Pyramid multi-loss vision transformer for thyroid cancer classification using cytological smear

Yu, B.; Yin, P.; Chen, H.C.; Wang, Y.F.; Zhao, Y.; Cong, X.L.; ... ; Cong, L.L.

Knowledge-Based Systems 275 (2023) 110721

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Pyramid multi-loss vision transformer for thyroid cancer classification using cytological smear

Bo Yu [a,b,c,1], Peng Yin [a,b,1], Hechang Chen [a,b,*], Yifei Wang [a], Yu Zhao [a], Xianling Cong [d], Jouke Dijkstra [c], Lele Cong [e,**]

[a] *School of Artificial Intelligence, Jilin University, Changchun, 130015, China*
[b] *Engineering Research Center of Knowledge-Driven Human–Machine Intelligence, Ministry of Education, China*
[c] *Department of Radiology, Leiden University Medical Center, Leiden, 2333ZA, Netherlands*
[d] *Tissue Bank, China-Japan Union Hospital of Jilin University, Changchun, 130033, China*
[e] *Department of Neurology, China-Japan Union Hospital of Jilin University, Changchun, 130033, China*

ARTICLE INFO

Article history:
Received 20 March 2023
Received in revised form 7 June 2023
Accepted 8 June 2023
Available online 14 June 2023

Keywords:
Pyramid feature
Cytological smear
Vision transformer
Multiple instance learning
Thyroid cancer

ABSTRACT

Multi-instance learning, a commonly used technique in artificial intelligence for analyzing slides, can be applied to diagnose thyroid cancer based on cytological smears. Since smears do not have multidimensional histological features similar to histopathology, mining potential contextual information and diversity of features is crucial for better classification performance. In this paper, we propose a pyramid multi-loss vision transformer model called PyMLViT, a novel algorithm with two core modules to address these issues. Specifically, we design a pyramid token extraction module to acquire potential contextual information on smears. The pyramid token structure extracts multi-scale local features, and the vision transformer structure further obtains global information through the self-attention mechanism. Furthermore, we construct multi-loss fusion module based on the conventional multi-instance learning framework. With carefully designed bag and patch weight allocation strategies, we incorporate slide-level annotations as pseudo-labels for patches to participate in training, thus enhancing the diversity of supervised information. Extensive experimental results on the real-world dataset show that PyMLViT has a high performance and a competitive number of parameters compared to popular methods for diagnosing thyroid cancer in cytological smears.

1. Introduction

Thyroid cancer is a relatively rare form of cancer that begins in the cells of the thyroid gland [1,2]. It is becoming a leading cause of disease burden worldwide, with its incidence and mortality registering rapid growth of 169% and 87%, respectively, in recent years [3]. Benefitting from early diagnosis and treatment, thyroid cancer patients have a nearly 98% five-year survival rate, and more than 95% survive a decade. As an easy, cost-effective, and minimally invasive technique, fine needle aspiration cytology (FNAC) has been a vital preoperative diagnostic modality in evaluating thyroid cancer [4]. The pathologist looks at the morphology of cells on a cytological smear collected by FNAC under a microscope and looks for lesions or abnormalities in the cells. However, its diagnosis is considered quite challenging given that its symptoms are very similar to those of other diseases, making precise diagnoses rely heavily on clinical experience and the medical knowledge of pathologists. Deep learning (DL), a form of artificial intelligence (AI) that uses algorithms to simulate aspects of human decision-making, has recently gained much attention [5,6]. With the development of DL-powered systems that can deliver a faster and more consistent computer-aided diagnosis, many recent breakthroughs in cancer diagnosis have been in the realm of slide-driven models [7–11]. Therefore, exploring DL-based methods on FNAC cytological smears is significant for developing thyroid cancer classification [12].

DL solutions for accurate and efficient cancer screening on slides or smears are typically divided into exact and inexact supervision [13]. For exact supervision, each sample has a label corresponding to itself, and slide-level screening is the most convenient and general method for diagnosis. It resizes the size of smears for dimensionality reduction, treating the process as a classification task in computer vision [14]. Then, a deep neural network, such as a convolution neural network (CNN), is used

[*] Corresponding author at: School of Artificial Intelligence, Jilin University, Changchun, 130015, China.

[**] Corresponding author at: Department of Neurology, China-Japan Union Hospital of Jilin University, Changchun, 130033, China.

E-mail addresses: chenhc@jlu.edu.cn (H. Chen), congll18@mails.jlu.edu.cn (L. Cong).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.knosys.2023.110721
0950-7051/© 2023 Elsevier B.V. All rights reserved.

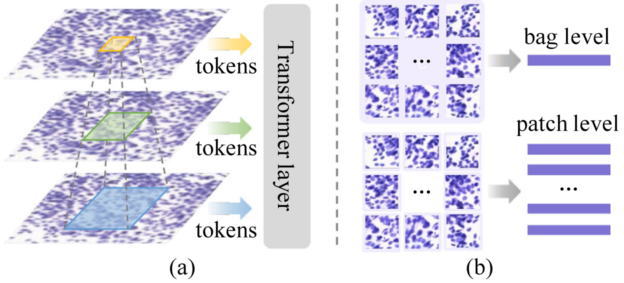

Fig. 1. (a) Multi-scale adjacency features are crucial for the ViT model. (b) Supervision information from patch-level features can further help optimize the parameters of networks.

to provide the category predictions. Due to the information loss with the resize operation, patch-level methods solve this problem by cutting the smear into patches as the input of the CNN. In addition, features from a CNN-based extractor are aggregated or directly used for analysis [15,16]. Limited by challenges in obtaining annotations of smears at the patch level, inexact supervision is appealing in medical image analysis scenarios, typically including multiple instance learning (MIL) [17,18]. A series of patches are cut from slides and formed into bags with the same slide-level label, and bag-level features are obtained from CNN with an attention network or addition operation [19]. Recent advances in Vision Transformers (ViTs) have shown that self-attention networks surpass conventional CNN models in most vision works [20,21]. Its improved variants have the tremendous potential to be an excellent feature extractor when applied to MIL models [22,23].

Despite the success that ViT-based MIL models achieved in histopathology classification tasks, there remain two challenges in classifying cytological smears: ***Challenge I:*** Extracting multi-scale features for analyzing smears is essential in the independent ViT model (Fig. 1(a)). Extracting multi-scale features has proven to be an excellent way of analyzing slides or smears [24]. Some recent works combine CNN and ViT to capture global and local information on images, thus improving the classification performance of models with multi-scale features [25,26]. However, CNN and ViT models can produce predictions through different mechanisms and representations, and it is difficult to understand why the model makes a particular prediction or diagnoses biases in the system. Furthermore, it will increase both models' computational complexity and memory requirements. Therefore, utilizing the complete ViT architecture to extract multi-scale features of the smear can not only clarify the effectiveness of the design pattern but also has the potential to reduce the complexity of the model. ***Challenge II:*** The training process of MIL lacks loss optimization for patch-level smears (Fig. 1(b)). Most of the existing MIL-based research on slide classification only focuses on bag-level supervision information [27,28]. However, compared to histopathology slides, cytological smears lack hierarchical histological features, resulting in insufficient supervised information for training. In contrast, due to the relatively uniform distribution of features in cytological smears, the patch-level smear has the potential information to supervise the training process with less noise. Therefore, simultaneously optimizing bag-level and patch-level losses during training is crucial for the smear classification task.

