{t\to t} + \alpha\mathcal{L}_{l\to t},$$
$$\mathcal{L}_l = (1-\alpha)\mathcal{L}_{l\to l} + \alpha\mathcal{L}_{t\to l}, \quad (5)$$

where α is the hyper-parameter weight of distillation loss and we set it as 0.2 in our experiments. It is worth noting that AMKD is an online distillation method. In the process of training, M_t and M_l are updated at the same time, by calculating $\mathcal{L}_t$ and $\mathcal{L}_l$.

IV. Experiments

A. Dataset and Experimental Setup

In this study, 595 cases of lymph node metastasis (LNM) of papillary thyroid carcinoma were collected from the Department of Pathology, Zhongshan Hospital, Xiamen University. This study has been approved by the institutional ethics review committee of Zhongshan Hospital and patients' consent was obtained. The whole slide images of each case are stained by Hematoxylin and Eosin and scanned by a scanner (EasyScan 102, Motic, China) at a 20× magnification with 0.5um/pixel resolution. Each whole slide image was annotated with positive and negative cancer labels. The positive and negative cases are 210 and 385 respectively. To train the feature extractor, 80 cases of WSIs were annotated with polygon outlines, which encircle the cancer regions by a pathologist with ten years of experience. For the remaining 515 WSIs with slide-level labels, we conduct five-fold cross-validation five times with different seeds (*i.e.*, five repeated cross-validation) to evaluate our WSI prediction method and report the average performance with standard deviation among the five runs of combined test folds. For the auxiliary papillary thyroid carcinoma (PTC) WSIs, we collected a relatively large dataset including 687 cases and the pathologist also annotated the cancer regions of 80 WSIs to train the feature extractor.

In the stage of feature extractor training, we divided the cancer region-annotated WSIs into patches with a size of 512 × 512 pixels and got the patch-level label according to the label of cancer regions. We also randomly divided these patches into the training set and test set along the patient-level to evaluate the accuracy of our patch-level classifier. The numbers of patches in the training set and test set of lymph node metastasis were 310000 and 51200, respectively.

B. Implementation Details

We implemented the whole pipeline, including Tiny-ViT, Transformer-MIL, and AMKD in Python with the PyTorch library. The tiling patches were resized into 224 × 224 pixels for network input. The Adam optimizer was used to train both the Tiny-ViT and Transformer-MIL. To tackle the class imbalance problem during the slide-level classification stage, we employed the class-wise weighted cross-entropy loss [34] to avoid ignoring the class with fewer WSIs in model training, as shown in Equ. 3, where $\beta_0 = 0.77$ and $\beta_1 = 1.43$.

C. Evaluation Metrics

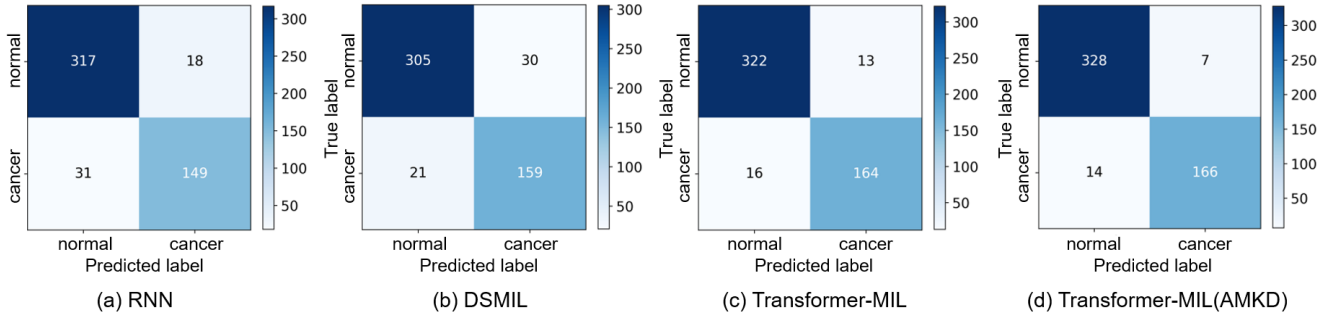
In our experiments, the area under the curve (AUC) together with the accuracy, precision, recall, and F1 score were used to evaluate the performance of our proposed method and the state-of-the-art methods. Among these, AUC is more comprehensive when comparing the performance of different methods.

