Computational and Structural Biotechnology Journal 24 (2024) 314–321

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

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

Research Article

Multi-scale window transformer for cervical cytopathology image recognition

Jiaxiang Yi [a,1], Xiuli Liu [a,1], Shenghua Cheng [b,*], Li Chen [c], Shaoqun Zeng [a]

[a] Britton Chance Center and MoE Key Laboratory for Biomedical Photonics, Wuhan National Laboratory for Optoelectronics-Huazhong University of Science and Technology, Wuhan, China
[b] School of Biomedical Engineering and Guangdong Provincial Key Laboratory of Medical Image Processing, Southern Medical University, Guangzhou, China
[c] Department of Clinical Laboratory, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

ARTICLE INFO

Keywords:
Cytopathology image recognition
Multi-scale window transformer
Convolutional feed-forward network
Cervical cancer screening

ABSTRACT

Cervical cancer is a major global health issue, particularly in developing countries where access to healthcare is limited. Early detection of pre-cancerous lesions is crucial for successful treatment and reducing mortality rates. However, traditional screening and diagnostic processes require cytopathology doctors to manually interpret a huge number of cells, which is time-consuming, costly, and prone to human experiences. In this paper, we propose a Multi-scale Window Transformer (MWT) for cervical cytopathology image recognition. We design multi-scale window multi-head self-attention (MW-MSA) to simultaneously integrate cell features of different scales. Small window self-attention is used to extract local cell detail features, and large window self-attention aims to integrate features from smaller-scale window attention to achieve window-to-window information interaction. Our design enables long-range feature integration but avoids whole image self-attention (SA) in ViT or twice local window SA in Swin Transformer. We find convolutional feed-forward networks (CFFN) are more efficient than original MLP-based FFN for representing cytopathology images. Our overall model adopts a pyramid architecture. We establish two multi-center cervical cell classification datasets of two-category 192,123 images and four-category 174,138 images. Extensive experiments demonstrate that our MWT outperforms state-of-the-art general classification networks and specialized classifiers for cytopathology images in the internal and external test sets. The results on large-scale datasets prove the effectiveness and generalization of our proposed model. Our work provides a reliable cytopathology image recognition method and helps establish computer-aided screening for cervical cancer. Our code is available at https://github.com/nmyz669/MWT, and our web service tool can be accessed at https://huggingface.co/spaces/nmyz/MWTdemo.

1. Introduction

Cervical cancer is the fourth most frequently diagnosed cancer, with an estimated 604,000 new cases and 342,000 deaths worldwide in 2020 [1]. The most commonly used cervical cancer screening method is the Papanicolaou (Pap) smear test, which involves examining cells collected from the cervix under a microscope. The widespread application of cytology screening in recent decades has been proven essential for the early detection and timely treatment of cervical cancer [2]. However, screening cervical smears under a microscope by cytologists consumes a significant amount of time and labor [3], and the accuracy of cervical cancer screening is limited by the subjective experiences of cytologists [4].

In recent years, artificial intelligence (AI) has made significant progress in medical imaging [5]. Computer-aided cervical cancer screening reduces the workload of cytologists and improves diagnostic accuracy. The key to computer-aided screening is accurate and robust cytopathology image recognition. Traditional cervical cell recognition methods mainly rely on the knowledge of cytologists to extract features such as nucleus size, shape, and nucleus-to-cytoplasm ratio [6]. Therefore, the accuracy of traditional rules-based methods is highly dependent on precise nuclear segmentation techniques and the rationality of feature engineering.

* Corresponding author.
E-mail address: chengsh2023@smu.edu.cn (S. Cheng).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.csbj.2024.04.028
Received 10 October 2023; Received in revised form 9 April 2024; Accepted 10 April 2024
Available online 16 April 2024

With the development of deep learning, convolutional neural networks (CNNs) are widely used to recognize cervical cytopathology images. Compared with traditional methods, CNNs can automatically extract features and learn mappings in an end-to-end way. Earlier, Shanthi et al. [7] and Zhang et al. [8] used shallow CNNs for cervical cell recognition. Chen et al. [9] and Li et al. [10] utilized deep CNNs such as ResNets to classify cervical cells. Later, Lin et al. [11] and Dong et al. [12] introduced cell morphology features in CNNs for cervical cell identification. Fang et al. [13] developed ordered loss with CNNs to achieve better classification accuracy. However, convolutional neural networks are based on sliding window operations for local perception and lacks in long-range modeling capability.

Transformer [14] can capture long-range information with self-attention and has been widely used in natural language processing. For the first time, Dosovitski et al. [15] established a Vision Transformer (ViT) model based on a self-attention module for image classification tasks. Some work [16–18] follows transformer encoders for computer vision tasks and achieve comparable or better performance than CNNs. Global modeling requires substantial computational resources since the computational and spatial complexity of multi-head self-attention (MSA) in Transformers is quadratic in image size rather than linear in CNNs. To alleviate this difficulty, Swin Transformer [19] proposes Window-based Multi-head Self-Attention (W-MSA) to compute Multi-head Self-Attention (MSA) within a small window rather than across the entire image. Therefore, Swin Transformer models local relationships and fuses information between windows using shifted windows as shown in Fig. 1(a). However, this approach requires two successive Swin Transformer Blocks to perform W-MSA and SW-MSA to integrate features across windows, increasing the parameter and computational cost.

To the best of our knowledge, multi-scale information is beneficial for identifying cervical cytopathology images. Large-scale information allows easier access to structural and domain information of the cell as a whole and the relationships between contexts. In contrast, small-scale information makes it easier to obtain detailed information such as cell texture and boundaries. However, most of the above work uses feature maps of different sizes at different stages without really fusing multi-scale information simultaneously.