To resolve these challenges in ViT-based MIL methods, we propose a novel method called **Py**ramid **M**ulti **L**oss **Vi**sion **T**ransformer (PyMLViT) that aims to achieve multi-scale representations from a single input and consider the effect of patch-level information in the thyroid cancer classification task. Specifically, for ***challenge I***, we designed the pyramid token extraction module to extract multi-scale features from a cytological smear using the variant ViT structure. A pyramid-shaped token generation unit is built using a selection mechanism with different receptive fields. Subsequently, a deep self-attention encoder converts tokens from different scales into high-dimensional feature vectors. This module realizes the function of extracting multi-scale features from a single input by improving the token selection mechanism in ViT. For ***challenge II***, we designed the multi-loss fusion module to trade off the bag-level and patch-level supervision information to better guide the optimized phase of network parameters. A prediction from the attention layer and slide-level annotation generates the bag-level loss. The patch-level loss consists of multiple scales, and the loss at each scale is computed by predictions from patches and annotations from the slide. The total loss is obtained by fusing the losses of these two levels according to reasonable weight values. In summary, the contributions of this paper are as follows:

- A novel model called PyMLViT is proposed for thyroid cancer classification based on cytological smears. It is a high-performance model with relatively fewer parameters, codriven by the pyramid token extraction and multi-loss fusion modules.
- In the pyramid token extraction module, we first design a pyramid-shaped receptive field selection structure, which can obtain tokens with various scale information. Then, deep self-attention networks are used to sufficiently extract multi-scale features from tokens.
- In the multi-loss fusion module, losses are first generated from bag-level and patch-level features. Then, these two losses are fused by different weights to obtain the total loss, thereby optimizing the model parameter more reasonably.
- Extensive experiments on the in-house dataset compared with the published popular methods demonstrate the effectiveness of PyMLViT. In addition, more in-depth analyses illustrate its parameter quantity advantage compared to the CNN model.

The remainder of this paper is organized as follows. Section 2 reviews previous work on related applications by existing AI technology. Section 3 presents our solution for classifying smears with pyramid token extraction and multi-loss fusion modules. Experiments and detailed analysis are presented in Section 4, and Section 5 concludes the paper.

2. Related work

This section briefly reviews some previous popular methods of intelligent disease classification, including vision transformer, multiple instance learning methods, and research on combining them.

2.1. Vision Transformer (ViT)

The Google Brain team first proposed the Transformer model, gradually replacing RNN models such as long short-term memory as the preferred model for solving natural language processing tasks [29]. Inspired by this, the vision transformer has been proposed recently, and it is a transformer targeted at vision processing tasks such as image classification [20]. An image is split into fixed-size patches, each of which is then linearly embedded, position embeddings are added, and the high-dimensional embedded representation is fed to a standard transformer encoder. Furthermore, a typical approach usually adds an extra learnable "classification token" to the sequence for performing classification. Many scholars have used it to analyze cytopathology or histopathology because of its outstanding ability for feature

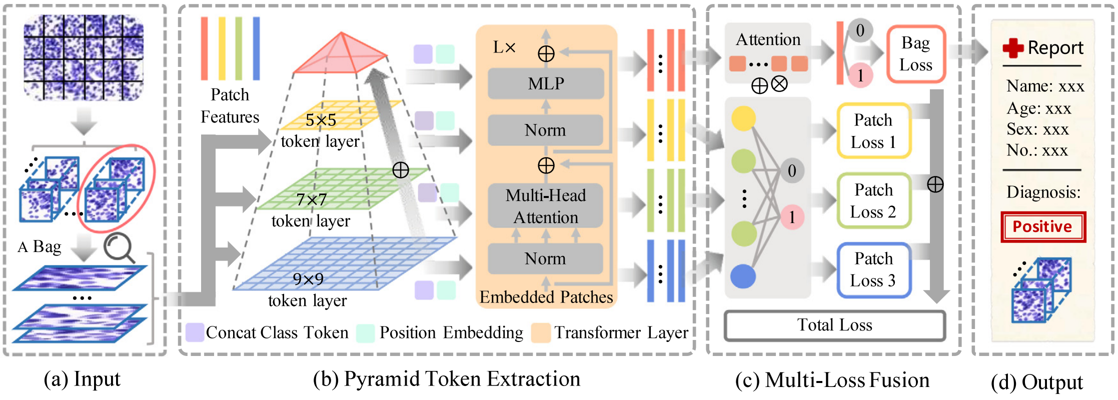

Fig. 2. An overview of the PyMLViT architecture. (a) Input: One of a bag consists of random patches from the cytological slide. (b) Pyramid token extraction module: The image is split and reconstructed using multiple scale-specific masks to form a new input, which is then converted into multi-scale features using a transformer encoder. (c) Multi-loss fusion module: The bag-level features generated by the attention mechanism are trained using slide-level annotations, while the patch-level features are trained using pseudo-labels from the slide. (d) Output: In the training phase, multiple losses are calculated and backpropagated; in the testing phase, only the bag-level prediction results are used.

extraction [30–33]. For example, a study implements a robust comparison of deep learning methods for multi-scale cytopathology cell image classification, and their research perspective was from convolutional neural networks to visual transformers [21]. With the development of this research, some variants of the ViT model for enhanced performance have also been proposed, such as T2T-ViT [34]. Accordingly, a study for improving cervical cancer classification on cytological smears is proposed with the T2T-ViT model [35]. These studies illustrate the significance of using ViT and its variants to analyze cytological smears.

2.2. Multiple Instance Learning (MIL)

Multi-instance learning (MIL) is explored as a practical mechanism, which aims at using coarse-level labels (e.g., slide-level) for learning fine-level (e.g., patch-level) images [36]. In general, it first selects multiple small patches from a whole input image with annotations according to specified rules. Then, the system randomly selects these patches to form multiple bags, and each bag has the same annotation as the original image. Finally, a multi-instance classifier is established by learning for representations of bags, which is used to predict the class of bags. Currently, this framework is used in many tasks that use DL methods to analyze cytopathology or histopathology [37,38]. For example, some studies use MIL to train a deep neural network and then apply it to diagnosing breast diseases [17,18]. Ilse et al. use the attention mechanism to train the MIL classifier, which further improved the classification performance of the model [19]. In the study of Hashimoto et al. a domain adaptation (DA) algorithm is used in MIL to balance the differences between images from various institutions [39]. Based on the above content, the MIL method is a practical approach for analyzing smears.