D. Comparison With State-of-the-Art Methods

1) Details of the Compared Methods: Table I demonstrates the comparison results of our proposed method and other state-of-the-art methods, including (1) MAXMIN-Layer [35], (2) GCN [15], (3) Attention-MIL [12], (4) Gated Attention-MIL [12], (5) DeepAttnMISL [13], (6) RNN [9], (7) DSMIL [16]. The MAXMIN-Layer [35] obtains the prediction score of each patch by convoluting the patch-level features and then selects several maximum and minimum scores to feed into the fully connected network to obtain the slide-level prediction result. Zhao *et al.* [15] first build a graph for each WSI by taking instances as graph nodes and calculating the Euclidean distance between these instances as the edge. They further embed the instance features into the representation of WSI through the graph convolution network (GCN). Attention-MIL and Gated Attention-MIL proposed by Ilse *et al.* [12] learn the relationship between instances using the attention mechanism. When learning the complex relationships of different instances, the non-linear $tanh(\cdot)$ in the Attention module is not efficient, so the Gating Mechanism is added to improve the ability of learning complex relationships. DeepAttnMISL proposed by Yao *et al.* [13] extracts all patch-level features of the same cluster as a cluster-level feature and then uses the attention module to learn the cluster relationship, to make the attention module play a better role by reducing the input scale. RNN-based method [9] can capture long-distance interdependence via calculating the input sequence in turn and accumulating multi-step information. The DSMIL proposed by Li *et al.* firstly selects the instance with the highest possibility of canceration as the critical instance by max pooling, then obtains the weight assigned to the instance by calculating the distance between other instances and the critical instance, and finally gives the bag-level prediction by combining the two methods.

TABLE I
COMPARISONS BETWEEN OUR PROPOSED METHOD AND OTHER STATE-OF-THE-ART APPROACHES [%]

Model	AUC	Precision	Recall	F1 score
MAXMIN-Layer [35]	90.90±0.87	78.97±0.77	82.00±0.95	80.27±1.08
GCN [15]	91.85±1.05	90.19±3.98	77.48±2.03	83.03±1.37
Attention-MIL [12]	94.70±0.15	80.80±0.81	84.49±0.44	82.96±0.89
GatedAttention-MIL [12]	94.57±0.12	80.77±0.53	85.11±0.89	82.90±0.61
DeepAttnMISL [13]	94.58±0.18	82.75±1.54	85.63±0.84	84.08±0.69
RNN [9]	95.58±0.75	89.38±0.63	83.54±0.67	86.25±0.35
DSMIL [16]	95.63±0.30	85.28±1.35	87.19±1.24	86.04±0.43
Transformer-MIL	97.26±0.27	90.94±0.95	90.92±0.15	90.86±0.46
Transformer- MIL(AMKD)	**98.35±0.12**	**94.82±1.87**	**91.51±0.52**	**92.97±0.83**



(a) RNN (b) DSMIL (c) Transformer-MIL (d) Transformer-MIL(AMKD)

Fig. 4. The confusion matrix of different methods for lymph node metastasis prediction from whole slide images.

