Original Article

MacNet: a mobile attention classification network combining convolutional neural network and transformer for the differentiation of cervical cancer

Yi An[1,2#], Yuanyuan Lei[3#], Zhenxing Huang[1], Yu Liu[1], Meiyong Huang[1], Zhou Liu[4], Wenbo Li[1,2], Dong Liang[1], Wenting Huang[3], Zhanli Hu[1]

[1]Research Center for Medical AI, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China; [2]University of Chinese Academy of Sciences, Beijing, China; [3]Department of Pathology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Shenzhen, China; [4]Department of Radiology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Shenzhen, China

Contributions: (I) Conception and design: Z Hu, W Huang, Z Huang; (II) Administrative support: Z Hu, W Huang; (III) Provision of study materials or patients: Y Lei, W Huang, Z Liu; (IV) Collection and assembly of data: Y Lei, Z Liu, Y An; (V) Data analysis and interpretation: Y An, Z Huang, Z Hu; (VI) Manuscript writing: All authors; (VII) Final approval of the manuscript: All authors.

[#]These authors contributed equally to this work as co-first authors.

Correspondence to: Wenting Huang, PhD. Department of Pathology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital & Shenzhen Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, No. 113 Baohe Avenue, Longgang District, Shenzhen 518116, China. Email: huangwt@cicams.ac.cn; Zhanli Hu, PhD. Research Center for Medical AI, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, No. 1068 Xueyuan Avenue, Shenzhen University Town, Xili, Nanshan District, Shenzhen 518055, China. Email: zl.hu@siat.ac.cn.

Background: Cervical cancer remains a critical global health issue, responsible for over 600,000 new cases and 300,000 deaths annually. Pathological imaging of cervical cancer is a crucial diagnostic tool. However, distinguishing specific areas of cellular differentiation remains challenging because of the lack of clear boundaries between cells at various stages of differentiation. To address the limitations of conventional clinical and deep learning (DL) methods, we developed a mobile attention classification network (MacNet) with multiscale features, aiming to increase the accuracy of differentiation classification and quantitatively analyze cervical cancer cell differentiation.

Methods: We investigated the application of MacNet for classifying non-background images into 3 stages of cervical cancer differentiation. The feature maps are processed through the Mobile Convolution Neural Network with Mobile Attention (MCMA) module, which integrates mobile convolutional blocks and mobile attention blocks. MacNet harnesses the benefits of the image pyramid structure and self-attention mechanism, enabling multiscale feature extraction and emulation of clinical pathologist analysis. The final prediction is generated by the adaptive fusion module, which aggregates features into a unified output.

Results: Comparative evaluations demonstrated that MacNet outperforms existing models. The proposed method achieved the best classification accuracy of 92.34% among all 7 DL-based models. Specifically, the result achieved by MacNet was 2.62% greater than that of Inception Version 3, 7.9% greater than that of vision transformer, 8.08% greater than that of the visual geometry group network, 3.21% greater than that of Densely Connected Convolutional Network, 2.85% greater than that of shifted window transformer (Swin transformer), 5.4% greater than that of Cross Stage Partial DarkNet, and 5.41% greater than that of Residual Neural Network. At the same time, MacNet also achieved superior results in recall, precision, and F1 score.

Conclusions: We have proposed a lightweight neural network method that innovatively combines attention mechanisms with convolutional neural networks (CNNs) to efficiently utilize multiscale information from histopathological images. This integration enables the precise quantitative display of different stages of differentiation in cervical cancer. By doing so, our method not only enhances diagnostic accuracy but also provides clinicians with a more effective tool for faster and more reliable diagnosis, representing a significant advancement in the field of pathological imaging.

Keywords: Cervical cancer detection; computer vision; deep learning model (DL model); image classification

Submitted Apr 21, 2024. Accepted for publication Oct 18, 2024. Published online Dec 30, 2024.
doi: 10.21037/qims-24-810
View this article at: https://dx.doi.org/10.21037/qims-24-810

Introduction

Cervical cancer is a malignant disease that significantly impacts women's health worldwide. In 2020, it was estimated that there were 604,127 cases of cervical cancer and 341,831 related deaths globally (1). According to recently released statistics on global cancer, cervical cancer ranks as the third most common cancer among women, with a mortality rate of 7.7% and a mortality rate as high as 6.5% (2).

In China, 61,579 patients died due to cervical cancer in 2022 (3,4). Therefore, it is crucial to protect women's health via reliable and accurate cervical cancer diagnosis. Effective medical screening and treatment can significantly reduce the mortality rate of patients with cervical cancer. However, the detection of cervical cancer is often neglected, leading to insufficient research in medical technology and clinical practice.

Histopathological images play a vital role in medical research (5). In traditional clinical methods, histopathological images obtained from a biopsy are commonly employed to determine whether patients have cervical cancer (6). The differentiation stage is confirmed by analyzing the cell features, which can identify patients with cervical cancer and can be used for cancer grading. Unfortunately, analyzing histopathological images is a time-consuming and subjective task because of their extremely high resolution, often comprising billions of pixels. It is easy to miss well-differentiated cancer cells, leading to serious consequences such as misdiagnosis. Additionally, pathologists need to combine ×100 or higher magnification lenses (×200 and ×400) to improve diagnostic accuracy, which requires even more time. Therefore, multiscale contextual information is critical for accurately classifying histopathological images containing cells at different stages of differentiation (7). In other scenarios, models that use multiscale features, such as fire and smoke you only look once (YOLO), have been efficient (8). In this study, we explore a multiscale strategy to analyze images, enabling our model to leverage multiscale information when examining histopathological images at ×400 magnification. This allows us to capture a wider feature range across multiple scales within less time.

Recent advancements in vision-based classification have been driven by state-of-the-art models such as convolutional transformer (ConvFormer) (9), triple attention transformer (TripleFormer) (10), stable diffusion XL (SDXL) (11), transferability of visual prompting (TVP) for multimodal large language models (12), and region-enhanced prototypical transformers (13). ConvFormer optimizes convolutional neural network (CNN)-based architectures by incorporating transformer-style attention mechanisms, significantly enhancing performance in image classification tasks, particularly in medical image analysis. TripleFormer introduces quadrangle attention to vision transformers, which effectively captures spatial relationships and improves classification accuracy. SDXL scales diffusion models to generate high-quality images from text, offering superior control in text-to-image generation, whereas TVP explores the transferability of visual prompts in multimodal large language models, excelling in few-shot and zero-shot image classification. Finally, the region-enhanced prototypical transformer advances few-shot medical image segmentation by refining regional prototypes, making it highly effective for classification with limited data. These models reflect the cutting edge of both natural and medical image analysis, providing a robust foundation for our study. However, most of these works have been applied to natural image scenarios, and the inference processes are slow. Currently, numerous machine learning-based approaches have been applied to the analysis and diagnosis of cervical cancer, including histopathological image segmentation and classification (14). For example, the gray level co-occurrence matrix (GLCM) was used for extracting the information of textural features, and then a support vector machine (SVM) method was applied for segmenting and recognizing cervical cancer (15). Song *et al.* (16) developed a superpixel algorithm (17) for segmenting cervical cancer nuclei, where the CNN-based rough template structure was used for detecting the nucleus region, and superpixels were applied to improve the accuracy of cytoplasm boundary segmentation. To classify moderate and poorly differentiated stages of cervical cancer, Li *et al.* (18) used a weakly supervised learning strategy for training a multilayer hidden conditional random field (MHCRF)-based classification model, achieving commendable performance. Similarly, Purwanti *et al.* (19) developed a learning vector quantization (LVQ) method based on an artificial neural network, which was used to detect cervical cancer cells automatically. These methods fail to provide accurate differentiation between stages of cervical cancer. This is partly due to their inability to capture and leverage multiscale information efficiently—a critical factor in medical image analysis. The absence of robust multiscale feature extraction mechanisms in these models often results in suboptimal performance. Inspired by the aforementioned studies, our proposed method bypasses the segmentation step, utilizing neural networks for direct patchwise classification, which leads to a substantial improvement in classification accuracy and other performance metrics.

CNN-based architectures have been widely used for different tasks in computer vision areas, such as classification, segmentation, and detection of histopathological images (20-23). Humans typically focus on distinct regions of an image when observing it (24). To incorporate the attention mechanism into a CNN, Sandler *et al.* (25) introduced the squeeze-excitation (SE) module (26) to mobile convolution (MBConv), which assigns different channel attention weights to capture global information. Considering the proposal of the powerful transformer layer, which uses a self-attention mechanism to capture global information efficiently, it is feasible to replace the SE module with transformer layers (27). Therefore, Yang *et al.* (28) proposed the mobile attention (MOAT) block. This architecture inherits MBConv's "invert bottleneck" design, which is designed for downsampling operations with strided depthwise convolution (29). Deep learning (DL) models are prone to overfitting, especially when trained on small datasets. This overfitting, coupled with the large number of parameters typical of these models, hampers their efficiency and limits their applicability in clinical settings. These limitations underscore the need for a more robust, scalable, and accurate approach to cervical cancer diagnosis. Our proposed model, a mobile attention classification network (MacNet), integrates MBConv and attention mechanisms, resulting in a MOAT classification network. The term "mobile" in the title reflects the use of MBConv modules, which employ an inverted bottleneck structure that is more parameter-efficient than traditional transformer modules. Consequently, despite the increased network depth, our model has fewer parameters (3M less) than the basic vision transformer (ViT), making it relatively lightweight.

Compared with widely used single-network models, fusing different models can sometimes achieve better performance (30). For the BACH dataset, the experimental results demonstrated modest accuracy rates of 79% and 81% for the SVM and CNN, respectively. However, fusing the SVM and CNN achieved an accuracy of 92% (31). Owing to the special features of histopathology images, which differ from those of natural images, it is crucial for our network to focus on both local cell morphology and the long-range arrangement of cells. Therefore, we fused a transformer with a CNN in the new network, which is more powerful and achieves better performance. We designed our network by combining the CNN with the transformer. However, most DL models for classification, such as MOAT and Residual Neural Network (ResNet), are designed primarily for natural image datasets such as ImageNet, which differ significantly in sample distribution compared with medical datasets. As a result, the original MOAT may not achieve optimal classification accuracy when it is directly applied to our medical dataset. To address this, we propose a MOAT classification network with multiscale features to enhance the performance on medical data.