2.3. MIL with transformer

Compared to CNN, transformer has more robust global feature extraction capabilities. Therefore, many MIL-based studies are beginning to introduce the transformer framework to improve the analysis ability of slides. Many prior studies drew inspiration from the self-attention mechanism to build relationships between multiple instances using transformers [40,41]. For example, a transformer-based MIL approach is being developed by Shao et al. [42]. The approach effectively incorporates morphological and spatial information, providing excellent visualization and interpretability. Li et al. [43] introduced a new MIL model that utilizes a deformable transformer architecture and convolutional layers in a latent space. It can update the features of each instance by simultaneously combining the features of all instances within a bag and encoding positional context information during the representation process. Recent research on histopathology has tended to combine vision transformers with MIL, which has also achieved significant advancements [22,23]. Cai et al. present a dual-stream MIL model that leverages self-supervised contrastive learning. The model employs the Swin Transformer as its backbone for feature extraction and demonstrates accurate classification of colorectal adenoma slides based on slide-level labels [44]. Since most studies use CNN to build the MIL framework, migrating ViT to MIL requires further optimization of its structure to suit smear classification.

The core idea of PyMLViT is inspired to solve ***challenges I*** and ***II*** for classifying thyroid cancer on cytological smears. Thus, we use the pyramid token extraction module to analyze the multi-scale feature of tokens on a single input. The multi-loss fusion module is designed to fuse the bag-level and patch-level losses to optimize the parameters during the training phase.

3. Methodology

In this section, we first illustrate the framework of PyMLViT. We then give a problem formulation of our model in the second subsection. Next, we describe details about the pyramid token extraction module in the third subsection. The last subsection describes more information about the multi-loss fusion module.

3.1. PyMLViT architecture

We introduce the components of PyMLViT according to Fig. 2, which consist of four parts. *Part I:* We use patch images from a cytological slide as input (Fig. 2(a)). *Part II:* We construct a pyramid token extraction module using the multi-token layer fusion unit and deep self-attention encoder (Fig. 2(b)). *Part III:* The multi-loss fusion module consists of a bag-level loss based on the attention mechanism and patch-level losses from pseudo-labels (Fig. 2(c)). *Part IV:* The training phase adjusts the network parameters by calculating multiple loss values and then directly uses the bag-level prediction results in the testing phase (Fig. 2(d)). Next, we will present the mathematical definition of the model and subsequently describe *Part II* and *Part III* in detail, which are the two essential components of PyMLViT.

Table 1
Key notations used in this paper.

Symbol	Meaning
X_n	One of the sample
Y_n	The sample's label
N	Number of samples
B_n	A set of bags from the sample
S	A variety of scales
$G_{s=1,2,3}$	Pyramid token layers
$H_{s=1,2,3}$	Pyramid token embeddings
$F_{s=1,2,3}$	Pyramid features
H_b	Fusion embedding
F_b	Fusion pyramid features
P_b	The bag-level prediction probability
$P_{s=1,2,3}$	Patch-level prediction probabilities
L_b	The bag-level loss
$L_{s=1,2,3}$	Patch-level losses
α	The bag-level loss weight
β	The patch-level loss weight
L_T	The total loss

3.2. Problem statement

This paper focuses on thyroid cancer classification by two core stages. For convenience, we denote the first and second phases as pyramid token extraction and multi-loss fusion, respectively.

Phase I: pyramid token extraction. Given a thyroid cancer dataset $\{X_n, Y_n, B_n\}$, $n \in N$ processed by the smear as the input, we are now interested in extracting the multi-scale features and representing them from these images. Where n is the sequence number of the current sample, and the total sample size of the dataset is N. X_n represents the input image of the n sample, Y_n represents the label of the sample and corresponding bags (0 or 1), and B_n represents a set of bags of X_n that will be used for MIL learning. S represents a variety of scales, $s \in S$. Next, B_n will be sent to the pyramid token extraction module, which consists of three scales of the token layers $G_{s=1}$, $G_{s=2}$, $G_{s=3}$, and generate three token embeddings $H_{s=1}$, $H_{s=2}$, $H_{s=3}$. Then the model adds these embeddings to obtain a new fusion embedding H_b and emits them to the transformer layers to obtain hierarchical pyramid features F_b, $F_{s=1}$, $F_{s=2}$, and $F_{s=3}$.

Phase II: multi-loss fusion. First, the fusion feature F_b is integrated by the attention layer, and this stage can obtain the prediction probability P_b at the bag level. Then, three pyramid features are processed by a fully connected layer (FC), and it can obtain three prediction probabilities $P_{s=1}$, $P_{s=2}$, $P_{s=3}$ at the patch level. Next, the model calculates the bag-level loss L_b between P_b and Y_n. The various losses at the patch level $L_{s=1}$, $L_{s=2}$, and $L_{s=3}$ are calculated by $P_{s=1}$, $P_{s=2}$, $P_{s=3}$ and Y_n. Finally, we design two weights α and β for combining losses at the bag level and patch level. This model can generate the total loss L_T for training the parameters.

In general, given a cytological smear, the model transforms it into various token embeddings by pyramid token layers. Next, the transformer layers are leveraged to extract crucial features from pyramid token embeddings. Then, the model can obtain the total loss according to the prediction probabilities at the bag-level and patch-level. In the end, the prediction process can give the classification result at the bag level. An overview of the notations used in this paper is provided in Table 1.

3.3. Pyramid token extraction module

Yuan et al. [34] argue that the sequential split of the input image in vanilla ViT cannot catch important local structures between adjacent tokens (such as edges and lines), affecting classification performance. Therefore, they proposed a new token-to-token split method. By aggregating several adjacent tokens into a new token, the surrounding tokens representing local structures can be reorganized and fused to give the model a better single receptive field. Accordingly, we design a pyramid token extraction module to obtain multi-scale potential features on histological smears with various receptive fields, and Algorithm 1 describes the modeling process of the pyramid token extraction module.

Algorithm 1: Pyramid Token Extraction Module

```
Input: $B_n$ from $X_n$
Output: $F_b, F_{s=1}, F_{s=2}, F_{s=3}$
Parameter Initialization: $\theta_{s=1,2,3}, \theta_T$;
for $n = 1$ to $N$ do
    $H_{s=1} = G_{s=1}(B_n; \theta_{s=1})$;
    $H_{s=2} = G_{s=2}(B_n; \theta_{s=2})$;
    $H_{s=3} = G_{s=3}(B_n; \theta_{s=3})$;
    $H_b = H_{s=1} + H_{s=2} + H_{s=3}$;
    $F_{b,s=1,2,3} = Transformer(H_{b,s=1,2,3}; \theta_T)$;
end
```

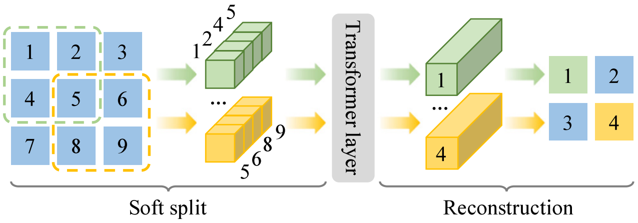

Fig. 3. First, the soft split operation recombines the tokens within the specific receptive field and then reforms smaller feature maps by the generated embeddings.

