IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

MSTD: A Multi-scale Transformer-based Method to Diagnose Benign and Malignant Lung Nodules

Xiaoyu Zhao[1,2,†], Jiao Li[1,†], Man Qi[1,†], Xuxin Chen[1], Wei Chen[1], Yongqun Li[1], Qi Liu[1], Jiajia Tang[1], Zhihai Han[1,2,*], and Chunyang Zhang[1,*]

[1]Department of Pulmonary and Critical Care Medicine, The Sixth Medical Center of Chinese PLA General Hospital, Beijing, 100048, China (email:419840635@qq.com)
[2]Graduate School of Hebei North University, Zhangjiakou,075000, Hebei, China

*Corresponding author: Chunyang Zhang and Zhihai Han (e-mail: zhcy101@126.com, hanzhihai@301hospital.com.cn).

† These authors contributed equally to this work and should be considered co-first authors.

ABSTRACT The identification of benign and malignant lung nodules is crucial for timely treatment to reduce the risk of the progression and metastasis of diseases. However, the varied sizes, diverse morphologies, non-fixed positions, and dynamic growth of lung nodules in computed tomography (CT) images make their accurate identification challenging. To address these issues, we propose a multi-scale transformer-based diagnosis (MSTD) method for benign and malignant lung nodules. To handle significant variations in the shapes and sizes of the lung nodules, we first design a multi-scale module based on parallel branches to extract multi-scale features. To make full use of these features, we then introduce a multi-scale transformer fusion (MSTF) module to integrate the information obtained at different scales. Unlike conventional vision transformers, our MSTF can simultaneously extract attention-based features from the spatial dimensions at different scales to enhance the accuracy of classification of lung nodules. We conducted extensive ablation experiments on multi-scale structures and transformer-based methods of fusion to explore the impact of features obtained at different scales on the accuracy of classification of lung nodules. The results of verification on the LUNA16 dataset showed that the average F1Score, Specificity, and Sensitivity of the proposed MSTD exceeded 90% (94.5%, 96.5%, and 91.1%, respectively), where this shows that it can accurately identify both benign and malignant lung nodules. Its average performance was superior to the state-of-the-art method by about 1%, 3.4%, and 3.6% in terms of the area under the curve (AUC), Accuracy, and F1Score, respectively.

INDEX TERMS MSTD, Multi-scale network, Multi-scale transformer fusion, Lung nodule diagnosis

I. INTRODUCTION

LUNG disease is one of the most common types of diseases worldwide, and lung cancer in particular poses a significant threat to human health and life [1]. Lung nodules are small masses or lesions in the lungs, with a diameter of less than 3 cm, that serve as early indicators of lung cancer [2]. Radiologists can diagnose the malignancy of nodules based on their size, density, edge-related characteristics, and shape as depicted in computed tomography (CT) images, and can then provide recommendations for treatment that are tailored to the patient. However, the manual diagnosis of lung disease is time consuming, costly, and susceptible to biases.

Deep learning algorithms can learn rich information from large volumes of data to significantly improve the efficiency of screening for diseases and minimize the rate of misdiagnosis [3]. With rapid advances in the relevant technologies in recent years, an increasing number of researchers have used deep learning models for identifying lung nodules [4]. The convolutional neural network (CNN) is sensitive to visual information, and has been widely used for tasks of recognition in medical imaging [5]–[8].

However, lung nodules can significantly vary in terms of their shapes and sizes, and this makes it difficult to distinguish between them and the surrounding tissues. This poses a daunting challenge in accurately identifying lung nodules, especially small or irregularly shaped nodules, in CT images [9]. The conventional CNN is not robust to variations in the scale of the input data [10].

To address the above issue, we design a multi-scale module based on multiple parallel branches (MSMPB) in this study to extract features from images of lung nodules. To this end, we design a network consisting of multiple, parallel branches to

extract features at different scales. Every branch at each scale is assigned different receptive fields through such operations as convolution and pooling. Such a multi-scale convolutional network can accurately capture information on the spatial structure of the input images of lung nodules by using receptive fields at different levels. This helps the network more accurately identify the different types of nodules appearing in CT images of the lung.

Once multi-scale features at different scales have been extracted by the MSMPB module, they need to be effectively fused. Scenarios may arise during feature fusion in which useful information is lost and redundant information is retained [11]. Effectively retaining and fusing key information obtained at various scales is thus an important issue in this context.

The vision transformer (ViT) [12] uses a self-attention mechanism to capture the global dependencies between different parts of a given image. This mechanism is useful for handling multi-scale features as it can establish connections between features at different scales to enhance our understanding of the overall structure of the image. Several recent studies have considered methods of image processing based on multi-scale strategies and the ViT [13]–[15]. Chen et al. [16] proposed a dual-branch ViT to integrate patches of images at different scales to generate more robust features. Shao et al. [17] replaced the feed-forward network in the ViT encoder with a mixed convolutional feed-forward module that enhanced the network's ability to capture local and multi-scale features. However, while most currently available methods for fusing multi-scale features are based on an analysis of their spatial dimension, few methods can simultaneously analyze features along the dimensions of both space and scale.

To address the above issues, we design a multi-scale transformer fusion (MSTF) module based on the ViT [12] to analyze and fuse features from branches at multiple scales. The ViT can capture the relevant information from images through a self-attention mechanism, where this makes it robust to changes in the size and dimensions of the input feature maps. Traditional ViT-based methods partition the spatial dimensions of images into smaller patches. However, our MSTF module performs patch partitioning and multi-head self-attention-based computations on a plane composed of the spatial and scale-related dimensions, which makes it more sensitive to changes in the scale of images. Therefore, our MSTF module can simultaneously extract critical scale and spatial information from multi-scale features and fuse them.

The fused features are then fed to a classifier for the binary classification of benign and malignant nodules. As the MSMPB and MSTF modules can extract sufficiently rich features from images of lung nodules, we directly use a fully connected network [18] to perform binary classification.

Finally, we integrate the MSMPB module, MSTF module, and classifier to construct a multi-scale transformer-based diagnosis (MSTD) method for the classification of benign and malignant lung nodules. We performed experiments on the Lung Nodule Analysis 2016 (LUNA16) dataset [19] to verify the performance of the proposed MSTD.

The LUNA16 [19] dataset contains over 1,000 CT scan images, each with detailed annotations of the lung nodules. The images of lung nodules in this dataset have diverse shapes, sizes, and densities, and thus require highly flexibility, adaptable, and automated diagnostic algorithms. The CT scan images in this dataset vary significantly in terms of their resolution and level of noise. This inconsistency poses a challenge to the robustness of the diagnostic algorithms. The environment of the lungs depicted in CT images is complex, and contains blood vessels, airways, and ribs in addition to lung nodules. These structures may exhibit morphological and density-related similarities with lung nodules to further complicate the algorithms used to distinguish between benign and malignant nodules.

Our proposed MSTD performed well on the LUNA16 dataset, with its average scores of the F1Score, Specificity, and Sensitivity all surpassing 90% (94.5%, 96.5%, and 91.1%, respectively). Its average diagnostic performance was superior to that of state-of-the-art methods by approximately 1% in terms of the area under the curve (AUC), 3.4% in terms of Accuracy, and 3.6% in terms of the F1Score. This shows that the proposed MSTD can accurately identify lung nodules of various sizes, shapes, and densities in clinical environments, and can adapt to images of different resolutions and levels of noise.

In summary, the main contributions of this study are as follows:

- We propose the MSTD for the precise identification of benign and malignant lung nodules. This method combines the advantages of multi-scale structures and scale-based attention mechanisms. Our experiments confirmed the superior performance of the MSTD in comparison with state-of-the-art methods.
- We use the MSMPB module to extract multi-scale information on lung nodules and their contextual background. The module is composed of multiple, parallel branches for feature extraction, each of which represents a scale-based information extractor.
- We design the MSTF module based on the transformer to fuse and filter information from features obtained at multiple scales. It not only performs self-attention operations in the spatial dimension of the image, but also executes them in the scale-based dimension.

The remainder of this article is organized as follows: Section II introduces recent research on medical imaging and the identification of lung nodules, while Section III provides a detailed description of the proposed MSTD, including the MSMPB module, MSTF module, and classifier. We report a series of ablation and comparative experiments in Section IV to verify the performance of our method, and summarize the conclusions of this study in Section V.

II. RELATED WORKS

A. THE ADVANCED METHODS IN THE FIELD OF MEDICAL IMAGE PROCESSING

With advances in artificial intelligence technology, deep learning has been used in tasks of medical image processing, including image diagnosis [20]–[22], pathological analysis [23], and image segmentation [24]–[27]. Deep learning be used to automatically learn and extract features from a large volume of data on medical images, which enables their automated processing and analysis to enhance the efficiency of medical diagnosis and treatment.

Guo et al. [28] proposed a composite network called RK-net that combines deep learning with an unsupervised K-means clustering algorithm to automatically process medical images. RK-net is more efficient than manual screening and annotation in refining such images. Zakareya et al. [29] proposed a deep learning model based on GoogLeNet and residual blocks to classify images obtained from patients of breast cancer. It uses granular computing, shortcut connections, and two learnable activation functions in place of traditional activation functions to improve the accuracy of diagnosis and reduce the workload of doctors. Granular computing can enhance diagnostic accuracy by capturing granular information from images of cancerous sites. Eltoukhy et al. [30] proposed a self-learning method that uses deep neural networks and residual learning to circumvent the requirement of a large number of labeled images to train deep learning models for classifying histopathological images of breast cancer tissues. This method provides a second opinion for radiologists using medical images.