There is no doubt that DL-based models require adequate images as inputs to guarantee that the models are capable of learning sufficient and valid features from specific objects (32-35). At present, studies on the direct assessment of cervical cancer severity on the basis of cell differentiation characteristics are limited, partly owing to the difficulty in obtaining annotated histopathological images and the limited number of datasets related to the classification of cervical cancer stages in publicly available datasets of histopathological images (36-38). Therefore, a labeled, high-quality dataset was also developed in this paper to classify the staging of histopathological cervical cancer images.

In summary, this paper presents a method to distinguish the stage of cervical cancer by integrating MBConv and MOAT. The attention mechanism has proven effective in image classification, especially in tasks that require extracting long-range relevant information (39-43). MacNet achieves a balance between local and long-range relevance by integrating a CNN-based model with an attention mechanism and employing an image pyramid to capture information at various scales. The backbone architecture is designed to extract multiscale features effectively. The block sequence is adjusted, and the model architecture fully leverages the model's multiscale capabilities; it emphasizes both local and long-range relevance, contributing to excellent performance in the field of medical image processing. Compared with other DL methods, such as Inception Version 3 (InceptionV3) (44), ViT (45), visual geometry group (VGG19) (46), Densely Connected Convolutional Network (DenseNet) (47), ResNet50 (48), Cross Stage Partial Darknet (CSPDarkNet) (49), and the shifted window transformer (Swin transformer) (50), the proposed method achieves outstanding performance in patchwise differentiation classification of cervical histopathological images. The proposed method is versatile and can be applied to other scenarios beyond classifying stages of differentiation in our future work.

The contributions of this paper are summarized as follows:

- ❖ Novel architecture and module distribution: MacNet integrates MBConv modules and attention mechanisms in a unique structure. The MBConv modules utilize an inverted bottleneck design, reducing the parameter count while effectively capturing local features. Additionally, MacNet is the first model to apply a multiscale strategy to the detection of cervical cancer in histopathological images. This approach mimics how pathologists observe samples under different magnifications, capturing features at multiple scales. This careful distribution of MBConv and MOAT blocks balances local and global feature extraction, enhancing the model's overall performance.
- ❖ High performance with fewer parameters: MacNet achieves high performance with significantly fewer parameters than traditional models do. The integration of lightweight MBConv modules contributes to this efficiency. MacNet demonstrated a superior classification accuracy of 92.34%, outperforming other DL models, such as VGG, ResNet, DenseNet, and ViT, while maintaining a lower parameter count. This makes the model more efficient and easier to deploy in clinical settings. A lightweight network is used to design the network architecture specifically. Background detection can reduce judgment time, making it more applicable to clinical devices.
- ❖ Fast processing speed and high clinical feasibility: MacNet offers rapid image processing, requiring less than 0.1 seconds per 1,024×1,024 image slice compared with the approximately 2 seconds needed by a pathologist. This speed significantly enhances clinical feasibility, allowing for quicker diagnoses. Additionally, MacNet includes attention heatmaps, providing visual explanations of the features the model has learned. This transparency increases clinicians' trust in the model's outputs, as they can see and understand the basis for the model's decisions. This method promotes the rapid assistance of doctors in making diagnoses, which has substantial clinical significance. The remainder of this article is arranged as follows.

Methods section provides an overview of the proposed method, including the overall workflow structure, MacNet structure, MOAT module, and MBConv module. *Experiments* section presents the experimental results, including the evaluation and analysis of the proposed method. The priorities of MacNet are presented in *Results* section, and the related discussion is presented in *Discussion* section. Finally, *Conclusions* section concludes this paper and discusses future work.

Methods

Method overview

In clinical settings, pathologists often cannot perform detailed classifications at the level that computer-aided design (CAD) systems can perform. To address the limitations of conventional clinical methods and DL approaches, we designed an end-to-end architecture for quantifying specific categories, including modules for background detection, multiscale feature extraction, classification, and concatenation. The overall workflow is illustrated in *Figure 1*.

Specifically, in the background detection module, histopathological images with extremely high resolution are

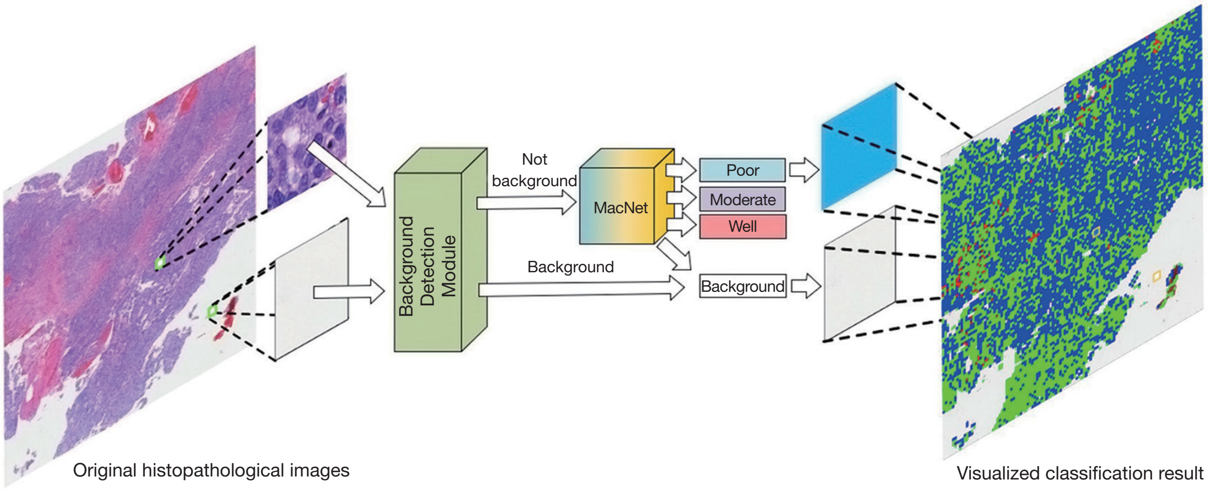

Not background
Background Detection Module
MacNet
Poor
Moderate
Well
Background
Background
Original histopathological images
Visualized classification result

Figure 1 Overall workflow of the proposed method. MacNet is the core classification network. MacNet, mobile attention classification network.

divided into millions of 256×256 patches. The patch size was designed to be 256×256 pixels. This size was selected because larger patches significantly slow network inference and training, whereas smaller patches lack sufficient cells for reliable differentiation and make it difficult to observe the cell arrangements necessary for accurate classification. These 256×256 patches are then processed by the background detection module, which employs a specialized technique to accelerate the process without relying on DL networks, effectively classifying all images into background and non-background categories. A detailed explanation of this module is provided in *Background detection* section. Following the background detection module, MacNet is employed to classify non-background images into 3 differentiated stages. Note that MacNet is the core network of the proposed method, which consists of an image pyramid structure, Mobile Convolution Neural Network with Mobile Attention (MCMA) modules, and an adaptive fusion module. The image pyramid structure is used to capture multiscale features, and the generated feature maps are processed via the MCMA module. The MCMA module consists of MBConv modules and MOAT modules. Inspired by Kerdvibulvech's works (51,52), we introduce an adaptive fusion module to produce the final classification results. Finally, during testing with ultrahigh-resolution original images, we cut the original images into multiple 256×256 patches. The different regions are colored based on the classification results to achieve visualization of the cell differentiation stages, including poorly, moderately, and well-differentiated stages and the background region. We create a canvas that matches the size of the original image and, after classification and colorization, paste these patches back onto the new image according to their original coordinates to form the final resulting image.

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). Approval for all ethical and experimental procedures and protocols was granted by the National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital & Shenzhen Hospital (No. IRB KYLX2022-44). Informed consent was provided by all patients.

Background detection

A cervical histopathological image typically contains billions of pixels; this means that a histopathological image can be divided into more than 200,000 256×256 patches. Therefore, a module was designed for efficiently making a preliminary prediction of whether the input image is a background image. As shown in *Figure 2*, background patches can be easily identified because they contain different pixel information to that of non-background patches. Therefore, the pixel value is the key to finding the background detection threshold, and if the gray value of a pixel is greater than a certain threshold, it is classified as a background pixel. In the first step, the final threshold is set to 210 on the basis of the lower bound of the pixel gray value of the pure background image. The numbers of pixels

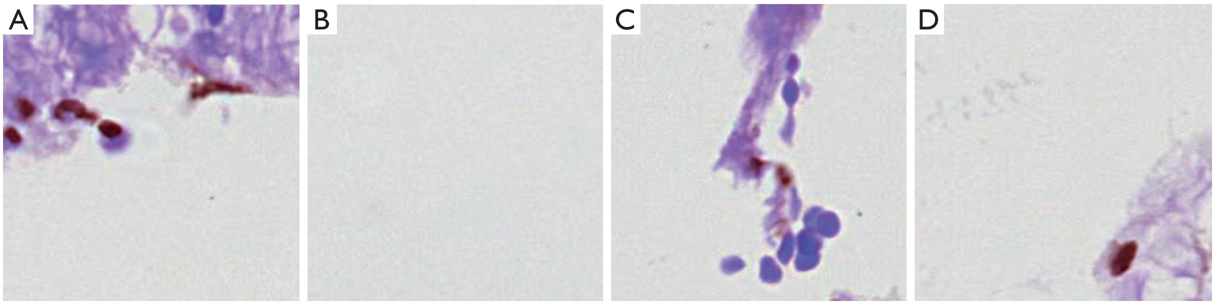


Figure 2 Background image examples (hematoxylin-eosin staining, ×400). (A-D) Examples. We define background images as those wherein more than half of the pixels in this category contain no cells.

