Hindawi
Complexity
Volume 2020, Article ID 6153657, 12 pages
https://doi.org/10.1155/2020/6153657

Research Article

Fine-Grained Lung Cancer Classification from PET and CT Images Based on Multidimensional Attention Mechanism

RuoXi Qin,[1] **Zhenzhen Wang**,[2] **LingYun Jiang**,[1] **Kai Qiao**,[1] **Jinjin Hai**,[1] **Jian Chen**,[1] **Junling Xu**,[2] **Dapeng Shi**,[2] **and Bin Yan**[1]

[1]*PLA Strategy Support Force Information Engineering University, Zhengzhou 450001, China*
[2]*Department of Radiology, Henan Provincial People's Hospital, Zhengzhou 450002, China*

Correspondence should be addressed to Bin Yan; ybspace@hotmail.com

Received 10 October 2019; Revised 24 December 2019; Accepted 30 December 2019; Published 20 January 2020

Academic Editor: Sergio Gómez

Lung cancer ranks among the most common types of cancer. Noninvasive computer-aided diagnosis can enable large-scale rapid screening of potential patients with lung cancer. Deep learning methods have already been applied for the automatic diagnosis of lung cancer in the past. Due to restrictions caused by single modality images of dataset as well as the lack of approaches that allow for a reliable extraction of fine-grained features from different imaging modalities, research regarding the automated diagnosis of lung cancer based on noninvasive clinical images requires further study. In this paper, we present a deep learning architecture that combines the fine-grained feature from PET and CT images that allow for the noninvasive diagnosis of lung cancer. The multidimensional (regarding the channel as well as spatial dimensions) attention mechanism is used to effectively reduce feature noise when extracting fine-grained features from each imaging modality. We conduct a comparative analysis of the two aspects of feature fusion and attention mechanism through quantitative evaluation metrics and the visualization of deep learning process. In our experiments, we obtained an area under the ROC curve of 0.92 (balanced accuracy = 0.72) and a more focused network attention which shows the effective extraction of the fine-grained feature from each imaging modality.

1. Introduction

In the 21st century, cancer is still considered a serious disease as the mortality rates are high. Among all cancer types, lung cancer ranks first regarding morbidity and mortality [1, 2]. There are two main categories of lung cancer: non-small-cell lung cancer (NSCLC) and small cell lung cancer (SCLC). For non-small-cell lung cancer, a subcategorization into lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD) is further used. These types of cancers account for approximately 85% of lung cancer cases [3]. Compared with the diagnosis of benign and malignant, further fine-grained classification of lung cancers such as LUSC, LUAD, and SCLC is of great significance for the prognosis of lung cancer. Accurately determining the category of lung cancer in the early diagnosis directly influences the effect of the treatment and thus the patients' survival rate [1, 4]. Positron emission tomography (PET) and computed tomography (CT) are both widely used noninvasive diagnostic imaging techniques for clinical diagnosis in general and for the diagnosis of lung cancer in particular [4]. Immunohistochemical evaluation is considered the gold standard for lung cancer classification. However, this procedure requires a tissue biopsy, an invasive procedure with the inherent risk of a delayed diagnosis and thus exacerbation of the patient's pain.

Advances in artificial intelligence research enabled numerous studies on the automatic diagnosis of lung cancer. The use of data in lung cancer-type classification is roughly divided into three categories: CT and PET image data as well as pathological images [5]. The well-known data science community Kaggle provides high-quality CT images for participants with the task to distinguish malignant or benign nodules from pulmonary nodules. Kaggle competitions repeatedly produce excellent deep learning approaches for these tasks [6, 7]. With the progresses in the

research of automatic lung cancer diagnosis, studies are no longer limited to the classification of benign and malignant nodules and data sets are no longer limited to CT images [8–12]. Wu et al. [9] use quantitative imaging characteristics such as statistical, histogram-related, morphological, and textural features from PET images to predict the distance metastasis of NSCLC, which shows that quantitative features based on PET images can effectively characterize intratumor heterogeneity and complexity. Two recent publications propose the application of deep learning to pathological images to classify NSCLC and SCLC [10] and to classify transcriptome subtypes of LUAD [11]. The complexity of the clinical diagnosis of lung cancer is also characterized by the wide range of imaging modality, which is employed in the diagnosis [13, 14]. Previous research already proved that deep learning approaches can not only use the feature distribution patterns from different pulmonary imaging modalities but even merging different features to achieve the computer-aided diagnosis. Liang et al. [15] employ multichannel techniques to predict the IDH genotype from PET/CT data using a convolutional neural network (CNN), while other approaches use a parallel CNN architecture to extract several features of different imaging modalities [16, 17].

Compared with the classification of the benign and malignant, the classification of the three types of lung cancer from medical images are more suitable to constitute a fine-grained image recognition problem as diverse distributions of features and potential pathological features need to be considered. Because the fine-grained features which need to extract in images, and meanwhile the lesion region is a small part of the whole image, the deep learning framework is susceptible to feature noise. At present, most methods based on various deep learning frameworks have proved to have certain bottleneck in fine-grained problems. In order to solve this problem, the previous research mainly implements the attention mechanism from the two dimensions (channel and spatial) of the feature representation. The channel attention mechanism models the relationship between feature channels [18], while the spatial attention mechanism ensures that noise is suppressed by weighting feature representation spatially [19–21]. So far, spatial attention mechanism has been used in medical image processing to enhance extracted features [20, 21]. The channel attention mechanism has been used in the detection and classification of pulmonary disease [22, 23]. The presentation of these attention mechanisms illustrates the source of characteristic noise from different perspectives. There are few related studies on how to use the attention mechanism more effectively on images with different imaging modalities, so the deep learning model based on the multimodality dataset still has problems in fine-grained problems.

In this paper, we use noninvasive clinical images to achieve the computer-aided diagnosis of fine-grained lung cancer on the basis of deep learning. Our network architecture consists of two parallel three-dimensional DenseNet, and each DenseNet corresponds to one input imaging modality. To more effectively extract the fine-grained features in different modalities, we combine the 3D DenseNet with a multidimensional (channel and spatial) attention mechanisms to further enhance the extraction of fine-grained features. This network architecture is used to extract features from different imaging modalities in parallel. Through the fusion of features, the fine-grained feature representation of different modalities is used to achieve the final classification. We evaluate our method and prove the effectiveness of our fine-grained lung cancer classification approach. Furthermore, the visualization experiment of deep learning network reveals the benefits of different attention mechanisms for different imaging modalities, demonstrating the effectiveness of multidimensional attention.

2. Methods

The network construction is mainly divided into the following two parts: (1) The multidimensional attention mechanism proposed in the single-path network architecture is the method of fine-grained feature extraction for each modality. (2) On the basis of the single-path network architecture, a parallel network architecture is established to achieve parallel extraction and fusion of multimodality features.

2.1. Fine-Grained Feature Extraction Network Based on Multidimensional Attention Mechanism. 3D CNNs [24] were used for early cancer detection to preserve the spatial relationship between neighboring CT slices [25, 26]. DenseNet [27] has been applied to numerous problems within the medical field [28, 29] because of its connectivity pattern and the small number of parameters needed. For these reasons, we use 3D DenseNet as our baseline model in the approach presented in this paper. To address the problem of noise in the extraction of fine-grained features, we proposed a multidimensional attention mechanism embedded in a single-path network. Our network structure consists of three main components: a 3D DenseNet block, SE block, and a spatial attention-gated module. Figure 1 shows the structure of our single-path model.