Cheung et al. [31] proposed a data-centric deep learning technique with big interpolated data, Interpolation-Split, to enhance the performance of airway tree segmentation.The method uses an ensemble learning strategy to aggregate the airway segments obtained at different scales. This approach has low requirements related to RAM/GPU usage and can be deployed on most 2D deep learning models. Cheung et al. [32] also explored the quantification of airway metrics and their impact and correlation with mortality in idiopathic pulmonary fibrosis (IPF). They found that the segmental inter-subsegmental tapering and segmental tortuosity measurements generated by airway measurement algorithm (AirQuant) were independently associated with mortality in IPF patients. Vijayakumar et al. [33] employed advanced segmentation techniques for lung cancer detection. The preprocessed data was segmented into different groups using UNet segmentation, and the segmented images were used in a Capsule Neural Network (CapsNet) to determine the exact condition of the original images.

During the training process, ViT leverages the self-attention mechanism to better handle noise and inconsistencies in images, thereby enhancing the robustness of the model. As a result, in recent years, ViT-based methods for medical image processing have become increasingly popular [34]–[37]. Chen et al. [38] proposed Mixblock, a hybrid encoder that effectively combines the strengths of CNNs and ViT to extract multidimensional high-level semantic segmentation information from images, moving beyond mere local and global spatial features representation. The method also innovatively incorporates frequency domain information into skip connections to eliminate semantic ambiguity between the encoder and decoder. Wang et al. [39] proposed a Single Encoder-Dual Decoder architecture called DBUNet, which integrates the ViT encoder framework. The ViT encoder is utilized as part of the decoder branches to enhance shallow features. A polarization amplification method for channel weights is used before the ViT encoder modules to optimize image segmentation. Fan et al. [40] introduced a method called ViT with Feature Recombination and Feature Distillation (ViT-FRD), which combines ViT and CNN through knowledge distillation to enhance the performance of cardiac structure segmentation in MRI images. The training process allows the student model (i.e., ViT) to learn from the teacher model (i.e., CNN) by optimizing the distillation loss.

Huo et al. [41] proposed a three-branch hierarchical multi-scale feature fusion network structure called HiFuse, which integrates the advantages of Transformers and CNNs at multiple scales, thereby enhancing the classification accuracy of various medical images. This approach introduces a parallel structure for local and global feature blocks to effectively extract local features and global representations across various semantic scales. Liu et al. [42] proposed an efficient medical image classification network called Eff-CTNet, based on an alternating hybrid series connection of CNN and Transformer. The method includes a Group Cascade Attention (GCA) module, which divides feature maps into different attention heads to further enhance attention diversity and reduce computational costs.

B. THE ADVANCED METHODS FOR LUNG NODULE DIAGNOSIS

With the extensive use of deep learning technology in medical image processing, a large number of methods have been developed for identifying lung nodules in images.

Amrita et al. [43] improved the accuracy of the lung nodule classification method by using Fractalnet architecture. This fractal structure is considered an effective alternative to residual structure. Fractalnet consists of 5 concatenated Fractal Blocks, each Fractal Block containing a multi-level convolution structure with a depth of 8. Halder et al. [44] proposed an Atrous Convolution-based Convolutional Neural Network (ATCNN) framework capable of segmenting and characterizing lung nodules by capturing multi-scale features from CT images. The method analyzed different variants of the ATCNN framework, among which the framework with a two-layer atrous pyramid and residual connections demonstrated the highest nodule characterization performance indices. Huang et al. [45] developed a self-supervised transfer learning based on domain adaptation (SSTL-DA) 3D convolutional neural network framework for the benign and malignant detection and classification of lung nodules. In the classification module, a series of 12 convolutional layers were directly adopted for feature extraction of lung nodule

images, followed by classification operations using three fully connected layers. Xu et al. [46] proposed a lung nodule classification method based on attribute privilege and capsule networks. The eight attribute features of lung nodules are used to enhance the discrimination ability between benign and malignant cases. The capsule structure helps to extract and understand the spatial relationships between different parts of lung nodule images.

Miao et al. [47] proposed a Transformer-based model for Ground-Glass Nodules (GGN) recognition. A 3D CNN is used as the backbone to automatically extract features of the 3D CT image of lung nodules. The positional encoding information is added to the extracted feature maps and inputted into the Transformer encoder layers to obtain higher-order asymmetric feature representations. These asymmetric features are inputted into a support vector machine, effectively improving the recognition accuracy of GGN. Kanipriya et al. [48] proposed a lung nodule abnormality classification algorithm based on CNN and Long Short Term Memory (LSTM). The lung nodules were located using a K-means clustering method and segmented using an automated active contour level set. An Improved Capuchin Search Algorithm (ICSA) optimized CNN and LSTM hybrid network was used to classify lung nodules into categories. Shi et al. [49] proposed a Semi-supervised Deep Transfer Learning (SDTL) framework for benign and malignant lung nodule diagnosis. They adopt a transfer learning strategy to distinguish lung nodules from pseudo-nodular tissue. Furthermore, they introduced a semi-supervised approach based on iterative feature matching to address the issue of limited samples with pathological features. Zhang et al. [50] proposed a 3D Feature Pyramid Network (FPN) for lung nodule detection, which solved the problem of small nodules that cannot be well detected in CT images. They also incorporated the Squeeze-and-Excitation (SE) attention module to enhance detection performance.

Unlike all the other methods mentioned above, our MSTD method combines the multi-scale strategy with space and scale Transformer, enhancing the network's comprehensive utilization capacity of nodule images at different scales.

III. METHODOLOGY

The proposed MSTD includes the MSMPB module, MSTF module, and classifier. Fig. 1 shows an overview of its workflow.

The MSMPB module is used to obtain multi-scale features from raw images of lung nodules. It is composed of multiple, parallel branches, each of which is responsible for extracting features at a specific scale. The features are extracted from branches at each scale through convolutional layers, and the scale-related transformation is accomplished through pooling operations.

The MSTF module is utilized to merge and process the Multi-Scale Features extracted by the MSMPB module. This module consists of a Space Transformer and a Scale Transformer. The Space Transformer is responsible for extracting crucial spatial information, while the Scale Transformer is used to extract essential scale information. The results of these two attention mechanisms are added together to form Space-Scale Features that simultaneously incorporate spatial and scale information.

The classifier module categorizes the spatial–scale-related features as either benign or malignant. It directly uses fully connected (FC) layers to compute the logit vector and applies the softmax function to obtain the probability distributions of the two categories.

A. MSMPB MODULE

Multi-scale feature extraction network structures can effectively handle variations in scale and spatial changes in images, thereby enhancing the robustness of the model [51]. In the context of lung nodule recognition, the nodule targets in input images may exhibit different scales and spatial shapes. Conventional convolutional networks are unable to adapt to such variations, due to their fixed receptive fields and kernel sizes.

To address the limitations of conventional convolutional networks when dealing with lung nodules of different shapes and sizes, we develop the MSMPB module to improve the adaptability of the network to multi-scale scenarios as illustrated in Fig.2. Our MSMPB module consists of multiple, parallel branches at different scales, where the branch at each scale comprises three convolutional layers and a pooling layer. Each convolutional layer is followed by a BN layer and an ReLU activation function. The magnitude of the operator of the pooling layer determines the receptive field of the convolution. We thus generate features at different scales by varying the size of this operator for branches at each scale.

The number of parallel branches is set to 6, and the number of convolutional kernels in all convolution layers is set to 8. After concatenating the feature maps from all scales along the channel dimension, the total number of channels is 48. Thus, we use the 2D linear interpolation layer, Interpolate, to uniformly scale the feature maps to a size of 48×48, creating Multi-Scale Features containing rich scale information. At this point, the number of dimensions in the scale, width, and height of Multi-Scale Features is exactly the same, which facilitates to perform the feature fusion across the scale and spatial dimensions in the subsequent MSTF module.

B. MSTF MODULE

The Multi-Scale Features obtained by the MSMPB module contain rich scale information, and it is important to fully utilize this information. During the fusion process, there may be redundancy of useless information and loss of valuable information [52]. The self-attention mechanism in ViT [12] enables automatic adjustment of attention weights based on the input image content, adapting to various image contents and task requirements. This dynamic adaptability allows ViT to perform exceptionally well in handling complex image scenes and accommodate images of different sizes and content [53].

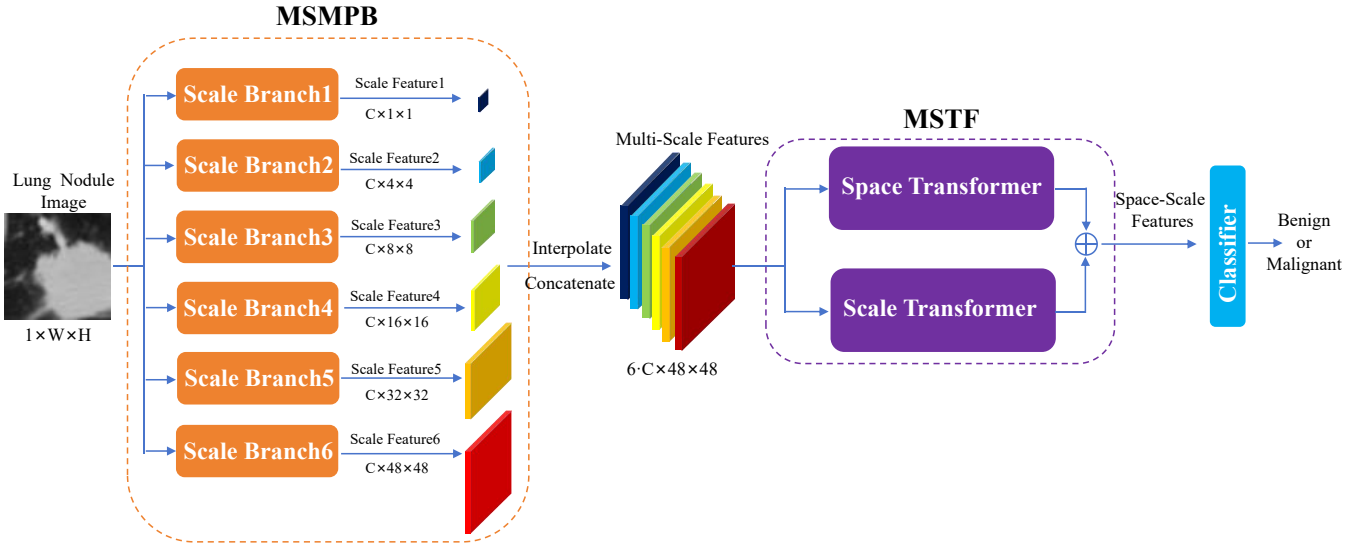


FIGURE 1. **The overview of the MSTD method. C is the number of channels for a scale feature, and C is set to 8.**

