Computers in Biology and Medicine 148 (2022) 105821

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Automated classification of cervical lymph-node-level from ultrasound using Depthwise Separable Convolutional Swin Transformer☆

Yanting Liu [a], Junjuan Zhao [a], Quanyong Luo [b], Chentian Shen [b], Ren Wang [c,*], Xuehai Ding [a,**]

[a] School of Computer Engineering and Science, Shanghai University, Shangda Rd, Shanghai, 200444, China
[b] Department of Nuclear Medicine, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Yishan Rd, Shanghai, 200233, China
[c] Department of Ultrasound Medicine, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Yishan Rd, Shanghai, 200233, China

ARTICLE INFO

Keywords:
Cervical ultrasound
Lymph-node-level classification
Depthwise separable convolution
Data imbalance
Visualization

ABSTRACT

There are few studies on cervical ultrasound lymph-node-level classification which is very important for qualitative diagnosis and surgical treatment of diseases. Currently, ultrasound examination relies on the subjective experience of physicians to judge the level of the cervical lymph nodes, which is easily misclassified. Unlike other automated diagnostic tasks, lymph-node-level classification needs to focus on global structural information. Besides, there is a large range of sternocleidomastoid muscles in levels II, III and IV, which leads to small inter-class differences in these levels, so it also needs to focus on key local areas to extract strong distinguishable features. In this paper, we propose the Depthwise Separable Convolutional Swin Transformer, introducing the deepwise separable convolution branch into the self-attention mechanism to capture discriminative local features. Meanwhile, to address the problem of data imbalance, a new loss function is proposed to improve the performance of the classification network. In addition, for the ultrasound data collected by different devices, low contrast and blurring problems of ultrasound imaging, a unified pre-processing algorithm is designed. The model was validated on 1146 cases of cervical ultrasound lymph node collected from the Sixth People's Hospital of Shanghai. The average accuracy precision, sensitivity, specificity, and F1 value of the model for the valid dataset after five-fold cross-validation were 80.65%, 80.68%, 78.73%, 95.99% and 79.42%, respectively. It has been verified by visualization methods that the Region of Interest (ROI) of the model is similar or consistent with the observed region of the experts.

1. Introduction

Lymph nodes are important immune organs found in the body, and abnormalities in lymph nodes such as enlarged lymph nodes, may be the clinical manifestations of lesions in the regions to which they belong. This has important clinical significance. In recent years, the prevalence of lymph node-related diseases has been increasing, and the neck is a high-prevalence area for lymph node-related diseases, so lymph nodes in the neck have become a common item for medical examination. With the development of high-frequency ultrasound imaging technology, the diagnostic level of ultrasounds for cervical lymph-node diseases has been significantly improved, and CT and MRI examinations are expensive, so ultrasound is now the preferred imaging method for cervical lymph nodes. Further CT or puncture examination is performed only when the lymph nodes are highly suspected. The ultrasound diagnosis of cervical lymph nodes is mainly performed using a color Doppler ultrasound scanner. The physician uses a probe to examine the patient's neck horizontally or longitudinally, and makes detailed records of the lymph-node size, boundary, shape and blood flow. During the imaging evaluation, the physician often marks the lymph-node boundaries on the ultrasound image to calculate the aspect ratio, observes the lymph-node boundaries, cortex and other characteristics, and determines the lymph-node benignity and malignancy using their subjective experience, as well as, determining the current lymph-node location by dynamically sweeping the probe to the peripheral position in real time. Cervical lymph-node-level classification is very helpful for the qualitative diagnosis of disease, and is an important basis for the grading and staging of malignant tumors, as well as having important value for the survival of tumor patients and the detection of local recurrence and distant metastases. For example, papillary thyroid cancer tends to metastasize to level VI but rarely to level I. Lymph

☆ This research obtained the ethics certificate and arose from the project about the establishment and application of an intelligent diagnosis system for thyroid nodules and cervical lymph nodes based on AI and multi-omics.

* Corresponding author.
** Corresponding author.
E-mail addresses: SylviaLau@shu.edu.cn (Y. Liu), dinghai@shu.edu.cn (X. Ding).

https://doi.org/10.1016/j.compbiomed.2022.105821
Received 10 April 2022; Received in revised form 13 June 2022; Accepted 3 July 2022
Available online 5 July 2022

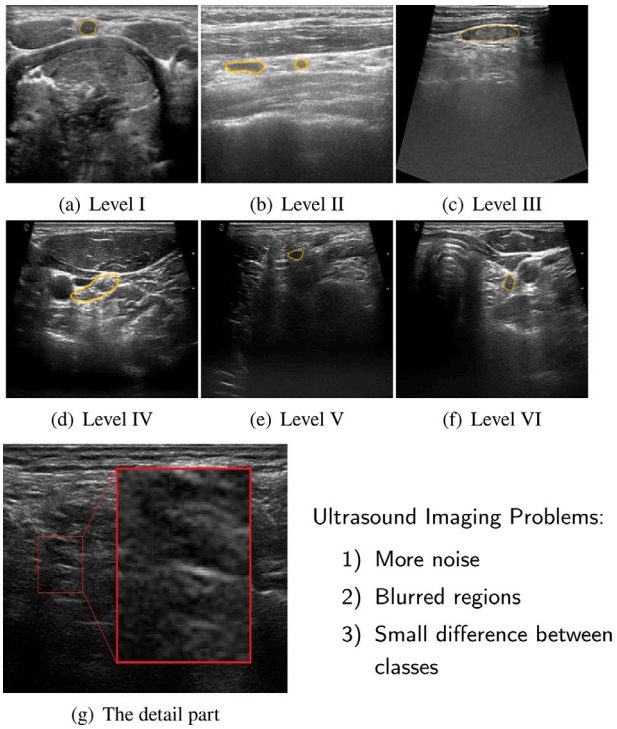

Fig. 1. Examples of cervical lymph-node-level classification on ultrasound images. (a)–(f) are selected ultrasound images of cervical lymph-node levels I through VI. (g) shows a detailed area of a cervical ultrasound image. Yellow markers denote position of lymph nodes, audited by one radiologist.

nodes in levels III and IV are common metastatic areas for laryngeal cancer, laryngopharyngeal cancer and thyroid cancer. In addition, the accurate assessment of lymph-node status and lymph-node-level determines whether lymph-node dissection should be performed in the neck and what extent dissection should be carried out. If the metastatic lymph nodes are missed, residual tumors may result, leading to the need for secondary surgery; if overdiagnosis occurs, damage by overtreatment may result. Therefore, the judgement of cervical lymph-node-level is closely related to the development of a surgical plan [1].