In this paper, we propose a Multi-scale Window Transformer (MWT) for cervical cytopathology image classification. We design multi-scale window multi-head self-attention (MW-MSA) to simultaneously integrate cell features of different scales. The attention heads were divided into three groups in each layer to perform large-scale window self-attention (LWSA), medium-scale window self-attention (MWSA), and small-scale window self-attention (SWSA). Small window self-attention is used to extract local cell detail features, and large window self-attention aims to integrate features from smaller-scale window attention to achieve window-to-window information interaction. Our design enables long-range feature integration but avoids whole image self-attention (SA) in ViT or twice local window SA in Swin Transformer. We fuse the self-attention information of three different scales of windows to represent cells better. The large-scale features allow easier access to cell overall structural and image domain information. In contrast, the small-scale features make it easier to obtain detailed information, such as cell textures and boundaries. Inspired by P2T [20], we find that the convolutional feed-forward network (CFFN) is more efficient than the original MLP-based FFN for representing cytopathology images. Following PVT [16] and Swin Transformer [19], our overall model adopts a pyramid architecture to gradually decrease the feature map size and increase the channel number at different stages like CNNs. We establish two large-scale multi-center cervical cell classification datasets of two-category 192,123 images and four-category 174,138 images. Extensive comparison experiments demonstrate that our MWT outperforms state-of-the-art (SOTA) general classification networks and specialized classifiers for cytopathology images in the internal and external test sets. Besides, a series of ablation experiments verify the effectiveness of our design.

The main contributions of this paper are summarized as follows:

- We propose a multi-scale window Transformer (MWT) for cervical cytopathology image classification. The multi-scale window multi-head self-attention design enables long-range feature integration but avoids whole image SA in ViT or twice local window SA in Swin Transformer.
- We construct two multi-center cervical cell classification datasets of two-category 192,123 images and four categories of 174,138 images and demonstrate the effectiveness and superiority of MWT through extensive experiments.

2. Related work

Here, we only review CNN-based methods for cervical cytopathology image recognition and vision Transformer methods for general image classification.

2.1. Cervical cytopathology image recognition methods based on CNN

In the last five years, convolutional neural networks have been applied to cervical cytopathology image recognition. For example, Zhang et al. [8] proposed a model Deep-Pap, which directly used convolutional networks to classify cells by transferring the weights on a natural image dataset. Wu et al. [21] focused on augmenting image datasets and constructed a model DCNN to identify the cervical cell images. Lin et al. [11] classified cervical cells in Pap smears by integrating cell morphology into CNN. Dong et al. [12] combined convolutional networks with artificial features to introduce prior features of cell images into an Inception network. Chen et al. [9] proposed a CytoBrain cell recognition method for cervical lesions by a compact VGG model. Fang et al. [13] proposed a lightweight classification method based on ShuffeNet and added channel attention. Some work further introduced attention mechanisms into CNNs. However, these CNN-based cytopathology image recognition methods have disadvantages in long-range modeling since the sliding window convolution operation for local perception.

2.2. Vision transformer methods

Transformer [14] has been widely used in natural language processing. In the last three years, Transformer's powerful long-distance modeling capability has attracted much attention from computer vision researchers. For example, Dosovitski et al. [15], for the first time, established a Vision Transformer (ViT) model based on a pure self-attention module for image classification tasks. Since ViT is whole image self-attention, its computation amount is enormous. Yuan et al. [17] proposed a T2T-ViT model to simulate local information by aggregating adjacent tokens and reducing the sequence length using a step-wise recursive approach. Wang et al. [16] proposed Pyramid Vision Transformer (PVT). They introduced a progressive shrink pyramid to gradually decrease the feature map size and increase the channel number at different stages, like CNNs. Liu et al. [19] designed Swin-Transformer architecture to solve the scaling problem and high computational complexity using a window shifting strategy. Wu et al. proposed CvT [22] and introduced convolution into the vision Transformer, thus improving performance and efficiency. Wu et al. [20] proposed a pyramid pooling Transformer (P2T) with high computational efficiency and context extraction capability. For cervical cytopathology image recognition, a few works involved a vision Transformer. Recently, Khan et al. [23] used Swin Transformer to classify cervical cells.