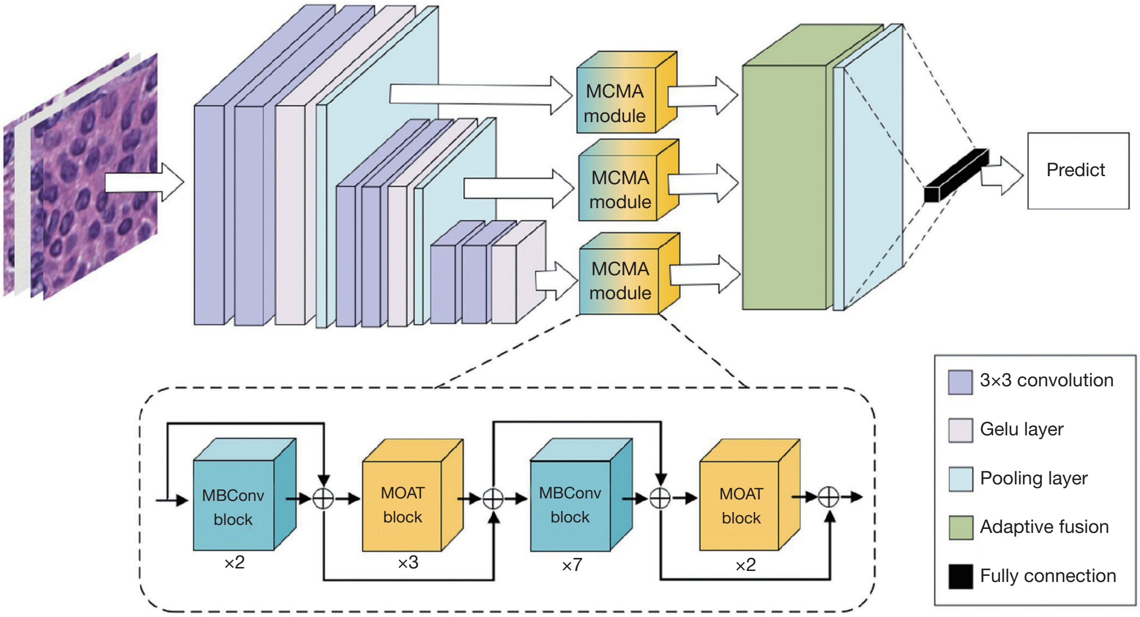


Figure 3 Illustration of the core of the network: MacNet. MCMA module integrates MBConv blocks and MOAT blocks. MacNet, mobile attention classification network; MCMA, Mobile Convolution Neural Network with Mobile Attention; MBConv, mobile convolutional; MOAT, mobile attention.

with a gray value of greater than 210 and less than 210 are subsequently calculated. Finally, the quantities of pixels in different categories are compared, and if the number of pixels greater than 210 is larger, the patch is classified as a background image.

The core of the network

MacNet

To extract multiscale information more efficiently, the core of the proposed method, including an attention mechanism and a CNN-based network, was designed, the details of which are shown in *Figure 3*. The attention mechanism is introduced to capture the long-range relationship between pixels, and the CNN-based network has inductive biases of locality, spatial invariance, and translation equivariance.

In this part, the original input images preprocessed to 256×256 images (patches) are fed into a multiscale head to extract spatial features at different magnifications, which can imitate pathologists analyzing histopathological images in clinical practice. Then, the data streams processed by the multiscale heads are sent separately to different modules

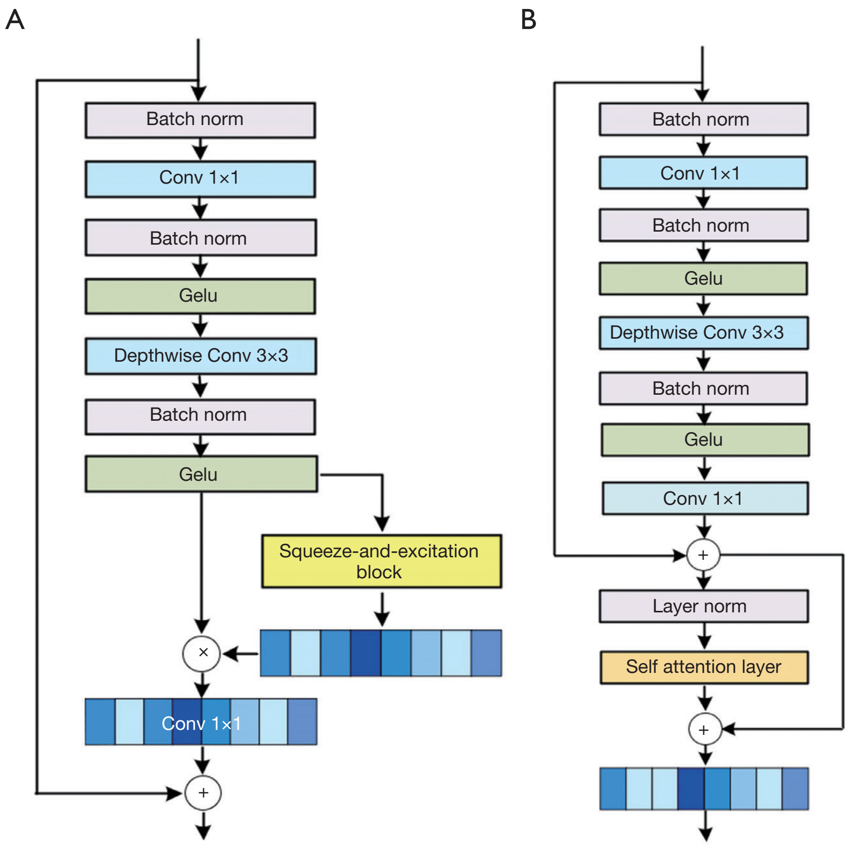


Figure 4 Network structure. (A) The structure of the MBConv block. (B) The structure of the MOAT block. MBConv, mobile convolution; MOAT, mobile attention.

in parallel. Finally, an adaptive fusion block is constructed for feature fusion. According to the output importance of different blocks, the model assigns weights to the outputs of different modules to obtain important features.

MCMA module (MBConv and MOAT block)

The MCMA module consists of two architectures, which are shown in *Figure 4*. In the MCMA module, these MBConv and MOAT blocks are arranged to balance local and global feature extraction effectively. The integration of the MBConv and MOAT blocks allows the model to capture multiscale features crucial for accurate classification of histopathological images. This combination ensures that the model can efficiently process high-resolution images while maintaining a high level of accuracy.

The MBConv blocks include a normalization layer, a depthwise separable convolution, and an SE head, and it employs the design of an "inverted bottleneck". It also uses a 1×1 convolution to expand channels, a 3×3 convolution to extract features, and another 1×1 convolution to project channels to their original size. This design allows the model to capture local spatial information more efficiently with fewer parameters. The number of channels is doubled, and the step length of the 3×3 convolution is set to 2 so that our model no longer requires additional layers of downsampling. The structure of the MBConv can be expressed as:

$$output_{MOConv} = Dropout\left(Conv\left(SE\left(DS\left(x\right)\right)\right)\right) + skip\left(x\right) \qquad [1]$$

where DS denotes depthwise separable convolution and where SE is the SE layer, which can capture global information and is used to assign different weights to different channels. After a 1×1 convolution ($Conv$), a dropout layer ($Dropout$) is added to randomly remove some of the fully connected layer paths because, in medical datasets, especially cervical cancer histopathological datasets, the model may overfit the training datasets.

Therefore, dropout is necessary. After the dropout layer, a residual connection ($skip$) is used to guarantee that the following layers will receive forward features.

The other architecture is MOAT, which combines a transformer with MBConv. The MOAT block integrates MBConv modules with transformer-like attention mechanisms. In the MOAT block, the MLP module in the standard transformer block is replaced with an MBConv block to leverage the efficiency of mobile convolutions. Unlike previous methods that stack these blocks separately, MOAT integrates them to enhance network representation capacity and produce better downsampled features. The MBConv block is placed before the self-attention operation, allowing the depthwise convolution to handle downsampling more effectively and capture better downsampled features. The input tensors are sent to an MBConv block that replaces SE layers with a normalizing layer and a Swin transformer block. Compared with traditional transformer layers, the Swin transformer, which is based on self-attention and shifted windows, can extract more global, long-range features and information about location relationships. The MOAT output can be represented as:

$$output_{MOAT} = Dropout\left(Swin\left(LN\left(MOConv\left(i\right)\right)\right)\right) \quad [2]$$

where $MOConv$ is the MBConv block and e LN denotes layer normalization. $Swin$ denotes the Swin transformer.

The proposed method connects 2 MBConv blocks, 3 MOAT blocks, 7 MBConv blocks, and 2 MOAT blocks to build an MCMA module. The skip connections between both the MBConv and MOAT blocks are used, which directly transmit forward features to prevent gradient explosion. The powerful MOAT blocks with attention mechanisms can receive diverse multiscale information extracted by MBConv blocks.

Classification and concatenation

Finally, the fused feature map fused by an adaptive fusion block is processed by a classifier. The classifier with fully connected layers generates the input image's expectation of different categories and the Softmax function to obtain the final result. The input and output of the unit can be expressed as follows:

$$out = pred\left(F_{sum}\left(f_{MCMA}s\left(i\right) + f_{MCMA}\left(s\left(p\left(i\right)\right)\right) + f_{MCMA}\left(s\left(p\left(i\right)\right)\right)\right)\right) \quad [3]$$

where i represents the input images, p represents the pooling layer, s represents the stem block that can extract shallow features, f_{MCMA} represents the MCMA module, F_{sum} represents the adaptive fusion layer, and $pred$ represents the classifier.