The main part of our network is composed by a 3D DenseNet [27]. Each of the dense block consists of a specific number of three-dimensional convolutional layers. The parameter quantity of DenseNet is determined by the feature channel (growth rate k) output by each convolutional layer. To ensure feature depth, the number of convolutional layers is set to 4 in different dense blocks. The k is set to 16, making the parameter in the network a small number to mitigate overfitting. The feature map after each dense-block contains all features of the previous convolutional layer. The SE block [18] is used after each dense block to employ channel attention. The SE block in the 3D model is calculated according to equation (1) as follows:

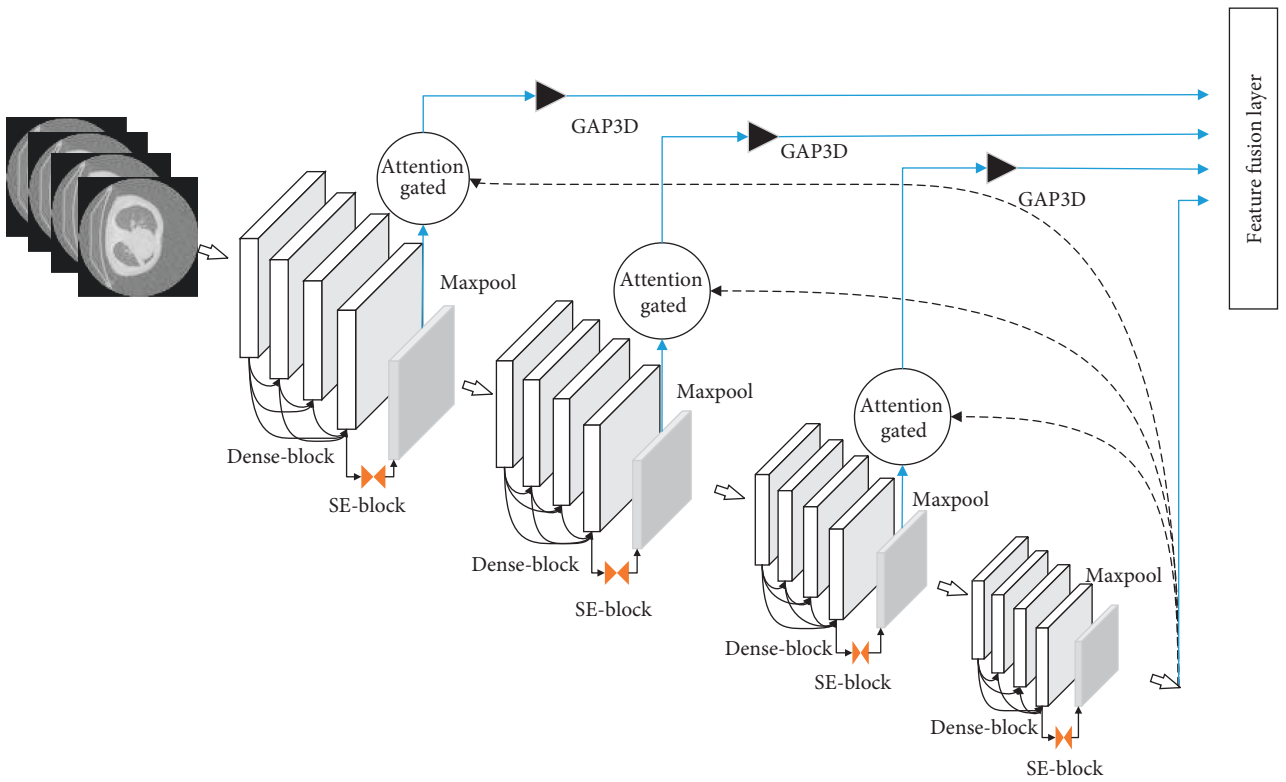


Figure 1: The structure of our single-path feature extraction network.

$$W_c(F) = \sigma(\mathrm{MLP}(\mathrm{Gap})(F)) = \sigma\left(W_1\left(W_0\left(\frac{1}{M \times N \times P}\sum_{i}^{M}\sum_{j}^{N}\sum_{k}^{P} F_c(i,j,k)\right)\right)\right). \quad (1)$$

Gap refers to the 3D global average pooling operation and σ refers to the sigmoid function. $W_0 \in \mathbb{R}^{C/r \times C}$ and $W_1 \in \mathbb{R}^{C \times C/r}$ compose a multilayer perceptron (MLP) with one hidden layer and r as the reduction ratio. The spatial attention-gated module [20] uses high-level semantic features and the feature after each SE block to generate the corresponding spatial mapping. Moreover, it weights all feature channels spatially to suppress noise originating from a nonlesion area. The spatial attention mechanism computes as defined in equation (2). The flow chart of the attention-gated module is shown in Figure 2:

$$\sigma\left(M_c\left(\delta\left(M_l(F_l) + U\left(M_g\left(F_g\right)\right)\right)\right)\right) \odot F_l, \quad (2)$$

where δ denotes the ReLU activation function, $\odot$ denotes the element-wise multiplication, and U denotes the upsampling operation. F_l refers to a feature map from different SE blocks. $M_l(F_l) \in \mathbb{R}^{H1 \times W1 \times Z1 \times G}$, $M_g(F) \in \mathbb{R}^{H2 \times W2 \times Z2 \times G}$, and $M_c \in \mathbb{R}^{H1 \times W1 \times Z1 \times G}$ describe the convolution operation with the specific channel output. As shown in Figure 2, the spatial attention mechanism generates the attention mapping through the element-wise add operation following the sigmoid activation function. Subsequently multiplied by F_l, the spatial weighted feature representation is generated. As F_l holds the feature map after SE block, we obtain the feature map weighted among the feature channel and spatial through this operation. The global average pooling operation used after each attention-gated module is to achieve the feature dimension reduction.

2.2. Parallel CNN Architecture Based on Multimodality Feature Fusion. In this section, we employ a parallel network architecture to extract and fuse features from multimodality data. The overall network structure and the flow chart of GMU are shown in Figure 3.

As illustrated in Figure 3, each single-path network equals the single-path network described before. We employ the gated multimodal unit (GMU) fusion strategy [30] for the fusion of different modality features. In contrast to the widely used connection operation, GMU allows to use hidden structures and gate controls to learn the intermediate representation of the multimodality features, thus enabling the prediction layer to assign weights to features that have intrinsic associations better. The calculation process of GMU is shown in the following equations:

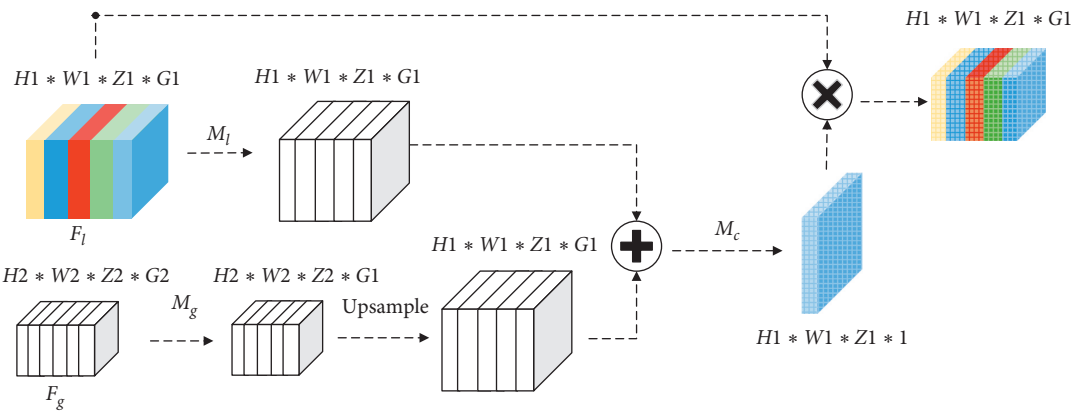


Figure 2: The flow chart of the attention-gated module.