The judgement of cervical lymph-node-level is often performed using the criteria of surgical zoning, i.e., six levels are defined according to the 2002 American Association for Head and Neck Surgery method of lymph-node-level delineation [2]. Level I is submandibular lymph nodes; levels II through IV are lymph nodes of the superior, middle, and inferior groups of the internal jugular vein chain, respectively; level V is the posterior triangular region of the lymph nodes and the supraclavicular fossa; and level VI is the central region of the lymph nodes. Lymph node levels are not all defined by tissues surrounding. For example, the border of levels II and III is horizontal extension of the central cervical hyoid muscle instead of the tissues in level II or III. Moreover, some tissues may not be displayed and identified on ultrasound. The ultrasound imaging effect of each level is shown in Fig. 1(a)–(f). Therefore, physicians cannot directly judge lymph-node-level by the Head and Neck Surgery method. They need to perform dynamic peripheral scans to look for clearly distinguishable peripheral tissues to identify the level. There are some imaging features summarized by physicians, such as the presence of the mandibular hyoid muscle usually in level I; and the presence of the thyroid gland as well as the hypoechoic trachea in level VI. However, the sternocleidomastoid muscle straddles levels II, III, and IV and takes up a large proportion of the area, leading to a decrease in the distinguishability of these levels, so the judgement of these levels is influenced by subjective factors and the clinical experience of the physician. In addition as shown in Fig. 1(g), ultrasound imaging with many noisy spots, blurred borders, and low contrast also makes classification difficult. These problems lead to low efficiency and low accuracy of cervical lymph-node-level classification as well as inconsistent results from different physicians. Therefore, the development of an intelligent assisted diagnosis and treatment system for cervical lymph-node-level classification can effectively assist physicians in pathological diagnosis and provide effective guidance on the need for further detailed diagnosis and subsequent surgical treatment and prognostic analysis. To summarize, this work addressed three challenges: (1) low ultrasound resolution, low contrast, unclear soft tissue boundaries, and a lot of noise; (2) small differences in categories between some levels, making it difficult to find distinctive features; and (3) small dataset and unbalanced categories.

This study considers the imaging features of cervical lymph nodes and regional texture features, so the Depthwise Separable Convolutional Swin Transformer is designed, which achieves global relationship modeling of the tissue structure around the lymph nodes, learns local regional features and carries out automatic lymph-node-level classification to assist doctor's clinical decision-making, which solves the subjective problem of traditional ultrasound diagnosis, and prevents excessive treatment and undertreatment to the greatest extent. In summary, the contributions of the article are as follows:

(1) A complete method for cervical lymph-node-level classification in ultrasound is proposed to assist doctors to make clinical decision and understand model decision by visualization methods.

(2) For the problems about large intra-class variation, small inter-class variation and data imbalance, we propose the Depthwise Separable Convolutional Swin Transformer model and new loss function.

(3) We design a unified ultrasound pre-processing method for different ultrasonic devices to effectively solve the ultrasound imaging problem as well as to improve the classification performance.

2. Related work

2.1. Medical imaging classification

There has much research on the automatic diagnosis of benign and malignant lymph nodes through ultrasound. However, there has been a lack of research on the classification of cervical lymph-node-level. This task can refer to classification research related to ultrasound. Ultrasound-related classification research methods are mainly divided into traditional statistical methods, machine learning methods and deep learning methods.

Traditional statistical methods and machine-learning methods rely on a large amount of medical knowledge, especially the research and analysis of imaging omics. Daoud MI [3] et al. introduced an improved texture analysis method, used the gray level co-occurrence matrix and support vector machine classifier to analyze each region of interest, and used a voting mechanism to combine all ROIs to estimate the breast cancer tumor category. CK Zhao [4] et al. quantified 10 types of ultrasound image features, introduced rough sets into support vector machines, and designed an auxiliary analysis system that integrates thyroid segmentation, thyroid nodule classification and recognition. Due to the problem of lymph nodes being small and the interference of tissues such as blood vessels around them, Kan Y [5] and others used radiomics and the physiological characteristics of the surrounding lymph nodes to first identify the lymph-node area, thereby reducing the interference from surrounding tissues and removing similar features present in benign and malignant lymph nodes from the ultrasound. Then, they used support vector machines to classify the distinguishable features, thereby improving the classification performance.

With the continuous development and application of deep-learning technology, research on ultrasound classification based on deep learning has received great attention. Cordes M [6] et al. collected ultrasound and clinical features, and designed a simple fully connected

network to classify and distinguish thyroid cancer. Lee JH [7] et al. proposed the CNN-GAP algorithm, which not only generates heatmaps to locate lymph nodes, but also replaces the fully connected layer in the VGG network with a global average pooling layer to train a high-precision classification network. Based on this algorithm, the CAD system was developed for the ultrasound positioning and diagnosis of metastatic lymph nodes in thyroid cancer, which has high sensitivity and relatively low specificity. Song R [8] et al. used a dual-branch network for the ultrasound classification of thyroid nodules. After feature extraction, the feature map was randomly cropped and the branch network was introduced to learn local detailed features and global features. This method extracts richer image features through multi-branch network training and feature fusion, which improves the classification accuracy.

Since the cervical lymph-node-level classification task is a multi-classification task that requires refined classification, this research also referred to relevant literature on the medical image multi-classification task. Xie H [9] and others used a multi-branch network based on ResNet to achieve the fine classification of multiple diseases about the fundus and introduced a hollow convolution module and a cross-CBAM attention mechanism for feature fusion between branches.

2.2. Deep learning model

Currently, deep learning is gradually applied in the medical field, where CNN model is a classical network structure including convolutional, normalization, and pooling layers from LeNet [10] and AlexNet [11] model. Most CNN models [12–15] are better at extracting local features by deep networks, but it is difficult to capture the global information. The SENet network model [16] was proposed, in which the SE block uses the global context to update the weights of different channels, but it does not make full use of the global contextual information.

Enhancing the global understanding requires Vision Transformer to capture long-range dependencies [17]. The pioneer work for the Visual Transformer in the field of image vision is known as ViT [18]. The area of innovation lies in the use of the image patch. The patch is used as the token in the NLP task, which is linearly mapped into an embedding vector. Then, the multi-scale feature hierarchical structure was designed into Transformer to reduce the spatial resolution in stages, which not only can produce high-dimensional semantic features, but also can make the model suitable for various visual tasks [19,20]. The second direction of improvement is the optimization of sparse self-attention to alleviate the large-scale calculation and number of parameters [21,22]. In the self-attention mechanism, k and v are compressed through convolution or the average pool. The third is optimization of training strategy. In [23], it was proposed that DeiT not only uses the method of mixed data enhancement, but also uses the method of knowledge distillation in the training process, which not only speeds up training and also improves the classification accuracy. In [24], the Swin Transformer network was used to solve the ViT problem. The self-attention calculation with shifted non-overlapping windows solves the problem of the extraction of local feature information. Designing a local self-attention structure also requires effective global information interaction capabilities; otherwise, it will weaken the feature representation extracted by the model. In [25], Shuffle Transformer was used to obtain input information from different windows through spatial mixing to improve information interactions between windows. In [26], Twins was proposed to achieve local and global information modeling through a layer of local self-attention and a layer of global self-attention. In [27], Cswin was proposed, where each query point related to the horizontal and vertical stripes of the point is designed to perform a self-attention calculation. In [28], Focal Transformer was proposed, and a coarse and fine-grained global interaction attention mechanism was used, giving more weight to neighboring parts and less weight to remote areas. The combination of local modeling with CNN and global modeling with Transformer can result in a good complementarity. In [29], the use of LocalViT to add local feature extraction capabilities to the Visual Transformer by introducing separable convolution into the feedforward network was proposed. CvT [30] maps tokens through convolution and uses convolution to generate q, k, and v of self-attention. These changes introduce the features of shift, scaling, and distortion invariance in CNN into the ViT architecture. The CoAtNet [31] model combines the convolutional layer and the self-attention layer to improve the learning ability and generalization ability of the model.

Since deep learning has obtained excellent results in medical diagnostic tasks, we believe it is also feasible to study deep learning-based cervical lymph-node-level classification. But current pure Transformers or with CNN are insufficient to extract discriminative local features. Therefore, our method including model is presented in the following chapters.