Experiments

Patient data

To test the effectiveness of the proposed method, a practical image dataset of cervical cancer histopathology was crafted by a professional team. A total of 69 patients were included in this study, from whom multiple biopsy pathology scans were collected at the Cancer Hospital Chinese Academy of Medical Sciences. Each original image had a resolution of approximately 100,000×100,000 pixels. Patient ages ranged from 30 to 83 years, with a median age of 54 years. Totals of 56 biopsy samples and 13 surgical samples were obtained. The samples were from different clinical stages: 14 from stage I patients, 40 from stage II patients, 8 from stage III patients, and 7 from stage IV patients. Owing to the varying differentiation regions occupying different areas of the entire image, it was not feasible to define the differentiation level of an entire large image. Therefore, during dataset annotation, the original images were cropped to a size of 256×256, and the differentiation level of the cells within these small image regions was determined. The images were labeled by 2 experienced doctors and classified into 4 categories: poor, well, and moderate differentiation, and background. To ensure consistency in labeling, the annotations were performed by 2 experienced doctors from the same hospital and department, who adhered to a unified set of standardized criteria, which are discussed in *Discussion* section. After annotation, a total of 196,306 images were obtained. Although the annotations were made on the basis of the ×400 magnified images, our model employs a multiscale strategy using an image pyramid structure to capture features at various scales. This technique allows the model to simulate the effects of observing the images at different magnifications without altering the physical resolution of the annotated images.

To ensure unbiased evaluation and prevent any potential model bias toward patient-specific characteristics, we split the data into training, validation, and test sets on the basis of individual patients. This means that all images from a single patient were assigned to the same subset. Specifically, the test set consisted of pathology image slices from new patients who were not included in the training or validation

Table 1 Distribution of the number of images in the training, validation, and test datasets

Stages	Training	Validation	Test
Background	11,095	1,584	3,139
Well	5,130	732	1,484
Poor	20,112	2,873	5,900
Moderate	101,079	14,439	28,739

Background: those wherein more than half of the pixels in this category contain no cells. Well: well-differentiated squamous cell carcinoma occupies more than half of the pixels in the image. Moderate: moderately differentiated squamous cell carcinoma occupies more than half of the pixels in the image. Poor: poorly differentiated squamous cell carcinoma occupies more than half of the pixels in the image.

sets. This approach ensured that the model's performance was evaluated on entirely unseen data, providing a more accurate measure of its generalization capabilities.

The data were normalized and input into the network for training with a horizontal flip probability of 0.3. The training set, validation set, and testing set were divided according to a ratio of 7:1:2, as shown in *Table 1*. The image distribution reflected a natural imbalance, with moderately differentiated images being the most prevalent. To address this imbalance and prevent the model from overfitting on the more abundant categories during training, we applied higher loss function weights to the less represented classes, such as the poorly and well-differentiated categories. This approach ensured that the model paid appropriate attention to all differentiation levels, improving its ability to generalize across the entire dataset.

Training details and platform

The proposed model trained 300 epochs with a batch size of 32, the initial learning rate was set as 0.001, and the adaptive moment estimation (ADAM) optimizer was used. Cross-entropy loss was selected as the loss function, which is defined as:

$$loss = -\sum_{1}^{l}\sum_{1}^{4} P(i)\log_2 Q(i) \qquad [4]$$

where $P(i)$ is the output of the model, which represents the probability of being classified into category i, and $Q(i)$ is the one-hot label of the dataset.

All operations were performed on a computer with a GeForce RTX 3090 GPU (NVIDIA, Sanat Clara, CA, USA) and the PyTorch neural network framework (https://pytorch.org/).

Evaluation metrics

The performance of the proposed method is quantified via accuracy, precision, recall, and F1 score metrics. Accuracy is the ratio of the number of samples correctly classified by the classifier to the total number of samples. Precision reflects the proportion of positive samples that are determined by the classifier to be positive samples. The recall rate reflects the proportion of total positive samples that are correctly identified by the classifier. The F1 score is an indicator that comprehensively considers accuracy and recall.

Table 2 describes these performance metrics. In this paper, positive samples refer to samples from the categories studied in this paper, whereas negative samples refer to samples from other categories. TP is the true positive (a positive sample is predicted to be positive), TN is the true negative (a negative sample is predicted to be negative), FP is the false-positive (a negative sample is predicted to be positive), and FN is the false-negative (a positive sample is predicted to be negative).

Results

Prediction accuracy

In this section, the comparison results with the state-of-the-art DL-based models, including VGG, ResNet, DenseNet, ViT, InceptionV3, Swin transformer, and CSPDarkNet, are presented to demonstrate the advantages of the proposed method in terms of performance, parameter counts, and model operation speed. *Figure 5* shows that the proposed method achieved the best classification accuracy of 92.34% among all 7 DL-based models. Specifically, the result achieved by MacNet was 2.62% greater than that of InceptionV3, 7.9% greater than that of ViT, 8.08% greater than that of the VGG19 network, 3.21% greater than that of DenseNet, 2.85% greater than that of Swin transformer, 5.4% greater than that of CSPDarkNet, and 5.41% greater than that of ResNet.

For the proposed method, when classifying the moderately differentiated stage, the recall rate was 0.9569, which was 0.0027–0.0354 higher than that of the other compared algorithms except for VGG19, but VGG19 overfit the moderately differentiated stage and performed poorly on other categories. The precision rates of the proposed method were 0.0248 to 0.0955, which were higher than those of the other methods, and the F1 scores were 0.0174 to 0.05, which were also higher than those of the other methods. For the poorly-differentiated stage, the proposed method achieved a recall rate of 0.77, which was 0.2744–0.431 higher than the recall rate of the compared models. The proposed method also achieved a precision rate of 0.8551, which was 0.0628–0.2159 higher than the precision rate of the compared models. The F1 score was 0.0661–0.3523 higher than the other models. For the well-differentiated stage, the proposed method achieved a recall rate of 0.7206, which was 0.2509–0.5238 higher than the recall rate of the compared models. The F1 score was 0.1471–0.4158 higher than the other models.

Table 2 Evaluation metrics and their definition details

Metrics	Accuracy on the test dataset
Accuracy	$\text{Accuracy} = \frac{TN + TP}{TP + TN + FP + FN}$
Precision	$\text{Precision} = \frac{TP}{TP + FP}$
Recall	$\text{Recall} = \frac{TP}{TP + FN}$
F1 score	$\text{F1 score} = \frac{2TP}{2TP + FP + FN}$

TN, true negative; TP, true positive; FP, false positive; FN, false negative.

Owing to the lack of explicit boundaries between moderately differentiated cells and other stages of differentiation, DL models may mistakenly classify this stage of differentiation. To show the classification performance of the proposed method more intuitively, these DL-based models were tested on the testing dataset to construct confusion matrices for visual analysis. As shown in *Figures 6,7*, the proposed method achieved superior accuracy, recall rates, and F1 scores when classifying poorly differentiated cells and well-differentiated cells. For classifying moderately differentiated cells, the proposed method also achieved better accuracy, precision, and F1 scores. Compared with the other DL-based models, the proposed model has a lower probability of error classification and greater stability.

For our dataset and an input size of 256×256, the performance versus parameters and 'floating point operations per second' (FLOPs) are displayed in *Table 3*. This finding demonstrates that the proposed method with many fewer parameters (only 5.4 million parameters) achieves outstanding performance compared with the ViT, InceptionV3, ResNet, Swin transformer, CSPDarkNet, and VGG.

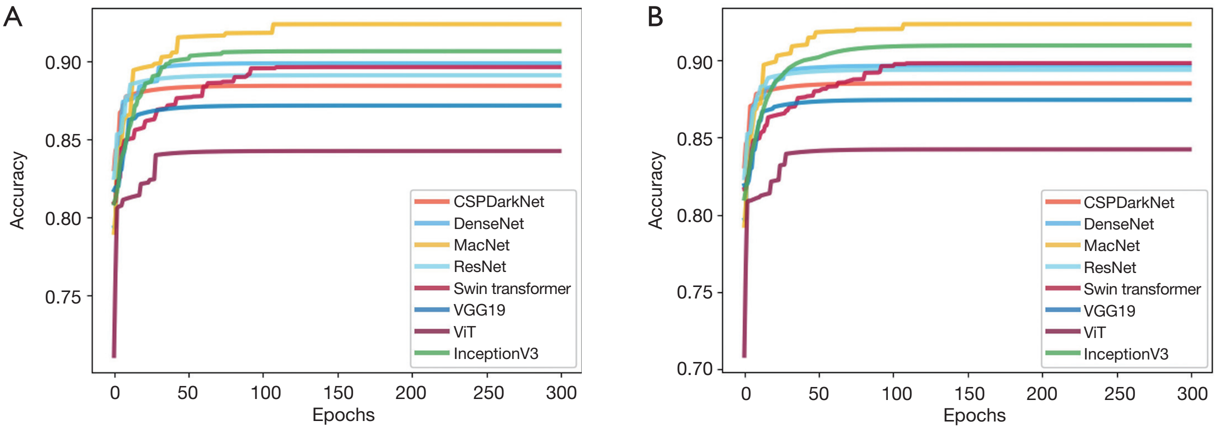
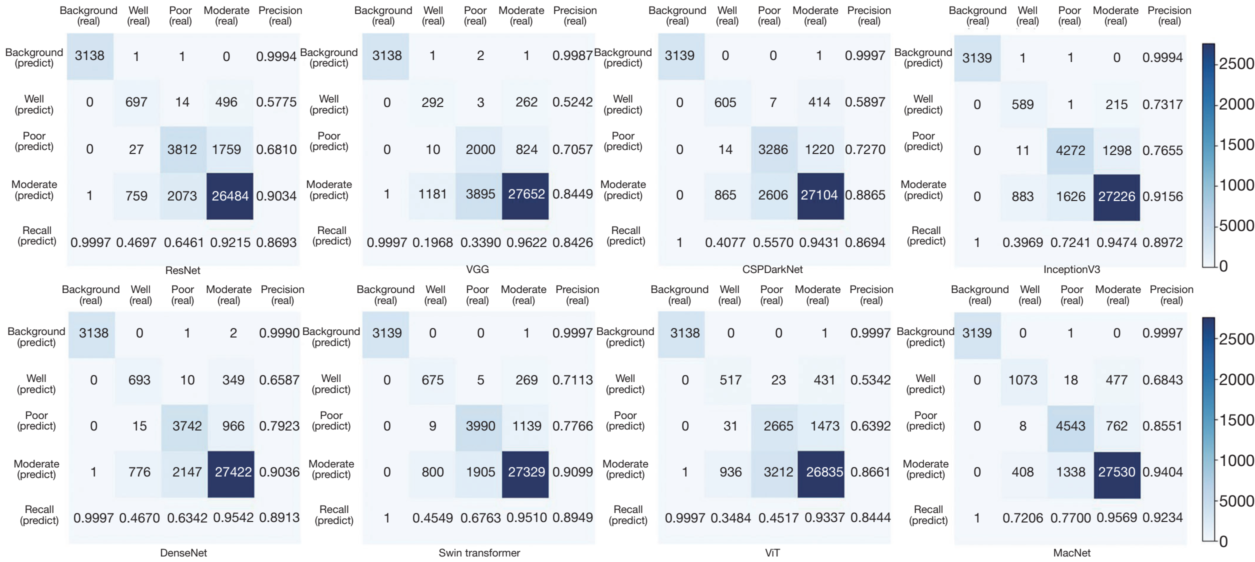