2) Comparison Analysis: We reimplemented all these previous methods based on literature and open-source codes and adopted the same feature extractor for a fair comparison. From the table, we can observe that of all the methods, MAXMIN-Layer [35] has the worst prediction performance with the AUC of 90.90%, showing that the patch relationship knowledge learned from the prediction scores is very limited. Compared with using the attention-based module to learn the relationship between instances, using Euclidean distance to determine the relationship in GCN [15] is limited, so it performs worse than attention-based MIL methods. Compared with the attention module used in Attention-MIL [12], Gated Attention-MIL [12], and DeepAttnMISL [13], the multi-head self-attention module used in Transformer-MIL can more effectively capture the characteristics of long-distance interdependence in the input instances, so it can achieve better results in multi-instance learning tasks than these three models. Although RNN [9] can also capture long-distance interdependence, its algorithm makes the effect worse than that of Transfomer-MIL when the scale of the input sequence is large, 97.26% vs. 95.58% for AUC. In addition to Transformer-MIL proposed in this paper, DSMIL [16] performs best in all other advanced methods with the AUC of 95.63% due to the innovative combination of max pooling and attention mechanism. It's worth noting that even without AMKD, our proposed Transformer-MIL still achieves better performance with AUC of 97.26%, indicating that the Transformer encoder blocks can better learn the patch relationship and aggregate the patch-level features. After using AMKD, the AUC of Transformer-MIL is further increased from the original to 98.35%. Moreover, to comprehensively compare the classification performance of all state-of-the-art methods, we show the confusion matrix of some methods and the ROC curves of all methods in Fig. 4 and 5. It can be seen that the Transformer-MIL (with AMKD) proposed in this paper is superior to other methods.

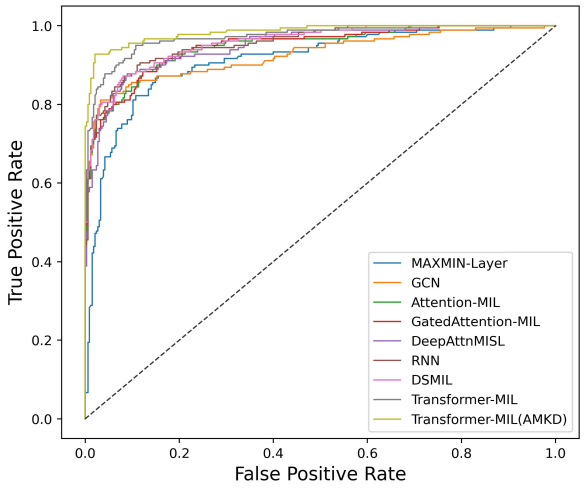


Fig. 5. The ROC curves of different methods for lymph node metastasis prediction from whole slide images.

E. Analysis of Our Framework

1) Qualitative Evaluation of Discriminative Instance Selection: Fig. 6 shows ten discriminative patches selected based on clustering (one patch for each category). The K-means clustering module divides the patch-level feature vectors into 10 categories and thus the selected patches are more diverse.

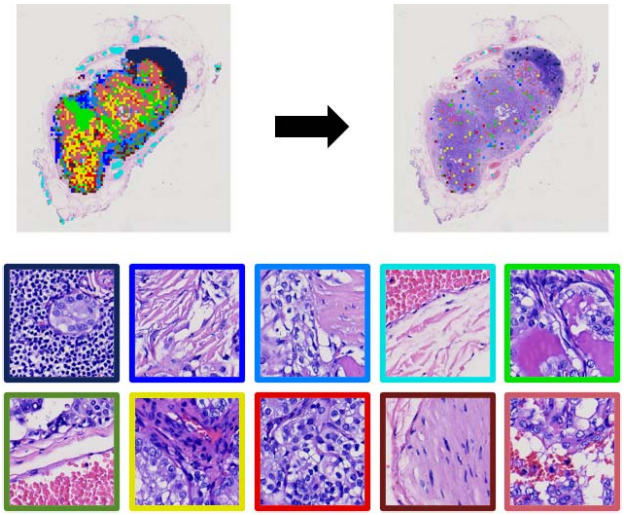


Fig. 6. Illustration of discriminative patches selected by the clustering-based instance selection strategy. Different colors represent clustering categories.