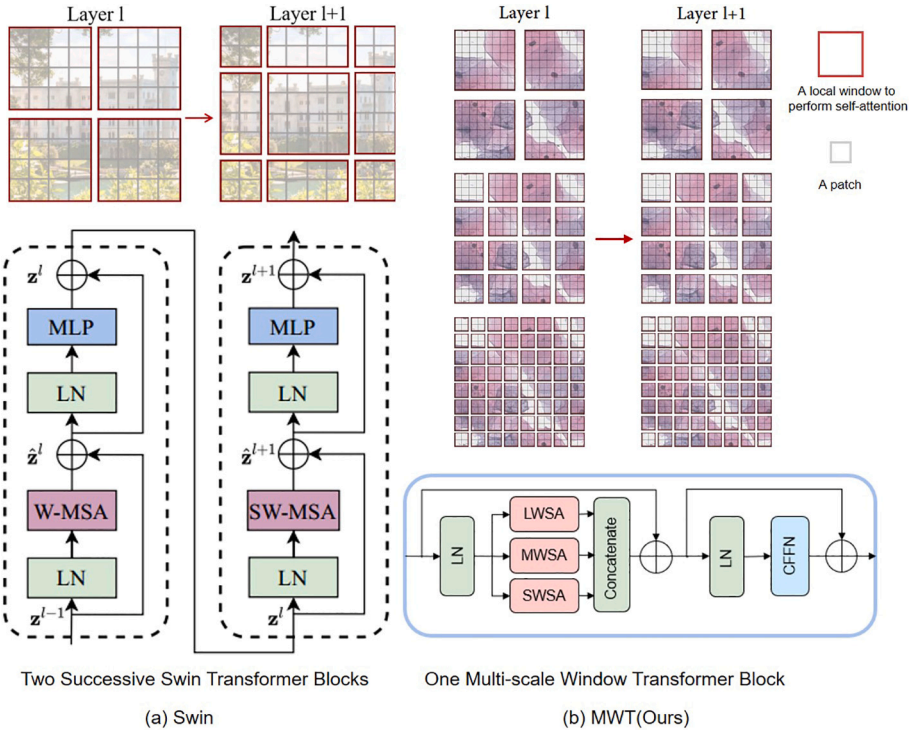

Fig. 1. (a) In contrast, the previous Swin Transformer [19] is performing window multi-head self-attention (W-MSA) and shifted-window multi-head self-attention (SW-MSA) in two adjacent layers. Two successive Swin Transformer Blocks fuse the information between windows. (b) The proposed Multi-scale Window Transformer performs self-attention computation in windows of different scales by dividing multiple self-attention heads into three groups in the same layer. MWT can simultaneously fuse large-scale, medium-scale, and small-scale information through LWSA, MWSA, and SWSA in a Multi-scale Window Transformer Block.

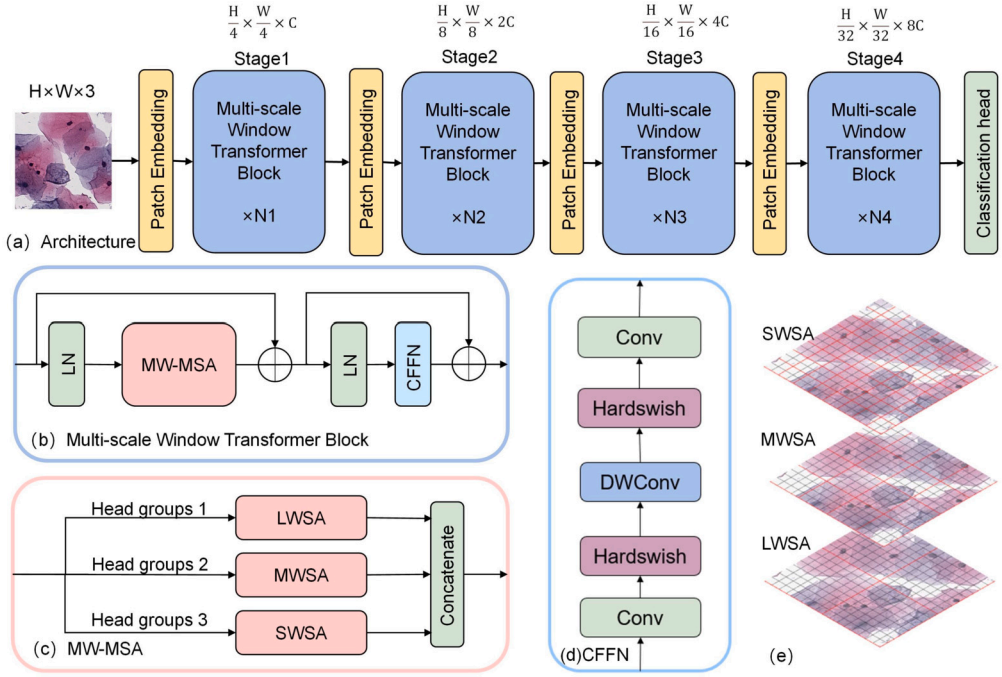

Fig. 2. (a) The overall architecture of our multi-scale window Transformer (MWT). (b) The multi-scale window Transformer block. (c) The multi-scale window multi-head self-attention (MW-MSA). (d) The convolutional feed-forward network (CFFN). (e) The schematic diagram of multi-scale window fusion.

3. Method

In this section, we elaborate on the three parts of our method: network architecture, multi-scale window-based self-attention, and convolutional feed-forward network.

3.1. Network architecture

The overall Architecture of our multi-scale window Transformer (MWT) is shown in Fig. 2(a). The input image is divided into $\frac{H}{4}\times\frac{W}{4}$ patches, with each patch treated as a "token". The original pixels of each patch are linearly projected to C by a patch embedding module.

Then, the proposed multi-scale window transformer blocks are stacked to form the network. The network consists of four stages with feature dimensions of C, 2C, 4C, and 8C, respectively. Before the last three stages, a stride of 2×2 convolution is used to down-sample the feature maps, and a linear projection doubles the number of channels. This pyramid framework allows for efficient use of computational resources while retaining vital information in the image. By down-sampling the feature maps and gradually increasing the number of channels, the network can learn features of different scales at different stages.

The patch embedding module (yellow part in Fig. 2(a)) is a crucial component of our MWT. It is responsible for dividing the input image into patches and projecting the pixels into feature space. A Conv2d implements the patch embedding module with a convolution kernel size of 4×4 and a stride of 4×4 and divides the input image into $\frac{H}{4}\times\frac{W}{4}$ patches. Each patch is projected into a C-dimension vector set to 96 in the proposed model. To achieve the fusion of adjacent patches, the embedding modules of the latter three transformer blocks are composed of Conv2d with a convolution kernel size of 2×2 and a stride of 2×2. This allows for gradual feature concentration as the network deepens. This design achieves the pyramidal structure of the overall model.

We replace standard multi-head self-attention Transformer blocks with multi-scale window Transformer blocks (blue part in Fig. 2(a)) by designing multi-scale window multi-head self-attention (MW-MSA), which will be described later. The block consists of a MW-MSA module and a CFFN module, as illustrated in Fig. 2(b). Before each MW-MSA and CFFN, a layer norm (LN) layer is applied, and a residual connection is used after each module to improve the flow of information through the network. The multi-scale window Transformer blocks are stacked N1, N2, N3, and N4 times in stage 1, stage 2, stage 3, and stage 4, respectively. The layers N1, N2, N3, and N4 are set to 2, 4, 4, and 2, respectively. MWT's modular design allows easy customization to suit different model size requirements.