Figure 5 The process curve of relationship between the number of training epochs and classification accuracy. (A) The classification results of different methods on the validation dataset. (B) The results of different methods on the test dataset. CSPDarkNet, Cross Stage Partial DarkNet; DenseNet, Densely Connected Convolutional Network; MacNet, Mobile Attention Classification Network; ResNet, Residual Neural Network; Swin transformer, shifted window transformer; VGG, visual geometry group; ViT, vision transformer; InceptionV3, Inception Version 3.



ResNet

	Background (real)	Well (real)	Poor (real)	Moderate (real)	Precision (real)
Background (predict)	3138	1	1	0	0.9994
Well (predict)	0	697	14	496	0.5775
Poor (predict)	0	27	3812	1759	0.6810
Moderate (predict)	1	759	2073	26484	0.9034
Recall (predict)	0.9997	0.4697	0.6461	0.9215	0.8693

VGG

	Background (real)	Well (real)	Poor (real)	Moderate (real)	Precision (real)
Background (predict)	3138	1	2	1	0.9987
Well (predict)	0	292	3	262	0.5242
Poor (predict)	0	10	2000	824	0.7057
Moderate (predict)	1	1181	3895	27652	0.8449
Recall (predict)	0.9997	0.1968	0.3390	0.9622	0.8426

CSPDarkNet

	Background (real)	Well (real)	Poor (real)	Moderate (real)	Precision (real)
Background (predict)	3139	0	0	1	0.9997
Well (predict)	0	605	7	414	0.5897
Poor (predict)	0	14	3286	1220	0.7270
Moderate (predict)	0	865	2606	27104	0.8865
Recall (predict)	1	0.4077	0.5570	0.9431	0.8694

InceptionV3

	Background (real)	Well (real)	Poor (real)	Moderate (real)	Precision (real)
Background (predict)	3139	1	1	0	0.9994
Well (predict)	0	589	1	215	0.7317
Poor (predict)	0	11	4272	1298	0.7655
Moderate (predict)	0	883	1626	27226	0.9156
Recall (predict)	1	0.3969	0.7241	0.9474	0.8972

DenseNet

	Background (real)	Well (real)	Poor (real)	Moderate (real)	Precision (real)
Background (predict)	3138	0	1	2	0.9990
Well (predict)	0	693	10	349	0.6587
Poor (predict)	0	15	3742	966	0.7923
Moderate (predict)	1	776	2147	27422	0.9036
Recall (predict)	0.9997	0.4670	0.6342	0.9542	0.8913

Swin transformer

	Background (real)	Well (real)	Poor (real)	Moderate (real)	Precision (real)
Background (predict)	3139	0	0	1	0.9997
Well (predict)	0	675	5	269	0.7113
Poor (predict)	0	9	3990	1139	0.7766
Moderate (predict)	0	800	1905	27329	0.9099
Recall (predict)	1	0.4549	0.6763	0.9510	0.8949

ViT

	Background (real)	Well (real)	Poor (real)	Moderate (real)	Precision (real)
Background (predict)	3138	0	0	1	0.9997
Well (predict)	0	517	23	431	0.5342
Poor (predict)	0	31	2665	1473	0.6392
Moderate (predict)	1	936	3212	26835	0.8661
Recall (predict)	0.9997	0.3484	0.4517	0.9337	0.8444

MacNet

	Background (real)	Well (real)	Poor (real)	Moderate (real)	Precision (real)
Background (predict)	3139	0	1	0	0.9997
Well (predict)	0	1073	18	477	0.6843
Poor (predict)	0	8	4543	762	0.8551
Moderate (predict)	0	408	1338	27530	0.9404
Recall (predict)	1	0.7206	0.7700	0.9569	0.9234

Figure 6 Confusion matrix for different models. Background: those wherein more than half of the pixels in this category contain no cells. Well: well-differentiated squamous cell carcinoma occupies more than half of the pixels in the image. Moderate: moderately differentiated squamous cell carcinoma occupies more than half of the pixels in the image. Poor: poorly differentiated squamous cell carcinoma occupies more than half of the pixels in the image. ResNet, Residual Neural Network; VGG, visual geometry group; CSPDarkNet, Cross Stage Partial DarkNet; InceptionV3, Inception Version 3; DenseNet, Densely Connected Convolutional Network; Swin transformer, shifted window transformer; ViT, vision transformer; MacNet, Mobile Attention Classification Network.

Additionally, real-time data processing is crucial in clinical settings, as demonstrated by methods such as YOLO, as well as technologies such as noninvasive, wearable multibiosensors, and flexible, wireless biosensor patches (53-55). Experts do not have the time to conduct extremely detailed scans. A single pathological image can contain nearly 10 billion pixels, which increases the risk of misdiagnosis and missed diagnoses. For a 1,024×1,024 image slice, a pathologist typically spends approximately 2 seconds (56), whereas our model requires less than 0.1 seconds, demonstrating its superior efficiency. Additionally, the clinical advantage of our model is its ability to quantitatively estimate the number of slices with varying differentiation levels within a region, which is more reliable than the empirically subjective judgments of doctors.

Visual analysis

Figure 8 presents a comparison between the input and output images. In the output images, the regions corresponding to different stages of differentiation are clearly discernible after being processed via the proposed method. This can aid pathologists in focusing on specific regions and quantifying the specific proportions of different differentiation levels. With the assistance of the proposed method, pathologists can make more reliable and accurate diagnoses without spending time analyzing billions of pixels.

Ablation studies

Effectiveness of data augmentation

Unlike natural images, histopathological images often face problems such as inconsistent staining and different light intensities. This means that image preprocessing is necessary. Therefore, an appropriate normalization method was selected to eliminate the impact of sampling variability. Moreover, cervical histopathology datasets cannot be easily acquired, and to increase our model's generalizability, we randomly rotated and flipped images to perform data augmentation. These steps help the following DL network extract more plentiful features.

Table 4 shows that the best classification accuracy of the proposed method with image preprocessing is 92.34%, and the accuracy is higher than 90.46%, which is achieved by the same model without preprocessing. Moreover,

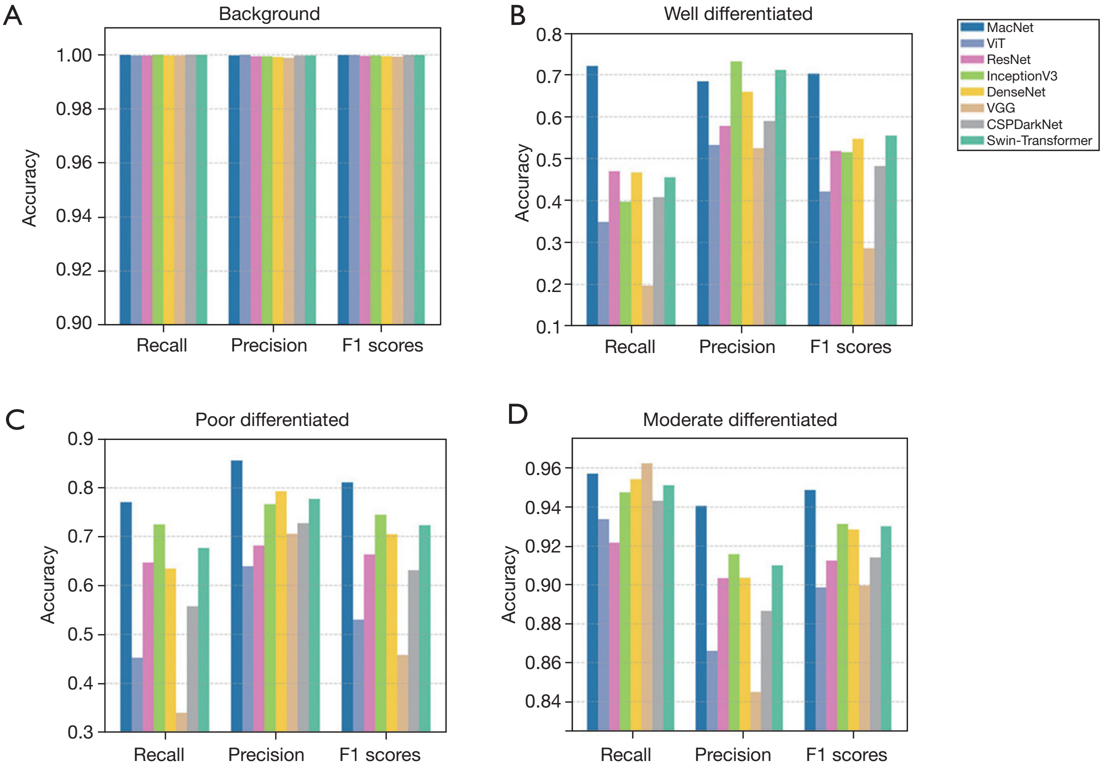


Figure 7 Comparison between different DL models in four differentiation stages. (A) Bar charts showing the recall, precision, and F1 scores of different models on background. (B) Bar charts of different models on data of well differentiation. (C) Bar charts of different models on data of well differentiation. (D) Bar charts of different models on data of well differentiation. Background: those wherein more than half of the pixels in this category contain no cells. Well: well-differentiated squamous cell carcinoma occupies more than half of the pixels in the image. Moderate: moderately differentiated squamous cell carcinoma occupies more than half of the pixels in the image. Poor: poorly differentiated squamous cell carcinoma occupies more than half of the pixels in the image. MacNet, Mobile Attention Classification Network; ViT, vision transformer; ResNet, Residual Neural Network; InceptionV3, Inception Version 3; DenseNet, Densely Connected Convolutional Network; VGG, visual geometry group; CSPDarkNet, Cross Stage Partial DarkNet; Swin transformer, shifted window transformer; DL, deep learning.