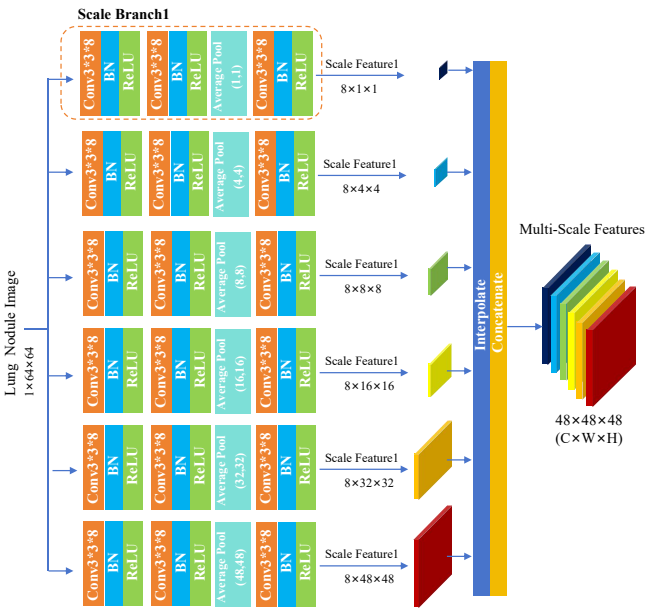


FIGURE 2. **The MSMPB module. The Conv3*3*8 represents a convolutional layer with 8 convolutional kernels, a kernel size of 3x3, and a step size of 1. The Average pool represents an Adaptive average pooling layer.**

To extract sensitive information at multiple scales, we use the MSTF module for feature fusion as illustrated in Fig. 4. This module consists of two parts: a Space Transformer for computing crucial spatial features of the input data, and a Scale Transformer for calculating their important scale-related features.

In the Space Transformer branch, the Multi Scale Features will be non overlapping decomposed into patches of size $C \times P_w \times P_h$ in the spatial dimension, where C represents the channel dimension (i.e. scale dimension), P_w and P_h respectively represent the width and height of each patch. The values of P_w and P_h are set to 4.

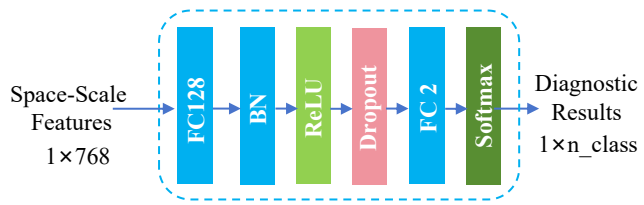


FIGURE 3. **The detailed structure display of classifier. The BN represents the Batch Normalization layer.**

Then, these patches will be embedded to obtain Patch Vectors, as shown in Equation (1).

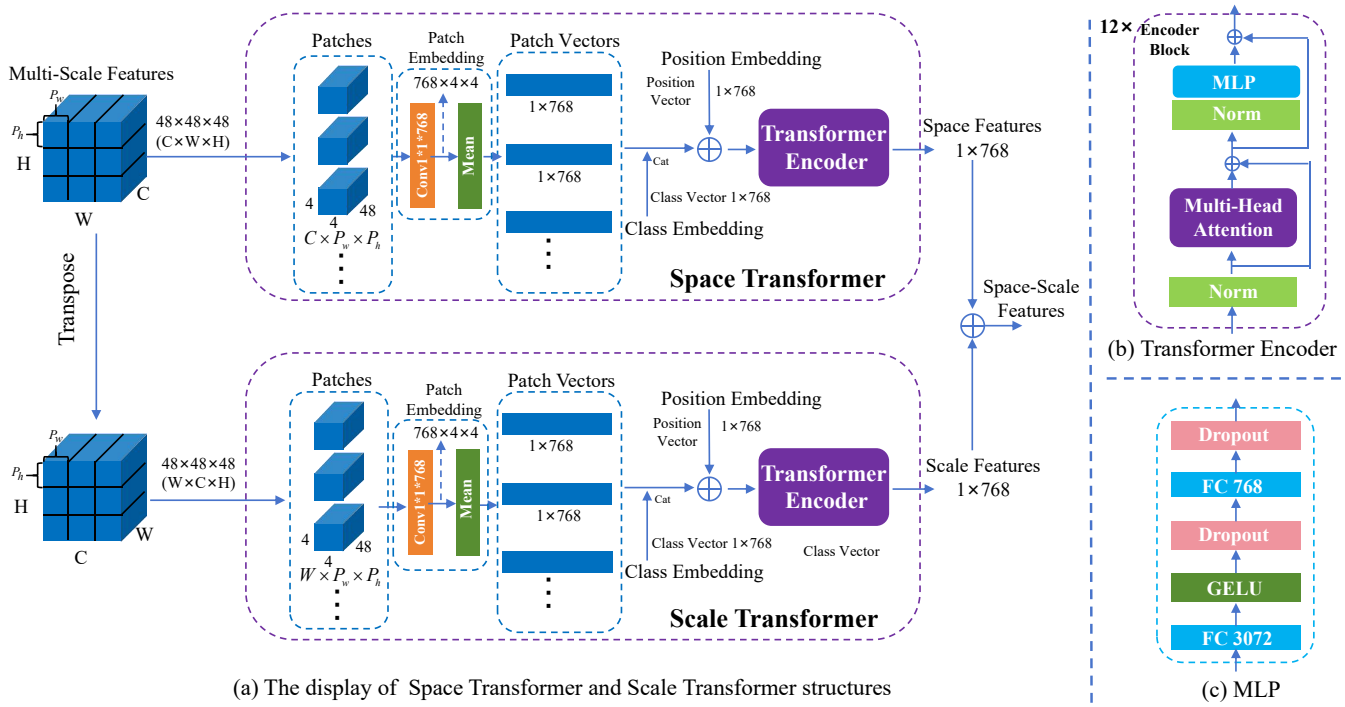
$$\begin{aligned} v_p &= \text{PatchEmbedding}(x_p), x_p \in \mathbb{R}^{C \times P_w \times P_h} \\ &= Mean(Conv2d(x_p)), v_p \in \mathbb{R}^{1 \times dim_{vit}} \end{aligned} \tag{1}$$

where x_p is a patch, v_p is a Patch Vector. $Conv2d(\cdot)$ represents a 2D convolution, the kernel size is 1×1 and the number of kernels is $dim_{vit} = 768$, dim_{vit} denotes the vector length of the Transformer Encoder. $Mean(\cdot)$ represents the operation of taking the average in the spatial dimension. These patches expand the channel dimension to 768 through 2D convolution, resulting in $768 \times 4 \times 4$ features. Then, the average value is taken across the spatial dimensions, yielding a Patch Vector of 1×768. At this point, each patch of the original image is embedded into a Patch Vector, which is a form suitable for processing by the Transformer Encoder.

For classification tasks, ViT usually creates an updated Class Vector that will be concatenated with all Patch Vectors, as shown in Equation (2).

$$\begin{aligned} V_{p,cls} &= \text{ClassEmbedding}(V_p) \\ &= Cat(V_p, v_{cls}), v_{cls} \in \mathbb{R}^{1 \times dim_{vit}} \end{aligned} \tag{2}$$

where $V_{p,cls}$ is Patch Vectors that incorporates the Class Vector, $V_{p,cls} \in \mathbb{R}^{(N_p+1) \times dim_{vit}}$. v_{cls} denotes the Class Vector, which is randomly initialized and can be updated during



(a) The display of Space Transformer and Scale Transformer structures

FIGURE 4. The detailed structure of MSTF module. The Transformer Encoder consists of 12 Encoder Blocks. Norm represents the Layer Normalization function, GELU represents the Gaussian Error Linear Units activation function, the value of Dropout is 0.5, Mean represents taking the mean in the spatial dimension, Conv represents convolutional layer, FC represents the Fully connected layer, Cat represents feature concatenation operation. P_h and P_w represent the height and width of each cropped patch, respectively. C, W, and H respectively represent the scale, width, and height dimensions of the Multi-Scale Features.

training. V_p represents the Patch Vectors of all patches, $V_p \in \mathbb{R}^{N_p \times dim_{vit}}$, $N_p = \lfloor W/P_w \rfloor \cdot \lfloor H/P_h \rfloor$, $\lfloor \cdot \rfloor$ denotes the downward rounding, the values of C, W, and H are all 48. $Cat(\cdot)$ denotes feature concatenation operation.