3.2. Multi-scale window-based self-attention

Most visual Transformer models [15–18] used a global self-attentive mechanism. Global computation leads to a considerable computation amount proportional to the number of patches. To address this issue, Swin-Transformer [19] developed a window-based self-attention mechanism to reduce the computational cost by designing a shifting window strategy to integrate features across windows. But this manner needs consecutive twice MSA and adds parameters and computation cost.

We design multi-scale window multi-head self-attention (MW-MSA) to integrate cell features of different scales. Self-attention heads were divided into three groups, then self-attention was performed on each of the three scales. As shown in Fig. 2(b-c), MW-MSA simultaneously consists of small-scale, medium-scale, and large-scale window self-attention in a self-attention module to extract local details and global information and then integrates these features across different scales. The small-scale window self-attention mechanism (SWSA) is used to extract local details such as cell texture, features of the nucleus cytoplasm, and cell boundaries. We can take full advantage of reduced computation by using smaller windows to save resources. This allows us to capture fine-grained details that more oversized windows may miss. At the same time, the large-scale window self-attention (LWSA) integrates the information of small-scale window attention at the same layer to achieve window-to-window information interaction for long-distance modeling. This is meaningful for representing the cell's overall structural and image domain information. We also introduce medium-scale window self-attention (MWSA), which transitions between the two scales. The multi-scale window Transformer block is computed as below:

$$\overline{z}^{l} = LN(z^{l-1}) \tag{1}$$

$$\overline{z}^{l} = [z_1, z_2, z_3] \tag{2}$$

$$\hat{z} = Concat(LWSA(z_1), MWSA(z_2), SWSA(z_3)) + z^{l-1} \tag{3}$$

$$z^{l} = FFN(LN(\hat{z})) + \hat{z} \tag{4}$$

where z^{l-1} and z^{l} are the input and output of the previous Transformer block.

Our design implements long-range feature integration but avoids the full-image SA in ViT or the two-local-window SA in Swin Transformer, thus reducing the computational complexity of self-attention to some extent. Supposing each window contains M × M patches, the window-based computational complexity of Swin [19] on an image of h × w patches is shown in Equation. (5).

$$\Omega(W-MSA) = 4hwC^2 + 2M^2hwC \tag{5}$$

The computational complexity of our proposed MWT is Equation. (6),

$$\Omega(MW-MSA) = 4hwC^2 + \frac{2hwC(L_w^2 + M_w^2 + S_w^2)}{3} \tag{6}$$

where C is the feature channel number, L_w is the large window size, M_w is the medium window size, and S_w is the small window size. Since our setting L_w is equal to M (default 7) and S_w, M_w are smaller than L_w, the computational complexity of our MWT is lower than that of Swin. Specifically, we divide the features into three groups: LWSA with a window size 7×7, MWSA with a window size 4×4, and SWSA with a window size 2×2.

As shown in the schematic diagram in Fig. 2(e), the black grid represents the divided image patches, and the red grid represents the split windows. The SWSA calculates the self-attention of the nearest patches, which can better focus on the similarity relationship between several image patches nearby and thus obtain the local details of cells, such as texture and edge information. The LWSA can better discriminate information from large-scale ranges such as background and cell context. The larger window also integrates the information on the self-attention of several smaller windows and plays the role of information interaction between smaller windows. The MWSA is a balance between SWSA and LWSA. In summary, the multi-scale window self-attention mechanism achieves the simultaneous integration of multi-scale windows at the same layer. When computing self-attention, we follow Swin [19], using relative position bias. The following is the formula for calculating self-attention.

$$Attention(Q,K,V) = SoftMax(\frac{QK^T}{\sqrt{d^k}} + B)V \tag{7}$$

Here, $Q, K, V \in R^{(M^2\times d)}$ are the query, key, and value matrices, d is the query/key dimension, and M^2 is the number of patches in a window.

3.3. Convolutional feed-forward network

Feed-forward networks are an essential part of the Transformer block and are often used to compensate for the possible lack of fitting ability in attention mechanisms. The classical visual Transformer [15] uses linear fully connected layers as a feed-forward neural network (FFN). Inspired by [20], we think the convolutional feedforward network is more accessible for modeling image feature spatial mapping in vision recognition tasks and reduces computation parameters. Thus, we use a convolutional feed-forward network to replace the classical MLP-based feed-forward network.

As shown in Fig. 2(d), the convolutional feed-forward network (CFFN) consists of two convolutional layers (Conv), one depthwise separable convolution layer (DWConv), and two Hard-Swish activation layers between them. The Hard-Swish activation function is proven to perform better than ReLU on deeper models.

4. Experimental analysis

We performed experiments on our private multi-center cervical cytopathology image classification datasets. In the following, we first introduce the datasets, training details, and evaluation metrics. Then, we

Table 1
Dataset of two-category cytopathology images.

	Training Set		Validation Set		Test Set		Total	
	pos	neg	pos	neg	pos	neg	pos	neg
data1	22,559	45,118	2,820	5,640	2,820	5,640	28,199	56,398
data2	7,862	15,725	983	1,966	983	1,966	9,828	19,656
data3(external)	0	0	0	0	3,376	6,752	3,376	6,752
data4	1,898	3,795	237	474	237	474	2,372	4,744
data5	3,211	6,422	401	803	401	803	4,014	8,028
data6	1,964	3,928	246	491	246	491	2,455	4,910
data7	6,242	12,483	780	1,560	780	1,560	7,802	15,604
data8(external)	0	0	0	0	2,450	4,900	2,450	4,900
data9	1,079	2,158	135	270	135	270	1,349	2,698
data10	1,757	3,514	220	439	220	439	2,196	4,392

Table 2
Dataset of four-category cytopathology images.