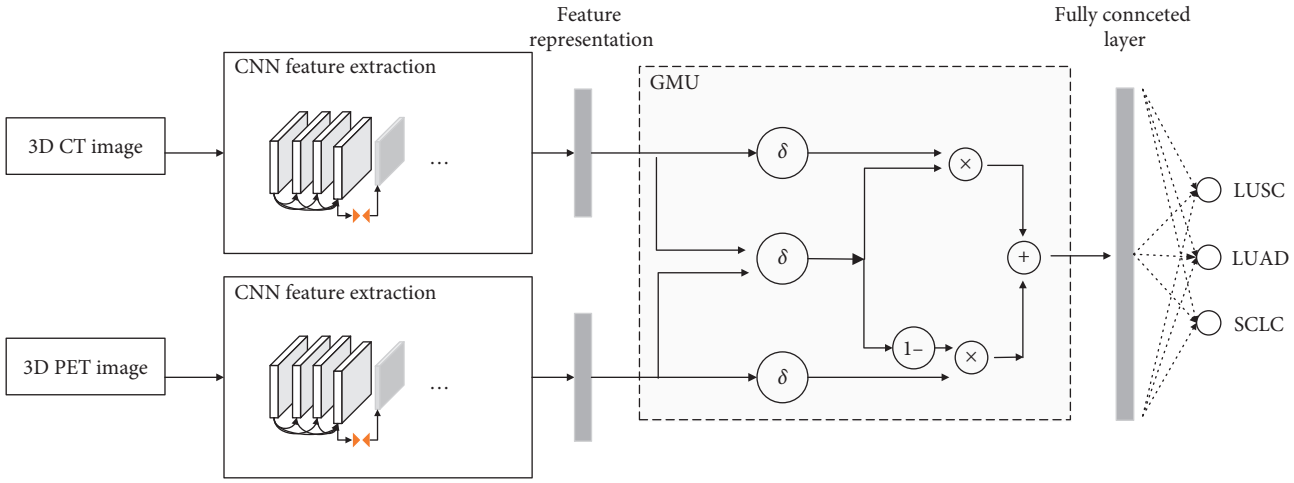


Figure 3: The structure of the parallel feature fusion network.

$$
\begin{aligned}
H_1 &= \delta\left(W_2\left(x_c\right)\right),\\
H_2 &= \delta\left(W_3\left(x_p\right)\right),\\
Z &= \sigma\left(W_4\left(\left[x_c, x_p\right]\right)\right),\\
H &= Z \odot H_1 + (1 - Z) \odot H_2.
\end{aligned}
\tag{3}
$$

where x_c refers to the feature extracted from the CT image, while x_p refers to the feature extracted from the PET image; H_1 and H_2 are the hidden states that are reached after the fully connected layer with ReLU activation function δ. Let the $W_2, W_3 \in \mathbb{R}^{C' \times C_{\text{int}}}$ and $W_4 \in \mathbb{R}^{C' \times 2C_{\text{int}}}$. $[\cdot, \cdot]$ refers to the connection operation and Z refers to the nonlinear weight learned from the combined features, which reveal the intrinsic relationship between the two modalities. The fused feature H is finally constructed by a linear weight between H_1 and H_2.

Inspired by the deep-supervision [31] method, the loss from each single-path network is, respectively, calculated and finally integrated with the loss from the feature fusion representation to obtain the joint optimization. Under the premise of a final classification, this training method forces the network to extract better high-level features from each modality for the generation of spatial attention to avoid trapping in a local minimum because of the use of feature from each level.

3. Experiments and Results

Our experiments mainly demonstrate the methods presented in this paper considering three aspects: (1) the validity of multimodality data. (2) The validity of multidimensional attention mechanism on each modality. (3) The validity of the feature fusion strategy. We evaluate the results under certain evaluation criteria to reflect the effectiveness of these methods. Regarding experimental details, batch normalization is employed prior to the Leaky-ReLU activation function [32]. The stochastic gradient descent (SGD) with a momentum of 0.9 is the optimizer. In the final fully connected layer of DenseNet, L1-regularization and dropout strategies are used to prevent overfitting. This framework is executed using Keras under a TITAN V 12 GB GPU.

To demonstrate the generalization power of our proposed network, we statistically analyzed the performance of the model in tenfold cross-validations. The area under the ROC curve (AUC), a metric that is widely used in medical image classification, objectively reflects the ability to classify positive and negative samples correctly. In addition,

accuracy is also used as a criterion of the model. The final performance of the model is given by the average value of ten cross-validations.

3.1. Data Preprocessing. The PET and CT data used in our experiments are provided by the Department of Radiology of the Henan Provincial People's Hospital, a governmental and public medical institution in China. In case patients explicitly requested that their data may not be shared for research purposes, the respective data samples were excluded when creating the dataset. For all data samples, the corresponding patient has a confirmed diagnosis. For the patients in our dataset, both CT and PET examinations are implemented in the same stage to ensure that the lesion's tissue morphology and metabolic levels are consistent. The datasets consist of data samples of 397 patients in total, 91 patients with SCLC, 103 patients with LUSC, and 203 patients with LUAD. Example lesion slices for three different types of lung cancer from CT and PET examinations are shown in Figure 4.