Multi-token layer fusion. The first step is called soft splitting. For a token layer, we use a square area of a specific scale to select adjacent multiple tokens, and each token has the same form as ViT. Using a receptive field of a specific scale to obtain overlapping image regions can make adjacent tokens more relevant, thereby avoiding information errors during feature extraction. The second step is to use a classical self-attention network to encode multiple adjacent tokens into a unified token. It can effectively reduce the feature dimension of multiple input tokens, making the model lightweight. The third step is reconstruction, and we restore the obtained tokens into a new feature map for the next operation. Therefore, the process can be formalized as follows:

$$\begin{aligned} T_s &= SS(I_s), \\ T'_s &= MLP(MSA(T_s)), \\ H_s &= reshape(T'_s), \end{aligned} \tag{1}$$

where I_s represents the input patches of the bag, and SS means soft split. T_s represents the token groups after the soft split operation, MSA and MLP mean the multi-head alternating self-attention layer and multilayer perceptron with layer normalization, which is G_s. θ_s and θ_T are trainable parameters that act on the multi-token layer and the transformer encoder of the model, respectively. T'_s are new token features. H_s represents the reconstructed token embedding. It converts tokens of $l \times c$ dimension into $h \times w \times c$ dimension, l represents the number of tokens, where h, w, and c represent the height, width, and channels of the new feature map, respectively. These steps can effectively reduce the size of the feature map, and the process is shown in Fig. 3.

Accordingly, the model follows the above steps with receptive fields of five, seven, and nine to obtain multi-scale feature maps. We define the process of extracting tokens using different soft split scales as a pyramid framework and perform three soft split operations on each scale to obtain the final reorganized token. The receptive field of the pyramid gradually decreases from the bottom to the top, and the model also extracts features from the global to the local. Finally, the model sums new tokens from different soft split scales to obtain a multi-scale token feature map. It contains rich semantic and detailed information, which can improve the prediction ability of the model at the bag level. The calculation process of fusion embedding H_b is as follows:

$$H_b = \sum_{s=1}^{S} H_s, \tag{2}$$

Deep self-attention encoder. We use a deep and narrow self-attention network in this unit to extract new multi-scale token features, and it consists of k layers with hidden neurons in each layer. For tokens with fixed length H_s from the multi-token layer fusion unit, a class token is added to it and then processed by sinusoidal position embedding (PE). Pyramid features F_s are obtained as follows:

$$\begin{aligned} H_s^0 &= [t_{cls}; H_s] + PE, \\ H_s^i &= MLP(MSA(H_s^{i-1})), i = 1...k \\ F_s &= FC(LN(H_s^i)), \end{aligned} \tag{3}$$

where LN is the layer normalization process, and i represents the serial number of layers in the current operation. The class token t_{cls} is a high-level representation of the entire input image, which the model learns to associate with the image's class or label. Therefore, both fusion and multi-scale pyramid features F_b, $F_{s=1,2,3}$ are obtained by the deep self-attention encoder.

3.4. Multi-loss fusion module

Most current research work based on MIL can be divided into two parts in the training phase. First, the model needs to divide large-scale slides into multiple patches, combine them into various bags, and then extract features. An aggregation model is then learned to integrate patch-level information to classify images at the bag level. This training method only considers the integrated bag-level features and lacks training of patch-level features. At the same time, the information contained in patches at specific scales is also different. Therefore, training on multi-scale patch-level features is also essential. Based on this, we further design the optimization process of the training phase, adding the supervision process at the patch level to the multi-scale feature generated by the pyramid token extraction unit. Moreover, the bag-level and patch-level training labels come from the annotations of the original slides, and there is no other labeling process. Finally, a fusion function is designed to combine the bag-level and patch-level losses to guide network training and improve classification performance. The pseudocode of this step is shown in Algorithm 2.

Algorithm 2: Multi-Loss Fusion

```
Input: $F_b$, $F_{s=1}$, $F_{s=2}$, $F_{s=3}$
Output: $Training : L_T$, $Testing : P_b$
Initialization: Attention, FC;
for $n$ = 1 to $N$ do
    $P_b = Attention(F_b; \theta_A)$;
    $P_{s=1,2,3} = FC(F_{s=1,2,3}; \theta_{FC})$;
    $L_b = CE(P_b, Y_n)$;
    $L_{s=1,2,3} = CE(P_{s=1,2,3}, Y_n)$;
    $L_T = \alpha L_b + \beta(L_{s=1} + L_{s=2} + L_{s=3})$;
end
```

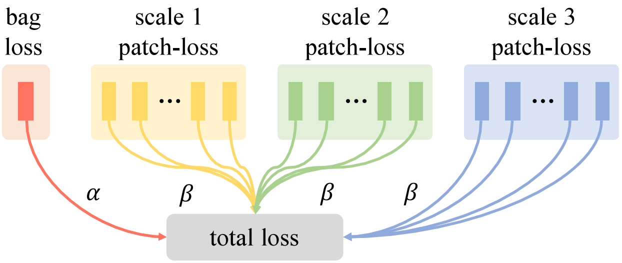

Fig. 4. The total loss is formed by weight fusion of bag-level and patch-level losses.

Bag-level Loss. The bag-level features provide global information about the bag, which can help the model identify the characteristics distinguishing positive bags from negative bags. This information can be useful in cases where the positive patches are spread out across the bag or are not easily identifiable based on their local features. In our model, each bag contains multiple patches, and we use a parameterized attention module to aggregate patch-level features in bags and generate bag-level representations. Specifically, we first reduce the dimension of the fusion features extracted by the pyramid token extraction module and use them as the input of the attention network. The attention network consists of two FC layer networks, which can provide an attention value for each patch-level feature. Then, each patch feature multiplies the corresponding attention value and performs a sum operation to obtain aggregated features for the bag. Finally, the model can give a bag-level prediction probability P_b by connecting the obtained aggregated features to an FC layer and using softmax activation. The specific operation method is as follows:

$$\begin{aligned} P_b &= softmax(FC(\sum_{m=1}^{M} a^m F_b^m)), \\ a &= softmax(FC(Tanh(FC(F_b)))), \end{aligned} \tag{4}$$

where M represents the number of patches in each bag. Next, the bag-level loss can be calculated by the cross entropy loss function (CE) as follows:

$$L_b = -\sum_{c=1}^{C} Y_n^c log P_b^c, \tag{5}$$

where C represents the total number of classes.