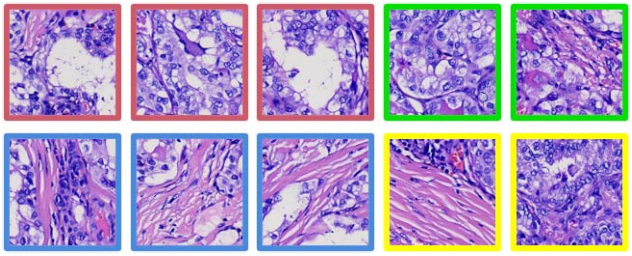


Fig. 7. Illustration of the top ten patches with the highest positive score. Different colors represent clustering categories.

In contrast, the appearance of the patches selected based on the prediction score is very similar and cannot well represent the whole slide image, as shown in Fig. 7.

2) The Impact of Different Feature Extractor Backbones: In the prediction-based instance selection method, we often need to employ a powerful feature extractor with high classification accuracy to select more discriminative patches. However, our experimental results show that in the clustering-based instance selection method, the feature extractor with high classification accuracy will not be conducive to selecting discriminative patches. We draw the clustering results of patch-level features extracted by different feature extractors, as shown in Fig. 8. ResNet50 has the highest patch-level prediction accuracy while the extracted patch-level features are mainly distributed into two categories after clustering, one of which is very similar to the cancer area. In this case, the selected patches will be very similar. As the patch-level prediction accuracy of the feature extractor decreases (*e.g.*, ResNet18), its clustering distribution will become more balanced. Although Tiny-ViT has lower patch-level prediction accuracy, the clustering distribution of the extracted patch-level features is most balanced and the selected patches are more diverse.

3) Efficiency of Lightweight Feature Extractor Tiny-ViT: To further prove the efficiency of lightweight feature extractor Tiny-ViT, we selected ResNet18 [36], ResNet50 [36], and Tiny-ViT proposed in this paper as the feature extractors to study the relationship between the prediction results of these network backbones and the complexity of the feature extractor. Table II shows the patch-level classification accuracy (Acc_P) and the slide-level AUC (AUC_S) of different feature backbones. It is observed that Acc_P of ResNet50 is obviously higher than ResNet18 and Tiny-ViT. However, when the features extracted by different feature extractors are input into Transformer-MIL, we find that AUC_S of patch-level features extracted by ResNet50 is the worst. Combined with the clustering results of features extracted by three feature extractors shown in Fig. 8, we think that Tiny-ViT can better retain the feature diversity to achieve better performance.

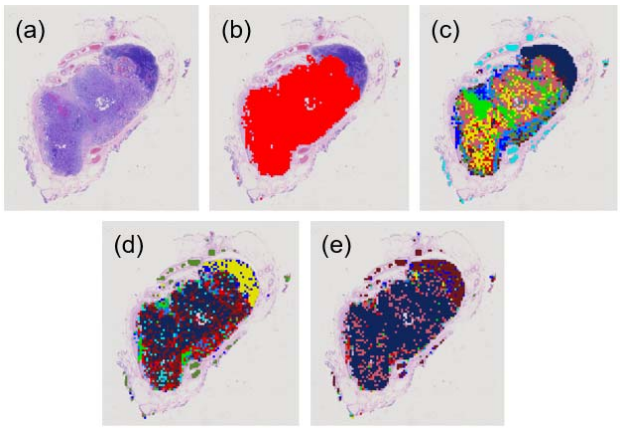


Fig. 8. The clustering results (clustered into 10 clusters) in WSI using different feature extractors: (a) original WSI from the LNM set, (b) prediction of cancer area generated by ResNet50 (feature extractor with the highest classification accuracy), (c) clustering results of Tiny-ViT, (d) clustering results of ResNet18, and (e) clustering results of ResNet50.