ViT is not sensitive to the position of patches, so it is necessary to add the position information to Patch Vector, as shown in Equation (3)

$$\begin{aligned} V_{p,cls,pos} &= \text{PositionEmbedding}(V_{p,cls}) \\ &= V_{p,cls} + v_{pos}, v_{pos} \in \mathbb{R}^{1 \times dim_{vit}} \end{aligned} \tag{3}$$

where $V_{p,cls,pos} \in \mathbb{R}^{(N_p+1) \times dim_{vit}}$, represents the $V_{p,cls}$ with position information added. v_{pos} stands for the Position Vector, which is randomly initialized and is updated along with the network's parameters during training.

Then, $V_{p,cls,pos}$ will go through the Transformer Encoder to perform the self-attention mechanism, resulting in Space Features. The Transformer Encoder [12] primarily consists of LayerNorm, Multi-Head Attention, and MLP .

In the Scale Transformer branch, unlike the Space Transformer, the dimension of Multi-Scale Features will undergo a transpose operation, transforming it into the shape $W \times C \times H$. Since Multi-Scale Features are specifically designed as a cube shape, the transposed features can still undergo the same processing steps as in the Space Transformer branch. In the multi-scale dimension ($C \times H$), Multi-Scale Features are decomposed into patches of size $W \times P_w \times P_h$. At this point, the spatial dimension ($P_w \times P_h$) of the patches already contains both spatial and scale information of the input features. After being embedded into Patch Vectors, they are integrated with the Position Vector and Class Vector. Subsequently, these patch-related features are fed into the Transformer Encoder for scale dimension feature extraction, to obtain the Scale Features.

Finally, Space Features and Scale Features will be added together to obtain Space-Scale Features, which will be fed into the classifier module for benign and malignant classification.

It is worth noting that, unlike traditional ViT classification models, our MSTF module not only performs attention mechanisms in the spatial dimension of images but also creatively applies attention weighting in the multi-scale dimension. This has significantly improved the performance of the entire MSTD method.

C. CLASSIFIER MODULE

After processing lung nodule images through the MSMPB module and MSTF module, we obtain Space-Scale Features that contain critical spatial and scale information. At this point, the two-dimensional image information has been compressed into a one-dimensional vector (1×768) after passing through the Transformer Encoder. Therefore, the classifier can directly use fully connected layers, which are sensitive to one-dimensional features, to accomplish the diagnostic task of distinguishing between benign and malignant lung nodules.

The detailed architecture of the classifier, as shown in

Fig.3, consists of two fully connected layers. Since the MSMPB and MSTF modules are connected in series, the network depth at this point is already sufficiently deep, and the Space-Scale Features already contain key and salient information relevant to the categories. Therefore, it is unnecessary for the classifier to significantly increase the number of network layers. We only require a limited number of fully connected layers to transform Space-Scale Features into category vectors.

The numbers of neurons in the first and second fully connected layers were set to 128 and two, respectively. The output of the fully connected layers was a logit vector that needed to be transformed into a probability distribution. We fed the logit vector to the softmax function to obtain the probability distributions of the given samples belonging to either of the two categories of benign and malignant lung nodules. Finally, the category with the higher probability was identified as the diagnostic result.

D. LOSS FUNCTION

The cross-entropy loss function has suitable mathematical properties that can be used to accelerate optimization and reduce the difficulty of training [54]. It is also sensitive to the probability distribution of the output of the model, such that it can accurately reflect the capability of the model to predict the category to which a given sample belongs [55].

Benign cases often outnumber malignant cases in datasets of images of lung nodules, which leads to a highly unbalanced distribution of samples in the two categories. We added weights to different categories when calculating the loss function to prevent the decisions of the model from favoring the category with a larger number of samples. This is illustrated in Equation (4):

$$\begin{aligned} L &= w_0 \cdot L_0 + w_1 \cdot L_1 \\ &= \sum_{i=1}^{N} \{ -w_0 \cdot [y_i log\hat{y}_i + (1-\hat{y}_i) log(1-\hat{y}_i)]_{y_i=0} \\ &\quad - w_1 \cdot [y_i log\hat{y}_i + (1-y_i) log(1-\hat{y}_i)]_{y_i=1} \} \\ &= \sum_{i=1}^{N} \{ -w_0 \cdot [log(1-\hat{y_i})]_{y_i=0} - w_1 \cdot [log\hat{y}_i]_{y_i=1} \} \end{aligned} \tag{4}$$

where L_0 is the L_0 is the loss of benign samples, L_1 is the loss of malignant samples, L is the overall loss of the two types of samples. w_0 is the weight of the benign category, and w_1 is the weight of the malignant category, and $w_0 = 1/n_0$, $w_1 = 1/n_1$, n_0 is the number of benign samples and n_1 is the number of malignant samples. That is, w_0 and w_1 are inversely proportional to the number of samples of the categories they represent. In this way, categories with fewer instances receive higher weights, while categories with more instances receive lower weights, thus balancing the model's attention to different categories.

E. EVALUATING INDICATORS

Several metrics are commonly used to assess the performance of methods of medical diagnostics, as shown in Equation (5-12).

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{5}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{6}$$

$$Specificity = \frac{TN}{TN + FP} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{10}$$

$$FPR = \frac{FP}{FP + TN} \tag{11}$$

$$AUC = Area\ Under\ the\ Curve\ of\ ROC \tag{12}$$

where where TN (true negative), TP (true positive), FN (false negative), and FP (false positive) represent the numbers of samples that are correctly predicted as negative by the model, correctly predicted as positive by it, incorrectly predicted as negative when they are actually positive, and incorrectly predicted as positive when they are actually negative. ROC represents Receiver Operating Characteristic. We obtain the ROC curve by calculating True Positive Rate (TPR, equals Recall) and False Positive Rate (FPR) at different classification thresholds and then plotting these points on a two-dimensional plane. The x-axis represents FPR, and the y-axis represents Recall. The closer the ROC curve is to the upper left corner, the better the model's performance. AUC stands for the Area Under the ROC Curve. The range of AUC values is between 0 and 1. The closer the AUC value is to 1, the better the model's classification performance.

Among the metrics used for evaluation, AUC, Accuracy, and F1Score are comprehensive indicators that can be independently compared numerically to assess classification performance. However, Specificity specifically measures the ability to correctly identify benign lung nodules, while Sensitivity measures the ability to correctly identify malignant lung nodules. These two metrics can be influenced by the model's bias towards a particular class. Therefore, only when both Specificity and Sensitivity are simultaneously high can it be concluded that the model's overall classification performance is good.

IV. EXPERIMENTS AND DISCUSSIONS

To ensure the fairness of the experiments, we conducted all of them on the same software and hardware platforms while maintaining the consistency of the hyperparameters of the model. The values of the hyperparameters and specifications of the computational platform are provided in Table.1. All

experimental results are presented as the mean ± standard deviation.

TABLE 1. The hyperparameters of the model and parameters of the computational platform

Hyperparameter	Value
Epoch	>200
Batch size	32
Optimizer	Adam
Learning rate	10^{-5}
Number of categories	2
Number of scales	6
ViT dim	768
Dropout	0.5
Computing platform	**Parameter**
CPU	AMD EPYC 7302
GPU	NVIDIA 3090 24G
Operating system	Ubuntu 18.04
Python	3.8.0
PyTorch	2.0.0
Cudatoolkit	11.3.0

A. DATASET

We used the LUNA16 dataset [19] for our experiments on the identification of malignant and benign lung nodules to verify the performance of the proposed algorithm. The dataset was derived from a larger dataset called LIDC-IDRI [56], which consisted of images acquired from 1,018 CT scans. The data in this dataset were acquired by seven academic institutions by using varying scanners and related parameters. This resulted in highly heterogeneous CT scan images. Identifying lung nodules in images from this dataset was thus challenging, because of which the results obtained by the model were considered to be appropriately generalizable.

We excluded CT images from the LIDC-IDRI dataset with slices of thickness greater than 3 mm and lung nodules smaller than 3 mm. This yielded a dataset containing 888 CT images, known as the LUNA16 dataset. Information on the nodules appearing in the CT images was manually annotated by four radiologists. The LUNA16 dataset was divided into 10 subsets, eight of which were used as the training set and the other two as the test set. We conducted 10-fold cross-validation as well. For more information about LUNA16, the interested reader can refer to [19].

B. MSTD PERFORMANCE EXPERIMENTS

This experiment is performed to verify the superior performance of our MSTD method in the lung nodule diagnosis task. Table.2 presents the performance metrics of the MSTD method in lung nodule recognition tasks, including AUC, Accuracy, F1Score, Specificity, Sensitivity, etc. All metrics exceed 90%, indicating that our MSTD method exhibits high robustness for lung nodule classification.

In the actual scenario of diagnosing lung nodules, the computational burden of a model is a critical factor to consider. If the model requires too much computational power, it may be impractical to deploy in a clinical setting. Therefore, in

TABLE 2. The classification performance of the MSTD method

Method	AUC	Accuracy	F1Score	Specificity	Sensitivity
MSTD	0.984±0.011	0.945±0.015	0.945±0.016	0.965±0.016	0.911±0.043

Table.3, we conduct a statistical analysis of the computational cost. To illustrate the lightweight advantage of the MSTD model, we compare it with the state-of-the-art methods using the same computing device. CrossViT [16] is one of the most advanced and widely used methods in the field of image classification. CrossViT also includes operations involving the ViT module, making it an excellent candidate for comparison with our approach.