3. Method

As shown in the pseudo code of Algorithm 1, the entire process of the cervical ultrasound lymph-node-level classification task can be divided into four stages. At first, the original cervical ultrasound image should be preprocessed, and we propose a complete process for the preprocessing of ultrasound images. The purpose of this is to remove irrelevant areas and interference markers to prevent the adverse effect of noise on classification performance. This step includes cutting out the ROI of an ultrasound image, performing histogram equalization and removing markers to complete the data preparation of the input model. Secondly, to address the problem of class imbalance, small differences in characteristics between classes, and excessive background noise, data augmentation including random cropping, mixup and cutmix, can be considered. These augmentation methods can weaken data noise and increase model stability. Then the ultrasound data can be trained by the model proposed in this paper. During this stage, we use a novel loss function and cosine annealing scheduler with the warm up to improve the robustness of the model and reduce intra-class variation. Finally, in order to understand the classification decision of the model, heatmaps are generated for each ultrasound image, and the logic of the model classification can be inferred from the sensitive areas. This step not only verifies the classification performance of the model, but also improves doctor's confidence in the classification of the model.

Algorithm 1: Overall Process

Data: ultrasound original images $\{\boldsymbol{U}\}$
Result: prediction labels $\{\boldsymbol{P}\}$ and heatmaps $\{\boldsymbol{H}\}$
1 get $\{\boldsymbol{U'}\}$ from ultrasound data preprocessing algorithm;
2 get $\{\boldsymbol{E}\}$ through data augmentation, including random crop, mixup and cutmix methods;
3 choose cosine annealing learning strategy with warm restarts;
4 train our model by dataset $\{\boldsymbol{E}\}$, and save model parameters;
5 get prediction labels $\{\boldsymbol{P}\}$ from trained model ;
6 get heatmaps $\{\boldsymbol{H}\}$ by visualization methods;
7 evaluate the performance of model;

3.1. Data pre-processing work

As mentioned above, the first stage of data preprocessing is divided into three main parts: cutting out the ultrasound image ROI, removing markers and performing histogram equalization. These parts help to improve the classification performance. The specific data preprocessing algorithm is shown in the pseudo code of Algorithm 2. Fig. 2 shows results of an example ultrasound image through steps 2, 3 and 4.

(1) Step 1: Desensitize the ultrasound in Dicom format (eliminate the patient's personal information), convert it into a lossless bmp

Algorithm 2: Ultrasound Data Preprocessing

Data: one ultrasound original image ***U***
Result: ultrasound image preprocessed ***R***
get grayscale image ***G*** related to ***U*** ;
get ***h_array*** ,***v_array*** by calculating the sum of variances between rows and columns;
smooth curves of ***h_array*** ,***v_array*** by one dimensional convolution;
get ***hgra_array*** ,***vgra_array*** by calculating the sum of variances between rows and columns;
get upper and lower borders ***top*** ,***bottom*** by finding foremost and last peaks of ***hgra_array*** ;
get ***vgra_peaks_mean*** by finding all peaks of ***vgra_array*** ;
if *max(* ***vgra_array****)-* ***vgra_peaks_mean*** *< threshold* ***D*** **then**
    calculate the second order derivative of ***vgra_array*** ;
end
get left and right borders ***left*** ,***right*** by finding foremost and last peaks of ***vgra_array*** ;
get location [***left,top,right,bottom***] of max connected region in ***G***;
crop out the ultrasound region ***GR*** from ***G*** by location [***left,top,right,bottom***] ;
suppose maximum and minimum area threshold ***MinC MaxC*** of criss-cross marker ;
dilate and erode ***GR*** by opencv and find all connected regions ***C***;
for ***c in C*** **do**
    if *c.area<* ***MaxC*** *and c.area>* ***MinC*** **then**
        use opencv inpaint method to eliminate criss-cross markers in ***U***
    end
end
crop out the ultrasound region ***R*** from ***U*** by location [***left,top,right,bottom***] ;

format, and delete color Doppler images from the dataset and multiple repeatedly saved ultrasound images.

(2) Step 2: Because the original ultrasound image contains the regions about device information which is not related to the task and represents a large enough proportion, it obviously affects the classification performance. In this work, we design an algorithm, which is suitable for all interfaces from different devices to allocate ultrasound region.

(3) Step 3: Although the irrelevant region has been removed, some markers are still located inside the ultrasound region. Doctors often need to make criss-cross markers on the border of the lymph node to measure the aspect ratio of the lymph node, which is very important for the identification of the lymph nodes. In addition, the supplier logo and ultrasonic control indicator also appear in the ultrasound region. These markers easily take the attention of the model, become noise areas, and cause interference in the model training. The positions of the supplier logo and ultrasonic control indicator are fixed, but the positions of the criss-cross markers are changeable. The algorithm designed in this study can recognize the positions of the criss-cross markers. Through the opencv area repair method, the supplier's logo, ultrasonic control indicator and the criss-cross markers manually made by the doctor are removed.

(4) Step 4: Since the ultrasound data came from different types of ultrasound equipment, and the imaging indicators of different types of ultrasound equipment are different, the contrast and brightness of ultrasound data differ. This study used the adaptive threshold histogram equalization (CLAHE) to adjust Ultrasound contrast and then normalized the ultrasonic pixel value, before converting it to the size of 224 × 224.

(5) Step 5: Divide and train the dataset by five-fold cross-validation.

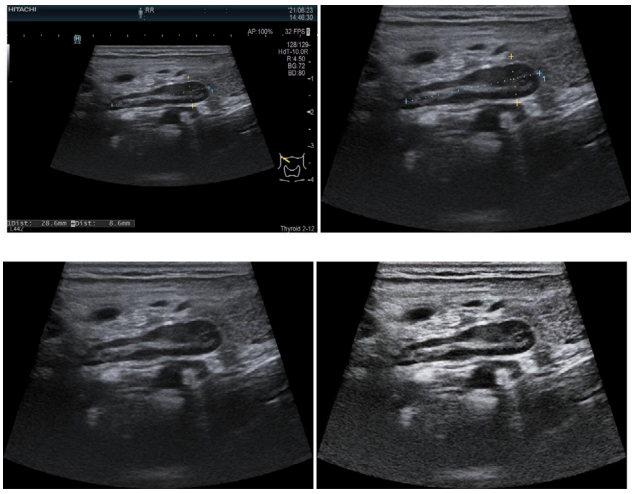

Fig. 2. Results of an example ultrasound image through steps 2, 3 and 4.

3.2. Depthwise separable convolutional swin transformer

For ultrasound images of the cervical lymph nodes, the classification task not only needs to pay attention to location information for the local area, but it also needs to obtain the overall structural information. As mentioned above, there are some problems about ultrasound imaging such as noise and low contrast, and the proportion of background information on an ultrasound is relatively large, while the discriminating target object may be small. Only using CNN to extract features may extract noise. As a result, distinguishing features cannot be extracted, resulting in a decline in classification performance. Moreover, the CNN convolution kernel is artificially set and only takes into account the information in the frame of the receptive field. It has insufficient ability to extract contextual information. The latest Visual Transformer takes self-attention as the key. Although it can capture remote dependencies and extract global information and rich context information with greater flexibility, it has several problems: due to its spatial self-attention calculations, a picture composed of pixels requires a huge amount of calculation that increases with the square of the input pixels, which greatly reduces its performance; secondly, overall spatial relationship modeling is performed while ignoring local feature information capture. The effective combination of Transformer and CNN can make up for their respective problems. The latest research shows that the introduction of the CNN convolution module before the transformer self-attention calculation can improve the peak performance of the model and the stability of training. However, this is still not enough to significantly improve the ability to extract local feature information.