TABLE II
THE PERFORMANCE OF FEATURE EXTRACTORS WITH DIFFERENT NETWORK BACKBONES

Model	FLOPs	params	Acc_P	AUC_S	F1 score
Tiny-ViT	0.968G	4.972M	86.44	**97.26±0.27**	**90.86±0.46**
ResNet18 [36]	1.820G	12.407M	90.44	96.46±0.66	89.57±0.71
ResNet50 [36]	**4.112G**	**26.311M**	**91.50**	94.57±0.50	83.28±1.03

TABLE III
COMPARISON OF DIFFERENT KNOWLEDGE DISTILLATION METHODS

KD method	AUC	Precision	Recall	F1 score
AKD	97.83±0.39	94.12±0.91	90.43±0.45	92.10±0.44
AMKD	**98.35±0.12**	**94.82±1.87**	**91.51±0.52**	**92.97±0.83**
PMKD	97.28±0.23	93.49±2.30	90.59±0.49	91.74±1.05

4) Effectiveness of Utilizing Related Datasets: Besides the proposed attention-based mutual knowledge distillation (AMKD) scheme, we also implement one-way attention-based knowledge distillation (AKD) and prediction-based mutual knowledge distillation (PMKD). As our task is to predict the lymph node metastasis of papillary thyroid carcinoma, only M_l obtains implicit knowledge from M_t in AKD. Table III shows the performance of different knowledge distillation strategies. It is observed that the AUC metric of AKD is 0.52% lower than AMKD. In the one-way knowledge distillation, M_l learns two classification tasks of papillary thyroid carcinoma and its lymph node metastasis concurrently, while M_t only learns the classification of papillary thyroid carcinoma, which would weaken the knowledge correlation

TABLE IV
PERFORMANCE OF PROPOSED ATTENTION-BASED MUTUAL KNOWLEDGE DISTILLATION FROM DIFFERENT AUXILIARY DATASETS

KD dataset	AUC	Precision	Recall	F1 score
PTC	**98.35±0.12**	**94.82±1.87**	**91.51±0.52**	**92.97±0.83**
PCam [37]	97.57±0.43	93.61±0.81	90.66±0.70	92.01±0.22

between the two models and cannot better improve the classification performance. In PMKD, we make the models learn the final prediction from each other to obtain implicit knowledge. The improvement of PMKD to Transformer-MIL is the weakest among the three methods. As compared with the patch relationship knowledge obtained from the attention map, the knowledge obtained only from prediction results is very limited. In addition, to explore the influence of the pathological relationship of "the morphology of cancer cells in the primary tumor and its lymph node metastasis is similar", we chose the public gastric cancer histopathological image dataset (PCam [37], [38]) and the papillary thyroid carcinoma dataset (PTC) as the auxiliary datasets to conduct the mutual knowledge distillation. As shown in Table IV, the model guided by gastric cancer data only increased by 0.31% in AUC, while the model guided by papillary thyroid carcinoma data increased by 1.09% in AUC. It can be seen that this pathological relationship plays an important role in our proposed AMKD.

V. DISCUSSION

In this paper, we propose a novel transformer-guided multi-instance learning framework to predict lymph node metastasis of papillary thyroid carcinoma by analyzing whole slide histopathological images, where the transformer module is employed holistically, including feature extraction, feature aggregation, and knowledge transfer