TABLE 3. The computational and parameter complexity of the MSTD and the state-of-the-art method

Method	FLOPs	Params	Times	GPU Memory
CrossViT [16]	16.079G	90.977M	21.342ms	1038M
Ours	6.467G	87.516M	18.813ms	812M

In Table.3, “FLOPs" stands for floating-point operations per second. “Params" represents the number of parameters in the model. “Times" indicates the computation time per sample. "GPU Memory" denotes the amount of GPU memory used. The comparison with the CrossViT method shows that our model slightly outperforms it in terms of computational time and the number of parameters. This is because we use the shallow multi-scale CNN branches to extract multi-scale features, whereas CrossViT employs a more computationally expensive multi-ViT branch architecture for the same purpose. Additionally, the diagnostic performance of MSTD is slightly better (see details in Table.7), indicating that MSTD does not sacrifice diagnostic performance due to its lightweight design. It’s noteworthy that although CrossViT has FLOPs that are 2.4 times higher than MSTD, its actual computation time is only 1.13 times that of MSTD. This discrepancy arises because FLOPs do not account for the parallel processing structure of CrossViT and its optimizations for GPU hardware acceleration. Overall, on our computing platform, the average time from inputting a sample into the MSTD model to obtaining results is only 18.813ms. Therefore, the MSTD model is lightweight and may perform rapid diagnosis of lung nodules without significantly increasing the time cost of the diagnostic system.

To help readers better understand the learning process of the model, we show the variation curve of the model’s diagnostic performance during training in Fig.5. We separate the curves of the three indicators: AUC, Accuracy, and F1Score into three subplots, in order to enable readers to clearly observe the changes of each indicator as the training progresses. From Fig.5, we can clearly see that during the 100th to 200th epochs, all indicators on the training set tend towards 100%. At the same time, the corresponding indicators on the test set also maintain a high level. This indicates that the MSTD

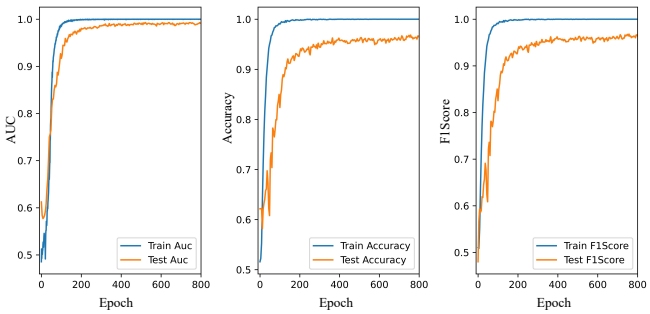


FIGURE 5. **The learning process curves of MSTD method.**

model is gradually converging and maintaining a good state at this stage.

C. MULTI-SCALE BRANCH EXPERIMENTS

To investigate the impact of different scale information on lung nodule diagnosis tasks, we conduct multi-scale branch experiments, the results are shown in Table.4.

The “MSMPB" represents the proposed MSMPB model with multiple parallel scale branches. The “No MSMPB" represents a single-scale model without the multi-scale structure. To avoid interference from the scale model comparison results by the MSTF module, we directly flatten the results obtained from the scale branches and input them into the fully connected layer of the classifier to obtain category results.

The results in Table.4 indicate that our MSMPB model with multi-scale information outperforms the single-scale model by a significant margin. Additionally, scales 32 and 48 are crucial for lung nodule recognition tasks. With the combined effect of 6 scales, the AUC performance of the MSMPB model is improved by 3.3% compared to the best single scale model.

D. TRANSFORMER FUSION MODULE EXPERMENTS

Our MSTF module is based on the transformer method and performs attention mechanism operations in both spatial and scale dimensions. In this experiment, we specifically conduct ablation experiments on the MSTF module to explore the most suitable scale fusion method.

We first construct a baseline method “No Transformer”, which removes all transformer operations and directly feeds the Multi-Scale Features outputted by the MSMPB module into the fully connected layer of the classifier. Secondly, we remove the Scale Transformer from the MSTF and only retain the Space Transformer to create a comparative module called “Only Space Transformer”. Lastly, we remove the Space Transformer from the MSTF and only retain the Scale Transformer to create a comparative module called “Only Scale Transformer”.

Table.5 shows the results of the ablation experiments for fusion methods. When the network model lacks the MSTF method, the AUC performance decreases by 5.9%. When the network model only retains the Space Transformer method, there is a slight improvement in classification performance compared to when there is no Transformer. This is because the Space Transformer method can appropriately enhance the model’s ability to model global spatial features of Multi Scale Features. However, this method lacks the ability to process critical scale information. When the network model only retains the Scale Transformer method, the classification performance is better than the No Transformer and Only Space Transformer methods. This indicates that the fusion process of scale information is crucial for Multi-Scale Features, as the classification task for lung nodules is more sensitive to scale variations. When both the Space Transformer and Scale Transformer methods operate simultaneously, the spatial and scale information of the input features complement each other, collectively promoting the improvement of classification performance.

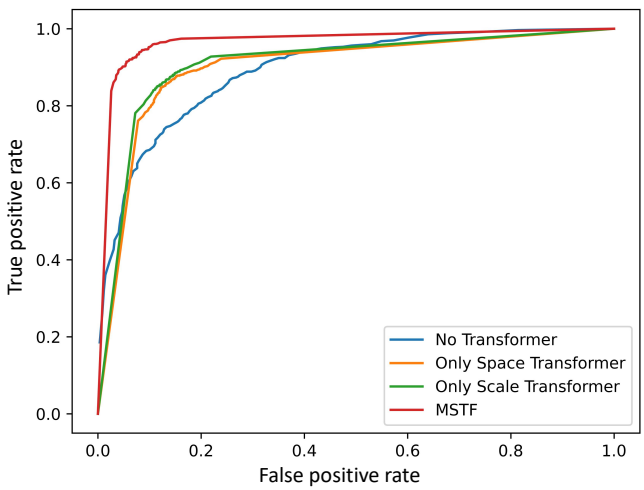


FIGURE 6. **Comparison results of ROC curves for MSTF modules.**

Fig.6 shows the ROC curve results of the MSTF experiments. The ROC curve of the MSTF method is located above the Space and Scale Transformer methods, which also validates the superior performance of our MSTF method.

The MSTF module enables the classification model to automatically focus on key information within the scale dimension. It is also meaningful to understand how MSTF internally pays attention to features of different scales. Therefore, we add an experiment to measure the contribution of different scale features within MSTF, to help readers gain a deeper understanding of the internal workings of MSTF and inspire more meaningful work.

When exploring the importance of a particular scale, we maintain the cube shape of the Multi-Scale Features unchanged and solely set the values of the feature maps of that scale to zero. This method cleverly eliminates the information of a single scale within the Multi-Scale Features, without disrupting the original structure of the model. This largely adheres to the principle of controlling variables.

We individually remove the information of six different scales within the Multi-Scale Features. Table.6 presents the

TABLE 4. The results of multi-scale branch experiments

Module	Method	AUC	Accuracy	F1Score	Specificity	Sensitivity
No MSMPB	Single scale (1×1)	0.647±0.045	0.386±0.057	0.228±0.053	0.014±0.013	1.000±0.000
	Single scale (4×4)	0.688±0.059	0.674±0.062	0.675±0.062	0.728±0.066	0.585±0.069
	Single scale (8×8)	0.788±0.038	0.694±0.043	0.698±0.042	0.621±0.052	0.815±0.045
	Single scale (16×16)	0.811±0.040	0.747±0.038	0.749±0.036	0.774±0.058	0.703±0.042
	Single scale (32×32)	0.901±0.034	0.825±0.036	0.826±0.036	0.830±0.038	0.817±0.051
	Single scale (48×48)	0.890±0.033	0.829±0.041	0.825±0.043	0.905±0.028	0.699±0.072
MSMPB	**Multi-scale**	**0.919±0.029**	**0.850±0.023**	**0.851±0.023**	0.862±0.020	**0.829±0.061**

TABLE 5. The results of transformer fusion experiments

Module	AUC	Accuracy	F1Score	Specificity	Sensitivity
No Transformer	0.919±0.029	0.850±0.023	0.851±0.023	0.862±0.020	0.829±0.061
Only Space Transformer	0.928±0.028	0.866±0.038	0.867±0.038	0.882±0.044	0.843±0.049
Only Scale Transformer	0.930±0.017	0.884±0.032	0.883±0.033	0.938±0.017	0.798±0.065
MSTF	**0.984±0.011**	**0.945±0.015**	**0.945±0.016**	**0.965±0.016**	**0.911±0.043**

TABLE 6. The experimental results on the contribution of different scales to MSTF performance

Module	Remove Scale	AUC	Accuracy	F1Score	Specificity	Sensitivity
MSTF	(48×48)	0.908±0.034	0.842±0.028	0.839±0.029	0.922±0.033	0.713±0.060
	(32×32)	0.917±0.025	0.853±0.042	0.854±0.041	0.856±0.053	0.848±0.039
	(16×16)	0.950±0.016	0.910±0.025	0.908±0.026	0.961±0.018	0.825±0.057
	(8×8)	0.955±0.028	0.896±0.025	0.896±0.025	0.918±0.027	0.860±0.040
	(4×4)	0.963±0.012	0.902±0.020	0.901±0.020	0.920±0.035	0.873±0.060
	(1×1)	0.975±0.012	0.928±0.028	0.927±0.028	0.953±0.029	0.883±0.054
	None	**0.984±0.011**	**0.945±0.015**	**0.945±0.016**	**0.965±0.016**	**0.911±0.043**

experimental results regarding the importance of different scale features. "Remove Scale" indicates the specific scale that is erased, while "None" represents no scale being erased, i.e., all scales are retained. Consequently, within MSTF, the higher the contribution of a scale, the lower the performance metrics will be after its removal. The results in Table.6 show that larger scales have a higher contribution to MSTF, while smaller scales have a lower contribution, which is consistent with the conclusions drawn from Table.4. This is because larger scales retain more comprehensive original information of lung nodule images, whereas smaller scales lose some effective information due to reduced resolution. However, each scale is indispensable for MSTF, and the classification performance of the model is highest only when all six scales are input into MSTF together.