In this study, we propose a model that fuses Transformer and CNN, and the overall structure is shown in Fig. 3. The model is improved based on the Swin Transformer model by incorporating depthwise separable convolution in the self-attention module. The window-shifting self-attention mechanism of Swin Transformer helps to reduce the redundant global spatial relationship modeling parameters and enhance local feature information acquisition. However, this structure is still not sufficient to improve the feature extraction capability of the model for local information because the smallest unit in the local self-attention is 4 × 4 patch, which may miss some fine features on the ultrasound. The self-attention module, which we design, calculate the self-attention mechanism by transforming the image blocks into two-dimensional features after embedding them into a moving window at the same time as well as carrying out the convolution calculation, after which not only the global feature information but also the local feature information on the ultrasound can be obtained, providing the model with information on different scales and different semantics.

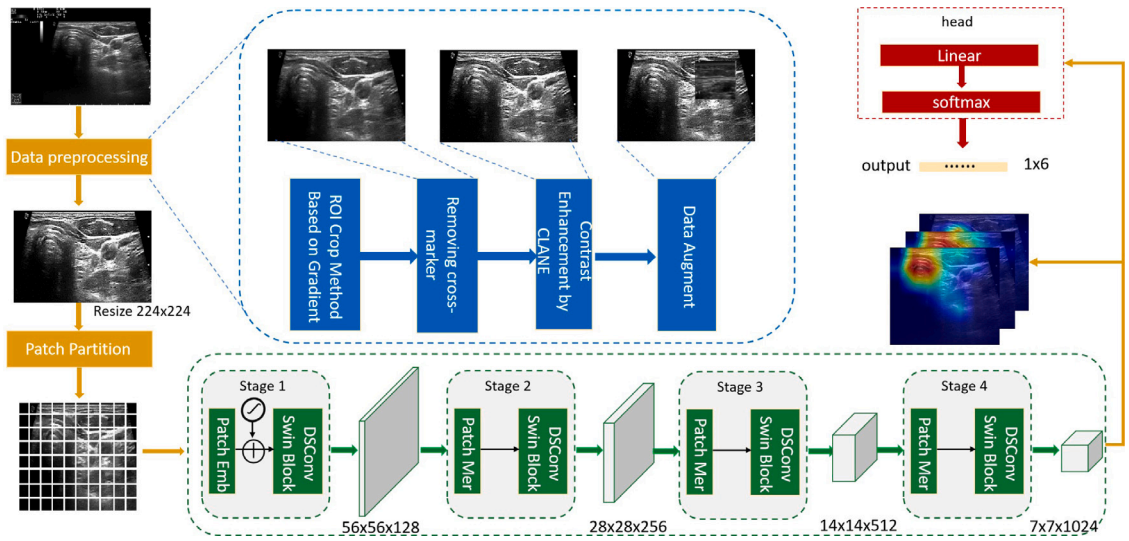

Fig. 3. Flowchart for the proposed Depthwise Separable Convolutional Swin Transformer. The 224 × 224pi ultrasound image can be obtained through preprocessing work and data enhancement, and after being input into the model, it will be cut into 4 × 4 patches. After the patch embedding module is carried out, it is flattened into a 3136 × 128 vector and goes through 4 stages. Every stage contains different number of DSConvSwinBlocks. The latter stages include the patch merging module instead of the patch embedding module used in the first stage. The output is a 1 × 6 vector and heatmap related to the input ultrasound image.

Depthwise Separable Convolutional Swin Transformer is mainly composed of Patch Partition, Linear Embedding, Patch Merging, and DSConvSwin Block. It is mainly divided into 4 stages. Each stage is composed of Patch Merging and multiple DSConvSwin Blocks (the first stage is Linear Embedding instead of Patch Merging). The following describes the principles of each layer :

(1) Patch Partition: The image is mainly divided into patches, which is similar to a token used in natural language processing, and the purpose is to represent the image with multiple image patches. If it is a three-channel image, the image block is generally cut into the $4 \times 4 \times 3$ size.

(2) Linear Embedding: Feature mapping is performed on the cut image blocks to generate a high-dimensional embedding vector, and the absolute position encoding corresponding to each image patch is added.

(3) Patch Merging: In order to realize applicability to downstream tasks and to extract high-dimensional semantic features, it is necessary to reduce the level of the resolution of the feature map while increasing the feature dimension. The feature maps calculated by self-attention are divided into a group by 2×2 adjacent blocks, spliced on the feature channel, and then compressed by the linear layer to double the number of dimensions.

(4) DSConvSwin Block: The self-attention block with the moving window is composed of two local self-attention layers. Each local self-attention layer is composed of a window self-attention mechanism and MLP, and both have LayerNorm, GeLU and residual connect. Besides, there has a depthwise separable convolution branch fused with each window self-attention mechanism.

Depthwise Separable Convolutional Swin Transformer passes through four stages, and high-dimensional feature vectors can be obtained, which can be used to complete image classification tasks and downstream tasks such as target detection and segmentation.

3.3. DSConvSwin block

The depthwise separable convolutional block (DSConvSwin Block) is the core module designed in this work. The module structure is shown in Fig. 4. Since the ultrasound classification of cervical lymph-node-level is different from benign and malignant identification, it is most important to capture the overall structural information rather than focusing on just one local region. Thus, the use of the Transformer with the self-attention mechanism is preferred over the simple CNN model. We believe that Swin Transformer with the shifting window can effectively capture the important structural features in the whole ultrasound, but the ability to capture small and low-level features, which are also more important local information for lymph-node-level classification, is still insufficient. Therefore, we propose an improved self-attention module based on the Swin Transformer Block. This paper introduces the depthwise separable convolution branch into the self-attention mechanism to enhance the capture capability of local information.

Each self-attention mechanism layer introduces a depthwise separable convolution branch. This branch and self-attention mechanism take the features reshaped into two dimensions, perform self-attention mechanism computation with the shifting window and depthwise separable convolution computation in parallel, and output the fusion feature maps. The depthwise separable convolution calculation process involves the calculation of the features by depthwise convolution with a 3×3 convolution kernel size, which is represented by dconv, and pointwise convolution with a 1×1 convolution kernel size, which is represented by pconv. Then, the features are expanded into one dimension, and layer-normalized by LN. After Dropout function, the fusion feature maps are added to the input with residual connect, and then go through MLP layer. Depthwise separable convolution can minimize the increase in the number of model parameters while retaining the advantages of convolution. By referring to and improving the formula of consecutive Swin Transformer block [24], two consecutive DSConvSwin Blocks can be expressed as follows:

$$\hat{z}^{l}_{att} = W - MSA\left(LN\left(z^{l-1}\right)\right), \tag{1}$$

$$\hat{z}^{l} = \mathrm{DP}\left(p_{\mathrm{conv}}\ \left(d_{\mathrm{conv}}\ \left(z^{l-1}\right)\right) + \hat{z}^{l}_{\mathrm{att}}\right) + z^{l-1}, \tag{2}$$

$$z^{l} = MLP\left(LN\left(\hat{z}^{l}\right)\right) + \hat{z}^{l}, \tag{3}$$

$$\hat{z}^{l+1}_{att} = SW - MSA\left(LN\left(z^{l}\right)\right), \tag{4}$$

$$\hat{z}^{l+1} = DP\left(p_{\mathrm{conv}}\ \left(d_{\mathrm{conv}}\ \left(z^{l}\right)\right) + \hat{z}^{l+1}_{\mathrm{att}}\right) + z^{l}, \tag{5}$$

$$z^{l+1} = \mathrm{MLP}\left(LN\left(\hat{z}^{l+1}\right)\right) + \hat{z}^{l+1} \tag{6}$$