Medical research shows that the 5-year survival rate of papillary thyroid carcinoma is more than 95% [1], [2]. However, the recurrence rate and mortality of patients with lymph node metastasis are significantly increased. Therefore, the study of lymph node metastasis of papillary thyroid carcinoma has great clinical value. Due to the large size of histopathological images, most of the current works mainly use multi-instance learning methods to analyze whole slide histopathological images. However, due to the small lesion area in many histopathological images, instance-space (IS) and bag-space (BS) MIL [23] cannot achieve good results. It is more effective to embed instance-level features into bag representation. Focusing on this problem, this paper proposes Transformer-MIL to ensure better embedding of instance-level features and uses attention-based mutual knowledge distillation (AMKD) to make the model learn more knowledge from the two datasets of cancer primary lesion and its lymph node metastasis. In addition, the proportion of positive and negative patches in WSI is extremely unbalanced. For example, there are only dozens of true-positive patches in some positive slides, *i.e.*, less than 1% of the total number of patches. There are also a few false-positive patches in the negative slides. The features of these false-positive patches are similar to true-positive patches. If we only focus on the patches with the highest and lowest scores in each slide, maximum and minimum scores [35] in each slide, the selected patches in the above two types of slides will be very similar, so it is difficult for the model to make an accurate diagnosis. Therefore, in the stage of discriminative patch selection, we propose a new strategy: lightweight feature extractor and clustering-based instance selection. Compared with the high-precision feature extractor and instance prediction-based selection strategy in other multi-instance learning methods, the lightweight feature extractor not only greatly saves training resources, but also makes the clustering distribution of the extracted patch-level features more balanced to select more representative and diverse patches. Fig. 9 shows the clustering results of Tiny-ViT on different WSIs, indicating the efficiency and robustness of the proposed lightweight feature extractor. As the selected patches are sparsely distributed on the WSI, the 2D location information cannot well reflect the relationship between patches. To solve this problem, we proposed clustering embedding to replace the traditional position embedding. From Table V, we can see that the performance of clustering embedding is better than position embedding in Transformer-MIL. Compared with other advanced multi-instance learning methods, Transformer-MIL achieves better results in our problem.

This is mainly because the multi-head self-attention module in Transformer-MIL can better learn the relationship among remote instances from many aspects compared with other methods. Moreover, using multiple repeated multi-head self-attention modules can make each module pay attention to a specific relationship and get a more accurate bag representation than other linear embedding methods of a single attention

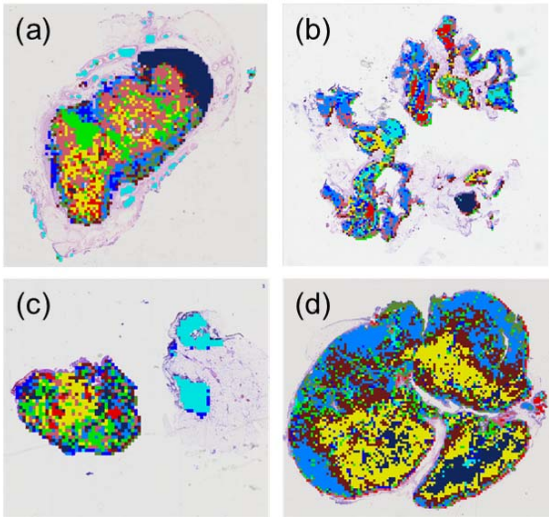


Fig. 9. The clustering results of Tiny-ViT (clustered into 10 clusters) on different WSIs: (a) True Positive WSI, (b) False Negative WSI, (c) False Positive WSI, (d) True Negative WSI.

TABLE V
THE PERFORMANCE OF DIFFERENT EMBEDDING METHODS IN TRANSFORMER-MIL

Method	AUC	Precision	Recall	F1 score
position embedding	96.61±0.49	87.75±1.82	87.98±0.80	87.76±0.89
clustering embedding	**97.26±0.27**	**90.94±0.95**	**90.92±0.15**	**90.86±0.46**

