IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 20 April 2024, accepted 19 June 2024, date of publication 5 July 2024, date of current version 12 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3424294

RESEARCH ARTICLE

Efficient Tumor Detection and Classification Model Based on ViT in an End-to-End Architecture

NING-YUAN HUANG AND CHANG-XU LIU
College of Computer and Information Engineering (College of Artificial Intelligence), Nanjing Tech University, Nanjing 211816, China
Corresponding author: Ning-Yuan Huang (ycnj78@163.com)

ABSTRACT Accurate tumor detection and classification are crucial for cancer diagnosis and treatment. Traditional medical image analysis methods face many challenges when dealing with highly heterogeneous tumor images, such as large differences in image quality and unclear or complex tumor features. Although breakthroughs have been made in image processing with deep learning techniques, there are still limitations in identifying small or irregular tumors. Existing tumor detection models often rely on local feature extraction, neglecting global information and subtle differences in the images, which limits their accuracy and robustness in practical applications. To address these issues, this paper proposes a deep learning model that integrates Feature Pyramid Network (FPN) and Vision Transformer (ViT) within an end-to-end architecture. Firstly, the model extracts rich features at multiple scales through FPN, covering various aspects from cellular structures to tissue layouts. Then, by introducing ViT, the model can effectively process and analyze global features, particularly achieving higher accuracy in recognizing ambiguous or complex tumor patterns. The self-attention mechanism further enhances the model's focus on critical regions of the image, improving its ability to detect subtle differences. Finally, the design of the end-to-end architecture enhances the overall efficiency and consistency of the model, facilitating global optimization and further improving detection and classification performance. The experimental results show that compared to existing techniques, this model demonstrates higher recognition accuracy on medical image datasets such as TCIA, BraTS, LUNA, and Camelyon17. The accuracy and F1 scores improved by 4.65% to 6.24%. These algorithmic improvements not only enhance the efficiency and accuracy of tumor detection but also provide new pathways for the application of deep learning in medical image analysis.

INDEX TERMS Feature pyramid network, vision transformer, self-attention mechanism, tumor detection, medical image.

I. INTRODUCTION

Cancer is a serious threat to human health, and early detection and accurate diagnosis are crucial for treatment and survival rates. With the continuous advancement of medical imaging technologies such as computed tomography, magnetic resonance imaging, and nuclear medicine scans, doctors are able to obtain rich image information to aid in the diagnosis of cancer. Early detection increases the likelihood of successful treatment, reducing treatment costs and physiological and psychological burdens on patients [1]. Medical imaging not only helps physicians detect the presence of tumors but also provides detailed information about their size, shape, and location, which is essential for formulating treatment plans and selecting the most appropriate treatment methods. For example, through precise image analysis, doctors can determine whether a tumor has spread and whether surgical removal is necessary. Additionally, medical imaging provides a reliable means for monitoring disease progression and evaluating treatment effectiveness [2]. During treatment, regular imaging monitoring helps doctors assess whether tumors are shrinking or if new lesions are

The associate editor coordinating the review of this manuscript and approving it for publication was Jiachen Yang.

appearing, thereby allowing timely adjustments to treatment strategies.

Currently, there are challenges in tumor detection and classification due to differences in image quality and the complexity of tumor characteristics, directly affecting the accuracy and efficiency of diagnosis. Firstly, images produced by different imaging techniques (such as CT, MRI, ultrasound, etc.) vary in terms of clarity, contrast, and resolution [3]. Moreover, even with the same imaging technique, differences in equipment, variations in operating conditions, or the characteristics of the patients themselves (such as body position, tissue density, etc.) can lead to differences in quality. These differences make automated image analysis and tumor recognition more complex. Secondly, tumors may vary greatly in shape, size, boundaries, and density. Some tumors may exhibit atypical features, making accurate diagnosis challenging even for experienced radiologists. Furthermore, the contrast between tumors and surrounding normal tissue may not be significant, especially in the early stages, further increasing the difficulty of detection. Addressing these challenges requires the development of more advanced image processing techniques and learning algorithms. These techniques and algorithms need to adapt to differences in image quality, accurately identify and classify tumors with different characteristics [4]. Therefore, this study proposes a tumor detection model that integrates ViT model to address these issues.

Detecting medical images through machine learning has gradually become one of the mainstream methods today. U-Net, as a classic convolutional neural network, performs well in medical image segmentation tasks, with its encoder-decoder structure effectively capturing local and global features in images. However, its performance may degrade when dealing with images with blurry edges or fine structures [5]. Faster R-CNN, as an excellent model for object detection, has advantages in fast and accurate lesion detection in medical images, but it may be slower and prone to information loss when processing large-scale images [6]. By reinforcing the feature pyramid network (R-FPN) and introducing channel space mixed attention (CSMA), the detection performance of the YOLOv5 algorithm can be significantly improved, which also has applications in tumor detection. ResNet, as a deep residual network, can learn features more deeply and has achieved good results in medical image analysis, but it requires more training samples and tuning efforts to achieve optimal performance. Attention U-Net, which introduces attention mechanisms on top of U-Net, dynamically adjusts the feature weights at different positions, thus improving the performance of medical image segmentation. However, it may increase computational complexity when processing large-scale images [7]. In tumor diagnosis, deep learning methods have been able to assist radiologists in more accurate image interpretation, sometimes even detecting subtle lesions that are difficult for the human eye to perceive. For example, in breast cancer screening, lung nodule detection, brain tumor segmentation, etc., deep learning has demonstrated performance comparable to or even better than that of expert radiologists [8]. Furthermore, deep learning is playing an increasingly important role in personalized medicine, treatment planning, and disease risk assessment. Significant progress has been made in brain tumor segmentation, with the multi-scale fractal feature network (MFFN) demonstrating excellent performance in terms of accuracy, sensitivity, and specificity. It is widely used in medical image processing, but there are also problems such as difficulty in data annotation and insufficient model interpretability. In tumor detection tasks, relying solely on local features sometimes makes it difficult to accurately distinguish between benign and malignant tumors, especially in cases where tumor features are not obvious or similar to normal tissues. Additionally, existing technologies lack robustness when facing practical issues such as differences in image quality and data imbalance [9]. This means that the performance of the model may significantly decrease when encountering new images different from the training data. The paper proposes a new method that integrates FPN, ViT, and self-attention mechanism, aiming to overcome the limitations of traditional techniques, achieve higher accuracy and stronger robustness in tumor detection and classification, and adapt to the diversity and complexity of tumors. It also reduces the false positive rate.

The organizational structure of this paper is as follows: The first section introduces the importance of medical imaging in tumor diagnosis and the challenges faced by existing tumor detection technologies are introduced, and discusses the research motivation. Section two introduces the application of deep learning in medical image processing, especially the application of FPN, ViT, and self-attention mechanisms in image classification and segmentation, providing a theoretical basis for the establishment of the research framework by emphasizing the innovation points of the research. Section three details the model we propose, including the use of FPN, integration of ViT, and application of end-to-end architecture. The fourth section presents the experimental results of the model on multiple medical image datasets, compares it with six other methods, and evaluates the model's accuracy, sensitivity, specificity, and other indicators. The fifth section discusses the advantages, limitations, and possible directions for improvement of the model. The sixth section summarizes the relevant findings of the paper, emphasizes the contributions of the research, and proposes future research directions.