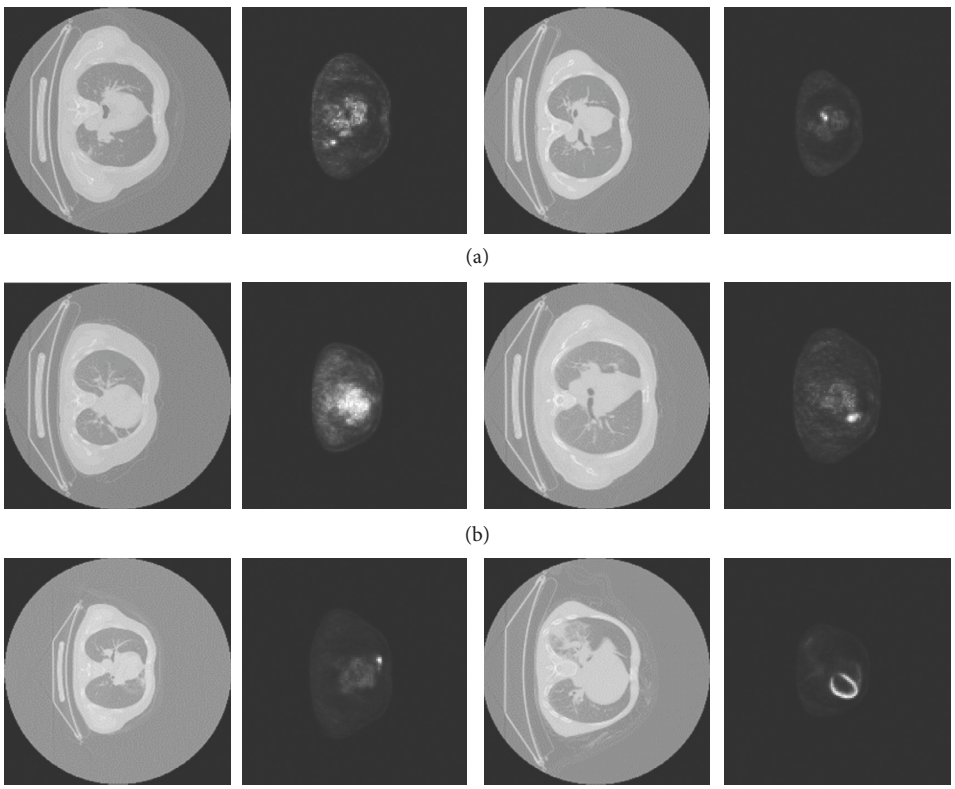
For three types of lung cancer, not only the lesion and its surrounding areas have important discriminative information but also some global information (such as location information), which is also helpful for classification in clinical diagnosis. So we use the whole image as input of CNN to preserve useful information and extract fine-grained features. For each patient's lesion, a varying number of slices (between 39 slices at maximum and 3 slices at minimum) were available in the direction of the vertical axis, which poses a variable input scale. We defined a fixed slice amount (P) for network input and provided the corresponding number of slices through sampling along the direction of the vertical axis in the 3D lesion area. PET and CT devices obtain images of different resolution: for CT images, the resolution of each slice in the direction of the vertical axis of the 3D image is 512×512 pixels; for PET images, this resolution is 256×256 pixels. We resize each slice to 112×112 pixels and normalize the range of pixel values to [0, 255]. Through this preprocessing, data samples of each modality are converted to a 3D image of size $112 \times 112 \times P$. As a small data set, we used data augmentation during the network training. Through the random combination of flipping up or flipping right, it is equivalent to expanding the data set by 4 times.

3.2. Analysis of Multimodality Data Validity. In the first experiment conducted, we use a single-modality model, a 3D DenseNet, on either a PET or a CT dataset. For the preliminary evaluation regarding the effectiveness of multimodality (both CT and PET) approach, we use the multimodality feature network named MF-DenseNet. Each parallel network of MF-DenseNet is equal to the 3D DenseNet with four dense blocks and uses the GMU as the feature fusion. The results are shown in Table 1. The variance term in the table reflects the variance of the AUC values between each round of cross-validation, and the average score term reflects the mean value of the AUC between each round of cross-validation. The balanced accuracy evaluates the balanced performance.

The experimental results show that the extraction and fusion of features from different modalities improve the performance considerably. The best average AUC score for single-modality model was reported as 0.678, which is achieved by the PET dataset. Comparing the performances achieved for CT and PET data shows that features from PET images even more facilitate for the classification. The combination of multimodal features has achieved the best performance under both AUC and accuracy verification metrics. The average AUC score of our MF-DenseNet is 0.810, and the accuracy is 0.68. Smaller variances between each round of cross-validation also show the effectiveness of multimodality data. They further demonstrate that different modalities show different feature distribution patterns, which need to be extracted. From the perspective of balanced accuracy indicators, our model also has a relatively balanced performance.

3.3. Analysis of Multidimensional Attention Mechanism Validity. To extract fine-grained features and further improve the network performance, we propose a multidimensional attention mechanism for MF-DenseNet. MFSE-DenseNet (r) consists of the MF-DenseNet with the SE-block, for which the parameter r indicates the reduction ratio of the SE-block. The MFSA-DenseNet on the other hand consists of MF-DenseNet with a spatial attention mechanism. The MFSCA-DenseNet employs both a spatial attention mechanism and a channel attention mechanism. The results are listed in Table 2, and the ROC curve of the different attention mechanisms is shown in Figure 5.

Comparing the MFSE-DenseNet ($r = 4$) and the MFSE-DenseNet ($r = 16$) with the MF-DenseNet shows that the channel attention mechanism has the ability to improve the overall performance. Although the AUC values of the model are similar between different reduction rates, it can be seen from the variance that the model has a more stable generalization performance when $r = 4$. The performance of the MFSA-DenseNet can be improved by the spatial attention, but the AUC value between classes remains unbalanced. Under these conditions, the best AUC score of 0.920 (accuracy = 0.82) was achieved using the MFSCA-DenseNet ($r = 4$). Although variance of 10-fold cross-validation is not the smallest (variance = 0.05), it is also close to variance of MFSA-DenseNet (variance = 0.04). In addition to the highest average AUC and accuracy score, this model also provides a more generalized performance through the smaller variance value of cross-validation. By analyzing the ROC curve, when the false-positive rate is reduced, that is, the misdiagnosis rate reduces, the SE module is less sensitive to LUAD than the other two types. When the reduction rate increases, the sensitivity of the model to SCLC increases simultaneously, but the sensitivity of the model to LUAD and LUSC decreases. On the contrary, the spatial attention module has a high sensitivity to LUAD when the misdiagnosis rate decreases. This also reflects the advantages of the two attention mechanisms in feature extraction for different categories. Through the combination of the two-dimensional



(a)

(b)

(c)

Figure 4: Corresponding PET and CT slices from different types of lung cancer. (a) Squamous cell carcinoma; (b) adenocarcinoma; (c) small cell lung cancer.

Table 1: Evaluation of different modality datasets.

Network architecture	Data	AUC			Balanced accuracy
		Score of each category	Average score	Variance	
3D DenseNet	CT	LUSC 0.667 LUAD 0.711 SCLC 0.485	0.621	0.08	0.52
3D DenseNet	PET	LUSC 0.722 LUAD 0.768 SCLC 0.544	0.678	0.12	0.53
MF-DenseNet	CT + PET	LUSC 0.863 LUAD 0.877 SCLC 0.690	**0.810**	**0.05**	**0.57**

Table 2: Evaluation of different attention mechanisms.

Network architecture	Data	AUC			Balanced accuracy
		Score of each category	Average score	Variance	
MFSE-DenseNet ($r = 4$)	CT + PET	LUSC 0.867 LUAD 0.876 SCLC 0.837	0.860	0.03	0.62
MFSE-DenseNet ($r = 16$)	CT + PET	LUSC 0.809 LUAD 0.774 SCLC 0.969	0.851	0.09	0.63
MFSA-DenseNet	CT + PET	LUSC 0.860 LUAD 0.961 SCLC 0.850	0.890	**0.04**	0.67
MFSCA-DenseNet ($r = 4$)	CT + PET	LUSC 0.938 LUAD 0.910 SCLC 0.913	**0.920**	0.05	**0.72**