TABLE VI
T-TEST BETWEEN OUR PROPOSED METHOD AND OTHER METHODS IN THIS PAPER

	Transformer-MIL(AMKD)
MAXMIN-Layer	1.42e-07
GCN	1.71e-06
Attention-MIL	2.61e-10
GatedAttention-MIL	07.60e-11
DeepAttnMISL	5.10e-10
RNN	8.73e-05
DSMIL	1.68e-07
Transformer-MIL	6.85e-05
AKD	3.08e-02
PMKD	3.26e-05
PCam	7.61e-03

module. The knowledge transferred from PTC data also plays a vital role in improving prediction accuracy. Since the original AUC of the proposed Transformer-MIL is already very high (*i.e.*, 97.26%), further improvement over Transformer-MIL is difficult. In this case, the proposed AMKD still improves about 1.09% on AUC, showing the effectiveness of acquiring implicit knowledge from existing related annotated WSI datasets. Since the patch-level annotation of WSI is time-consuming, it is generally difficult to collect enough training samples to train a discriminative network to generate high accuracy in WSI diagnosis problems. In past works, the strategies for data expansion need to meet the condition of the same cancer type, while the proposed knowledge distillation framework utilizes the morphological similarity between primary tumor and metastatic tumor cells and provides another paradigm for data expansion by utilizing the existing related WSI datasets. In this case, it is worth using the proposed knowledge distillation paradigm to alleviate the annotation scarcity and this paradigm provides a new choice for researchers to expand histopathological data for other WSI diagnosis problems. We conduct t-test on the comparison of our method and other methods, as shown in Table VI. It is observed that the AUC improvement of our proposed method is significant (p-value < 0.05) compared with other methods in this paper.

In the future, we will test our method on other public datasets to evaluate the generalization capability of the model. At the same time, we will continue to alleviate the shortcomings of the method. This method trains two stages of feature extraction and multi-instance model separately. Although we select discriminative patch-level features by K-means clustering, these features may not be the best choice for multi-instance learning models. We will study how to design the end-to-end training strategy for our multi-instance learning framework to get better prediction results.

VI. CONCLUSION

In this paper, we propose an effective transformer-guided multi-instance learning method to detect lymph node metastasis of papillary thyroid carcinoma from whole slide histopathological images. Experimental results demonstrate that our method benefits from our proposed components, including (1) lightweight feature extractor Tiny-ViT and clustering-based instance selection, (2) Transformer-MIL, and (3) attention-based mutual knowledge distillation (AMKD). In the evaluation process, our method achieves 98.35% AUC on the test images. Compared with other advanced methods, our method shows excellent performance, which indicates that our model provides a more efficient multi-instance learning method for future histopathological image analysis. The future work includes exploring new knowledge distillation methods to learn more knowledge on limited datasets.

REFERENCES

[1] Y. Yu *et al.*, "Clinical implications of TPO and AOX1 in pediatric papillary thyroid carcinoma," *Transl. Pediatrics*, vol. 10, no. 4, p. 723, 2021.
[2] S. Ortiz *et al.*, "Extrathyroid spread in papillary carcinoma of the thyroid: Clinicopathological and prognostic study," *Otolaryngol.-Head Neck Surg.*, vol. 124, no. 3, pp. 261–265, 2001.
[3] O. Sertel, J. Kong, H. Shimada, U. V. Catalyurek, J. H. Saltz, and M. N. Gurcan, "Computer-aided prognosis of neuroblastoma on whole-slide images: Classification of stromal development," *Pattern Recognit.*, vol. 42, no. 6, pp. 1093–1103, Jun. 2009.
[4] T. J. Fuchs and J. M. Buhmann, "Computational pathology: Challenges and promises for tissue analysis," *Comput. Med. Imag. Graph.*, vol. 35, nos. 7–8, pp. 515–530, Oct./Dec. 2011.
[5] R. Li, J. Yao, X. Zhu, Y. Li, and J. Huang, "Graph CNN for survival analysis on whole slide pathological images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 174–182.
[6] J. Yao, X. Zhu, and J. Huang, "Deep multi-instance learning for survival prediction from whole slide images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 496–504.
[7] C. L. Srinidhi, O. Ciga, and A. L. Martel, "Deep neural network models for computational histopathology: A survey," *Med. Image Anal.*, vol. 67, Jan. 2021, Art. no. 101813.
[8] D. Mahapatra, S. Kuanar, B. Bozorgtabar, and Z. Ge, "Self-supervised learning of inter-label geometric relationships for Gleason grade segmentation," in *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*. Cham, Switzerland: Springer, 2021, pp. 57–67.
[9] 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, 2019.
[10] G. Quellec, G. Cazuguel, B. Cochener, and M. Lamard, "Multiple-instance learning for medical image and video analysis," *IEEE Rev. Biomed. Eng.*, vol. 10, pp. 213–234, 2017.
[11] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie, "Deep multi-instance networks with sparse label assignment for whole mammogram classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2017, pp. 603–611.
[12] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2127–2136.
[13] J. Yao, X. Zhu, J. Jonnagaddala, N. Hawkins, and J. Huang, "Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks," *Med. Image Anal.*, vol. 65, Oct. 2020, Art. no. 101789.
[14] T. Wang *et al.*, "Microsatellite instability prediction of uterine corpus endometrial carcinoma based on H&E histology whole-slide imaging," in *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2020, pp. 1289–1292.
[15] Y. Zhao *et al.*, "Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 4837–4846.
[16] B. Li, Y. Li, and K. W. Eliceiri, "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14318–14328.
[17] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–21.