II. RELATED WORK

The field of medical image processing has undergone rapid development, with early computer-aided diagnosis systems emerging to assist physicians in analyzing medical images. These systems typically use simple processing techniques such as image enhancement, edge detection, and contrast adjustment to improve image readability and the visibility of information [10]. However, traditional methods face challenges in handling complex image analysis tasks, such as

accurately locating and classifying pathological features [11]. With the development of deep learning methods, the field of medical image analysis has undergone revolutionary changes. Deep learning algorithms can automatically learn complex patterns in image data, significantly improving the accuracy and efficiency of diagnosis. These technologies are not only good at identifying lesions but can also perform more complex tasks such as pathology grading and disease prediction. The DeepLab semantic segmentation model uses dilated convolution and multi-scale processing technology to improve the accuracy of segmentation by capturing detailed information and contextual relationships in the image. However, it may require longer training time and more computing resources when processing large datasets [12]. Pix2Pix is used in the field of medical imaging for tasks such as image registration, image enhancement, and image generation. Able to learn complex mapping relationships between medical images and generate high-quality medical images. However, a large amount of training data and precise parameters are required to obtain satisfactory results when processing medical images. In image segmentation, deep learning techniques play a crucial role in accurately identifying and labeling specific structures (such as organs, tumors, etc.) in medical images. This is particularly important in fields such as surgical planning and radiotherapy plan formulation. With the continuous advancement of deep learning technology, more advanced network architectures, such as residual networks [13], dense connection networks [14], and attention mechanisms [15], have been introduced into medical image analysis, further improving the performance and generalization ability of models.

In recent years, FPN, ViT, and self-attention mechanisms have played an important role in medical image processing tasks. FPN is mainly used for feature extraction and image segmentation in medical image processing. Its core advantage lies in its ability to capture image features at multiple scales, which is particularly important for processing medical images because tumors and other lesions may exhibit different features on images of different scales. For example, in a study on lung CT scans, FPN was used to effectively identify and segment lung nodules, demonstrating higher accuracy than traditional methods [16]. However, FPN also has certain limitations. Since it needs to process feature maps at multiple scales, it may lead to increased computational complexity and memory requirements. Additionally, in some cases, FPN may not be fine-grained enough when integrating features from different scales, affecting the final segmentation or classification accuracy. ViT models have brought a fresh perspective to the field of medical image analysis by introducing the Transformer architecture. Unlike traditional CNNs, ViT processes the entire image through self-attention mechanisms, thereby capturing more complex and global image features [17]. In medical image applications, such as brain MRI image analysis, ViT demonstrates better recognition capabilities for irregular shapes and blurry boundaries of lesions. Although ViT performs well in handling global information, it also faces some challenges, especially in dealing with the computational burden of large amounts or high-resolution medical images. Moreover, ViT requires relatively large amounts of training data, which may be a limiting factor in medical image tasks with limited data. The self-attention mechanism uses weighted calculations to make the model pay more attention to important areas in the image to more accurately identify and focus on the lesion area [18]. In pathological image analysis of breast cancer, self-attention mechanisms are used to highlight key pathological features, such as clusters of cancer cells, thereby improving the accuracy of classification and diagnosis. However, the self-attention mechanism also increases the complexity of the model and its sensitivity to data. This requires careful tuning of the model in order to balance performance improvements and computational efficiency [19]. It is important to note that in the process of image processing, we also need to segment medical images. Zhang X et al. proposed an improved squeeze-and-excitation residual network (SERNet), which combines the squeeze-and-excitation residual module (SERM) and the refined attention module (RAM). By modeling long-range dependencies and focusing on global features, it demonstrated superior segmentation performance. In enhancing image continuity and clarity, Lu J et al. proposed a method combining classifier-guided StyleGAN with AdaIN GAN. By introducing a conditional classifier-guided module and a linear weighted image stitching method, they overcame the limitations of style diversity and generation quality. Additionally, stereoscopic imaging technology significantly improves doctors' understanding of patients' anatomical structures in medical diagnosis and imaging. However, stereoscopic imaging faces issues of low resolution and blurry images. M Hayat et al. proposed a method combining channel and spatial attention blocks with a parallax attention module (PAM), as well as a super-resolution (SR) model specifically for endoscopic images to enhance the super-resolution of endoscopic images, thereby greatly improving the accuracy and effectiveness of medical diagnosis and surgery. The paper aims to improve the accuracy of tumor or other lesion detection by utilizing FPN for multi-scale feature extraction, ViT for global information processing, and employing self-attention mechanism to enhance the recognition of critical regions. This makes the model more suitable for clinical applications, especially in cases where tumor features are not obvious or similar to surrounding tissues, and enhances its robustness and generalization ability when facing practical issues such as differences in image quality and data imbalance.

III. METHOD

Figure 1 shows the overall algorithm for tumor image processing used in this article.

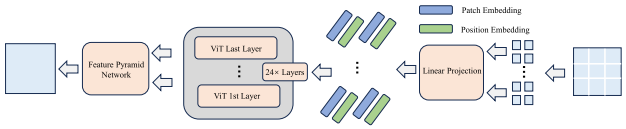

FIGURE 1. Overall Algorithm Flowchart. First, we divide the medical images into patches and apply a linear projection operation to each patch. Through Patch Embedding and Position Embedding, we extract and combine image features and positional information. Next, these features are fed into the ViT model, where the self-attention mechanism captures global features. Then, the extracted multi-scale features are processed through the FPN, generating high-resolution feature maps to improve detection and classification accuracy. This entire process is completed in an end-to-end architecture, simplifying data processing and model training, and ensuring the model's efficiency and consistency.

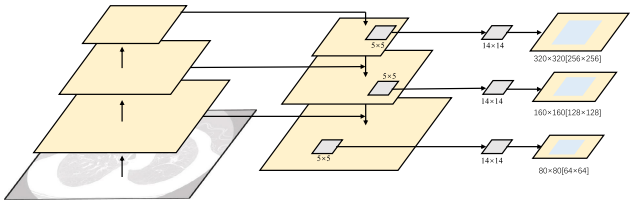

FIGURE 2. Structure diagram of FPN.

A. FPN ARCHITECTURE

FPN: a deep learning architecture used for effectively extracting image features at multiple scales. This is particularly important when dealing with medical images, as subtle anatomical structures and pathological features may only be prominent at specific scales. FPN integrates the details and overall semantic information in the image, enabling the model to obtain local features and global background at the same time, thereby improving the ability of image understanding [20]. The FPN architecture is depicted in Figure 2.