Table 3 Performance versus parameters and FLOPs on the dataset

Methods	Parameter counts (M)	FLOPs (B)	Accuracy (%)
DenseNet	0.8	0.6	89.23
VGG	91.3	255	88.12
ResNet	21.3	4	88.86
Vision transformer	8.6	0.5	85.53
InceptionV3	22.3	4.1	90.02
CSPDarkNet53	26.6	6.6	86.94
Swin transformer	27.5	5.6	89.49
MOAT (only)	9.4	4.2	91.73
MBConv (only)	2.6	0.4	87.62
Proposed method	5.4	0.9	92.34

FLOPs, floating point operations per second; DenseNet, Densely Connected Convolutional Network; VGG, visual geometry group; ResNet, Residual Neural Network; CSPDarkNet, Cross Stage Partial Darknet; MOAT, mobile attention; MBConv, mobile convolution.

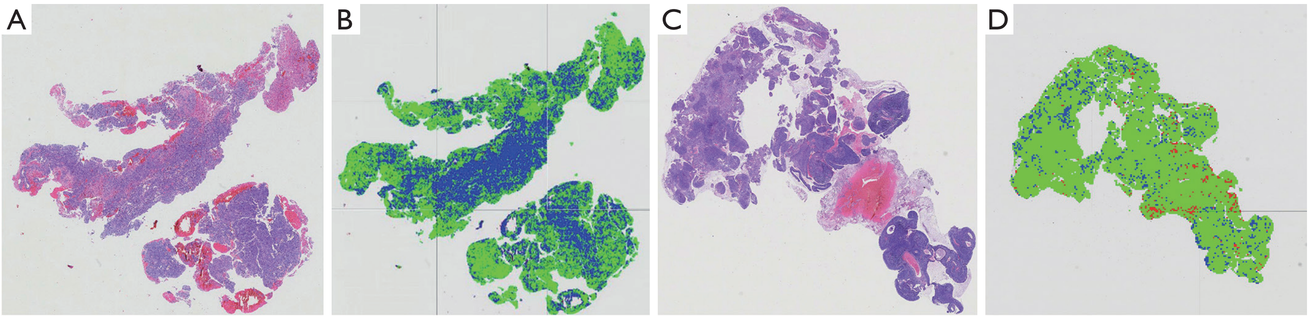


Figure 8 Input histopathological images and visualized output results. Red pixels indicate well-differentiated regions, green represents moderately differentiated regions, and blue represents poorly differentiated regions. (A,C) Example of complete pathology image. (B,D) Distribution map of cell differentiation levels of images (A,C). Cells in (A,C) have been subjected to hematoxylin-eosin staining, ×400. Background: those wherein more than half of the pixels in this category contain no cells. Well: well-differentiated squamous cell carcinoma occupies more than half of the pixels in the image. Moderate: moderately differentiated squamous cell carcinoma occupies more than half of the pixels in the image. Poor: poorly differentiated squamous cell carcinoma occupies more than half of the pixels in the image.

Table 4 Experiment on whether to use data augmentation

Strategies	Accuracy on the test dataset (%)
The proposed method	92.34
Without data augmentation	90.46

when analyzing the training process, the proposed method with preprocessing greatly increased model stability. Preprocessing reduced the training difficulty and led to better network generalization and robustness.

Effectiveness of feature extraction

To evaluate and validate the proposed method, we constructed a MacNet heatmap. Under normal conditions, clinical pathologists and patients typically refuse to trust a DL model because DL models lack interpretability. When histopathological images are analyzed, human pathologists can explain the features they capture. However, DL models are 'black boxes' and cannot explain the reasoning process to interpret their judgment. In this classification task, pathologists classify differentiated levels by the cell nuclear-cytoplasmic ratio and arrangement. Therefore, the likelihood of locations attracting the attention of human pathologists is their cellular morphology and arrangement. The attention heatmaps of the network show the regions of interest and visualize whether the model learns the correct features of specific regions. If the attention heatmap of the proposed method is similar to that of a human, it can be assumed that the method correctly captures specific features. The work in this paper has clinical significance. The proposed method not only outputs the classification results but also informs pathologists and patients about the areas of interest of the model, and the user is more likely to trust the results.

To construct the heatmap, the final feature map is separated before being sent to the classifier, and the Softmax layer is used to obtain the weights of different regions. The tensor is subsequently normalized to integers between 0 and 255. Finally, the tensors are projected back to 256×256 and fused with the original input images to generate heatmaps of specific areas.

In the attention heatmap, areas marked in red receive more attention from the model, and blue areas receive less attention. As *Figure 9* shows, when analyzing the background images, the proposed method tends to focus on white pixels, which corresponds with the definition of background images. When classifying areas other than the background, the proposed method identifies these abnormal cells and classifies them according to their differentiation level. When analyzing regions with a greater proportion of well-differentiated cells, the proposed method focuses on the morphological features that distinguish them from poorly differentiated and moderately differentiated cells. Furthermore, the model not only considers the morphological features but also concentrates on the arrangement of cervical cells, which validates the rationality of the proposed method.

Effectiveness of the multiscale strategy

When histopathological images are analyzed, pathologists usually zoom in and out on the target image to obtain

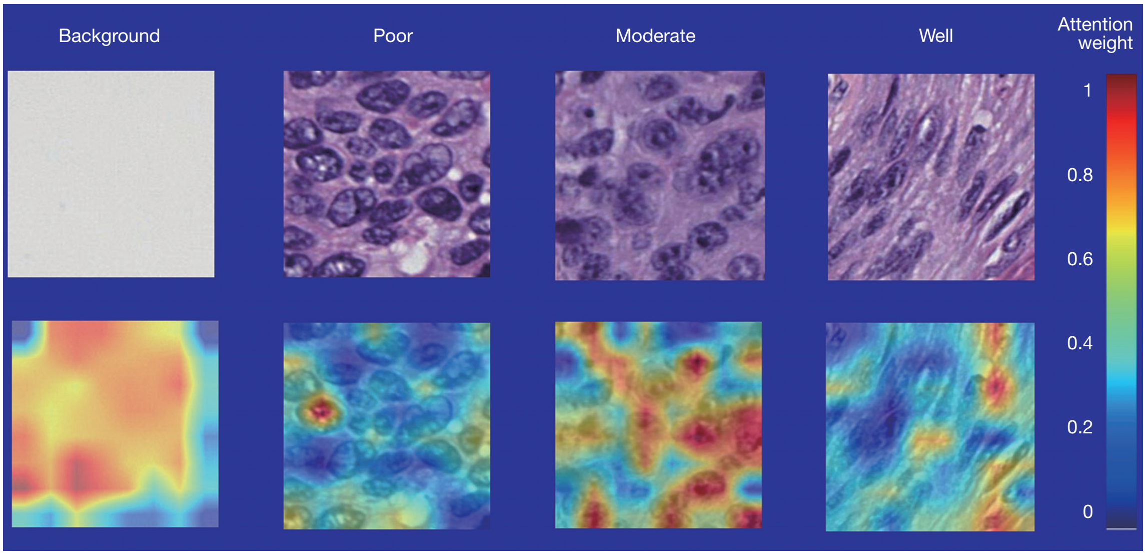


Figure 9 Examples of attention heatmaps. Red indicates regions that receive more attention, and blue indicates that the model will not focus on these areas of the images when classifying. Cells in images above have been subjected to hematoxylin-eosin staining, ×400. Background: those wherein more than half of the pixels in this category contain no cells. Well: well-differentiated squamous cell carcinoma occupies more than half of the pixels in the image. Moderate: moderately differentiated squamous cell carcinoma occupies more than half of the pixels in the image. Poor: poorly differentiated squamous cell carcinoma occupies more than half of the pixels in the image.

Table 5 Experiment on different multiscale strategies

Strategies	Accuracy on the test dataset (%)
The proposed method	92.34
Without image pyramid	91.99
With original sequence	91.89

information from different resolutions. To extract information effectively at different magnifications, a pyramid structure with a global pooling operation is used to generate low-resolution images from the input high-resolution images. In addition, 2 parallel MCMA modules are used to extract information from low-resolution images at different magnifications and fuse their feature maps to make respective predictions. *Table 5* shows that the best classification accuracy of the proposed method with the pyramid structure was 92.34%, which was higher than the 91.99% achieved by the same model without multiscale branches.

Although this improvement may undoubtedly be related to the increase in parameters, simply adding more feature map channels or blocks cannot improve the performance and, counterproductively, will lead to overfitting of the training dataset. This means that the improvement of using an image pyramid is not the result of stacking parameters, and the importance of multiscale information in this classification task motivates concentrating on appropriately extracting these features more efficiently when designing our network.