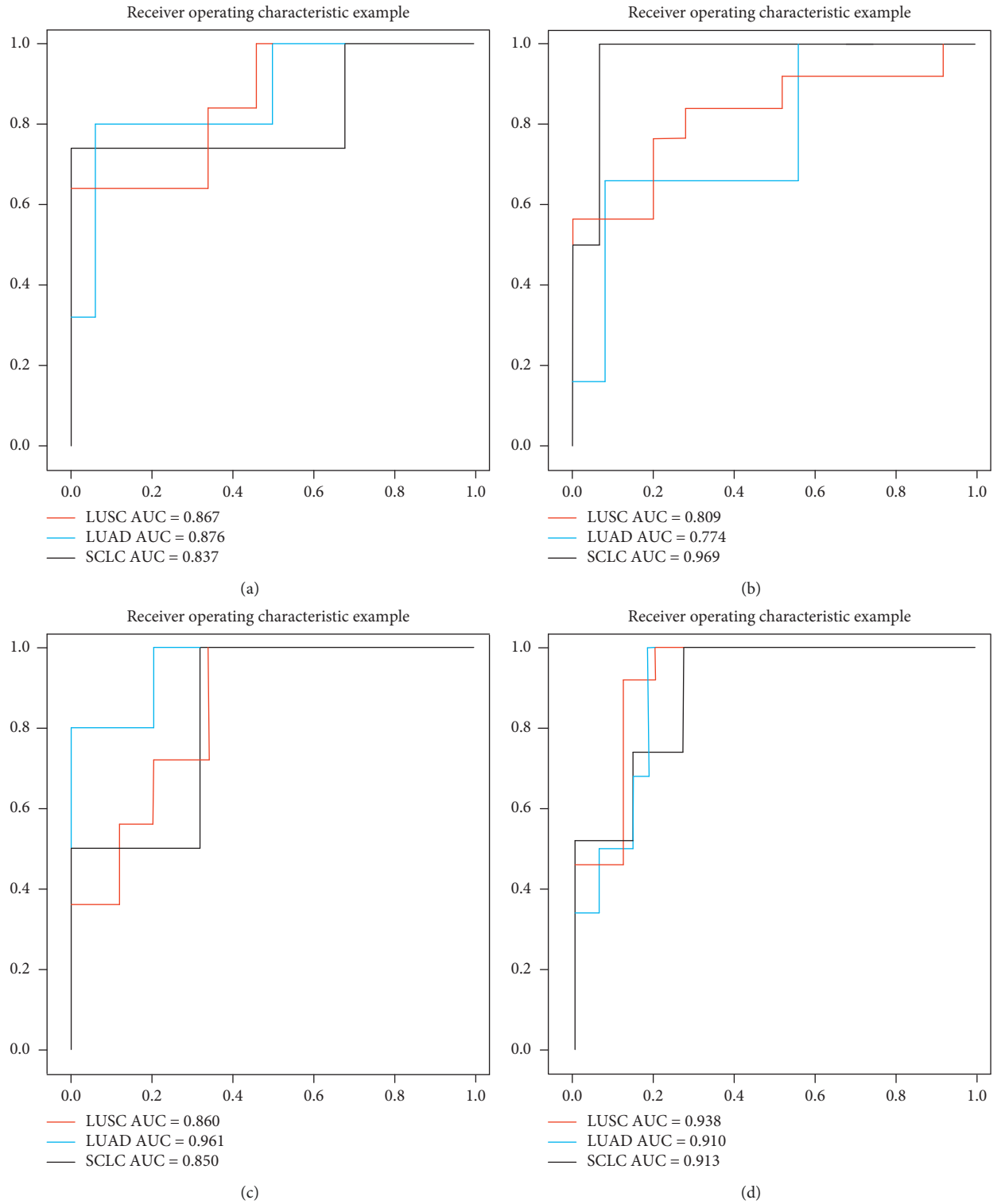


Figure 5: The ROC curve obtained with different attention mechanisms. (a) MFSE ($r = 4$); (b) MFSE ($r = 16$); (c) MFSA; (d) MFSCA ($r = 4$).

attention mechanisms, our model relieves the sensitivity difference between categories, which also balances the constraints of the fine-grained feature extraction between categories. We will conduct a more detailed discussion and analysis in the Discussion section.

3.4. Evaluating Different Feature Fusion and Loss Supervision Strategies. We compare the following feature fusion strategies: (1) optimization strategy and (2) fusion strategies. Deep-supervision strategy [31] has been introduced as an effective method for fine-grained feature extraction from a

Table 3: Evaluation of different fusion strategies.

Network architecture	Data	AUC				Balanced accuracy
		Score of each category		Average score	Variance	
MFSCA-DenseNet (GMU)	CT + PET	LUSC	0.843	0.840	0.07	0.67
		LUAD	0.899			
		SCLC	0.868			
MFSCA-DenseNet (GMU + joint optimization)	CT + PET	LUSC	0.938	**0.920**	**0.05**	**0.72**
		LUAD	0.910			
		SCLC	0.913			
MFSCA-DenseNet (connection)	CT + PET	LUSC	0.771	0.801	0.12	0.63
		LUAD	0.703			
		SCLC	0.885			
MFSCA-DenseNet (connection + joint optimization)	CT + PET	LUSC	0.788	0.815	0.10	0.65
		LUAD	0.759			
		SCLC	0.897			

single modality. The idea is to implement loss supervision on different feature outputs to achieve deeper optimization of the network. Inspired by this idea, we apply separate loss supervision on the output of each modality and integrate with the loss of final fusion feature to achieve joint optimization. The loss supervision strategy employed after the high-level semantic features of each modality enhances the effectiveness of the spatial attention mechanism. For the multimodality feature fusion, we employ GMU as the fusion strategy. The quantitative results of this experiment are listed in Table 3. In this table, joint optimization refers to the loss supervision that includes each single-path network integrated with the loss of final fusion feature to obtain the joint optimization. The connection refers to the connection operation of features to achieve the feature fusion, while the GMU models the relevant features of the two modalities. As can be seen from the results in the table, the performance obtained by GMU (variance = 0.05) has smaller fluctuations in the model prediction than the feature fusion mode of the connection operation (variance = 0.12). The result shows that both GMU and joint optimization provide the best end-to-end prediction for multimodality data.

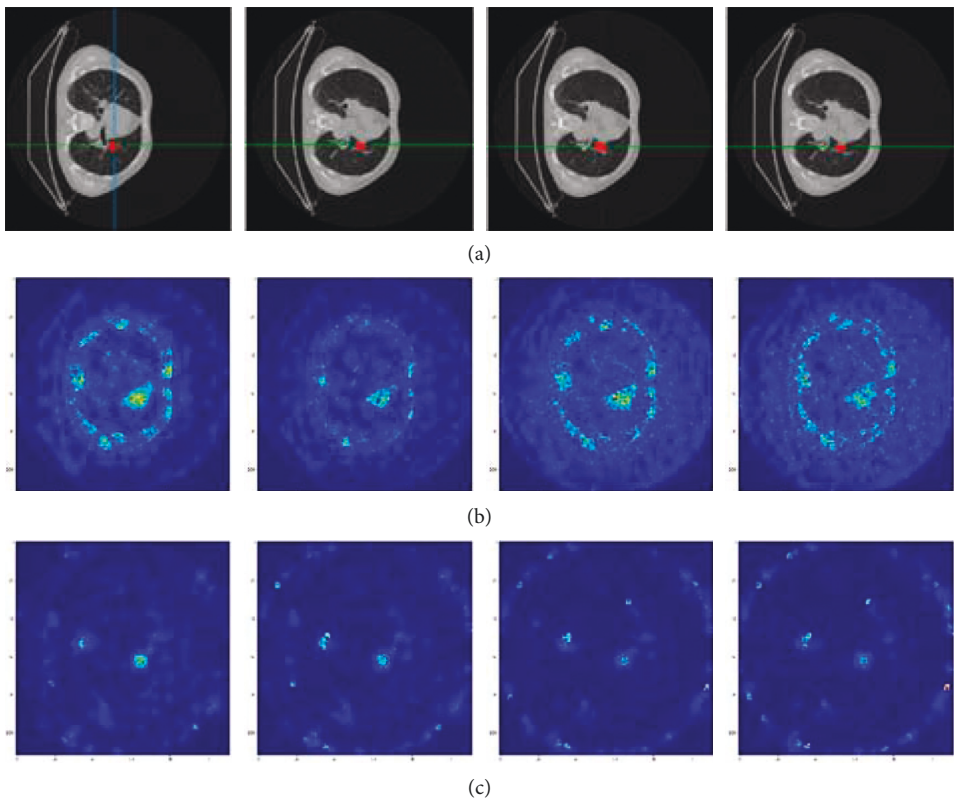
4. Visualization Experiment and Discussion