Bottom-up Pathway: In this pathway, the network extracts features from the input image through a series of convolution operations. These features are typically generated at different convolutional layers, each containing varying degrees of semantic information. Let the input image be denoted as I, and the feature maps obtained after convolution and activation functions as C_i, where i represents the level of the feature map. The output of the bottom-up pathway is a pyramid consisting of a series of feature maps, represented as $P_1, P_2, \ldots, P_n$.

$$C_i = Conv(I), \quad P_i = Conv(C_i), \quad i = 1, 2, \ldots, n \tag{1}$$

Top-down Pathway: In this pathway, feature maps are propagated from higher levels to lower levels to enhance the semantic expression capability of lower-level feature maps using semantic information from higher levels. This is achieved through upsampling operations, such as bilinear interpolation or transpose convolution. Let P_i represent the feature map at the i-th level of the bottom-up pathway, and $U(\cdot)$ denote the upsampling operation. Then, the output of the top-down pathway is denoted as F_i.

$$F_i = U(P_{i+1}) + P_i \tag{2}$$

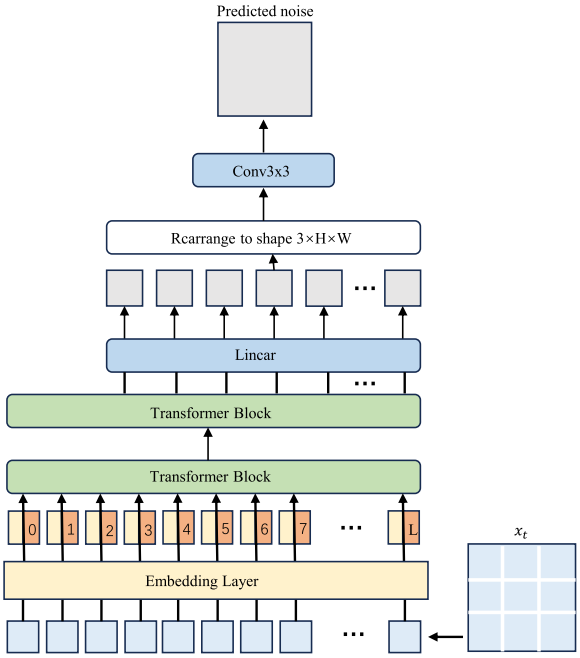

FIGURE 3. Structure diagram of ViT.

Among them, the range of i is $1 \leq i \leq n$. The top-level feature map P_n usually does not require upsampling and is output directly.

Lateral Connection: In order to fuse the feature maps of bottom-up and top-down paths, lateral connection operations need to be performed at each level to combine high-level semantic information with low-level feature maps. This is achieved through a series of convolution operations to generate the final feature map F_i' for each level:

$$F_i' = Conv(F_i) \tag{3}$$

Finally, the FPN module converts the input image into a multi-scale feature pyramid to provide rich semantic information for subsequent tasks.

B. VIT ARCHITECTURE

Vision Transformer (ViT) is a novel image processing model that employs self-attention mechanisms to handle image data. Compared to traditional convolutional neural networks (CNNs), ViT views images as sequential data and processes them through Transformer structures, enabling it to capture global information and long-range dependencies within images [21]. The architecture diagram of ViT is shown in Figure 3.

Input Representation: ViT divides the input image into a series of fixed-size image blocks, and then flattens each image block into a one-dimensional vector. Let x_i represent the i-th image block and n represent the total number of image blocks, then the input sequence $X = x_1, x_2, \ldots, x_n$.

$$x_i \in \mathbb{R}^d, \quad i = 1, 2, \ldots, n \tag{4}$$

Positional Encoding: Since Transformer does not contain convolution operations, it cannot automatically capture

positional information in the sequence. In order to introduce positional information, ViT introduces Positional Encoding to embed positional information into the input vector. Position encoding vectors are generated from sine and cosine functions:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{\omega_i}\right)$$
$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{\omega_i}\right) \quad (5)$$

Among them, $\omega_i = 10000^{2i/d_{\text{model}}}$, pos represents the position, i represents the dimension in the position encoding vector, and d_{model} represents the dimension of the input vector.

Embedding Layer: The input sequence X is added with positional encoding to obtain the embedded sequence X_{emb}. This embedded sequence contains both the image features of image patches and positional information, which serves as the input to the Transformer model.

$$X_{emb} = X + PE \quad (6)$$

Transformer Encoder: The ViT model adopts a Transformer encoder structure, which includes multiple Transformer Encoder layers. Each Transformer Encoder layer consists of multiple self-attention mechanisms and feedforward neural network layers. The self-attention mechanism is used to capture dependencies between elements in the sequence, as well as the contextual information at each position. The feedforward neural network is used to perform nonlinear transformations on the features at each position.

$$H = TransformerEncoder(X_{emb}) \quad (7)$$

where H represents the output feature representation of the ViT model.

C. SELF-ATTENTION ARCHITECTURE

This architecture analyzes and processes images by capturing the dependencies between different positions in sequential data, enabling the model to better understand the global structure and local correlations in the image, thus achieving more accurate feature extraction and task execution [22]. The algorithm architecture diagram is shown in Figure 4.

First, a weight vector w_i is calculated for each position i, which is used to weight and sum the elements at other positions in the sequence. For each position i, compute the similarity score between it and the other positions in the sequence. This can be achieved by taking the dot product of the feature vectors at each position with the feature vectors at the other positions in the sequence:

$$\text{Score}(x_i, x_j) = x_i^T x_j \quad (8)$$

where x_i^T represents the transpose of x_i, and x_j represents the feature vector at the other positions in the sequence.

Next, the scores are normalized using the softmax function to obtain the weight of each position i relative to other positions.

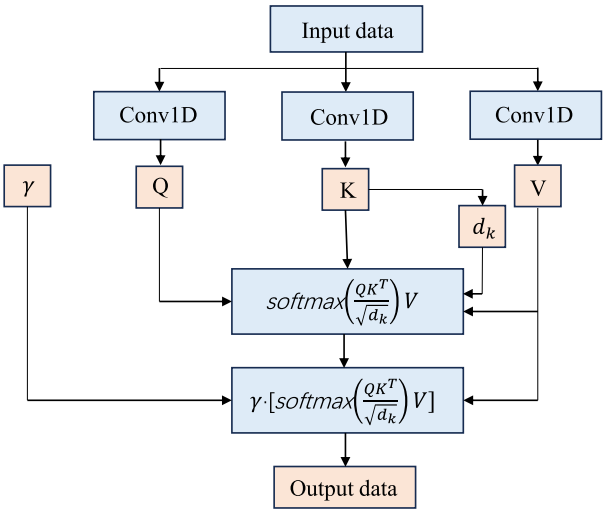

FIGURE 4. Structure diagram of self-attention mechanism.