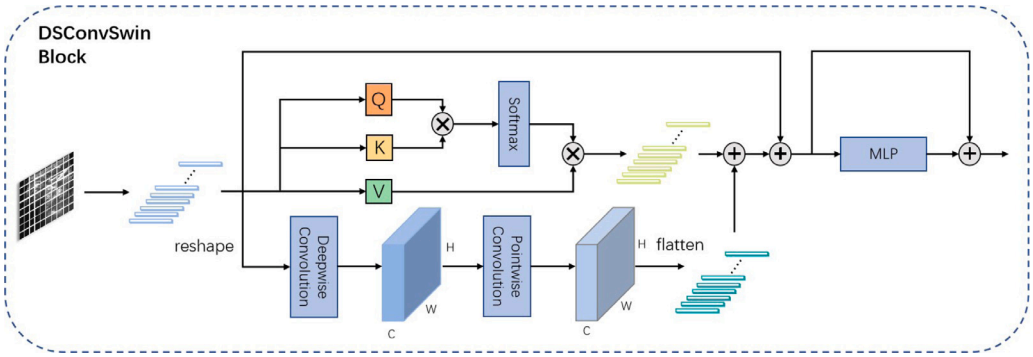

Fig. 4. The detailed structure of DSConvSwinBlock. 'Reshape' means that (wxh,c) data are transferred to (w,h,c) data. 'Flatten' means that (w,h,c) data are transferred to (wxh,c) data.

where $\hat{z}^l_{att}$ $\hat{z}^l$ and z^l denote the output features of the (S)W-MSA module, depthwise separable convolution module and the MLP module for block l, respectively; W-MSA and SW-MSA denote window based multi-head self-attention using regular and shifted window partitioning configurations; LN denotes LayerNorm layer, MLP denotes multi-layer perceptron neural network and DP denotes dropout layer; p_{conv} and d_{conv} denote pointwise convolution and depthwise convolution.

3.4. Loss function

After analyzing the ultrasound data of cervical lymph nodes, we found that the amounts of data in each level are extremely unbalanced; there are too many data in the level VI and there are fewer data in levels I, II. The reason for these factors is that the level VI is the high-risk area for thyroid cancer, so it contains more data. Meanwhile, levels I and II are lower-risk areas. Therefore, in model training, model tends to learn the data in level VI, while ignoring the feature information of other levels. This study uses a new loss function based on CB Loss [32] to alleviate sample imbalance, reduce intra-class difference and let the model focus on learning difficult samples.

The Focal CB Loss function formula is defined in (7) (8):

$$p_i^t = \mathrm{softmax}\left(z_i^t\right) = \frac{\exp\left(z_i^t\right)}{\sum_{j=1}^{C}\exp\left(z_j^t\right)} \tag{7}$$

$$\mathrm{FocalCBLoss}(z,y) = \frac{1-\beta}{1-\beta^{n_y}}\sum_{i=1}^{C}\left(1-p_i^t\right)^{\gamma}\log\left(p_i^t\right) \tag{8}$$

The FCCMix Loss function proposed in this paper is defined in (9):

$$\mathrm{FCCMixLoss}(z,y,x) = FocalCBLoss(z,y) + \lambda\sum_{i=1}^{m}\left\|x_i - c_{y_i}\right\|_2^2 \tag{9}$$

where z denotes the output of model, y denotes the target labels, C denotes the class number, n_y denotes the number of samples corresponding to the label y, x denotes the feature vector through the last DSConvSwin Block and global average pooling, and c_{yi} denotes average feature vector of x corresponding to the label y in one batch. In formula (8), the hyper-parameter β denotes the degree of re-weighting with effective number of samples and γ is an adjustable factor of Focal Loss. In formula (9), the hyper-parameters λ is the weight of FCCMix Loss.

To sum up the above, we proposed a complete method for cervical lymph-node-level classification in ultrasound, including data preprocessing algorithm, DSConvSwin Transformer model and loss function. Our innovation work helps to solve ultrasound imaging problem, data imbalance as well as intra-class and inter-class difference. By the following experiments, it can be proved that our work effectively improves classification performance of lymph-node-level.

Table 1
The data distribution of cervical lymph-node-level classification.

Class	Total data	Number of cases
1	227	148
2	343	191
3	364	153
4	512	221
5	234	144
6	588	289
All	2268	1146

4. Experimental results

This study proposed an optimized and novel network model based on Swin Transformer and applied it to the task of cervical lymph-node-level classification. We validated the collected ultrasound dataset of cervical lymph nodes from real patients. By the score indexes commonly used in medical classification tasks, we compared the latest CNN series, Transformer series and MLP series benchmark models. The network proposed in this article achieved the best results in this task.

4.1. Dataset

The number of patients participating in the study was 1146. The collected dataset included 2268 ultrasound images of cervical lymph nodes in six levels correctly marked by the radiologists. Detailed descriptions of the lymph-node ultrasound types in the six levels involved in this study are given in Table 1. According to the data distribution of the whole data set, about 6% 7% of the data in each level are randomly sampled as the test set. The remaining data were trained through 5-fold cross-validation. The dataset was collected from October 2020 to August 2021, and the data were acquired during routine ultrasound examinations performed by the Imaging Department of Shanghai Sixth People's Hospital. Ultrasound imaging is performed with a Siemens ultrasound system. All the images have been processed to filter out sensitive information. This study has been approved by the Ethics Committee of Shanghai Sixth People's Hospital.

4.2. Experimental environment

All the experiments described in this paper used the GPU server provided by the machine learning platform of Shanghai University. The GPU configuration is one NVIDIA Tesla V100, the operating environment is CentOS7, the graphics card memory is 32 GB, the programming language environment is Python 3.6, and the deep-learning framework is Pytorch.

4.3. Evaluation

The cervical lymph-node-level task is a multi-class task problem. In this study, we selected common multi-class evaluation indicators. The

Table 2
The classification performances of different models (%).

Model	Param	AUC	ACC	PPV	NPV	TPR	TNR	F1
CNN								
ResNet152 [14]	58.16	92.00	73.64	73.70	94.64	71.86	94.57	72.49
DenseNet201 [15]	18.10	90.33	69.85	71.10	93.94	67.62	93.72	68.46
EfficientNetb7 [33]	63.80	93.17	76.60	76.26	95.25	74.87	95.19	75.22
MLP								
MlpMixerB16 [34]	59.12	92.67	73.94	73.39	94.72	72.13	94.68	72.41
ResMlp36 [35]	44.31	94.17	75.57	75.60	95.03	74.14	95.00	74.55
Transformer								
ViTB [18]	85.80	93.17	75.15	76.42	94.96	72.78	94.84	74.05
VisformerS [20]	39.46	91.67	72.37	73.08	94.39	70.23	94.26	71.16
SwinB [24]	86.75	94.67	77.96	77.76	95.57	75.90	95.44	76.54
Ours(DscSwinB)	93.86	95.33	78.53	78.00	95.66	76.86	95.58	77.19

Param means model parameter size. AUC means the area under the ROC curve.
ACC = Accuracy, PPV = Positive Predictive Value, NPV = Negative Predictive Value, TPR = Sensitivity, TNR = specificity.