In the original MOAT network (28), 2 MBConv blocks are followed by 2 MOAT blocks. Considering that the original MOAT network delegates the downsampling operation to depthwise convolution, there are no extra downsampling layers, such as average pooling layers. They do not use skip connections to deliver residual features between blocks. The most powerful block in the network, the MOAT block, is only used after downsampling. This means that the original MOAT structure cannot efficiently extract multiscale information. However, MOAT blocks with an attention mechanism can make a difference in the whole network. The attention mechanism enables MOAT to perform better in processing global multiscale information. Therefore, the skip connection layers are used in blocks and between both MBConv and MOAT blocks, which directly transmit forward features to prevent gradient explosion, adding branches to extract multiscale information and adjusting the connecting methods of blocks and block sequences to achieve the best performance on differentiated images. Finally, the classifier predicts the differentiated level of the input regions on the basis of mixing information at different magnifications. *Table 5* shows that after the rearrangement of blocks and the addition of residual connections between blocks, the accuracy of classification increased from 91.89% to 92.34%. We conducted ablation experiments that included models with only the MOAT module and only the MBConv module. The results showed that the model with only MOAT achieved an accuracy of 91.73%, and the model with only MBConv achieved 87.62%, both of which are lower than the accuracy achieved by our proposed MacNet.

Discussion

In this work, a DL fusion network was applied to classify the differentiation level of cervical cells. The proposed model fuses a mobile CNN with an attention mechanism and introduces a multiscale strategy to imitate human pathologists' observation habits. Compared with other algorithms, the proposed model achieves the best classification performance with the help of background detection models and probably decreases the overall classification time because of the end-to-end system. Moreover, an attention heatmap is used to validate that the proposed method focuses on proper features when classifying an image. In this case, pathologists are more willing to trust the output result of the proposed DL neural network.

In the background detection module, a grayscale value of 210 is selected as a threshold, which categorizes the 256×256 image as background according to the number of pixel grayscale values is above 210 or below 210. The proposed method processes a single image in 0.09 seconds, and after the proposed method is utilized, the system can detect a background image within 0.00001 seconds, which greatly accelerates the processing of histopathological images with billions of pixels. Moreover, judging whether an image should be classified as background is extremely subjective. For a stricter model, increasing the threshold is more effective than training a new model. It is more flexible and controllable than the traditional DL method.

In addition, the rationality of the dataset should be considered. There are few publicly available datasets on the level of differentiation of cervical cancer cells, and in this case, the quality of the dataset directly affects the results of the DL model (36). The dataset used in this study was labeled by several experienced pathologists, which mitigated, to some extent, the bias associated with the subjectivity of a single pathologist. Although most researchers use 3 classification levels for cervical cancer cells, there are practical problems in adopting this classification standard. In histopathological sections, except for cervical cancer cells with different degrees of differentiation, red blood cells occupy some areas. To address this issue, a new category called red blood cells should be added to datasets.

It is necessary not only to clarify the classification but also to clarify the classification criteria. When experienced pathologists were asked to label the proportions of different categories, the quantitative results returned by the pathologists were different. Therefore, there is no clear line between poorly differentiated, moderately differentiated, and well differentiated cervical cancer cells. In clinical diagnosis, the differentiation of cervical cancer cells is characterized by overlapping features without clear boundaries. Well differentiated squamous carcinoma cells of the cervix are a mix of basal and squamous cells with slight increases in the nuclear-plasma ratio and irregularly shaped nuclei with thick, deeply stained chromatin. Nuclear pyknosis and division are rare, with few atypical divisions and multinucleated cells. Moderately differentiated squamous carcinoma cells of the cervix exhibit less keratinization, unclear intercellular bridges, and pronounced nuclear and cellular polymorphisms. These cells have a high nuclear-plasma ratio and exhibit intermediate polymorphisms and nuclear division, including abnormal nuclear division. Poorly differentiated squamous carcinoma cells of the cervix is characterized by immature cells with minimal keratinization and nearly undetectable intercellular bridges. These cells have highly irregular nuclei with coarse chromatin and small nucleoli, indicating significant nuclear and cellular pleomorphism. The nuclear-plasma ratio is very high and often inverted, with frequent normal and abnormal nuclear division (57,58). To improve the proposed method, it is necessary for the pathologist to define specific boundaries with certain features before labeling the target pathological images. Understanding these features will improve the quality of the dataset. Similar to FaNet, which was proposed by Huang *et al.* (23), suitable features that can be identified as prior knowledge are selected to improve our model.

Ensuring patient data privacy is paramount in medical research. In our study, all patient data were anonymized. Strict data security measures, such as encrypted storage and controlled access, were implemented to protect against unauthorized access and breaches. Integrating MacNet into existing diagnostic workflows involves several key considerations. The implementation should be seamless and compatible with current histopathology imaging systems. Training programs for clinicians will be essential to facilitate the adoption of MacNet, ensuring that clinicians understand how to use the technology to improve diagnostic accuracy and efficiency. Additionally, ongoing technical support will be provided to address any issues during integration. Addressing patient data privacy and ensuring seamless integration into diagnostic workflows are critical for the ethical and practical application of MacNet. These steps, combined with transparency and continuous validation, enhance the utility of MacNet in improving cervical cancer diagnosis.

Clinicians can use MacNet to quickly prescreen or assist in the analysis of pathology slides, significantly reducing the time required to identify areas of interest. The model's ability to quantitatively estimate differentiation levels within a region enhances the objectivity of diagnoses, providing a valuable second opinion that complements the pathologist's expertise. Moreover, MacNet's use of an attention mechanism with visual explanations (attention heatmaps) increases transparency, helping clinicians understand the basis for the model's decisions and thereby fostering trust in the technology.

To facilitate practical usage, methods such as U2Net should be deployed within existing information technology (IT) infrastructures in health care facilities (59). MacNet requires only standard computational resources when making diagnoses. The integration of MacNet into existing diagnostic workflows has been designed to be extremely user friendly. Clinicians will simply need to input pathology images into the program, and MacNet will return the analysis results with minimal interaction needed.

The proposed method does not use a specific module to obtain higher scores on the training dataset, and more experiments will be performed on other classification datasets to fine-tune the method to verify its generalization effect, which will show that the proposed method has the potential to complete other classification tasks in other fields.

Conclusions

This paper proposes a multiscale DL model called MacNet, which combines a mobile CNN with an attention mechanism. The attention fusion with the CNN allows MacNet to not only focus on local features but also extract global and remote information. Given the importance of multiscale information in the task, the proposed approach employs strategies used by human pathologists to efficiently utilize information at different magnifications. In addition, the end-to-end system outputs a labeled image of the same size as the input after the histopathological image, which can quantitatively show the proportion of different differentiation levels and label them with different colors. This will help doctors make faster, more convenient, and more accurate diagnoses. In future work, the accuracy and time consumption of the proposed method will be further improved so that the method can be applied to other classification tasks.

Acknowledgments

Funding: This work was supported by the National Natural Science Foundation of China (Nos. 82372038 and 62101540), the Shenzhen Excellent Technological Innovation Talent Training Project of China (No. RCJC20200714114436080), the Key Laboratory for Magnetic Resonance and Multimodality Imaging of Guangdong Province (No. 2023B1212060052), the Shenzhen Science and Technology Program (Nos. JCYJ20220818101804009 and RCBS20210706092218043), the Central Guidance for Local Science and Technology Development Project (No. ZYYD2023D02).

Footnote

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-24-810/coif). D.L. serves as an unpaid editorial board member of *Quantitative Imaging in Medicine and Surgery*. The other authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). Approval for all ethical and experimental procedures and protocols was granted by the National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital & Shenzhen Hospital (No. IRB KYLX2022-44). And informed consent was provided by all the patients.

References

1. Singh D, Vignat J, Lorenzoni V, Eslahi M, Ginsburg O, Lauby-Secretan B, Arbyn M, Basu P, Bray F, Vaccarella S. Global estimates of incidence and mortality of cervical cancer in 2020: a baseline analysis of the WHO Global Cervical Cancer Elimination Initiative. Lancet Glob Health 2023;11:e197-206.
2. Cohen PA, Jhingran A, Oaknin A, Denny L. Cervical cancer. Lancet 2019;393:169-82.
3. Xia C, Dong X, Li H, Cao M, Sun D, He S, Yang F, Yan X, Zhang S, Li N, Chen W. Cancer statistics in China and United States, 2022: profiles, trends, and determinants. Chin Med J (Engl) 2022;135:584-90.
4. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA Cancer J Clin 2023;73:17-48.
5. Eluf-Neto J, Booth M, Muñoz N, Bosch FX, Meijer CJ, Walboomers JM. Human papillomavirus and invasive cervical cancer in Brazil. Br J Cancer 1994;69:114-9.
6. Ngelangel C, Muñoz N, Bosch FX, Limson GM, Festin MR, Deacon J, Jacobs MV, Santamaria M, Meijer CJ, Walboomers JM. Causes of cervical cancer in the Philippines: a case-control study. J Natl Cancer Inst 1998;90:43-9.
7. Komura D, Ishikawa S. Machine Learning Methods for Histopathological Image Analysis. Comput Struct Biotechnol J 2018;16:34-42.
8. Phan DT, Yap KH, Garg K, Han BS. Vision-Based Early Fire and Smoke Detection for Smart Factory Applications Using FFS-YOLO. 2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP). Poitiers: IEEE; 2023.
9. Lin X, Yan Z, Deng X, Zheng C, Yu L. ConvFormer: Plug-and-Play CNN-Style Transformers for Improving Medical Image Segmentation. Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. Cham: Springer; 2023:642-51.
10. Zhang Q, Zhang J, Xu Y, Tao D. Vision Transformer With Quadrangle Attention. IEEE Trans Pattern Anal Mach Intell 2024;46:3608-24.
11. Li H, Zou Y, Wang Y, Majumder O, Xie Y, Manmatha R, Swaminathan A, Tu Z, Ermon S, Soatto S. On the Scalability of Diffusion-based Text-to-Image Generation. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE; 2024:9400-9.
12. Zhang Y, Dong Y, Zhang S, Min T, Su H, Zhu J. Exploring the Transferability of Visual Prompting for Multimodal Large Language Models. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE; 2024:26552-62.
13. Zhu Y, Wang S, Xin T, Zhang H. Few-Shot Medical Image Segmentation via a Region-Enhanced Prototypical Transformer. Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. Cham: Springer; 2023:271-80.
14. Srinidhi CL, Ciga O, Martel AL. Deep neural network models for computational histopathology: A survey. Med Image Anal 2021;67:101813.
15. Wei L, Gan Q, Ji T. Cervical cancer histology image identification method based on texture and lesion area features. Comput Assist Surg (Abingdon) 2017;22:186-99.
16. Song Y, Zhang L, Chen S, Ni D, Li B, Zhou Y, Lei B, Wang T. A deep learning based framework for accurate segmentation of cervical cytoplasm and nuclei. 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Chicago, IL, USA: IEEE; 2014:2903-6.
17. Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Süsstrunk S. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans Pattern Anal Mach Intell 2012;34:2274-82.
18. Li C, Chen H, Zhang L, Xu N, Xue D, Hu Z, Ma H, Sun H. Cervical Histopathology Image Classification Using Multilayer Hidden Conditional Random Fields and Weakly Supervised Learning. IEEE Access 2019;7:90378-97.
19. Purwanti E, Bustomi MA, Aldian RD. Applied Computing Based Artificial Neural Network for Classification of Cervical Cancer. CISAK 2013 : The 6th Conference of Indonesian Students Association in Korea. 2013.
20. Hou L, Samaras D, Kurc TM, Gao Y, Davis JE, Saltz JH. Patch-based Convolutional Neural Network for Whole Slide Tissue Image Classification. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2016;2016:2424-33.
21. Krizhevsky A, Sutskever I, Hinton GE. ImageNet