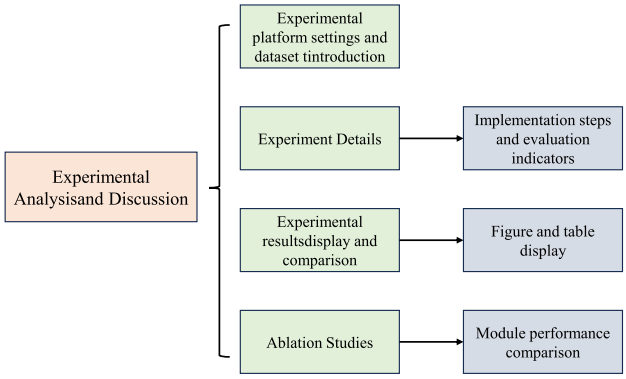

FIGURE 5. Experimental flowchart.

$$\text{Attention}(x_i, x_j) = \frac{\exp(\text{Score}(x_i, x_j))}{\sum_{k=1}^{N} \exp(\text{Score}(x_i, x_k))} \quad (9)$$

Here, $Attention(x_i, x_j)$ represents the attention weight of position i towards position j. The purpose of normalization is to ensure that the sum of all weights is 1.

Finally, the feature vector at each position in the sequence is multiplied by the corresponding attention weight at that position, and then all positions are weighted summed to obtain the self-attention vector z_i for position i:

$$z_i = \sum_{j=1}^{N} \text{Attention}(x_i, x_j) \cdot x_j \quad (10)$$

The self-attention mechanism can generate a new feature representation z_i for each position, which not only considers the position information of the position itself, but also considers the information of other positions in the sequence. This helps capture the global dependencies within the sequence.

IV. EXPERIMENT

Figure 5 shows the experimental method used in this article.

A. LAB ENVIRONMENT

This study uses a computing server to facilitate the research of image detection models. The server uses Intel Xeon E5-2690 v4 @ 2.60GHz CPU, whose powerful computing power suitable for handling the complex requirements of deep learning tasks. The server is equipped with 512GB large memory to ensure sufficient memory resources for model training and data processing. The server is also equipped with eight Nvidia Tesla P100 16GB GPUs, which accelerates the training and inference stages of the model and improves experimental efficiency.

We chose the Adam optimizer to optimize the parameters, with an initial learning rate set to 0.0001, which gradually decays in the later stages of training to improve model performance and stability. The batch size is set to 16, and the total number of epochs is 20,000, ensuring sufficient training time for the model. The model is saved every 2,000 iterations, and the best-performing model is selected for validation. We use Python as our primary programming language, combined with PyTorch as the deep learning framework for our research implementation. Python, known for its simplicity and power, offers a rich ecosystem of third-party libraries and tools, providing flexibility and convenience. PyTorch, an open-source deep learning framework, offers rich APIs and efficient computational capabilities, enabling easy construction, training, and deployment of deep learning models. Such a software environment provides a robust foundation for our research, allowing us to focus on model design and experimental exploration.

B. EXPERIMENTAL DATA

- TCIA Dataset
  The Cancer Imaging Archive (TCIA) is a publicly available medical imaging database supported by the National Cancer Institute (NCI), specifically designed for the cancer research community. The primary goal of TCIA is to advance cancer-related imaging research by providing researchers with a rich and diverse data source, including medical imaging data from thousands of cancer patients covering various types of cancer and imaging modalities. This archive includes data from 422 non-small cell lung cancer patients collected by the Tumor Radiology section, comprising 290 male and 132 female patients with an average age of 68 years. All patients underwent lung scans using Siemens Biograph scanners, with CT image slice thickness of 3mm and spatial resolution of 0.997mm. Additionally, it includes data from 211 non-small cell lung cancer patients, mainly consisting of two clinical cohorts, R01 and AMC. The R01 clinical cohort comprises 124 male and 38 female patients, while the AMC clinical cohort includes 16 male and 33 female patients. Due to the use of different scanners for lung scans, there are variations in CT image slice thickness, ranging from 0.625 to 3.0mm. The core of the dataset includes various imaging types such as CT, PET, and MRI, along with clinical information related to the imaging data. These high-quality imaging data cover all stages of cancer from early to late stages, including some rare and common cancer types. There are over 130 cancer imaging datasets covering more than 40,000 patients and over one million images. These images are provided in various formats, including DICOM, NIfTI, and JPEG. Additionally, TCIA ensures comprehensive patient privacy protection measures, respecting patient privacy when using the data.
- BraTS Dataset
  The Brain Tumor Segmentation (BraTS) challenge is an important medical imaging dataset focused on brain tumor segmentation, providing researchers with a valuable platform. The dataset primarily consists of multi-parameter magnetic resonance imaging (MRI) scans from different brain regions of various patients. These images encompass different types of brain tumors, including 369 training cases and 125 validation cases. Each sample is composed of four brain tumor MRIs: T1, T1CE, T2, and FLAIR. The volume of each modality image is $240\times240\times155$, and they are aligned to the same spatial coordinate system. The image labels include four categories: background (label 0), necrotic and non-enhancing tumor (label 1), peritumoral edema (label 2), and enhancing tumor (label 4). These labels are used for segmenting the enhancing tumor region (ET, label 4), the tumor core region (TC, labels 1 and 4), and the whole tumor region (WT, labels 1, 2, and 4). The dataset includes gliomas of different types and stages, featuring multi-grade tumors. A key characteristic of BraTS is its high-quality annotated data. The imaging data for each case is manually annotated by experts, accurately delineating tumor boundaries and different sub-regions such as the tumor core and peritumoral edema. This provides researchers with a precise benchmark for training and evaluating automatic segmentation models.
- LUNA16 Dataset
  The Lung Nodule Analysis (LUNA) dataset is designed specifically for the detection and analysis of lung nodules, aiming to promote research and development in early lung cancer detection technologies. Lung cancer often manifests as lung nodules in its early stages. LUNA16 is based on another widely used public dataset, LIDC-IDRI, which includes 1,018 low-dose lung CT images. LUNA16 removes CT images with a slice thickness greater than 3mm and lung nodules smaller than 3mm, resulting in a collection of approximately 1,000 lung CT scan samples from LIDC-IDRI. These high-resolution samples include lung nodules of various sizes and shapes, some of which are benign while others may indicate early-stage lung cancer. The original images are three-dimensional, with each image consisting of a series of axial slices of the thorax. Each sample in

the LUNA16 dataset has been meticulously reviewed and annotated by multiple radiologists, ensuring data quality and reliability. These experts provide detailed annotations regarding the presence, location, size, and other characteristics of the nodules, serving as a valuable "ground truth" benchmark for automatic detection algorithms.