	Training Set				Validation Set			
	ASCUS	HSIL	LSIL	neg	ASCUS	HSIL	LSIL	neg
data1	4,188	14,378	3,994	45,118	524	1,797	499	5,640
data2	4,082	3,238	542	15,725	510	405	68	1,966
data3(external)	0	0	0	0	0	0	0	0
data4	1,316	419	162	3,795	165	52	20	474
data5	3,014	22	175	6,422	377	3	22	803
data6	1,711	10	243	3,928	214	1	30	491
data7	2,435	2,587	1,219	12,483	304	323	152	1,560
	Test Set				Total			
	ASCUS	HSIL	LSIL	neg	ASCUS	HSIL	LSIL	neg
data1	524	1,797	499	5,640	5,235	17,972	4,992	56,398
data2	510	405	68	1,966	5,103	4,048	677	19,656
data3(external)	2,365	709	302	6,752	2,365	709	302	6,752
data4	165	52	20	474	1,645	524	203	4,744
data5	377	3	22	803	3,767	28	219	8,028
data6	214	1	30	491	2,139	12	304	4,910
data7	304	323	152	1,560	3,044	3,234	1,524	15,604

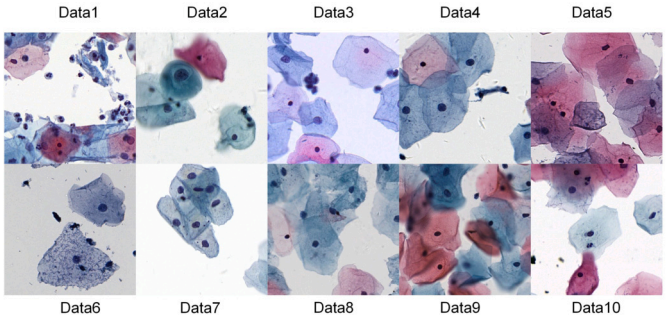

Fig. 3. Examples of the multi-center datasets of cervical cytopathology images.

compare the proposed multi-scale window Transformer network with the previous state-of-the-art CNNs and Transformers. Finally, we ablate the essential design elements of our method.

4.1. Datasets

As shown in Table 1 and Table 2, we established two multi-center cervical cell classification datasets of two-category 192,123 images (ten cohorts) and four-category 174,138 images (seven cohorts). The dataset was obtained from Hubei Maternal and Child Health Hospital and Tongji Medical College of Huazhong University of Science and Technology in Hubei Province, China. The images were captured by different instruments, including instruments from 3DHISTECH (with 0.234 μm/pixel, under × 20 magnification), Wuhan National Laboratory for Optoelectronics (WNLO) homemade equipment (with 0.293 μm/pixel, under × 20 magnification), and Shenzhen Shengqiang Technology Co., Ltd. (with 0.180 μm/pixel, under × 40 magnification).

The datasets encompass various variables, such as diverse hospitals, production batches, and imaging instruments. These factors inherently influence image attributes like brightness, contrast, and resolution. These variations effectively reflect the heterogeneity typically encountered in real-world image acquisition scenarios. As shown in Fig. 3, there are differences in the staining style and resolution of cervical cell images due to the differences in staining protocols and imaging scanners of different hospitals. Therefore, model generalization is critical for multi-center cervical cell datasets. Using the multi-source data, we can better train our model with diverse data distributions and validate its generalization by setting an external test set.

4.1.1. Two categories

The two-category dataset includes positive and negative cells. The specific data composition and division are shown in Table 1. To facilitate training our model on diverse data and evaluating its generalization, datasets 3 and 8 were randomly selected from our ten datasets as the external test set, and divided the internal data set into training, validation, and test sets in the ratio of 8:1:1. There are a total of ten cohorts and 192,123 images. The ratio of negative images to positive images is 2:1.

4.1.2. Four categories

The four-category dataset includes ASCUS, HSIL, LSIL, and negative. ASCUS represents atypical squamous cells of undetermined significance. HSIL refers to high-grade squamous intraepithelial lesions, and LSIL refers to low-grade squamous intraepithelial lesions. The specific dataset composition and data division are shown in Table 2. As with the

Table 3
Comparison of experimental results of two-category classification.

Models	Params(M)	GFLOPs	Internal Test Acc(%)	Data3 Acc(%)	Data8 Acc(%)
ConvNeXt-Tiny	27.8	4.45	93.3	91.0	83.2
Swin-Tiny	27.5	4.37	95.2	91.8	85.7
CvT-13	19.6	4.06	95.1	93.6	87.4
T2T-ViT_t-14	21.1	4.35	96.0	92.1	83.2
PVT_v2_b2	24.9	3.90	95.5	93.2	88.3
P2T-Small	23.6	3.65	94.8	92.4	88.6
Deeppap	13.4	1.09	93.5	89.8	85.7
Wu et al	58.3	1.13	94.1	91.0	85.6
Fang et al	1.3	0.15	95.3	91.6	86.0
MWT	19.8	3.53	**96.1**	**94.3**	**88.9**

Table 4
Comparison of experimental results of four-category classification.

Models	Params(M)	GFLOPs	Internal Test Acc(%)	Data3 Acc(%)
ConvNeXt-Tiny	27.8	4.45	85.0	72.9
Swin-Tiny	27.5	4.37	88.4	77.9
CvT-13	19.6	4.06	88.8	77.8
T2T-ViT_t-14	21.1	4.35	**90.3**	77.3
PVT_v2_b2	24.9	3.90	85.9	78.5
P2T-Small	23.6	3.65	88.7	75.9
Deeppap	13.4	1.09	87.4	74.5
Wu et al	58.3	1.13	86.7	76.3
Fang et al	1.3	0.15	88.1	75.7
MWT	19.8	3.53	89.1	**79.1**

two-category classification, we use dataset 3 as an external test set. We divided the internal dataset into training, validation, and test sets in the ratio of 8:1:1. There are seven cohorts and 174,138 images. The ratio of negative to positive images is 2:1, and the ratio of the three positive categories is natural.