Patch-level Loss. The patch-level features provide local information about the patches within the bag, which can help the model identify the relevant patches that contribute to the bag's label. This information can be helpful in cases where positive patches are rare or difficult to distinguish from negative patches. Therefore, we also perform dimensionality reduction on the patch-level features and then use an FC layer to classify the dimensionality-reduced features directly. After activation by the softmax function, the patches on each scale will give a prediction result $P_{s=1,2,3}$. The labels at the patch level are the same as those at the bag level, and they all come from Y_n. Therefore, the multi-scale loss at the patch level can also be calculated using the CE function, and the process is as follows:

$$L_{s=1,2,3} = -\sum_{c=1}^{C} Y_n^c log P_{s=1,2,3}^c. \tag{6}$$

Fusion Loss. By combining bag-level and patch-level features, the model can capture local and global information about the bag,

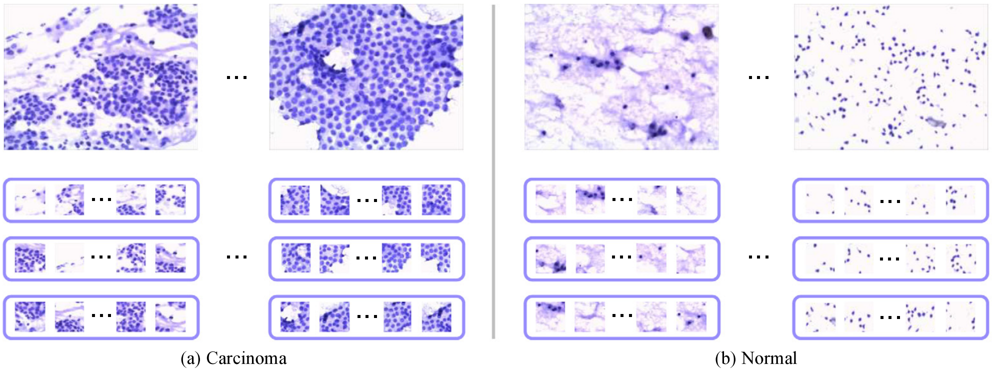

Fig. 5. Some examples of thyroid cytological smears and the process of bag integration.

leading to a more accurate classification. The model can weigh the importance of different levels based on their contribution to the training process. Therefore, we designed two hyperparameters, α and β, to balance the impact of bag-level and patch-level losses for training the neural network. Finally, the fusion loss function L_T is defined as follows:

$$L_T = \alpha L_b + \beta(L_{s=1} + L_{s=2} + L_{s=3}). \quad (7)$$

Furthermore, the fusion pattern can increase the model's robustness to noisy or irrelevant features. This is because the model can ignore features that are not useful for classification by assigning patch-level information low weights, enhancing better feature representation capability for the bag-level branch. We aim to enhance the bag-level feature representation capabilities by employing auxiliary supervision of patch-level information. Therefore, we only use the bag-level predictions and do not use patch-level results in the inference phase. The bag-level embedding is composed of multiple patches in the image, allowing for a diverse set of bags to represent sample features and thereby enhancing the generalization of algorithm evaluation. This process is shown in Fig. 4.

4. Experiments

In this section, we first describe the experimental setups and then conduct four groups of experiments to answer the following questions: **Q1**: How does our proposed method perform compared with popular techniques? **Q2**: How are different scales of tokens transformed and represented with the pyramid token extraction module? **Q3**: Is it necessary to combine patch-level loss for the training phase in the multi-loss fusion module? **Q4**: How do the number of model parameters and computational complexity of PyMLViT compare to other models?

4.1. Experimental settings

4.1.1. Dataset

Our thyroid cell smear database consisted of 560 samples from the China-Japan Union Hospital of Jilin University. The pathologist used FNAC technology to obtain diseased cells, stained them with hematoxylin, and generated digital slides through a microscope. At the same time, they conducted an ethical review and expert confirmation. The dataset only has annotations for the category, and no other annotations are used for training. The slide size of the dataset is approximately 2048 × 1536 pixels. Fig. 5 gives some examples of thyroid cancer and normal samples. To maintain the balance of positive and negative samples in the dataset, we selected 280 patient slides of thyroid cancer and 280 slides of healthy people. The above dataset is divided into 60% training data, 20% validation data, and 20% testing data.

4.1.2. Baselines and evaluation

To comprehensively compare the test results, we divide the relative research algorithms into **exact** and **inexact supervision**. Exact supervised algorithms mainly include the following:

- **slide level.** It uses the entire compressed smear as input to classify whether there is a disease [14].
- **patch level.** All patches are obtained from a smear as input, and the patch label is the same as the slide label [45].

Inexact supervised algorithms mainly include the following:

- **vanilla MIL.** It combines CNNs and attention mechanisms to extract features for classification [19].
- **DA-MIL.** It adds a domain adaptive network and attention mechanism to the CNN [39].
- **ViT-MIL.** The extractor used to generate features is changed from CNN to the vision transformer [20].
- **T2T-ViT-MIL.** It expands the local receptive field of the ViT and makes it lightweight [34].

We use three evaluation metrics for testing: accuracy (ACC), precision (P), and recall (R). In addition, we use floating point operations (FLOPs) and parameters (Params) metrics to compare the complexity of the deep learning models.

4.1.3. Implementation

The implementation of the model in this paper is based on the bag-level MIL framework. We randomly select 100 patches with a size of 224 × 224 from a smear to obtain a bag, and the maximum number of bags in each smear is limited to 50. Furthermore, since our datasets are from the same medical institution, there is no need to consider the effect of staining on the results. In the training phase, we set 30 epochs to update the network parameters and randomly generate patches in each bag during each iteration. The dimension of multi-scale and integrated features is 384, and the attention module consists of two layers of fully connected networks with 384 and 128 neurons. The model uses stochastic gradient descent to optimize network parameters with a learning rate of 0.0005 and a momentum set of 0.9.

To balance the performance and complexity of the model, as well as to maintain consistency with previous relevant research, we design a three-layer pyramid structure. Moreover, the token receptive field of T2T-ViT is seven, and we expand it upward and downward to four additional receptive fields: five, six, eight, and nine. Furthermore, we empirically validate the effects of various combinations in our experiments.

4.2. Overall experimental results (Q1)

To assess the efficacy of our model, we compare the performance of PyMLViT to related algorithms on an in-house thyroid

Table 2
The results of different algorithms on the in-house thyroid dataset.

Method type	Models	R (%)	P (%)	ACC (%)
Exact supervision	Slide level	79.76	77.90	78.57
	Patch level	68.72	70.95	70.29
Inexact supervision	Vanilla MIL	85.41	86.11	85.41
	DA-MIL	85.71	85.76	85.71
	ViT-MIL	85.71	84.21	84.82
	T2T-ViT-MIL	86.90	85.39	86.01
Ours	PyMLViT	**88.69**	**86.62**	**87.50**

Table 3
Ablation results of different token layer combinations.

Models	R (%)	P (%)	ACC (%)
PyViT(5+6+7)	88.69	84.73	86.30
PyViT(6+7+8)	83.93	**87.50**	85.71
PyViT(7+8+9)	88.09	85.71	**86.60**
PyViT(5+7+9)	**90.47**	84.02	**86.60**

dataset (see Table 2). First, we discuss some results of exact supervised methods. Slide-level methods simplify preprocessing of the original image, such as rescaling, which can compress the smears and lead to information loss. As a result, the indices of slide-level methods do not exceed 80%. On the other hand, patch-level approaches only consider independent patches, missing global information, and perform poorly, with each index at approximately 70%.

Second, inexact supervised algorithms consider bag-level information perform well, although they will be distracted by noise labels during training. The vanilla MIL method reduces the interference of noise labels at the bag level by aggregating patch-level information through the attention mechanism, increasing by more than 15% in each index. The DA-MIL algorithm introduces a domain adaptive module, but the classification result has not been significantly improved due to the lack of significant changes in the dataset. ViT methods usually perform well in general classification tasks, but when transferred to MIL, the ViT-MIL method fails to consider contextual information between different scales, and the T2T-ViT-MIL method does not consider the multi-scale case. At the same time, they do not consider that the bag-level and patch-level supervision information also plays a crucial role in the training process.