- Camelyon17 Dataset
  Camelyon17, following Camelyon16, is a significant medical dataset focused on the analysis of breast cancer pathology images. It aims to challenge and advance automated techniques in pathology image analysis, particularly in the context of breast cancer. Camelyon17 provides a large collection of whole-slide images (WSI) from breast cancer patients across multiple medical centers. It includes 1,399 annotated whole-slide images of lymph nodes, encompassing both metastatic and non-metastatic lymph nodes, with a total data volume of 3TB. The data were collected from five different medical centers, covering various image appearances and staining variations. Each complete slide image is labeled to indicate whether it contains metastases (including macro-metastases, micro-metastases) or isolated tumor cells. The dataset includes detailed hand-drawn contours of metastatic lesions for 209 whole-slide images. Camelyon17 is larger in scale and involves multiple centers, providing researchers with a more challenging and diverse testing environment. This dataset features detailed annotations of lymph node metastases carefully examined and confirmed by pathology experts. These high-resolution WSIs contain rich information about tumor cells and surrounding tissue structures.

C. EVALUATION INDICATORS

- Accuracy: Accuracy is one of the most basic and intuitive indicators to evaluate the performance of a classification model. It measures the proportion of instances correctly classified by the model to the total number of instances. In the problem of tumor detection, accuracy is the proportion of cases in which the model correctly identifies tumors and non-tumors among all cases. The formula of accuracy can be expressed as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{11}$$

  where TP is the number of cases in which the presence of tumor was correctly identified. TN is the number of cases in which the model correctly identified no tumors. FP is the number of cases where the model incorrectly identifies healthy tissue as a tumor. FN is the number of cases in which the model failed to detect actual tumors.
- Sensitivity: Sensitivity is a measure of a classification model's ability to identify positive instances. In the context of tumor detection, sensitivity refers specifically to the model's ability to correctly identify the presence of a tumor. The formula for sensitivity can be expressed as:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

  In tumor detection, high-sensitivity representative models can more accurately identify patients with tumors and reduce the risk of missed diagnosis, thus improving the reliability and effectiveness of detection.
- Specificity: Specificity is a measure of the ability to identify negative instances in a classification model task. In fields such as medical diagnostics or tumor detection, specificity refers to a model's ability to correctly identify tumors or the absence of disease. The calculation formula is as follows:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{13}$$

  High specificity means that the model performs well in identifying negative cases, i.e., it does not incorrectly diagnose healthy individuals as having the disease too often.
- F1 Score: The F1 score is a statistical metric that measures the balance of accuracy and sensitivity of a classification model. This metric is especially suitable for those cases where the class distribution is unbalanced. The formula of F1 score is the harmonic mean of precision and sensitivity:

$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \tag{14}$$

  The F1 score takes Precision and Sensitivity into consideration, providing a more comprehensive and balanced performance evaluation indicator.

D. EXPERIMENTAL DATA ANALYSIS

In the experimental part, we will conduct a multi-faceted evaluation of the tumor detection and classification models. To ensure the breadth and depth of the evaluation, we have selected four different public medical imaging datasets: TCIA, BraTS, LUNA, and Camelyon17. These datasets cover a variety of cancer types and imaging modalities, providing our model with diverse testing scenarios. To evaluate the overall performance of the model across multiple dimensions, we will use several metrics: accuracy, sensitivity, specificity, and F1 score. These metrics not only assess the model's overall accuracy but also reveal its ability to identify true tumors (sensitivity), reduce misdiagnosis (specificity), and consider both accuracy and coverage comprehensively (F1 score).

As can be seen from Table 1, on the TCIA dataset, our method achieved an accuracy of 93.27%, which is significantly higher than the highest value of 91.73% among other methods. This suggests that our model is overall more accurate in identifying tumor and non-tumor cases. Furthermore, for the two indicators of sensitivity and specificity, our method achieved 95.35% and 94.42%, respectively, which were 2.89% and 2.21% higher than

TABLE 1. Comparison of indicators of various models under TCIA Dataset and BraTS dataset.

TCIA Dataset					BraTS Dataset				
Model	Accuracy	Sensitivity	Specificity	F1-Score	Model	Accuracy	Sensitivity	Specificity	F1-Score
Gupta et al.	91.13	90.88	89.80	87.84	Gupta et al.	91.73	88.46	92.18	90.22
Khairandish et al.	89.96	88.05	92.06	90.24	Khairandish et al.	90.40	92.55	88.54	89.96
Rahman et al.	87.12	92.86	92.89	87.61	Rahman et al.	91.31	91.13	91.04	91.60
Dalia et al.	86.76	90.95	89.21	90.44	Dalia et al.	89.18	89.83	90.31	90.62
Hossain et al.	88.49	91.04	91.11	90.19	Hossain et al.	91.94	92.67	88.26	91.60
Khan et al.	89.11	92.87	89.31	88.14	Khan et al.	89.05	91.06	89.14	89.60
Ours	93.27	95.35	94.42	94.97	Ours	94.65	96.64	95.68	94.21

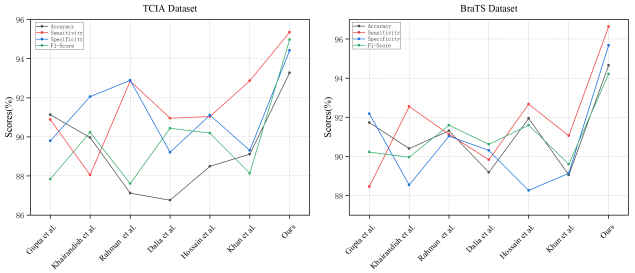

FIGURE 6. Comparative visualization of each model indicator under the TCIA dataset and the BraTS dataset.

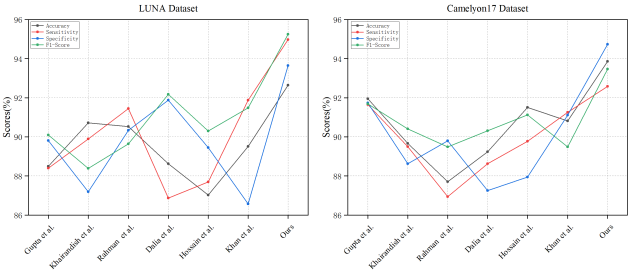

FIGURE 7. Comparative visualization of each model indicator under the LUNA dataset and camelyon17 dataset.

the highest values among other methods. This suggests that our model can more effectively detect true tumors while reducing the misdiagnosis of healthy individuals. On the BraTS dataset, our method achieved an accuracy of 94.65%, also surpassing the highest value among other methods, which was 91.94%. This indicates that our model demonstrates similar advantages on another dataset. In terms of sensitivity and specificity, our method achieved 96.64% and 95.68%, respectively, exceeding the highest values among other methods by 3.97% and 3.42%. This indicates that our model's performance in identifying true tumors and reducing misdiagnosis is also superior. Our method exhibited higher accuracy, sensitivity, specificity, and F1 score on both datasets, further validating the superiority of the model. Figure 6 shows the comparison of various indicators, which can more intuitively see the differences in the models.