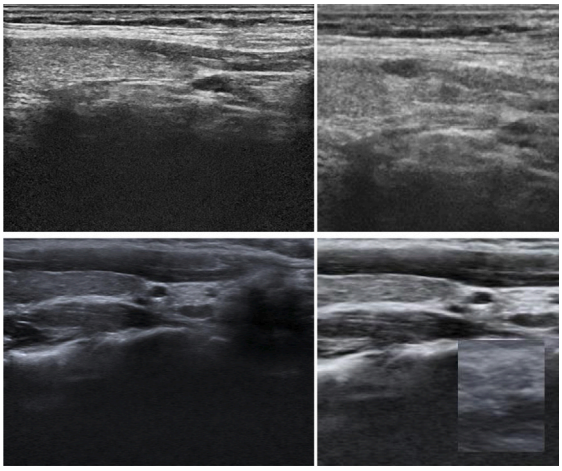

Fig. 5. Raw images and data-enhanced images. The first column contains two raw images. The top right is the image after preprocessing and mixup. The bottom right is the image after preprocessing and cutmix.

formulas of these indicators are as follows:

$$ACC = \frac{TP + TN}{\text{Total}} \tag{10}$$

$$PPV = \frac{TP}{TP + FP} \tag{11}$$

$$NPV = \frac{TN}{TN + FN} \tag{12}$$

$$TPR = \frac{TP}{TP + FN} \tag{13}$$

$$TNR = \frac{TN}{TN + FP} \tag{14}$$

$$F_1 = \frac{\left(1 + \beta^2\right) \cdot \text{ Precision } \cdot \text{ Recall}}{\beta^2(\text{Precision } + \text{ Recall})}, \beta = 1 \tag{15}$$

The indicators includes ACC, PPV, NPV, TPR, TNR and F1. ACC, or accuracy, refers to the overall classification performance of the model; PPV is the proportion of true positives in the data predicted as positive by the model; NPV is the proportion of true negatives in the data predicted as negative by the model. TPR represents Sensitivity and Recall, which means true positive rate. TNR represents Specificity, which means true negative rate. F1 is a mixed indicator of accuracy and recall.

4.4. Test result

In this study, the collected cervical lymph-node dataset is first subjected to data cleaning as well as data preprocessing as described above. As shown in Table 2, in this study, our model is experimentally compared with current mainstream models and uniformly initializes with the pre-training parameters on the ImageNet dataset. The model is optimized by SGD optimizer and cosine annealing scheduler in which we set T_mult to 2, T_cur to 100 and learning rate to 0.001 with warm up and early stopping. The images are resized into $3 \times 224 \times 224$ and the batch size is set to 8. For the FCCMix loss, the hyper-parameters β γ λ are set to 0.99, 0.5, 0.001 respectively in our experiment. In addition, after train data is randomly cropped, mixup and cutmix enhancement methods are used in this paper, which ratio is 8 to 2(see Fig. 5).

As shown in Table 2, Figs. 6 and 8, by five-fold cross-validation, our model obtained the best results in each index, followed by the Swin Transformer model. From the ROC curve of Fig. 6, it can be concluded that our model has the best results in terms of the reconciliation of true positive rate and false positive rate. The average accuracy, precision, recall and F1 value on test set is 72.75%, 72.73%, 74.07% and 73.29%. Under the first round of cross validation, the confusion matrix in Fig. 7 shows the satisfying result in our model on the test set. We can conclude that our model has the best results in levels I and VI and a slightly lower classification performance in levels II, III and IV, which verifies our assumption about small differences between levels II, III and IV. Our model works as well as the latest swin transformer in levels II, III and IV and slightly better in level III. The experimental results demonstrate that our model can extract the local feature information more effectively and fuse better with the self-attention mechanism, which effectively improves ultrasound classification of cervical lymph-node-level.

4.5. Visualization

Since deep-learning is a black box, it may be possible that the classification is correct but the model decision is wrong, making it difficult for imaging physicians to trust the classification results of the models, thus hindering the development of deep learning for medical application. Therefore, we also investigated the model interpretability methods, and we analyzed and selected three methods to apply to the decision interpretation of ultrasound image classification models.

We believe that CAM and Grad-CAM using gradient to obtain feature weights to interpret the classification decisions of ultrasound images is unreliable because the gradient exists noise and has saturation problem, while ultrasound is blurred, which leads CAM/Grad-CAM to focus on noisy regions. Score-CAM, XGrad-CAM, and Eigen-GradCAM are CAM-based improved methods. They are gradient-free, and have higher resolution and discrimination extent in ultrasound image classification models.

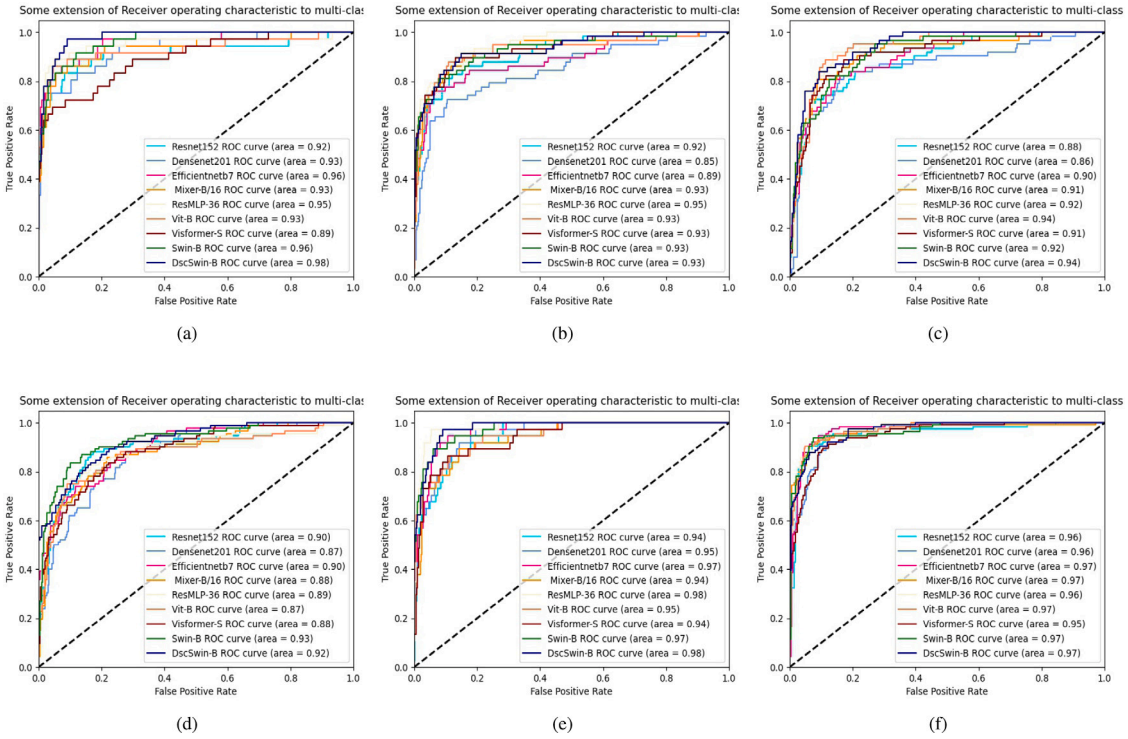

Fig. 6. ROC curves on each cervical lymph-node-level with the compared models and proposed model. (a)–(f) indicate the classification performances of different models through level I to VI respectively.

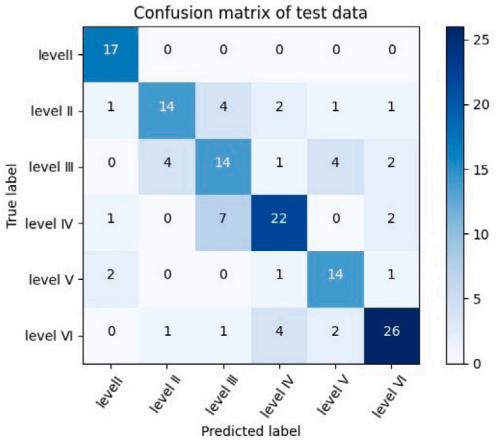

Fig. 7. The confusion matrix of test data.