Therefore, we consider both multi-scale features and multiple types of supervisory information. Our model can obtain multi-scale structure features on single-scale input by combining the pyramid token extraction module. Moreover, we adopt the multi-loss fusion module, which can significantly improve the classification performance by combining the bag-level and patch-level losses. As shown in Table 2, PyMLViT achieves the best results in three indices, with a recall of 88.69%, a precision of 86.62%, and an accuracy of 87.50%. PyMLViT obtains the highest classification accuracy by comparing it with other algorithms. Moreover, it illustrates that the model proposed in this paper is effective for classifying thyroid cancer on cytological smears.

4.3. Validity of pyramid token extraction (Q2)

In this section, we verify the effectiveness of the model by adding the pyramid token extraction module. First, we perform ablation experiments with single token layer changes. In order to ensure that the dimension of the output feature vector is unchanged, we use five different sizes of token layers to verify the effect on this module. As shown in Fig. 6, the token layer of medium size has the highest recall with almost the same accuracy, which proves that the size of the token layer does not increase or decrease blindly to improve the performance of the model. Therefore, in the following experiments, we choose the token layer of medium size (7) as a reference to further select the combination strategy of multi-scale tokens further.

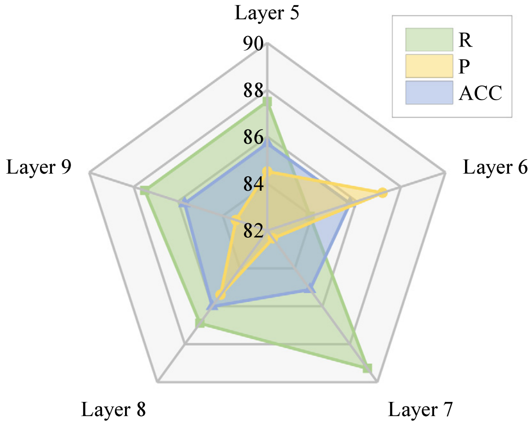

Fig. 6. The impact of different single-scale token layers on model performance.

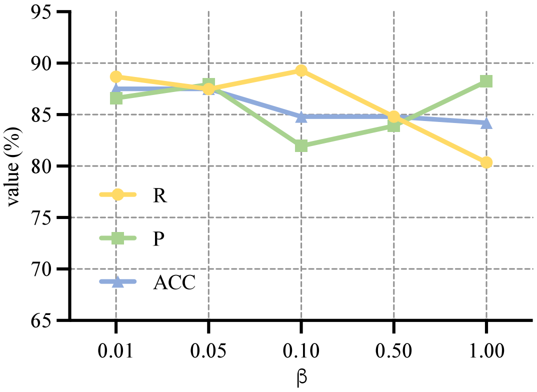

Fig. 7. Impact of β weight changes on performance.

To further verify the effectiveness of multi-scale token layer fusion, we perform ablation experiments with different combinations of token layers. We refer to the model that only combines the pyramid token extraction module as the PyViT framework. As shown in Table 3, all the PyViT models with the token layer of medium size achieve more than 88% recall, but the precision of PyViT $(6+7+8)$ increases by approximately 2% compared to other models. The accuracy of each model in this experiment is similar. PyViT $(5+7+9)$ achieves the highest recall of 90.47%; PyViT $(6+7+8)$ achieves the highest precision of 87.50%; and the PyViT $(5+7+9)$ model and PyViT $(7+8+9)$ have the highest accuracy of 86.60%.

In conclusion, the performance of the model with multi-scale token layer fusion is generally higher than that of the model with a single token layer. The pyramid token extraction module can significantly improve the classification performance of the model.

4.4. Benefit of multi-loss fusion (Q3)

In this section, we will analyze the outcomes of the multi-loss fusion module. To determine the impact of bag-level and patch-level information on model performance, we introduce two hyperparameters (α and β) to modify the importance of each. We conduct model comparison experiments, setting the hyperparameter controlling the bag-level information α to a constant value of 1. As shown in Fig. 7, the value of the hyperparameter

Table 4
Comparison of PyViT and PyMLViT.

Models	R (%)	P (%)	ACC (%)
PyViT	**90.47**	84.02	86.60
PyMLViT	88.69	**86.62**	**87.50**

governing the weight of patch-level information β determines how well the classification model performs. Given that the patch-level information is an auxiliary for bag-level training, the value of β should be less than the value of α. This demonstrates that when the patch-level and bag-level information is combined, the model's classification performance is inferior to the model's classification performance when only considering the bag-level information. This means that the patch-level information will correct the bag-level information and subsequently produce noise to interfere with the model's performance. If we define β as 1, the precision might reach 88.24%. When β is set to 0.1, the recall is 89.29%, and the accuracy is 84.82%; when β is set to 0.01, we obtain the most remarkable accuracy, 87.50%. Notably, the optimal ACC values are observed when β is set to 0.01 and 0.05. In auxiliary diagnosis, the risk of false-negatives is more significant than that of false-positives, rendering the recall more clinically significant than precision. Consequently, we ultimately opt for the results with a higher recall, corresponding to the β value of 0.01.

To further verify the effectiveness of this module, we compare the PyViT model without patch-level supervision information with the PyMLViT model. This module weights bag-level and patch-level loss information with $\alpha = 1$ and $\beta = 0.01$ so that PyMLViT can better use the available features to generalize performance on unseen details. Table 4 shows that PyMLViT achieves better performance in precision and accuracy, outperforming PyViT by 2% and 0.90%, respectively. The addition of patch-level supervision information can allow the model to better distinguish between objects and backgrounds, thereby screening meaningful information to help model classification. For example, the background of carcinoma and normal cases are repeatedly labeled as positive and negative. This conflicting supervision information makes the model give less attention to them because they are not helpful for diagnosis. Furthermore, the optimization process of PyMLViT prioritizes the reduction of false-positives, leading to high precision. PyMLViT typically indicates that it can correctly classify most positive examples and has few false-positives with high accuracy and precision. However, it might still miss some positive instances, resulting in false-positives. This situation can lead to a lower recall. The situation with PyViT is the opposite. Due to insufficient precision, PyViT will predict more true and false-positive examples, thereby increasing the value of recall. Therefore, the PyViT and PyMLViT have opposite performances on R and P & ACC.

The model performance steadily improves as β is gradually reduced, demonstrating that the patch-level information is supposed to supplement the bag-level information or that the patch-level information aids the model in considering what the bag-level information lacks.

4.5. Parameters and complexity of PyMLViT(Q4)

In this section, we will compare the parameters and complexity of the model presented in this research to previous methods. Table 5 expresses the performance of various models on our thyroid cytological smear dataset.