[18] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[19] H. Lin, H. Chen, S. Graham, Q. Dou, N. Rajpoot, and P.-A. Heng, "Fast scannet: Fast and dense analysis of multi-gigapixel whole-slide images for cancer metastasis detection," *IEEE Trans. Med. Imag.*, vol. 38, no. 8, pp. 1948–1958, Jan. 2019.
[20] L. Wang, T. Song, T. Katayama, X. Jiang, T. Shimamoto, and J.-S. Leu, "Deep regional metastases segmentation for patient-level lymph node status classification," *IEEE Access*, vol. 9, pp. 129293–129302, 2021.
[21] H. R. Roth *et al.*, "Improving computer-aided detection using convolutional neural networks and random view aggregation," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1170–1181, May 2016.
[22] Y. Song, W. Cai, J. Kim, and D. D. Feng, "A multistage discriminative model for tumor and lymph node detection in thoracic images," *IEEE Trans. Med. Imag.*, vol. 31, no. 5, pp. 1061–1075, May 2012.
[23] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," *Pattern Recognit.*, vol. 74, pp. 15–24, Feb. 2018.
[24] Y. Yan, X. Wang, X. Guo, J. Fang, W. Liu, and J. Huang, "Deep multi-instance learning with dynamic pooling," in *Proc. Asian Conf. Mach. Learn.*, 2018, pp. 662–677.
[25] Y. Zhou, X. Sun, D. Liu, Z. Zha, and W. Zeng, "Adaptive pooling in multi-instance learning for web video annotation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Oct. 2017, pp. 318–327.
[26] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 36–46.
[27] Y. Gao, M. Zhou, and D. N. Metaxas, "UTNet: A hybrid transformer architecture for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2021, pp. 61–71.
[28] Z. Shao *et al.*, "TransMIL: Transformer based correlated multiple instance learning for whole slide image classification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 2136–2147.
[29] K. Li, L. Yu, S. Wang, and P.-A. Heng, "Towards cross-modality medical image segmentation with online mutual knowledge distillation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 1, pp. 775–783.
[30] N. Komodakis and S. Zagoruyko, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–13.
[31] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, 2010.
[32] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *Stat*, vol. 1050, p. 21, Jul. 2016.
[33] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *Stat*, vol. 1050, p. 9, Mar. 2015.
[34] S. Panchapagesan, S. Ming, A. Khare, S. Matsoukas, and S. Vitaladevuni, "Multi-task learning and weighted cross-entropy for DNN-based keyword spotting," in *Proc. INTERSPEECH*, 2016, pp. 760–764.
[35] 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*.
[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[37] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling, "Rotation equivariant CNNs for digital pathology," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 210–218.
[38] 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.