We investigated the interpretability of the model for lymph-node-level classification. This will help physicians to understand the inference decisions of the deep learning model. We use three visualization methods to study the model decisions made with ultrasound images. As shown in Fig. 9, after training our model, randomly selected ultrasound images were input to the model, and then the prediction labels and corresponding heatmaps were obtained. It shows that the ultrasound data model of level I focuses on the mandibular region of the hyoid muscle, level V focuses on some muscle tissue, level VI focuses on the hypoechoic tracheal part. This verifies that our model is more consistent with imaging physicians' judgment in classification decisions, increasing the credibility of automated classification of cervical lymph-node-level, while bringing different diagnostic ideas to the physicians. For two misclassified data, we confirmed with physicians that the areas focused by the model did not match what they considered important, or those areas are too broad.

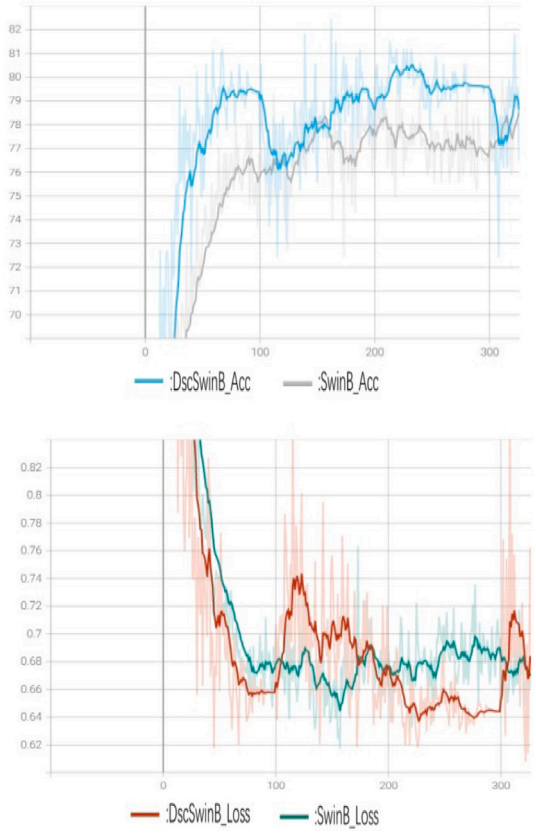

Fig. 8. Two line graphs represent the accuracy curve and loss curve of our model comparison with swin Transformer on the validation set of the first fold during training, in which y-axis represent accuracy score and loss score, respectively, and the x-axis represent the ith epoch.

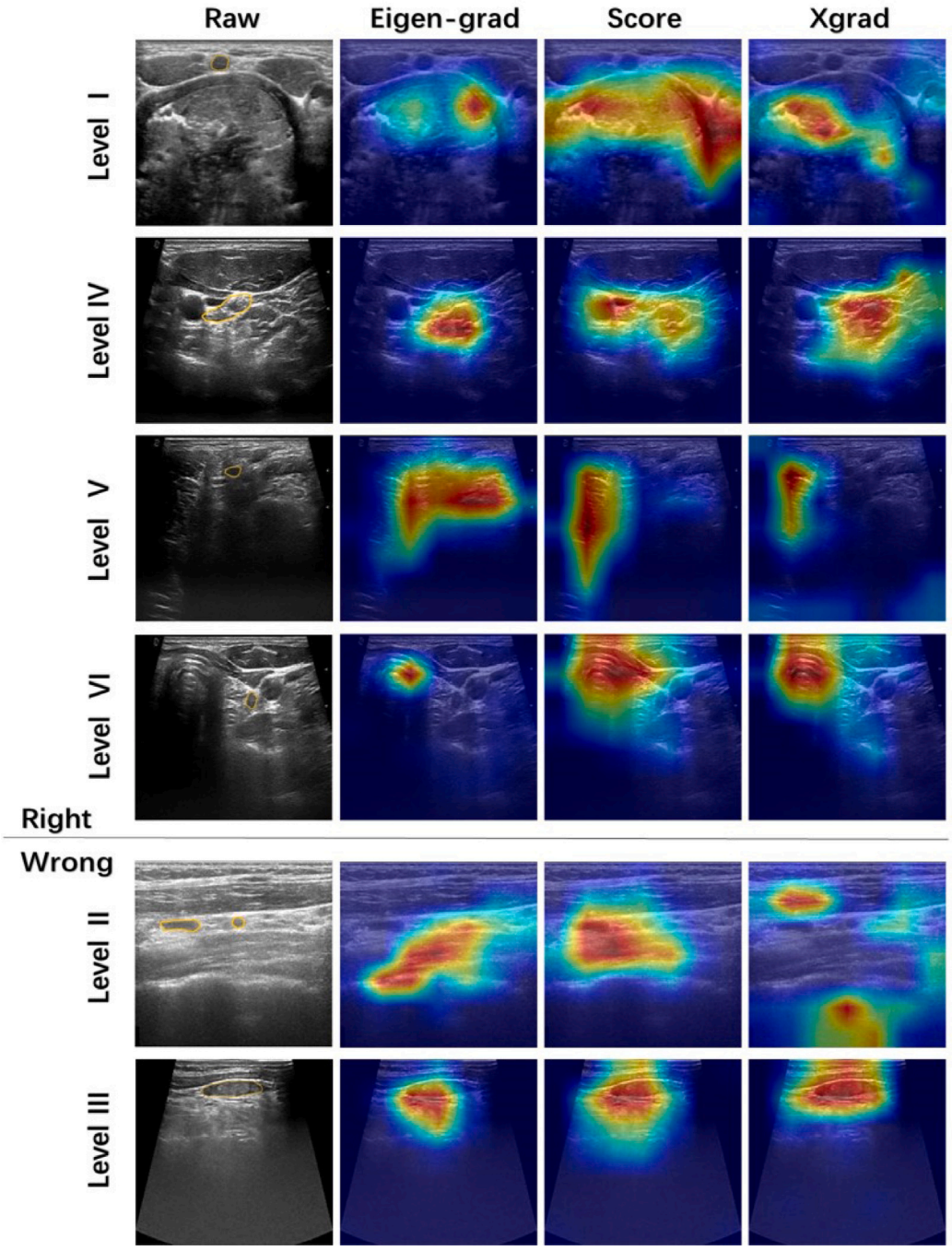

Fig. 9. Examples of heatmaps of each lymph node level. Row indicates levels I to VI. The first column shows ultrasound images, where yellow markers denote position of lymph nodes, audited by one radiologist. The last three columns show the heatmaps by the eigen-grad-cam, score-cam and xgrad-cam methods.

5. Discussions

5.1. Ablation study

In this study, ablation experiments are conducted, as shown in Table 3. It can be proved that the pre-processing work of ultrasound image, the CLAHE algorithm, the data augmentation methods and the removal of interface markers can effectively remove noise regions that are irrelevant to the classification and solve the ultrasound imaging problem. Second, FCCMix Loss is found to be more effective than Multiclass Cross-entropy Loss for improving data imbalance and reducing intra-class difference. In addition, the DSConvSwinBlock module is shown to be effective for extracting important local region information that is lost when using Swin Transformer.

6. Conclusion

In this study, we propose the Depthwise Separable Convolutional Swin Transformer to implement the task of cervical lymph-node-level classification. A data preprocessing method and a loss function are designed in our work, which helps to optimize performance of the model; and we use three visualization methods to explain the model's decisions. Through discussion with the physicians, we found that the

Table 3
Comparison of different obfuscations in terms of their transformation capabilities.