Regarding the number of parameters of the model, the related exact supervised methods have a slight advantage. Their parameters are more than 1000, requiring more GPU memory for calculation and increasing the development cost. In contrast, the inexact supervised methods have reduced the number of parameters, especially the T2T-ViT-MIL method using token transformation, which has minor parameters. Based on improving the classification performance, our model keeps the characteristics of low parameter quantity as much as possible. PyMLViT dramatically reduces the number of parameters compared to exact supervised models and some inexact supervised models. For example, except for T2T-ViT-MIL, the number of parameters in PyMLViT falls by a factor of three to roughly eight compared to other models. Moreover, it has approximately five times fewer parameters than the slide-level model.

Table 5
Comparison of model parameters and complexity.

Method type	Models	Params(M)	FLOPs(G)
Exact supervision	Slide level	134.27	15.50
	Patch level	134.27	15.50
Inexact supervision	Vanilla MIL	67.21	15.42
	DA-MIL	174.52	15.52
	ViT-MIL	85.71	16.85
	T2T-ViT-MIL	21.12	4.36
Ours	PyViT	21.91	4.95
	PyMLViT	21.88	17.10

In terms of complexity, the performance of exact supervision and inexact supervision is comparable, except for the T2T-ViT-MIL method. We designed the PyViT architecture based on the T2T-ViT-MIL model, which adds a multi-scale token extraction function and improves some performance while maintaining low complexity. Furthermore, we added patch-level supervision information based on PyViT, which improved the classification performance and only sacrificed a small amount of complexity compared with other algorithms. The accuracy rate of PyMLViT is higher (as shown in Table 2), which means that the model in this research is relatively light.

It shows that the strategy described in this research increases the cytological classification performance of the model with nearly no increase in the number of parameters and complexity while reducing the number of parameters significantly more than the CNN approach.

5. Conclusion

This paper proposes a novel PyMLViT model with two modules, which implements the diagnosis process of thyroid cancer with cytological smears. A pyramid-shaped multi-scale token selection mechanism is designed, and tokens from it can be further converted into high-dimensional features by a deep self-attention network. Furthermore, we design a weight-based loss function for bag-level and patch-level fusion, which can optimize the training process of the network. The study provides sound experimental evidence that PyMLViT has ultimate performance in classifying thyroid cancer and outperforms the currently popular methods. This shows that it is feasible to extract multi-scale features from a single input to improve classification accuracy, and patch-level supervision information is crucial for analyzing cytological smears. In follow-up work, an optimized PyMLViT will be adopted and applied to other cancers for classification to verify its applicability further.

CRediT authorship contribution statement

Bo Yu: Writing – original draft, Validation, Methodology, Conceptualization. **Peng Yin:** Validation, Methodology, Investigation, Conceptualization. **Hechang Chen:** Writing – review & editing, Supervision, Investigation. **Yifei Wang:** Conceptualization. **Yu Zhao:** Validation. **Xianling Cong:** Data curation. **Jouke Dijkstra:** Supervision. **Lele Cong:** Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request

Acknowledgments

This work is partially supported in part by the International Cooperation Project of Jilin Province (20220402009GH); the Special Project for Health Research Talents in Jilin Province (2020SCZ39); the Key Laboratory Project of Basic Research Division of Jilin Provincial Department of Science and Technology, China (20200601010JC); the Key R&D Projects of the Pharmaceutical and Health Industry Division of the Department of Science and Technology of Jilin Province (20210204150YY); the National Natural Science Foundation of China (61976102 and U19A2065); the National Key R&D Program of China (2021ZD0112501 and 2021ZD0112502); the China Scholarship Council Project (202206170090); the Guangdong Key Disciplines Project (2022ZDJS139).

References

[1] C. Clinic, Thyroid disease, 2022, URL: https://my.clevelandclinic.org/health/diseases/8541-thyroid-disease.

[2] M. Clinic, Thyroid disease: symptoms and treatment, 2019, URL:: https://www.mayoclinichealthsystem.org/hometown-health/speaking-of-health/thyroid-disease-symptoms-and-treatment.

[3] Y. Deng, H. Li, M. Wang, N. Li, T. Tian, Y. Wu, P. Xu, S. Yang, Z. Zhai, L. Zhou, et al., Global burden of thyroid cancer from 1990 to 2017, JAMA Netw. Open 3 (6) (2020) e208759.

[4] V. Hsiao, E. Massoud, C. Jensen, Y. Zhang, B.M. Hanlon, M. Hitchcock, N. Arroyo, A.S. Chiu, S. Fernandes-Taylor, O. Alagoz, et al., Diagnostic accuracy of fine-needle biopsy in the detection of thyroid malignancy: A systematic review and meta-analysis, JAMA Surg. (2022).

[5] Z. Li, F. Liu, W. Yang, S. Peng, J. Zhou, A survey of convolutional neural networks: analysis, applications, and prospects, IEEE Trans. Neural Netw. Learn. Syst. (2021).

[6] V.S. Lalapura, J. Amudha, H.S. Satheesh, Recurrent neural networks for edge intelligence: A survey, ACM Comput. Surv. 54 (4) (2021) 1–38.

[7] B. Yu, H. Chen, Y. Zhang, L. Cong, S. Pang, H. Zhou, Z. Wang, X. Cong, Data and knowledge co-driving for cancer subtype classification on multi-scale histopathological slides, Knowl.-Based Syst. 260 (2023) 110168.

[8] S. Prabhu, K. Prasad, A. Robels-Kelly, X. Lu, AI-based carcinoma detection and classification using histopathological images: A systematic review, Comput. Biol. Med. (2022) 105209.

[9] S. Fremond, S. Andani, J.B. Wolf, J. Dijkstra, S. Melsbach, J.J. Jobsen, M. Brinkhuis, S. Roothaan, I. Jurgenliemk-Schulz, L.C. Lutgens, et al., Interpretable deep learning model to predict the molecular classification of endometrial cancer from haematoxylin and eosin-stained whole-slide images: a combined analysis of the PORTEC randomised trials and clinical cohorts, Lancet Digit. Health 5 (2) (2023) e71–e82.

[10] O. Ciga, T. Xu, A.L. Martel, Self supervised contrastive learning for digital histopathology, Mach. Learn. Appl. 7 (2022) 100198.

[11] C.L. Srinidhi, S.W. Kim, F.-D. Chen, A.L. Martel, Self-supervised driven consistency training for annotation efficient histopathology image analysis, Med. Image Anal. 75 (2022) 102256.

[12] K.-S. Lee, H. Park, Machine learning on thyroid disease: a review, Front. Biosci.-Landmark 27 (3) (2022) 101.

[13] H. Jiang, Y. Zhou, Y. Lin, R.C. Chan, J. Liu, H. Chen, Deep learning for computational cytology: A survey, Med. Image Anal. (2022) 102691.

[14] Y. Hou, Breast cancer pathological image classification based on deep learning, J. X-Ray Sci. Technol. 28 (4) (2020) 727–738.

[15] H. Lin, H. Chen, X. Wang, Q. Wang, L. Wang, P.-A. Heng, Dual-path network with synergistic grouping loss and evidence driven risk stratification for whole slide cervical image analysis, Med. Image Anal. 69 (2021) 101955.