In order to verify the role our attention mechanism plays for each modality, we use Grad-CAM [33] to generate class activation mapping (CAM) of the network. More concentrated and precise CAM responses mean a higher reduction of noise in the feature extraction. Figures 6 and 7 show the CAM on each 2D slice.

We visualize the CAM in different modalities and conclude that the multidimensional attention forces the network to focus on the lesion area and to reduce feature noise in this way. Thus, the multidimensional attention mechanism accurately extracts features from the lesion area while excluding interference of the feature with the surrounding tissue. This is important especially when the entire image instead of a segmented image is used in the automatic diagnosis. To optimally use these intensively distributed features for the classification, it is necessary to ensure an accurate extraction of these characteristics from the lesion area. It can be seen from the results that our proposed attention mechanism can better concentrate the features extracted by the network in some areas.

Due to the differences in the feature distributions of various imaging modalities, we further investigate the impact that the two attention mechanisms have on the network's CAM. The results of this comparison are shown in Figures 8 and 9. Our observations show that the type of attention mechanisms has a great influence on the CAM of different modalities. For CT images, the CAM generated by a spatial attention mechanism appears similar to the CAM of models without attention mechanism. In contrast, the application of the channel attention mechanism leads to a concentrated CAM with the least amount of feature noise, but it lacks in localization accuracy and deviates from the true position of the lesion area. For PET images, the channel attention mechanism cannot focus the attention map; however, the spatial attention mechanism proved itself useful in this regard. CT images illustrate intricate structures which are represented by complex spatial features. While their identification poses a challenging task, the modelling of the feature weights on the basis of the channel dimension is more effective. In contrast, PET images, as a binarization image, hold less feature types that are difficult to distinguish based on the feature channel. However, the spatial dimension facilitates the modelling of feature weights in PET images. This experiment and the different performances we observed for PET and CT images demonstrate the complementarity of two attention mechanisms.

Through the LUAD which greatly fluctuates in sensitivity, Figure 5, we try to explain the reasons why attention mechanisms in deep learning network are effective. The metabolic level of the lesion area is measured in PET images using the standardized uptake value (SUV). This measurement plays an important role in the clinical diagnosis as, in general, LUSC and SCLC have higher SUV values than LUAD. Because the channel attention mechanism has poor ability to locate features in PET images, models based on this are not sensitive to LUAD. The spatial attention mechanism has a better feature extraction effect on the PET image, which

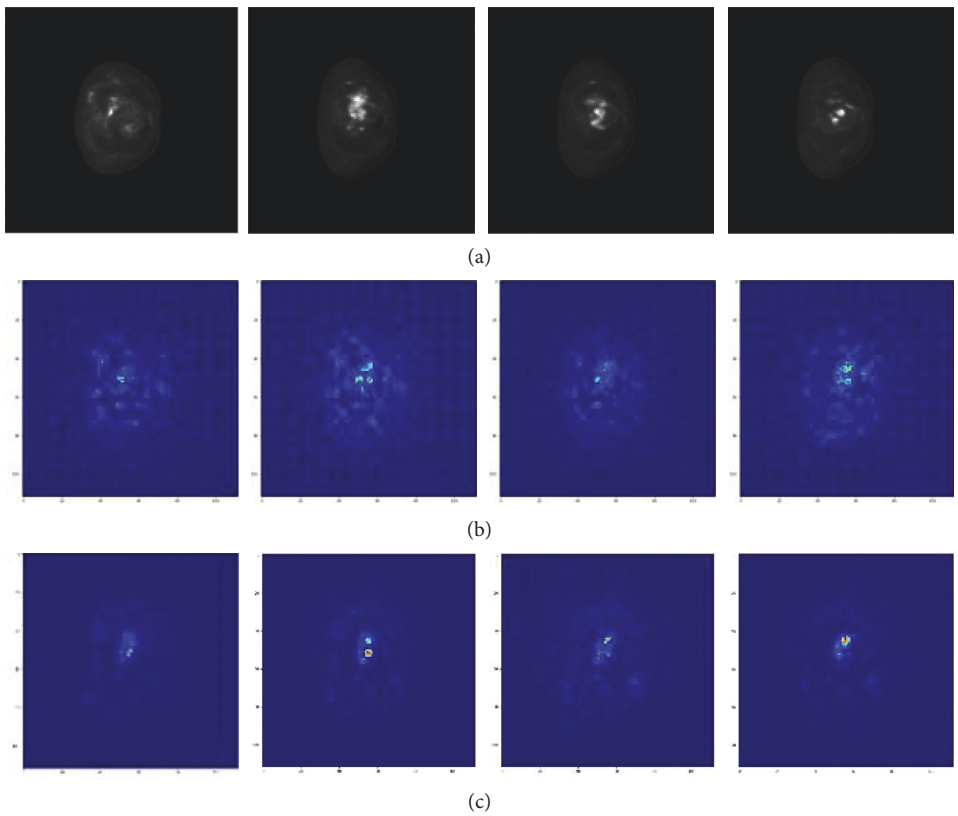


(a)

(b)

(c)

Figure 6: Grad-CAM of CT images. (a) Four slices from the original CT image—the red area highlights the lesion labeled by the radiologist; (b) the Grad-CAM generated without attention mechanism; (c) the Grad-CAM generated by the multidimensional attention mechanism.



(a)

(b)

(c)

Figure 7: Grad-CAM of PET images. (a) Four slices from the original PET image; (b) the Grad-CAM generated without attention mechanism; (c) the Grad-CAM generated by the multidimensional attention mechanism.