Pre-processing	Loss function		Mixup+CutMix	Block		ACC	PPV	TPR	F1
	BCE	FCCMix		Swin	DSConvSwin				
×	✓	×	×	✓	×	75.95	76.10	74.58	75.02
✓	✓	×	×	✓	×	76.65	77.68	74.82	75.70
✓	×	✓	×	✓	×	77.96	77.76	75.90	76.54
✓	×	✓	✓	✓	×	79.96	80.23	77.88	78.67
✓	×	✓	✓	×	✓	80.65	80.68	78.73	79.42

model did effectively focus on relevant lymph node peripheral tissues consistent with the regions physicians observed, as in the example results analyzed in Section 4.5. In a number of correctly classified and similar ultrasound images, the model focuses on the same regions, demonstrating that the model learns discriminative features. For lymph node level, the consistency of judgement results among physicians is low. Visualizing the decision process of the model and summarizing the specific patterns and imaging features learned by the model can provide complementary opinions as reference for physicians, not only to improve the accuracy and efficiency of lymph-node-level classification in ultrasound diagnosis, but also to provide effective information for lymph node metastatic diagnosis and subsequent debulking surgery.

The existing Vision Transformer builds the long-range relationship between global patches as a way to obtain global correlations. It is more capable of global feature extraction, but, at the same time, ignores detailed information from the local regions. In this study, a model fusing transformer and cnn is proposed, and the effectiveness of the model is fully demonstrated through our experiments. In the cervical lymph-node-level classification task, the Transformer model itself ignores semantic information from local regions, and the current tandem-type convolutional and transformer fusion networks or the pyramidal transformer variants are insufficient to compensate for the lack of local information extraction. So we enhanced the extraction ability of local feature information by adding a depthwise separable convolution branch to the self-attention mechanism and fusing information on different semantics and different scales into the self-attention mechanism which can effectively extract discriminative features to achieve the fine classification of six levels. In future, our trained model and the pre-processing algorithm can be merged into a unified architecture and be deployed online, which can generate lymph-node-level result and heat maps related to model decision in real time for application to ultrasound diagnostic report.

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.

Acknowledgements

This work was supported by the Shanghai Foundation for Development of Science and Technology, China (No. 21142202400). The authors gratefully acknowledge this support. This work was performed in cooperation with the Ultrasound Medicine and Nuclear Medicine Departments of the Sixth People's Hospital, and approved by the Ethics Committee of Shanghai Sixth People's Hospital.(Approval code: SH6H-2021-YS-218).

References

[1] B.R. Haugen, et al., 2015 American Thyroid Association management guidelines for adult patients with thyroid nodules and differentiated thyroid cancer: the American Thyroid Association guidelines task force on thyroid nodules and differentiated thyroid cancer, J. Thyroid 26 (1) (2016) 1–133.

[2] V. Grégoire, et al., Delineation of the neck node levels for head and neck tumors: A 2013 update. DAHANCA, EORTC, HKNPCSG, NCIC CTG, NCRI, RTOG, TROG consensus guidelines, J. Radiother. Oncol. 110 (1) (2014) 172–181.

[3] M.I. Daoud, A.A. Atallah, F. Awwad, et al., Automatic superpixel-based segmentation method for breast ultrasound images, J. Expert Syst. Appl. 121 (2019) 78–96.

[4] C.K. Zhao, et al., A comparative analysis of two machine learning-based diagnostic patterns with thyroid imaging reporting and data system for thyroid nodules: diagnostic performance and unnecessary biopsy rate, J. Thyroid 31 (3) (2021) 470–481.

[5] Y. Kan, et al., Radiomic signature as a predictive factor for lymph node metastasis in early-stage cervical cancer, J. Magn. Reson. Imaging 49 (1) (2019) 304–310.

[6] M. Cordes, et al., Advanced thyroid carcinomas: neural network analysis of ultrasonographic characteristics, J. Thyroid Res. 31 (3) (2021) 470–481.

[7] J.H. Lee, et al., Deep learning–based computer-aided diagnosis system for localization and diagnosis of metastatic lymph nodes on ultrasound: A pilot study, J. Thyroid 28 (10) (2018) 1332–1338.

[8] R. Song, et al., Thyroid nodule ultrasound image classification through hybrid feature cropping network, J. IEEE Access 8 (2020) 64064–64074.

[9] H. Xie, et al., Cross-attention multi-branch network for fundus diseases classification using SLO images, J. Med. Image Anal. 71 (2021) 102031.

[10] Y. LeCun, LeNet-5, convolutional neural networks, 20 (5) (2015) 14. http://yannlecun.com/exdb/lenet.

[11] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, J. Adv. Neural Inform. Process. Syst. 25 (2012) 1097–1105.

[12] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.

[13] C. Szegedy, et al., Going deeper with convolutions, in: C. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1-9.

[14] K. He, et al., Deep residual learning for image recognition, in: C. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770-778.

[15] G. Huang, et al., Densely connected convolutional networks, in: C. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700-4708.

[16] J. Hu, et al., Squeeze-and-excitation networks, in: C. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132-7141.

[17] T. Lin, et al., A survey of transformers, 2021, arXiv preprint arXiv:2106.04554.

[18] A. Dosovitskiy, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.

[19] H. Fan, et al., Multiscale vision transformers, 2021, arXiv preprint arXiv:2104.11227.

[20] Z. Chen, et al., Visformer: The vision-friendly transformer, 2021, arXiv preprint arXiv:2104.12533.

[21] W. Wang, et al., Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, 2021, arXiv preprint arXiv:2102.12122.

[22] W. Wang, et al., Pvtv2: Improved baselines with pyramid vision transformer, 2021, arXiv preprint arXiv:2106.13797.

[23] H. Touvron, et al., Training data-efficient image transformers & distillation through attention, in: C. International Conference on Machine Learning, PMLR, 2021, pp. 10347–10357.

[24] Z. Liu, et al., Swin transformer: Hierarchical vision transformer using shifted windows, 2021, arXiv preprint arXiv:2103.14030.

[25] Z. Huang, et al., Shuffle transformer: Rethinking spatial shuffle for vision transformer, 2021, arXiv preprint arXiv:2106.03650.

[26] X. Chu, et al., Twins: Revisiting the design of spatial attention in vision transformers, 1 (2) (2021) 3. arXiv preprint arXiv:2104.13840.

[27] X. Dong, et al., Cswin transformer: A general vision transformer backbone with cross-shaped windows, 2021, arXiv preprint arXiv:2107.00652.

[28] J. Yang, C. Li, P. Zhang, X. Dai, B. Xiao, L. Yuan, J. Gao, Focal self-attention for local–global interactions in vision transformers, 2021, arXiv preprint arXiv:2107.00641.

[29] Y. Li, K. Zhang, J. Cao, R. Timofte, L.V. Gool, Localvit: Bringing locality to vision transformers, 2021, arXiv preprint arXiv:2104.05707.

[30] H. Wu, et al., Cvt: Introducing convolutions to vision transformers, 2021, arXiv preprint arXiv:2103.15808.
[31] Z. Dai, H. Liu, Q.V. Le, M. Tan, CoAtNet: Marrying convolution and attention for all data sizes, 2021, arXiv preprint arXiv:2106.04803.
[32] Class-balanced loss based on effective number of samples, in: C. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9268-9277.
[33] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: C. International Conference on Machine Learning, PMLR, 2019, pp. 6105–6114.
[34] I.O. Tolstikhin, et al., Mlp-mixer: An all-mlp architecture for vision, J. Adv. Neural Inform. Process. Syst. 34 (2021).
[35] H. Touvron, et al., Resmlp: Feedforward networks for image classification with data-efficient training, 2021, arXiv preprint arXiv:2105.03404.