From Table 2 it is obvious that our method outperforms other models on both LUNA and Camelyon17 datasets. On the LUNA dataset, the accuracy of this method is 92.64%, significantly higher than the highest value among other methods, which was 90.72%. In terms of sensitivity and specificity, our method achieved 94.97% and 93.64%, respectively, surpassing the highest values among other methods by 4.08% and 2.79%. The results show that the model can more accurately identify true tumors while reducing the likelihood of misdiagnosis. On the Camelyon17 dataset, our method achieved an accuracy of 93.86%, also surpassing the highest value among other methods, which was 91.95%. In terms of sensitivity and specificity, our method achieved 92.58% and 94.73%, respectively, exceeding the highest values among other methods by 0.93% and 2.01%.The model showed similar advantages on another dataset, better identifying tumors and reducing misdiagnoses. Figure 7 shows the comparison of various indicators, which can more intuitively see the differences in the models.

From Table 3, it can be seen that our approach demonstrates advantages in multiple indicators such as the number of parameters in the four datasets. On the TCIA dataset, our method possesses fewer parameters, totaling 336.58M, compared to other approaches. Additionally, our model exhibits higher efficiency in both inference and training times, with values of 236.57 milliseconds and 174.36 seconds, respectively. Compared to other methods, our model achieves faster inference speeds and shorter training times. Similarly, our approach demonstrates comparable advantages on the LUNA and Camelyon17 datasets. On the LUNA dataset, our model has 326.45M parameters, with inference and training times of 225.41 milliseconds and 171.57 seconds, respectively. On the Camelyon17 dataset, our model has 344.27M parameters, with inference and training times of 214.63 milliseconds and 144.35 seconds, respectively. The model performs excellently across all four datasets. Figure 8 shows the comparison of various indicators, which can more intuitively see the differences in the models.

Table 4 presents the comparative analysis of ablation experiments conducted on the TCIA and BraTS datasets. On the TCIA dataset, We evaluate the performance of the model by gradually adding different modules, including the baseline model, addition of FPN, incorporation of ViT, and simultaneous inclusion of FPN and ViT. The results demonstrate a gradual improvement in accuracy, sensitivity, and F1-score as the components are added. Specifically, after incorporating FPN and ViT, the model achieved the highest values for accuracy, sensitivity, and F1-score, reaching 93.27%, 95.35%, and 94.97%, respectively. Similar experiments conducted on the BraTS dataset yielded comparable results. With the stepwise addition of components, there

TABLE 2. Comparison of indicators of various models under LUNA Dataset and Camelyon17 dataset.

LUNA Dataset					Camelyon17 Dataset				
Model	Accuracy	Sensitivity	Specificity	F1-Score	Model	Accuracy	Sensitivity	Specificity	F1-Score
Gupta et al.	88.49	88.40	89.81	90.10	Gupta et al.	91.95	91.65	91.73	91.68
Khairandish et al.	90.72	89.89	87.19	88.39	Khairandish et al.	89.66	89.50	88.62	90.41
Rahman et al.	90.53	91.45	90.33	89.64	Rahman et al.	87.71	86.94	89.80	89.49
Dalia et al.	88.63	86.87	91.87	92.17	Dalia et al.	89.24	88.62	87.25	90.31
Hossain et al.	87.02	87.69	89.45	90.30	Hossain et al.	91.51	89.77	87.95	91.12
Khan et al.	89.51	91.87	86.57	91.48	Khan et al.	90.81	91.25	91.11	89.49
Ours	92.64	94.97	93.64	95.24	Ours	93.86	92.58	94.73	93.46

TABLE 3. Metrics of multiple models on four datasets.

TCIA Dataset				BraTS Dataset			
Model	Parameters(M)	Inference Time(ms)	Trainning Time(s)	Model	Parameters(M)	Inference Time(ms)	Trainning Time(s)
Gupta et al.	416.95	276.80	195.99	Gupta et al.	410.10	280.00	201.04
Khairandish et al.	354.59	241.86	211.50	Khairandish et al.	379.28	286.90	189.85
Rahman et al.	397.77	305.39	217.64	Rahman et al.	353.58	284.02	209.55
Dalia et al.	458.28	287.15	209.22	Dalia et al.	387.46	252.23	170.31
Hossain et al.	442.69	282.10	212.49	Hossain et al.	360.76	241.01	183.69
Khan et al.	450.18	268.01	210.57	Khan et al.	389.97	271.07	201.53
Ours	336.58	236.57	174.36	Ours	342.94	228.34	163.45
LUNA Dataset				Camelyon17 Dataset			
Model	Parameters(M)	Inference Time(ms)	Trainning Time(s)	Model	Parameters(M)	Inference Time(ms)	Trainning Time(s)
Gupta et al.	466.85	233.51	181.36	Gupta et al.	432.95	296.40	181.17
Khairandish et al.	489.38	272.52	203.62	Khairandish et al.	459.43	284.29	148.57
Rahman et al.	427.59	276.30	179.38	Rahman et al.	410.24	246.80	173.76
Dalia et al.	486.16	241.29	208.37	Dalia et al.	444.54	221.50	153.08
Hossain et al.	443.66	288.38	192.45	Hossain et al.	415.93	281.57	153.43
Khan et al.	350.92	271.84	201.59	Khan et al.	434.48	231.15	197.87
Ours	326.45	225.41	171.57	Ours	344.27	214.63	144.35

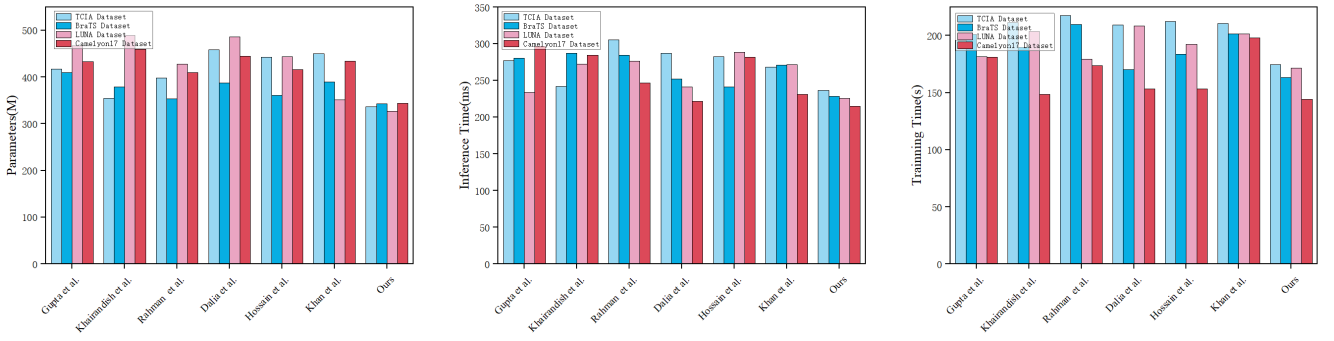

FIGURE 8. Visual comparison of indicators of multiple models on four datasets.

TABLE 4. Ablation experiments of this model on the TCIA dataset and BraTS dataset.