[16] S. Takahama, Y. Kurose, Y. Mukuta, H. Abe, M. Fukayama, A. Yoshizawa, M. Kitagawa, T. Harada, Multi-stage pathological image classification using semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 10702–10711.

[17] K. Das, S. Conjeti, A.G. Roy, J. Chatterjee, D. Sheet, Multiple instance learning of deep convolutional neural networks for breast histopathology whole slide classification, in: 2018 IEEE 15th International Symposium on Biomedical Imaging, 2018, pp. 578–581.

[18] P. Sudharshan, C. Petitjean, F. Spanhol, L.E. Oliveira, L. Heutte, P. Honeine, Multiple instance learning for histopathological breast cancer image classification, Expert Syst. Appl. 117 (2019) 103–111.

[19] M. Ilse, J. Tomczak, M. Welling, Attention-based deep multiple instance learning, in: International Conference on Machine Learning, PMLR, 2018, pp. 2127–2136.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021, URL: https://openreview.net/forum?id=YicbFdNTTy.

[21] W. Liu, C. Li, M.M. Rahaman, T. Jiang, H. Sun, X. Wu, W. Hu, H. Chen, C. Sun, Y. Yao, et al., Is the aspect ratio of cells important in deep learning? A robust comparison of deep learning methods for multi-scale cytopathology cell image classification: From convolutional neural networks to visual transformers, Comput. Biol. Med. 141 (2022) 105026.

[22] Z. Qian, K. Li, M. Lai, E.I.-C. Chang, B. Wei, Y. Fan, Y. Xu, Transformer based multiple instance learning for weakly supervised histopathology image segmentation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part II, Springer, 2022, pp. 160–170.

[23] J. Zhang, C. Hou, W. Zhu, M. Zhang, Y. Zou, L. Zhang, Q. Zhang, Attention multiple instance learning with transformer aggregation for breast cancer whole slide image classification, in: 2022 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2022, pp. 1804–1809.

[24] N. Marini, S. Otálora, D. Podareanu, M. van Rijthoven, J. van der Laak, F. Ciompi, H. Müller, M. Atzori, Multi_Scale_Tools: a python library to exploit multi-scale whole slide images, Front. Comput. Sci. 3 (2021) 684521.

[25] H. Chen, C. Li, G. Wang, X. Li, M.M. Rahaman, H. Sun, W. Hu, Y. Li, W. Liu, C. Sun, et al., GasHis-transformer: A multi-scale visual transformer approach for gastric histopathological image detection, Pattern Recognit. 130 (2022) 108827.

[26] B. Fu, M. Zhang, J. He, Y. Cao, Y. Guo, R. Wang, StoHisNet: A hybrid multi-classification model with CNN and transformer for gastric pathology images, Comput. Methods Programs Biomed. 221 (2022) 106924.

[27] H. Zhang, Y. Meng, Y. Zhao, Y. Qiao, X. Yang, S.E. Coupland, Y. Zheng, 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, 2022, pp. 18802–18812.

[28] P. Sandarenu, E.K. Millar, Y. Song, L. Browne, J. Beretov, J. Lynch, P.H. Graham, J. Jonnagaddala, N. Hawkins, J. Huang, et al., Survival prediction in triple negative breast cancer using multiple instance learning of histopathological images, Sci. Rep. 12 (1) (2022) 14527.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[30] T. Stegmüller, B. Bozorgtabar, A. Spahr, J.-P. Thiran, Scorenet: Learning non-uniform attention and augmentation for transformer-based histopathological image classification, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 6170–6179.

[31] Y. Wang, J. Guo, Y. Yang, Y. Kang, Y. Xia, Z. Li, Y. Duan, K. Wang, CWC-transformer: a visual transformer approach for compressed whole slide image classification, Neural Comput. Appl. (2023) 1–13.

[32] Z. Li, Y. Cong, X. Chen, J. Qi, J. Sun, T. Yan, H. Yang, J. Liu, E. Lu, L. Wang, et al., Vision transformer-based weakly supervised histopathological image analysis of primary brain tumors, IScience 26 (1) (2023).

[33] U. Zidan, M.M. Gaber, M.M. Abdelsamea, SwinCup: Cascaded swin transformer for histopathological structures segmentation in colorectal cancer, Expert Syst. Appl. 216 (2023) 119452.

[34] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F.E. Tay, J. Feng, S. Yan, Tokens-to-token vit: Training vision transformers from scratch on imagenet, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 558–567.

[35] C. Zhao, R. Shuai, L. Ma, W. Liu, M. Wu, Improving cervical cancer classification with imbalanced datasets combining taming transformers with T2T-ViT, Multimedia Tools Appl. (2022) 1–36.

[36] O. Maron, T. Lozano-Pérez, A framework for multiple-instance learning, Adv. Neural Inf. Process. Syst. 10 (1997).

[37] J.-G. Yu, Z. Wu, Y. Ming, S. Deng, Y. Li, C. Ou, C. He, B. Wang, P. Zhang, Y. Wang, Prototypical multiple instance learning for predicting lymph node metastasis of breast cancer from whole-slide pathological images, Med. Image Anal. (2023) 102748.

[38] Y. Yang, Y. Tu, H. Lei, W. Long, HAMIL: Hierarchical aggregation-based multi-instance learning for microscopy image classification, Pattern Recognit. 136 (2023) 109245.

[39] N. Hashimoto, D. Fukushima, R. Koga, Y. Takagi, K. Ko, K. Kohno, M. Nakaguro, S. Nakamura, H. Hontani, I. Takeuchi, 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, 2020, pp. 3852–3861.

[40] Y. Zhao, Z. Lin, K. Sun, Y. Zhang, J. Huang, L. Wang, J. Yao, SET-MIL: spatial encoding transformer-based multiple instance learning for pathological image analysis, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part II, Springer, 2022, pp. 66–76.

[41] F. Yu, X. Wang, R. Sali, R. Li, Single-cell heterogeneity-aware transformer-guided multiple instance learning for cancer aneuploidy prediction from whole slide histopathology images, IEEE J. Biomed. Health Inf. (2023).

[42] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, et al., Transmil: Transformer based correlated multiple instance learning for whole slide image classification, Adv. Neural Inf. Process. Syst. 34 (2021) 2136–2147.

[43] H. Li, F. Yang, Y. Zhao, X. Xing, J. Zhang, M. Gao, J. Huang, L. Wang, J. Yao, DT-MIL: deformable transformer for multi-instance learning on histopathological image, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24, Springer, 2021, pp. 206–216.

[44] H. Cai, X. Feng, R. Yin, Y. Zhao, L. Guo, X. Fan, J. Liao, MIST: multiple instance learning network based on swin transformer for whole slide image classification of colorectal adenomas, J. Pathol. 259 (2) (2023) 125–135.

[45] S. Graham, M. Shaban, T. Qaiser, N.A. Koohbanani, S.A. Khurram, N. Rajpoot, Classification of lung cancer histology images using patch-level summary statistics, in: Medical Imaging 2018: Digital Pathology, Vol. 10581, SPIE, 2018, pp. 327–334.