4.1.3. Mendeley LBC dataset

The Mendeley LBC dataset [24] includes NILM, LSIL, HSIL, and SCC. NILM represents negative for intraepithelial lesion or malignancy. HSIL refers to high-grade squamous intraepithelial lesions. LSIL refers to low-grade squamous intraepithelial lesions. SCC refers to Squamous Cell Carcinoma. Consistent with the previous setup, we divided the internal dataset into training, validation, and test sets in the ratio of 8:1:1. There are a total of 963 images with a size of 2048 × 1536 pixels.

4.2. Evaluation metrics

We calculate the accuracy to evaluate our proposed model and other current SOTA models, and the accuracy is defined as follows.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

True positive (TP) refers to the number of positive samples predicted correctly. True negative (TN) refers to the number of negative samples predicted correctly. False positives (FP) and false negatives (FN) are the number of negative and positive samples classified incorrectly.

In addition, we also calculated the Params and FLOPs of the model using Python's "thop" package, as shown in Table 3.

4.3. Experiment setup

All models were run on the PyTorch framework, and the comparison models used the official codes. For training all models, we used the data augmentation methods [25] for improving model generalization, and then we resized the input images to 224×224 pixels. We randomly selected 64,000 positive and 64,000 negative images from the training set as an epoch with a batch size 64. We trained our model and all the comparison models from scratch for 120 epochs without inheriting any weights when we trained two-class classification. To save training time, we transferred the last model parameters from two-class classification to four-class training and then further trained 40 epochs. For the Mendeley LBC dataset, we trained all models from scratch for 120 epochs except ConvNeXt-Tiny (ConvNeXt-Tiny only achieved 68.4% accuracy with the default 120 epochs, thus we trained it with more epochs). All models were trained using the AdamW [26] optimizer using a cosine decay learning rate scheduler [27]. A batch size of 64, an initial learning rate of 0.0001, and a weight decay 0.05 were used. We uniformly selected the model with the best performance on the validation set over all models as the final model and then evaluated its performance on the test sets.

4.4. Experiment results

To validate the superiority of MWT, we compared it with SOTA general classification Transformers networks including Swin Transformer [19], CvT [22], PVT [16], T2T-ViT [17] and P2T [20], and recent SOTA convolutional network ConvNeXt [28], and specialized classifiers for cytopathology images including DeepPap [8], Wu's method [21] and Fang's method [13]. For the general classification network, we selected a suitable configuration for each model to achieve similar parameter numbers and computation amounts. For classifiers specialized for cytopathology images, we used the default configuration of these methods.

In the two-category classification task, we tested all models' performance using internal and external test sets. The results are presented in Table 3. For fairness, we compared our model's accuracy with state-of-the-art networks with comparable parameter numbers and FLOPs. The results show that our MWT outperforms ConvNeXt-Tiny, Swin-Tiny, CvT-13, PVTv2-b2, T2T-ViT_t-14, and P2T-Small on the external test dataset data3, with improvements by 3.3%, 2.5%, 0.7%, 1.1%, 2.2% and 1.9%, respectively. Similarly, on external test dataset data8, our model's accuracy is significantly higher than these models by 5.7%, 3.2%, 1.5%, 0.6%, 5.7%, and 0.3%, respectively. Our MWT also achieves the best accuracy on the internal test set. For specialized

Table 5
Comparison of experimental results of four-category classification on the Mendeley LBC dataset.

Models	Params(M)	GFLOPs	Acc(%)
ConvNeXt-Tiny	27.8	4.45	88.4
Swin-Tiny	27.5	4.37	92.6
CvT-13	19.6	4.06	87.4
T2T-ViT_t-14	21.1	4.35	90.5
PVT_v2_b2	24.9	3.90	89.5
P2T-Small	23.6	3.65	91.6
Deeppap	13.4	1.09	84.2
Wu et al	58.3	1.13	84.2
Fang et al	1.3	0.15	91.6
MWT	19.8	3.53	**94.7**

classifiers for cytopathology images, our MWT has a more noticeable improvement of about three percentage points on the external test sets and 0.8-2.6 percentage points on the internal test set.

In the four-class classification task, we also tested all models' performance using internal and external test sets. The results are presented in Table 4. The results show that our MWT outperforms ConvNeXt-Tiny, Swin-Tiny, CvT-13, PVTv2-b2, T2T-ViT_t-14, and P2T-Small on the external test set Data3, with improvements of 6.2%, 1.2%, 1.3%, 0.6%, 1.8% and 3.2%, respectively. On the internal test set, T2T-ViT obtains the highest accuracy of 90.3%. Our MWT achieves the sub-optimal accuracy of 89.1% and outperforms other methods by an accuracy improvement of 0.3%-4.1%. For specialized classifiers for cytopathology images, our MWT has a noticeable improvement of 2.8%-4.6% on the external test set and 1.0%-2.4% on the internal test set. These results demonstrate the superiority of our model compared to these state-of-the-art models. The higher accuracy of our model on both internal and external test sets indicates that our model can generalize well to new data and is robust enough to perform well in multi-center scenarios.

The Mendeley LBC dataset classification task results are presented in Table 5. The results show that our MWT outperforms ConvNeXt-Tiny, Swin-Tiny, CvT-13, PVTv2-b2, T2T-ViT_t-14, and P2T-Small with improvements of 6.3%, 2.1%, 7.3%, 5.2%, 4.2% and 3.1%, respectively. For specialized classifiers for cytopathology images, our MWT has a noticeable improvement of 3.1%-10.5%. These results also demonstrate the superiority of our model compared to these state-of-the-art models.