TCIA Dataset				BraTS Dataset			
Model	Accuracy	Sensitivity	F1-Score	Model	Accuracy	Sensitivity	F1-Score
baseline	82.64	81.87	83.54	baseline	82.28	81.59	82.37
+FPN	86.34	83.97	84.38	+FPN	84.32	85.94	84.33
+ViT	87.52	88.64	86.53	+ViT	88.69	89.21	86.43
+FPN ViT	93.27	95.35	94.97	+FPN ViT	94.65	96.64	94.21

was an improvement in accuracy, sensitivity, and F1-score. Notably, after integrating FPN and ViT, the model exhibited the best performance, with accuracy, sensitivity, and F1-score reaching 94.65%, 96.64%, and 94.21%, respectively. These experimental findings underscore the significant enhancement in model performance with the introduction of FPN and ViT. They demonstrate the superior performance achieved on both datasets by incorporating these components. In figure 9 we can intuitively see the indicator trend of the ablation experiment.

Table 5 shows the comparison of various indicators in the ablation experiments performed on the two datasets. On the LUNA dataset, we systematically added different components to assess the model's performance, including the baseline model, addition of FPN, integration of ViT, and simultaneous incorporation of both FPN and ViT. The findings indicate a progressive improvement in accuracy, sensitivity, and F1-score with the addition of these components. Specifically, after incorporating FPN and ViT, the model achieved its highest values for accuracy, sensitivity,

TABLE 5. Ablation experiments of this model on the LUNA Dataset and Camelyon17 dataset.

LUNA Dataset				Camelyon17 Dataset			
Model	Accuracy	Sensitivity	F1-Score	Model	Accuracy	Sensitivity	F1-Score
baseline	80.56	81.42	80.19	baseline	81.61	80.24	81.34
+FPN	83.72	84.76	82.64	+FPN	83.49	82.15	84.63
+ViT	86.18	87.52	85.24	+ViT	87.65	85.63	86.37
+FPN ViT	92.64	94.97	95.24	+FPN ViT	93.86	92.58	93.46

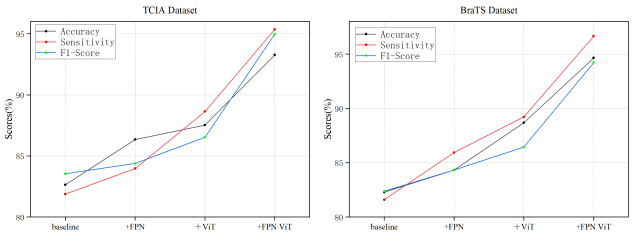

FIGURE 9. Comparative visualization of ablation experiments on TCIA Dataset and BraTS dataset.

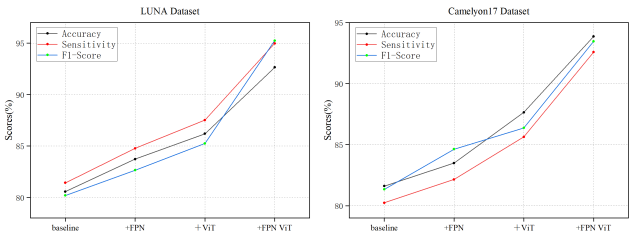

FIGURE 10. Comparative visualization of ablation experiments on LUNA dataset and Camelyon17 dataset.

and F1-score, reaching 92.64%, 94.97%, and 95.24%, respectively. Similarly, the indicators on the Camelyon17 dataset also show this trend. As we incrementally added components, there was an enhancement in accuracy, sensitivity, and F1-score. Notably, after integrating FPN and ViT, all indicators of this model reach the optimal level, with accuracy, sensitivity, and F1-score reaching 93.86%, 92.58%, and 93.46%, respectively. These experimental results underscore the significant performance improvement achieved with the introduction of FPN and ViT. They illustrate the superior performance attained on both datasets through the incorporation of these components. In figure 10 we can intuitively see the indicator trend of the ablation experiment.

V. DISCUSSION

The performance of our model was thoroughly examined across four distinct public medical imaging datasets: TCIA, BraTS, LUNA, and Camelyon17. These datasets, encompassing a variety of cancer types and imaging modalities, provided robust testing environments for our model. Compared with other methods in the field, our model shows excellent performance on multiple key metrics. For instance, on the TCIA dataset, our model achieved an outstanding accuracy of 93.27%, outperforming other methods by a significant margin. Similarly, on the BraTS dataset, the model's accuracy rose to 94.65%, showcasing its effectiveness in different scenarios. This trend was consistently observed across the LUNA and Camelyon17 datasets as well. Furthermore, the ablation studies conducted added valuable insights. These studies revealed the individual contributions of the FPN and ViT to the overall performance. By incrementally adding these components, we observed a marked improvement in model performance. This was particularly evident in the accuracy, sensitivity, and F1-score improvements seen on all datasets when both FPN and ViT were integrated into the model. This model has fewer parameters than other models, which improves the calculation speed. These technologies have shown significant advantages in improving the performance of medical image detection and classification, but they still have some limitations. The FPN excels at multi-scale feature extraction but performs less effectively when dealing with tumors of extreme size differences and has high computational complexity. The ViT, while enhancing global feature capturing capabilities, has high computational costs and relies on large-scale datasets, which may affect its performance given the limited data in the medical imaging field. Furthermore, the end-to-end architecture improves system efficiency and consistency but increases system complexity, requiring extensive parameter tuning and model adjustments to ensure stability and reliability across different application scenarios.

VI. CONCLUSION

The paper presents the development and testing of a novel tumor detection and classification model. Through FPN, our model is able to extract rich features at different scales, allowing for a more comprehensive understanding of tumor morphology and structure. This multiscale approach aids in capturing the microscopic details of tumors as well as their context within larger tissue structures, thereby enhancing detection accuracy. The introduction of ViT enables the model to leverage self-attention mechanisms to process global information. This global perspective is crucial for understanding complex patterns and relationships in the images, particularly in cases where tumor features are atypical or ambiguous. The additional self-attention layer enables the model to focus more on key regions within the image, further enhancing the sensitivity and identification capability of tumor features, particularly in noisy backgrounds. Experimental results demonstrate that the model outperforms other detection methods in terms of accuracy, sensitivity, specificity, and F1 score. It requires fewer parameters compared to other methods, and it achieves faster inference and training times. This efficiency is particularly valuable in clinical

settings, where rapid and accurate diagnosis can significantly impact patient outcomes. In future research directions, we can consider the following aspects. First, optimizing multi-scale feature extraction by introducing more complex multi-scale feature fusion techniques or combining pyramid attention mechanisms to enhance FPN's ability to extract features from very small or very large tumors. Second, in terms of data augmentation and generation techniques, using data augmentation techniques and Generative Adversarial Networks (GANs) to expand medical image datasets to improve ViT's feature learning capabilities, and exploring lightweight self-attention mechanisms to reduce computational complexity. Additionally, modular architecture design is also an important direction. By introducing a modular design in the end-to-end architecture, each part of the model can be independently optimized and adjusted, thereby reducing system complexity and enhancing flexibility and robustness. Meanwhile, cross-domain learning can utilize large-scale datasets from other fields for feature transfer, enhancing the model's generalization and feature learning capabilities to compensate for the scarcity of medical image data. Finally, in the area of real-time processing and resource optimization, researching techniques such as model compression, quantization, and acceleration algorithms can reduce computational resource consumption and improve the efficiency and feasibility of the model in practical applications.