E. COMPARISON WITH OTHER ADVANCED METHODS

In this experiment, we reproduce other advanced image classification methods based on multi-scale strategies and ViT in recent years, to conduct comparative tests with the MSTD method, thereby verifying the strong competitiveness of our approach.

In terms of CNN multi-scale strategies, the advanced methods used for comparison include ATCNN [44], Res2Net [58], and Fractalnet [43]. Notably, ATCNN and Fractalnet methods are specifically designed to address lung nodule classification problems. In the realm of advanced image classification methods based on ViT, the comparison methods include HRViT [57], HiFuse [41], CrossVit [16], and MViTv2 [59]. These methods also employ multi-scale architectures to enhance classification performance,which are suitable for comparison with our MSTD method. To ensure fairness in the comparative experiments, all comparison methods are tested on the same experimental platform and dataset. The experimental results are presented in Table.7.

It can be clearly seen from Table.7 that the average diagnostic performance of our MSTD method has increased by about 1%, 3.4%, and 3.6% in AUC, Accuracy, and F1Score indicators compared to the most advanced methods MViTv2. This indicates that our method based on multi-scale branch feature extraction and multi-scale transformer fusion has reached a leading level in the field of lung nodule diagnosis.

Fractalnet [43] uses a fractal structure, which is considered an effective alternative to the residual structure. However, this method is not more commonly used than the residual structure because it does not deliver strong generalization-related performance on various tasks, like ResNet does. Furthermore, Fractalnet consists of five fractal blocks arranged in series, each of which contains a multi-level convolutional structure with a depth of eight. In other words, the convolutional depth of Fractalnet can reach 40 layers. However, the scale-related branches in our proposed model contain only include three convolutional layers, and all its branches are arranged in a parallel structure. This means that our MSMPB module required

TABLE 7. The comparison results with other advanced methods

Author	Method	AUC	Accuracy	F1Score	Specificity	Sensitivity
Amrita et al. [43] 2021	Fractalnet	0.573±0.046	0.607±0.042	0.505±0.047	0.940±0.027	0.061±0.025
Gu et al. [57] 2022	HRViT	0.788±0.045	0.740±0.046	0.739±0.046	0.792±0.058	0.656±0.075
Dosov et al. [12] 2020	ViT	0.865±0.026	0.810±0.025	0.811±0.025	0.827±0.043	0.784±0.054
Gao et al. [58] 2019	Res2Net	0.933±0.022	0.867±0.030	0.867±0.029	0.897±0.050	0.815±0.048
Huo et al. [41] 2024	HiFuse	0.936±0.014	0.886±0.027	0.886±0.028	0.913±0.027	0.840±0.067
Chen et al. [16] 2021	CrossViT	0.943±0.021	0.878±0.030	0.879±0.030	0.882±0.039	0.870±0.042
Halder et al. [44] 2023	ATCNN	0.957±0.017	0.892±0.033	0.892±0.033	0.919±0.045	0.847±0.035
Li et al. [59] 2022	MViTv2	0.975±0.007	0.911±0.014	0.909±0.014	0.982±0.010	0.795±0.039
Ours	**MSTD**	**0.984±0.011**	**0.945±0.015**	**0.945±0.016**	0.965±0.016	**0.911±0.043**

only a shallow convolution to obtain the necessary features of the image. Therefore, the excessively deep network of Fractalnet may lead to overfitting or the distortion of crucial features of the images of lung nodules.

The ATCNN [44] uses atrous spatial pyramid pooling (ASPP) and residual convolutional structures to classify lung nodules in images. The pyramid structure of ASPP is similar to the multi-scale structure of our MSMPB module. However, after extracting multi-scale feature maps, ASPP uses 1x1 convolutions to reduce the number of channels for feature fusion. This approach relies solely on information from the spatial dimension for feature fusion, and neglects useful information on the scale of the image. On the contrary, we have specifically designed the MSTF to simultaneously perform feature fusion in both the spatial and scale-related dimensions. It thus outperformed the ATCNN.

Res2Net [58] is a mainstream model based on the multi-scale strategy of the CNN, and can construct multi-scale features by building multi-level residual connections within a single residual block. Res2Net mainly extracts local features through local convolution operations. However, the ViT can leverage its self-attention mechanism to process long-range dependencies in images within a global scope. Therefore, the diagnostic performance of Res2Net was inferior to that of CrossViT, HiFuse, MViTv2, and MSTD, which combine multi-scale strategies of fusion with the ViT architecture.

The ViT [12] method is one of the most popular approaches in the field of image classification, utilizing the self-attention mechanism of Transformers to capture dependencies between different spatial regions in images. However, ViT lacks the ability to capture and analyze features at different image scales. Our MSMPB method can capture features at different scales through multiple parallel scale branches, and the MSTF method can analyze features of different scales in both spatial and scale dimensions simultaneously. Therefore, our approach achieves better classification results.

The HiFuse [41] method primarily consists of a local feature branch and a global feature branch, both of which contain scale information across four stages. The local feature branch uses depthwise separable convolutions to extract features and employs linear layers for information interaction between channels. The global feature branch uses the Windows Multi-head Self-Attention (W-MSA) [60] module based on ViT to extract features. Features between the local and global branches are fused through the Hierarchical Feature Fusion (HHF) block, which includes channel attention, spatial attention, and an MLP. However, the HHF method can only fuse features at the same level, i.e., the same scale, from the two branches and cannot simultaneously fuse features at different scales. In contrast, our MSTF method can selectively and analyze features at multiple scales simultaneously, providing stronger scale fusion capabilities. Therefore, our method achieves better diagnostic performance compared to the HiFuse method.

The HRViT [57] model consists of four progressive Transformer stages, each representing a scale level. Within each stage, finer-grained scale features are obtained through upsampling and downsampling, and fine-grained scale interactions are achieved using multiple repeated augmented local self-attention blocks (HRViTAttn). However, the scale features across different stages are concatenated layer by layer without the design for fusion between them. Additionally, the full size of lung nodule images (64x64) is relatively small, making overly fine-grained scale partitioning and repeated self-attention blocks redundant. Therefore, HRViT quickly falls into overfitting during training, leading to poor diagnostic performance.

The CrossViT [16] method employs two independently ViT branches to process small and large sized patches, encoding them into tokens and fusing these tokens through multiple cross attention mechanisms. Cross attention treats the tokens from a single branch as Query (Q) and the tokens from the other branch as Key (K) and Value (V), thereby obtaining fused features. However, each dual-branch transformer block can only handle features of two sizes simultaneously. To fuse multiple different scale features, this dual-branch structure needs to be stacked multiple times, which can result in a complex model structure prone to overfitting issues. In contrast, our MSTF module, through clever feature dimension conversions, can process multiple scale features simultaneously with just two ViT operations. Therefore, our MSTD method exhibits stronger anti-overfitting capabilities and is more lightweight.

MViTv2 [59] is a powerful image processing model that possesses capabilities for image classification, object detection, and video recognition. In MViTv2's Pooling Attention, Q, K, and V undergo pooling layers to reduce the spatial resolution of backbone features, thereby obtaining a broader

field of view. This method then expands the channel dimensions of the backbone features through an MLP to acquire more abstract feature understanding capabilities. MViTv2 uses Pooling Attention to construct four stages of scales, extracting each scale individually, which can naturally integrate with feature pyramid networks for image classification tasks. The multi-scale features constructed through Pooling Attention possess stronger spatial long-range correlations, making MViTv2's diagnostic performance robust. However, these multi-scale feature fusions heavily rely on spatial features and ignore scale dimension analysis. Since our MSTF module has the capability to analyse both spatial and scale dimensions simultaneously, our method's classification performance is slightly higher.

Although the proposed method has improved diagnostic performance compared to state-of-the-art approaches, it still has some limitations. The MSMPB module extracts features of different scales through parallel scale branches to obtain the Multi-Scale Features. However, the shape of the Multi-Scale Features must be a cube rather than a cuboid, limiting the flexibility of the model's hyperparameters. Therefore, in future work, we will upgrade the MSTF module to flexibly handle Multi-Scale Features of different shapes.

Additionally, the MSMPB and MSTF modules are connected in series, which may result in a deeper network and loss of some useful information from the original input data. The scale branches of MSMPB are shallow, avoiding the issue of excessive network depth. However, this also limits the network depth of MSMPB, preventing it from extracting high-level semantic features. Therefore, in future work, we will attempt to establish appropriate residual connections between the MSMPB and MSTF modules to allow for a moderate increase in network depth.

V. CONCLUSION

The MSTD method proposed in this paper is a novel approach specifically designed for the diagnosis of benign and malignant lung nodules. This method obtains Multi-Scale Features through a multiple parallel scale branch structure, MSMPB. The scale branches are lightweight, consisting of only 3 convolutional layers with 8 kernel channels and an average pooling layer. The features from all scales are fused through the MSTF module, which includes spatial and scale transformers, to generate Space-Scale Features. Unlike conventional Vision Transformers, the introduction of the scale transformer in this paper is a novel contribution of this study. The MSMPB module is cleverly designed to maintain consistent lengths in spatial and scale dimensions for the Multi-Scale Features, facilitating the MSTF module to perform attention operations simultaneously in scale and spatial dimensions. The classifier, composed of fully connected layers, outputs the final diagnosis of benign or malignant nodules. Through experiments, we have demonstrated the effectiveness of the lightweight MSMPB module and highlighted the enhancement of classification performance achieved by the Space and Scale Transformers.