4.5. Ablation studies

In this section, we performed ablation studies to analyze the effectiveness of our design and hype-parameter choice in MWT. We explored the influence of the number of scales, window size in self-attention, convolutional feed-forward network, and stage distribution configuration. We set up a baseline model of MWT, which integrates window self-attention of three scales with CCFN. The window sizes are 2, 4, and 7, respectively. The depths N1, N2, N3, and N4 of the four stages of the baseline model are set to 2, 4, 4, and 2, respectively. The number of channels C is set to 96. We changed these configurations to analyze the influence.

4.5.1. Number of scales

To investigate the effect of multi-scale windows, we conducted the following ablation experiments. LW represents the self-attention within a window size of 7×7; MW represents the self-attention within a window size of 4×4; and SW represents the self-attention computation within a window size of 2×2. To ensure the fairness of comparison, we controlled the same number of channels for the three models as 96. As a result, the number of channels for each scale of self-attention for the two-scale windows is C/2, and the number of channels for each scale of self-attention for the three-scale windows is C/3. The experimental results in Table 6 show that multi-scale windows, especially three-scale

Table 6
Ablation study results of the number of scales.

Scales	Internal	Data3 Acc(%)	Data8 Acc(%)
LW	95.9	93.8	88.3
LW+MW	96.2	94.0	88.8
LW+MW+SW	96.1	94.3	88.9
MW+SW	95.8	94.0	87.9
SW	95.9	94.0	88.5

Table 7
Ablation study results of the window sizes.

Window_Sizes	Internal	Data3 Acc(%)	Data8 Acc(%)
(4,7,12)	96.4	94.5	88.8
(2,7,12)	96.2	93.8	88.4
(2,4,7)	96.1	94.3	88.9

Table 8
Ablation study results of the convolutional feed-forward network.

FFN	Internal	Data3 Acc(%)	Data8 Acc(%)
MLP	95.3	92.3	86.8
CFFN	96.1	94.3	88.9

windows, significantly improve the classification accuracy of external test data. The experimental results also reveal different contributions from windows of different scales, and we explore the effect of window size in 4.5.2. These results demonstrate that multi-scale windows can improve the model's performance and generalization. Compared with other Transformers, our designed multi-scale window multi-head self-attention enables long-range feature integration but avoids whole image SA in ViT or twice local window SA in Swin Transformer. Thus, MWT has fewer parameter numbers and FLOPs.

4.5.2. Window size

To explore suitable configurations for the window size, we compared different window sizes in the Transformer. As shown in Table 7, we set different window size combinations:(4,7,12), (2,7,12), and default (2,4,7). On the external test sets, the window size combination of (2,4,7) has higher accuracy than the combination of (2,7,12) and has lower accuracy than the combination of (4,7,12). This suggests that 1) continuous window sizes are more reasonable; 2) bigger window sizes have a slight advantage but also increase computation cost. Therefore, we set the window size combination of (2,4,7) as our final configuration.

4.5.3. Convolutional feed-forward network

To verify the effectiveness of the convolutional feed-forward network (CFFN), we compared it with the traditional MLP-based FFN that uses fully connected layers. The results in Table 8 show a significant improvement in the performance of MWT with CFFN. CFFN is more conducive to capturing local spatial mapping for cervical cytopathology image recognition than traditional FFN. Moreover, using convolution effectively reduces computation costs and improves training efficiency.

4.5.4. Stage distribution

To obtain a better distribution of layers per stage, we controlled the total number of layers in the MWT model to be 12 and adjusted the number of layers N1, N2, N3, and N4 in different stages as shown in Table 9. The results of the ablation experiment show that the best performance is achieved when the number of layers in each stage is (2,4,4,2). The results indicate that 1) shallower layers with more detailed information are critical for cervical cytopathology image recognition; 2) a relatively balanced distribution of layers in each stage is essential. This

Table 9
Ablation study results of the stage distribution.

(N1,N2,N3,N4)	Internal	Data3 Acc(%)	Data8 Acc(%)
(2,3,4,3)	96.2	94.0	88.8
(1,1,8,2)	95.1	93.0	87.2
(2,2,6,2)	95.9	94.0	88.7
(2,4,4,2)	96.1	94.3	88.9

suggests that a reasonable network architecture design is vital for high performance.

5. Conclusion

In this paper, we propose a multi-scale window Transformer (MWT) for cervical cytopathology image recognition. We design multi-scale window multi-head self-attention to simultaneously integrate cell features of different scales. Unlike other vision Transformers, our multi-scale window self-attention design enables long-range feature integration. Still, it avoids whole image SA in ViT or twice local window SA in Swin Transformer. On two multi-center large-scale cervical cell classification datasets of two and four categories, we prove the superiority of MWT compared with SOTA vision Transformers and demonstrate the generalization of MWT on the external test sets. In conclusion, our work provides a reliable cytopathology image recognition method and helps establish computer-aided screening for cervical cancer.

Currently, self-supervised pretraining of Transformers on large-scale natural images is developing rapidly. Self-supervised pretraining is proven to help achieve label-efficient image recognition. Massively unlabeled cell images are easy to obtain in computational cytopathology, but establishing large-scale annotated cell images with doctor knowledge is costly. Thus, studying self-supervised pretraining of cytopathology images and label-efficient cell recognition is meaningful and an opportunity in this field. In this work, we used about 6,4000 positive cell annotations to train the recognition model. In the future, we will explore how to establish accurate and robust cervical cell recognition models with 10^2 - 10^3 annotations.

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.

Acknowledgement

This work is supported by the National Natural Science Foundation of China project (grant 62201221, 62375100), China Postdoctoral Science Foundation (grant 2021M701320, 2022T150237), and Southern Medical University Research Start-up Foundation.