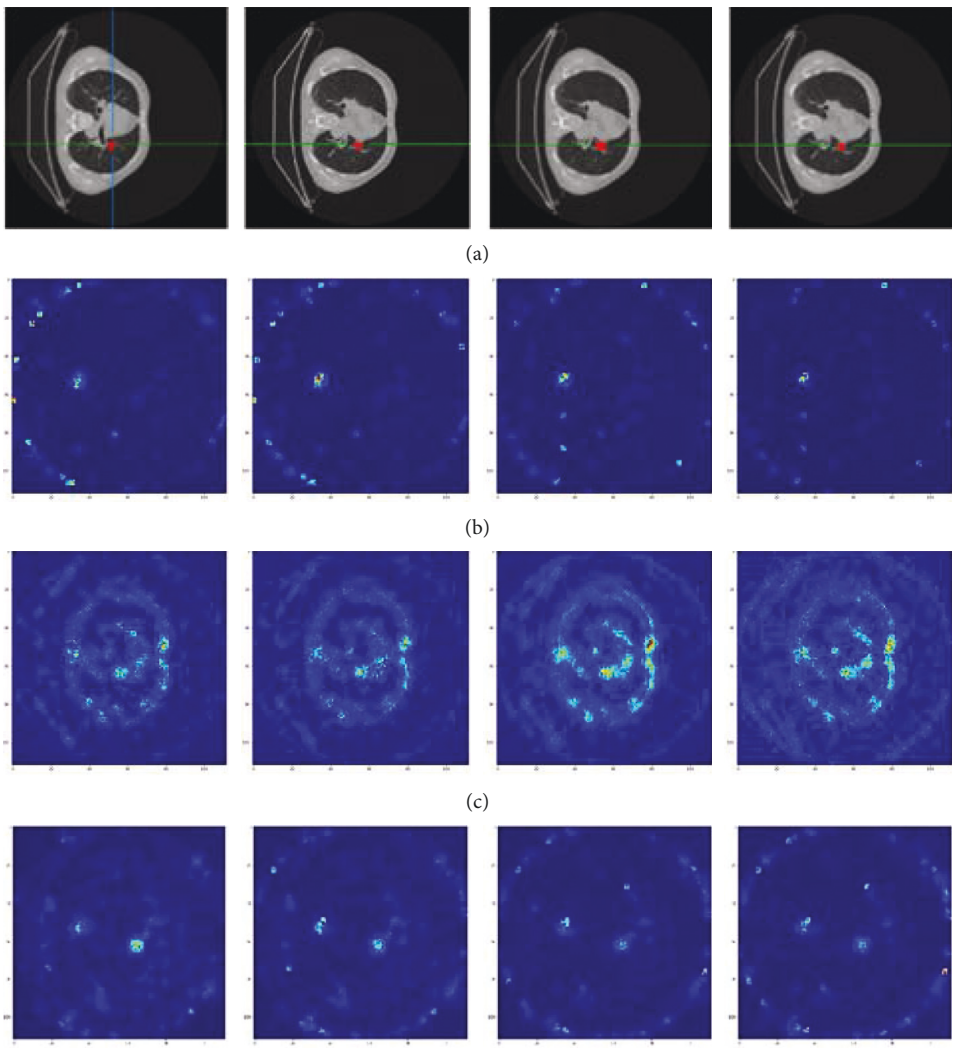


Figure 8: The Grad-CAM generated by different attention mechanisms in CT images. (a) Four slices from the original CT image; (b) Grad-CAM in MFSE; (c) Grad-CAM in MFSA; (d) Grad-CAM in MFSCA.

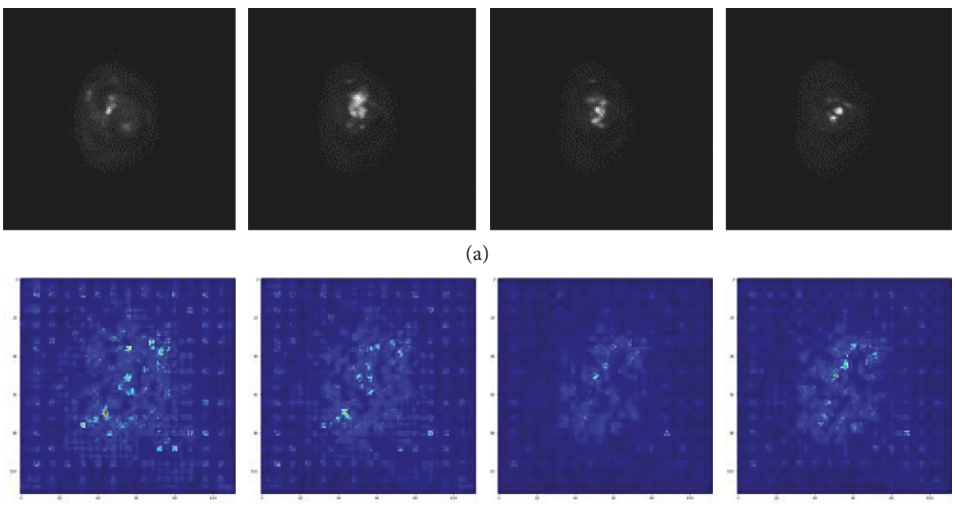


Figure 9: Continued.

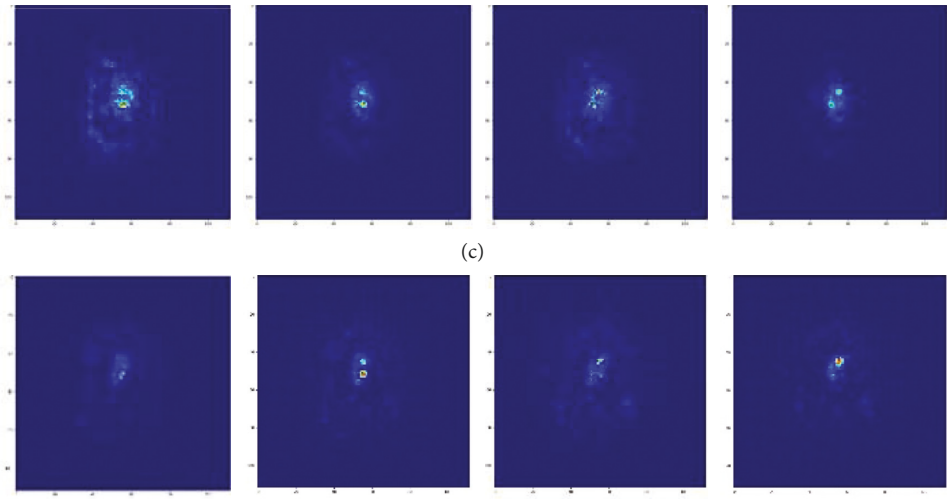


Figure 9: The Grad-CAM generated by different attention mechanisms in PET images. (a) Four slices from the original PET image; (b) Grad-CAM in MFSE; (c) Grad-CAM in MFSA; (d) Grad-CAM in MFSCA.

greatly improves the sensitivity of the model to LUAD. Compared with LUAD, SCLC lesions have low density, and there is no clear edge information in CT image. The clinical features of SCLC make it less sensitive on models based on spatial attention mechanisms. Channel attention can more effectively extract complex features on CT images and reduce feature noise, so it has better performance on SCLC. Different types of lung cancer have different characteristics on different modality images, thus demonstrating the necessity of multimodal image application. On the other hand, the complexity of feature extraction from different modalities also illustrates the necessity of multidimensional attention mechanism for different image feature extractions.

5. Conclusions

In this paper, we propose an approach for the classification of lung cancer using multimodality noninvasive clinical images (CT and PET). A parallel network for automatic lung cancer diagnosis is proposed. Furthermore, we optimize the network regarding the extraction of fine-grained features in both channel and spatial dimensions and utilize the GMU to consider the intrinsic correlation between different modalities. We consider two attention mechanisms in different modality images and visualize the results to provide a comparison between them. In future work, we will address the following topics to improve our approach furtherly: we plan to expand the dataset used in the training to achieve a clinical application level. In addition, we will collect more segmentation labels for the data in our dataset and complete the objective evaluation of our weakly supervised detection approach.