DECLARATIONS

The study was approved by the Ethics Committee of the Sixth Medical Center of the Chinese PLA General Hospital. The data on all patient were sourced from publicly available databases. All methods were applied in accordance with relevant guidelines and regulations.

None of the authors has a conflict of interest to report regarding the publication of this manuscript.

REFERENCES

[1] S. J. Adams, E. Stone, D. R. Baldwin, R. Vliegenthart, P. Lee, and F. J. Fintelmann, "Lung cancer screening," *The Lancet*, vol. 401, no. 10374, pp. 390–408, 2023.
[2] P. Sengodan, K. Srinivasan, R. Pichamuthu, and S. Matheswaran, "Early detection and classification of malignant lung nodules from ct images: an optimal ensemble learning," *Expert Systems with Applications*, vol. 229, p. 120361, 2023.
[3] C. de Margerie-Mellon and G. Chassagnon, "Artificial intelligence: A critical review of applications for lung nodule and lung cancer," *Diagnostic and Interventional Imaging*, vol. 104, no. 1, pp. 11–17, 2023.
[4] S. H. Hosseini, R. Monsefi, and S. Shadroo, "Deep learning applications for lung cancer diagnosis: a systematic review," *Multimedia Tools and Applications*, vol. 83, no. 5, pp. 14 305–14 335, 2024.
[5] J. Xu, H. Ren, S. Cai, and X. Zhang, "An improved faster r-cnn algorithm for assisted detection of lung nodules," *Computers In Biology And Medicine*, vol. 153, p. 106470, 2023.
[6] D. Zhao, Y. Liu, H. Yin, and Z. Wang, "An attentive and adaptive 3d cnn for automatic pulmonary nodule detection in ct image," *Expert Systems with Applications*, vol. 211, p. 118672, 2023.
[7] A. A. Shah, H. A. M. Malik, A. Muhammad, A. Alourani, and Z. A. Butt, "Deep learning ensemble 2d cnn approach towards the detection of lung cancer," *Scientific reports*, vol. 13, no. 1, p. 2987, 2023.
[8] Z. UrRehman, Y. Qiang, L. Wang, Y. Shi, Q. Yang, S. U. Khattak, R. Aftab, and J. Zhao, "Effective lung nodule detection using deep cnn with dual attention mechanisms," *Scientific Reports*, vol. 14, no. 1, p. 3934, 2024.
[9] A. E. Prosper, M. N. Kammer, F. Maldonado, D. R. Aberle, and W. Hsu, "Expanding role of advanced image analysis in ct-detected indeterminate pulmonary nodules and early lung cancer characterization," *Radiology*, vol. 309, no. 1, p. e222904, 2023.
[10] J. Shi, Y. Wang, Z. Yu, G. Li, X. Hong, F. Wang, and Y. Gong, "Exploiting multi-scale parallel self-attention and local variation via dual-branch transformer-cnn structure for face super-resolution," *IEEE Transactions on Multimedia*, 2023.
[11] Q. Wang, K. Jiang, Z. Wang, W. Ren, J. Zhang, and C.-W. Lin, "Multi-scale fusion and decomposition network for single image deraining," *IEEE Transactions on Image Processing*, vol. 33, pp. 191–204, 2023.
[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
[13] M. Xie and N. Ye, "Multi-scale and multi-factor vit attention model for classification and detection of pest and disease in agriculture." *Applied Sciences (2076-3417)*, vol. 14, no. 13, 2024.
[14] G.-I. Kim and K. Chung, "Vit-based multi-scale classification using digital signal processing and image transformation," *IEEE Access*, 2024.
[15] Z. Xu and Z. Wang, "Mcv-unet: a modified convolution & transformer hybrid encoder-decoder network with multi-scale information fusion for ultrasound image semantic segmentation," *PeerJ Computer Science*, vol. 10, p. e2146, 2024.
[16] C.-F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 357–366.
[17] R. Shao, X.-J. Bi, and Z. Chen, "Hybrid vit-cnn network for fine-grained image classification," *IEEE Signal Processing Letters*, 2024.
[18] S. S. Basha, S. R. Dubey, V. Pulabaigari, and S. Mukherjee, "Impact of fully connected layers on performance of convolutional neural networks for image classification," *Neurocomputing*, vol. 378, pp. 112–119, 2020.
[19] A. A. A. Setio, A. Traverso, T. De Bel, M. S. Berens, C. Van Den Bogaard, P. Cerello, H. Chen, Q. Dou, M. E. Fantacci, B. Geurts *et al.*, "Validation,

comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge," *Medical image analysis*, vol. 42, pp. 1–13, 2017.

[20] T. Liu, E. Siegel, and D. Shen, "Deep learning and medical image analysis for covid-19 diagnosis and prediction," *Annual Review of Biomedical Engineering*, vol. 24, pp. 179–201, 2022.

[21] Q. Zeng, W. Sun, J. Xu, W. Wan, and L. Pan, "Machine learning-based medical imaging detection and diagnostic assistance," *International Journal of Computer Science and Information Technology*, vol. 2, no. 1, pp. 36–44, 2024.

[22] D. Kollias, A. Arsenos, and S. Kollias, "Domain adaptation, explainability & fairness in ai for medical image analysis: Diagnosis of covid-19 based on 3-d chest ct-scans," *arXiv preprint arXiv:2403.02192*, 2024.

[23] L. Schneider, S. Laiouar-Pedari, S. Kuntz, E. Krieghoff-Henning, A. Hekler, J. N. Kather, T. Gaiser, S. Froehling, and T. J. Brinker, "Integration of deep learning-based image analysis and genomic data in cancer pathology: A systematic review," *European journal of cancer*, vol. 160, pp. 80–91, 2022.

[24] Y. Zhang, Z. Shen, and R. Jiao, "Segment anything model for medical image segmentation: Current applications and future directions," *Computers in Biology and Medicine*, p. 108238, 2024.

[25] J. Wu, R. Fu, H. Fang, Y. Zhang, Y. Yang, H. Xiong, H. Liu, and Y. Xu, "Medsegdiff: Medical image segmentation with diffusion probabilistic model," in *Medical Imaging with Deep Learning*. PMLR, 2024, pp. 1623–1639.

[26] J. Wu, W. Ji, H. Fu, M. Xu, Y. Jin, and Y. Xu, "Medsegdiff-v2: Diffusion-based medical image segmentation with transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 6030–6038.

[27] Z. Xing, T. Ye, Y. Yang, G. Liu, and L. Zhu, "Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation," *arXiv preprint arXiv:2401.13560*, 2024.

[28] J. Guo, W. Cao, B. Nie, and Q. Qin, "Unsupervised learning composite network to reduce training cost of deep learning model for colorectal cancer diagnosis," *IEEE journal of translational engineering in health and medicine*, vol. 11, pp. 54–59, 2022.

[29] S. Zakareya, H. Izadkhah, and J. Karimpour, "A new deep-learning-based model for breast cancer diagnosis from medical images," *Diagnostics*, vol. 13, no. 11, p. 1944, 2023.

[30] M. M. Eltoukhy, K. M. Hosny, and M. A. Kassem, "Classification of multiclass histopathological breast images using residual deep learning," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[31] W. K. Cheung, A. Pakzad, N. Mogulkoc, S. H. Needleman, B. Rangelov, E. Gudmundsson, A. Zhao, M. Abbas, D. McLaverty, D. Asimakopoulos *et al.*, "Interpolation-split: a data-centric deep learning approach with big interpolated data to boost airway segmentation performance," *Journal of big Data*, vol. 11, no. 1, p. 104, 2024.

[32] W. K. Cheung, A. Pakzad, N. Mogulkoc, S. Needleman, B. Rangelov, E. Gudmundsson, A. Zhao, M. Abbas, D. McLaverty, D. Asimakopoulos *et al.*, "Automated airway quantification associates with mortality in idiopathic pulmonary fibrosis," *European radiology*, vol. 33, no. 11, pp. 8228–8238, 2023.

[33] S. Vijayakumar, S. Aarthy, D. Deepa, and P. Suresh, "Sustainable framework for automated segmentation and prediction of lung cancer in ct image using capsnet with u-net segmentation," *Biomedical Signal Processing and Control*, vol. 99, p. 106873, 2025.

[34] A. A. Asiri, A. Shaf, T. Ali, M. A. Pasha, M. Aamir, M. Irfan, S. Alqahtani, A. J. Alghamdi, A. H. Alghamdi, A. F. A. Alshamrani *et al.*, "Advancing brain tumor classification through fine-tuned vision transformers: A comparative study of pre-trained models," *Sensors*, vol. 23, no. 18, p. 7913, 2023.

[35] C. Ma, L. Zhao, Y. Chen, S. Wang, L. Guo, T. Zhang, D. Shen, X. Jiang, and T. Liu, "Eye-gaze-guided vision transformer for rectifying shortcut learning," *IEEE Transactions on Medical Imaging*, vol. 42, no. 11, pp. 3384–3394, 2023.

[36] H. Feng, B. Yang, J. Wang, M. Liu, L. Yin, W. Zheng, Z. Yin, and C. Liu, "Identifying malignant breast ultrasound images using vit-patch," *Applied Sciences*, vol. 13, no. 6, p. 3489, 2023.

[37] H. Shin, S. Jeon, Y. Seol, S. Kim, and D. Kang, "Vision transformer approach for classification of alzheimer's disease using 18f-florbetaben brain images," *Applied Sciences*, vol. 13, no. 6, p. 3453, 2023.

[38] Y. Chen, X. Zhang, Y. He, L. Peng, L. Pu, and F. Sun, "Mixunet: A lightweight medical image segmentation network capturing multidimensional semantic information," *Biomedical Signal Processing and Control*, vol. 96, p. 106513, 2024.