References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021;71:209–49. https://doi.org/10.3322/caac.21660.

[2] Cao L, Yang J, Rong Z, Li L, Xia B, You C, et al. A novel attention-guided convolutional network for the detection of abnormal cervical cells in cervical cancer screening. Med Image Anal 2021;73:102197. https://doi.org/10.1016/j.media.2021.102197.

[3] Lin H, Chen H, Wang X, Wang Q, Wang L, Heng PA. Dual-path network with synergistic grouping loss and evidence driven risk stratification for whole slide cervical image analysis. Med Image Anal 2021;69:101955. https://doi.org/10.1016/j.media.2021.101955.

[4] Sarenac T, Mikov M. Cervical cancer, different treatments and importance of bile acids as therapeutic agents in this disease. Front Pharmacol 2019;10:484. https://doi.org/10.3389/fphar.2019.00484.

[5] Jiang H, Zhou Y, Lin Y, Chan RCK, Liu J, Chen H. Deep learning for computational cytology: a survey. Med Image Anal 2023;84:102691. https://doi.org/10.1016/j.media.2022.102691.

[6] Lin CH, Chan YK, Chen CC. Detection and segmentation of cervical cell cytoplast and nucleus. Int J Imaging Syst Technol 2009;19:260–70. https://doi.org/10.1002/ima.20198.

[7] P BS, Faruqi F, K SH. Deep convolution neural network for malignancy detection and classification in microscopic uterine cervix cell images. Asian Pac J Cancer Prev 2019;20:3447–56. https://doi.org/10.31557/APJCP.2019.20.11.3447.

[8] Zhang L, Le L, Nogues I, Summers RM, Liu S, Yao J. Deeppap: deep convolutional networks for cervical cell classification. IEEE J Biomed Health Inform 2017;21:1633–43. https://doi.org/10.1109/JBHI.2017.2705583.

[9] Chen H, Liu J, Wen QM, Zuo ZQ, Liu JS, Feng J, et al. Cytobrain: cervical cancer screening system based on deep learning technology. J Comput Sci Technol 2021;36:347–60. https://doi.org/10.1007/s11390-021-0849-3.

[10] Li J, Dou Q, Yang H, Liu J, Fu L, Zhang Y, et al. Cervical cell multi-classification algorithm using global context information and attention mechanism. Tissue Cell 2022;74:101677. https://doi.org/10.1016/j.tice.2021.101677.

[11] Lin H, Hu Y, Chen S, Yao J, Zhang L. Fine-grained classification of cervical cells using morphological and appearance based convolutional neural networks. IEEE Access 2019;7:71541–9. https://doi.org/10.1109/access.2019.2919390.

[12] Dong N, Zhao L, Wu CH, Chang JF. Inception v3 based cervical cell classification combined with artificially extracted features. Appl Soft Comput 2020;93. https://doi.org/10.1016/j.asoc.2020.106311.

[13] Fang S, Yang J, Wang M, Liu C, Liu S. An improved image classification method for cervical precancerous lesions based on shufflenet. Comput Intell Neurosci 2022;2022:9675628. https://doi.org/10.1155/2022/9675628.

[14] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in neural information processing systems; 2017. p. 6000–10.

[15] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Houlsby N. An image is worth 16x16 words: transformers for image recognition at scale. In: Proceedings of the ninth international conference on learning representations (ICLR); 2021.

[16] Wang W, Xie E, Li X, Fan DP, Shao L. Pyramid vision transformer: a versatile backbone for dense prediction without convolutions. In: IEEE/CVF international conference on computer vision (ICCV); 2021.

[17] Yuan L, Chen Y, Wang T, Yu W, Shi Y, Tay FE, et al. Tokens-to-token vit: training vision transformers from scratch on imagenet. In: IEEE/CVF international conference on computer vision (ICCV); 2021.

[18] Touvron H, Cord M, Douze M, Massa F, Jégou H. Training data-efficient image transformers & distillation through attention; 2021.

[19] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: IEEE/CVF international conference on computer vision (ICCV); 2021.

[20] Wu YH, Liu Y, Zhan X, Cheng MM. P2t: pyramid pooling transformer for scene understanding. IEEE Trans Pattern Anal Mach Intell 2022. https://doi.org/10.1109/TPAMI.2022.3202765.

[21] Wu M, Yan C, Liu H, Liu Q, Yin Y. Automatic classification of cervical cancer from cytological images by using convolutional neural network. Biosci Rep 2018;38. https://doi.org/10.1042/BSR20181769.

[22] Wu H, Xiao B, Codella N, Liu M, Dai X, Yuan L, et al. Cvt: introducing convolutions to vision transformers. In: IEEE/CVF international conference on computer vision (ICCV); 2021.

[23] Khan A, Han S, Ilyas N, Lee YM, Lee B. Cervixformer: transformer-based cervical pap-smear wsi classification framework. SSRN Electron J 2022. https://doi.org/10.2139/ssrn.4266652.

[24] Hussain E, Mahanta LB, Borah H, Das CR. Liquid based-cytology pap smear dataset for automated multi-class diagnosis of pre-cancerous and cervical cancer lesions. Data Brief 2020;30:105589.

[25] Cheng S, Liu S, Yu J, Rao G, Xiao Y, Han W, et al. Robust whole slide image analysis for cervical cancer screening using deep learning. Nat Commun 2021;12:5639. https://doi.org/10.1038/s41467-021-25296-x.

[26] Diederik PK. Adam: a method for stochastic optimization; 2014 (No Title).

[27] He T, Zhang Z, Zhang H, Zhang Z, Xie J, Li M. Bag of tricks for image classification with convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019. p. 558–67.

[28] Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR); 2022.