Data Availability

The part of data used in the research can be obtained from https://pan.baidu.com/s/1FBH7WZ5PoeggvcJrvX_0ug.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported by the National Key R&D Program of China under grant 2018YFC0114500.

References

[1] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, "Global cancer statistics," *CA: A Cancer Journal for Clinicians*, vol. 61, no. 2, pp. 69–90, 2013.

[2] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2018," *CA: A Cancer Journal for Clinicians*, vol. 68, no. 1, pp. 7–30, 2018.

[3] L. A. Jemal, R. L. Siegel, and A. Jemal, "Lung cancer statistics," in *Lung Cancer and Personalized Medicine*, vol. 893, pp. 1–19, Springer, Berlin, Germany, 2016.

[4] C. I. Henschke, D. I. Mccauley, D. F. Yankelevitz et al., "Early lung cancer action project: overall design and findings from baseline screening," *The Lancet*, vol. 354, no. 9173, pp. 99–105, 1999.

[5] A. K. Alzubaidi, F. B. Sideseq, A. Faeq, and M. Basil, "Computer aided diagnosis in digital pathology application: review and perspective approach in lung cancer classification," in *Proceedings of the New Trends in Information & Communications Technology Applications*, pp. 219–224, IEEE, Baghdad, Iraq, March 2017.

[6] W. Sun, B. Zheng, and Q. Wei, "Computer aided lung cancer diagnosis with deep learning algorithms," in *Proceedings of the Medical Imaging: Computer-Aided Diagnosis*, vol. 9785, p. 97850Z, San Diego, CA, USA, March 2016.

[7] K. Kuan, M. Ravaut, G. Manek et al., "Deep learning for lung cancer detection: tackling the kaggle data science bowl 2017 challenge," 2017, https://arxiv.org/abs/1705.09435.

[8] G. Litjens, C. I. Sánchez, N. Timofeeva et al., "Deep learning as a tool for increased accuracy and efficiency of

histopathological diagnosis," *Scientific Reports*, vol. 6, no. 1, Article ID 26286, 2016.

[9] J. Wu, T. Aguilera, D. Shultz et al., "Early-stage non-small cell lung cancer: quantitative imaging characteristics of 18F fluorodeoxyglucose PET/CT allow prediction of distant metastasis," *Radiology*, vol. 281, no. 1, pp. 270–278, 2016.

[10] A. Teramoto, T. Tsukamoto, Y. Kiriyama, and H. Fujita, "Automated classification of lung cancer types from cytological images using deep convolutional neural networks," *BioMed Research International*, vol. 2017, Article ID 4956063, 9 pages, 2017.

[11] V. A. A. Antonio, N. Ono, A. Saito, T. Sato, M. Altaf-Ul-Amin, and M. Kanaya, "Classification of lung adenocarcinoma transcriptome subtypes from pathological images using deep convolutional networks," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, no. 12, pp. 1905–1913, 2018.

[12] S. Lakshmanaprabu, S. N. Mohanty, K. Shankar, N. Arunkumar, and G. Ramirez, "Optimal deep learning model for classification of lung cancer on CT images," *Future Generation Computer Systems*, vol. 92, pp. 374–382, 2019.

[13] A. Teramoto, H. Fujita, O. Yamamuro, and T. Tamaki, "Automated detection of pulmonary nodules in PET/CT images: ensemble false-positive reduction using a convolutional neural network technique," *Medical Physics*, vol. 43, no. 6, pp. 2821–2827, 2016.

[14] A. Teramoto, M. Tsujimoto, T. Inoue et al., "Automated classification of pulmonary nodules through a retrospective analysis of conventional CT and two-phase PET images in patients undergoing biopsy," *Asia Oceania Journal of Nuclear Medicine Biology*, vol. 7, no. 1, pp. 29–37, 2019.

[15] S. Liang, R. Zhang, D. Liang et al., "Multimodal 3D DenseNet for IDH genotype prediction in gliomas," *Genes*, vol. 9, no. 8, p. 382, 2018.

[16] D. Nie, H. Zhang, E. Adeli, L. Liu, and D. Shen, "3D deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients," in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016*, pp. 212–220, Springer, Berlin, Germany, 2016.

[17] F. Ye, P. Jian, J. Wang, Y. Li, and H. Zha, "Glioma grading based on 3D multimodal convolutional neural network and privileged learning," in *Proceedings of the IEEE International Conference on Bioinformatics & Biomedicine*, pp. 759–763, IEEE, Kansas City, MO, USA, November 2017.

[18] H. Jie, S. Li, and S. Gang, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, Salt Lake City, UT, USA, June 2018.

[19] W. Fei, M. Jiang, Q. Chen et al., "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156–3164, Honolulu, HI, USA, July 2017.

[20] J. Schlemper, O. Oktay, C. Liang et al., "Attention-gated networks for improving ultrasound scan plane detection," 2018, https://arxiv.org/abs/1804.05338.

[21] M. Al-Shabi, B. L. Lan, W. Y. Chan, K.-H. Ng, and M. Tan, "Lung nodule classification using deep local–global networks," *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 10, pp. 1815–1819, 2019.

[22] L. Gong, S. Jiang, Z. Yang, G. Zhang, and L. Wang, "Automated pulmonary nodule detection in CT images using 3D deep squeeze-and-excitation networks," *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 11, pp. 1969–1979, 2019.

[23] C. Yan, J. Yao, R. Li, Z. Xu, and J. Huang, "Weakly supervised deep learning for thoracic disease classification and localization on chest X-rays," in *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 103–110, ACM, Washington, DC, USA, 2018.

[24] J. Shuiwang, Y. Ming, and Y. Kai, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 35, pp. 221–231, 2013.

[25] T. Jin, C. Hui, Z. Shan, and X. Wang, "Learning deep spatial lung features by 3D convolutional neural network for early cancer detection," in *Proceedings of the International Conference on Digital Image Computing: Techniques & Applications*, pp. 1–6, Sydney, Australia, November 2017.

[26] R. Dey, Z. Lu, and Y. Hong, "Diagnostic classification of lung nodules using 3D neural networks," in *Proceedings of the IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 774–778, IEEE, Washington, DC, USA, April 2018.

[27] H. Gao, L. Zhuang, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition*, pp. 4700–4708, Honolulu, HI, USA, July 2017.

[28] J. D. Fauw, J. R. Ledsam, B. Romeraparedes et al., "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature Medicine*, vol. 24, no. 9, pp. 1342–1350, 2018.

[29] W. Guo, Z. Xu, and H. Zhang, "Interstitial lung disease classification using improved DenseNet," *Multimedia Tools and Applications*, vol. 78, no. 21, pp. 30615–30626, 2019.

[30] J. Arevalo, T. Solorio, M. Montes-Y-Gómez, and F. A. González, "Gated multimodal units for information fusion," in *Proceedings of the ICLR (Workshop)*, Toulon, France, April 2017.

[31] C. Y. Lee, S. Xie, P. Gallagher et al., "Deeply-supervised nets," in *Proceedings of the Artificial Intelligence and Statistics*, pp. 562–570, San Diego, CA, USA, May 2015.

[32] X. Zhang, Y. Zou, and S. Wei, "Dilated convolution neural network with LeakyReLU for environmental sound classification," in *Proceedings of the International Conference on Digital Signal Processing*, pp. 1–5, IEEE, London, UK, August 2017.

[33] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, Seoul, South Korea, October 2019.