[39] Y. Wang, X. Yu, X. Guo, X. Wang, Y. Wei, and S. Zeng, "A dual-decoding branch u-shaped semantic segmentation network combining transformer attention with decoder: Dbunet," *Journal of Visual Communication and Image Representation*, vol. 95, p. 103856, 2023.

[40] C. Fan, Q. Su, Z. Xiao, H. Su, A. Hou, and B. Luan, "Vit-frd: A vision transformer model for cardiac mri image segmentation based on feature recombination distillation," *IEEE Access*, 2023.

[41] X. Huo, G. Sun, S. Tian, Y. Wang, L. Yu, J. Long, W. Zhang, and A. Li, "Hifuse: Hierarchical multi-scale feature fusion network for medical image classification," *Biomedical Signal Processing and Control*, vol. 87, p. 105534, 2024.

[42] S. Liu, W. Yue, Z. Guo, and L. Wang, "Multi-branch cnn and grouping cascade attention for medical image classification," *Scientific Reports*, vol. 14, no. 1, p. 15013, 2024.

[43] A. Naik, D. R. Edla, and V. Kuppili, "Lung nodule classification on computed tomography images using fractalnet," *Wireless Personal Communications*, vol. 119, no. 2, pp. 1209–1229, 2021.

[44] A. Halder and D. Dey, "Atrous convolution aided integrated framework for lung nodule segmentation and classification," *Biomedical Signal Processing and Control*, vol. 82, p. 104527, 2023.

[45] H. Huang, R. Wu, Y. Li, and C. Peng, "Self-supervised transfer learning based on domain adaptation for benign-malignant lung nodule classification on thoracic ct," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 3860–3871, 2022.

[46] Y. Xu, Q. She, S. Sun, X. Xi, and S. Du, "Attribute-enhanced capsule network for pulmonary nodule classification," *Journal of Medical and Biological Engineering*, pp. 1–11, 2024.

[47] J. Miao, M. Zhang, Y. Chang, and Y. Qiao, "Transformer-based recognition model for ground-glass nodules from the view of global 3d asymmetry feature representation," *Symmetry*, vol. 15, no. 12, p. 2192, 2023.

[48] M. Kanipriya, C. Hemalatha, N. Sridevi, S. SriVidhya, and S. J. Shabu, "An improved capuchin search algorithm optimized hybrid cnn-lstm architecture for malignant lung nodule detection," *Biomedical Signal Processing and Control*, vol. 78, p. 103973, 2022.

[49] F. Shi, B. Chen, Q. Cao, Y. Wei, Q. Zhou, R. Zhang, Y. Zhou, W. Yang, X. Wang, R. Fan *et al.*, "Semi-supervised deep transfer learning for benign-malignant diagnosis of pulmonary nodules in chest ct images," *IEEE transactions on medical imaging*, vol. 41, no. 4, pp. 771–781, 2022.

[50] M. Zhang, Z. Kong, W. Zhu, F. Yan, and C. Xie, "Pulmonary nodule detection based on 3d feature pyramid network with incorporated squeeze-and-excitation-attention mechanism," *Concurrency and Computation: Practice and Experience*, vol. 35, no. 16, p. e6237, 2023.

[51] S. Zhang, J. Wu, E. Shi, S. Yu, Y. Gao, L. C. Li, L. R. Kuo, M. J. Pomeroy, and Z. J. Liang, "Mm-glcm-cnn: A multi-scale and multi-level based glcm-cnn for polyp classification," *Computerized Medical Imaging and Graphics*, vol. 108, p. 102257, 2023.

[52] H. Xu, S. Zhong, T. Zhang, and X. Zou, "Multi-scale multi-level residual feature fusion for real-time infrared small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[53] T. Yao, Y. Li, Y. Pan, Y. Wang, X.-P. Zhang, and T. Mei, "Dual vision transformer," *IEEE transactions on pattern analysis and machine intelligence*, 2023.

[54] S. Bruch, "An alternative cross entropy loss for learning-to-rank," in *Proceedings of the web conference 2021*, 2021, pp. 118–126.

[55] A. Mao, M. Mohri, and Y. Zhong, "Cross-entropy loss functions: Theoretical analysis and applications," in *International Conference on Machine Learning*. PMLR, 2023, pp. 23 803–23 828.

[56] L. M. Pehrson, M. B. Nielsen, and C. Ammitzbøl Lauridsen, "Automatic pulmonary nodule detection applying deep learning or machine learning algorithms to the lidc-idri database: a systematic review," *Diagnostics*, vol. 9, no. 1, p. 29, 2019.

[57] J. Gu, H. Kwon, D. Wang, W. Ye, M. Li, Y.-H. Chen, L. Lai, V. Chandra, and D. Z. Pan, "Multi-scale high-resolution vision transformer for semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 094–12 103.

[58] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 2, pp. 652–662, 2019.

[59] Y. Li, C.-Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer, "Mvitv2: Improved multiscale vision transformers for classification

and detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4804–4814.

[60] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

XIAOYU ZHAO obtained a Bachelor's degree in Clinical Medicine from Jining Medical University in 2022 and is currently pursuing a Master's degree at Hebei North University and the Sixth Medical Center of Chinese PLA General Hospital. Her main research areas include lung cancer and pulmonary nodules.

JIAO LI obtained a Bachelor's degree in Clinical Medicine from Southwest Medical University in 2022. She is currently pursuing a Master's degree at Anhui Medical University and the Sixth Medical Center of Chinese PLA General Hospital, with a primary research focus on the treatment of advanced lung cancer.

MAN QI Obtained a Doctor of Medicine degree from Tongji University. Engaged in post-doctoral research at the State Key Laboratory of Cardiovascular Research, Fuwai Hospital, Chinese Academy of Medical Sciences. Currently employed at the Sixth Medical Center of Chinese PLA General Hospital. Her research interests include respiratory system diseases and pulmonary diseases.

XUXIN CHEN Doctor of Medicine, Associate Chief Physician, Master's supervisor, and supervisor for American BLS and ACLS. Enjoys a Class III allowance for outstanding professional technical talents in the military. Serves as a member of the Respiratory Disease Branch of the Geriatric Health Medicine Society, a member of the Technical Committee on Respiratory Equipment of the Medical Equipment Association, and a member of the Critical Care Group of the Respiratory Committee of the Armed Forces.

WEI CHEN Associate Chief Physician in the Department of Respiratory and Critical Care Medicine at the PLA General Hospital, with a Master of Medicine degree. He has long been engaged in the research of infectious diseases of the respiratory system and interstitial lung disease. He serves as a member of the Infection Group of the Respiratory Internal Medicine Professional Committee of the Medical Science and Technology Committee, a member of the Infection Group of the Beijing Medical Association, a member of the Respiratory Internal Medicine Professional Committee of the Beijing Association of Traditional Chinese and Western Medicine, and a member of the Marine Medicine Branch of the Chinese Medical Association. He has participated in three research projects both within and outside the military, led two projects, and received two third-class military medical achievement awards.

YONGQUN LI Associate Chief Physician in the Department of Respiratory and Critical Care Medicine at the Sixth Medical Center of the PLA General Hospital, serves as a member of the Lung Cancer Group of the Respiratory Professional Committee of the PLA, a member of the Precise Medicine Professional Committee of the China Association for Rehabilitation Technology Transformation and Development Promotion, and a member of the Lung Cancer Immunotherapy Committee of the Beijing Cancer Prevention and Treatment Society. He has long been engaged in the diagnosis and treatment of respiratory system diseases and pulmonary tumors, particularly in the field of precise comprehensive treatment and minimally invasive interventional treatment for lung cancer, gaining rich experience.

QI LIU obtained a bachelor's degree from China Three Gorges University in 2016; worked as a radiologist at the Fourth Medical Center of Chinese PLA General Hospital from 2017 to 2021; and obtained a master's degree in medicine from Anhui Medical University in 2024, with a major research direction of lung cancer.

JIAJIA TANG obtained her bachelor's degree from Xi'an Medical University in 2022. She is currently pursuing a master's degree at South China University of Technology and the Sixth Medical Center of the PLA General Hospital, with a research focus on acute lung injury and related topics.

ZHIHAI HAN is a Chief Physician in the Department of Respiratory and Critical Care Medicine at the Chinese People's Liberation Army General Hospital, holding a Doctor of Medicine degree and a professorship. He serves as a doctoral supervisor at the PLA Medical Academy, South China University of Technology, Anhui Medical University, and Southern Medical University. He is the director of the Department of Respiratory and Critical Care Medicine at the Sixth Medical Center and the director of the Tuberculosis Specialty Center of the Chinese People's Liberation Army. He has previously worked as a visiting physician at Advocate Christ Medical Center in Chicago and as a visiting scholar at Jefferson University in Philadelphia. Han has been engaged in basic and treatment research on respiratory critical illnesses (such as sepsis and acute respiratory distress syndrome) for over 30 years, with particular expertise in combat-related respiratory critical conditions.

CHUNYANG ZHANG is an Associate Chief Physician in the Department of Respiratory and Critical Care Medicine at the PLA General Hospital, holding a Doctor of Medicine degree. He is a member of the Lung Interstitial Disease Group of the 10th Military Respiratory Professional Committee, a youth member of the 2nd Beijing Medical Association Respiratory Professional Committee, a member of the Pleurisy and Mediastinal Disease Group of the 9th Beijing Medical Association Respiratory Division, and an executive member of the Respiratory Internal Medicine Professional Committee of the 5th Beijing Association of Integrative Medicine. He serves as an editorial board member for the journal "Chinese Journal of Drug Applications and Monitoring." He has received two third-class military medical achievement awards and one third-class military scientific and technological progress award.