classification with deep convolutional neural networks. Communications of the ACM 2017;60:84-90.

22. Xu J, Luo X, Wang G, Gilmore H, Madabhushi A. A Deep Convolutional Neural Network for segmenting and classifying epithelial and stromal regions in histopathological images. Neurocomputing (Amst) 2016;191:214-23.
23. Huang Z, Liu Z, He P, Ren Y, Li S, Lei Y, Luo D, Liang D, Shao D, Hu Z, Zhang N. Segmentation-guided Denoising Network for Low-dose CT Imaging. Comput Methods Programs Biomed 2022;227:107199.
24. Huang Z, Liu X, Wang R, Chen Z, Yang Y, Liu X, Zheng H, Liang D, Hu Z. Learning a Deep CNN Denoising Approach Using Anatomical Prior Information Implemented With Attention Mechanism for Low-Dose CT Imaging on Clinical Patient Data From Multiple Anatomical Sites. IEEE Journal of Biomedical and Health Informatics 2021;25:3416-27.
25. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: Inverted Residuals and Linear Bottlenecks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE; 2018:4510-20.
26. Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE; 2018:7132-41.
27. Bahdanau B, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. 2015 The 3rd International Conference on Learning Representations (ICLR). San Diego, CA, USA, 2015.
28. Yang C, Qiao S, Yu Q, Yuan X, Zhu Y, Yuille A, Adam H, Chen LC. MOAT: Alternating Mobile Convolution and Attention Brings Strong Vision Models. The Eleventh International Conference on Learning Representations 2022.
29. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv:1704.04861. 2017. Available online: https://doi.org/10.48550/arXiv.1704.04861
30. Huang Z, Liu X, Wang R, Chen J, Lu P, Zhang Q, Jiang C, Yang Y, Liu X, Zheng H, Liang D, Hu Z. Considering anatomical prior information for low-dose CT image enhancement using attribute-augmented Wasserstein generative adversarial networks. Neurocomputing 2021;428:104-15.
31. Zerhouni E, Lányi D, Viana M, Gabrani M. Wide residual networks for mitosis detection. 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). Melbourne: IEEE; 2017:924-8.
32. McKinley ET, Shao J, Ellis ST, Heiser CN, Roland JT, Macedonia MC, Vega PN, Shin S, Coffey RJ, Lau KS. MIRIAM: A machine and deep learning single-cell segmentation and quantification pipeline for multi-dimensional tissue images. Cytometry A 2022;101:521-8.
33. Hu Z, Xue H, Zhang Q, Gao J, Zhang N, Zou S, Teng Y, Liu X, Yang Y, Liang D, Zhu X, Zheng H. DPIR-Net: Direct PET Image Reconstruction Based on the Wasserstein Generative Adversarial Network. IEEE Transactions on Radiation and Plasma Medical Sciences 2021;5:35-43.
34. Huang Z, Liu X, Wang R, Zhang M, Zeng X, Liu J, Yang Y, Liu X, Zheng H, Liang D, Hu Z. FaNet: fast assessment network for the novel coronavirus (COVID-19) pneumonia based on 3D CT imaging and clinical symptoms. Appl Intell (Dordr) 2021;51:2838-49.
35. Huang Z, Chen Z, Chen J, Lu P, Quan G, Du Y, Li C, Gu Z, Yang Y, Liu X, Zheng H, Liang D, Hu Z. DaNet: dose-aware network embedded with dose-level estimation for low-dose CT imaging. Phys Med Biol 2021;66:015005.
36. Aresta G, Araújo T, Kwok S, Chennamsetty SS, Safwan M, Alex V, et al. BACH: Grand challenge on breast cancer histology images. Med Image Anal 2019;56:122-39.
37. Spanhol FA, Oliveira LS, Petitjean C, Heutte L. A Dataset for Breast Cancer Histopathological Image Classification. IEEE Trans Biomed Eng 2016;63:1455-62.
38. Araújo T, Aresta G, Castro E, Rouco J, Aguiar P, Eloy C, Polónia A, Campilho A. Classification of breast cancer histology images using Convolutional Neural Networks. PLoS One 2017;12:e0177544.
39. Ji H, Zhu Q, Ma T, Cheng Y, Zhou S, Ren W, et al. Development and validation of a transformer-based CAD model for improving the consistency of BI-RADS category 3-5 nodule classification among radiologists: a multiple center study. Quant Imaging Med Surg 2023;13:3671-87.
40. Khan RF, Lee BD, Lee MS. Transformers in medical image segmentation: a narrative review. Quant Imaging Med Surg 2023;13:8747-67.
41. Tolstikhin IO, Houlsby N, Kolesnikov A, Beyer L, Zhai X, Unterthiner T, Yung J, Steiner A, Keysers D, Uszkoreit J, Lucic M, Dosovitskiy A. MLP-Mixer: An all-MLP Architecture for Vision. Advances in Neural Information Processing Systems 2021;34:24261-72.
42. Huang Z, Wu Y, Fu F, Meng N, Gu F, Wu Q, Zhou Y, Yang Y, Liu X, Zheng H, Liang D, Wang M, Hu Z. Parametric image generation with the uEXPLORER total-

body PET/CT system through deep learning. Eur J Nucl Med Mol Imaging 2022;49:2482-92.

43. Huang Z, Chen Z, Quan G, Du Y, Yang Y, Liu X, Zheng H, Liang D, Hu Z. Deep Cascade Residual Networks (DCRNs): Optimizing an Encoder–Decoder Convolutional Neural Network for Low-Dose CT Imaging. IEEE Transactions on Radiation and Plasma Medical Sciences 2022;6:829-40.
44. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE; 2016:2818-26.
45. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929. 2020. Available online: https://doi.org/10.48550/arXiv.2010.11929
46. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556. 2014. Available online: https://doi.org/10.48550/arXiv.1409.1556
47. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE; 2017:2261-9.
48. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE; 2016:770-8.
49. Bochkovskiy A, Wang CY, Liao HYM. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv:2004.10934. 2020. Available online: https://doi.org/10.48550/arXiv.2004.10934
50. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin transformer: Hierarchical vision transformer using shifted windows. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal: IEEE; 2021:9992-10002.
51. Kerdvibulvech C. Real-time Adaptive Learning System using Object Color Probability for Virtual Reality Applications. The 1st International Conference on Simulation and Modeling Methodologies, Technologies and Applications. 2011:200-4.
52. Kerdvibulvech C. A methodology for hand and finger motion analysis using adaptive probabilistic models. J Embedded Systems 2014;2014:18.
53. Phan DT, Ta QB, Huynh TC, Vo TH, Nguyen CH, Park S, Choi J, Oh J. A smart LED therapy device with an automatic facial acne vulgaris diagnosis based on deep learning and internet of things application. Comput Biol Med 2021;136:104610.
54. Phan DT, Phan TTV, Huynh TC, Park S, Choi J, Oh J. Noninvasive, Wearable Multi Biosensors for Continuous, Long-term Monitoring of Blood Pressure via Internet of Things Applications. Computers and Electrical Engineering 2022;102:108187.
55. Phan DT, Nguyen CH, Nguyen TDP, Tran LH, Park S, Choi J, Lee BI, Oh J. A Flexible, Wearable, and Wireless Biosensor Patch with Internet of Medical Things Applications. Biosensors (Basel) 2022;12:139.
56. Fu B, Zhang M, He J, Cao Y, Guo Y, Wang R. StoHisNet: A hybrid multi-classification model with CNN and Transformer for gastric pathology images. Comput Methods Programs Biomed 2022;221:106924.
57. Diede C, Walker T, Carr DR, Shahwan KT. Grading differentiation in cutaneous squamous cell carcinoma: a review of the literature. Arch Dermatol Res 2024;316:434.
58. Broders AC. Carcinoma: Grading and practical application. Arch Pathol 1956;2:376-81.
59. Phan DT, Ta QB, Ly CD, Nguyen CH, Park S, Choi J, Hwi O S, Oh J. Smart Low Level Laser Therapy System for Automatic Facial Dermatological Disorder Diagnosis. IEEE J Biomed Health Inform 2023. [Epub ahead of print]. doi: 10.1109/JBHI.2023.3237875.

Cite this article as: An Y, Lei Y, Huang Z, Liu Y, Huang M, Liu Z, Li W, Liang D, Huang W, Hu Z. MacNet: a mobile attention classification network combining convolutional neural network and transformer for the differentiation of cervical cancer. Quant Imaging Med Surg 2025;15(1):55-73. doi: 10.21037/qims-24-810