REFERENCES

[1] C. Singh, "Medical imaging using deep learning models," *Eur. J. Eng. Technol. Res.*, vol. 6, no. 5, pp. 156–167, Aug. 2021.
[2] R. Goel, "Image processing implementation for medical images to detect and classify various diseases on the basis of MRI and ultrasound images," in *Advanced Sensing in Image Processing and IoT*. Boca Raton, FL, USA: CRC Press, 2022, pp. 277–296.
[3] P. Kora, C. P. Ooi, O. Faust, U. Raghavendra, A. Gudigar, W. Y. Chan, K. Meenakshi, K. Swaraja, P. Plawiak, and U. R. Acharya, "Transfer learning techniques for medical image analysis: A review," *Biocybernetics Biomed. Eng.*, vol. 42, no. 1, pp. 79–107, 2022.
[4] C. Kanumuri and C. R. Madhavi, "A survey: Brain tumor detection using MRI image with deep learning techniques," in *Smart and Sustainable Approaches for Optimizing Performance of Wireless Networks: Real-Time Applications*. Hoboken, NJ, USA: Wiley, 2022, pp. 125–138.
[5] H. Jiang, Z. Diao, T. Shi, Y. Zhou, F. Wang, W. Hu, X. Zhu, S. Luo, G. Tong, and Y.-D. Yao, "A review of deep learning-based multiple-lesion recognition from medical images: Classification, detection and segmentation," *Comput. Biol. Med.*, vol. 157, May 2023, Art. no. 106726.
[6] M. Liang, Q. Zhang, G. Wang, N. Xu, L. Wang, H. Liu, and C. Zhang, "Multi-scale self-attention generative adversarial network for pathology image restoration," *Vis. Comput.*, vol. 39, no. 9, pp. 4305–4321, Sep. 2023.
[7] Z. Xu, S. Liu, D. Yuan, L. Wang, J. Chen, T. Lukasiewicz, Z. Fu, and R. Zhang, "Ω-Net: Dual supervised medical image segmentation with multi-dimensional self-attention and diversely-connected multi-scale convolution," *Neurocomputing*, vol. 500, pp. 177–190, Aug. 2022.
[8] H. Li, Y. Nan, J. Del Ser, and G. Yang, "Large-kernel attention for 3D medical image segmentation," *Cognit. Comput.*, vol. 16, no. 4, pp. 2063–2077, Jul. 2024.
[9] X. Jiang, Z. Hu, S. Wang, and Y. Zhang, "Deep learning for medical image-based cancer diagnosis," *Cancers*, vol. 15, no. 14, p. 3608, Jul. 2023.
[10] C. Panigrahy, A. Seal, C. Gonzalo-Martín, P. Pathak, and A. S. Jalal, "Parameter adaptive unit-linking pulse coupled neural network based MRI–PET/SPECT image fusion," *Biomed. Signal Process. Control*, vol. 83, May 2023, Art. no. 104659.
[11] Z. Yang and S. Yang, "Multimedia image evaluation based on blockchain, visual communication design and color balance optimization," *Heliyon*, vol. 9, no. 12, Dec. 2023, Art. no. e23241.
[12] A. Saboor, J. P. Li, A. Ul Haq, U. Shehzad, S. Khan, R. M. Aotaibi, and S. A. Alajlan, "DDFC: Deep learning approach for deep feature extraction and classification of brain tumors using magnetic resonance imaging in E-healthcare system," *Sci. Rep.*, vol. 14, no. 1, p. 6425, Mar. 2024.
[13] W. Xu, Y.-L. Fu, and D. Zhu, "ResNet and its application to medical image processing: Research progress and challenges," *Comput. Methods Programs Biomed.*, vol. 240, Oct. 2023, Art. no. 107660.
[14] J. Zhang, Y. Zhang, Y. Jin, J. Xu, and X. Xu, "MDU-Net: Multi-scale densely connected U-net for biomedical image segmentation," *Health Inf. Sci. Syst.*, vol. 11, no. 1, p. 13, Mar. 2023.
[15] X. Li, M. Li, P. Yan, G. Li, Y. Jiang, H. Luo, and S. Yin, "Deep learning attention mechanism in medical image analysis: Basics and beyonds," *Int. J. Netw. Dyn. Intell.*, vol. 2, no. 1, pp. 93–116, Mar. 2023.
[16] Z. Xu, T. Li, Y. Liu, Y. Zhan, J. Chen, and T. Lukasiewicz, "PAC-Net: Multi-pathway FPN with position attention guided connections and vertex distance IoU for 3D medical image detection," *Frontiers Bioeng. Biotechnol.*, vol. 11, Feb. 2023, Art. no. 1049555.
[17] K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang, J. Zhang, and D. Shen, "Transformers in medical image analysis," *Intell. Med.*, vol. 3, no. 1, pp. 59–78, 2023.
[18] R. Azad, R. Arimond, E. K. Aghdam, A. Kazerouni, and D. Merhof, "DAE-former: Dual attention-guided efficient transformer for medical image segmentation," in *Proc. Int. Workshop PRedictive Intell. Med.* Cham, Switzerland: Springer, 2023, pp. 83–95.
[19] K. Li, Z. Qian, Y. Han, E. I.-C. Chang, B. Wei, M. Lai, J. Liao, Y. Fan, and Y. Xu, "Weakly supervised histopathology image segmentation with self-attention," *Med. Image Anal.*, vol. 86, May 2023, Art. no. 102791.
[20] Y. Chen, X. Zhu, Y. Li, Y. Wei, and L. Ye, "Enhanced semantic feature pyramid network for small object detection," *Signal Process., Image Commun.*, vol. 113, Apr. 2023, Art. no. 116919.
[21] X. Yue, S. Sun, Z. Kuang, M. Wei, P. Torr, W. Zhang, and D. Lin, "Vision transformer with progressive sampling," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 377–386.
[22] H. Yang, L. Wang, Y. Xu, and X. Liu, "CovidViT: A novel neural network with self-attention mechanism to detect COVID-19 through X-ray images," *Int. J. Mach. Learn. Cybern.*, vol. 14, no. 3, pp. 973–987, Mar. 2023.

NING-YUAN HUANG is currently pursuing the bachelor's degree with the College of Computer and Information Engineering (College of Artificial Intelligence), Nanjing Tech University, Nanjing, China. His research interests include computer vision, deep learning, and embedded development.

CHANG-XU LIU is currently pursuing the bachelor's degree with the College of Computer and Information Engineering (College of Artificial Intelligence), Nanjing Tech University, Nanjing, China. His primary research interests include deep learning, computer vision, and computer networks.